Computational Discovery of Transition-Metal Complexes: From High-throughput Screening to Machine Learning

Aditya Nandy[1,2,#], Chenru Duan[1,2,#], Michael G. Taylor[1], Fang Liu[1], Adam H. Steeves[1], and Heather J. Kulik[1,*]

[1]*Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139*

[2]*Department of Chemistry, Massachusetts Institute of Technology, Cambridge, MA 02139*

*email: hjkulik@mit.edu phone: 617-253-4584

#These authors contributed equally.

ABSTRACT: Transition-metal complexes are attractive targets for the design of catalysts and functional materials. The behavior of the metal–organic bond, while very tunable for achieving target properties, is challenging to predict and necessitates searching a wide, and complex space to identify needles in haystacks for target applications. This review will focus on the techniques that make high-throughput search of transition-metal chemical space feasible for the discovery of complexes with desirable properties. The review will cover the development, promise, and limitations of "traditional" computational chemistry (i.e., force field, semi-empirical, and density functional theory methods) as it pertains to data generation for inorganic molecular discovery. The review will also discuss the opportunities and limitations in leveraging experimental data sources. We will focus on how advances in statistical modeling, artificial intelligence, multi-objective optimization, and automation accelerate discovery of lead compounds and design rules. The overall objective of this review is to showcase how bringing together advances from diverse areas of computational chemistry and computer science have enabled the rapid uncovering of structure–property relationships in transition-metal chemistry. We aim to highlight how unique considerations in motifs of metal–organic bonding (e.g., variable spin and oxidation state, and bonding strength/nature) set them and their discovery apart from more commonly considered organic molecules. We will also highlight how uncertainty and relative data scarcity in transition-metal chemistry motivate specific developments in machine learning representations, model training, and in computational chemistry. Finally, we will conclude with an outlook of areas of opportunity for the accelerated discovery of transition-metal complexes.

Table of Contents

1. Introduction.

Transition-metal complexes have properties that can be carefully tuned through ligand design and metal selection for the control of spin/oxidation state and electronic or optical properties. As a result, tailoring the properties of transition-metal complexes has led to significant advances in catalysis and in the design of functional (e.g., magnetic) materials. Computational chemistry and, in particular, quantum chemical modeling have long played essential roles in explaining structure–property relationships and in designing this class of inorganic molecules. More recently, the accelerated discovery and design of transition-metal complexes has greatly benefitted from developments in accelerated computer hardware, larger data sets from theory and experiment, and machine learning models. Advances in artificial intelligence are changing all areas of chemistry at an increasingly rapid pace, and the combinatorial challenges that metal–organic bonding at open-shell transition-metal centers make this materials space particularly ripe for data-driven, accelerated methods.

1.1. Scope of the Review.

This review focuses on cases where computational chemistry and machine learning play a leading role in accelerated inorganic complex discovery. To understand this role, we address the intersection of developments in electronic structure theory, molecular modeling, cheminformatics, and machine learning. To describe progress in large-scale, automation of transition-metal chemical space exploration, we consider efforts in both experimental and computational high-throughput screening. We also address specific challenges for open-shell, first-row (i.e., 3d) transition-metal chemistry in terms of data availability and reliability, both

from conventional computational modeling and from experiment. We touch briefly upon heavier elements (i.e., with open 4d, 5d, or lanthanide 4f shells), but the discussion is more limited in scope, in part due to the smaller amount of available experimental data.

This review highlights recent developments over the past few years in machine learning models capable of reproducing electronic structure property predictions at low cost in high-throughput screening and accelerated discovery in transition-metal chemistry. It discusses closely related efforts in data mining, semi-empirical modeling, and molecular mechanics over the past fifty years. This discussion is necessary to provide context on where machine learning models can supersede or, alternatively, learn from these methods. Other topics in the scope of this review include data-driven efforts in extracting quantities from experimental literature or in developing linear quantitative structure–property and scaling relationships from experimental and literature data.

The central focus of this review is on methods and software for property prediction and accelerated chemical discovery in metal–organic complexes and closely related materials (e.g., porous metal–organic frameworks), at the interface with more traditional domains of electronic structure theory. Other inorganic materials, such as most solid-state materials or biological systems (e.g., metalloenzymes) are not discussed. This review also largely excludes the special considerations associated with neural network potentials and representations for carrying out molecular dynamics. Where we cover catalytic applications, our emphasis is on the design of catalysts rather than on the discovery of reaction mechanisms.

1.2. Previous Reviews.

Several previous reviews have been written that are of broad use to the reader and provide complementary discussions of topics not extensively covered in this review. Li and

Merz[1] have written a comprehensive review of molecular modeling for metal ion bonding. Pidko and coworkers[2] have extensively discussed the combined computational and experimental considerations for transition-metal catalysis. Fey and coworkers[3] have examined the relationship between chemical properties and relevant descriptors from cheminformatics for interpreting the properties of transition-metal complexes. Jensen and coworkers recently discussed some of the challenges for in silico catalyst design[4]. A number of reviews have addressed specific challenges of density functional theory in its application to transition-metal chemistry[5-7] as well as opportunities for going beyond DFT in catalysis[8]. Additional reviews are mentioned in the context of specific topics, as noted in relevant sections throughout this review.

1.3. Note on Machine Learning Nomenclature.

This review is written primarily from the perspective of computational and theoretical chemistry with some reference to concepts from chemical bonding and transition-metal chemistry as well as related experimental data. Applications of machine learning will be introduced with some needed background, but concepts behind the terminology may be less familiar to a reader without prior exposure to machine learning. To understand machine learning's relevance to transition-metal complex discovery, we will primarily discuss key aspects of statistical machine learning. We will discuss *descriptors* or *features*, which are the inputs to machine learning models. We will address *supervised learning*, where the model learns the mapping between inputs and outputs (here, property prediction). This review will also introduce applications of *unsupervised* or *semi-supervised learning*, where the model chooses the structure of data without being provided any (or most) property output labels. The review will cover models of varying complexity or interpretability and what role model choice plays in discovery. This review will discuss how researchers determine when machine learning models are

confident, in an area of statistical learning known as *uncertainty quantification*. This aids in the identification of model *domain of applicability*, which refers to where a machine learning model should be applied. The model can either be *exploited* by being used only in its domain of applicability, or we can engage in *model exploration*, preferentially acquiring promising data in regions of high uncertainty. Finally, we will address some efforts focused on addressing scarce or biased data, including through so-called active learning approaches that iteratively seek out new information for machine learning models. For additional background on the fundamental statistical learning concepts beyond the scope of this review, we suggest consulting reference texts, such as *The Elements of Statistical Learning* by Hastie, Tibshirani, and Friedman[9] or the recent primer developed by Janet and Kulik for applications of machine learning in chemistry.[10]

2. Molecular Modeling for Transition-Metal Chemistry.

There has long been significant interest in developing tractable, low-cost computational models of transition-metal complexes. The special challenges arising from the diverse range of bonding, spin, and oxidation states in open-shell transition-metal complexes have limited the development of general, accurate models. Nevertheless, semi-empirical[11-14] and classical molecular mechanics[15-20] models have been developed for understanding the properties of transition-metal complexes, particularly prior to the advent of practical density functional theory (DFT) functionals and wide applicability of DFT through increases in computational power in the 1990s. The earliest extended Huckel[21] and other semi-empirical Hamiltonian[11-14] calculations on transition-metal complexes were carried out over sixty years ago. This section describes developments in both force field (Sec. 2.1) and semi-empirical (Sec. 2.2) methods along with advancements and obstacles in the theory and practice of DFT (Sec. 2.3). Challenges for correlated wavefunction theory (WFT) methods specific to transition-metal chemistry (Sec 2.4)

are also briefly reviewed. This section concludes with a brief description of how statistical learning methods and molecular modeling methods have been simultaneously leveraged and where opportunities remain.

2.1. Force Fields.

Despite the somewhat decreased focus on such methods in recent years, affordable but reasonably accurate tools for structure generation and property prediction are often needed for efficient computational workflows. While many force fields and semi-empirical theories were primarily developed for main group chemistry, a number of semi-empirical methods[22-25] or force fields[16-17, 26-28] were developed with transition-metal chemistry in mind. The best performance is typically obtained by focusing on a subset of properties (e.g., structure[25] or spectra[22]) or materials[29-30]. Here, we briefly review some of the key historical advances in force field development with a focus on open-shell transition-metal chemistry.

2.1.1. General-Purpose Force Field Modeling.

Force fields capable of predicting properties across the periodic table have been devised[26, 31-33]. The most widely used and well known of these that is suitable for transition-metal chemistry is the universal force field (UFF)[26]. In UFF, a heuristic set of rules based on mixing of elements and on Badger's rules for relating bond length and strength enables the generation of parameters for a large number of elements. For transition-metal complexes, the ability of UFF to predict experimental metal–ligand bond lengths varies from good (0.02 Å) to fair (ca 0.08 Å) or poor, with average errors around 0.05 Å[20] to 0.10 Å[34]. For open-shell transition-metal chemistry in particular, it is problematic that a single oxidation and spin state is used in parameter generation for each metal[26], and the force field may not be accurate for alternative oxidation or spin states (Figure 1).[34-35] The appropriate bond order for metal–ligand interactions must also be

manually adjusted for ligands that involve backbonding (e.g., carbonyl or phosphine) to avoid over- or underestimating bond lengths.[20]

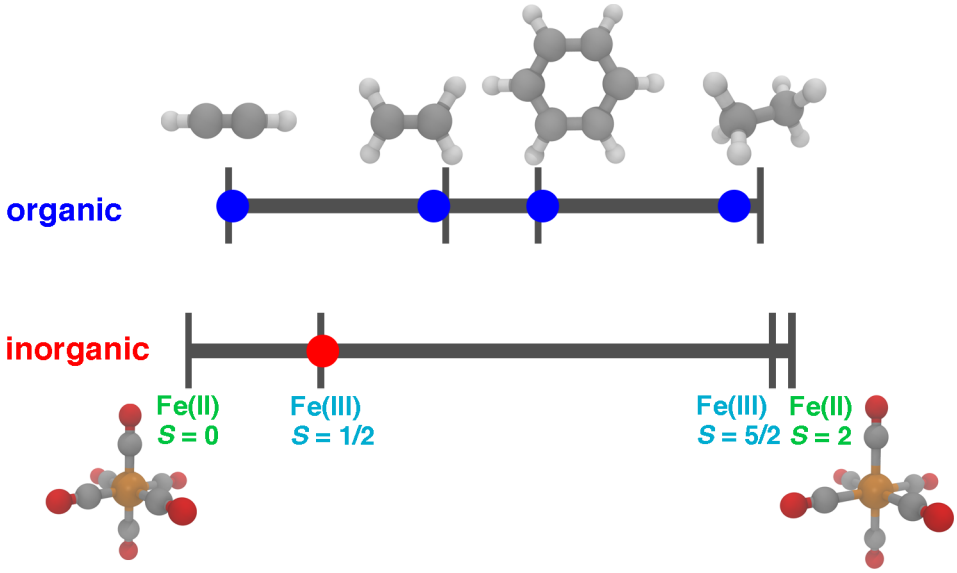

Figure 1. Comparison of relationships between bond length and chemical composition in organic molecules (top) and inorganic chemistry (bottom) obtained from UFF (circles) to those from DFT (vertical lines). For organic chemistry, the lengths of C–C bonds in acetylene, ethylene, benzene, and ethane are compared. For inorganic chemistry, the Fe–C bond in four spin and oxidation states of the homoleptic $Fe(CO)_6$ complex are compared. The relative bond lengths obtained from first-principles calculation are shown to scale across the two data sets: the organic molecules span a 0.34 Å range from acetylene (1.20 Å) to ethane (1.54 Å), and the inorganic complex bond lengths span a 0.37 Å range from singlet Fe(II) (1.94 Å) to quintet Fe(II) (2.31 Å). Reprinted with permission from ref. 35. Copyright 2020 John Wiley and Sons.

To overcome some limitations, these general force fields have been specifically tailored for common materials targets in inorganic chemistry, such as porous metal–organic framework (MOF) materials (i.e., with UFF4MOF[29]). The MM3[36] force field of Allinger and coworkers is another widely used general-purpose force field. It incorporates Morse potentials for bond stretching to better describe distorted structures. In a similar spirit to the tailoring of UFF, a number of groups have manually adjusted the MM3 parameters for a narrow range of transition-metal complexes[15] or materials (e.g., for MOFs with MOF-FF[30]). To make this approach straightforward, a number of groups have employed genetic algorithms[19, 30, 37] to optimize metal-

specific force field parameters.

It is challenging for any of these general force fields to describe the relative stability of isomers, e.g., tetrahedral versus square planar complexes. To address this limitation, Landis and coworkers developed a Fourier series approach to handling angular potentials in the SHAPES[38] and VALBOND[27, 39-40] force fields. These force fields address the fact that there are multiple minima in the L–M–L angular potential (e.g., 90 and 180° for an octahedral complex), and achieve bond length errors ca. 0.01–0.03 Å on training complexes or up to 0.1 Å for some higher-error points. Recent extensions by Meuwly[28] were added to predict other electronically driven features (i.e., trans influence, see Sec. 2.1.2). Other efforts to tailor force fields to describe aspects of metal–ligand bonding include adaptations for the Jahn–Teller effect[18, 41] and the trans effect[42]. A more thorough discussion of some of these considerations, especially for bioinorganic systems is carried out in Refs. 1, 19.

A critical outstanding area is in the generation of conformers, where even for organic molecules a force field conformer ranking often differs from that of more accurate methods.[43] In catalysis, this can be critical, as choosing a different conformer for a reaction can change the predicted activation energies[44-45]. For transition-metal chemistry, the question of how to address the metal–ligand bond (i.e., with fixed bond order or non-bonded, electrostatic terms)[46] adds additional challenges to assessing and improving force field accuracy as well.

For conformer searching and other tasks where standard force fields fail but higher-level methods remain cost-prohibitive, a divide-and-conquer approach may be useful. Errors from general molecular mechanics force fields are dramatically smaller on organic components of transition-metal complexes than on the metal–organic bond (Figure 2).[34] Thus, in one strategy, the metal and its coordinating atoms are described with QM, whereas the organic components

are treated with a force field and allowed to move during the conformer search.[47] Another family of approaches called hierarchically improved methods, in which the structure is pre-optimized with a force field or semi-empirical method (see Sec. 2.2) and then refined with DFT (see Sec. 2.3), has frequently been pursued in transition-metal chemistry (Figure 3).[48-49] More recent developments include systematic construction, e.g., from CSD fragments[50-51] (see Sec. 5) or databases of DFT data[52-53]. In combination with machine learning model predictions[54-55], these approaches have emerged as preferable alternatives (see Sec. 4).

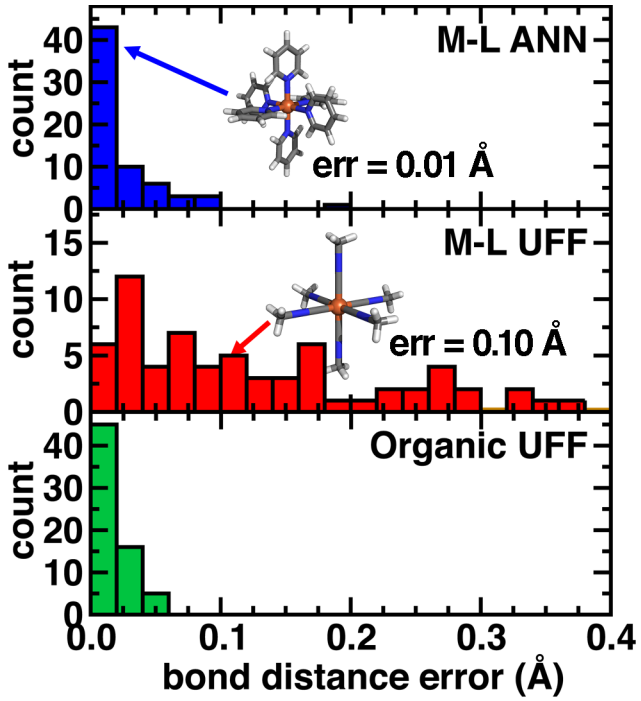

Figure 2. Comparison of average organic (bottom) and metal–ligand (M–L) bond (middle) errors (in Å) for 66 M(II/III) homoleptic octahedral complexes with M = Cr, Mn, Fe, or Co. The top pane shows the performance of an ANN trained in ref. 54. Representative compounds with median errors for M–L prediction are shown in inset: 0.01 Å abs. error for the ANN example of quintet Fe(II)$(pyr)_6$ and 0.10 Å abs. error for UFF example of singlet Fe(II)$(misc)_6$. Reprinted with permission from ref. 34. Copyright 2019 American Chemical Society.

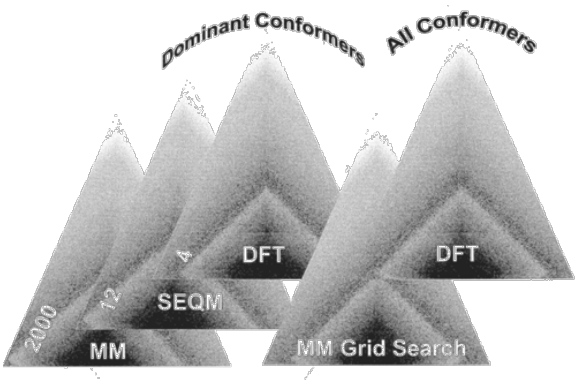

Figure 3. Modified *de novo* design pyramid suggested by Cundari and coworkers, and by Gillespie et al. wherein many MM calculations are refined with semi-empirical quantum mechanics (SQM, or SEQM in figure) followed by DFT. Further refinement with additional MM and DFT-based conformer searching has also been proposed to ensure exhaustive searches are completed. Reprinted with permission from ref. 48. Copyright 2002 American Chemical Society.

2.1.2. Ligand Field Effects and Molecular Mechanics Models.

The ligand field molecular mechanics (LFMM) approach first introduced in the 1990s[16] incorporates a distance-dependent ligand field stabilization energy (LFSE) into the force field that attempts to model the crystal field splitting. Each individual metal–ligand interaction is parameterized using standard terms (e.g., stretch, bend, torsion, vdW) along with a ligand field stabilization interaction.[17] Notably, LFMM derives its terms from the widely employed empirical Hamiltonian approach known as the angular overlap model (AOM) that was first introduced in the 1960s[56]. In the AOM approach, sparse matrix elements describing the ligand field are parameterized based on experimental structural and spectroscopic data. These methods then aid the interpretation of spectroscopic observations. Accurate structures are important inputs into the AOM model for prediction of spectroscopic properties.[57] Comba and coworkers developed the multi-level MM-AOM[57] approach for cases where an accurate experimental structure is unknown for input into the AOM model. In MM-AOM, a molecular mechanics force field parameterized[58-59] to reproduce structures of relevant compounds (e.g., Co(III) ammines) is applied to the molecules being studied. One key feature of the AOM model and the related

LFMM approach is the assumption of additivity of ligand field effects[60]. By comparing the ability of AOM models to predict experimental spectra, the qualitative assumption of ligand field additivity was confirmed[60], although minor exceptions can be noted.

In recent years, LFMM has become the most widely used implementation of AOM within MM-based modeling. The LFMM field-splitting parameters for σ and π bonding[17] are tuned to reproduce a combination of experimental structural data from the CSD and DFT-calculated distances[61], typically for a narrow range of spin and oxidation states (e.g., Ni(II) or Cu(II)[16] and Fe(II)[61]). As a result of inclusion of the LFSE term, LFMM can be parameterized to predict spin state energies[61-62], the preferred coordination geometry of a given spin or oxidation state[63], ligand exchange or dissociation energies[64], and the presence of Jahn–Teller distortions[16, 41]. Training errors on structural properties are typically small in these models for a single set of force field terms (around 0.01–0.02 Å), but application to diverse test molecules[61] can yield larger errors of ca. 0.06–0.10 Å. LFMM-like terms have also been added to polarizable force fields (i.e., in SIBFA-LF) to enable modeling of Jahn–Teller distortions.[65]

The LFMM parameters must be determined for each metal, oxidation, spin state, and ligand field strength, but the developed parameters can be used in a way that exploits ligand additivity. For example, LFMM parameters derived from homoleptic complexes can reproduce general trends of some heteroleptic compounds[62], as exemplified by tests on the spin-splitting energies of $Co(III)F_n(CN^-)_{6-n}$ complexes (Figure 4). The mixed LFMM parameters overestimate the effect of each ligand addition on the spin splitting with respect to DFT, but they are able to capture subtle differences in cis versus trans or fac versus mer isomers (Figure 4).

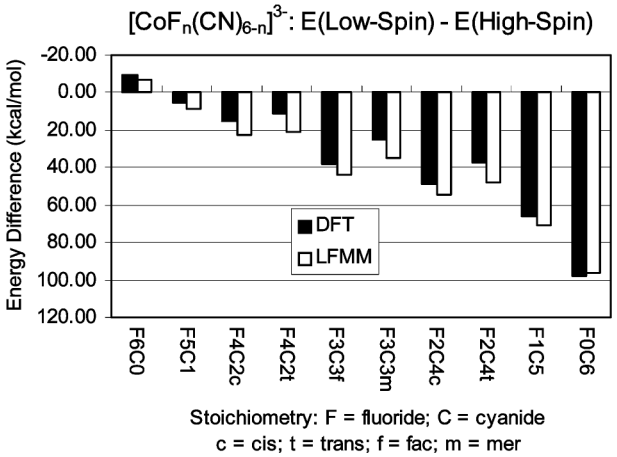

Figure 4. Comparison of LFMM-predicted spin state energetics for isomers of $[CoF_n(CN)_{6-n}]^{3-}$ with BP86 DFT calculations, both in kcal/mol. The LFMM parameters were derived from DFT on the homoleptic complexes. Reprinted with permission from ref. 62. Copyright 2003 Royal Society of Chemistry.

While implemented in a number of codes (e.g., Tinker and DommiMOE)[63, 66], widespread adoption of the LFMM approach has predominantly been limited by the lack of demonstrated general performance of the parameter set. Nevertheless, promising directions of the LFMM approach include recent improvements in automated optimization of parameters. Using multi-objective optimization with a genetic algorithm, Deeth and coworkers simultaneously minimized standard LFMM parameter spin-splitting energy errors (ca. 1–7 kcal/mol) and metal–ligand bond length errors (ca. 0.02–0.03 Å) with respect to DFT (i.e., with the OPBE functional) results for Fe(II) complexes including spin-crossover (SCO) complexes.[61] These continued methodological developments of LFMM have led to applications that include modeling host–guest interactions of water in SCO Fe(II) MOFs[67]. Using LFMM, Paesani and coworkers[67] performed molecular dynamics and Monte Carlo simulations to observe changes in the spin-crossover transition of a Fe(II) SCO MOF in the presence of additional water molecules in the unit cell that alter the ligand field environment around the Fe(II) metal (Figure 5). Other recent extensions of LFMM have been developed for metalloenzyme modeling[68].

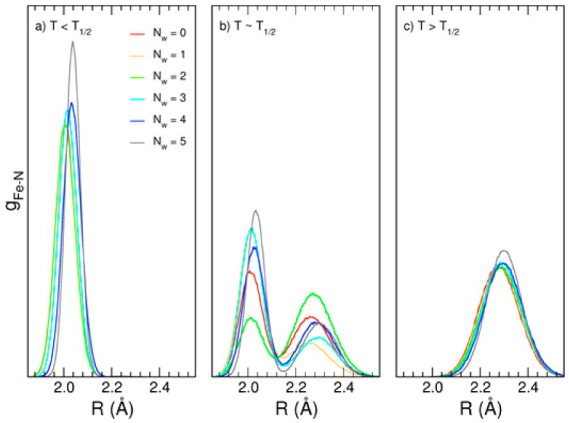

Figure 5. Distribution of bond lengths between Fe(II) and N atoms of pyrazine linkers in a spin-crossover MOF evaluated from Monte Carlo/molecular dynamics with the LFMM force field for the cases of a) below the transition temperature, $T_{1/2}$, b) close to the transition temperature, and c) above the transition temperature. Each curve corresponds to the number of water molecules, N_W, adsorbed per unit cell. Reprinted with permission from ref. 67. Copyright 2016 American Chemical Society.

2.1.3. Alternative Tailored Force Fields for Transition States.

Thus far, we have described a number of cases where specific force field parameters are developed for a narrow range of metals and materials in their equilibrium structures. Transition states play a critical role in transition-metal catalyst design, motivating a special example of force fields tailored to optimize and characterize saddle points. The transition state force field (TSFF) approach was pioneered by Houk and coworkers[69] and popularized by Norrby, Wiest, and coworkers[70-71] through combination with systematic parameterization. For the TSFF family of approaches, a force field is generally fit to the quantum mechanical (i.e., DFT) Hessian of a known transition state that is the target for ligand design.[72] The negative eigenvalue of this Hessian is shifted to become positive such that a minimization algorithm normally used for geometry optimization locates the transition state.

While parameterized on the transition state for a single ligand, this approach can then be used to screen a large number of ligands or carry out conformational searches in the TS geometry. Chief limitations here are the difficulty of accurately parameterizing the force field

and the prediction of changes in steric contributions along the reaction coordinate.[70] Norrby and coworkers have focused on identifying optimal parameterization objectives and weightings as well as systematic, automated optimization of this objective from quantum data in the Q2MM approach.[73] Despite potential limitations in the TSFF formalism including sensitivity to parameterization, these methods have been shown to successfully discern vanishingly small energetic differences between ligands that are needed to predict enantiomeric excess.[71] The predictive capabilities of Q2MM/TSFF on experimental enantioselectivities also suggests they may be able to identify errors in reported experimental values.[74] Alternative methods to the TSFF for characterizing transition states at low cost are summarized in Ref. 72.

2.2. Semi-Empirical Methods.

Over the past fifty to sixty years, semi-empirical quantum mechanical (SQM) methods have been developed and applied to the study of transition-metal complexes.[11-14, 21] As in organic chemistry where these methods are most well developed[75-77], early developments included parameterization with the complete neglect of differential overlap (CNDO)[78-79], followed by improvements through less severe approximations in the intermediate neglect of differential overlap (INDO)[23, 80-84], and neglect of differential diatomic overlap (NDDO)[76, 85] flavors of semi-empirical theory. Semi-empirical methods were most widely used following their development in the 1970s–1980s prior to the advent of predictive DFT methodology. Nevertheless, the need for rapid structural pre-optimization prior to DFT calculations, especially in data-driven workflows, as well as possible improvements arising from larger available data sets and more sophisticated parameterizations have motivated their continued development.

2.2.1. Early SQM Methods.

While SQM methods have long been demonstrated for main group chemistry, special

considerations are needed for handling d electrons. INDO, which includes one-center exchange integrals neglected in CNDO, can be expected to be the minimum level of theory capable of capturing essential aspects of transition-metal chemistry, such as d–d orbital ordering and transition characteristics.[82] The most successful demonstration of these early methods were Zerner's INDO[23] and INDO/S[22, 86] techniques for predicting ground state and spectral properties (e.g., vertical spin transition energies[87] of $[Fe(H_2O)_6]^{3+}$). These methods were applied on static structures of open-shell transition-metal complexes, typically obtained from experiment. Highly successful semi-empirical methods developed for organic molecules (e.g., the PM6 and PM7 methods[88-89] or AM1[90]) have modifications (e.g., AM1/d or PM3(tm))[91-92] to enable treatment of d electrons. Nevertheless, benchmarking of these methods on first-row transition-metal complexes has been limited and not very promising.[76, 93-94] The most glaring limitations of semi-empirical methods are the inability to predict the existence of stable structures[93-94] and the prediction of the incorrect ordering preference of conformers with respect to DFT benchmarks (Figure 6).

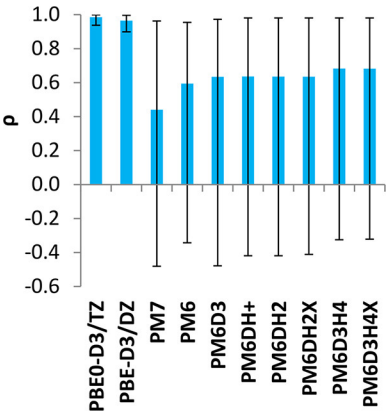

Figure 6. The Pearson correlation coefficient, ρ, between single-point energies from DFT with PBE0-D3/def2-TZVP (TZ) and PBE-D3/def2-SVP (DZ) as well as semi-empirical (PM6, PM6-D3, PM6-DH+, PM6-DH2, PM6-DH2X, PM6-D3H4, PM6-D3H4X, and PM7) methods in comparison to a M06/def2-TZVP reference for transition-metal complexes studied in ref. 93. The solid blue bars correspond to average values over 27 transition-metal complexes with 3d, 4d, and 5d metals studied in ref. 93. The black bars correspond to the lowest and highest values, with negative values corresponding to an anticorrelation between the method and the reference, which

is observed for at least some compounds for all semi-empirical methods. Reprinted with permission from ref. 93. Copyright 2018 American Chemical Society.

Other observed errors for SQM methods include incorrect metal–ligand bond lengths (by ca. 0.1 Å[94]), incorrect ground state spin prediction[95], and both over- and underestimation of ligand dissociation energies[76]. When tested on MOF crystal lattice optimizations, PM6 and PM7 generally underestimate lattice parameters by around 0.5–1.0 Å (ca. 3%) even for relatively simple (e.g., Mg-MOF-74) materials.[96] In comparison to 3d metals and main group elements, few parameter sets have been developed for 4d or 5d transition metals.[83, 97-98] Poor performance for metal–ligand bond lengths (ca. 0.1 Å on 48 training compounds) for an INDO-based model was attributed to the inflexibility of parameters derived to reproduce atomic ionization energies.[83]

Unlike in organic chemistry, errors for semi-empirical methods in transition-metal chemistry are overall unsuitably large.[76, 93] Parameterizations of these methods typically involve fitting to experimental atomic data as well as to metal–ligand bond distances from experimental structures of mononuclear complexes.[99] However, the lack of explicit encoding of the effect of ligand properties (i.e., field strength or contributions beyond the direct coordinating atoms) has limited the predictive capabilities of these methods. The application and performance of semi-empirical theory to polynuclear transition-metal complex properties (e.g., exchange coupling) has been even less extensively explored.[100] For example, Reiher and coworkers[76] showed that PM6[88] or PM7[89] ligand dissociation energies not only over- or underestimate the dissociation energies from local coupled cluster theory references by as much as 150 kJ/mol (40 kcal/mol) but also that modest changes from PM6 to PM7 parameterization significantly increase the errors (ca. 10 kcal/mol or more) for some compounds (Figure 7). The current parameter sets are

unlikely to be sufficiently flexible to eliminate these errors. Detailed, critical reviews of developments and remaining limitations for semi-empirical theory are provided in Refs. 76-77, 99.

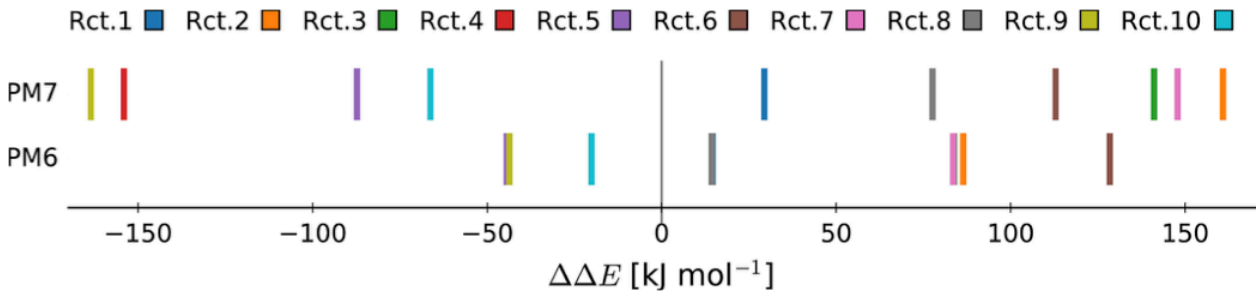

Figure 7. Deviation of electronic ligand dissociation energies ΔE in kJ/mol between either PM6 or PM7 and DLPNO-CCSD(T) for the 10 reactions in the WCCR10 set. Reaction 4 is not shown for PM6 due to the error being high. Reprinted with permission from ref. 76. Copyright 2018 John Wiley and Sons.

The best performance has been obtained when the semi-empirical parameters for the metal center were optimized systematically (i.e., with a genetic algorithm) on experimental (i.e., CSD structure) data for a specific metal (e.g., Tc[101]) and/or application (e.g., Ru dyes[102]) to achieve errors on the order of 0.02–0.04 Å in metal–ligand bond lengths. For example, by both focusing on a narrow range of materials and re-optimizing the Ru parameters in PM6, metal–ligand bond length errors were reduced from 0.04 Å to less than 0.02 Å on molecules both in and held back from the training set[102].

2.2.2. Recent Developments in Semi-Empirical Theory.

Hierarchical strategies proposed by Cundari, Gillespie and coworkers[48-49] that were critical in the early 1990s, due to the demanding nature of DFT at the time, remain relevant for modern uses of semi-empirical methods for geometry per-optimization in high-throughput workflows. These authors proposed the *de novo* structure prediction approach of MM-then-SQM-then-DFT, occasionally requiring some retracing of steps in the hierarchy to ensure all

minima were located in conformational searches (Figure 3). However, the large geometric errors typically observed with early semi-empirical methods present a barrier to their use in modern workflows. Improvements in the treatment of non-covalent interactions and thermochemistry, e.g., with the orthogonalization-corrected OMx methods of Thiel and coworkers[103], have not been extended to transition-metal chemistry. Readily available semi-empirical methods (e.g., PM6[88] or PM7[89]) sometimes optimize structures to the incorrect coordination environment favored by a metal, but constraining the local coordination environment does not generally improve prediction accuracy.[93]

These challenges have motivated the development by Grimme and coworkers of the GFN-xTB series of semi-empirical methods better suited to the prediction of structures[25, 104] of large transition-metal complexes. The GFN-xTB family generally outperforms PM6 and PM7 variants (Figure 8). However, these newer SQM methods have some limitations, in that they cannot predict spin state energetics[75], and they generally perform poorly for predicting properties of open-shell transition-metal complexes such as pK_as[105] or redox potentials[106]. For use in characterizing transition states, the GFN-xTB methods slightly outperform PM6 or PM7 on prediction of barrier heights and structures with respect to DFT reference values on the MOBH35 set[107] (Figure 9). Nevertheless, GFN-xTB cannot locate some transition states and others have significantly different geometries (ca. 0.5–1 Å RMSD) or energies (ca. 10–30 kcal/mol) in comparison to DFT reference values (Figure 9).[107]

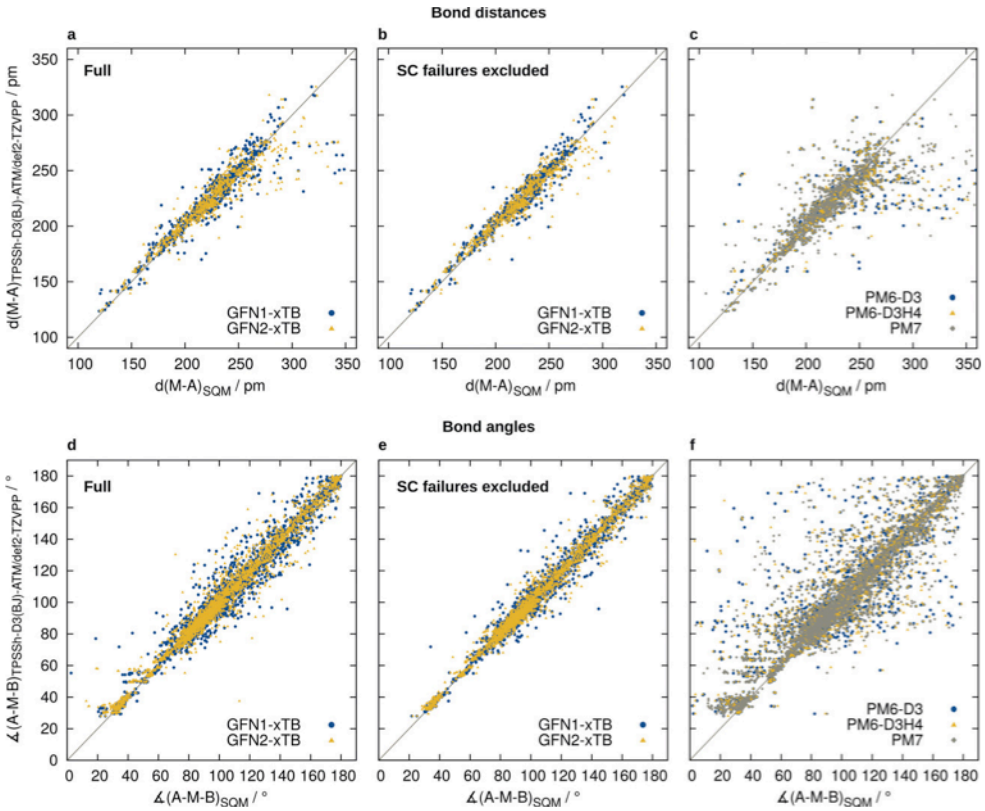

Figure 8. Performance of semi-empirical methods (i.e., GFN1-xTB, GFN2-xTB, PM6-D3, PM6-D3H4, and PM7) in terms of bond lengths and angles on a benchmark set of closed-shell singlet mononuclear transition-metal complexes (i.e., TMG145) optimized with TPSSh-D3(BJ)/def2-TZVPP. Bond lengths in pm (a)–(c) and angles in ° (d)–(f). Parts (b) and (e) show the results obtained upon exclusion of structurally incorrectly optimized structures (i.e., labeled "SC failures excluded"). Reprinted with permission from ref. 104. Copyright 2019 John Wiley and Sons.

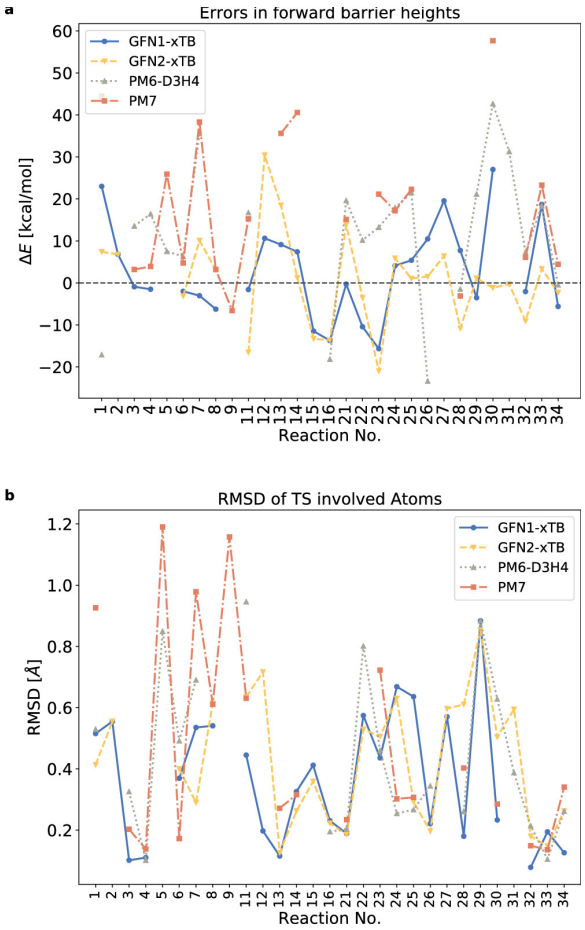

Figure 9. Deviations of semi-empirical methods GFN1-xTB, GFN2-xTB, PM6-D3H4, and PM7 from a DFT reference (TPSS-D3(BJ)/def2-SVP) for the MOBH35 transition-metal complex reaction set a) barrier heights in kcal/mol and b) structural RMSDs in Å for transition states. Reprinted with permission from ref. 107. Copyright 2020 American Chemical Society.

The self-consistent charge (SCC) density functional tight binding (DFTB) method represents a distinct family of semi-empirical techniques based on reproducing results from DFT in the context of a minimal basis.[24, 108-109] Because the method is largely parameterized to reproduce results from semi-local DFT, it can suffer from the same challenges as DFT in describing transition-metal complexes (see Sec. 2.3). Nevertheless, parameters have been developed for some transition metals with DFTB (e.g., Ni[110], Cu[111], and earlier 3d transition metals[112]). Reasonably promising results have been obtained especially for structural properties of transition-metal complexes (ca. 0.06–0.10 Å errors), but energetic errors are still significant

(ca. 10–20 kcal/mol or higher) especially for charged ligands.[110, 112] More parameters, including those that can account for multiple spin and oxidation states in a balanced fashion are needed to make DFTB a generally useful tool.

2.3. Density Functional Theory Modeling.

While the theoretical groundwork for density functional theory (DFT) was first developed over 50 years ago, development of accurate functionals and increases in computing power in the 1990s led to its rapid adoption for application to transition-metal chemistry. Much like for the MM and SQM theories that preceded it, special considerations must be made in the application of DFT to transition metals. In this section, we briefly review the most relevant theoretical challenges (Sec. 2.3.1) and practical performance (Sec 2.3.2) of conventional DFT functionals in this domain. We also describe the most widely employed semi-empirical corrections that have been developed specifically for transition-metal chemistry (Sec. 2.3.3).

2.3.1. Theoretical Challenges for Conventional DFT Functionals

Although DFT is widely employed for its balance of cost and accuracy in main group chemistry, most exchange-correlation (xc) approximations in DFT suffer from one- and many-electron self-interaction errors.[113-117] The self-interaction error has been associated with a number of missing constraints in DFT, including the lack of piecewise linearity or derivative discontinuity with respect to electron removal, which leads to underestimations of fundamental gaps and incorrect hybridization.[118-120] As reviewed elsewhere (e.g., in Refs. 6, 121-122), a popular approach to improving DFT accuracy in main group methods is to systematically incorporate higher-order terms in an expansion of the density, which have been described as rungs on a Jacob's ladder to chemical accuracy heaven, while satisfying constraints on model systems or against experimental data. This paradigm is challenged in open-shell transition-metal

chemistry. The partially filled d shell is strongly sensitive to both absolute self-interaction error, also referred to as delocalization error[113, 123], as well as imbalances in delocalization across a reaction coordinate.

Exchange-correlation functionals that reduce delocalization error tend to increase the static correlation error[124-126] (i.e., poor treatment of multi-determinantal solutions), important for transition-metal chemistry, with few exceptions[126-130]. Recommendations of best-in-class functionals for transition-metal chemistry differ from the main group: pure semi-local generalized gradient approximations (GGAs) tend to be favored[131] in transition-metal chemistry when static correlation error is manifest, whereas hybrid GGAs and "higher rung" functionals are typically believed to improve over GGAs for organic molecules. Similarly, DFT errors are often significantly larger in transition-metal chemistry, and practical correction schemes tend to differ from those used in main group chemistry.

2.3.2. Practical Performance on Structures and Energetic Properties.

Some properties are disproportionately sensitive to self-interaction errors in DFT functionals. For example, adiabatic spin state splitting[132-139], barrier heights[140-141], frontier orbital energies and absorption spectra[142-144] are all very sensitive to functional choice. For complexes that are known to exhibit near-degenerate spin states (i.e., spin-crossover complexes or SCOs), incorporating a significant amount of HF exchange usually biases functionals toward overpredicting high-spin ground states, whereas semi-local functionals predict too many complexes to have low-spin ground states.[132-139] Isomerization or conformer energies[145-146] as well as ligand dissociation energies[147] represent an intermediate case for sensitivity to self-interaction error, generally increasing when there is an imbalance in the degree of delocalization between intermediates.[148] Most observations have been made on 3d complexes, and

comparatively less is known about the relative accuracy of electronic structure methods for 4d complexes[149-151]. However, a combination of the strong low-spin shift of 4d metals with respect to their 3d counterparts and the more diffuse nature of orbitals suggests that some properties that are normally quite sensitive to functional choice (i.e., spin-splitting energies) in 3d metals are considerably less sensitive with 4d metals.[152]

Experimental gas-phase ligand dissociation energies from primarily closed-shell, late transition-metal (Cu, Pd, Ru, Pt, and Ag) complexes were collected by Reiher and coworkers in the WCCR10 dataset[147]. The WCCR10 set was designed to isolate potential effects of a condensed-phase environment on ligand dissociation energies. The errors obtained with DFT functionals on this set are moderate (ca. 20–40 kJ/mol or 5–10 kcal/mol) [147], outside the reported error margin of the experiments. Reiher and coworkers[153] also evaluated these ligand dissociation energies with more accurate DLPNO-CCSD(T)[154] reference energies and used diagnostics of multi-reference character (MR) to estimate when the DLPNO-CCSD(T) results should be most trustworthy. They found that discrepancies between DLPNO-CCSD(T) and experiment persisted in these cases, suggesting some uncertainty in the experimental interpretation.[153] In choosing a DFT functional that reproduced DLPNO-CCSD(T), PBE0 with a D3 dispersion correction was the best performer[153], albeit not agreeing particularly well with all experimental results[147]. Recent tests on 151 transition-metal complex and small-molecule reaction energies known as TMC151[155], which is a superset of previously proposed test sets[156-158], have also favored dispersion-corrected functionals but in combination with range-separated hybrids.

Structures obtained from geometry optimizations (i.e., local minima) are generally less sensitive to DFT functional than many other properties. Exceptions to this tend to occur only for

isomers with small energetic differences that are sensitive to delocalization imbalances and can lead to method disagreement[145, 159]. A representative example of where delocalization errors from DFT can influence structure is in the inserted dioxide structure of transition metal dioxides, which progresses from bent to linear experimentally across the first-row transition metals. Semi-local GGAs predict all structures to be bent[146, 159] in contrast with CCSD(T) or experiment, whereas delocalization corrections (i.e., DFT+U[159] or hybrids[146]) tend to improve structures (Figure 10).

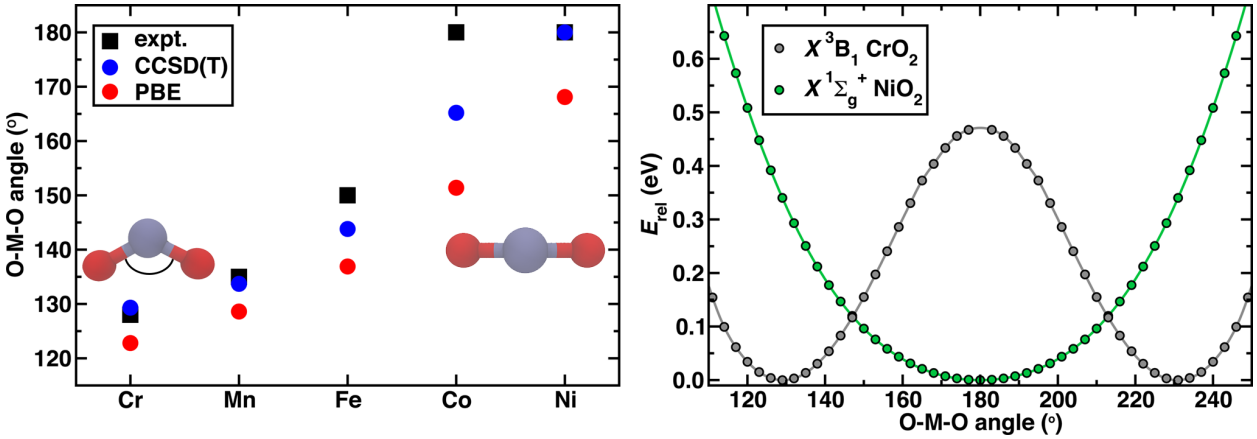

Figure 10. (Left) Comparison of favored angles (in °) for the inserted dioxide MO_2 (M = Cr, Mn, Fe, Co, or Ni) molecules from experiment (black squares), CCSD(T) (blue circles), and PBE (red circles). (Right) Overlay of the potential energy curves (in eV) from CCSD(T) for a bent $X\ ^3B_1\ CrO_2$ (gray circles) and linear $X\ ^1\Sigma_g^+\ NiO_2$ (green circles). Reprinted with permission from ref. 146. Copyright 2020 American Chemical Society.

The most abundant source of geometric data is structures obtained from X-ray diffraction of molecular complexes in the crystal field of surrounding ions and complexes (e.g., that are deposited in the CSD[160]). The condensed-phase, crystalline environment could be expected to be distinct from the *in vacuo* calculations used to benchmark DFT. Thus, the effect of an implicit solvent on DFT structures has been directly probed[161] and found to be influential particularly for complexes with a mixture of ionic ligands. As an alternative, Bühl and coworkers[131, 150, 162] collected structures from gas-phase neutron diffraction for which more direct comparisons could

be made. The data sets for these 3d[162], 4d[150], and 5d[131] transition-metal complexes were comparatively modest in size (ca. 20–30 complexes). Across all transition-metal complexes, most functionals overestimated bond lengths, but errors were modest for all cases, and several pure GGA functionals (e.g., PBE or BP86) performed well.[131] This divergence suggests that DFT should provide suitable structures even when energetics cannot be predicted, whereas force fields or semi-empirical methods[93] may not be sufficient for structural predictions.

2.3.3. Semi-Empirical Corrections to DFT

A number of semi-empirical, albeit physically motivated, corrections have been developed and applied as additive corrections for deficiencies in DFT functionals in a manner that is not dissimilar from semi-empirical wavefunction theories. While several, such as empirical dispersion corrections, are widely used across the periodic table, we focus here on those that have been developed especially for open-shell transition metals. Corrections are often motivated here by the need to address self-interaction errors to which open-shell (e.g., 3d) electrons tend to be disproportionately sensitive. The so-called LDA+U or DFT+U approach[163-165], which is reviewed in Refs. 6, 166, is one such method. While introduced for the solid state[163-165], DFT+U was later demonstrated for open-shell transition-metal molecules[6, 142, 167]. DFT+U incorporates a Hubbard U correction term to localized (e.g., 3d) states. This correction can be calculated from the electronic structure[167-170] but is more frequently tuned[171-173] to reproduce target properties. Due to this association, we classify DFT+U as a semi-empirical correction, despite having desirable fundamental origins in the Hubbard model and demonstrating the ability to recover[169] exact conditions (e.g., the derivative discontinuity[118], Figure 11). The limits of DFT+U in practice are that while it can improve some properties (e.g., band gaps or spin state ordering), it may do so at the cost of worsening others (e.g., geometries) due to the limited

flexibility of a single parameter applied to the 3d states.

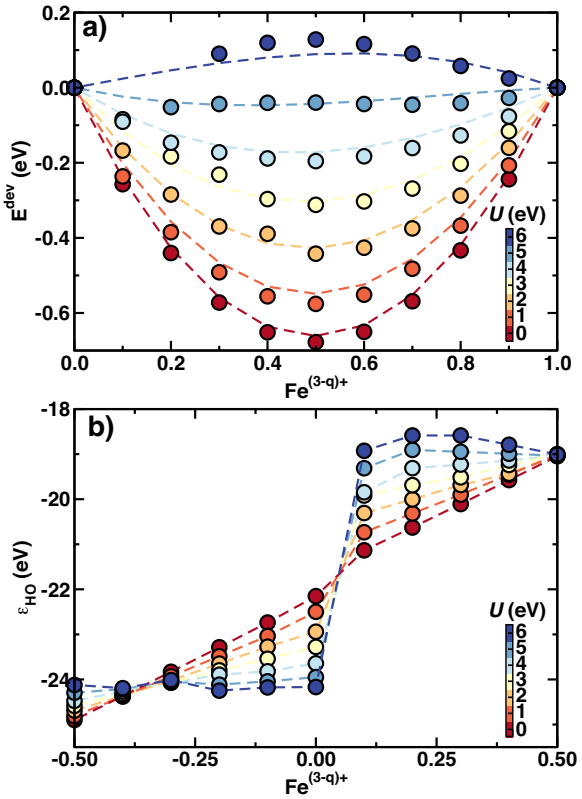

Figure 11. a) Comparison of explicit (circle symbols) and interpolated (dashed lines) deviation from linearity (E^{dev}) for low-spin $[Fe(CO)_6]^{(3-q)+}$ for q = 0 to 1. b) Highest occupied eigenvalue (ε_{HO}) with charge centered around $Fe(CO)_6^{3+}$. The lines for both graphs are colored by the value of U applied following the color bar shown in inset (i.e., red is a pure GGA result and blue is a U = 6 eV result). Reprinted with permission from ref. 169. Copyright 2016 American Institute of Physics.

A promising, holistic empirical approach is Friesner and coworkers' DBLOC method[174-176]. Force-field-like energy corrections, which depend on the metal and electron configuration (i.e., spin and oxidation state) as well as the ligand field strength, are added to DFT (typically B3LYP) energetics. The DBLOC method has been demonstrated to correct hybrid B3LYP DFT spin-splitting[177] or redox properties[178] (Figure 12). The approach's empiricism comes in part from the least-squares minimization of errors to obtain parameters[177] that increase correspondence between hybrid DFT properties and curated experimental training data (e.g., 57 spin-splitting energies[177] or 95 redox potentials[178]). The total parameter set (ca. 5–12 per

metal/ligand combination[177-178]) can approach the size of the available training set. Given the limited training data, bondwise corrections due to ligand field strength are thus assumed to be additive[60] and also classified only in terms of a weak, intermediate, or strong ligand field.[176-177]

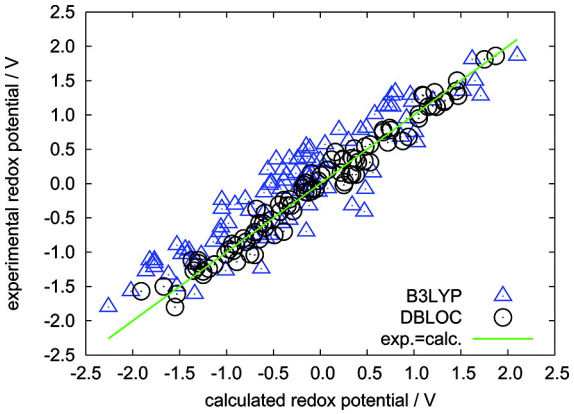

Figure 12. Experimental redox potentials (in V) versus calculated redox potentials for B3LYP (blue triangles) and after DBLOC correction (black circles) for 95 redox couples. Reprinted with permission from ref. 178. Copyright 2012 American Chemical Society.

The DBLOC approach reduces absolute errors on both training and test[177] molecules as well as the standard deviation of errors with respect to curated experimental data sets, for example, reducing B3LYP MAEs on redox potentials from 0.4 V to 0.12 V[178] (Figure 12). In addition to spin splitting or redox, the method has also been demonstrated on ligand dissociation energies[179], overall typically reducing errors from moderate values (ca. 4[179] to 9[178] kcal/mol) to near-chemical accuracy (ca. 1–3 kcal/mol). It has been noted that DBLOC can more systematically reduce errors for few-parameter functionals in comparison to more complex DFT functionals.[180] Nevertheless, the method has not reached widespread use due in part to the difficulty with obtaining parameters or training data[176].

A common systematic, albeit empirical approach, is to optimize DFT functional parameters for specific DFT properties, such as redox potential[181] or spin-splitting energy[182]. In these approaches, the optimal exchange fraction is typically sought by minimizing error with

respect to experiment, leading to distinct values (e.g., 12% in B4XLYP[181] or 15% in B3LYP*[182]) from the parent functional and decreasing errors (e.g., ca. 0.05–0.06 V in redox potential [181]) on test molecules. While tailoring the HF exchange fraction[181-182] has been demonstrated to work well when optimizing for a single class of materials and properties, e.g., spin-splitting energies of Fe(II)/N spin-crossover complexes[182], this approach can fail across broader ranges of materials[133, 177]. The optimal exchange fraction in a hybrid GGA or meta-GGA is strongly metal- or ligand-dependent in spin state ordering (Figure 13).[135-136, 138-139, 182-186] Differences in exchange sensitivity for ligand dissociation energies or steps in a catalytic cycle create additional complications for typically studied properties of transition-metal complexes.[148, 187] One type of physically motivated tuning (i.e., like DFT+U) is the use of optimally tuned range-separated hybrids[188-190], wherein the range-separation parameter for incorporating HF exchange is used to eliminate deviations from piecewise linearity for each specific molecule with some improvement noted for transition-metal complex properties.[125, 191-193]

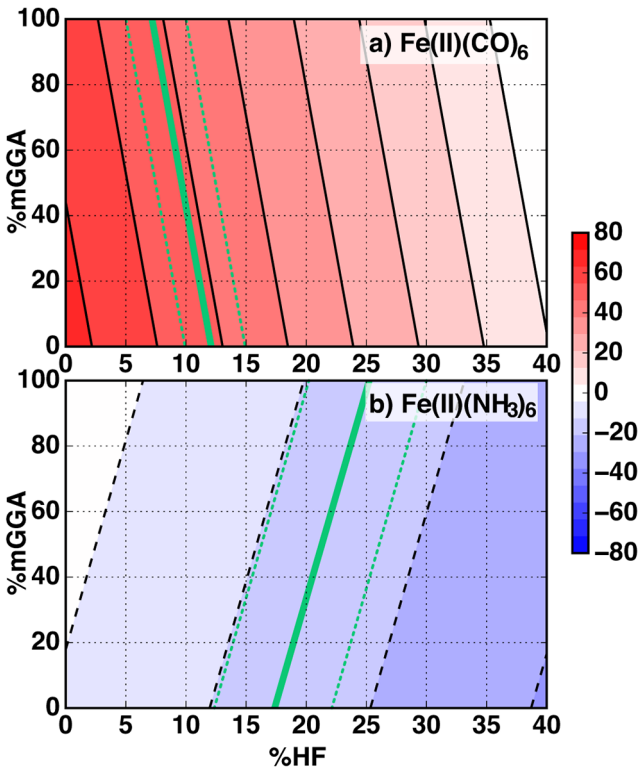

Figure 13. Spin-splitting energies (in kcal/mol) as a function of % HF exchange (%HF, *x*-axis) and % meta-GGA exchange (%mGGA, *y*-axis) in a PBE/TPSS (meta)-GGA with hybrid HF exchange for Fe(II)$(CO)_6$ (top) and Fe(II)$(NH_3)_6$ (bottom). The meta-GGA exchange from TPSS has the opposite effect on the hexa-carbonyl complex to that of the weaker-field hexa-ammine complex. The CASPT2 value is indicated with a green solid line along with a ±3 kcal/mol confidence interval (green dotted lines) in both cases. The plane is colored according to the colorbar shown in inset, and dashed or solid lines correspond to increments of 10 kcal/mol. Reprinted with permission from ref. 139. Copyright 2017 American Chemical Society.

An alternative approach is to leverage error cancellation by quantifying relative or referenced[194-195] energetics. Batista and coworkers[194] demonstrated reduction of redox potential errors to within the margin of error on experimental redox potential values (ca. 64 mV) by calculating redox potentials relative to the value of a couple in the same row of the periodic table and under the same solvent conditions (Figure 14). These results provided a slight improvement over earlier suggestions of using the well-characterized one-electron Fc/Fc^+ couple[195]. Interestingly, it was more straightforward to reduce errors from modest basis sets (i.e., 6-311G*) than larger ones[195]. Similar observations have been made in referenced (i.e., isodesmic) pK_a

predictions involving transition-metal complexes[196] and in heterogeneous catalysis where modified reference adsorbate energies (e.g., O atom from H_2O - H_2) have been suggested to be less sensitive to the computational method[197].

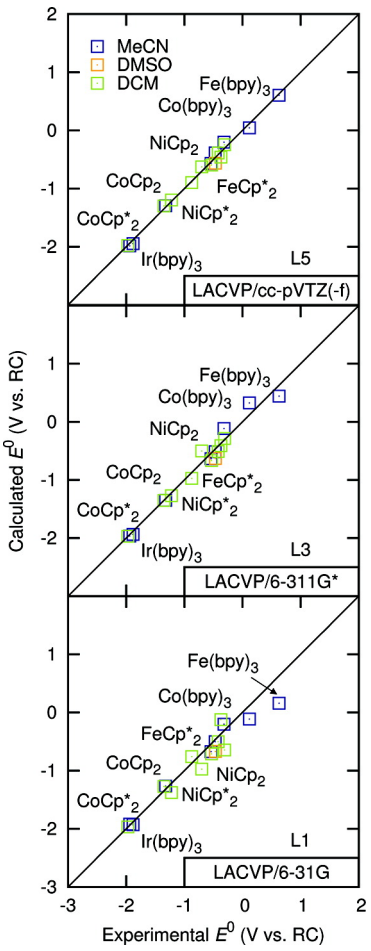

Figure 14. Correlation between measured E_{exp}^0 and calculated E_{Ln}^0 redox potentials in V vs RC, where RC = $[FeCp_2]^{0/+}$, $[Ru(bpy)_3]^{2+/3+}$, and $[Ir(acac)_3]^{0/+}$ for first-, second-, and third-row transition-metal complexes, respectively, using n = 1 (i.e., L1 = B3LYP/6-31G), n = 3 (i.e., L3 = B3LYP/6-311G*), and n = 5 (i.e., L5 = B3LYP/cc-PVTZ) levels of theory. Reprinted with permission from ref. 194. Copyright 2012 American Chemical Society.

2.4. Moving Beyond DFT in Transition-Metal Chemistry.

As computing power increases and reduced-scaling algorithms have been developed[154, 198-202], correlated wavefunction theory (WFT) methods have become increasingly tractable. In return for the high formal scaling (i.e., > O(N^5)), single-reference (SR) WFT methods such as

coupled cluster theory are widely regarded as being able to achieve chemical accuracy (i.e., predictions to within 1 kcal/mol of experiment) in the realm of organic chemistry. However, the challenges for DFT posed by transition-metal chemistry are not easily overcome by carrying out SR WFT. Both the larger size of transition-metal complexes (100s–1000s of electrons and 10s–100s of atoms) and the nature of electron correlation in open d and f shells play a role in this distinction. The SR methods that work well for organic chemistry are less predictive in open-shell transition-metal chemistry.[203] This has been attributed to the greater contribution of static correlation in open-shell transition-metal complexes as well as a greater sensitivity to the degree of recovery of dynamic correlation.[204-205] When two spin states or a reactant and intermediate are compared, even a slight difference in the ability of a method to recover the dynamic correlation from a single Slater determinant can lead to an outsized effect on relative energetics.

Among the outstanding challenges for WFT in transition-metal chemistry are that the local approximations[154, 202] to single-reference CCSD(T) that make it tractable on larger transition-metal complexes have a strong influence on spin state predictions[205] as does the choice of orbitals or determinants.[206] These observations have led some to question[203] or alternatively defend[207] the proposition that coupled cluster methods can outperform DFT functionals for transition-metal chemistry in terms of agreement with experiment. These comparisons are complicated by the potentially greater uncertainty associated with experimental reference data for transition-metal chemistry.

While multireference (MR) WFT methods appear to be promising, most such methods remain intractable for high-throughput screening due to cost and typical requirements of manual tuning of parameters and active spaces. Alternatives to overcome limitations imposed by manual active space selection include rule-based[208], automated[209-210], and machine learning (ML) model-

based[211] approaches. Automated active space selection in transition-metal chemistry, e.g., with measures of orbital entanglement[212-214] often reveals a large number of participating orbitals for systems with significant covalency (e.g., CrF_6 in Figure 15).[209] Lower-scaling, multireference perturbation theory (i.e., complete active space second-order perturbation theory or CASPT2) can be applied to moderately sized transition-metal complexes. CASPT2 spin state ordering is sensitive to both the zeroth-order Hamiltonian shift[215] with predictions alternately biased[186] towards low-spin or high-spin states depending on the choice of shift and the treatment of semi-core states[216] (Figure 16). Recent work has motivated a composite CASPT2/CC[217] approach for open-shell transition-metal chemistry.

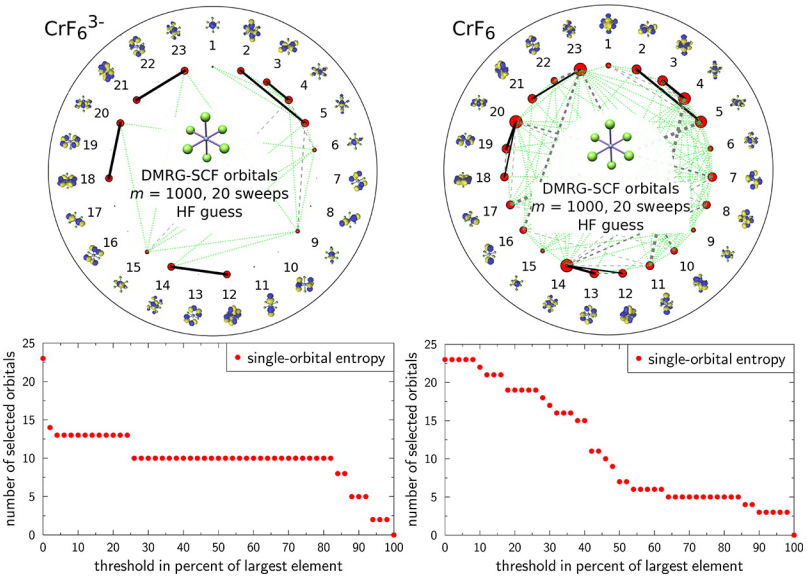

Figure 15. Entanglement analysis for active space selection for all 23 DMRG(N,23)[1000]-SCF valence orbitals of CrF_6^{3-} (N = 39) and CrF_6 (N = 36) (left and right, respectively). (top) Diagram of orbital-entropy-selected active spaces from DMRG calculations described in ref. 209. All orbitals are numbered and arranged on a circle. The area of the red circles assigned to each numbered orbital is proportional to the single-orbital entropy of the respective orbital. The line connecting two orbitals denotes their mutual information value. The lines in black indicate a value of at least 0.1 for the mutual information, whereas dashed gray and green lines represent mutual information values of at least 0.01 and 0.001, respectively. (bottom) Single-orbital entropy threshold diagrams for the determination of the number of orbitals to be included in the active space. Reprinted with permission from ref. 209. Copyright 2016 American Chemical Society.

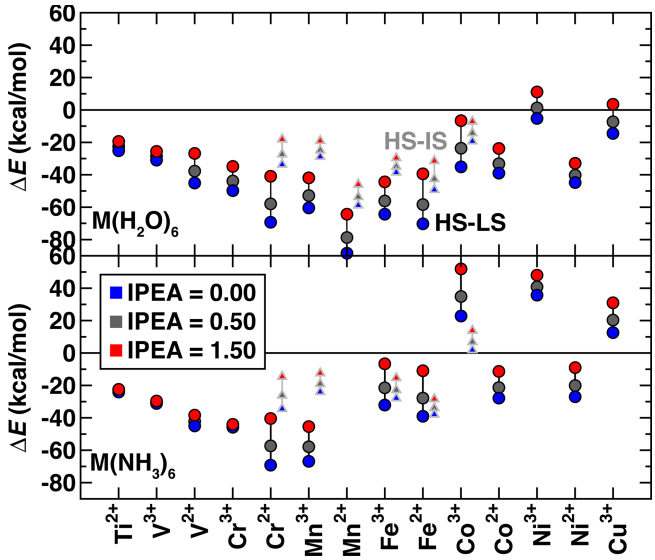

Figure 16. CASPT2 spin-splitting energies, Δ*E* in kcal/mol, for HS-LS (circles) and HS-IS (triangles), as indicated in inset, of hexa-aqua (top) and hexa-ammine (bottom) transition-metal complexes. Results are shown for 3 IPEA shifts: 0.00 (blue symbols), 0.50 (dark gray symbols), and 1.50 (red symbols) a.u.. Both M(II) and M(III) complexes are shown arranged by the number of 3d electrons, from Ti^{2+} to Cu^{3+}. Reprinted with permission from ref. 186. Copyright 2019 Frontiers Media SA.

Overall, MR WFT remains challenging to apply in high-throughput screening or data generation for machine learning, with SR WFT and DFT being easier to automate. While this could change in the near future with improvements in methods and computing power, it will remain useful to know which compounds in a screening pool have the strongest MR character. Numerous diagnostics with heuristic cutoffs have been proposed.[218-229] These diagnostics can be classified by whether they probe the propensity of frontier orbitals to become partially occupied[224-225, 227-231] or whether they indicate a strong sensitivity of a property (e.g., the atomization energy) to the method employed as well as whether they are derived from a SR WFT or DFT calculation versus directly from an MR WFT (e.g., CASSCF leading weights[218-221]) calculation. Many of the most predictive diagnostics require carrying out computationally demanding correlated WFT calculations (e.g., with CCSD for the T_1 diagnostic[218]) that would be too costly for high-throughput screening. DFT-based diagnostics[224, 226-230] that are more tractable

have been proposed, including those based on the atomization energy[224, 226] as well as measures of fractional occupation[227-230] in a finite-temperature DFT calculation. Appropriate diagnostic cutoffs for identifying when MR character is large enough to erode SR WFT predictions differ by material, typically with larger values motivated for transition-metal chemistry in comparison to those for organic molecules.[232-233] Diagnostics often disagree, leading to reliance on intuition or a subset of the proposed cutoffs. The ability of machine learning tools to overcome some of these limitations is described in Sec. 5.3.

2.5. Bringing Statistical Learning to Molecular Modeling.

The molecular mechanics, semi-empirical, and quantum mechanical methods described in Secs. 2.1–2.4 provide essential training data for machine learning model development, as described in detail in the next sections. Here, we briefly touch upon some key efforts that have been taken to systematically improve transition-metal chemical modeling techniques with statistical learning.

MM force fields for transition-metal chemistry are traditionally parameterized by hand against a small number of experimental crystal structures. Although more systematic parameterization, e.g., with GAs[37, 61] or Newton–Raphson[15] has been demonstrated, the small sizes of the data sets limit testing or validation of the force fields. Still, opportunities for incorporating machine learning into this approach were recognized early on.[19] In 2002, ANNs were applied to transition-metal chemistry force field development by Marques and Cukrowski[234], who used the ANN to interpolate estimated errors for the bond distance and force constant of Co(II) porphyrins, reducing the time required to test individual force field parameters (Figure 17). A single trained ANN was selected for this interpolation, and further improvements could be achieved by training on multiple subsets of the data in an approach known as cross-

validation. In cross-validation, the subset of data not used in the training of each model is in the validation set, and the lowest average error over the validation set is used to select parameters to avoid overfitting to noisy data.

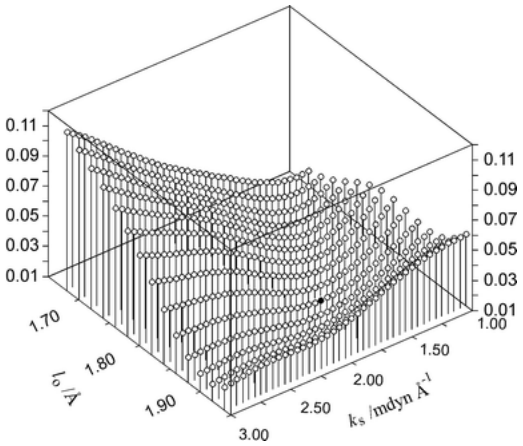

Figure 17. Error response surface generated by an ANN with a single hidden layer with 10 neurons trained on 574 pairs of bond length, l_0, and the Co–N bond stretch constant, k_s, for the modeling of four-coordinate Co(II) porphyrins. The error (vertical axis) is the mean absolute difference between the crystallographically observed and molecular-mechanics-calculated Co–N bond lengths with the respective force constants. The global minimum is indicated by a filled circle. Reprinted with permission from ref. 234. Copyright 2002 Royal Society of Chemistry.

In organic chemistry, SQM modeling and machine learning models have been combined with some success. Integration of the DFTB Hamiltonian into deep learning models as an additional layer has been shown to impart greater physics to the ML model and improve model performance on small organic molecules.[235] Because correlations between SQM and DFT structures or properties are poor[93], the transfer learning approaches that have been successful in organic chemistry[236-237] may fail in transition-metal chemistry.

The use of machine learning to accelerate DFT-level property prediction for transition-metal chemistry at low cost is discussed in Sec. 4. Here, we note some early efforts to systematically reduce errors in DFT functionals with statistical learning. The standard approach to exchange-correlation development typically involves small datasets of organic molecules or closed-shell atoms as well as constraints imposed by satisfying exact bounds of model

systems[121-122]. For transition-metal chemistry, common guidance on functional selection tends to fail, motivating correction schemes with a range of first-principles and empirical roots (Sec. 2.3.3). The "Minnesota" functionals developed by Truhlar and coworkers (e.g., the pure M06-L[238] or hybrid M06[239]) are an exception to this and have gained wide popularity in modeling inorganic complexes.[240] To develop these, a much larger training set that includes transition-metal chemistry is used[239]. These functionals perform well when the chemistry being modeled is sufficiently similar to the training data, but researchers have encountered difficulty in systematically reducing their errors.[180] This can be interpreted in the context of statistical learning, where a high number of parameters and paucity of training data lead to overfitting.[241] Although not targeted at open-shell transition-metal chemistry, researchers have addressed this in part by exploring ways to incorporate regularization and parameter selection[242] into DFT functionals.

In complementary efforts, researchers have leveraged statistical tools, such as uncertainty quantification (UQ) and sensitivity analysis, to inform DFT property prediction. UQ refers to a Bayesian framework where an ensemble of functionals are employed to produce a credible interval on an average prediction.[243-245] While UQ of DFT has primarily been employed in the solid state[243-244, 246], Reiher and coworkers have demonstrated its value in identifying functional-sensitive steps in a molecular reaction network.[247] In closely related sensitivity analysis, derivatives of model predictions with respect to one or more functional parameters are employed.[138, 148] For open-shell transition-metal chemistry, the single most relevant functional parameter is the fraction of exact exchange, which is often associated with key variations of the transition-metal complex (e.g., ligand field, see Sec. 2.3.3). Both linear[138, 148] and non-linear ML models[54] (e.g., artificial neural networks or ANNs) of the relationship between chemical

structure and exchange-correlation functional sensitivity have been developed. These models were used to identify how spin-crossover (SCO) complexes (i.e., with near-degenerate HS and LS states) changed in composition with varied HF exchange (Figure 18).

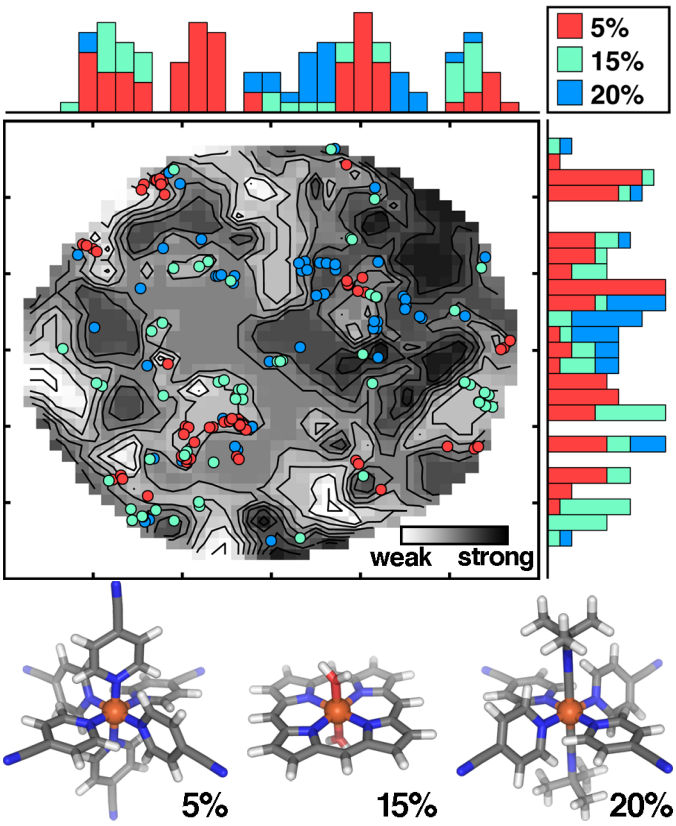

Figure 18. (top) A t-SNE plot of 5,600 complexes colored by connecting atom ligand field from weak (white) to strong (black). Lead spin-crossover complexes (i.e., $|\Delta E_{\text{H-L}}| < 5$ kcal/mol) are shown as circle symbols at three HF exchange fractions: 5% (red circles), 15% (green circles), and 20% (blue circles). One-dimensional stacked histograms of lead compounds are shown for a projection of the two t-SNE dimensions with the same coloring by exchange fraction as the symbols and shown in inset legend. (bottom) Example Fe(II) complex leads at different percentages of HF exchange: a homoleptic 4-cyanopyridine complex at 5% (left), equatorial porphine with two axial water ligands at 15% (middle), and a heteroleptic complex with equatorial 4-cyanopyridine and axial t-butyl-isocyanide ligands at 20% (right). Reprinted with permission from ref. 34. Copyright 2019 American Chemical Society.

ML models that reduce the cost or complexity of recovering the correlation energy in wavefunction theory have yet to be demonstrated in transition-metal chemistry, but some have been developed to aid identification of when correlated WFT is needed (See Sec. 5.3.2).

Statistically informed (e.g., with mutual information between orbitals[214]) methods of active space selection[204, 210, 214] have also demonstrated promise. ML models for active space selection[211] have yet to be demonstrated on large transition-metal complexes but could be improved to enable rapid assessment of the effect the active space on property prediction. The ways in which data science is informing and enriching chemical discovery efforts by reducing barriers to the application of expert-level computational chemistry decisions is discussed in greater detail in Sec. 5.3.2.

3. Physicochemical Properties of Transition-Metal Complexes.

The tunable nature of metal centers in open-shell transition-metal complexes gives rise to a range of applications. Transition-metal complexes are most well established in the field of homogeneous catalysis, including their use as precatalysts.[248-249] Molecular transition-metal complexes have also long been studied for their potential as sensors[250-251], machines[252], molecular magnets[253-254], redox couples for energy storage[255-256], and dye sensitizers[257-258]. As a result of their nature and applications, the target properties for prediction by machine learning or traditional computational (e.g., quantum mechanical or molecular mechanical) techniques differ from those in other domains, such as heterogeneous catalysis, solid-state materials, or in discovery of organic molecules as therapeutic drugs. We thus briefly review these key focus areas for property evaluation with experiment and their relationships to computational predictions. Discussion of these properties provides useful background on the development of statistical and deep learning methods for their prediction (Sec. 4) and in materials discovery (Sec. 5).

3.1. Catalytic Properties.

The most widespread application of transition-metal catalyst design is likely in the

selective functionalization of fine chemicals (e.g., therapeutic drugs).[259] While heterogeneous catalysts are favored in many large-scale operations[259], molecular catalysts remain superior for select industrial applications such as hydroformylation (e.g., with Co or Rh)[248, 260], olefin metathesis (e.g., with Ru)[261], or selective C–H activation (e.g., with Pd)[262]. Transition-metal complexes also have been used widely in photocatalysis[263] and in molecular electrocatalysis[264] (e.g., for water oxidation[265]). Identification of optimal catalyst ligand structures requires determination of the mechanism and turnover frequency (i.e., how quickly the catalytic cycle completes). Catalysts also must be optimized for their selectivity for the preferred reaction without generating undesired intermediates. For a designed catalyst to be useful in practice, the catalyst structure must also be determined to be stable (e.g., not prone to auto-oxidation, dimerization, or ligand dissociation) and not easily deactivated.[266] Observations of catalyst activity will be sensitive to the reaction conditions, including temperature, solvent, and pH. Frequently, experimental efforts tune the Hammett parameters of ligands[267] (i.e., through electron-donating and electron-withdrawing group addition) or otherwise vary ligand properties in order to deduce structure–property relationships. Such kinetic studies can reveal linear free energy relationships (LFERs) that identify limits of how reaction steps may be tuned.[268-269]

Challenges for interfacing data-driven or computational models with experiment occur when the key reactive catalyst structure is poorly characterized, e.g., if the added species is effectively a precatalyst[270] that aggregates into catalytically active nanoparticles[271]. Furthermore, flexible ligand structure and noncovalent interactions have been increasingly adopted as useful tools in catalyst design[272], but their contributions *in operando* can be challenging to determine. Experimentally, many characterization techniques can only be applied to the spent catalyst or reaction products, with short-lived intermediates sometimes trapped through freeze-quenching

X-ray absorption spectroscopy[273], a technique more widely employed in enzymology[274-275]. Other useful spectroscopic techniques for catalyst characterization include ultrafast vibrational (e.g., infrared) spectroscopy[276], electrochemical characterization with cyclic voltammetry[277], isotopic labeling[278], and mass spectrometry[279-280]. Reactions involving hydrogen (e.g., hydrogen atom transfer) can be sensitive to quantum mechanical tunneling effects. Experimentally, kinetic isotope effects can be used to probe the relevance of such contributions in catalysis.[278, 281]

Given the fleeting nature of intermediates, computational tools are an essential complement to experimental catalyst studies. Computationally, relative energetics of intermediates can be identified and activation energies used to obtain rate constants, with trends in rate constants being less sensitive to method errors (see Sec. 2). Combining rate constants with microkinetic[282] modeling or the simplifying degree of rate control[283] or energy span models[284] can help to refine the identification of rate-determining steps, rate laws, and to relate computational predictions to experimentally observable turnover frequencies and numbers (Figure 19).[285] For many open-shell transition-metal complexes, the identity of the spin state is believed to be important, whether by confining the reaction to a single spin state or by multiple spin states being invoked to explain reactivity patterns.[286-287] In these cases, calculation of minimum energy crossing points between spin surfaces as well as coupling between spin surfaces becomes important.[288] All of these efforts are sensitive to errors (Sec. 2) in the electronic structure methods typically employed (i.e., DFT). The experimental resolution of spin-state-dependent reactivity (i.e., with distinct kinetics for each spin state) is challenging but has been corroborated through spectroscopic and kinetic isotope effect data (e.g., for C–H activation).[289-290]

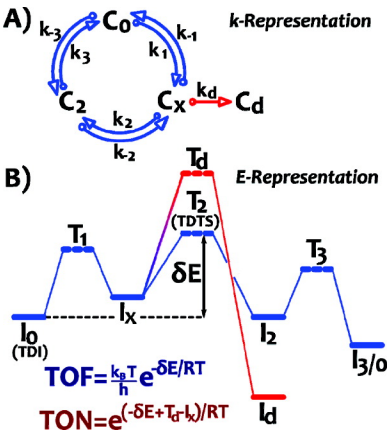

Figure 19. Depiction of the use of the energetic span model to obtain turnover frequency (TOF) and turnover number (TON) from (a) a simplified kinetic model in conjunction with (b) computational activation energies and reaction energies. Deactivation of a catalyst by a slow, irreversible step (red) will cause eventual deactivation. The turnover frequency on the target catalytic cycle (blue) is defined by the energetic span, δE. Reprinted with permission from ref. 285. Copyright 2011 American Chemical Society.

Computational modeling of reaction mechanisms can aid these experimental studies. However, exhaustive search of intermediates and evaluation of transition states are time-consuming. This motivates the reliance on the Brønsted–Evans–Polanyi principle to use reaction thermodynamics as a proxy for activation energies, a particularly widely exploited approach in computational heterogeneous catalysis.[291] The benefit of such analysis has also been realized in experiments through the observation that simple-to-obtain properties can predict catalytic activity.[292] The experimental characterization of some properties of the metal center (e.g., transition-metal hydricity[293]) has been used to predict the effect of solvent on the reactivity of transition-metal catalysts (e.g., for CO_2 reduction[294]).

The direct or indirect roles of the solvent and counterions in homogeneous catalysis are increasingly areas of research focus.[295-296] Computationally, the role of solvent, solubility, and entropic corrections (i.e., both harmonic and anharmonic contributions) are often neglected or treated only approximately (e.g., with implicit models)[297] with studies that consider these effects in combination being rare exceptions.[298] Correction schemes for computational electrochemistry

that may work well for redox potential refinement (Sec. 2) are expected to be challenged by significant rearrangement across an electrocatalytic cycle[299-300].

3.2. Redox Properties.

Redox activity and electron transfer processes are useful for catalysis and energy storage. Cyclic voltammetry is used to determine redox potentials, including for electrocatalyst and functional materials characterization.[277] Well-established redox processes, e.g., the one-electron oxidation of ferrocene, can be useful as an internal reference both computationally[194-195] and experimentally[301-302], especially when varying the solvent in which the redox potential is measured. Nevertheless, only a few moderately sized (ca. 100+ points) data sets of experimental redox potentials for mononuclear transition-metal complexes[178, 303] are available across a range of solvents.

Experimentally, challenges exist in identifying if the structure changes (e.g., by complexation of distinct solvent molecules or counterions[304]) when a material is oxidized or reduced. When multiple metal centers are present in close enough proximity to form metal–metal bonds, unexpected changes in the number of redox events and in redox potential values have been observed (Figure 20).[305] Some of these changes can be traced to the degree of metal–metal bonding (i.e., distance estimated from the formal shortness ratio) and the resulting spin state (Figure 20).[305] An added challenge for both computation and experiment is the difficulty in knowing *a priori* when ligand non-innocence will lead to ligand oxidation or reduction rather than the electron transfer event occurring at the metal.[306-308] A well-established example of this effect is the valence tautomerism between catecholate and semiquinonate forms of chelating ligands, where a reversible equilibrium between the two ligand states can be observed at room temperature, depending on the metal ion identity, charge, and spin state.[309]

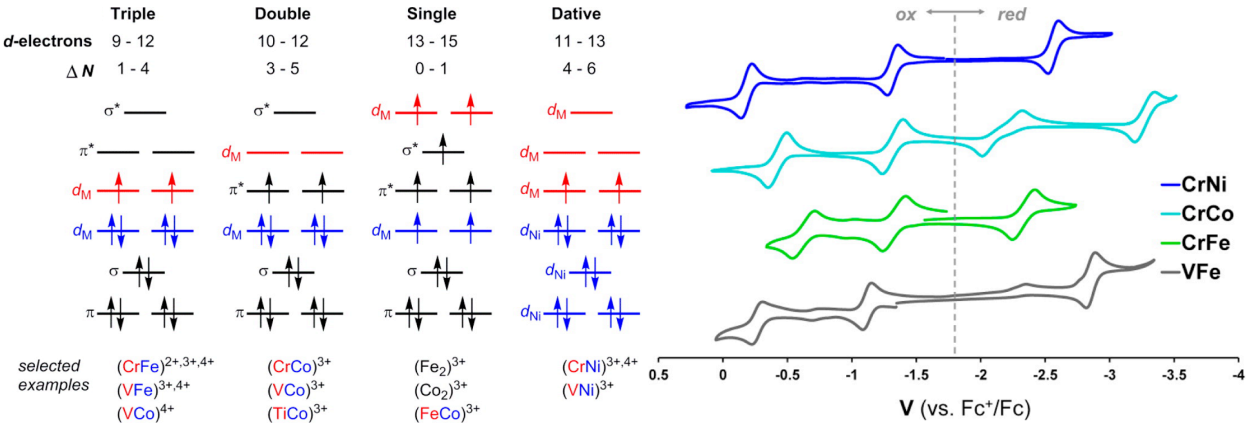

Figure 20. (Left) Depiction of qualitative d-orbital manifolds of select bimetallic complexes formed with double-decker ligands, with the first metal in the lower binding site. The coloring of the energy levels and electrons denotes whether they are delocalized (black) or localized (red and blue) to specific metal sites. The difference in electron configuration between the two metals is indicated as Δ*N*. (Right) Cyclic voltammograms of selected bimetallic complexes with the same ligands, some of which have up to 4 reversible one-electron transfers (e.g., CrNi or CrCo). Adapted with permission from ref. 305. Copyright 2015 American Chemical Society.

For redox couples, even simple cases with a single metal and no anticipated ligand non-innocence, accurate calculation of a redox potential can be a challenge because errors can be sensitive to the relative accuracy of the evaluation of properties of the reduced and oxidized species. Computationally, referenced Born–Haber cycles have been proposed[194-195] as a way to reduce errors (see Sec. 2 and Figure 14), but additional corrections are often necessary for highly charged transition-metal complexes.[310-311] Explicit treatment of solvent, especially in aqueous solution, has been shown to be critical for reproducing experimental trends[312-314], whereas implicit treatment of the solvation environment on redox processes is more computationally tractable for screening and for training data-driven models.

3.3. Optical Properties.

Optical spectroscopic interrogation of transition-metal complexes is useful both for their characterization[315-316] (e.g., electronic and oxidation state identification) and for understanding how their optical properties can be exploited for various applications. For example, harnessing

light to enable chemical transformations is widely exploited in photocatalysis.[263] Transition-metal complexes have long been used as dyes[317], dye sensitizers in optoelectronic devices[257], and have recently seen even broader application in molecular electronics[318]. The spectrochemical series of strong-field (e.g., CN^- or CO) to weak-field (e.g., H_2O or oxalate) ligands is named for the observed relationship between optical absorption of homoleptic octahedral complexes and the effect of the ligands on the metal d states.[319] In some cases, tuning optical absorption by a complex across the visible range (i.e., as observed qualitatively or with UV/Vis absorption spectra) is also possible via solvent adjustment or by tuning the field strength of the axial ligands alone[320] (e.g., tuning L in $Rh_2(OAc)_4(L)_2$, Figure 21). Depending on the ligand, the absorption range of transition-metal complexes can be tuned (e.g., into the visible[321] or into the near infrared[317]). While a key advantage of optical characterization is that the absorption spectrum often provides useful insight into the electronic structure and magnetic properties (e.g., spin state), the relationship between trends in these quantities is not always consistent[322].

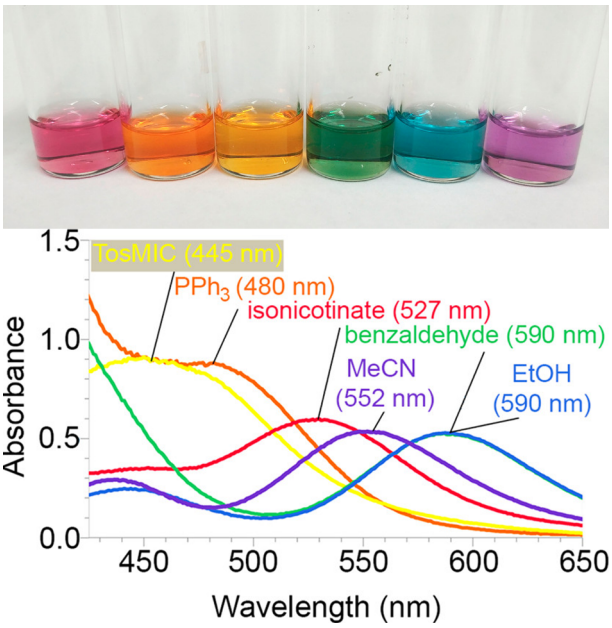

Figure 21. (top) $Rh_2(OAc)_4(L)_2$ solutions (OAc = acetate or $^-O_2CCH_3$) from left to right with: isonicotinate, triphenylphopshine (PPh_3), p-toluenesulfonylmethyl isocyanide (TosMIC), benzaldehyde, ethanol (EtOH), and acetonitrile (MeCN). (bottom) Quantitative absorption

spectra of complexes shown at top, with the lowest-energy band observed to shift to higher energy (i.e., lower wavelength) for increasing field strength. Adapted with permission from ref. 320. Copyright 2019 American Chemical Society.

In addition to the energy gaps between ground and excited states, other properties of excited states can be designed such as lifetimes, e.g., for applications in small-molecule sensing.[321] The interplay of optical and redox or spin state properties gives rise to transition-metal complexes with switchable absorption features[323]. Modulation of absorption in the near infrared (NIR) has been demonstrated with potential applications include in telecommunications.[323] Transition-metal complexes frequently act as emissive compounds, most typically from triplet excited states[324] (e.g., in Ru or Ir polypyridyl compounds). Recently, some of these complexes have been reported to have intense fluorescence with long lifetimes.[324] When the intersystem crossing between excited states is faster in one direction than the other, the resulting phenomenon is called thermally activated delayed fluorescence (TADF), which has potential applications in lighting and molecular electronics[318, 325-326]. Because the metal center and ligand chemistry of the transition-metal complex can be tailored to adjust these excited state properties, they have significant potential for design. Luminescent properties of transition-metal complexes can be readily tuned by adjusting solvent[327] or ligand chemistry. In the context of lighting applications, they can be tailored for emission ranges (e.g., blue) that are currently limited in commercial light-emitting diodes (LEDs), leading to advances beyond existing LED technologies.[318]

Broadly, numerous transition-metal complexes have been observed[328] to have unique non-linear optical (NLO) properties[329], making them amenable for switching in response to chemical, photochemical, or electrochemical stimuli. Different architectures have been developed, wherein metal centers act as donors, acceptors, or bridging moieties.[329] This

tunability and modularity has led to a range of designs of NLO materials based on transition-metal complexes.[329-331]

3.4. Magnetism and Other Switchable Functional Materials Properties.

A change in electronic state of a transition-metal complex in response to stimuli that in turn changes its properties presents opportunities for materials design. As a paradigmatic example, open-shell metal centers (e.g., octahedral, N-coordinated Fe(II) complexes) frequently exhibit spin-crossover (SCO) behavior.[251, 332-333] The transition between a low-spin (LS) singlet and high-spin (HS) quintet state is associated with changes in optical and structural properties that can then be leveraged, e.g., in host–guest[334] sensing. This phenomenon also governs SCO behavior in MOFs[334], including the material $Co[Cr(CN)_6]_{2/3}{\cdot}zH_2O$, which changes both size and color as it is reversibly dehydrated (Figure 22). The change in the spin state typically is associated with a change in the metal–ligand bond length and thus entropic contributions to stability.[334] In these cases, the high-spin state is often higher in energy but becomes stabilized with increasing temperature[332], in the presence of guest (e.g., solvent[335]) molecules[336], change in solvent[337] or lattice packing[338], or can be trapped via light-induced excitation to the high-spin state (i.e., LIESST)[339]. This behavior has also been observed for other metals[333] (e.g., Co(II)[340]) and in non-octahedral[341-342] environments.

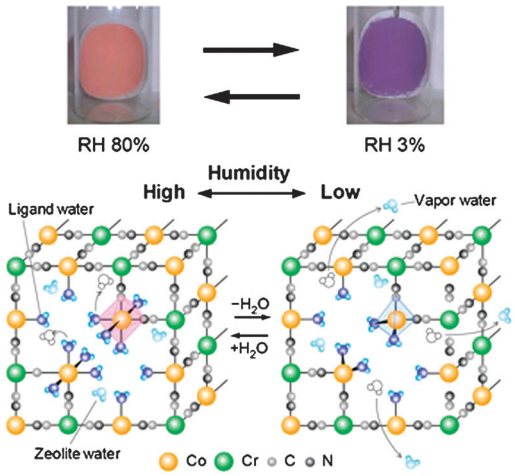

Figure 22. Depiction of color (top) and structural (bottom) changes when $Co[Cr(CN)_6]_{2/3}{\cdot}zH_2O$ is reversibly dehydrated. The pink solid is a ferromagnet that reversibly transforms into a blue antiferromagnet on dehydration, due to the presence or absence of water molecules at Co(II) centers. Reprinted with permission from ref. 334. Copyright 2013 Royal Society of Chemistry.

SCO complexes can be characterized a number of ways. The magnetic susceptibility as a function of temperature typically reveals the nature of the crystalline or solution environment, either having a gradual or abrupt switch along with differing hysteresis widths. SCO is also typically associated with changes in the UV/Vis spectrum (see Sec. 3.1 and Figure 22). For iron-containing complexes, Mössbauer spectroscopy provides key insight into electron density at the metal center, electron configuration, and magnetic state[343]. In fact, Mössbauer spectroscopy was developed in part through observations on SCO complexes.[343] Numerous metals (e.g., Mn, Co) are also amenable to characterization with electron paramagnetic resonance (EPR) spectroscopy[344].

The magnetic properties of transition-metal complexes have also been tailored for device miniaturization in data storage/memory and quantum computing. For quantum computing, molecular qubits[345-347] containing a range of V and Cu centers have demonstrated promise and have been characterized with both magnetic susceptibility and EPR. For memory storage, single-molecule magnets[254, 348-350] containing either multiple Mn centers or mononuclear metal

centers[253], often with heavier lanthanide elements, have also been designed and characterized for their magnetic susceptibility. From a machine learning or computational chemistry perspective, properties of SCOs, single-molecule magnets, and qubits are all influenced by spin state ordering, entropic contributions and spin–orbit coupling between states, with each having some challenges in its description with first-principles methods (see Sec. 2).

4. Statistical Modeling for Transition-Metal Chemistry.

Statistical modeling techniques have long been applied in transition-metal chemistry. However, the availability and diversity of data in transition-metal chemistry imparts unique challenges and opportunities. We describe the available data sets and analysis derived from data mining of experimental and computational results in Sec. 4.1. Next, we briefly describe descriptor development and application in quantitative structure–property relationships (QSPRs) on both experimental and molecular modeling data in Sec 4.2. We conclude this section with early and more recent demonstrations using machine learning algorithms such as kernel ridge regression as well as deep learning (i.e., with artificial neural networks) in Sec. 4.3.

4.1. Data Mining Structures for Chemical Trends.

Data sets of transition-metal complexes tend to be smaller, more varied, and more subject to uncertainty (i.e., noise) in both experimental (see Sec. 3) and computational (see Sec. 2) properties than organic molecule data sets. The largest repository of metal–organic complexes is the Cambridge Structural Database (CSD)[160], which consists of approximately 140,000 unique mononuclear transition-metal complexes in a larger number of solved X-ray crystal structures. The value of mining geometric properties to understand trends and structure–property relationships was recognized decades ago, despite the fact that the set of structures in the CSD was considerably smaller at that time.[351-352] User-friendly graphical tools[353], improved annotation

of metal oxidation state and complex charge[354], as well as a python API have made it increasingly straightforward to mine experimental structures. We briefly review opportunities for using CSD data in data-driven and machine learning models. Reviews by Fey[3] and Orpen[355] provide greater detail of how CSD data mining has been used in transition-metal chemistry.

Subsets of the CSD have been mined to identify trends in transition-metal complex structures, including the propensity of a metal for specific coordination numbers.[356] In the 1990s, Orpen and coworkers started curating subsets from the CSD to interpret "structural systematics" that related ligand structure and bonding, especially in phosphine ligands.[357-361] Avnir, Alvarez, and coworkers analyzed large sets (i.e., over 23,000) of transition-metal complex structures from the CSD and used this data to develop continuous shape measures that describe the relationships between different idealized coordination symmetries and numbers.[362-365] These symmetry measures have been employed to rationalize and interpret changes in SCO complex structure observed with ultrafast spectroscopy.[366] Although developed for molecular crystals, similar local measures of topology based on Voronoi polyhedral volumes and deviations from idealized values have also recently shown promise, e.g., for porous MOFs.[367] Analysis of metal–ligand bond lengths in the CSD has been used to estimate the effective size of atoms, including spin-state-dependent covalent radii of transition metals[368] and the degree of covalency in metal–ligand bonds[369]. The CSD has also been widely used to develop an understanding of structure–property relationships among known classes of materials (e.g., structural asymmetry in Dy-containing single molecule magnets[370]).

The CSD has been mined to uncover other structural trends. This has led to confirmation of expected relationships between bond length and strength (e.g., for late Cu, Zn compounds with common O/N-coordinating ligands)[371], identification of metal coordination number

preference[372], quantification of metal-dependent trans effects (see Sec. 2)[373], and validation of the expectation of longer bond lengths for higher coordination numbers[372]. The large number of structural variations in synthesized ligands also motivated data mining to elucidate the relationship between metal–ligand bonding and steric effects[358] or conformer preference in well-studied macrocycles[374].

Although catalytic intermediates are often fleeting and cannot be captured crystallographically, structure–reactivity relationships have been inferred from CSD data, e.g., by studying the relationship between complex structure and small-molecule binding.[375] For example, a difference in preference of dinitrogen binding to mononuclear (i.e., end-on) versus binuclear transition-metal complexes (i.e., side-on) was observed[375] as were shifts in the degree of metal-dependent (i.e., 3d vs 4d and early vs late) back-bonding in metal–carbonyl bonds[376] (Figure 23). Structures from the CSD have been shown to have good correspondence[377] with typically challenging to crystallize/characterize metalloenzyme active sites.[377] This data set has been leveraged in a number of other ways, including to curate fragments for improved structure generation of *de novo* complexes[50-51] (see Sec. 5).

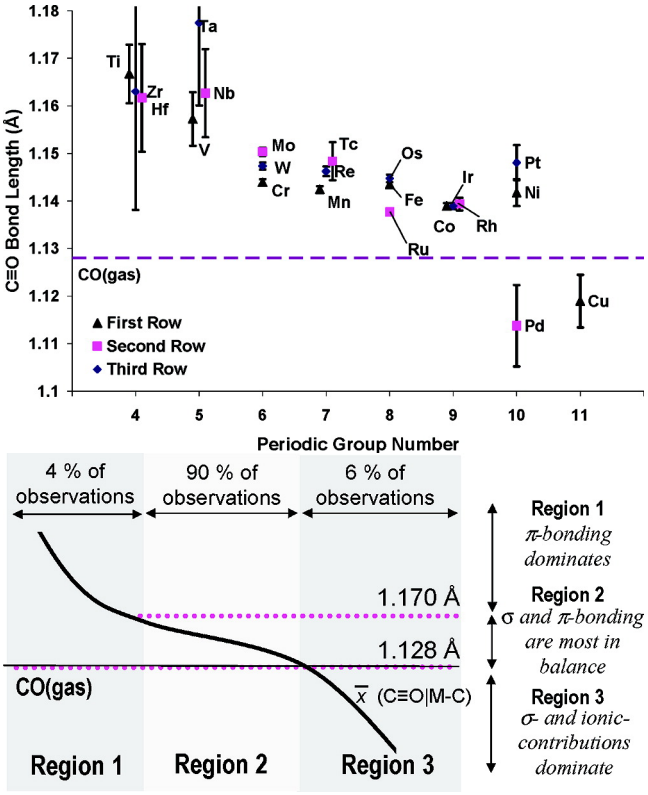

Figure 23. (top) Data of average C–O bond lengths vs periodic group for all transition-metal complexes with known CO bonds. The error bars shown represent 95% confidence intervals around the mean. (bottom) Schematic illustration of regimes of bonding for CO ligands in transition-metal complexes from data-mined CSD structures. Reprinted with permission from ref. 376. Copyright 2007 American Chemical Society.

Despite the value the CSD data set provides, limitations on the utility of the data are notable. Most structures are obtained from X-ray diffraction, which by definition includes the influence of the crystal field environment. Researchers have analyzed packing effects and found them to significantly influence structures of relatively flexible metal–organic complexes.[361] In less pathological cases, these effects were observed to be small (ca. 0.01–0.02 Å).[355, 361] In 2006, Fey, Orpen, and coworkers proposed semi-automated workflows to detect structural outliers and use DFT calculations to determine if the disagreement between DFT and the crystal structure was evidence of unusual chemistry versus a potentially erroneously solved crystal structure (Figure 24).[378] Other complications include the fact that structures in the CSD may have

unknown or erroneous oxidation state assignment, which has led to recent development of efficient heuristic models that can automatically assign oxidation state with good (ca. 85%) accuracy.[354]

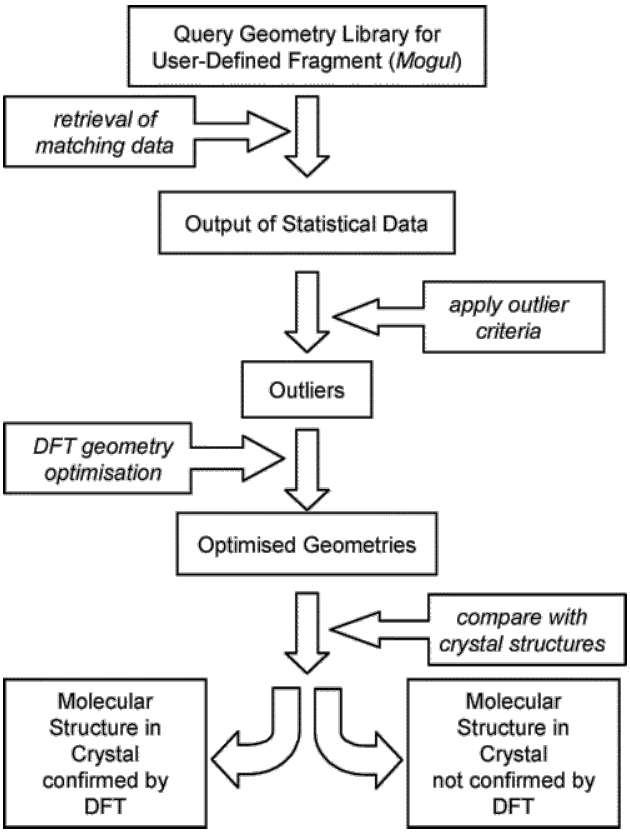

Figure 24. Schematic of partially automatic identification and computational confirmation of outliers in the CSD based on DFT calculations of unexpected transition-metal complex structures. Reprinted with permission from ref. 378. Copyright 2006 American Chemical Society.

The intense experimental focus on metal–organic frameworks (MOFs) has been leveraged in the development of the computation-ready, experimental MOF (CoRE MOF)[379-380] database first introduced in 2014[379]. CoRE MOF consists of most 3D MOF structures deposited in the CSD. In 2019, the database was refined[380] to expand the original set of 5,109 3D MOFs to a larger set of 14,142 MOFs along with an improved procedure for identifying and removing free solvent molecules, recovering disordered structures, and mapping between literature references

and CSD structural files. This type of database has been exploited to accelerate high-throughput screening of MOFs for C–H activation[381] and to build machine learning models to predict properties related to gas separations and storage, such as surface area[382] and adsorption isotherms[383]. These sets have also been instrumental in revealing mismatches between hypothetical and synthesized materials[384]. Given that the CoRE MOF resource provides a mapping between the structures and the literature associated with them, natural language processing (NLP) has begun to be exploited to map structure to properties in MOFs.[385-386] NLP has been more widely demonstrated for other tasks, including understanding materials synthesis recipes[385] (see also Sec. 5.3.4).

New tools are also being developed as part of the CSD to allow for expansion and analysis of experimental structures.[387] In a similar fashion to the MOF or solid-state datasets, closed-shell molecular transition-metal complexes have recently been the subject of high-throughput computational screening with a combination of semi-empirical methods (see Sec. 2) and DFT.[388] The generation of DFT properties on experimental structures is considerably better established for solid-state materials, where a number of repositories are available (e.g., AFLOWlib[389], Materials genome project[390], and Open Quantum Materials Database[391]) that share semi-local-DFT-computed properties for materials design. An outstanding limitation is that many of these sets of extracted structures are primarily used as a starting point for simulation, rather than mapping to experimental results (e.g., for extraction of spectroscopic or catalytic quantities), which may not be systematically reported. For example, to curate experimental data for their empirical DBLOC corrections (see Sec. 2), Friesner and coworkers manually analyzed graphs for key peak features.[175]

CSD structural data has also been subject to multiple linear regression (see Sec. 4.2) and

supervised ML models (see Sec. 4.4) and analyzed with unsupervised machine learning (see Sec. 4.3). In particular, Fey, Orpen, and coworkers have leveraged the CSD data to build "ligand knowledge base" (LKB)[392-393] subsets and analyze them with PCA and MLR.

4.2. Building Structure–Property Relationships: Quantitative and Mechanistic Inquiry.

Recent years have seen the development of machine learning models (Sec. 4.4) that encode structure–property relationships and enable large-scale chemical exploration (Sec. 5). We briefly summarize here complementary approaches that bear similarity to machine learning efforts but typically rely more on intuition and human intervention. First, we briefly review quantitative structure–property relationships (QSPRs) for transition-metal chemistry, including frequently invoked descriptors (Sec. 4.2.1) and examples of multiple linear regression QSPR models (Sec. 4.2.2). Throughout, we use the term QSPR to refer broadly to simple, multivariate or heuristic mapping of structure to either property or activity of molecules (i.e., both QSPR and QSAR). Finally, we discuss how these approaches can be used quantitatively or qualitatively to guide iterative and mechanistic inquiry (Sec. 4.2.3). Some of these approaches remain in widespread use in conjunction with machine learning models, as will be discussed in subsequent sections, and the performance of these transparent models can be used to determine when moving to more sophisticated machine learning models is beneficial.

4.2.1. Common Descriptor Types for QSPRs.

Numerous descriptors have been developed as inputs to linear QSPR models or as guides for iterative design of catalysts. Recent reviews in Refs. 2-3, 394-396 provide more detail on the range of descriptors that have been used in the mechanistic study of transition-metal complexes. We briefly highlight here some representative QSPR descriptors to enable comparison to representations more typically applied in deep learning in Sec. 4.4. The descriptors typically used

in QSPR models are either derived from quantities that are easy to measure experimentally (e.g., from vibrational spectroscopy or UV/Vis) or to compute. To be predictive across modest (ca. 40–50) transition-metal complex sets, the descriptors are chosen to represent key chemical variations that determine an activity or property.

For ligand design in homogeneous catalysis, where the approach of substrate to the catalyst is strongly directed by ligand shape, a focus has been placed on the sterics of either the individual ligands or of the complex. An early and prominent example of steric descriptors is the cone angle for phosphine ligands devised by Tolman[397-398] (Figure 25). By approximating the steric bulk of the substituted phosphine ligands, the Tolman cone angle was shown to approximately quantify experimental measures of the ease of displacing ligands in catalytic Ni(0) complexes[397-398] (Figure 26). For N-heterocyclic carbene (NHC) complexes, the fraction of volume of the first coordination sphere that is "buried" by overlap with atoms from the NHC ligand has been proposed as an alternative quantitative descriptor of steric bulk (Figure 25).[396, 399-400] Because this volume represents the space around the metal atom that must be shared by the different ligands upon coordination, it has been shown to be a good linear predictor of the experimentally measured relative cost to displace ligands from Cr-containing transition-metal complexes (Figure 27).[399] These descriptors have been made accessible to users through the development of an easy-to-use web interface.[400] First developed in the 1970s, sterimol parameters of ligand length and asymmetry have recently been demonstrated[401] to be useful descriptors for QSPRs in the design of enantioselective catalysts[396]. Steric descriptors have also been obtained from the molecular mechanics ligand repulsion energy (i.e., van der Waals terms only) associated with the cost to displace a ligand from a metal center.[402]

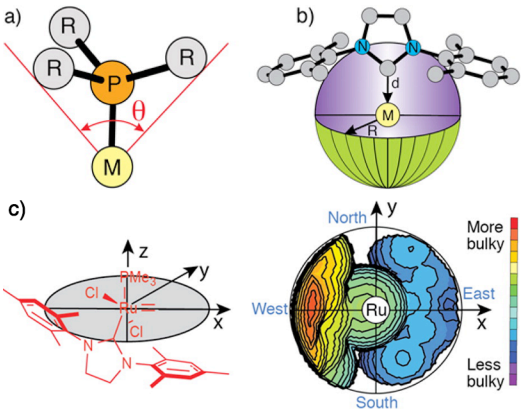

Figure 25. Depictions of (a) Tolman cone angle, (b) buried volume steric descriptors, and (c) steric map of a (NCH)(Cl_2)Ru=CH_2 complex (right) oriented according to the structure shown on the left. Adapted with permission from ref. 400. Copyright 2016 American Chemical Society.

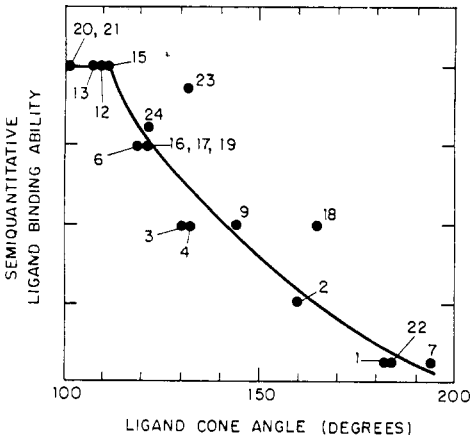

Figure 26. Stability of Ni(0) tetrahedral complexes measured through semiquantitative ligand binding ability obtained from ligand competition experiments as a function of Tolman ligand cone angle. Reproduced with permission from ref. 397. Copyright 1970 American Chemical Society.

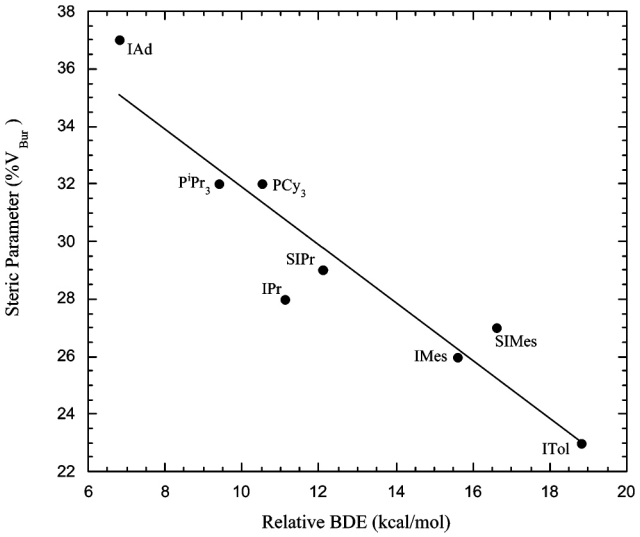

Figure 27. Relative experimental bond disruption enthalpy (BDE, kcal/mol) vs steric parameter (percent buried volume, %V_{Bur}) of N-heterocyclic carbene ligands, L, in the Cp*Ru(L)Cl system (slope: −1.01; R: 0.94). Reproduced with permission from ref. 399. Copyright 2003 American Chemical Society.

Geometric properties were widely used as descriptors in early QSAR/QSPR studies because they can be readily extracted from crystal structures[403] (see Sec. 4.1) or reasonably estimated from moderately accurate molecular mechanics force fields[404] (see Sec. 2). Experimental or calculated chelating atom distances were used to classify spin-crossover complexes.[405] Similarly, molecular mechanics and DFT were both shown to be predictive of experimental bidentate bite angles that can inform reactivity, thus demonstrating how tuning ligand chemistry can be used to rationally design ligands.[404, 406] Analysis of known Ziegler–Natta catalyst structures revealed the importance of metal–ligand dihedrals as descriptors.[403] Other measures of the planarity of the ligand around the metal (e.g., twisting and puckering) have been proposed from semi-empirical or force field modeling to be useful in predicting enantioselectivity of catalysts (e.g., Jacobsen–Katsuki epoxidation catalysts).[407-408]

In addition to numerical descriptors, 3D QSPR descriptors originally developed for organic chemistry were adapted to ligand design in transition-metal complexes. One of the most

widely used approaches in 3D QSPR is comparative molecular field analysis (CoMFA).[409] In CoMFA, a molecular electrostatic potential or steric repulsion is evaluated using a probe molecule on a grid around the structure. Although initially designed with protein–ligand binding in mind, CoMFA descriptors found use in ligand design for transition-metal chemistry.[410-413]

In close analogy to CoMFA, specialized projections of steric bulk and crowding proximal to the metal center in steric maps have been proposed to provide richer insight into potentially anisotropic contributions relevant for enantioselective catalysis (Figure 25).[400] Lipkowitz and coworkers pioneered[414-415] an approach tailored for enantioselective catalyst design that they referred to as stereocartography. In this approach, the probe is a transition state displaced along a grid to identify regions of maximum stereoinduction (Figure 28). Using this stereocartography technique in conjunction with semi-empirical energy evaluations, most (ca. 95% or 17 of 18) catalysts were found to conform to the hypothesis that enantioselectivity is made possible only when the point of greatest discrimination between enantiomers is close to the reacting center.[414-415]

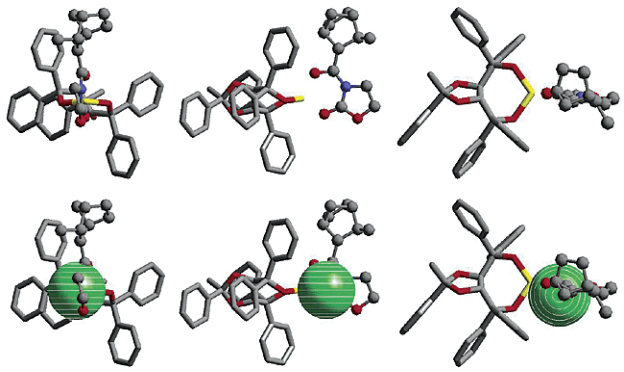

Figure 28. Location of transition-state probe and region of maximum stereoinduction for a catalyst with a docked transition-state probe (left: front, middle: side, and right: top views). The heavy atoms of the catalyst are depicted as sticks, the transition-state probe atoms are shown in ball-and-stick, and hydrogen atoms are omitted for clarity. The bottom row shows the most enantiodiscriminating region within a sphere of 1 Å radius centered about the most enantiodiscriminating grid point. Reproduced with permission from ref. 414. Copyright 2002 American Chemical Society.

Through-bond electronic effects are also known to have a significant influence on the properties of transition-metal complexes and have therefore been invoked in QSPR models. For example, Hammett parameters have long been used to quantify and predict effects of tuning on catalysis.[416-417] Both parameterized semi-empirical methods and DFT functionals became increasingly accessible over the past 20 years (see Sec. 2), leading to the increased use of QM-computed electronic descriptors in QSPR/QSAR modeling.[3, 394] Electronic parameters easily extracted from DFT were observed to correlate to experimentally measured structural determinants of reactivity that could only be obtained after a catalyst was crystallized and characterized.[418] Widely employed descriptors include metal partial charges and quantities derived from frontier orbital energies.[419] Reactivity descriptors such as estimates of electrophilicity[419] or softness[420] of the metal have been proposed. Similarly, explicitly computed ionization potentials, electron affinities, or the analogous orbital energies (e.g., HOMO or LUMO) have been widely used as descriptors owing to the expectation of their influence on reactivity.[420-421] Other DFT-derived descriptors include simpler-to-compute quantities (e.g., LUMO, pK_a) representative of the overall catalytic activity but obtained only on the organic ligands in the complex.[422] The application of these descriptors in iterative catalyst design with linear models is discussed next in Sec. 4.2.2, and the relationship of these descriptors to those developed more recently for machine learning applications will be discussed in Sec. 4.4.

4.2.2. Examples of QSPR Models for Transition-Metal Chemistry.

QSPR modeling in transition-metal chemistry has provided important insight into the design of transition-metal complexes. Here, we distinguish QSPRs from other supervised learning (Sec. 4.4) models by referring specifically to models developed on small data sets (ca. 50–100) points that are typically used to predict an experimental measure of catalyst activity,

redox potential, or other property using (multiple) linear regression (MLR).[3, 394, 410] While QSPR models are typically fit to all available data given the small data set sizes, there are exceptions where cross-validation error[423] or other error estimations from bootstrapping were demonstrated.[406]

Prediction and tuning of redox potentials is an example application to which QSPR modeling has long been applied because the absolute experimental redox potential is challenging to predict even from accurate theoretical chemistry methods (see Sec. 2). At the same time, prediction of relative redox properties to identify opportunities for tuning is an easier task, and corrections to DFT can be easily fit with simple linear models[194, 424]. For example, the difference in strain energy between reduced and oxidized states of redox couples (e.g., Co(III/II)[425-426]) from a tailored (i.e., MM2 with tuned terms for the metal–ligand bond) force field has been demonstrated to estimate experimental redox potentials with reasonable fidelity (i.e., $R^2 = 0.78$). This agreement is impressive considering the neglect of solvent, entropic, and other electronic effects not included in the methods. Descriptors from DFT, such as orbital energies of electrons involved in oxidation, have also been used to identify how to tune redox potentials (e.g., in W-containing complexes[427]).

For predicting spin-state-dependent behavior, Phan et al. showed that the experimental nitrogen–nitrogen separation of a free ligand in 33 homoleptic Fe(II) diimmine complexes could be used to classify when a complex would exhibit spin-crossover behavior (Figure 29).[405] Geometric descriptors were also found to be important in iterative ligand design for spin state properties of Fe(II) polypyridines.[428] Data mining and linear correlation analysis has also been carried out on substituted Mn cubanes to identify the most important structural properties that influence magnetic behavior.[429]

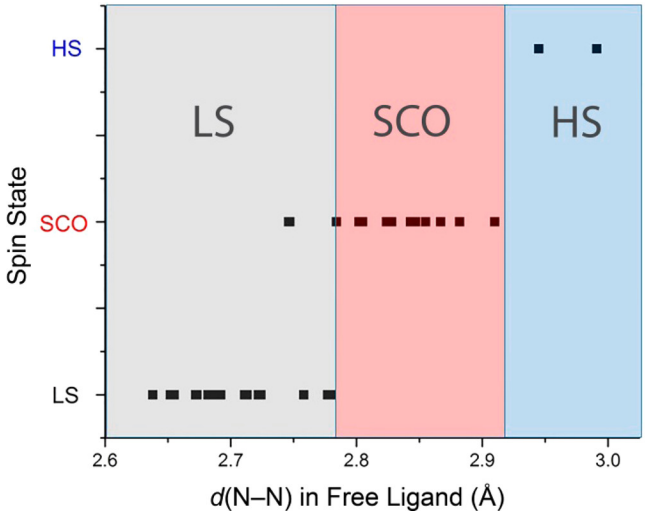

Figure 29. Comparison of N–N separations (in Å) for free ligands from 33 tris-diimine Fe(II) complexes with their known spin state preference. Reprinted with permission from ref. 405. Copyright 2017 American Chemical Society.

QSPRs have been most widely employed for experimental catalyst screening. Much of this work is outlined in recent reviews such as Refs. 394, 413, 430. Often, the target predicted quantity (i.e., experimental enantioselectivity) in these cases is sufficiently challenging to obtain from computation such that even relatively expensive descriptors derived from DFT calculations represent an efficiency improvement over direct screening. In nearly all such cases, a single metal and oxidation state is employed, meaning that organic-focused descriptors developed for organic chemistry QSPRs are often suitable. Sigman and coworkers have demonstrated both quantum mechanical and steric descriptors across data sets of numerous catalysts to predict experimental activity[431-433], as reviewed in detail in Refs. 401, 430. In most cases, descriptors are evaluated on a single ligand or substrate conformer, but for highly flexible ligands and substrates, weighted evaluation of 3D descriptors has been demonstrated.[434] In many of these examples, the QSPR models themselves serve to guide improvements to ligand chemistry.

At the same time, models capable of predicting DFT-derived catalytic properties have also been pursued in cases where experimental data was limited or as a means to understand

trends in DFT-derived reactivity. For example, Truhlar and coworkers investigated[419] how linear models could predict selectivity of Ru catalysts with NHC ligands toward olefin metathesis over cyclopropanation. To simplify screening, they correlated DFT estimates of catalyst selectivity (i.e., relative reaction free energy barriers of the competing reactions) to electronic structure descriptors (Figure 30).[419] They showed that both metal partial charge and electrophilicity could explain DFT selectivities, highlighting a useful trend for experimental ligand design.[419] While correlations to experimental data may be preferred, identification of predictive quantities from the DFT calculation provides important insights into design principles where experimental data is limited or not robust due to contributions from competing processes, such as side reactions.

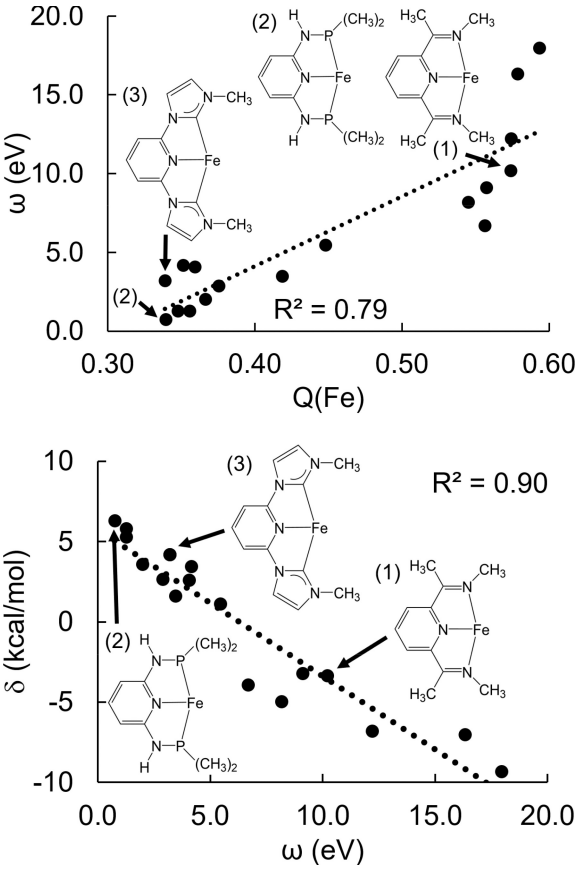

Figure 30. (top) Electrophilicity index, ω, as a function of the CM5 partial atomic charge of the Fe ion in the metallocyclobutane complex, *Q*(Fe). (bottom) Free energy selectivity, δ, from the difference of free energies of activation of reactions 1 and 2 as a function of ω. Resulting linear

fits are shown inset. In both panels, each data point is for a different ligand environment. Reproduced with permission from ref. 419. Copyright 2018 American Chemical Society.

Parveen et al. investigated whether through-space geometric descriptors or through-bond electronic effects could better explain differences in 38 sandwich Zr complexes for polymerization.[435] The developed MLR models were predictive of reactivity trends (R^2 = 0.86) and performed nearly as well as more sophisticated (e.g., ANN) machine learning models.[435] Jensen and coworkers employed a combination of cheminformatics-derived topological (see Sec. 4.4) and DFT-calculated descriptors to predict DFT-calculated estimates of activity for 82 ligands in Grubbs Ru olefin metathesis catalysts, where the DFT-calculated activity was also known to correlate to relative experimental activity.[436] The final MLR model consisted of 14 features and had a modest cross-validation error after manual pruning of descriptors on a representative test set below 1 kcal/mol.[436] Analysis of the most important features revealed the dominant role of geometric and electronic properties of the Ru=CH_2 bond for iterative ligand design (see Sec. 4.2.3).[436]

Linear models have also been built on DFT data to identify substrate-dependent factors that favor different (e.g., homolytic vs heterolytic bond cleavage) catalytic mechanisms.[437] Nazemi and Cundari used this analysis to identify that substrate pK_a could play a more significant role than homolytic bond dissociation free energies in C–H activation by metal-oxo catalysts.[437] Correlations of ligand pK_a have also been demonstrated across far-ranging applications from ligand precursors in quantum dot synthesis[438] to CO_2 conversion[422] kinetics. Relationships between pK_a and a target activity/property are closely related those observed by Hammett on organic compounds[416-417] and the concepts of linear free energy relationships described in Sec. 4.2.3.

Fey et al. used QSPR descriptors and MLR models as analysis tools for understanding ligand chemistry for design.[406] They observed that stability of Cr complexes with chelating bidentate P,P or P,N ligands, as judged by DFT ligand removal energies, could be well predicted (R^2 = 0.95 and mean squared error of 3.2 kcal/mol, Figure 31).[406] This model used a combination of only five descriptors that included the proton affinity of the coordinating phosphorus in the ligand, two from the binding energy of the ligand to other metals (Zn or Pd), and two structural properties of the Zn complex analogue (i.e., the Zn–Cl bond in a $Zn(L)Cl_2$ complex and the change in ligand angle upon complexation).[406] Addition of other descriptors, including the HOMO energy or other geometric properties, did not substantially improve model performance. This analysis highlights strategies to tune ligand chemistry and the extent to which these design principles might be transferable across multiple metals.

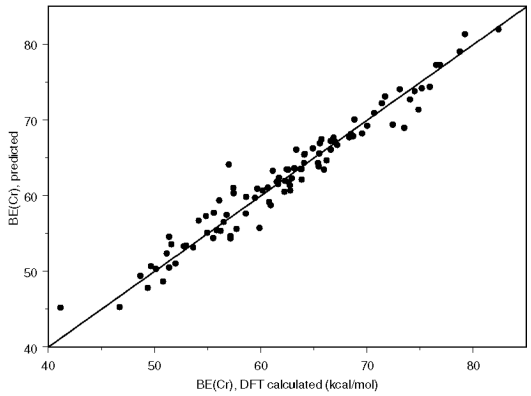

Figure 31. Prediction of the binding energy of a bidentate phosphine ligand, L, to Cr from DFT (BE(Cr), in kcal/mol) in a $Cr(CO)_4L$ complex. The model R^2 is 0.95 and descriptors include the binding energy for ligands to Zn and Pd complexes computed with DFT as well as proton affinity of one side of the ligand and change in a ligand angle upon complexation. Adapted with permission from ref. 406. Copyright 2008 American Chemical Society.

As an alternative approach, Maseras and coworkers[439] used singular value decomposition (SVD) to uncover the predominant descriptors to build QSPRs for ligand binding energy (Figure 32). They determined that the number of "hidden" descriptors needed to predict DFT-level bond dissociation energies (BDEs) from transition-metal complexes was quite small, with a model that

used only six able to predict the DFT BDE to within 1.4 kcal/mol. While application of the model to new complexes requires additional calculations to generate the relevant matrix elements, model generalization error was low[439] (Figure 32). They then identified the correlation between the six predominant descriptors obtained from the SVD to individual descriptors from a set of over 660 standard quantities (e.g., charges, volumes, ionization potential, etc.).[439]

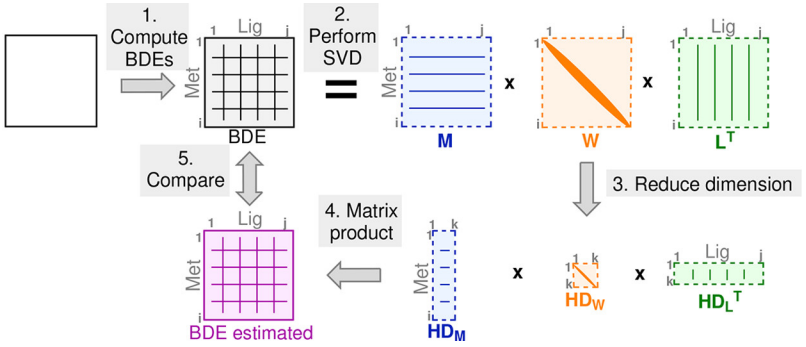

Figure 32. Schematic of singular value decomposition approach to analyzing hidden variables in DFT-computed bond dissociation energies between a matrix of metals and ligands in transition-metal complexes. The reduced dimensionality enables generation of a simplified matrix product to obtain estimated BDEs. Reproduced with permission from ref. 439. Copyright 2018 American Chemical Society.

4.2.3. Descriptors in Linear Models and Iterative Design.

Many transition-metal complex design studies leverage individual descriptors normally employed in combination in QSPRs to instead develop, test, and validate hypotheses. Such studies may evaluate effects one at a time or employ descriptors to guide the update of the set of compounds to evaluate with DFT or experiments. For example, orbital energies found to correlate to C–H activation were used to accelerate the screening of ligands for candidate methane-to-methanol catalysts with metals (i.e., Rh) that have been less frequently studied experimentally.[440] Such relationships tend to hold well within a given spin state and metal, but catalysts with varied ligand chemistry (e.g., denticity) or spin states often require the development of multiple linear relationships when a single descriptor is used. As an example,

Sautet and coworkers showed multiple correlation lines were needed to explain the relationship between the LUMO σ* orbital of open-shell Fe=O moieties and the C–H activation energy when comparing monodentate and multidentate catalysts (Figure 33).[441]

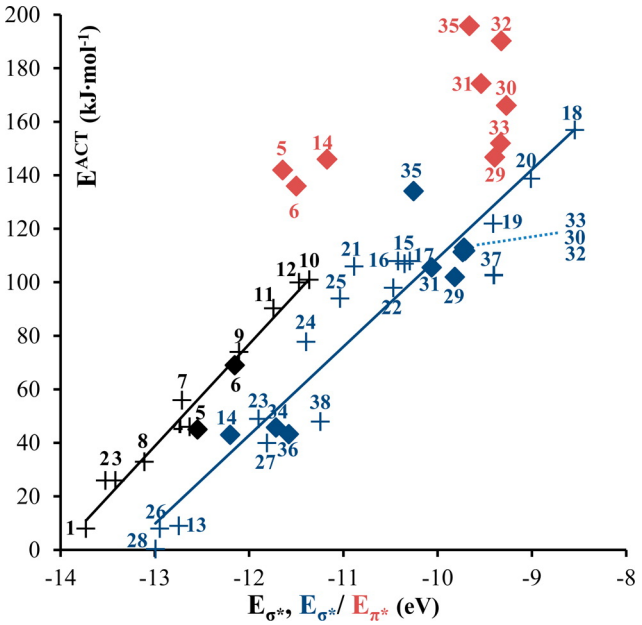

Figure 33. Correlation between the DFT (i.e., OPBE) energy of the σ* orbital of the Fe–oxo complex in eV and the DFT barrier height of the C–H abstraction step in kJ/mol for model catalysts in the HS-σ path. Two subsets are distinguished: (**1**–**12**, monodentate model catalysts) in black and (**13**–**38**, macrocyclic, experimentally motivated catalysts) in blue. Crosses indicate complexes with a quintet ground state, while diamonds indicate those with a triplet ground state. For the Fe═O complexes with a triplet ground state, the activation energy for C–H abstraction along the LS-π path is also plotted as a function of the energy of the π* orbital shown as red diamonds without any correlation indicated. Reproduced with permission from ref. 441. Copyright 2015 American Chemical Society.

Common phosphine ligand descriptors related to angle and bonding strength were used by Cooney et al. to computationally identify novel ligands that exhibited behavior distinct from those that had been previously synthesized (Figure 34).[442] While less frequently applied, this descriptor-guided screening approach has been employed to identify ways to improve catalyst stability (i.e., by strengthening the bond with the weakest ligand) without affecting activity.[443] When sterics are believed to play an essential role, this screening and iterative design process has been accelerated by using molecular-mechanics-derived quantities such as the ligand repulsion

energy[402] to predict relative stereoselectivity of substrate binding.[444]

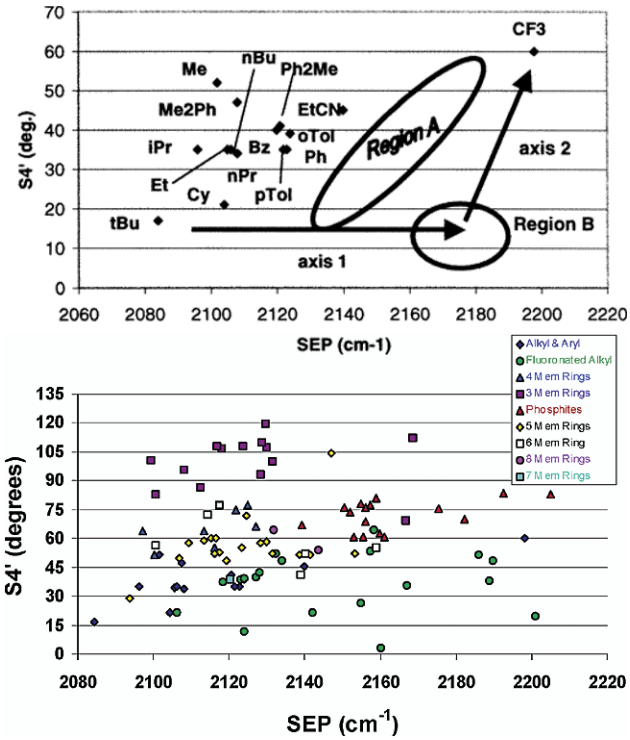

Figure 34. (top) Calculated stereoelectronic map for phosphines derived from semi-empirical (i.e., PM3(tm)) calculations on *trans*-Rh(PR_3)$_2$(CO)Cl with ligands from the CSD. (bottom) Proposed properties of novel phosphines that fill in the gap identified in the top graph. SEP corresponds to the stretching frequency of the bound CO and S4' is a modified Tolman cone angle. Adapted with permission from ref. 442. Copyright 2003 American Chemical Society.

Exploiting linear free energy relationships (LFERs) between activation energies and reaction energies or selectivities as well as relating energies to a specific descriptor is a special example of descriptor-based design. This approach has been most widely applied and popularized in heterogeneous catalysis, where it has been shown that bulk metal alloy reaction energetics are closely related to descriptors (e.g., the d-band center), enabling estimation of overall optimal activity (i.e., a maximum in a volcano plot) from a single descriptor.[445] The natural bridges between heterogeneous and homogeneous catalysis have become more evident[446], as LFERs have been widely demonstrated in single-site MOF catalysis[381, 447-448] and single-atom catalysts (SACs)[449-450], a family of materials with isolated metal sites coordinated by

N-doped graphene substrates that resemble and behave like macrocyclic catalysts (e.g., porphyrins)[451].

Within homogeneous catalysis, the potential role of LFERs for catalyst design has thus been investigated with increasing interest in recent years.[422, 441, 446, 452-458] Corminboeuf and coworkers[452-455] have demonstrated how LFER-based volcano plot development can be used to accelerate homogeneous catalyst screening. They have demonstrated their approach to scaling relations for a range of reactions including cross-coupling[454-455], hydroformylation[453], and CO_2 hydrogenation[452], among others. Recognizing the potential variability of mechanism and scaling relationship as ligand chemistry changes, they proposed[453] developing and applying scaling relations within families of ligand types (e.g., phosphines) to then construct multiple volcano plots and identify the optimal ligand chemistry for a specific reaction (Figure 35).

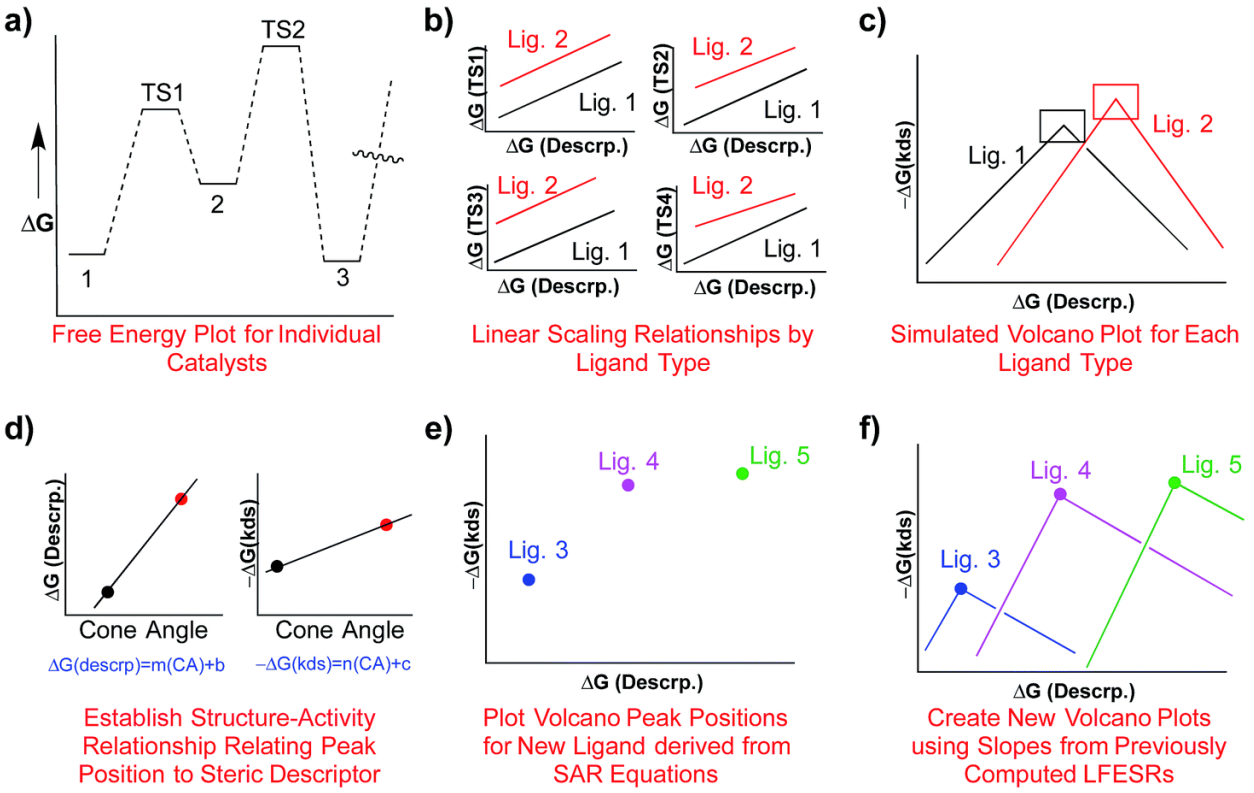

Figure 35. Summary of an approach to generate volcano plots from scaling relationships. a) The free energy profiles for several catalysts bearing the same ligand are computed. b) Linear scaling relationships between relative free energies of the transition states and the free energy of the

descriptor variable are established. c) Simulated volcano plots for each ligand type are derived from the linear scaling relationships. d) The peak positions are then related to a steric parameter, such as the cone angle, and equations are derived that relate the steric parameter to a new peak position. e) The peak positions of the hypothetical volcanoes for new catalysts are plotted. f) The final hypothetical volcanoes are created by using the volcano slopes from the computed data. Reproduced with permission from ref. 453. Copyright 2016 Royal Society of Chemistry.

In cases where a mechanism and scaling relationship is known, iterative experimental-computational efforts have been carried out in which calculations are used to predict functionalizations of ligands that will lead to improved performance on established catalyst scaffolds.[459-461] In another example, computed relative free energies were correlated to experimental selectivities (i.e., for 1-octene over 1-hexene formation) for ethylene oligomerization.[462] This correlation was exploited to successfully computationally screen new ligand chemistry with the desired selectivity and validated experimentally (Figure 36).[462] This type of relative free energy barrier screening has also been beneficial in identifying, for a fixed Ni phosphine catalyst chemistry, which leaving groups on substrates were compatible with the desired functionalization (Figure 37).[463] Numerous researchers have exploited computed[464-465] transition-metal hydricities[293-294] as a descriptor for activity in Fe(II) and Co(III) CO_2 hydrogenation catalyst screening, owing to the scaling relationships to other governing steps in the catalytic cycle, including CO binding. The energetic span model[284] represents a generalization of this approach and has been fruitfully employed to iteratively vary pincer ligand design to maximize activity for carboxylation with CO_2 by identifying ligand features that increase turnover frequency[466] or to analyze the effects of functional groups on Rh sandwich catalysts[467].

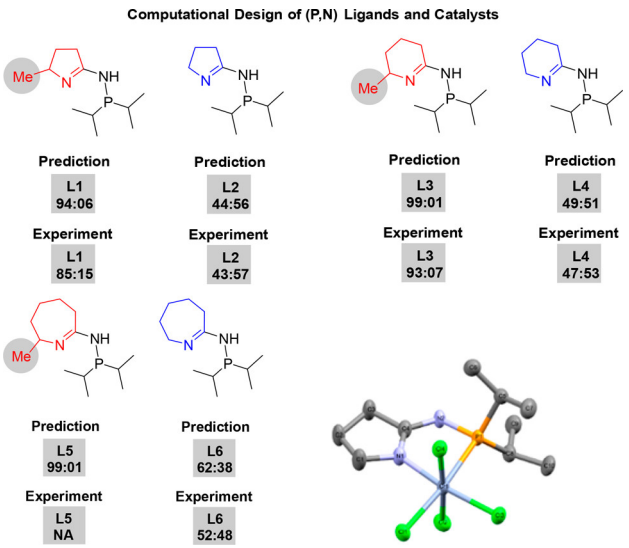

Figure 36. Predicted 1-hexene:1-octene weight % ratio for Cr-catalyzed ethylene oligomerization with phosphine monocyclic imine ligands from relative free energy barriers for the two reactions compared to the experimentally observed result. NA = catalyst not synthesized. The experimental weight % ratios are normalized for C_6+C_8 fractions and include all C_6 or C_8 components. Bottom right corner shows the X-ray structure of $[Li(THF)_2]^+[(\mathbf{L2})Cr^{III}Cl_4]^-$. Reproduced with permission from ref. 462. Copyright 2018 American Chemical Society.

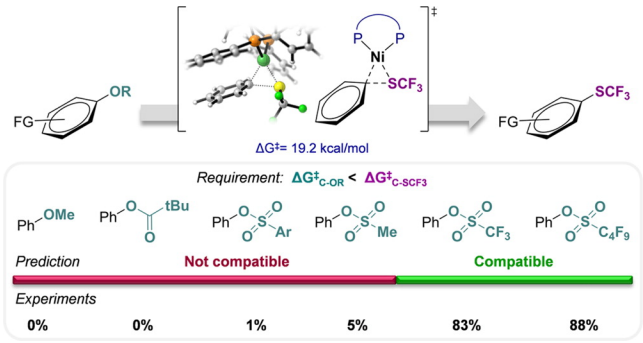

Figure 37. Computational prediction and validation of the leaving groups compatible with Ni-catalyzed trifluoromethylthiolation by the catalyst depicted at top. Reproduced with permission from ref. 463. Copyright 2016 American Chemical Society.

Nevertheless, the LFER-based, iterative screening approach has been most successfully demonstrated in closed-shell complexes with late transition metals. Larger-scale exploration of multiple oxidation states and spin states typically reveals the limitations of LFERs in catalyst screening. For example, the HOMO level of the resting state of a catalyst has been invoked as a good descriptor for how favorably a high-valent metal-oxo species capable of C–H activation

will form[448], but numerous screening efforts have revealed exceptions.[152, 441, 456, 468] Gani et al. observed[456] that geometric distortions about Fe(II) metal centers in minimal model catalysts broke the strong scaling between HAT and oxo formation central to the design of direct methane-to-methanol conversion catalysts (Figure 38). Tuning the ligand field, however, resulted in shifts along a standard scaling relationship, and analysis of experimentally characterized catalysts revealed that many macrocyclic catalysts indeed had the favored metal–ligand plane distortion. Distortion was also observed to be naturally occurring and beneficial in a class of Fe(II) single-molecule magnets due to Jahn–Teller distortions.[469] Nandy et al. showed[457] that over a set of nearly 1200 catalysts, scaling relationships between metal-oxo formation and HAT differed widely with metal, oxidation, and spin state (Figure 39). The observed scaling relations also differed from those that had been proposed in heterogeneous[381, 447] (i.e., including MOFs[381]) catalysis. These wide variations indicated that use of a single scaling relationship across the entire data would fail to be predictive for catalyst screening, motivating more flexible machine learning models (see Sec. 4.4).

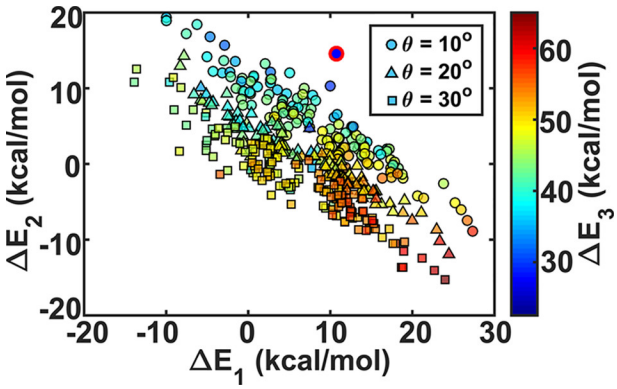

Figure 38. Oxo formation energy (ΔE_1, in kcal/mol) versus HAT energy (ΔE_2, in kcal/mol) for minimal model Fe(II) complexes with square planar distorted geometries. The minimal models are grouped by metal–ligand plane dihedral angle: 10° (circles), 20° (triangles), and 30° (squares). Symbols are colored according to the corresponding CH_3OH release energy (ΔE_3, in kcal/mol), as indicated in the inset color bar. A single outlier is indicated by a red border. Reproduced with permission from ref. 456. Copyright 2018 American Chemical Society.

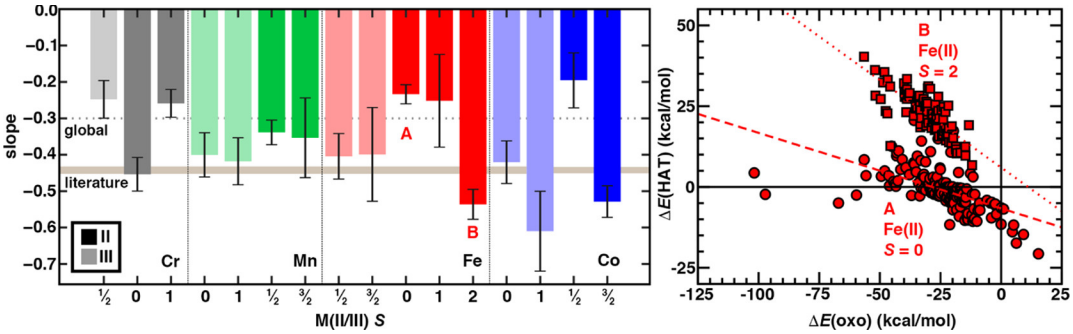

Figure 39. Δ*E*(oxo) vs Δ*E*(HAT) LFER slopes and standard errors obtained across catalysts in 16 metal/oxidation/spin states (left) along with representative data sets used to determine LFER slopes (right). Data is colored by metal, with Cr in gray, Mn in green, Fe in red, and Co in blue, and shown opaque for M(II) and translucent for M(III) resting states. The range of heterogeneous catalysis (i.e., literature) slopes[381, 447] is shown as a shaded light brown area, and the global LFER over all 16 metal/oxidation/spin states studied in the transition metal complex set is shown as a gray dotted line. The data is labeled A for LS singlet Fe(II) and B for HS quintet Fe(II), both left and right, with LFER lines shown at right as dashed and dotted lines, respectively. Reproduced with permission from ref. 457. Copyright 2020 American Chemical Society.

When the guiding descriptor for ligand design is unknown and the computational cost of determining reaction energetics is relatively high, a combination of descriptor-based screening and variation of ligand or metal chemistry in a small-scale screen is often beneficial. These screens are often used to determine the relative magnitude of effects, first by varying metal and then substrate or ligand chemistry. Pickup et al. built a computational database to observe what effects governed the experimental stability of 60 alkyne/vinylidene isomers bound to a transition-metal complex (Figure 40).[470] By comparing metals, substituents on the alkyne, and ligands in the rest of the complex, they determined[470] that the metal played a far more limited role in comparison to the alkyne chemistry in favoring a vinylidene structure. Shiekh et al. studied biomimetic Mo/W CO_2 hydrogenation catalysts to determine the relative magnitude of changing metal oxidation state, functional groups distant from the metal, and axial coordinating ligands (Figure 41).[471] They observed a wide range of feasible H_2 binding energies that could be used to tune the catalysts, with oxidation state playing a pivotal role in comparison to functional group

substitution.[471] Similarly, factors governing ligand redox non-innocence are poorly understood, motivating a screen of how altering functional groups on o-benzoquinone ligands affects the behavior of Co(II) complexes.[472]

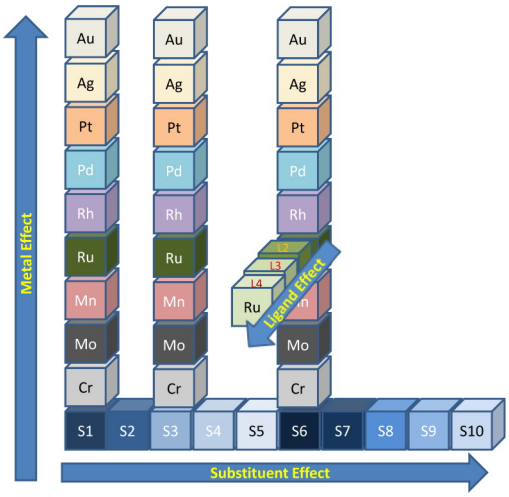

Figure 40. Schematic illustrating the exploration of different variables in analyzing substituent effects on vinylidene/acetylene tautomer stability in transition-metal complexes from experimental data. The multivariate analysis involves analyzing multiple metals for different substituents on the binding ligands along with studying ligand effects on select substituents. Reproduced with permission from ref. 470. Copyright 2014 American Chemical Society.

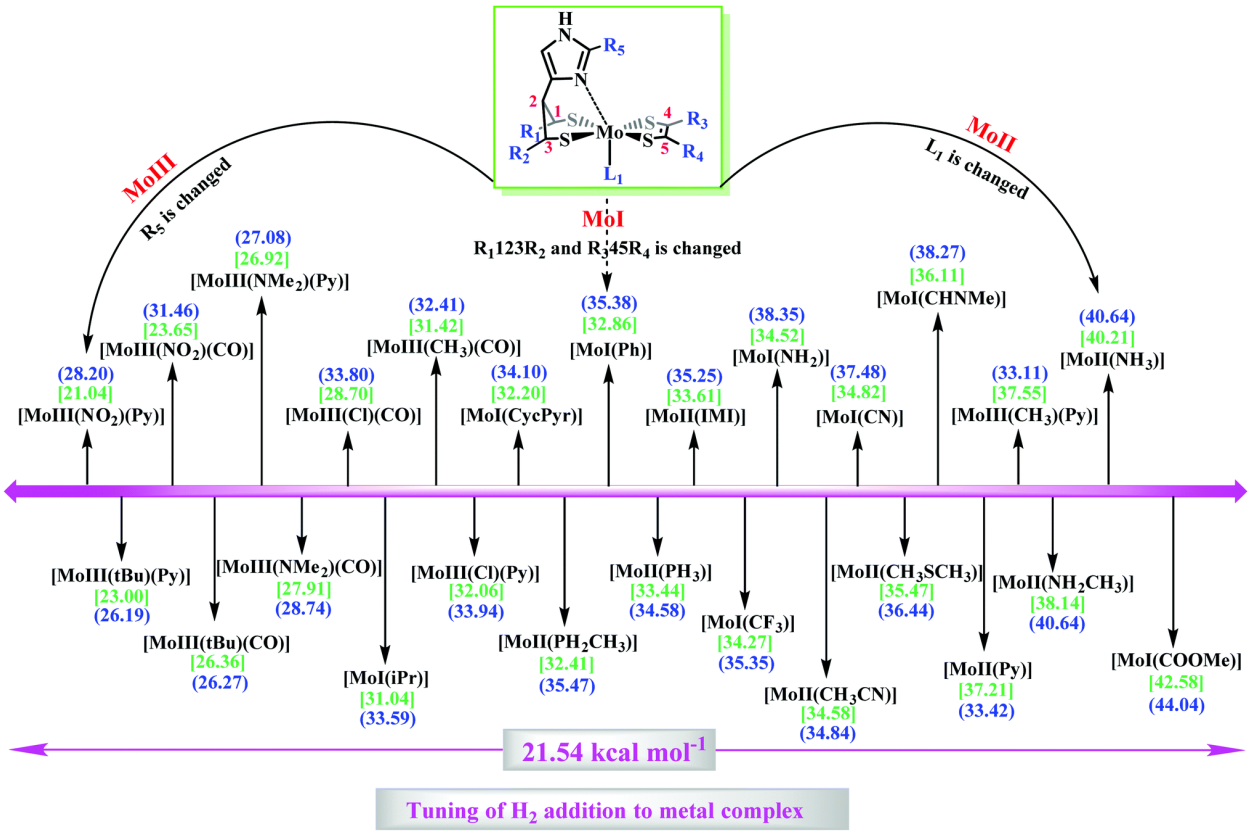

Figure 41. Free energy tuning of H_2 addition (green colored values) over the metal center in Mo-based catalysts that act as formate dehydrogenase mimics along with their free energy of activation values (blue colored values) all in kcal/mol. The choice of ligands is indicated schematically in blue text and arrows highlight the effect of changing substituents versus changing the axial ligand. Reproduced with permission from ref. 471. Copyright 2019 Royal Society of Chemistry.

Such studies are also widely used in inferring and confirming metal- or spin-state-dependent reactivity and mechanistic trends. Examples include comparisons of the role of metal identity, oxidation state, and row on the relative preference for proton-coupled electron transfer versus uncoupled reaction steps[473] and examinations of the influences on reactive intermediate (e.g., metal-fluoro vs. metal-oxo) formation.[474] These studies can also reveal when a descriptor that is pivotal in one class of reactions or catalysts bears less importance in others or when a single descriptor will prove insufficient. For water oxidation in bimetallic complexes, varying the metal identities revealed the numerous possible spin states and their cooperative effect on

relative reaction barriers or favored mechanisms, giving rise to complex patterns that could not be captured by a single chemical descriptor.[475] Similarly, while bite angle is often invoked in complexes with bidentate phosphine ligands, for reductive elimination at Pd(II) catalysts, the effect of electron-withdrawing groups on the ligands was found to be more significant.[476] Iterative ligand design has also been demonstrated in cases where key aspects of an established catalyst are preserved, such as the presence of a multidentate ligand, while designing other less exhaustively explored ligands to propose improvements beyond known best-in-class catalysts.[457, 477] A key caveat in computational design is that ligand flexibility can alter reaction mechanism or influence predicted activation energies, often requiring a tight coupling between computation and experiment to realize such suggestions.[478]

While the design of transition-metal complexes for homogeneous catalysis has been the focus of much of the effort in computational chemistry, the rise of porous metal–organic frameworks or zeolites for gas separation has also spurred interest in mechanistic and iterative design with some parallels to catalyst design. Given the distinct scale and nature of MOF properties, the descriptors favored are often more focused on geometry or topology, as we will describe in more detail in Sec. 4.4. For MOFs, iterative design has been carried out by holding fixed metal secondary building units (SBUs) expected to be needed for stability and key features (e.g., open sites for gas binding) while replacing and varying linker molecules to adjust for other properties such as surface area for gas storage.[479-480] Alternately, researchers can study the effect of substituting metal sites in well-established frameworks to reveal metal-dependent gas adsorption without confounding factors from varied pore geometry.[481] Analysis of the relationship between the resulting topology of the material and properties for gas adsorption has been used to reveal design principles and notably to identify a hypothetical material that was

experimentally verified to have high gas uptake characteristics.[480]

4.3. Unsupervised Learning.

The concept of distance in chemical space plays an essential role in the interpretation, visualization, and manipulation of data sets and machine learning models (Sec. 4.4). Principal component analysis (PCA) was among the first methods to leverage this concept and thus has been the most widely employed technique.[3] PCA reduces a high-dimensional feature set to the set of normal vectors that encode as much of the variance in the data as possible. PCA has been widely applied to analyze trends in structural properties of common moieties (e.g., Mo=O) in the CSD with changing ligand chemistry.[482] Orpen, Fey, and coworkers have developed the most extensive PCA-based analysis of CSD structures, including through the application of PCA to electronic descriptors from DFT on DFT-augmented "ligand knowledge bases" (LKB).[3, 483] They have exploited PCA on these LKBs to analyze relationships among ligand classes from the CSD (e.g., monodentate[359, 392, 484] and bidentate[485] phosphine or carbene[486] ligands). For these curated sets, Fey and coworkers computed a range of steric and DFT-derived descriptors, and they then simplified the visualization of the ligand space by PCA dimensionality reduction (Figure 42). The suitability of the PCA visualization was highlighted by i) good separation and clustering of families of ligand types, ii) the two components typically represented the majority (> 65%) of the data variance, and iii) the descriptors were used to build predictive multiple linear regression (see Sec. 4.2) models (Figure 42). Using PCA on other LKBs, they have also related PCs back to key ligand variations, such as functional groups[470], and steric or electronic factors[3, 406].

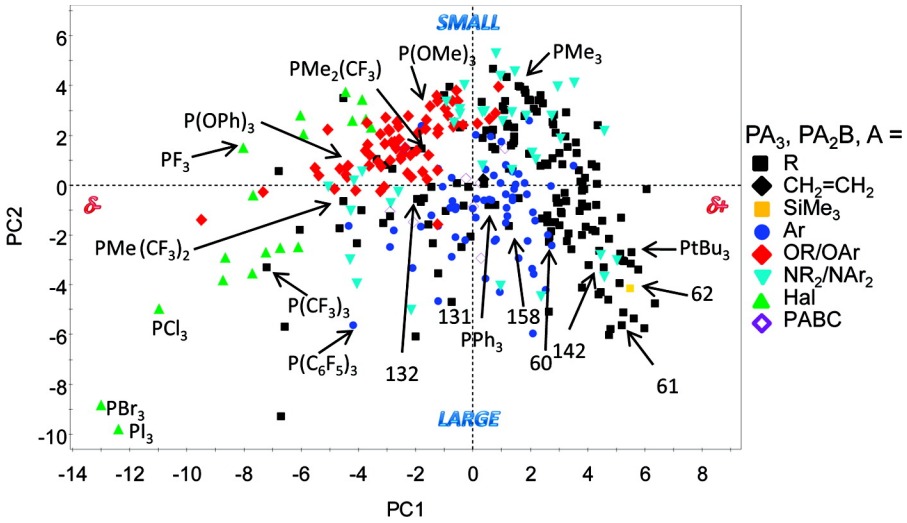

Figure 42. Principal component analysis of the LKB-P ligand set. The first two principal components from a 28-descriptor set are shown and capture 65% of the variance in the data. PC1 is dominated by steric and σ-bond electronic descriptors along with bonding terms, whereas PC2 has a greater emphasis on steric effects. Reprinted with permission from ref. 484. Copyright 2010 American Chemical Society.

As another early demonstration in the 1990s, Beyreuther et al. demonstrated both structural-descriptor-based multiple linear regression and unsupervised learning approaches, including PCA and self-organizing maps, which are a type of ANN used for unsupervised dimensionality reduction (here, 10 neurons in a Kohonen map).[487] Using such techniques, they were able to predict tripodal complex conformations based on local descriptors such as bite angles using 82 CSD complexes.[487] Within heterogeneous catalysis, PCA has recently been applied to analyze which DFT binding energy descriptors explain experimental observations[488] or to guide search for new catalysts with improved properties[489]. Rothenberg and coworkers highlighted the importance of dimensionality reduction in interpreting the three high-dimensional spaces most relevant to catalyst optimization: continuous feature space, discrete ligand space, and the ultimate property space being optimized (Figure 43).

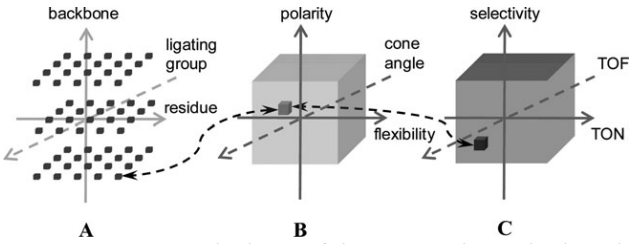

Figure 43. Depiction of how catalyst design involves three types of multidimensional space: discrete catalysts (space A), continuous descriptor values from representations (space B), and figures of merit (space C). Reproduced with permission from ref. 489. Copyright 2010 Royal of Chemistry.

While PCA is primarily carried out on feature spaces, dimensionality reduction of the space of molecular properties/activities for a set of compounds can also be useful. Such an approach has been used to identify activity cliffs (i.e., abrupt changes in properties with small changes in features) to reveal discontinuities in structure–activity relationships in therapeutic drugs.[490] When points are distant in a feature space but similar in property, they also provide evidence that shallow, transparent models (e.g., MLR or KRR) would struggle to use these features to obtain a suitable prediction of properties. Thus, selection of optimal features by PCA has been carried out[491] as a feature selection technique, including in an iterative fashion[492], for ML model training on data sets, as discussed further in Sec. 4.4.

PCA remains the most widely used approach and performs best when visualizing the first 2–4 PCs that capture the most data variance (i.e., ideally the majority). Alternative dimensionality reduction techniques have recently been proposed, including t-distributed stochastic neighbor embedding (t-SNE)[493] and uniform manifold approximation and projection (UMAP)[494-495], and these techniques have seen increased use in materials and catalyst discovery. Both aim to generate the most informative dimensions for visualizing data differences. The t-SNE algorithm approximately preserves pairwise distances between each data point in a set[493], but it is stochastic in nature and thus cannot be used as a tool to visualize the effect of adding

new data. UMAP aims to preserve local structure in a manner similar to t-SNE with slightly relaxed requirements on the pairwise separation that nevertheless allows for data addition. Self-organizing maps[487] and other clustering algorithms have been used in chemical discovery, for example k-means clustering is often used in large chemical spaces to select diverse leads for property optimization (see Sec. 5.2). An additional use of dimensionality reduction techniques is in the visualization of the latent space of deep learning models or to identify when new data points (i.e., either in feature or property space) are distinct from prior training data (see Sec. 4.4).

4.4. Deep Learning and Artificial Intelligence.

In recent years, more highly parameterized supervised and semi-supervised models have been employed in transition-metal chemistry property prediction and for molecular discovery. These include kernel methods (e.g., Gaussian processes and kernel ridge regression or KRR) as well as deep artificial neural networks (ANNs) with multiple hidden layers. In this section, we review some of the representations used for such models (Sec. 4.4.1) and discuss applications of supervised regression (Sec. 4.4.2) and classification (Sec. 4.4.3) models to transition-metal chemistry. In comparison to small-molecule organic chemistry, smaller data sets and more varied properties introduces distinct challenges for ML in transition-metal catalysis.

4.4.1. Machine Learning Representations.

The representations introduced as QSPR descriptors in Sec. 4.2.3 (e.g., heuristic Hammett parameters of ligands) can also be used[491, 496] in conjunction with deep learning models. Here, we briefly review additional recent representation developments that are more widely used in supervised deep learning models. For open-shell transition-metal chemistry, the limitations of force fields and semi-empirical theories were discussed in Sec. 2.2. As a result, force field or semi-empirical pre-optimization widely exploited in organic chemistry machine

learning to predict DFT-level energetics[236] are typically avoided in transition-metal chemistry. As an example of geometry-free representations in transition metal chemistry, Janet et al. proposed[54] an *ad hoc* set of 25 mixed continuous discrete metal-local (MCDL-25) descriptors for supervised learning (i.e., ANN) prediction of properties of octahedral complexes with mid-row transition metals. The MCDL-25 descriptors included one-hot encoded metal and coordinating atom identities, metal-local ligand atom electronegativity and bond order, all selected via LASSO feature selection on the data set. The most non-local measure included was the truncated Kier shape index[497] that measured the rigidity and branching of the ligands via the number of two-bond paths formed by the atoms.

Graph-based representations such as Moreau-Broto autocorrelations[498] were first introduced 40 years ago for use in organic chemistry cheminformatics. In the mid-2000s, Burello et al. proposed[499] path-based descriptors tailored for bidentate phosphine ligands to be used in machine learning models (Figure 44). Examples of these tailored, *ad hoc* descriptors included the minimum path distance between phosphorus atoms, the number of rotatable bonds, and the mass of atoms within three bonds of the chelating phosphorus atoms (Figure 44). Systematic sets of graph-based descriptors were proposed by Janet et al.[500] as extensions of Moreau-Broto autocorrelation (AC) functions[52, 498, 500-501] to develop the revised autocorrelation (RAC) representation for transition-metal chemistry. Standard ACs have the form:

$$P_d = \sum_i \sum_j P_i P_j \delta\left(d_{ij}, d\right) \tag{1}$$

where P is a heuristic, tabulated property and d_{ij} is the distance separating two atoms on the molecular graph. Five heuristic properties typically used in ACs are: i) nuclear charge, Z; ii) Pauling electronegativity, χ, motivated by its importance in MCDL-25[54]; iii) topology, T, which is the atom's coordination number; iv) identity, I, that is 1 for any atom, as suggested in Ref. [501];

and v) covalent atomic radius, *S*. Other heuristic properties that have been used include the atomic polarizability.[384] These atomic properties can be correlated but recent analysis has indicated they can be useful to indicate distinct aspects of elements in the periodic table.[502]

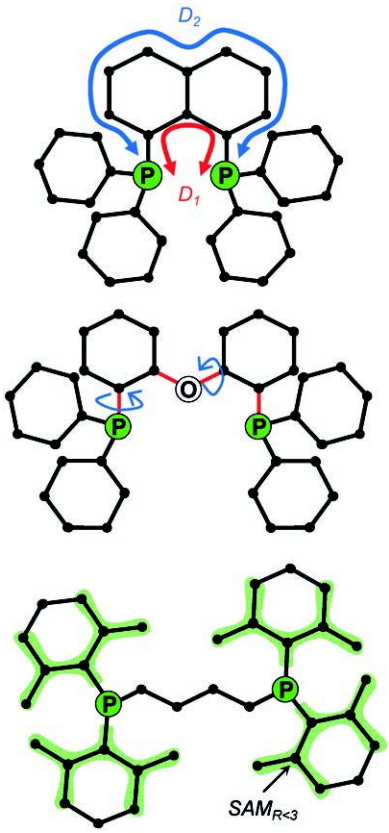

Figure 44. Representative topological descriptors computed on backbone and R groups of bidentate phosphorus ligands. (top) The red path indicates the minimum P_1-P_2 connectivity path, D_1. The alternative longer P_1-P_2 path, D_2, is shown in blue. (middle) The number of rotatable bonds is indicated in red. (bottom) The R group descriptor, $SAM_{R<3}$, is shown in green and corresponds to the sum of mass of atoms that are within three bonds of the P atoms. Reproduced with permission from ref. [499]. Copyright 2005 WILEY‐VCH Verlag GmbH & Co.

RACs are an extension of ACs to include both products and differences on the molecular graph. In order to capture the centrality of the metal and its coordination environment in determining properties of a transition-metal complex, the convention of RACs also introduced the selection of scope[52], i.e., by choosing which atoms to sum over, such as equatorial versus axial ligands (Figure 45). RACs also include metal-centered (mc) or ligand-centered (lc) RACs

in which one of the atoms in the AC expression is always a metal center or ligand coordinating atom, respectively (Figure 45). Finally, a chief distinction of RACs is the inclusion of property differences rather than products, which are non-trivial only for a minimum d of 1 and apply only to lc or mc RACs:

$$ {}^{lc/mc}_{ax/eq/all}P'_d = \sum_{i}^{lc \text{ or } mc\, scope} \sum_{j} \left(P_i - P_j\right)\delta\left(d_{ij}, d\right) \tag{2} $$

where scope is chosen to be all ligands or averaged (i.e., equatorial or axial). Extension of RACs to other materials has also motivated definition of new scopes, as described next. In total, there are six types of standard ACs and RACs (i.e., full averaged over all, axial, or equatorial; lc averaged over axial or equatorial; and mc averaged over all). For a given cutoff d and five heuristic properties, there are $42d$+10 RACs. Observations on small organic molecule data sets (i.e., QM9[503]) has typically motivated[500] the truncation of the expansion of standard ACs and RACs to a depth of three bond paths (i.e., d = 3, RACs). Unlike MCDL-25 or other metal-nearsighted representations, information about the full molecule is still included in the expansions of ACs and RACs. However, specific properties of atoms more than three bonds apart are not included in most RAC descriptors. To this set of graph-based descriptors, a number of one-hot descriptors are added[500], including ligand denticity and spin or oxidation state of the metal. Eliminating RACs that are invariant in the data set typically (e.g., for mononuclear octahedral transition metal complexes) produces a 155-dimensional feature set known as RAC-155.[500]

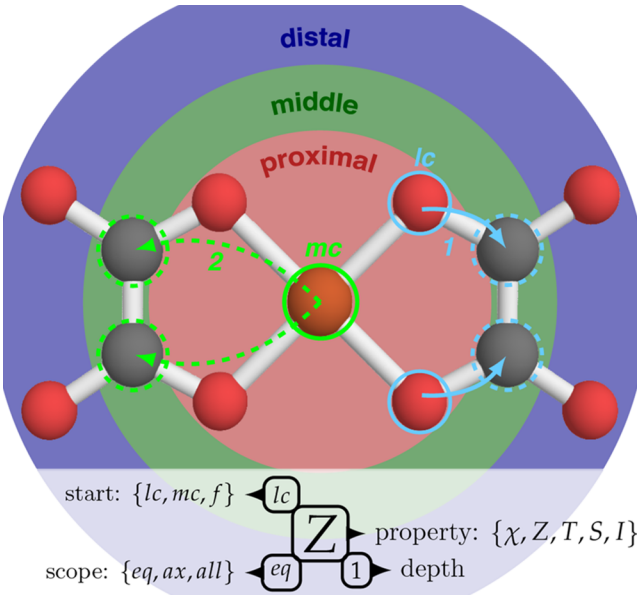

Figure 45. Schematic of RACs in the equatorial plane of an iron octahedral complex with two equatorial oxalate ligands shown in ball-and-stick representation. Regions of the molecule used to classify descriptors are designated as proximal (metal and first coordination shell, in red), middle (second coordination shell, in green) and distal (third shell and beyond, in blue). Light green circles and arrows depict terms in a 2-depth *mc* RAC, and the light blue circles and arrows depict terms in a 1-depth *lc* RAC. Reproduced with permission from ref. [500]. Copyright 2017 American Chemical Society.

A number of three-dimensional representations have been widely used in organic chemistry and occasionally applied to open-shell transition-metal chemistry, and we briefly review some of them here. One popular whole-molecule representation is the Coulomb Matrix (CM) [504]. The off-diagonal elements of the CM are the products of the *i*th and *j*th atoms' nuclear charges, Z, scaled by their separation (i.e., Z_iZ_j/r_{ij}), and the diagonal terms are heuristically fit to $0.5Z_i^{2.4}$. The CM representation has strong size-dependence, requiring padding with zeros up to number of atoms in the largest molecule studied, which can challenge chemical discovery efforts where the system size is not known *a priori*. This size-dependence is particularly problematic for transition metal complex properties (e.g., spin, catalytic) that depend much more on the metal-local environment and less on the size. The CM also depends on the order of atoms in the molecule. Alternative 3D descriptors have been proposed and used particularly in organic

chemistry. These include the Bag of Bonds (BoB) descriptor, which groups interactions by the pairwise atom-types and distances involved and computes CM-like interactions among them[505]. Other functions have explicit distance decay around local atomic environments, making them amenable to development of force fields not discussed in this review, and avoid the strong size-dependence of the CM. These include the smooth overlap of atomic position (SOAP)[506] representation, the SLATM[507] representation, and the many-body tensor representation (MBTR)[508]. Since all of these representations require geometries before their calculation, their utility in transition metal chemistry is limited to closed shell transition metal complexes where low-cost (e.g., semi-empirical) models provide an adequate initial geometry.

Distance-dependent modifications to RACs have also been proposed to partially account for structure dependence.[509] The Coulomb-decay RAC (CD-RAC) representation was proposed for prediction of properties of both equilibrium and distorted molecular structures[509] as follows:

$$P_{d,\mathrm{CD}} = \frac{1}{n}\begin{cases} \sum_{i}^{n}\sum_{j}^{n}\frac{P_i P_j}{r_{ij}}\delta\left(d_{ij},d\right), & d>0 \\ \frac{1}{2}\sum_{i}^{n}P_i^{2.4}, & d=0 \end{cases} \tag{3}$$

where the $d = 0$ term adopts the functional form of the CM diagonal, $d > 0$ terms incorporate explicit internuclear dependence, and both types of terms are normalized by the number of atoms, n, to limit size dependence. The heuristic properties encoded by CD-RACs can be the same as standard RACs, but modified heuristics have also been proposed[509] to increase transferability, including the number of valence electrons (e.g., five for N and six for O) and the number of bonds the neutral element should form based on the octet rule (e.g., four for C and three for N).

Tailored representations have been developed for periodic MOFs. Geometric descriptors

of pore volume and surface area[510-512] or energetic descriptors from evaluation with probe molecules[513] have been employed in gas separation models. An extension to SMILES or InChI has been proposed[514] to generalize these string expressions for more universal naming and categorization of MOFs (Figure 46). Both string types preserve the domains of SBU and linker that are normally used in describing MOF chemistry (Figure 46). Molecular graph-based RACs[498, 500-501] have been generalized for periodic MOFs[384] with an expanded heuristic set that included atomic polarizabilities. MOFs are divided into the metal node SBUs, linkers, and functional groups (FGs). Metal-centered RACs are computed on the SBUs, linker-only RACs are computed both over the full linker and centered on coordinating atoms, and FG (i.e., non-C or H-atoms in the linker) RACs are computed centered on the FG heavy atoms.[384]

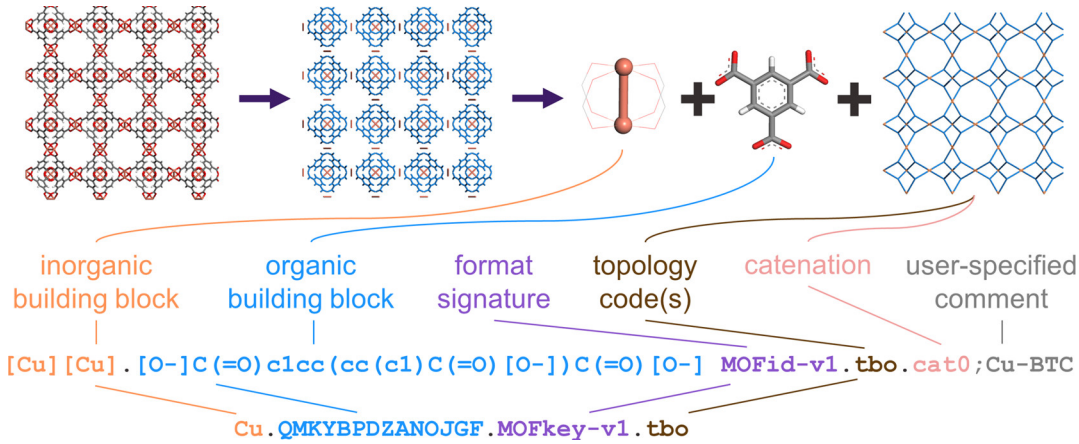

Figure 46. Algorithms are developed to break MOFs down into chemical components and an underlying topological net. This is then used to generate text-based MOFid and MOFkey strings. An example of identifying the Cu-BTC MOF using the SMILES-derived MOFid format and InChI-based MOFkey format is shown. Reproduced with permission from ref. [514]. Copyright 2019 American Chemical Society.

Representations based on atom-counting, compositional[515] features or one-hot encoding of bond/atom types[516-517] and functional groups only[510] have also been developed for MOFs. While 3D geometric representations have been less widely employed in the context of MOFs, atomic-property-weighted radial distribution functions (AP-RDFs)[518] have been proposed:

$$\mathrm{RDF}^{P}(R) = f \sum_{i,j}^{\text{all atom pairs}} P_i P_j e^{-B(r_{ij}-R)^2} \quad (4)$$

where B is a smoothing parameter, r_{ij} is the minimum distance of atoms within a certain radius R inside the unit cell, and f is a normalization factor. These descriptors are evaluated over all pairs in the entire MOF unit cell, and the properties P evaluated typically include tabulated electronegativity, polarizability, and van der Waals volume.

A challenge in representations for inorganic bonding is the extent to which models should be transferable across electronic states, including molecular charge or metal oxidation and spin states, as well as across groups or rows of the periodic table. While this remains an outstanding challenge for machine learning representation development as a whole, it is particularly relevant for transition-metal chemistry. One approach that researchers have adopted to sidestep this challenge is to use electronic-structure-based descriptors as inputs to ML models, typically improving transferability but at the computational cost of an electronic structure calculation as part of ML model evaluation. Many such descriptors are similar to those described for QSPRs (see Sec. 4.2.1). In addition, tailored representations applied to machine learning in transition-metal chemistry have included properties of the metal-centered gradient[519] and metal charge[519-521], orbital coefficients and matrix elements from approximate electronic structure[522-523], HOMO–LUMO gaps[491, 521], vibrational frequencies[521], or components from energy decomposition analysis[524].

4.4.2. Supervised Learning Regression Models.

Although supervised ML has been receiving increasing interest in recent years, the earliest ML regression models applied to transition-metal chemistry were developed nearly 20 years ago.[491, 520, 525] Burello and Rothenberg used QM descriptors (e.g., frontier orbital energies and structural descriptors) to train models to predict experimental catalytic activity of Pd-based

cross-coupling catalysts, including with awareness of solvent effects. Available training sets were around 500 experimental results (e.g., TOF and TON) and could be predicted well by the models (Figure 47).[491, 525] They showed that while a range of models were suitable (e.g., decision trees, linear regression, ANNs), considerably smaller residuals were obtained with the non-linear ANN models (Figure 47).[491] Once trained, the ANN models could then be used to screen a large space of candidate catalysts at the cost of only DFT-level evaluation[491] (see Sec. 5).

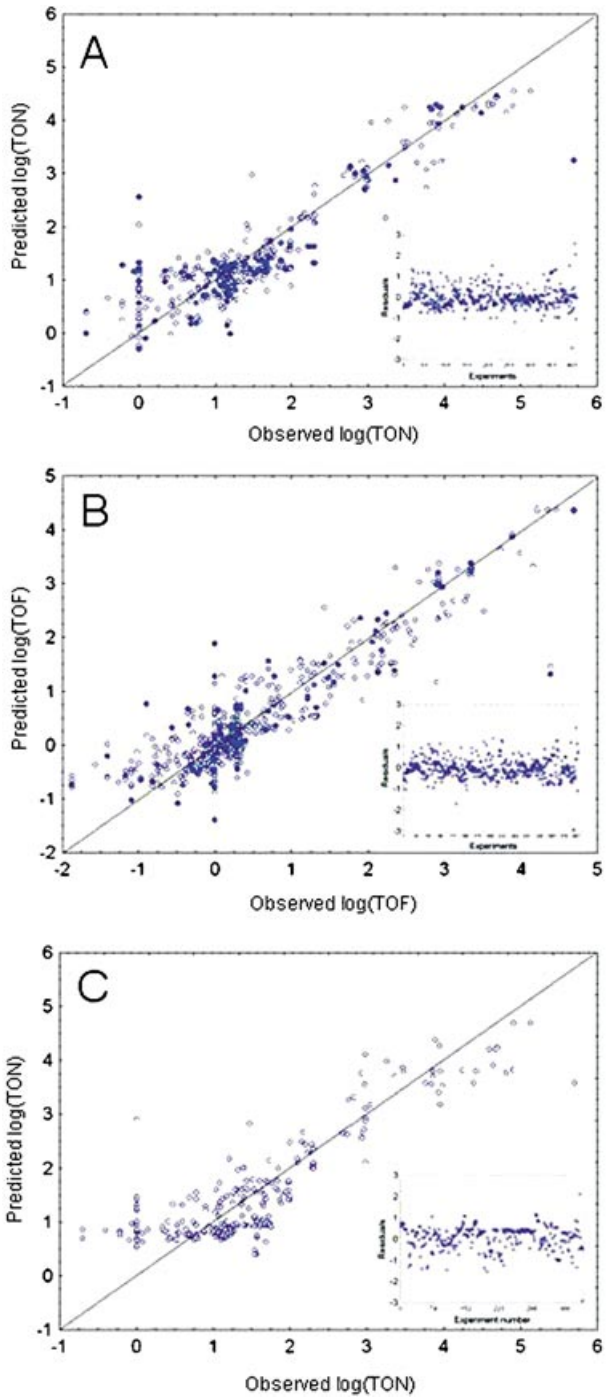

Figure 47. Predicted vs observed (A) TON and (B) TOF values for Heck reactions obtained by ANN regression, and (C) TON by multiple linear regression. The inset shows the residuals for each model. Training data is shown as solid circles and test data is shown as open circles. Differences are observed in the scattered residuals for A and B in comparison to the more structured residuals for C, suggesting that linear regression fails to capture higher-order effects for some reactions. Reproduced with permission from ref. 491. Copyright 2004 WILEY‐VCH Verlag GmbH & Co.

Janet et al. trained ANNs to predict the spin-splitting energies of mid-row transition-metal complexes.[54] They used an ad hoc, metal-local descriptor set (MCDL-25, see Sec. 4.4.1) to increase transferability of the models across a range of transition-metal complex sizes.[54] The performance of the models was balanced across the Cr–Ni M(II)/M(III) complexes studied.[54] The ANN models accurately predict trends in spin-splitting with ligand field strength (e.g., as in the spectrochemical series) for the full range of M(II)/M(III) (M = Cr–Ni) metals studied (Figure 48).[54] The models were trained on DFT data (ca. 1000 complexes using common ligands from inorganic chemistry) from a range of Hartree–Fock exchange fractions to account for the sensitivity of spin-splitting predictions to DFT functional choice. Models were both trained to predict the exchange sensitivity and to enumerate different lead compounds with changed exchange fraction (see Figure 18).[34, 54] ANN test set performance (MAE: 2.5 kcal/mol, RMSE: 3.0 kcal/mol) was generally within 1 credible interval, which they estimated from Monte Carlo dropout[526], of the DFT ground-truth result (Figure 48). The models also outperformed whole-molecule descriptors (e.g., the Coulomb matrix) by at least an order of magnitude for the same data set size.[54] Tests on out-of-distribution (i.e., larger complexes from the CSD) complexes however revealed an increase in errors due to differences from training data.[54] Given the high degree of feature engineering in MCDL-25, Janet et al. showed feature-space distance could be used to distinguish when ANN performance would be poor on out-of-distribution points.[54]

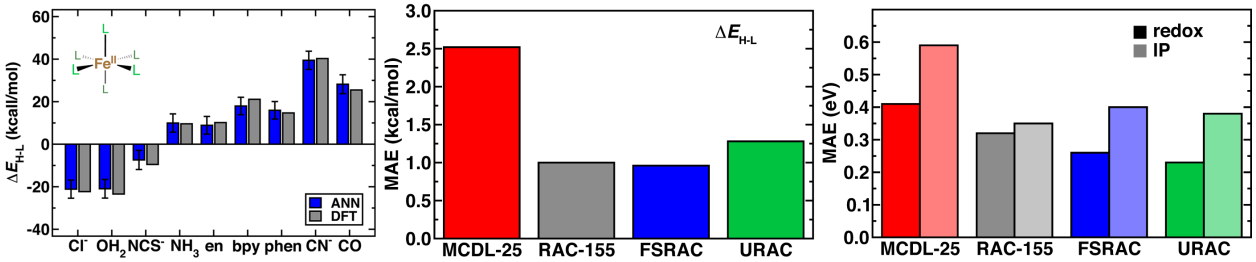

Figure 48. (Left) High-spin/low-spin splitting (ΔE_{H-L}, in kcal/mol) for representative ligands from left to right by increasing spectrochemical series ligand field strength obtained from the MCDL-25/ANN with credible intervals and from DFT training data. (Middle and right) MAEs for ΔE_{H-L} (in kcal/mol) and redox or adiabatic ionization potential (in eV) MAEs for both

MCDL-25/ANN and RAC KRR models (full RAC-155, feature-selected, FS RAC, on a single property, and URAC selected on multiple properties). Adapted with permission from ref. 34. Copyright 2019 American Chemical Society.

Janet et al. trained ANN[527] and KRR[500] models to predict spin-splitting energies with the graph-based, RAC representation (see Sec. 4.4.1). The RAC-155/ANN[527] or RAC-155/KRR[500] models improved upon the performance of MCDL-25 (Figure 48).[500] The best-performing RAC-155 models[500] achieved sub-kcal/mol test MAE on comparably sized data sets to the MCDL-25/ANN model (Figure 48). Feature selection preserved the good performance of the RACs while reducing the number of descriptors to a size comparable to MCDL-25 (Figure 48).[500] PCA of this feature-selected subset indicated that by eliminating non-local features from the full RAC set, complexes with distinct sizes due to distal variation but otherwise comparable spin splitting were more proximal in PC space of the reduced representation (Figure 49).[500] In comparison, the full RAC-155 or a poor-performing whole-molecule representation placed the same points very far from each other, making it more challenging for ANNs to learn the structure–property mapping (Figure 49).[500] An approach to predict dynamic correlation energies (e.g., MP2 corrections to HF) from KRR models trained on orbital-based descriptors was also demonstrated on transition-metal complexes from this dataset.[523]

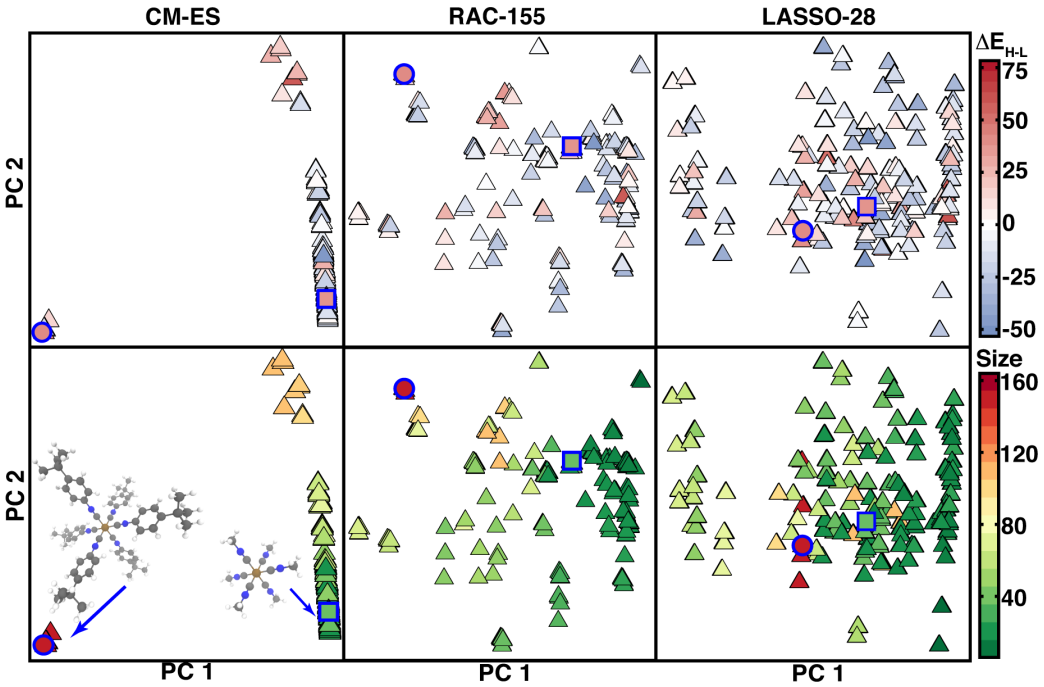

Figure 49. Projection of a spin-splitting data set onto the first two principal components (arbitrary units) for the Coulomb matrix eigenspectrum (CM-ES, left), full revised AC set (RAC-155, center), and the LASSO-selected (i.e., on spin splitting) RAC subset (LASSO-28, right). The PCA plots are colored by DFT-calculated spin-splitting energy (top, scale bar in kcal/mol at right) and size (bottom, scale bar in number of atoms at right). Ball and stick structures of representative complexes are inset in the bottom left, and the associated data points are highlighted with a blue circle and square in each plot. Reproduced with permission from ref. [500]. Copyright 2017 American Chemical Society.

Other applications focused on open-shell systems have additionally used geometry-dependent descriptors. ML models (i.e., ridge regression) have been trained to interpolate the potential energy surface for mononuclear Fe single-molecule magnet (SMM) candidates.[528] These models reduced the computational cost to identify optimal structural properties in designing single-ion magnetic anisotropy.[528] Using experimental data sets, LASSO models have also been developed to design SMMs with a large isothermal magnetic entropy change.[529] Based on a set of 60 data-mined experimentally synthesized SMMs, heuristic descriptors, such as size, number and type of metal ions, were used as inputs to the LASSO model for both feature selection and model prediction.[529] Gaussian processes have also been used to predict the

exchange–spin coupling between copper centers in a small set (ca. 250) of dicopper complexes, with good performance achieved on a range of distance-dependent descriptors from heuristics (e.g., Cu-distance) to more sophisticated ML descriptors (e.g., SOAP, see Sec. 4.4.1).[530]

As a demonstration of the benefit of avoiding structure-based descriptors, graph-based descriptors have been used[34, 54, 499-500] to rapidly predict geometric properties relevant to transition-metal catalysis and to overcome limitations in the cost and accuracy of physics-based models (see Sec. 2.5). The bite angle or flexibility (i.e., range of low-energy bite angles) of bidentate phosphine ligands were predicted by ML models trained on a tailored set of topological descriptors.[499] Given a small training set of 65 ligands from the CSD and limited number (ca. 19 after PCA selection) of input features, these early ANN demonstrations consisted of a small number of nodes (ca. 4–7) and only one hidden layer.[499] Despite the simplicity, these models exhibited good performance on validation data (R^2=0.84–0.9).[499]

Janet et al. trained ANNs to predict DFT-derived equilibrium bond lengths of transition-metal complexes in a spin- and oxidation-state-dependent manner.[54, 500] The data set consisted of 1,350 hybrid-DFT-geometry-optimized mononuclear octahedral transition-metal complexes with mid-row, 3d transition metals (Cr–Ni) and common ligands used in inorganic chemistry. Using the sparse, *ad hoc* MCDL-25[54] descriptor set to train a fully connected ANN with two hidden layers, bond lengths were predicted to 0.02–0.03 Å MAE on set-aside test molecules for both low-spin and high-spin states. ANNs trained with 2–3 hidden layers on a 155-dimensional set of RACs plus denticity and oxidation/spin state information[500] improved these test errors to sub-0.01 Å on average (Figure 2).[34] These models were implemented in the molSimplify[53, 55] code's structure generation module[531] including in a web frontend[532] to enable rapid structure prediction by force fields for organic components and the ANNs for the metal–organic bond length (Figure

2). KRR models trained on feature-selected subsets achieved comparable errors to the ANNs.[54, 500] Consistent with the symmetry of transition-metal complexes in the training data, the RAC-155/ANN models predict both an averaged equatorial bond length and two independent axial bond lengths. As a result, these ML models encoded the metal and ligands where Jahn–Teller distortion will be present by accurately predicting asymmetry in these bond lengths.[34] In a related approach, neural network potentials (NNPs) have been developed for some solid-state materials that can exhibit redox or spin state switching, in which case, changes in bond length for this single NNP model have been ascribed to changes in oxidation or spin state.[533]

ML regression models (e.g., ANNs and Gaussian process regression or GPRs) have been trained on moderately sized (ca. 2000 points) data sets (e.g., the MNSol database[534]) for predicting solvation free energies[535-536] (i.e., in numerous solvents) and redox potentials[537] predominantly for organic molecules. ANNs trained to predict DFT-derived transition-metal complex redox potentials have demonstrated good performance (ca. 0.2–0.3 eV) even on small (ca. 200–300 points) data sets (Figure 48).[34, 500] Models trained on RACs or on MCDL-25 have been used to predict the adiabatic ionization potential as well as the solvent- and thermo-corrected redox potential of mid-row transition-metal complexes (Figure 48). As in the case of spin splitting, the best performance has been obtained with feature-selected subsets of RACs[500], and performance for predicting redox potential and ionization potential were comparable (Figure 48).

Using a similar dataset, Nandy et al. developed both KRR and ANN models on RACs for predicting the HOMO, LUMO, and HOMO–LUMO gap of mid-row transition-metal complexes.[55] They observed comparable (ca. 0.2–0.3 eV) model errors from a RAC-155/ANN and KRR models trained on feature-selected subsets.[55] For the KRR models, they used random-

forest-ranked recursive feature addition (RF-RFA) to a KRR model both to improve model performance and to analyze important features.[55] Liu et al. built models across a larger set of 4500 transition-metal complexes with a greater diversity in ligand chemistry and size, achieving comparable errors on HOMO level and HOMO–LUMO gap prediction.[538] They generalized[538] this ML approach to fractional occupation number (FON) DFT properties to accurately predict the degree of MR character with a RAC-155/ANN (Figure 50). Enumeration of a large space of 187k candidate transition-metal complexes was used to identify ligand and metals that corresponded to strong MR character (Figure 50).[538]

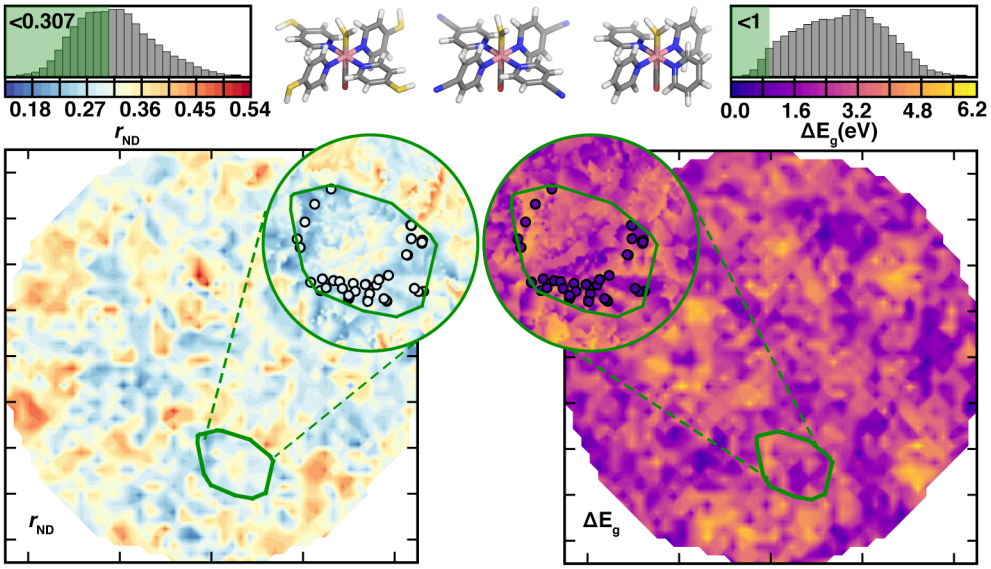

Figure 50. Dimensionality reduction with t-distributed stochastic neighbor embedding (t-SNE) of ANN-predicted descriptions of MR character (r_{ND}, unitless) and HOMO–LUMO gap (right, ΔE_g in eV) highlights differences in the two quantities over a theoretical space of 187k complexes. The 1D histograms of properties are shown at top with low MR character and small-gap complexes shaded in green. A convex hull of a family of complexes with functionalized pyridinyl ligands that have low MR character but small HOMO–LUMO gap is shown inset. Reprinted with permission from ref. [538]. Copyright 2020 American Chemical Society.

Redox potential and frontier orbital properties are also frequently used to rationalize and design catalysts. Nandy et al. trained ML models on RACs to predict open-shell transition-metal catalyst reaction energies (i.e., oxo formation) for methane-to-methanol conversion.[468] The

models achieved good accuracy (ca. 4 kcal/mol) for predicting the reaction energetics, enabling both enumerative study of 37k catalysts as well as screening for design principles. This combined ML model prediction and feature analysis approach with RACs has also been demonstrated on iridium catalysts for H_2 splitting (Figure 51).[539] In this case, RAC-trained models performed well for predicting barrier heights, and feature analysis revealed opportunities for design of catalysts with lowered barriers (Figure 51).[539]

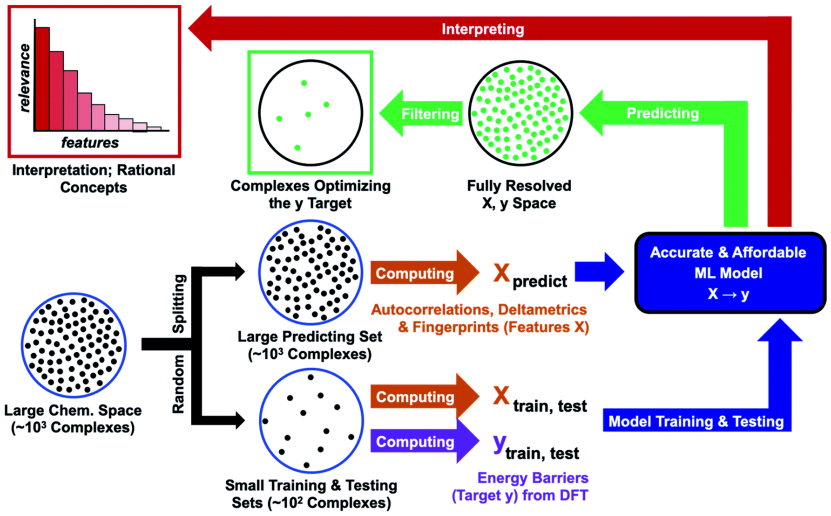

Figure 51. Approach to catalyst screening that uses RACs[500] to accelerate both prediction of DFT-calculated reaction energetics and extraction of design principles for accelerated catalyst discovery.[500] Reproduced with permission from ref. [539]. Copyright 2020 Royal Society of Chemistry.

The HOMO level is often suggested[448] as a heuristic in methane-to-methanol catalyst screening, but ML models have revealed a poor correlation between these two quantities when the range of metals and ligands considered is large.[468] Analysis of RF-RFA-selected features with KRR models for oxo formation energy and HOMO level prediction revealed differences in feature importance for these two properties (Figure 52).[468] The HOMO level depended much more on metal-distant RACs, highlighting its expected size dependence, whereas the oxo formation energy was metal-local (Figure 52).[468] The selected features for oxo formation energy shared more commonalities with those essential for predicting spin splitting, which was

consistent with observations of strong spin state dependence in catalyst reactivity (Figure 52 and see Sec. 4.2). Despite some limitations for large-scale screening, HOMO prediction models have been proposed for use in catalyst design involving a single metal and oxidation state. Chang et al.[496] trained ML models to predict the HOMO level for tungsten benzylidyne photoredox catalysts since this quantity has been related to experimental redox potentials (Figure 53). After comparing several ML model types (e.g., LASSO, Gaussian processes, and ANNs) trained on Hammett parameter heuristics, ANNs were observed to perform best at predicting DFT-level HOMO energetics (Figure 53).[496]

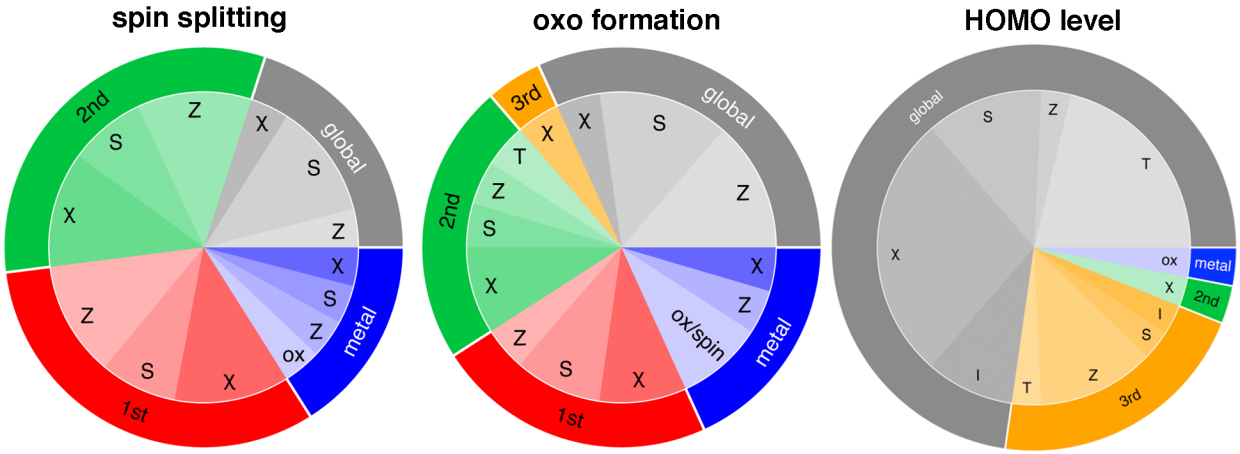

Figure 52. Pie charts of the spin-splitting-selected features (26, left) compared to features selected for oxo formation (22, middle), and the HOMO level (33, right). Features are grouped by the most distant atoms present: metal in blue, first coordination sphere in red, second coordination sphere in green, third coordination sphere in orange, or global features in gray. Within each distance category, the property (i.e., χ, S, T, Z, or I) is also indicated, and oxidation state (ox) and spin are assigned as metal-local properties. Oxo formation features are more similar to those for spin splitting than the HOMO level. Adapted with permission from ref. 468. Copyright 2019 American Chemical Society.

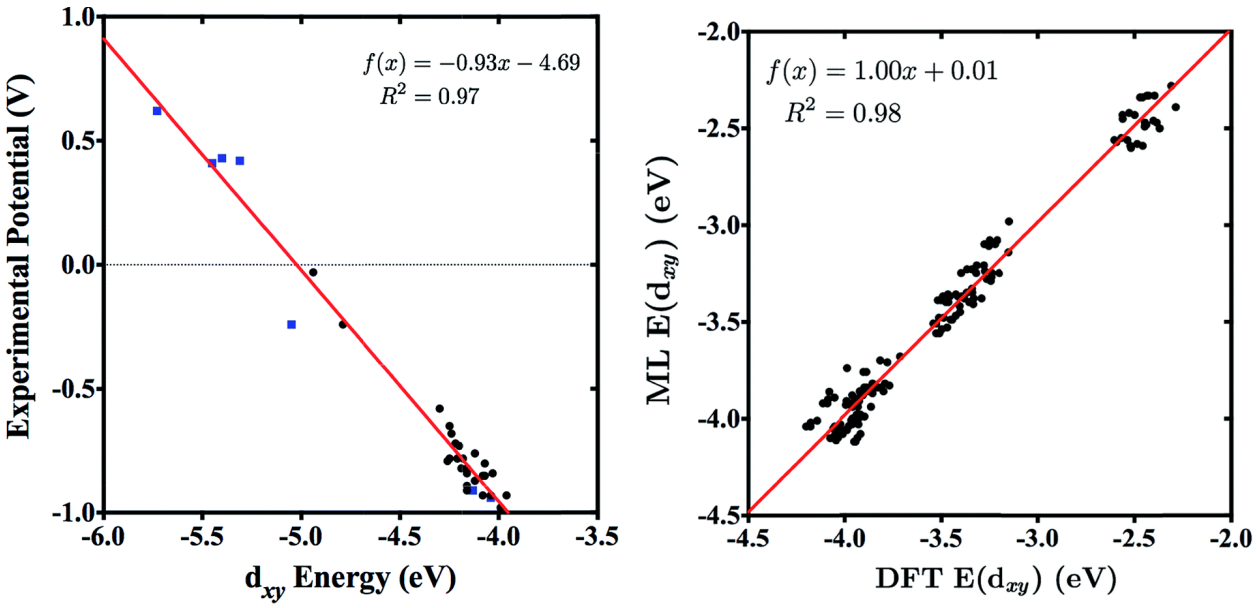

Figure 53. (left) Experimental potentials (V) of 32 W complexes plotted against their corresponding DFT (B3LYP) d_{xy} frontier orbital energies (eV). The method was chosen to reproduce experimentally observed potentials from reversible (black circles) and quasi-reversible (blue squares) experimental potentials. (right) Artificial neural network prediction of DFT orbital energies. Reproduced with permission from ref. [496]. Copyright 2019 Royal Society of Chemistry.

Using structure-dependent descriptors, Meyer et al.[540] devised an ML-model-accelerated approach for volcano-based (see Sec. 4.2) screening of cross-coupling catalysts. First, they trained KRR models on a range of conventional organic chemistry descriptors to predict a representative free energy in the cross-coupling reactions (Figure 54). On these complexes with late transition metals (e.g., Pd, Cu, and Au), they used semi-empirically optimized geometries to evaluate the structure-dependent features. Consistent with observations from Janet et al. on spin-splitting energies[54, 500], they observed poor performance of the Coulomb matrix representation in learning curves up to around 7,000 catalysts (Figure 54). The bag of bonds (BoB, see Sec. 4.4.1) and similar fragment-oriented structural descriptors achieved the highest accuracies up to 2.5 kcal/mol MAE, with some metal dependence likely attributable to poorer structure generation for earlier metals (e.g., Ni, Figure 54).[540] This approach has also been used for screening other Ni

catalysts for reductive cleavage in aryl ether compounds.[541]

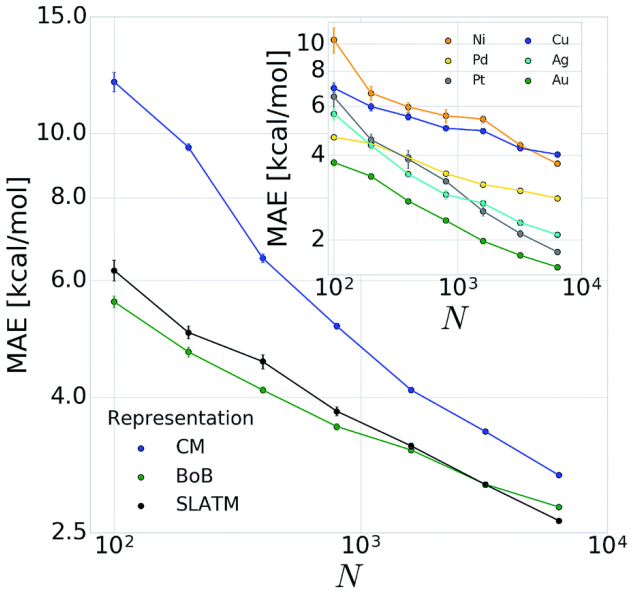

Figure 54. Learning curves, i.e., test error of catalytic descriptor values as a function of training set size (*N*), for oxidative addition of vinyl bromide for the Coulomb matrix (blue line), bag of bonds (BoB, green line), and SLATM potential (black line). Error bars correspond to standard deviation in cross-validation. The inset shows the corresponding learning curves for individual metals for BoB. Reproduced with permission from ref. [540]. Copyright 2018 Royal Society of Chemistry.

In recent years, ML regression models have also been extensively applied to experimental data sets as a generalization of QSPRs (Sec. 4.2) often using similar QM descriptors. In these cases, the DFT calculations and ML models tend to outperform standard physics-based modeling but require considerable calculation time in comparison to models trained on heuristic descriptors. Denmark and coworkers used descriptors of sterics and the electrostatic potential (i.e., similar to those proposed by Lipkowitz and coworkers[414]) to train supervised ML models (i.e., ANNs and SVMs) to predict enantioselective organocatalysts for chiral phosphoric acid-catalyzed thiol addition to N-acylimines.[542-543] To avoid extrapolation, the training subset was selected to uniformly cover the space spanned by the input features. The model's performance was sensitive to the data balance in the necessarily small training set accessible to a synthetic chemist.[543] The predicted free energy changes from the best-performing

ANN models exceeded what could typically be obtained from DFT-based screening alone and enabled prediction of new catalysts with up to 99% enantioselectivity.[542-543] Others have shown that graph-based descriptors (e.g., Kier shape index used in MCDL-25) that have been demonstrated for open-shell transition-metal complex ML are as important as steric- or charge-based descriptors (e.g., in prediction of experimental ethylene polymerization catalysts).[544]

Doyle and coworkers predicted the yield of catalysts[545-546] for cases (e.g., Pd C–N cross-coupling on substrates containing five-membered rings) where traditional approaches to performance improvement had failed. Using high-throughput experimentation, they trained random forest models on DFT-derived descriptors to predict reaction yields.[545] Feature selection in conjunction with non-linear models (i.e., the random forest model) outperformed linear models, but ANNs provided limited benefit on these small data sets.[545] It has been debated the extent to which the descriptors used allowed the models to generalize to unseen chemical entities.[547-548] LASSO models have also been trained on a similar series of QM descriptors (e.g., HOMO–LUMO gaps, charges, and frequencies) to predict the reaction yields of tungsten epoxidation catalysts to good accuracy.[521] The benefit of QM charge descriptors to improve predictive capabilities on small data sets echo earlier observations made for ML models (e.g., decision trees and SVMs) trained on a series of ligand descriptors (e.g., partial charges and shape or connectivity) to predict the molecular weight of polymers generated by polymerization catalysts.[520]

Iterative approaches to catalyst design that merge experimental, DFT-based, and data-driven models have been pursued. For example, Maley et al. trained[549] random forest models with heuristic properties of transition state geometries to predict DFT barrier heights in Cr-catalyzed olefin oligomerization. They interpreted the feature importance in these models to

identify key ligand factors correlated strongly to barrier height and exploited them to identify ligands that would enable selective formation of octene over hexene (Figure 55). Through this process, they improved the catalyst selectivity to high levels[549] (Figure 55). Others have used interaction energies from energy decomposition analysis to accurately predict experimental activation energies with ML models and have observed these features to be as important as conventional (e.g., steric, frontier orbital energy, or charge) descriptors.[524]

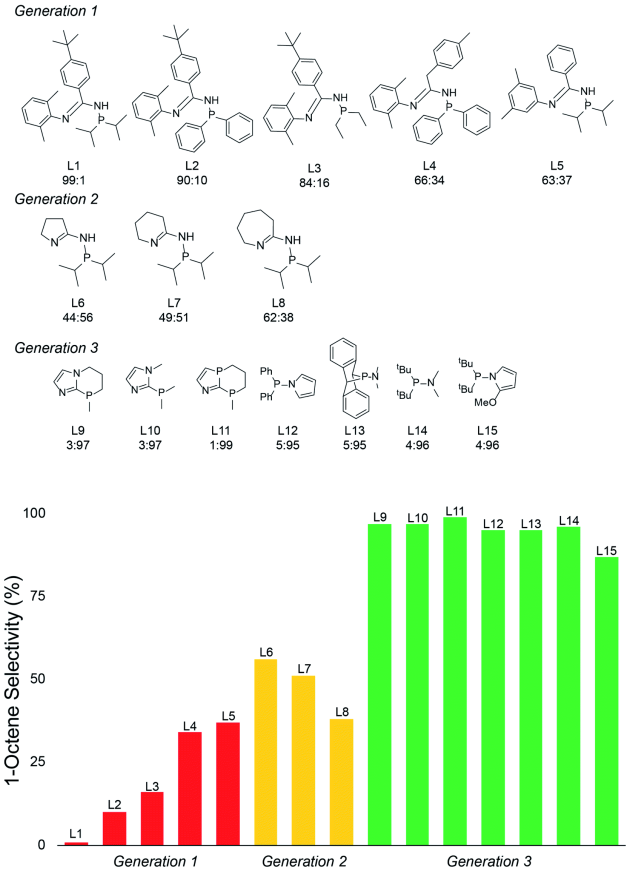

Figure 55. (top) Structures for previous (P,N) ligand generations and new proposed ligands (generation 3) based on DFT and machine-learning-identified features. The 1-hexene:1-octene selectivity (predicted) is given below each structure. (bottom) Plot of 1-octene selectivity for previous (P,N) ligand generations and proposed ligands in generation 3 obtained from a combination of DFT and machine learning. Reproduced with permission from ref. [549]. Copyright 2020 Royal Society of Chemistry.

Data-driven models have also been developed to predict a combination of experimental

catalytic quantities that would be challenging to obtain from DFT-based energetic analysis. Using kinetic characteristics of the catalysts as inputs to train a multi-task ANN, Rizkin et al.[550] built a model to predict the average molecular weights, polydispersity, and reactor temperature for zirconocene-catalyzed olefin polymerization. Siebert et al.[551] built random forest models to identify optimal reaction temperature, time, catalyst loading, and concentrations for CO_2 reduction with ruthenium catalysts.

Considerable effort has been made towards supervised learning of the properties of solid-state, inorganic materials owing to the large datasets of DFT calculations available. For example, early efforts used these data sets and *ad hoc* descriptors to predict thermodynamic stability[552] and electronic properties[553]. More recently, graph-convolutional neural networks and graph-based representations have been used to predict properties with improved accuracy.[554-555]

We next focus on MOF ML model predictions given the close connections of MOFs to molecular transition-metal complexes. A number of ML regression models of MOF properties have been developed using tailored representations described in Sec. 4.4.1. A range of models (e.g., SVR, ridge regression, and random forest) were trained on one-hot compositional features for very small experimental data sets (ca. 100 points) of MOF CO_2 and H_2 adsorption curves, but small data set sizes have limited performance.[517] Instead, most predictive MOF models have focused on predicting grand canonical Monte Carlo (GCMC)-simulated gas uptake characteristics on hypothetically enumerated MOFs from compositional and structural information in order to accelerate exhaustive screening efforts.

Support vector regression (SVR) models were trained on a subset of 120k hypothetical MOFs featurized with geometric descriptors (i.e., AP-RDFs) to predict uptake of CO_2 and methane from GCMC.[518] The SVRs outperformed standard MLR models. While whole-

structure, geometric descriptors were sufficient to yield predictive models for methane, the AP-RDFs that encode chemical composition were needed to yield improved (R^2 = 0.75) models for predicting CO_2 uptake. Work with training random forest and KRR models on graph-based descriptors across a range of hypothetical MOF data sets has reaffirmed the observation that chemical composition information is key to prediction of CO_2 uptake at low pressure but less essential with high pressure or for predicting methane uptake.[384, 516] Analysis of feature importance in these predictions in particular highlighted the importance of the metal SBU chemistry[384], emphasizing limited transferability of MOF ML models to new SBUs not in the training set (Figure 56).[518] Indeed, earlier feature importance analysis applied to the role of topology and functionalization in MOFs (e.g., electron-donating and withdrawing groups) in a carefully constructed data set that limited variation of the metal SBU had instead indicated that geometry played a dominant role in determining relative gas uptake of CO_2, H_2, and N_2.[510]

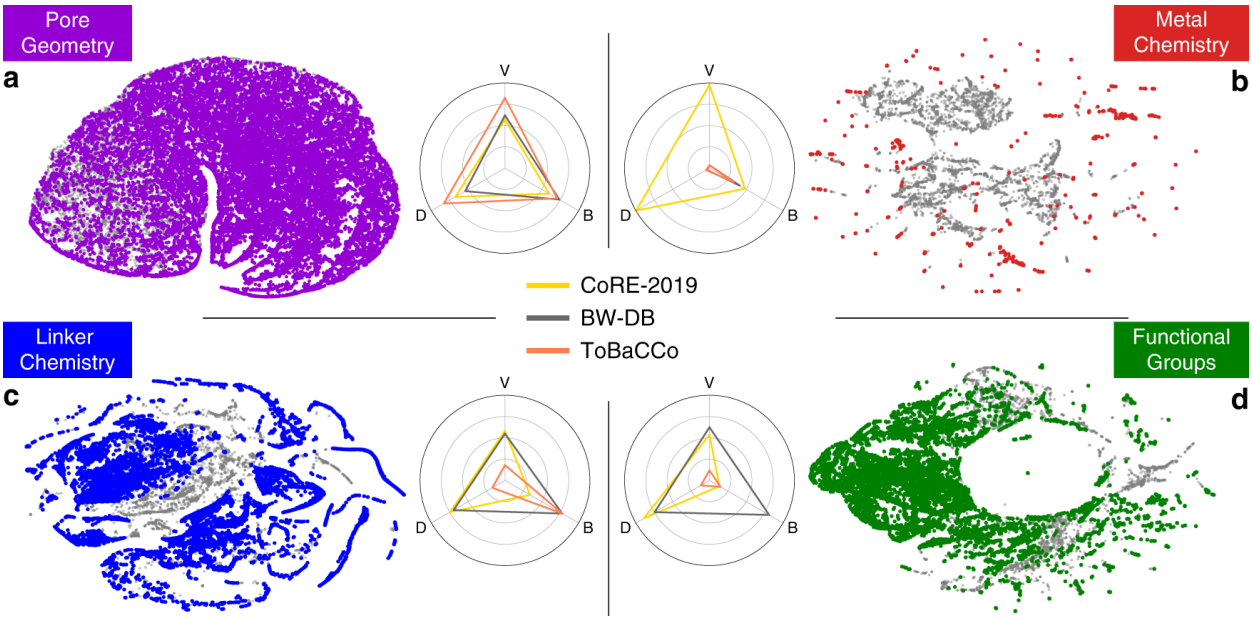

Figure 56. t-SNE on subsets of MOF RACs and geometric descriptors including: (top, left) pore geometry in purple, (top, right) metal chemistry RACs in red, (bottom, left) linker RACs in blue, and (bottom, right) functional group RACs. All hypothetical and experimentally characterized (i.e., CoRE-2019) datasets are shown in gray, whereas only hypothetical data is colored, highlighting the lack of diversity in metal chemistry in hypothetical sets. The radar charts show the three diversity metrics: variety (V), balance (B) and disparity (D), for the three

databases. Reproduced with permission from ref. 384. Copyright 2020 Springer Nature.

Despite the good performance of geometric descriptors on methane uptake, Pardakthi et al.[515] showed that compositional information further improved predictions, with very good performance for methane uptake prediction (R^2 = 0.97) on the HMOF data set[556] by random forest models trained on both structural and chemical (i.e., elemental composition, bond types) information (Figure 57).[515] Random forest models outperformed alternatives such as SVRs or a single binary decision tree (Figure 57).[515]

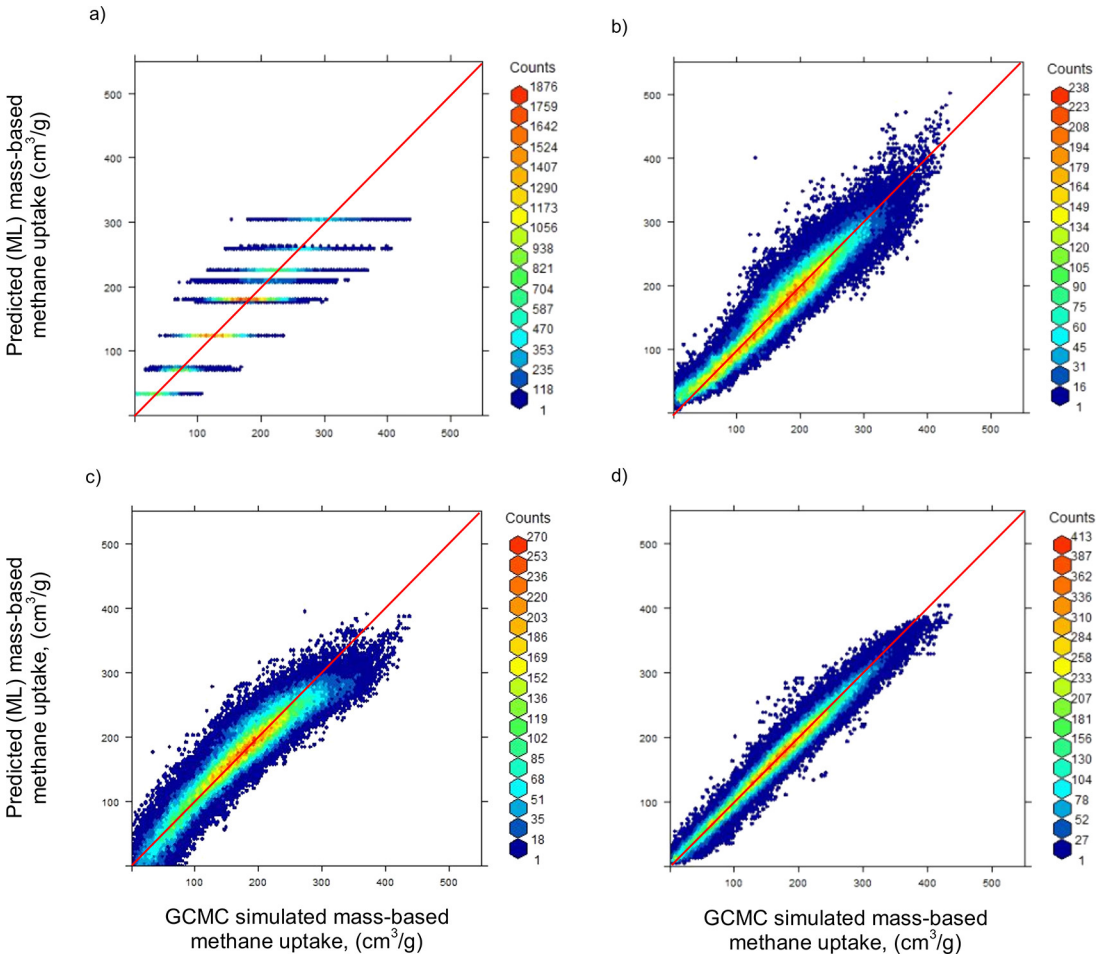

Figure 57. Parity plots for predicted (ML) vs GCMC-simulated mass-based methane uptake (cm^3/g) using structural and chemical variables applied on (a) single decision tree (DT), (b) Poisson, (c) support vector machine (SVM), and (d) random forests (RF) models for hypothetical MOFs (HMOFs). The parity line indicating perfect correspondence between ML predictions and GCMC simulation results is shown in red in each plot. The color scale indicates the data density in the plots (i.e., number of counts). Reproduced with permission from ref. [515]. Copyright 2017

American Chemical Society.

While most described models for MOF property prediction have favored random forests or ANNs and use heuristic compositional and structural descriptors, alternatives have also been pursued. In a hybrid approach, Bucior et al.[514] developed a descriptor-based screening with a probe H_2 molecule, in close analogy to the steric map approaches used for transition-metal complex QSPR. These models involved evaluation of H_2 adsorption energies on a grid in the MOF materials and used these energies as inputs to a LASSO regression model.[514] The evaluation of interaction energies could be carried out on a coarse 1-Å grid but still required more time to compute than heuristic descriptors.

Models are typically judged based on prediction errors on a set-aside test set, and an outstanding challenge is in quantifying uncertainty of a model for generalization to new compounds in chemical discovery (see Sec. 5). For highly engineered feature sets, distances of new compounds to the training data in feature space has been used in ML property prediction.[54, 557] The popularity, flexibility, and performance of ANNs with numerous hidden layers that effectively carry out feature engineering motivate alternative approaches to uncertainty quantification (UQ). Ensembles of models trained on distinct partitions are sometimes used[527, 558-560], and the Monte Carlo dropout[526] approach where uncertainty is estimated from different ANN architectures (i.e., with zeroed out nodes) has also been applied in transition-metal complex ML models.[54]

Janet et al. showed[527] the distance in the latent space of an ANN to be a superior, low-cost strategy to detecting high-error points, especially for discovery applications where standard methods (e.g., ensembles) would be overconfident. Calibration of this distance with maximum likelihood estimation enables conversion of the distance to a confidence interval on the property

prediction. Optimization of the properties of organic molecules in continuous (i.e., latent space) representations and inversion back to discrete candidates has also been demonstrated as useful tool for molecular discovery.[561] More recently, the distance in latent space has been shown to be useful for experimental catalyst optimization[562] and demonstrated for accelerated discovery with active learning[563-564] (see Sec. 5.2).

4.4.3. Classification and Semi-Supervised Learning Models.

Although the majority of machine learning efforts in transition-metal chemistry have been focused on regression for property prediction, there are a number of cases where alternative models have been developed for classification. We briefly summarize these approaches, including those focused on binary classification as well as emerging techniques in semi-supervised learning for both regression and classification of materials in which only some of the data labels are known beforehand. Classifiers and semi-supervised learning models have been developed for artificial-intelligence-informed workflow decisions[519, 565], as presented in more detail in Sec. 5.3.

In binary classification tasks, the model output is a score ranging from 0 (false) to 1 (true). The models most widely employed are support vector machines (SVMs) and ANN classifiers. Typically, a threshold of 0.5 on the output score is used to assign each data point to a class. The overall performance of the model is captured by the receiver operating characteristic (ROC) curve that indicates the rate of false positives as the threshold is adjusted to increase the rate of true positives. When the area under the curve (AUC) of the ROC approaches 1, the model is robust in a manner that is less sensitive to the threshold chosen to assign the classes. In certain cases, classification model scores can be interpreted as probabilities and therefore can be used for uncertainty quantification. Duan et al. observed[519] ANNs to generalize better than SVMs in

some classification tasks, but application of ANN classifiers to diverse molecules suggested raw probability scores were overconfident. This observation led them to introduce[519] an alternative uncertainty quantification metric they termed the latent space entropy (LSE):

$$s_i = -\sum_j^J \tilde{p}_i^{(j)} \ln(\tilde{p}_i^{(j)}) \quad (5)$$

where the LSE measures the proximity of the training data in each class, *j*, and ranges from 0 for perfect confidence to ln(2) = 0.693 for no confidence. The confidence is low when the point is close to the model's decision boundary in latent space or when it is distant from all points even those within the same class.

An early example of classification in transition-metal chemistry is the development of models that could predict transition-metal binding sites in metalloenzymes from the protein's structure without the metal present (i.e., the apoenzyme).[566] This approach was motivated by the much higher availability of structures of a protein without the metal resolved in the apoenzyme in comparison to those with the metal bound. Decision trees and SVM classifiers were trained on a series of geometric and sequence-based descriptors for 125 protein chains with 367 known metal binding sites, and the models achieved a low rate of false positives (i.e., 95% selectivity).[566]

The phenomenon of spin-crossover poses a straightforward question for classification models, and SCO complexes in the CSD represent a rich source of data for ML models because crystal structures solved at multiple temperatures give access to strongly spin-state-dependent bond lengths.[567] Comparison of approximate DFT bond lengths to experimental bond lengths usually yields good agreement[131, 150, 162] (see Sec. 2). Taylor et al. leveraged ML-predicted DFT bond lengths to assess and classify experimental spin states in the CSD.[568] This model improved upon purely energetic predictions of ground state spins from approximate DFT functionals and

correctly classified (i.e., high-spin vs low-spin) 96% of experimental structures based on the similarity of the CSD metal–ligand bond length to the ML-predicted DFT value.[568] Cundari et al. employed bond lengths from around 100 CSD complexes to build small neural network classifiers, but over the small data set size, classification models based on Bayesian inference performed better to determine relationships between M–N bond lengths and M–N–C angles in transition-metal imidos.[569]

Classifiers have also been developed to separate good and bad materials characteristics for CO_2 uptake by MOFs.[570] The classification task consisted of distinguishing MOFs capable of greater than 1 mmol/g CO_2 uptake at low pressure (i.e., 0.15 bar CO_2). The SVM models for this classification task were trained on 10% of 324,000 hypothetical MOFs generated for the study and featurized with AP-RDFs[518]. This approach could be used to accelerate pre-screening of MOFs for improved gas uptake characteristics in hypothetical databases (see Sec. 5).[570]

Semi-supervised learning techniques have yet to be widely employed in materials or transition-metal chemistry. The term "semi-supervised" refers broadly to methods where a mixture of initially labeled and unlabeled data is employed, and the remainder of labels get assigned during model training.[571] In one example, researchers have combined clustering with supervised learning in a two-step semi-supervised approach to image classification for labeling X-ray diffraction images relevant to high-throughput experimentation in materials science[572]. In analysis of the materials literature with natural language processing, where explicit labeling of extracted results can be challenging, a two-step semi-supervised approach was also recently demonstrated for extraction of magnetic phase transition temperatures[573] and for materials synthesis procedures (e.g., with a random forest classifier).[574] Still more applications of semi-supervised learning are anticipated as a way to overcome limitations in the availability high-

quality, unambiguous data in transition-metal chemistry, some of which are discussed in detail in Sec. 5.3.

Within semi-supervised learning models, such as for training of ANN classifiers with partially labeled data[565], a related concern is the effect of the amount of labeled data on model prediction. For chemical problems, models have been shown to be fairly insensitive to this labeling threshold[565], and the high cost of data generation may guide more when lower thresholds are employed in labeling. For some supervised and semi-supervised learning models where the classification task is based on an arbitrary cutoff[565, 570], the ROC may also not be the optimal method for assessing model performance. Instead, some have qualified how well the model separates the two classes[565] using Bhattacharyya distance to measure the differences in the distributions of the two classes.

5. Transition-Metal Chemical Space Exploration.

Thus far, we have described iterative, mechanistic search (Sec. 4.2) where a transition-metal complex that addresses most design objectives may be known but requires some improvements. Conversely, when good design principles are not known, alternative approaches have been used to evaluate a larger space of feasible compounds as candidate materials or catalysts. If the scoring and evaluation of materials is efficient (e.g., with low-cost computation or surrogate models), expansive, combinatorial evaluation of all materials on a large scale (ca. 1000s to millions of compounds) is often carried out (Sec. 5.1). When seeking to identify materials that satisfy design objectives without carrying out unproductive experiment or computation, single- and multi-objective optimization with evolutionary algorithms (Sec. 5.2) has been widely employed to guide experiment, computation, or ML-accelerated discovery.

5.1. Combinatorial Enumeration.

In contrast with the functionalization and small-scale ligand permutation studies discussed in Sec. 4.2, we refer here to combinatorial enumeration that generates a search space of at least several thousand transition-metal complexes or materials. The reticular nature of metal organic frameworks (MOFs) combined with the relatively low cost of evaluating their properties with classical force fields (i.e., for gas separations characteristics) has led to a large number of enumerative studies on MOFs over the past two decades[510, 556, 575-581], as reviewed in detail in Ref. [578]. When previously synthesized secondary building units (i.e., metal nodes) are combined with synthetically accessible linkers and then functionalized, they can easily yield large sets (ca. 100k or more) of hypothetical MOF materials.[556] Early efforts in this area were limited to straightforward-to-enumerate combinations that led to primitive cubic unit cells[556, 575], whereas later studies expanded to generalize to non-cubic unit cells.[577]

Given the possible ways to combine linkers, functional groups, and metal SBUs, more focused approaches have also been used. Functional group variation has been used to create a combinatorially large (ca. 10k or more) space for screening materials with improved characteristics (e.g., for CO_2 capture) even when the metal SBU is held fixed.[580] For gas separation, it is also expected that the pore structure and overall topology should play key roles, motivating a focus on the underlying MOF connectivity[577, 579, 581]. Gómez-Gualdrón et al. focused on the relationship between MOF topology and gas adsorption characteristics, identifying overall structures most suited for hydrogen storage in the ToBACCo MOF data set[579, 582]. Similar strategies were previously applied to generating hypothetical porous polymer network materials[583] and, even earlier, zeolites[584]. By sampling distinct crystal structures and ring size distributions in porous zeolite materials[584], Deem and coworkers built a database of 2.7 million hypothetical zeolites, a dramatically larger number than have ever been successfully

synthesized.[585]

These enumerated sets (e.g., of MOFs) are often used for large-scale screening (i.e., with grand canonical Monte Carlo) of gas uptake properties. For example, Boyd et al.[576] screened a space of 325k hypothetical MOFs for CO_2 working capacity and selectivity of binding relative to water. They identified design characteristics of the top 8k materials, finally identifying the MOFs that looked synthetically accessible and were subsequently demonstrated to experimentally have the computationally-predicted performance.[576] Given the large number of materials that can be generated in this way, it is possible to identify numerous design leads and use intuition to address questions regarding stability or synthetic accessibility. More recently, machine learning models have been increasingly applied to the development of interpretable design rules from these sets (see Sec. 5.2.3).

One concern for the use of enumerative sets is the extent to which design leads or design principles are sensitive to the data set employed for their extraction. Moosavi and coworkers observed[384] that the CoRE-MOF database of experimentally synthesized materials exhibits key distinctions from most of these hypothetical MOF sets (Figure 56). Using t-SNE maps of MOF materials in graph-based representations (i.e., RACs) extended for metal-organic frameworks in combination with geometric descriptors, they analyzed the similarity of experimental and hypothetical MOFs over specific domains (i.e., the SBU vs linker or functional group). Some observed distinctions could be expected, such as that pores were larger in some hypothetical data sets (i.e., ToBaCCo[582]) that had emphasized distinct topologies not synthetically accessible. While greater diversity of hypothetical over synthesized materials was expected, the hypothetical MOFs also interestingly lacked some diversity observed in experimental MOFs (Figure 56). In particular, the SBUs sampled in hypothetical sets are often much narrower than those that have

been experimentally synthesized. While demonstrated on MOFs, these observations highlight the general challenges for combinatorial enumeration of inorganic materials in avoiding bias when proposing new materials while maintaining some realism.

In comparison to MOFs or small organic molecules, fewer efforts have been made to elucidate rules for enumeration of feasible transition-metal complexes. In organic chemistry, the generated database (GDB)[586] series has enumerated molecules up to discrete numbers of heavy (i.e., non-H) atom sizes under the constraints of obeying the octet rule and being in closed-shell singlet states with no net charge. Subsets of the GDB with seven to nine heavy atoms have been widely employed to generate data sets[503, 587] for ML of properties of organic molecules and for high-throughput screening. Enumerative strategies for transition-metal complexes are less well developed. In one example, a small-scale study of hypothetical modified cyclopentadienyl ligands (e.g., all-nitrogen or all-phosphorus rings) was used to understand the mechanism of mononuclear single-molecule magnet behavior.[588] Because the smallest non-trivial octahedral complex has at least seven heavy atoms due to the ligands and has more electrons than a typical organic molecule due to the metal center, exhaustive generation of very small transition-metal complexes has motivated a distinct strategy.[589]

Gugler et al. constrained the search of individual ligands to a small number (ca. one or two) of heavy atoms per metal-coordinating site (e.g., four heavy atoms in a bidentate ligand) and relaxed typically applied constraints (e.g., for the GDB) in generating ligands with one to four heavy (i.e., C, N, O, P, or S) atoms.[589] They assigned more favorable scores to neutral and octet-rule-obeying molecules but still allowed for exceptions, producing an octahedral homoleptic ligand database, OHLDB[589], with DFT properties of 1250 complexes formed with mid-row transition metals (i.e., Cr–Co). The enumeration of lower-symmetry heteroleptic

complexes directly with these ligands or by exploiting ligand additivity[60] leads to hundreds of thousands of complexes. These data sets have been used to improve ML model performance on larger complexes[589].

Exhaustive enumeration of transition-metal complexes has also been carried out in a stepwise fashion for larger bidentate ligands. In 2006, Rothenberg and coworkers hypothesized that generating bidentate ligands from a series of rings, bridging, and functional groups could rapidly lead to vast (ca. 1.7 billion) transition-metal complex spaces infeasible for full-scale exploration (Figure 58).[590] Instead, they focused on the properties of 600 representative complexes from this larger space for which they carried out force field optimization and ultimately further characterization of figures of merit (i.e., experimental turnover frequency, Figure 43). More recently, Janet et al.[563] started from well-studied five- and six-membered ring motifs with common coordinating atoms (i.e., N or O), modified (i.e., with changes in conjugation or additional heteroatoms) the rings, and bridged them to generate nearly 800 unique core bidentate ligand types. Sequential functionalization and homoleptic complexation with mid-row transition metals generated a space of 2.8 M candidate complexes for ML-accelerated multi-objective optimization (see Sec. 5.2.3) validated with DFT calculations.

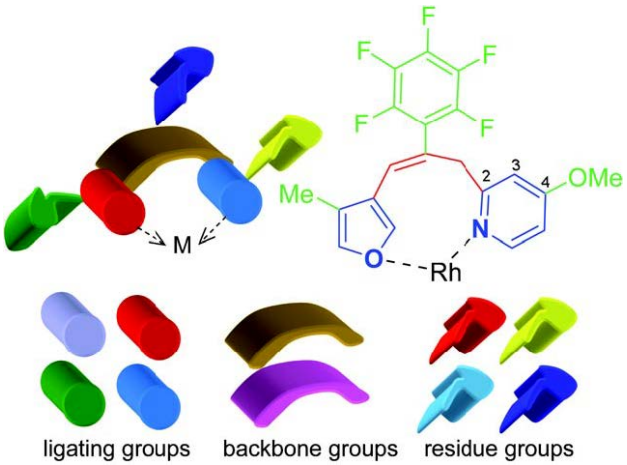

Figure 58. Cartoon showing the construction of a bidentate ligand (top left) and an example resulting structure (top right) from building blocks comprised of ligating, backbone, and residue

groups shown at bottom. Reproduced with permission from ref. [590]. Copyright 2006 WILEY‐VCH Verlag GmbH & Co.

5.2. Optimization Approaches.

When aiming to discover materials or catalysts with improved properties, exhaustive enumeration of large chemical spaces is impractical and unnecessary. If we have a single, well-defined objective, evolutionary algorithms such as standard genetic algorithms (Sec. 5.2.1) can accelerate the identification of transition-metal complexes and materials that satisfy that objective. In practice, multiple objectives must typically be satisfied (Sec. 5.2.2). Machine learning models can accelerate single- or multi-objective optimization but introduce special challenges related to model uncertainty (Sec. 5.2.3). Here, we briefly highlight efforts in discrete optimization with evolutionary algorithms for transition-metal complex and materials design, especially with ML-accelerated discovery. Details beyond the scope of this review regarding design, including for efforts beyond transition-metal chemistry, are provided in Refs. [591-592].

5.2.1. Genetic Algorithms.

Genetic algorithms (GAs) are widely employed in transition-metal complex and broader materials discovery efforts to accelerate identification of lead compounds that satisfy a design objective.[591-592] GAs are applied to a pool of candidate materials (e.g., transition-metal complexes) with components that are represented by discrete genes. A fitness function that is a score approaching one when the material approaches the ideal property (e.g., catalyst activity) is used to determine which materials (i.e., combinations of genes) continue on to subsequent generations in a GA. Operations at each generation involve crossover of genes between parents and random mutations to avoid too rapid collapse to a local minimum. Although also used in other areas of molecular discovery, GAs are a natural fit for the modularity of transition-metal

complexes, in which the genes will often represent the metal, one or more ligand scaffolds, and functional groups attached to the ligand scaffold. This set of genes is evolved by scoring each complex with a fitness function and carrying out crossover and mutation on the top-scoring subset for a number of generations.

The value of GAs in accelerating screening of combinations of reaction conditions in experimental catalyst and materials screening as well as to explore large chemical spaces was recognized over 20 years ago.[593-594] Within experimental homogeneous catalysis, advances in microfluidics and high-throughput screening (e.g., with a 96-well plate) has enabled rapid catalyst optimization. Early studies carried out sequential screens of catalysts (e.g., for copolymerization), with a fast primary stage followed by slower secondary and tertiary screens in a narrowed space of candidates.[595] This high-throughput experimentation approach also enabled rapid GA evolution of catalyst chemistry and reaction conditions for methane oxidation (Figure 59).[596] A GA was used to explore genes representing variations in the catalyst metal from 8 choices, 11 co-catalyst types for oxygen activation, and 13 ligands.[596] To accelerate scoring, colorimetric assays were developed, to avoid time-consuming but likely more conclusive methods (e.g., NMR) of characterization (Figure 59).[596]

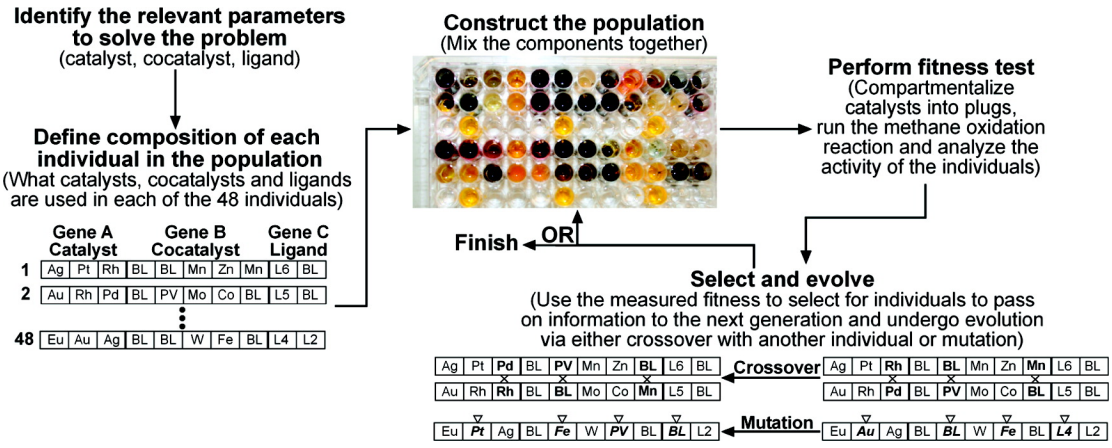

Figure 59. Schematic illustration of genetic algorithm implemented with microfluidics to search

for catalysts. Each vessel contained three genes to optimize the different parameters: catalyst (gene A), cocatalyst (gene B), and ligand (gene C). Each gene was composed of multiple chemical species, examples of which are shown in this schematic: BL stands for blank solution, PV for $H_5PMo_{10}V_2O_{40}$ POM-V2, L*n* (*n* = 1, 2, 3, etc.) for one ligand type, and elemental symbols represent the metal present. After the first generation was produced, the catalytic activity of each individual was tested, and the results were analyzed to determine fitness and generate a new population of 48 catalysts/co-catalysts/ligands through a combination of crossover or mutation. The genetic algorithm optimization was repeated over eight generations. Reproduced with permission from ref. [596]. Copyright 2010 American Chemical Society.

In molecular discovery, when the objective function is explicitly calculated from electronic structure methods, it is also necessary to avoid high computational cost in evaluating each generation. For instance, in heterogeneous catalysis, GAs have been applied to search for semiconductor alloys[597] or core-shell nanoparticle catalysts[598] with favorable band gaps or catalytic activity. Here, approximate (i.e., d-band theory) methods were used in scoring, rather than more computationally demanding approaches. Within transition-metal chemistry, Chu et al. built an objective function for Grubbs-type Ru metathesis catalysts based on semi-empirical PM6 descriptors in conjunction with a QSPR model capable of making DFT-quality activity predictions.[599] Although representing only a proof of principle due to the small scope of ligands considered (i.e., two types of functionalized phosphines or NHCs), they showed that crossover, mutation, and selection of fittest compounds led to the rapid selection of experimentally favored NHC ligands over alternatives, such as phosphines (Figure 60).

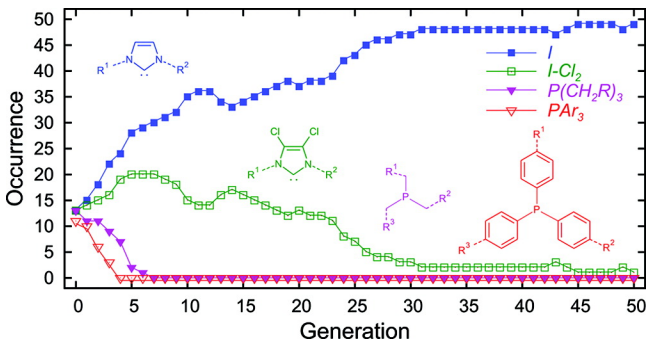

Figure 60. A genetic algorithm for optimization of ligand chemistry in metathesis. The objective function is obtained from partial least-squares regression of semi-empirical geometries and DFT molecular descriptors. Examples of how ligands are selected with the evolutionary (i.e., genetic) algorithm grouped by whether it selects: an imidazol-2-ylidene scaffold with two substitution points and two hydrogen atoms (termed I) or two chlorine atoms (I–Cl_2), as well as a triphenyl phosphine skeleton with substitution points at the phenyl para positions (PAr_3) and a trialkyl phosphine skeleton with three substitution points ($P(CH_2R)_3$). Substitution points are indicated by R^n (enumeration by *n*) and dashed lines. Reproduced with permission from ref. [599]. Copyright 2012 American Chemical Society.

Foscato et al. demonstrated a more general approach to *de novo* complex discovery by developing ring-closure operations embedded within a genetic algorithm for the design of spin-crossover complexes.[600] Starting from an initial population of 50 random complexes that were evolved for 100 generations, the fitness of candidate Fe(II)/N octahedral complexes was evaluated at each step with an objective function based on the adiabatic spin-splitting enthalpy (Figure 61).[600] Both the use of an enthalpy and its evaluation with LFMM[16-17] kept the computational cost of objective evaluation suitably low. Notably, because both graph-based and 3D information was used to introduce operations that altered the denticity of the ligands, GA optimization identified the preference for tridentate ligands (e.g., terpy) over a starting population that consisted predominantly of combinations of tetradentate and bidentate ligands (Figure 61). The specific computational prediction of spin-crossover behavior for $[Fe(II)(tame_2)]^{2+}$ from this work was later experimentally verified, demonstrating the potential for computational design to reveal functional transition-metal complexes.[601]

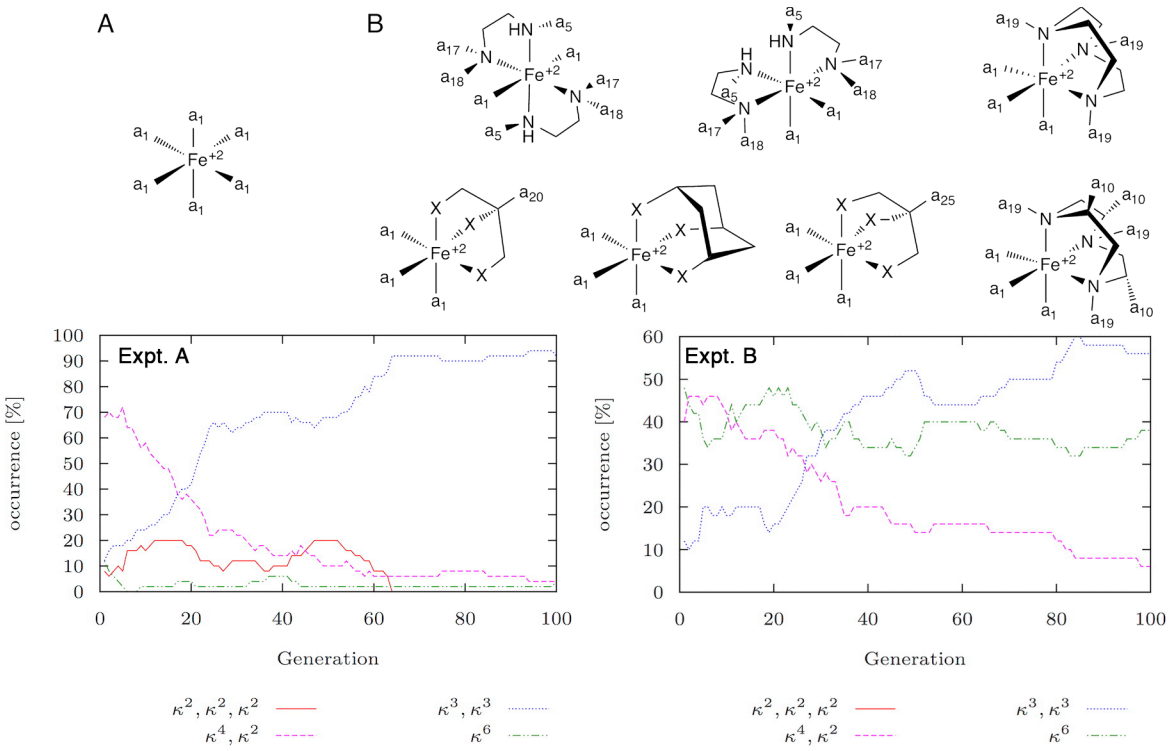

Figure 61. (top) Library of 3D fragments (scaffolds) used to initiate the construction of Fe(II) complexes in experiments of types A and B. The addition of fragments can lead to further ring closures. A fitness function of spin-splitting energies is evaluated using the LFMM force field in experiments A (bottom, left) and B (bottom, right). The final surviving denticities of ligands are indicated in the legends at bottom. Adapted with permission from ref. [600]. Copyright 2015 American Chemical Society.

In a similar fashion, genetic algorithms have been used to accelerate exploration of candidate MOF materials (e.g., to maximize methane storage capacity).[602] Bao et al. focused on using candidate linkers from commercially available molecules in combination with known SBUs and MOF topologies.[602] They devised a series of operations to replace and mutate linker chemistry, filtered the candidates based on geometric observations from classical molecular dynamics, and then scored and retained structures with high methane storage capacity (Figure 62).[602] This procedure simultaneously increased the feasibility of the materials studied for simulation (i.e., by filtering) and reduced the total number of materials that needed to be screened by GCMC simulation.[602] Similarly, Collins et al. used a GA to identify how functional

group variation could alter CO_2 uptake characteristics, searching a space of nearly 96k viable structures and observing significant improvement in the best-performing materials after only a few generations of the GA.[603]

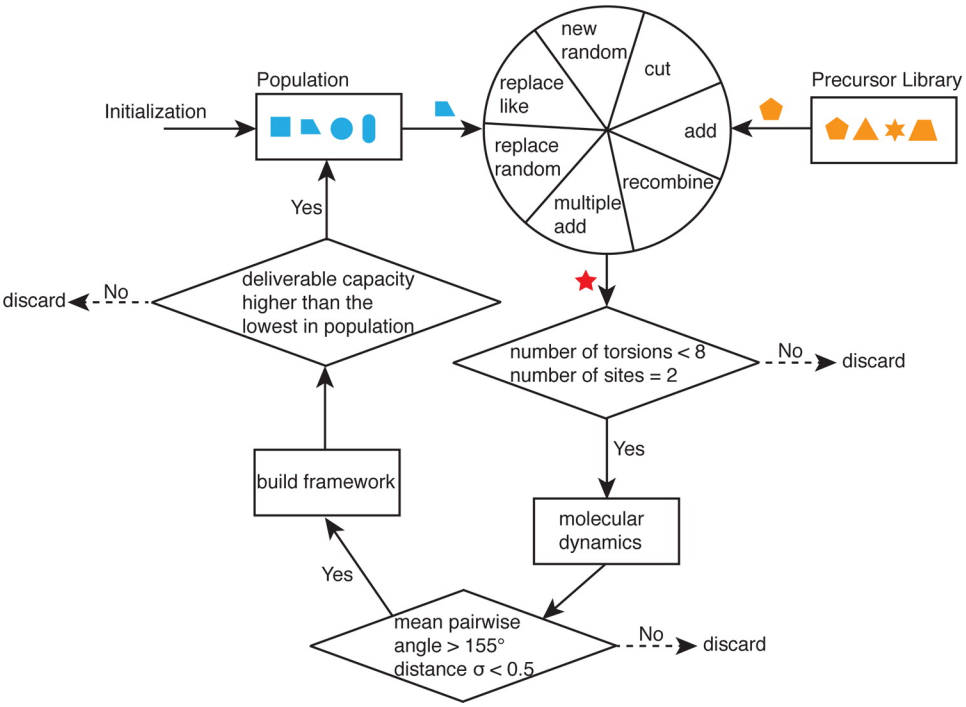

Figure 62. Approach for discovering new MOF materials with high deliverable capacity. Illustration of an algorithm for initializing a population with add or multiple add operations from a precursor library. One randomly chosen linker from the population undergoes one of seven evolution operations. The produced linker is evaluated by two filters: the number of torsions is < 8, and number of sites = 2. If the linker passes these filters, molecular dynamics is used on the linker to produce a set of conformations. These linker conformations are evaluated by the two additional filters of mean pairwise angle ≥ 155° and a standard deviation of pairwise distance < 0.5 Å. If the linker passes all filters, it is used to build a MOF of the chosen network. Finally, if the constructed MOF has a greater deliverable capacity than the lowest one in the current population, the linker is inserted into the population in rank order. Reproduced with permission from ref. [602]. Copyright 2015 American Chemical Society.

The large number of purely hypothetical MOFs that have been generated[556] also creates a large combinatorial space to search for new materials and guide new simulations. A GA was used to accelerate discovery of hypothetical MOFs with high selectivity (i.e., over H_2) for CO_2 uptake (Figure 63).[604] Because some properties, such as surface area or working capacity were

known for all of the MOFs, the GA strategy could be validated based on its ability to find the best performers for those properties (Figure 63).[604] Once validated, the GA was then used to reduce the computational effort required for the selectivity studies to only 1% of what a brute force effort would have required.[604] Given their suitability for accelerating search of large combinatorial spaces, GAs have been used to optimize candidate arrays of MOFs as sensing materials by varying the individual MOF genes (i.e., from the CoRE-MOF database) and favoring MOF arrays that maximize the differences in gas uptake properties among the MOFs.[605]

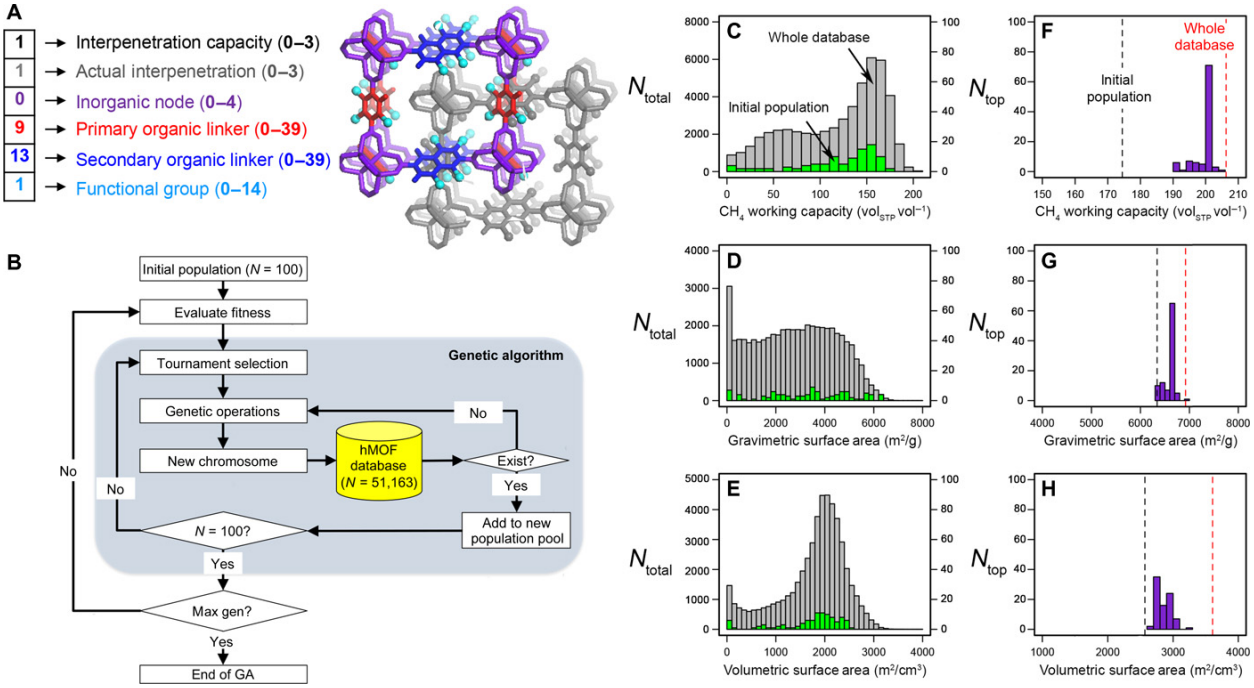

Figure 63. (A) An example set of genes and the corresponding hypothetical MOF (HMOF) structure. Colors illustrate the correspondence between the genes and the MOF structural features. (B) Workflow of GA. (C to E) Histograms for all HMOFs (gray) and for the initial population used in the GA runs (green). (C) Methane working capacity. (D) Gravimetric surface area. (E) Volumetric surface area. (F to H) Histograms collected from 100 GA runs show the fitness of the top-performing MOF at the end of each run: (F) Methane working capacity. (G) Gravimetric surface area. (H) Volumetric surface area. The vertical lines in (F) to (H) correspond to the fitness of the top performer from the initial population (black) and from the whole database (red). Reproduced with permission from ref. [604]. Copyright 2016 AAAS.

5.2.2. Multi-Objective Optimization.

Most design challenges in materials or catalysis discovery require satisfying multiple

objectives. For catalysis, these concerns could be activity (i.e., TOF) and stability, selectivity, solubility, or cost. As an example from heterogeneous catalysis, satisfying the objectives of cost and activity of methanation catalysts revealed an inherent trade-off in optimal properties (Figure 64).[606] The hypersurface known as the Pareto front, when assessed for a two-objective optimization, consists of the points that represent the best first property (e.g., activity) for a given second property (e.g., cost), and depending on the primary concern[606], the most desirable materials may more strongly emphasize one of these two objectives. Experiment or computation should thus aim to enrich or move beyond known Pareto front compounds with multi-objective optimization.

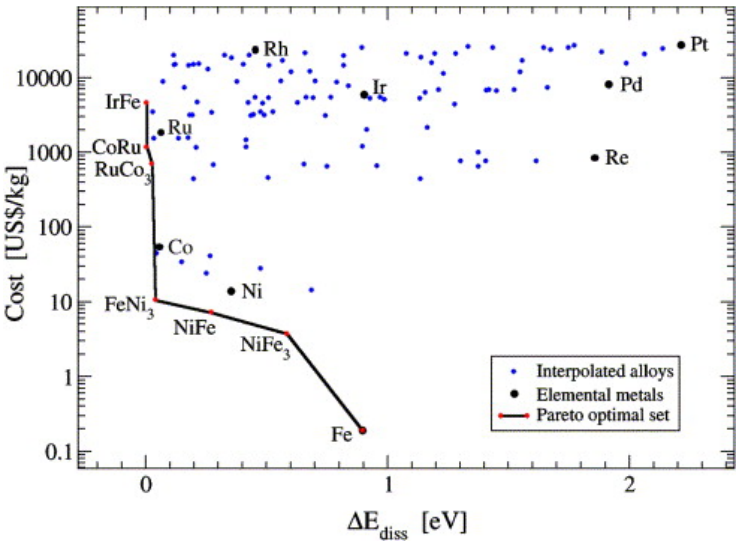

Figure 64. Pareto plot of methanation catalysts using the activity measure and the cost for 117 elemental metals and bimetallic alloys of the form A_xB_{1-x} (x=0, 0.25, 0.50, 1). Each blue point corresponds to a particular alloy. The elemental metals are shown (black), and the Pareto optimal set is also indicated (red). The cost of the bimetallic alloys has been approximated by the commodity price of the constituent elemental metals. Reproduced with permission from ref. [606]. Copyright 2006 Elsevier.

Although GAs with composite fitness functions (i.e., by multiplying the satisfaction of multiple individual objectives) have been successfully applied to transition-metal complex

discovery[468], multi-objective optimization typically benefits from algorithms less prone to finding local minima due to the complexity of the structure–property landscape along a higher-dimensional Pareto front. Multi-objective optimization has been carried out to obtain stable, low band gap oxide polymorphs of TiO_2. An evolutionary algorithm (i.e., multi-objective differential evolution or MODE) was employed to rank each generation of points based on whether they sat on the current Pareto front of these two properties evaluated with DFT.[607] Multi-objective optimization has also been applied[608] to the simultaneous optimization of two types of surface area measures for gas uptake and identification of the relevant MOF topologies that maximize these two quantities.

Multi-objective optimization methods have also been used in conjunction with high-throughput experimentation to simultaneously optimize conversion and selectivity of heterogeneous catalysts for propane dehydrogenation.[609] Llamas-Galilea et al.[609] employed a multi-objective algorithm (i.e., the strength Pareto evolutionary algorithm-2, SPEA-2) that assigns the highest fitness scores to points along the Pareto front and maximizes diversity in the Pareto set to evolve catalyst genes consisting of elements and promoters. They observed that in comparison to a single-objective optimization of yield (i.e., a composite of selectivity and conversion), the multi-objective approach did a better job of improving both variables along the Pareto front over six generations (Figure 65).[609] Recently, Bayesian optimization algorithms have been pursued in chemical problems[610] especially to account for multi-objective optimization when the evaluation of fitness is relatively costly.[611]

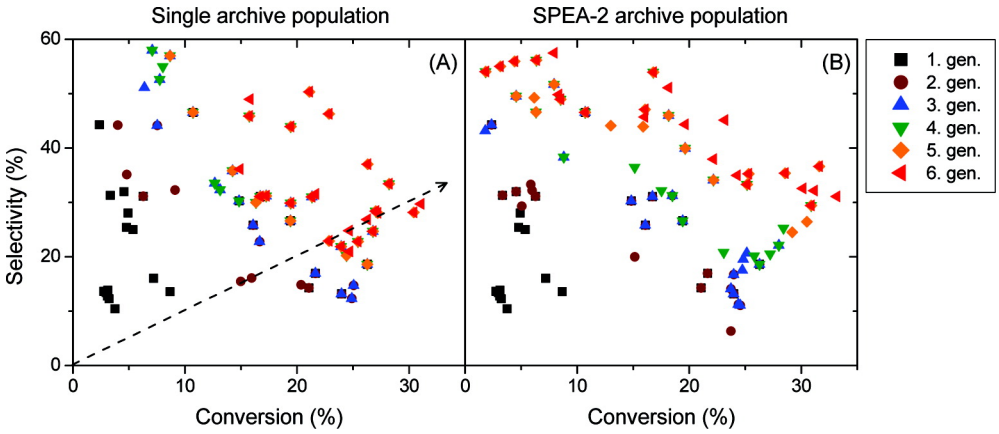

Figure 65. Visualization of the evolution of the populations of catalysts for selective dehydrogenation of propane in the objective space using a single-objective approach (left) and a multi-objective strength-Pareto evolutionary algorithm (SPEA-2, right). Reproduced with permission from ref. [609]. Copyright 2009 American Chemical Society.

5.2.3. ML-Accelerated Discovery.

Applying evolutionary algorithms to the optimization of properties in large materials spaces requires efficient evaluation of fitness. Beyond low-cost approximate computation or cheap proxies for experimental observables, a key way to achieve faster fitness evaluation is to employ a surrogate (i.e., machine learning regression) model. When employing an ML model, optimization algorithms should be adapted to address the approximate nature of the surrogate either by restricting its use to its domain of applicability or by identifying high uncertainty points with promise in active learning. Both approaches have advantages in accelerated transition-metal complex or materials discovery, as highlighted below.

Artificial neural networks (ANNs) were first used for bulk heterogeneous catalyst optimization nearly twenty years ago.[612-617] These early efforts in heterogeneous catalysis were focused on encoding compositional descriptors, such as element identity and stoichiometric contributions as well as adjustable scalars such as synthesis temperature. Corma et al.[613] trained an ANN to identify optimal reaction conditions at each generation in combination with high-throughput experimentation to optimize Ti zeolite catalysts for olefin epoxidation (Figure 66).

The ANN predictions were used to guide GA fitness evaluation, and the optimal results predicted by the ANN-GA loop were then validated experimentally (Figure 66).[613] The diversity in the initial and subsequent pools was used to aid iterative ANN retraining to predict the rank of the catalysts for epoxidation starting from a modest initial training set of 38 catalysts.[613] GAs were frequently used in combination with ANNs because the ANN optimum is not invertible back to realistic catalyst conditions. Other applications in heterogeneous catalyst screening included the design of catalyst compositions with a multitask network that predicted both selectivity and conversion for methane coupling[614] or propane dehydrogenation as well as catalyst composition and catalyst treatment temperature for methane oxidation[616] or synthesis[615].

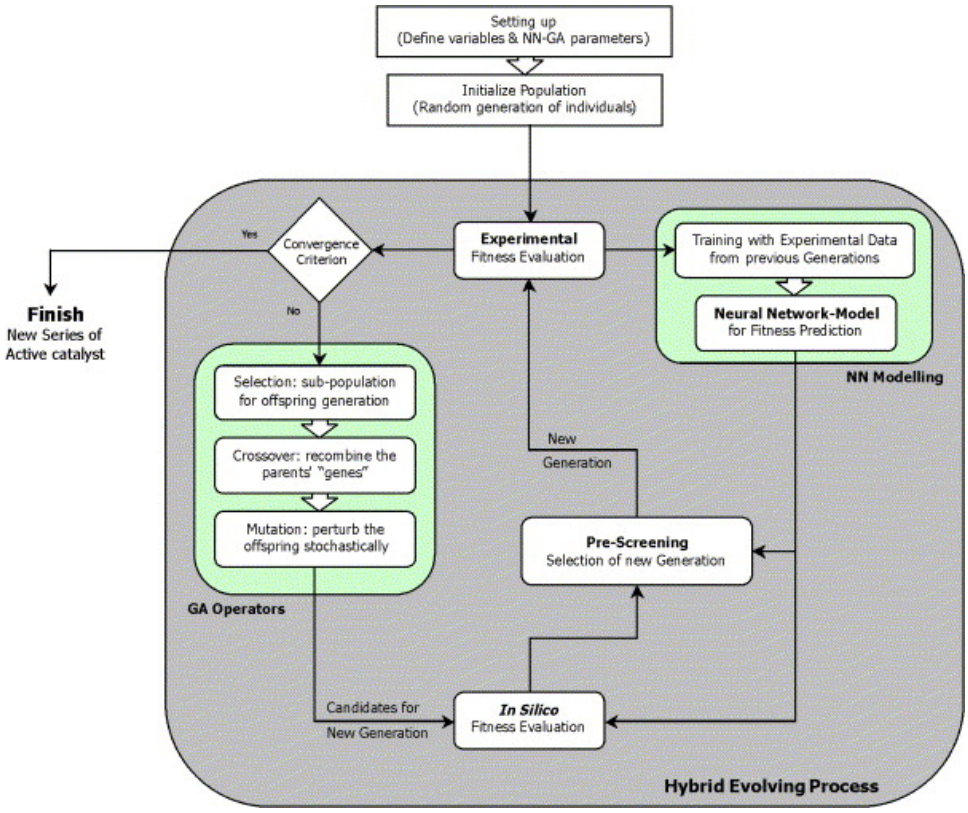

Figure 66. Scheme of a hybrid optimization algorithm comprising a genetic algorithm assisted by an artificial neural network for optimization of olefin epoxidation catalysts. High-throughput characterization and synthesis is used to optimize pH, surfactant content, and Ti content in Ti silicate materials. Reproduced with permission from ref. [613]. Copyright 2005 Elsevier.

ML-accelerated discovery has similarly been applied in multi-objective optimization of

materials and catalysts. In an early example, Scott et al.[618] used the multi-objective NSGA-II[619] algorithm in conjunction with an ANN that predicted materials properties based on a binary compositional elemental representation. They used the multi-objective optimization approach to simultaneously optimize candidate quaternary electroceramic materials for i) their permittivity, ii) whether the selected materials composition had a neutral, balanced charge, and iii) the distance to training data of the ANN.[618] Experimentally, NSGA-II was used in combination with a trained ANN that detected relationships between SCO colors as inputs and the temperature to identify the minimal SCO array for sensing applications.[620]

In comparison to efforts in heterogeneous catalysis, computational or experimental screening of molecular transition-metal complexes for homogeneous catalysis and materials discovery with deeper ML (i.e., beyond MLR QSPRs) has been a more recent development. Janet et al.[557] employed ANNs trained on the sparse MCDL-25 *ad hoc* feature set[54] to predict adiabatic DFT-quality spin-splitting energies in conjunction with a GA to explore a space of around 5k feasible spin-crossover complexes, only about 2% of which had been seen during ANN model training.[54] To ensure that the ANN prediction errors did not rise rapidly on new compounds, they introduced[557] a composite fitness function into the single objective GA that penalized compounds distant in feature space from training data (Figure 67). Most (ca. 2/3) ANN-selected leads were validated with DFT evaluation, with a modest increase in error (ca. 2x) over test set errors.[557] In comparison, using DFT as the objective function in the GA would have taken weeks instead of minutes the ANN required.[557] This approach has also been demonstrated for exploring 10–40k transition-metal complex spaces, including through the targeted design of HOMO–LUMO gaps[55] and discovery of C–H activation catalysts[468] with stable metal-oxo moieties.

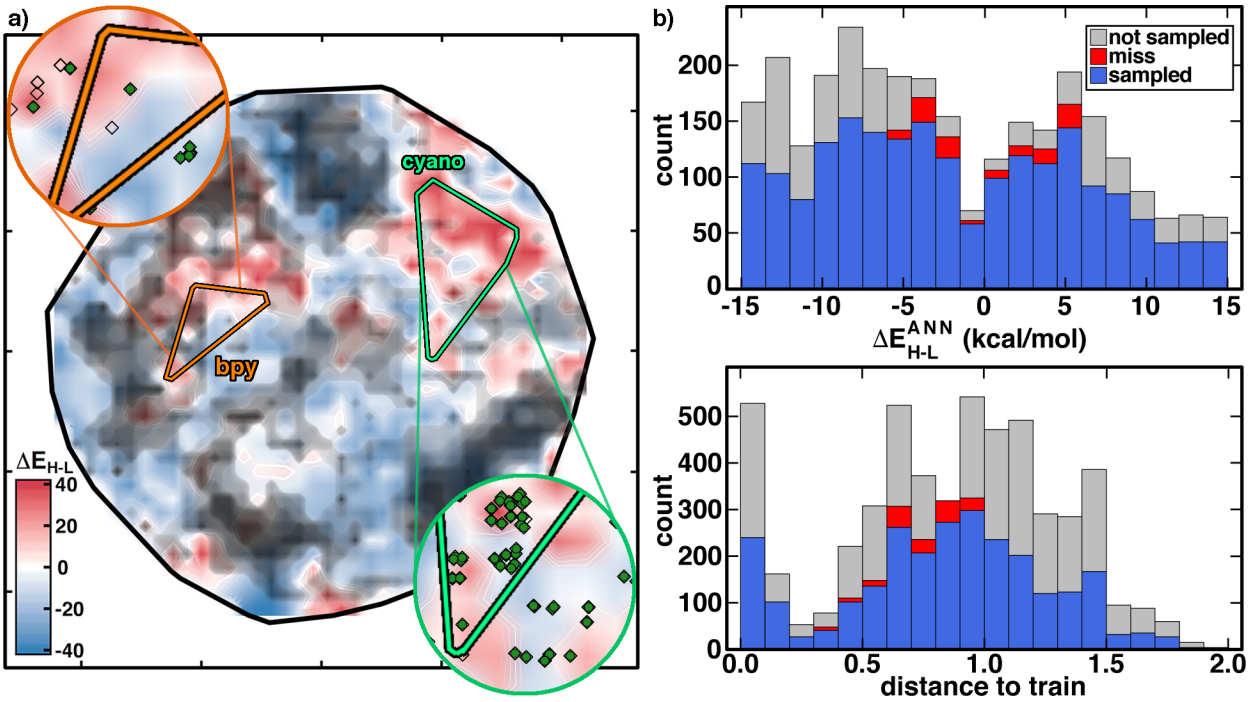

Figure 67. (a) t-SNE plot of 5600 SCO candidates colored by MCDL-25/ANN-predicted ΔE_{H-L} (in kcal/mol as indicated in inset color bar) with increasingly high distance-to-train regions indicated in darker shades of gray. The convex hulls of two families of ligands are indicated by orange and bright green triangles, respectively, with inset zooms showing the discrete hits as filled green circles. (b) 1D histograms of the MCDL-25/ANN-predicted ΔE_{H-L} (top) and distance to training data (bottom) using distance and diversity control in a GA with a stacked bar graph consisting of all sampled points (blue), non-sampled, non-hits (gray), and non-sampled hits (red). Reproduced with permission from ref. [557]. Copyright 2018 American Chemical Society.

In addition to model exploitation, where a trained surrogate is used to accelerate solving an optimization problem, ML-accelerated discovery of new materials and transition-metal complexes motivates active learning algorithms that focus on promising compounds where ML models are uncertain, also known as model exploration. For example, Jennings et al. used both a semi-empirical approximation to DFT (i.e., effective medium theory) and an ML-accelerated GA to optimize nanoparticle alloy shape prediction and reduce the full electronic structure evaluations by around 50-fold with active learning.[621] This active learning approach is particularly useful in multi-objective optimization of materials where prior training data may be limited and search spaces are large, leading to a number of recent applications in chemical and

materials discovery[563, 622-628]. Several of these efforts[563, 623] have used the expected improvement[629-630] criterion in efficient global optimization (EGO), wherein the likelihood of a new point residing beyond the Pareto front is balanced by the uncertainty of the model prediction. Because Gaussian processes (GPs) inherently provide a measure of uncertainty, they have been most widely employed in such active learning studies[621, 623-624, 627] including recently in experimental catalyst design[631] or in the selection of conditions (e.g., solvents)[632] for catalysis.

Given the high cost of generating data within an active learning framework, many studies have only been applied to evaluating algorithms in toy problems with data sets where the ground-truth best material was already known. Such strategies have enabled comparison of how leveraging surrogate models compares against heuristics in active learning or other autonomous discovery strategies.[633] As an exception in which high-cost DFT data was acquired at each generation, Janet et al.[563] applied the 2D expected improvement criterion in EGO for the simultaneous optimization of multiple properties of redox couples in redox flow batteries. Starting from a space of 2.8 M bulky (ca. 100–200-atom) transition-metal complexes, they simultaneously optimized the DFT-evaluated solubility in polar, non-aqueous electrolytes (i.e., with implicit-solvent calculated logP) and the solvent-corrected adiabatic redox potential.[563] Rather than using a GP, they noted that a multi-task ANN exhibited better lookahead errors, i.e., the model's performance on future generations of molecules it has not yet seen in training.[563] This suggested the ANN would generalize better in active learning.[563] By using a calibrated uncertainty metric based on distance to training data in the model's latent space[527], they employed the 2D-EI to select promising and uncertain points likely to expand the Pareto front (Figure 68).[563] In five generations, the Pareto front was expanded dramatically, with each generation of 2D-EI requiring nearly a week of parallelized DFT computation to obtain 20–50

new transition-metal complex property results (Figure 68). In comparison to random search, the algorithm accelerated discovery of new leads by 500-fold, which corresponded to obtaining the same result in weeks that would have taken fifty years by brute force search (Figure 68). These accelerations generally are expected to increase as more objectives are optimized simultaneously.

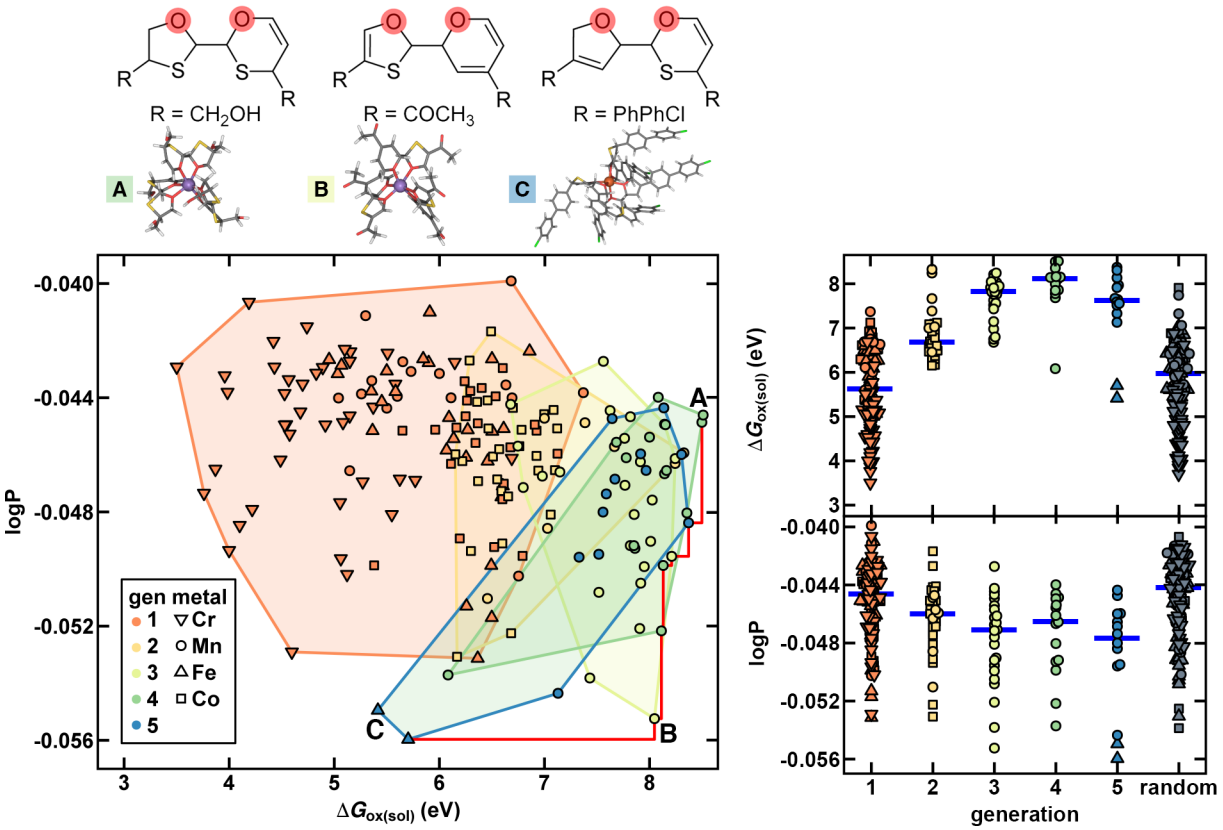

Figure 68. (Left) Redox (i.e., $\Delta G_{ox(sol)}$ in eV) and logP values for complexes simulated during 5 generations of the design algorithm, colored by generation and with unique symbols for each metal center (as indicated in inset legend). The range of values sampled in each generation is indicated by a translucent convex hull, and the final Pareto front is indicated by a red line. Three Pareto complexes along this front are labeled and shown at top. (Right) Distribution of $\Delta G_{ox(sol)}$ (top) and logP (bottom) values for each generation (colors and symbols as in left pane) alongside a random sample (gray symbols). Reproduced with permission from ref. [563]. Copyright 2020 American Chemical Society.

All methods discussed thus far have relied on introducing realism into the predictions by enumerating a hypothetical space of materials based on feasible building blocks. An alternative promising approach is using generative machine learning models to learn the distribution of

known materials (e.g., from a database of experimental structures) and predict new hypothetical materials representing this distribution. Such an approach has been developed for bulk inorganic materials[634] and porous materials such as zeolites[635] or metal–organic frameworks[636] but not yet extended to transition-metal complexes.

5.3. Automated and Autonomous High-Throughput Workflows

In large part, the QSPR (Sec. 4.2) and ML models (Sec. 4.4) as well as advances in property optimization (Sec. 5.2) have depended heavily on new strategies to share and access large data sets. In comparison to open-shell transition-metal chemistry that is the focus of this review, high-throughput screening workflows and strategies are considerably more mature in crystalline, solid-state materials as well as in small-molecule organic chemistry. Although distinct challenges are apparent for bulk periodic materials in how they are characterized or screened, some parallels are evident in the need to screen large chemical spaces for both the solid state and transition-metal complexes. Representative reviews of how high-throughput screening, large databases, and machine learning have advanced solid-state materials design or in organic reaction discovery are provided in Refs. [390, 637-638] and Ref. [639], respectively.

5.3.1. Computational High-Throughput Screening and Exploration.

A number of toolkits are available for the automation of high-throughput DFT calculations for discovery and ML model generation. Some general-purpose tools are available that are most frequently used for the screening of solid-state materials. For example, the atomic simulation environment (ASE) is a python toolkit that enables unit cell generation and interfaces to a wide number of codes, especially those that carry out periodic DFT and the analysis or control of those calculations[640-641]. The AFLOW[642] code software framework and web interface are even more tailored for the solid state, streamlining high-throughput calculation of crystalline

alloys, including through permutations of experimental structures.[643] The pymatgen python library[644] was designed to perform initial calculation setup and structure generation for solid-state materials as well as for the post-processing analysis of materials properties from raw data. The FireWorks[645] and AiiDA[646] codes are designed to optimize the procedure of automating and distributing calculation workflows as well as ensuring the provenance and reproducibility of calculations, also with a focus on solid-state DFT.

While not originally tailored necessarily for DFT calculations, a number of cheminformatics toolkits are available for studying small organic molecules. Small-molecule organic chemistry codes such as RDkit and OpenBabel[647] are widely used for high-throughput screening in organic chemistry. These tools carry out 3D structure generation from string-based molecule identifiers such as the simplified molecular input line entry system (SMILES)[648] and have integrated force field (e.g., with MMFF94 or UFF) optimization. These codes are also useful for computing molecular fingerprints and using them to quantify molecular similarity (e.g., using the Tanimoto coefficient). Both codes have interfaces to large databases of organic molecules (e.g., ChEMBL[649]).

Motivated by the distinct challenges in describing and modeling metal–organic bonding, tailored toolkits have also been developed for transition-metal complexes. One early example developed by Hay et al. was the HostDesigner code that automated construction of candidate metal ion hosts.[650] The FORTRAN code modules (i.e., LINKER and OVERLAY) were used to construct macrocycles from a fragment library in which terminal hydrogen atoms were replaced with adjacent fragments. These components were force field optimized with MM3 to generate candidate chelators of metals that were scored by the proximity of the metal-coordinating sites to the optimal position for a metal center.[650]

The molSimplify code[53] was developed to enable divide-and-conquer structure generation especially for high-throughput screening of open-shell transition-metal complexes. The code enables commandline generation of a range of transition-metal complex symmetries with ligands built from SMILES strings, large molecular databases, or a built-in dictionary of common organic ligands for inorganic complexes that users can update (Figure 69).[53] The code uses OpenBabel as a backend to force field preoptimize the individual ligands and then attach them to points around a metal center with rotations and translations to minimize steric clashes (Figure 69). The initial, spin-state-dependent metal–ligand bond lengths selected for structure generation were derived from a database of DFT values and later updated to be obtained from an ANN trained to predict[54, 527] metal–ligand bond lengths (Figure 69). This ANN improves upon general (i.e., UFF) force field bond length predictions by an order of magnitude (Figure 2).[34] The molSimplify code also interfaces to ChEMBL via OpenBabel for ligand selection. These tools have been used to carry out both pattern matching (i.e., with SMARTS) and diversity-oriented screens of organic molecules (e.g., for transition-metal complex[52] or inorganic molecule generation[438]). A custom core and decorator module can be used to start from a pre-built complex structure and then replace or modify aspects of the structure with functional groups.[53] The molSimplify code also generates input files and scripts for automating electronic structure calculations on generated individual and supramolecular complexes.

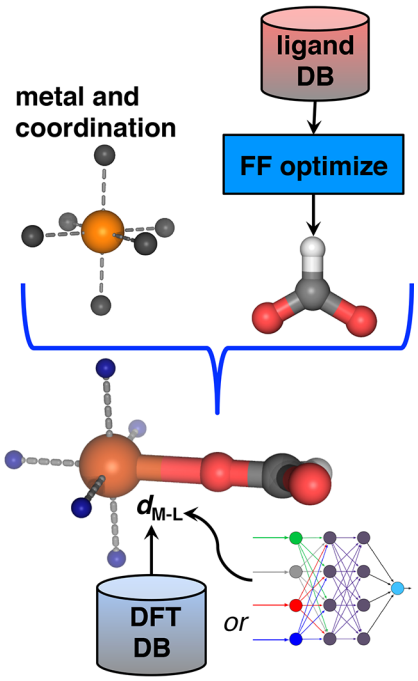

Figure 69. Schematic of molSimplify structure building: (top) the user selects a metal and coordination environment as well as ligands (e.g., from a database, DB) that are force field preoptimized; (bottom) the ligand is aligned to the metal coordination site and the metal−ligand bond distance (d_{M-L}) is set from a database of values or from an ANN. This can be followed by constrained optimization of the full complex. Reproduced with permission from ref. 52. Copyright 2017 American Chemical Society.

Foscato et al. developed methods to fragment existing transition-metal complexes and reassemble them to make new complexes[50] (Figure 70). Fragments were most readily derived from structures obtained from the CSD or PubChem. Cutting rules were developed based on pattern matches (i.e., with SMARTS) to identify where complexes should be fragmented.[50] Attachment points (APs) were then defined and used for the reassembly of new transition-metal complexes from available building blocks (Figure 70).[50] First demonstrated for building 2D structures[50], these methods were generalized through a series of fragment rotations to generate 3D structures of acceptable quality[51]. This approach was further extended to include operations that closed rings to form ligands of increased denticity from fragments[600], generalizing beyond fragments available in the CSD. In a similar fashion, the MOLASSEMBLER code is a tool that

was developed to build molecular graphs and then identify stereocenters and conformational permutations during automated 3D structure generation for both organic and inorganic molecules.[651]

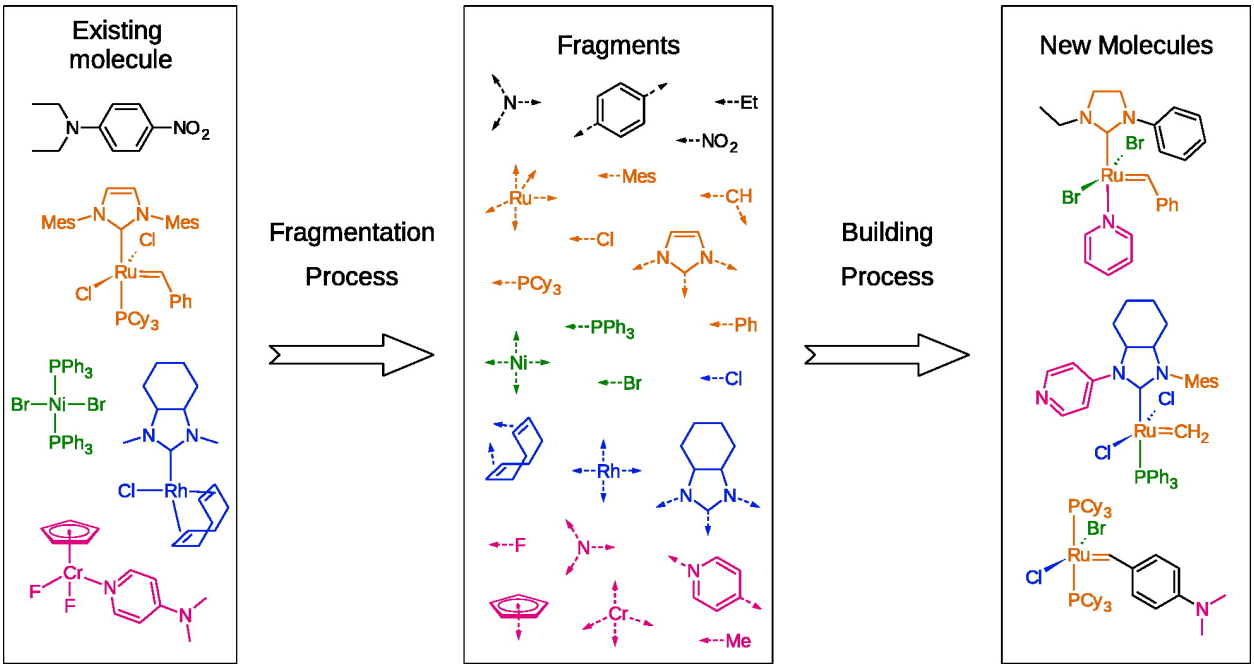

Figure 70. Schematic representation of fragmenting experimental structures of complexes and building new ones for transition-metal complex structure generation. Reproduced with permission from ref. 50. Copyright 2014 American Chemical Society.

Other codes have been specifically designed to focus on catalyst screening. The AARON code[652] focuses on computational prediction of the stereoselectivity of catalysts. AARON is applicable to reactions with well-established mechanisms but multiple possible stereocontrolling transition states to enable systematic screening and optimization of catalysts with quantum chemical calculations (Figure 71).[652] AARON relies on the existence of a good template library of prior catalysts and ligands to build catalyst structures but can detect rotatable bonds for conformer searching and transition state generation, and it automates a series of DFT calculations needed to evaluate catalysts[652] (Figure 71). These tools have been primarily demonstrated on widely studied Pd- and Rh-containing catalysts.[652-653] Like AARON, the CatVS

code uses a user-defined complex and substrate and screens a library of candidate ligands.[654] CatVS differs from AARON in that it employs force fields (i.e., TSFF and MM3, see Sec. 2) to accelerate catalyst screening.[654] Notably, the CatVS program was recently used to make computational predictions of highly enantioselective rhodium hydrogenation catalysis that were then verified experimentally to display the predicted behavior.[654] The autodE code[655] was developed to overcome some challenges in structure generation for transition-metal catalysts by fully automating catalytic intermediate structure generation from input SMILES strings.[655] The graphs of the reactants and products are used to generate guess transition states, which are refined with electronic structure calculations.[655]

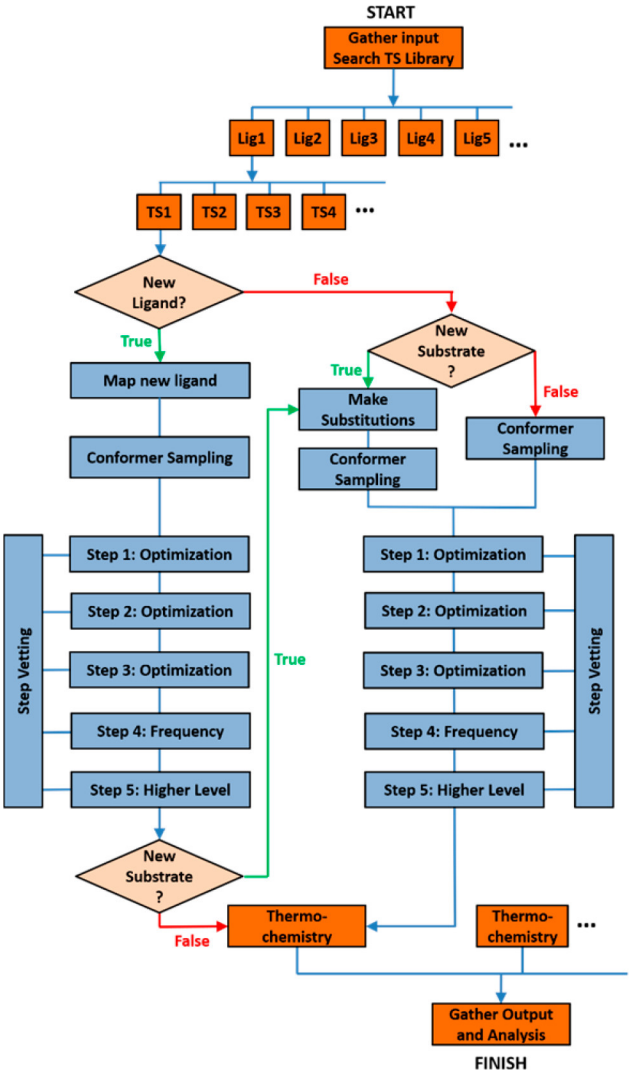

Figure 71. Overall workflow of AARON: From a library of candidate structures in the template library, AARON constructs each possible catalyst/substrate combination and locates all TS structures for these combinations, following the same workflow for each combination. It performs a series of constrained and unconstrained optimizations and frequency calculations that are analyzed. Transition state structures that have been located are used as templates for new substrates (shown at right). Reproduced with permission from ref. [652]. Copyright 2018 American Chemical Society.

Given the high computational cost of carrying out explicit DFT calculations on a large number of complexes in a screening context, toolkits have also been developed for evolutionary optimization[55, 656] (see Sec. 5.2.1). The molSimplify automatic design (mAD) module is an extension to molSimplify that carries out evolutionary (i.e., single and multi-objective genetic algorithms) optimization of transition-metal complexes represented as a series of genes. The fitness function can be evaluated directly from DFT calculations (i.e., with automated generation of input and queue script files for running electronic structure calculations) or by one of molSimplify's artificial neural networks (e.g., for spin-splitting prediction[54, 527]), and the set of genes is evolved to best satisfy the design objectives[55] (Figure 72).

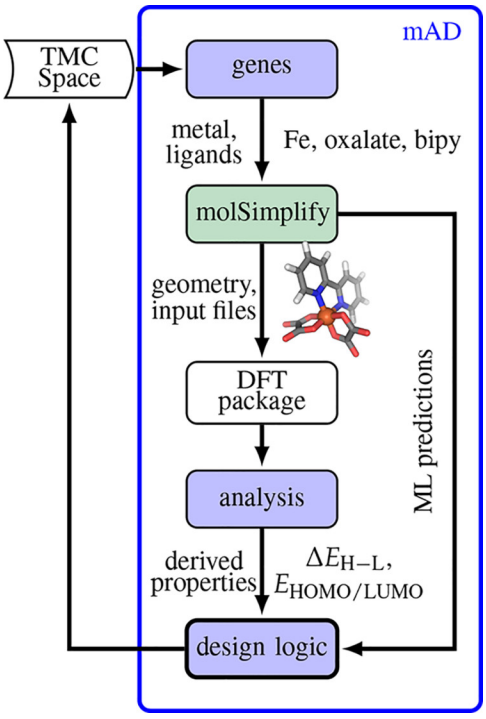

Figure 72. Workflow for mAD with key mAD components shown in blue: transition-metal chemical (TMC) space is discretized into genes and converted to 3D geometries and optional input files using molSimplify (shown in green), and properties are either calculated using an external DFT package using molSimplify-generated input files or internal ML models. Results are analyzed and guide iterative discovery using the design logic in mAD. An example 3D geometry is shown in the inset. Reproduced with permission from ref. 55. Copyright 2018 American Chemical Society.

Jensen and coworkers developed the DENOPTIM code[656] to enable evolutionary-algorithm-based design of transition-metal complexes from fragments, building upon several other earlier tools (e.g., the fragment explorer[50]) with wrapper codes (Figure 73). DENOPTIM represents the molecule as a graph[656], and the code analyzes fragments in the molecule by comparison to available databases (e.g., ChEMBL and the CSD). Mutations to the fragments of the complex are carried out through a series of routines that ensure synthetic accessibility while optimizing for the user-provided fitness function (Figure 73).[656] Because the fitness function is a user-provided external call in a BASH script, DENOPTIM supports[656] force field, ML model, or DFT calculations for the fitness evaluation if the user has these available for the problem at hand (Figure 73). The globally optimal catalyst (GOCAT) framework is another toolkit that has been developed to optimize an abstraction of the catalyst environment (i.e., the electric field) with low-cost semi-empirical theory and to use genetic algorithms to identify discrete candidates that best match the optimal field.[657] Further development is needed (i.e., including in low-cost molecular modeling) to demonstrate this approach for open-shell transition-metal catalyst design.

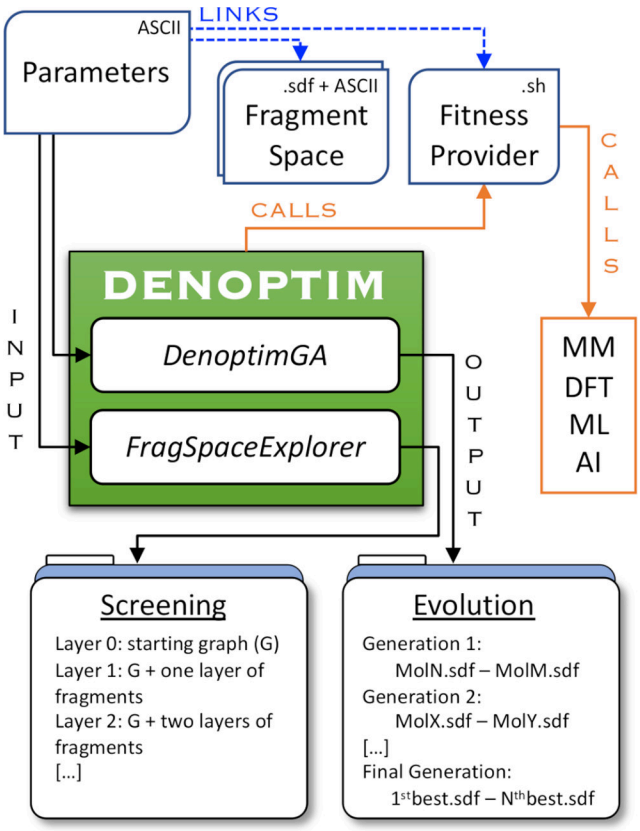

Figure 73. Schematic of the DENOPTIM code, which consists of a series of modules, including input parameters for the main interface, a genetic algorithm that uses fragments to explore chemical space, and a fitness function provider. The output from DENOPTIM is an organized series of SDF files, each containing a candidate molecule, its Cartesian coordinates, connectivity, and associated data fields such as the fitness, SMILES/InChI encoding, and other properties that the user may choose to include via the Fitness Provider. Reproduced with permission from ref. [656]. Copyright 2019 American Chemical Society.

5.3.2. Autonomous and ML-Accelerated Computational Chemistry.

The interface of data science tools with traditional modeling was described briefly in Sec. 2.5. We expand upon that section to discuss how artificial intelligence has begun to be employed in a way that can guide and accelerate the high-throughput screening workflows described in Sec. 5.3.1 for transition-metal chemistry.

For transition-metal complex and catalyst screening, a unique challenge is the high rate of failure of calculations carried out with semi-empirical methods and DFT (see Sec. 2.5) that can be difficult to predict *a priori*. At the same time, the large size and associated high computational

cost of calculations of transition-metal complexes relative to small organic molecules further highlights the importance of only initiating calculations that are likely to be successful. To automate and inform such a procedure, Duan et al. first defined[55] a series of geometric health metrics to determine when transition-metal complex geometry optimizations resulted in distortion of ligands or loss of ligand bonding to the metal. They also employed heuristic cutoffs for deviations of $<S^2>$ from its expected value of $S(S+1)$ and lower than expected localization of the spin to the metal (e.g., from Mulliken spin analysis).[55] To address the challenges of calculation failure, Duan et al. trained[519] ANN classifiers to accurately (ca. 88–95%) predict the likelihood of calculation failures using graph-based RAC descriptors.[498, 500-501] Because calculations take longer to fail than to succeed[519], they showed that these models both avoid generating spurious data for property prediction models and explore most (ca. 88%) of the feasible chemical space in 33% of the computational time.[519] Because the models employed graph-based descriptors, they required no prior calculation to predict calculation success, but their accuracy dropped rapidly on materials (e.g., high-valent metal-oxo catalysts) distinct from training data. Use of uncertainty quantification[527] improved performance on the smaller fraction of calculations for which the model was confident. Compositional descriptors have also demonstrated utility in training machine learning models to predict the computational cost of electronic structure calculations to guide calculation selection.[658]

As a complementary approach, Duan et al.[519] trained convolutional neural networks (CNNs) trained on a set of 30 transferable descriptors of the electronic structure of the complex during the first 2–40 steps of a geometry optimization without including any explicit connectivity or chemical information. This information typically corresponded to around the first 6 hours of GPU-accelerated calculation time (e.g., with TeraChem[659]) and no more than one third

of the total average time.[519] Given the strong path dependence of calculation outcomes, analysis of the CNN model's focus[519] with gradient-weighted class-activation mapping[660] revealed that some calculations took a significant number of steps before the model could determine if failure would occur. An extension[661] was incorporated into workflows for active learning[563] of methane-to-methanol catalysts[457] in order to reduce calculation attrition rates in catalyst discovery (Figure 74). This extended dynamic multi-task network model predicted three measures of calculation health (i.e., metal spin, deviations of $\langle S^2 \rangle$ from expected values, and geometry optimization outcome) based solely on properties of the wavefunction and the energetic gradient, representing an artificial intelligence model that avoided unfruitful calculations during active learning with limited prior training data (Figure 74).

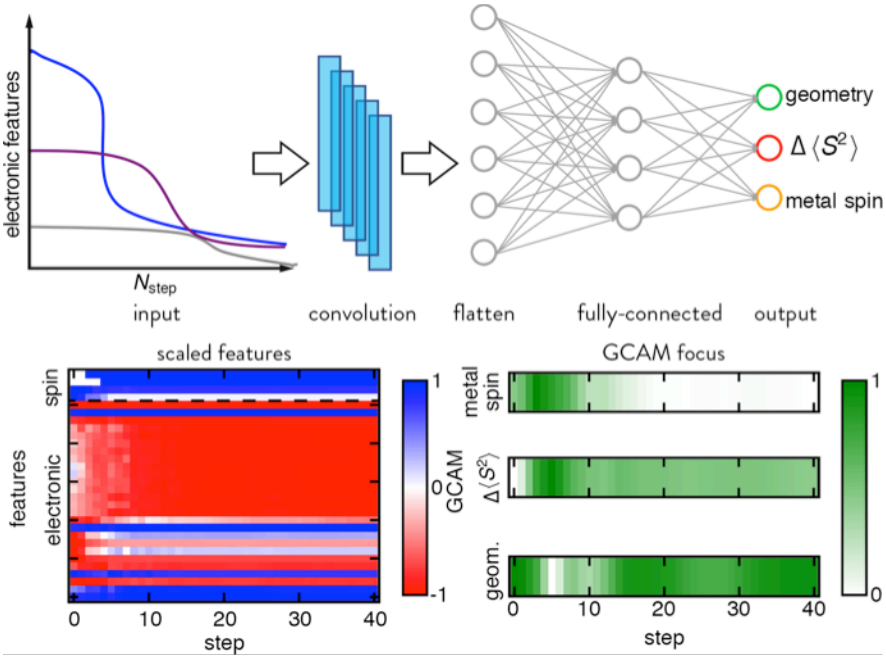

Figure 74. (top) Schematic of how electronic features are collected from the properties of the wavefunction during a geometry optimization and used as input to an ANN classifier with a convolutional layer followed by a fully connected layer. The model is a multi-task ANN classifier that predicts calculation success with respect to geometry, deviations of $\langle S^2 \rangle$ from the expected value, and the deviation of the spin on the metal from the total spin. (bottom) Gradient-weighted class-activation map (GCAM)[660] analysis of the multi-task model. In this example, the geometry optimization leads to a good result for all three properties, but features oscillate in early stages of the optimization (left). For each of the three prediction models, the GCAM focus is on different phases of the geometry optimization (right). Reproduced with permission from

ref. [661]. Copyright 2021 American Chemical Society.

In addition to sensitivity analysis and Bayesian techniques for uncertainty quantification in DFT as described in Sec. 2.5, Bayesian optimization has recently been proposed as a tool to optimize DFT functional forms.[662] To increase the domain of applicability of functionals, they proposed[662] an automated approach to functional fitting against a dataset by stochastic sub-sampling and iterative improvement of the functional coefficients with the tree-structured Parzen estimator, the same algorithm widely employed in ML architecture selection (e.g., HyperOpt). The automatically selected functionals were observed to outperform functionals developed with expert knowledge but also could fail when suitable benchmark data was insufficient to guide coefficient selection.[662]

As an alternative strategy to address the challenge of balancing cost and accuracy, McAnanama-Brereton et al.[663] demonstrated a proof-of-concept application of game theory to the selection of the appropriate density functional and basis set for studying a new molecule. Within game theory, the optimal choice for all players is selected in a so-called Nash equilibrium, and the goal of method selection was to balance expected performance, similarity to known data, and calculation cost. Using the Decider interface[663], the Tanimoto similarity of an input molecule to benchmark organic molecules with known basis set/functional performance is calculated (Figure 75). Using this information, the optimal basis set and functional are chosen to balance performance and complexity (i.e., computational cost, Figure 75).[663] While not yet developed for strongly correlated systems or beyond a limited range of DFT functionals, such an approach could be highly beneficial in the screening of transition-metal complexes.

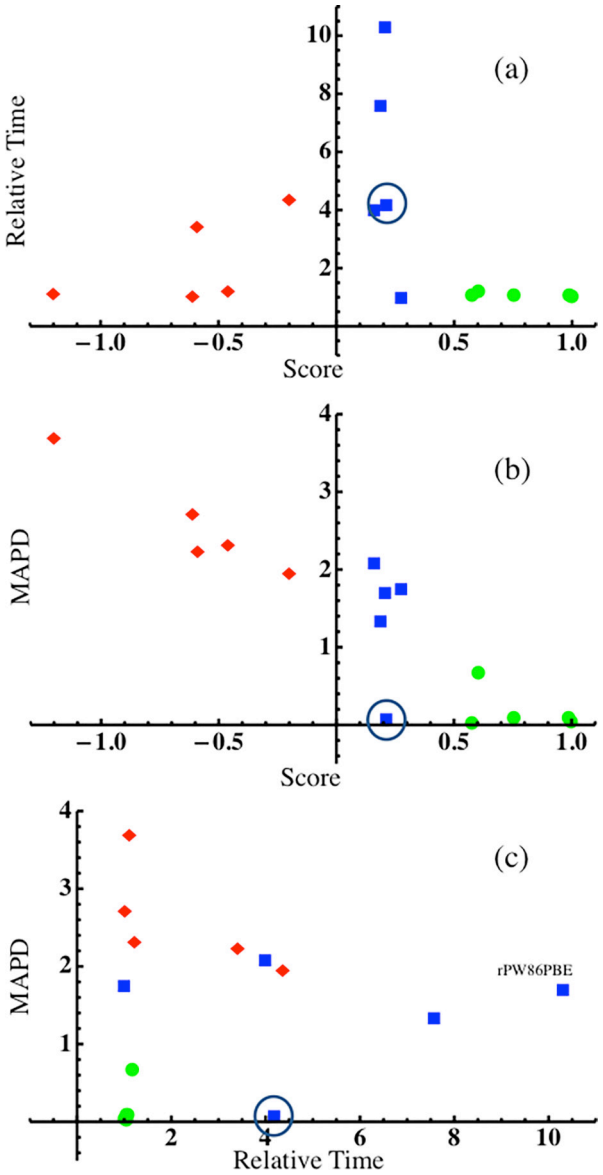

Figure 75. (a) Game theory model (i.e., *Decider*) score vs relative timing (i.e., times are relative to the fasted calculation performed). (b) *Decider* score vs mean absolute percent deviation (MAPD) to the reference GMTKN24 benchmark set (i.e., the lower the MAPD, the closer the value is to the reference value). (c) Relative timing vs MAPD for the same points. One functional, rPW86PBE, is labeled, and M06-2X is circled in blue in all graphs. The overall ranking by the *Decider* model of the 15 points are indicated as top five (green circles), middle five (blue squares), and bottom five (red circles). Reproduced with permission from ref. [663]. Copyright 2018 American Chemical Society.

A particularly relevant issue for transition-metal complexes is that the presence of strong multi-reference (MR) character could call into question results from a high-throughput screen or ML-accelerated exploration. Liu et al.[538] developed ML models to predict diagnostics of multi-

reference (MR) character at low cost, enabling rapid assessment of strong correlation.[538] Using these models, they visualized regions of an enumerated space containing complexes with relatively low- or high-MR character and identified that traditional measures such as the HOMO–LUMO gap provide an incomplete picture of what gives rise to MR character[538] (Figure 50). These models can be used alongside chemical discovery algorithms to avoid regions where low-cost methods cannot be trusted.

Although a number of MR diagnostics have been proposed[218-231, 664] in combination with heuristic cutoffs (see Sec. 2.5), their application in high-throughput screening has nevertheless been limited. Because the diagnostics have a range of relative cost and reliability, it can be challenging to know beforehand which diagnostics are most suitable for a class compounds to be screened. Sprague et al. proposed a classification analysis[665] to define MR effect as having a large error on predicted bond dissociation energies and determined that most diagnostics and cutoffs produced more false positives than negatives, with %TAE(T) providing the best performance, in part due to the close relationship between the diagnostic form and the predicted quantity[665] (Figure 76). Comparison of diagnostic pairs (e.g., geometric mean, minimum, or maximum) showed that no combination improved upon the single best-performing diagnostic[665], despite suggestions by Wilson and coworkers[232-233] that a combination of WFT-based diagnostics with adjusted thresholds was useful in predicting MR character in transition-metal complexes.

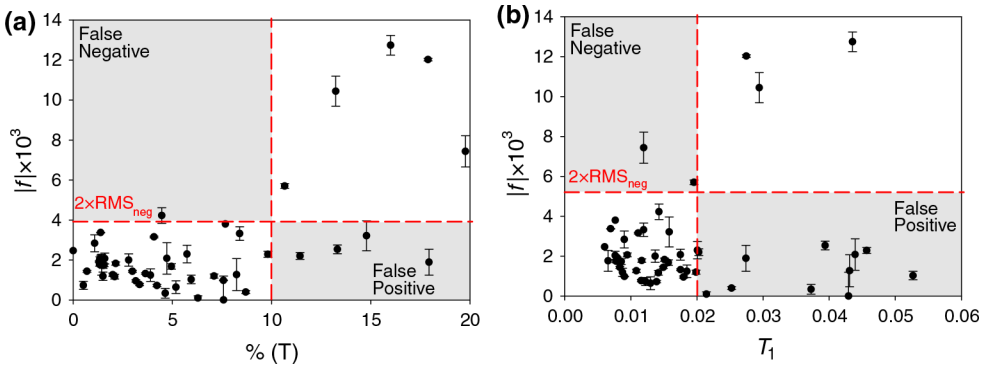

Figure 76. Relationship of unsigned fractional errors on atomization energies, |*f*|, with diagnostics %TAE(T) (left) and T_1 (right) on small organic molecules. The vertical red dashed lines correspond to the recommended diagnostics. The horizontal dashed lines correspond to twice the root mean square error used to assign false positives. Reproduced with permission from ref. [665]. Copyright 2015 Springer Nature.

Most studies to date on MR character have necessarily been limited to small molecules. Duan et al. confirmed[509] over a large data set of small organic molecules in equilibrium and distorted structures[587] that established diagnostics disagree. Using the recovery of the correlation energy as a figure of merit for diagnostic performance, they showed[509] that lower-cost diagnostics from DFT were outperformed by higher-cost diagnostics obtained directly from MR WFT (e.g., the leading weight of the CASSCF wavefunction[218-221]) that would be intractable in a high-throughput screening context. They developed ML models to overcome these limitations first by building regression models on size-independent, structure-dependent[509] descriptors (see Sec. 4.4.1) in combination with the low-cost, DFT-level diagnostics to predict high-cost, WFT-based diagnostics.

To address the disagreement of multiple WFT-based diagnostics in combination with heuristic cutoffs, Duan et al.[565] used virtual adversarial training (VAT) to train an ANN classifier of MR character. In this semi-supervised learning approach, an ANN was trained with a modified loss function that contained both a supervised and an unsupervised term, starting with labels only for the most extreme points where all diagnostics agreed. Because VAT models are

robust to noisy inputs (e.g., in image classification), the ML-model-predicted WFT-based MR diagnostics could be used as inputs to the VAT model, reducing the overall cost of MR classification to that of DFT in a manner that robustly distinguished SR from MR molecules at a level of accuracy that surpassed the conventional cutoff-based approach.[565] Such methods can be used in large-scale chemical space exploration to determine where DFT is reliable and adapt to automated MR WFT where it is not.

Within correlated wavefunction theory, ML models have been developed to predict the correlation energy[522-523, 666] or corrections on top of lower-level theories or in transfer learning models[236, 667]. A number of representations have been developed for this and related approaches, including transferable orbital-derived densities[522-523] and energies/matrix elements[666], in addition to standard 2D and 3D ML representations. With few preliminary demonstrations on transition-metal complexes[523], such techniques have largely been applied to closed-shell singlet organic molecules with unambiguous ground states at their equilibrium structures. Automated, statistical[209] and ML-informed[211] methods for active space selection discussed in Sec. 2.5 likely require further advances to enable their application in high-throughput screening in transition-metal chemistry. For example, a preliminary demonstration has shown that artificial neural networks can accelerate convergence of selected configuration interaction on small molecules by selecting the most essential configurations.[668]

For mechanism discovery, automated and semi-autonomous tools to identify the most important reaction mechanisms have been proposed to minimize computational cost or expert intervention. The artificial force-induced reaction (AFIR) of Morokuma and coworkers enables the automated identification of approximate local minima and transition states, including for reactions catalyzed by transition-metal complexes.[669] Bergeler et al. proposed[670] a computational

protocol for generation of candidate reactive transition-metal complexes from heuristic rules derived from conceptual DFT. By adding reactive protons, they analyze changes to the catalyst electronic structure to identify candidate reactive intermediates. They retain only those below an energy cutoff, and finally they refine transformations of similar intermediates with path-based transition state search (Figure 77).[670] This approach remains too computationally demanding to apply in combination with computational catalyst design.

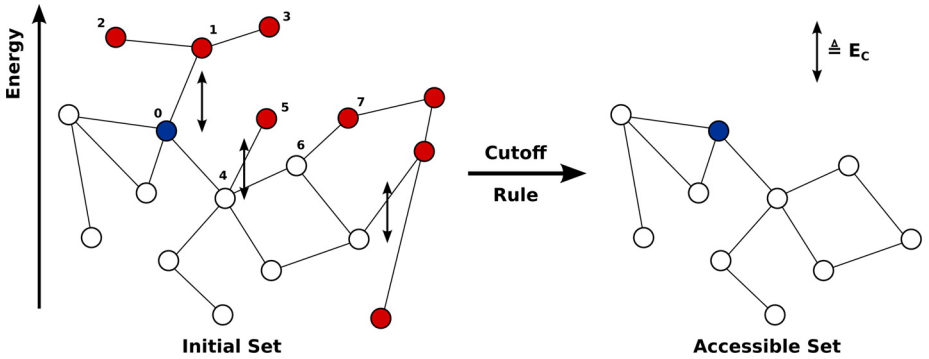

Figure 77. Illustration of the process of removing intermediates (shown as vertices) from a chemical reaction network by applying the energy cutoff E_C. The vertex representing the substrate colored blue; vertices to be removed are colored red. Reproduced with permission from ref. [670]. Copyright 2015 American Chemical Society.

Other researchers have developed fully automated workflows that leverage accelerated high-temperature *ab initio*[671-672] (e.g., Hartree–Fock) or semi-empirical[673] molecular dynamics as well as enhanced sampling[674] (e.g., metadynamics) for mechanism discovery primarily in small organic molecules. The nanoreactor[671-672] and tsscds2018[673] tools automate the normally challenging task of identifying reactive events by using heuristics to detect bond rearrangement[672-673] and then post-process these structures as starting points for automated transition state searches. Caveats remain in the application of these semi-empirical or low-cost ab initio methods in transition-metal chemistry where intermediates and transition states could be missed by lower-accuracy methods (see Sec. 2).

ML models have also been demonstrated to reduce the number of explicit calculations required during time-consuming geometry optimization and transition state search in mechanism discovery. These techniques include using surrogate models (e.g., Gaussian processes or ANNs) to estimate the local potential energy surface[675-680], to reduce the time to self-consistent energy evaluation[681], or to estimate when a candidate mechanistic step in a reaction network will be too high in energy to contribute to a reaction mechanism[682-683].

5.3.3. Machine Learning Toolkits.

The majority of tools used in machine learning model training are general-purpose in nature and developed by the computer science community. Given the fast-changing nature of the field, we only briefly mention some of these tools and their purpose. We also briefly mention some tailored tools useful for property prediction in chemistry and materials science. ML model training is most frequently carried out with scikit-learn or for more intensive models (e.g., neural networks) with Keras or PyTorch as a frontend to TensorFlow. Hyperparameter selection in ML model training, including selection of model architectures for neural networks, is too computationally demanding for exhaustive search. As a result, Bayesian optimization has been demonstrated as an approach to identify optimal architectures, e.g., in Hyperopt[684] and GPtune[685] toolkits. Recently, the COMBO toolkit[686] was developed specifically for materials science applications and includes Bayesian optimization with Thompson sampling both for the optimization of hyperparameters in neural networks as well as specifically for optimization of materials objectives. In materials science, frameworks for data mining from external databases[687] and automatic ML model building with a diverse set of *ad hoc* attributes have also been demonstrated[688-689].

Within the chemical sciences, efforts have been made to make ML and data visualization

and feature selection more user friendly. Examples of these efforts include the automated training and optimization of ML models. An early example is the DeepChem open-source project, including the MoleculeNet[690] dataset of benchmarks and model training wrappers for organic chemistry. More recent developments include the application of chemistry-specific featurizations with the ML4Chem[691] toolkit. The ChemML[692] toolkit specializes in providing a high-level wrapper for automated ML model training and visualization. Within the QSPR/QSAR community, automated tools have been developed for some time[693-695]. Special attention here has been paid to whether pharma-relevant quantities are reproducibility predicted by models (e.g., in AMPL[694]) or automated feature selection and to estimating the "learnability" of a data set[693]. None of these tools have been tailored for open-shell transition-metal chemistry to date, in part due to the smaller data sets and few available benchmarks.

5.3.4. Data Extraction and Automation of Experiments.

This review has focused on accelerated discovery of transition-metal complexes with high-throughput screening and machine learning from a computational perspective. Here, we briefly review synergistic efforts in extraction from the literature and autonomous experimentation that could inform computationally driven efforts. For discussion beyond the scope of this review, readers are encouraged to consider the more detailed reviews of Refs. [696-699]. In comparison to the outstanding challenges for autonomous computational chemistry, self-driving laboratory demonstrations have matured substantially over the past 30 years[696]. Advances include integration of AI with robotics and flow synthesis to enable autonomous organic molecule synthesis[700-704]. Both neural networks for reaction outcome prediction from literature data[705-706] and reinforcement learning with microdroplet reactions for experimental design and mechanism discovery[707] along with improvements in online characterization[702, 704, 708] have

enabled the retrosynthetic advances necessary for this autonomous experimentation (Figure 78). While somewhat less mature, robotics and AI have also been demonstrated in the synthesis and characterization of materials, such as quantum dots[709]. In a related, manual approach, recommendations from an ANN have been scored by experimentalists to guide synthesis of organic light-emitting diodes[710].

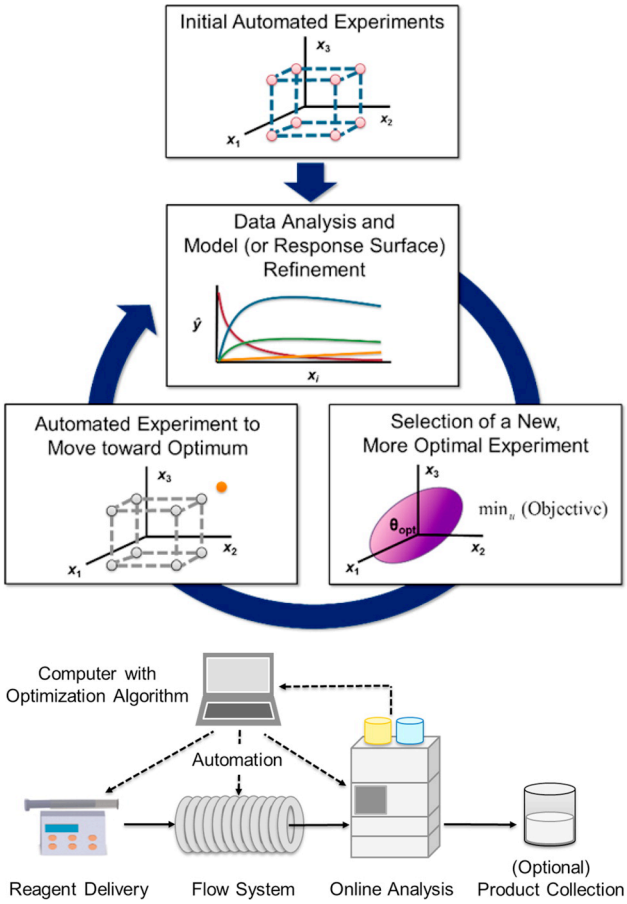

Figure 78. Computational and experimental feedback loops for automated reaction optimization with flow synthesis. Reproduced with permission from ref. [708]. Copyright 2016 American Chemical Society.

Machine learning in the form of natural language processing (NLP) has extended[385, 711-712] text-mining tools to enable extraction of information about materials properties and synthesis from the literature, as recently reviewed in Ref. [713]. Although examples have been demonstrated

for extraction of uniformly reported properties (e.g., surface area of a MOF[386]), natural language processing has been demonstrated to be essential to capture complexities in sentence structures (Figure 79). Toolkits such as chemdataextractor[712] aim to tailor NLP to common chemical phrases and properties (Figure 79). Extraction of synthesis recipes has generated sufficiently large data sets on a range of solid-state materials (e.g., metal oxides[385] or zeolites[714]) to enable prediction of which conditions improve synthesis outcomes. These large data sets have enabled unsupervised (e.g., variational autoencoder) methods to develop representations of influences on synthesis[711, 715], semi-supervised models where only a fraction of synthesis outcomes are labeled[574], as well as regression models where more chemically informed descriptors of materials properties are known[714].

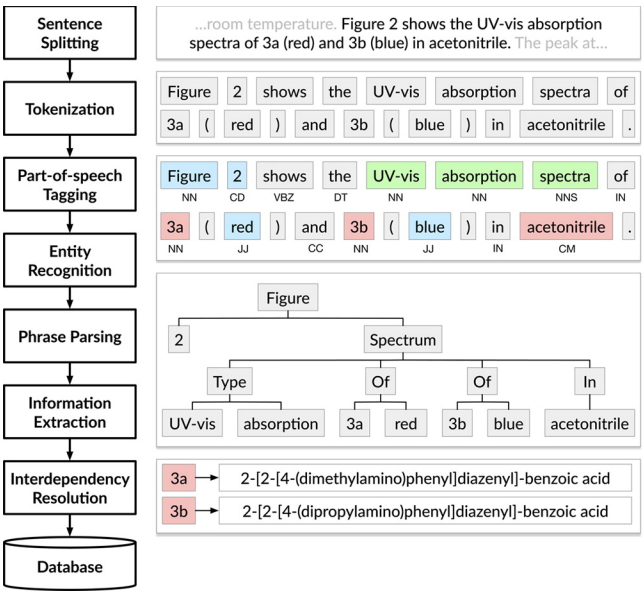

Figure 79. Schematic of a pipeline for natural language processing applied to a materials chemistry manuscript. The manuscript text is split into sentences and then tokens followed by parsing with rule-based grammar to extract individual chemical records. Example tags shown include nouns (singular: NN, plural: NNS), cardinal number (CD), verb (VBZ), determiner (DT), preposition (IN), adjective (JJ), and chemical mention (CM). Reproduced with permission from ref. [712]. Copyright 2016 American Chemical Society.

Still, limitations in extending NLP extraction efforts to transition-metal complexes and metal–organic materials property predictions remain. This is largely due to data sparsity or lack

of uniformity in the reporting of properties (e.g., in the supplemental information versus main text) as well as publication and experimentation bias toward successful results[716-719]. Nevertheless, such data should significantly augment currently available experimental data sets in transition-metal chemistry that are either small or consist only of X-ray diffraction structures (see Sec. 4.1).

6. Conclusions and Outlook.

In recent years, machine learning has emerged as an essential complement to traditional physics-based models and experiment. The advances in this area have presented unique opportunities and challenges in transition-metal chemistry due to the diversity of chemical bonding possible in a transition-metal complex as well as the interplay between spin, oxidation state, and metal electron configuration. This review has covered how machine-learning-accelerated discovery in this space has benefitted from key advances in automation, ligand chemistry, experimental data sets, and physics-based modeling. These advances have enabled improvement upon best-in-class experimental catalysts and materials. The development of uncertainty-aware search with surrogate ANN models of large spaces have led to identification of compounds with optimal properties and revealed new design principles. Further improvements have come with the automation and acceleration of computational modeling techniques.

Despite this progress, several outstanding challenges in the field remain for fully autonomous machine-learning-accelerated transition-metal chemical discovery to be realized. We will briefly outline some of these now.

A practical approach to systematically improve electronic structure method accuracy is not as well established for transition-metal chemistry as it is for organic chemistry. While automated or rule-based active space selection methods have demonstrated some promise, they

often suggest active spaces that are beyond reach of high-throughput screening workflows. Furthermore, the cost–accuracy trade-off is not well established here. There is a need still for development of low-cost methods that can robustly predict properties of open-shell transition-metal complexes, and when such a method (e.g., MM, SQM, or DFT) will perform well or fail needs to be more broadly understood in transition-metal chemistry for these choices to be easily incorporated into automated workflows. Similarly, robust benchmarks from experiment or those that eliminate uncertainty in the identity of the reacting species are an outstanding challenge in transition-metal chemistry. Extraction of larger, robust data sets will benefit both benchmarking of physics-based and data-driven models.

Transferability remains an open challenge for machine-learned representations and models. This refers both to generalization to new regions of chemical space and to better prediction from smaller data sets. For transition-metal chemistry, the balance between near- and far-sighted properties has been established to be distinct from organic chemistry. This distinction can be encoded in learned representations or in feature-selected representations. Just as force fields are often developed for a single or small range of metals, present machine learning models often only attempt to generalize to a small portion of the periodic table. Building models that truly generalize between rows of metals or across different ligand types is essential. Furthermore, interpreting the appropriate feature or latent space that accurately describes similarity of isovalent species will be essential for developing new approaches to design in transition-metal chemical space.

A perhaps even greater question only starting to be addressed is the extent to which machine learning models should simultaneously capture multiple Born-Oppenheimer potential energy surfaces. Should such models learn not just the ground state but also when the ground

state changes and in what manner, either for photoexcited species or multiple low-lying spin states?

A combinatorial challenge for data generation and chemical space exploration also remains. While enumerative strategies have been exploited for early successes in machine learning for subsets of small-molecule organic chemistry, the path to enumeration in transition-metal chemical space is less clear. Furthermore, it is not obvious to what extent chemical space should be explored in neighborhoods around known hits versus through larger variations in chemistry to overcome the multiple constraints in most practical materials design. While not the focus of this review, the twin challenge of materials discovery is mechanism discovery, especially in catalysis. Often, one must be simplified to address the other in screening, but at present there are no systematic approaches to identifying when changing the material changes the mechanism or how to design a catalyst or material with the intent of improving its stability.

Despite some of these limitations, machine-learning-accelerated discovery of transition-metal complexes is showing rapid progress across a range of applications. The inherently challenging to describe quantum mechanical nature of open-shell metal centers has particularly benefitted from non-linear machine learning models in their accelerated discovery. It is expected the next few years will bring ever more changes to the way we think about carrying out materials and molecular discovery on a computer, especially in these challenging materials spaces.

ASSOCIATED CONTENT

AUTHOR INFORMATION

Notes

The authors declare no competing financial interest.

Biographies

Aditya Nandy is a Ph.D. candidate in Chemistry in the group of Professor Heather Kulik at MIT. Prior to that, he obtained his B.S. in Chemical Engineering from UC Berkeley in 2017, with a minor in Chemistry.

Chenru Duan is a Ph.D. candidate in Chemistry in the group of Professor Heather Kulik at MIT. Prior to that, he obtained his B.S. in Physics from Zhejiang University in China in 2017.

Michael G. Taylor is a Postdoctoral Researcher in the group of Professor Heather J. Kulik at MIT. Prior to that, he obtained his Ph.D. in Chemical Engineering from the University of Pittsburgh in the group of Prof. Giannis Mpourmpakis in 2019. He obtained his B.S. in Chemical Engineering from the University of Nebraska-Lincoln in 2013.

Fang Liu is an Assistant Professor in Chemistry at Emory University. Prior to that, she did postdoctoral research in the group of Professor Heather J. Kulik at MIT (2017–2020) and obtained her Ph.D. in Chemistry at Stanford in the group of Todd J. Martínez in 2017.

Adam H. Steeves is a Research Associate in the group of Professor Heather J. Kulik at MIT. He was previously an Assistant Professor of Chemistry at Ithaca College. Prior to that, he received his Ph.D. in Physical Chemistry from MIT in 2009 and was postdoctoral fellow at UCSF.

Heather J. Kulik is an Associate Professor in the Department of Chemical Engineering at MIT. She received her B.E. in Chemical Engineering from the Cooper Union in 2004. She obtained her Ph.D. from the Department of Materials Science and Engineering at MIT in 2009 in the group of Nicola Marzari. She completed postdocs at Lawrence Livermore with Felice Lightstone and Stanford with Todd Martínez, prior to joining MIT as a faculty member in 2013.

ACKNOWLEDGMENT

The authors acknowledge support by the Office of Naval Research under grant numbers N00014-17-1-2956, N00014-18-1-2434, and N00014-20-1-2150, DARPA grant D18AP00039, the Department of Energy under grant numbers DE-NA0003965, DE-SC0019112, DE-SC0018096, and DE-SC0012702, and the National Science Foundation under grant numbers ECCS-1449291, CBET-1704266, and CBET-1846426. F.L. was partially supported by a MolSSI fellowship (grant no. ACI-1547580). A.N. was partially supported by a National Science Foundation Graduate Research Fellowship under Grant #1122374. This work was supported by an AAAS Marion Milligan Mason Award. H.J.K. holds a Career Award at the Scientific Interface from the Burroughs Wellcome Fund, which partially supported this work.

References

1. Li, P.; Merz Jr, K. M., Metal Ion Modeling Using Classical Mechanics. *Chem. Rev.* **2017**, *117*, 1564-1686.
2. Vogiatzis, K. D.; Polynski, M. V.; Kirkland, J. K.; Townsend, J.; Hashemi, A.; Liu, C.; Pidko, E. A., Computational Approach to Molecular Catalysis by 3d Transition Metals: Challenges and Opportunities. *Chem. Rev.* **2018**, *119*, 2453-2523.
3. Durand, D. J.; Fey, N., Computational Ligand Descriptors for Catalyst Design. *Chem. Rev.* **2019**, *119*, 6561-6594.
4. Foscato, M.; Jensen, V. R., Automated in Silico Design of Homogeneous Catalysts. *ACS Catal.* **2020**, *10*, 2354-2377.
5. Cramer, C. J.; Truhlar, D. G., Density Functional Theory for Transition Metals and Transition Metal Chemistry. *Phys. Chem. Chem. Phys.* **2009**, *11*, 10757-10816.
6. Kulik, H. J., Perspective: Treating Electron over-Delocalization with the DFT+U Method. *J. Chem. Phys.* **2015**, *142*, 240901.
7. Swart, M.; Gruden, M., Spinning around in Transition-Metal Chemistry. *Acc. Chem. Res.* **2016**, *49*, 2690-2697.
8. Gaggioli, C. A.; Stoneburner, S. J.; Cramer, C. J.; Gagliardi, L., Beyond Density Functional Theory: The Multiconfigurational Approach to Model Heterogeneous Catalysis. *ACS Catal.* **2019**, *9*, 8481-8502.
9. Hastie, T.; Tibshirani, R.; Friedman, J., *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*; Springer Science & Business Media, 2009.
10. Janet, J. P.; Kulik, H. J., *Machine Learning in Chemistry*, 2020.
11. Wolfsberg, M.; Helmholz, L., The Spectra and Electronic Structure of the Tetrahedral Ions Mno4−, Cro4−−, and Clo4−. *J. Chem. Phys.* **1952**, *20*, 837-843.
12. Fenske, R. F., Semiempirical Molecular Orbital Theory for Transition Metal Complexes. *Inorg. Chem.* **1965**, *4*, 33-36.

13. Hall, M. B.; Fenske, R. F., Electronic Structure and Bonding in Methyl- and Perfluoromethyl(Pentacarbonyl)Manganese. *Inorg. Chem.* **1972**, *11*, 768-775.
14. Clack, D. W.; Hush, N. S.; Yandle, J. R., All-Valence Electron CNDO Calculations on Transition Metal Complexes. *J. Chem. Phys.* **1972**, *57*, 3503-3510.
15. Norrby, P.-O.; Brandt, P., Deriving Force Field Parameters for Coordination Complexes. *Coord. Chem. Rev.* **2001**, *212*, 79-109.
16. Burton, V. J.; Deeth, R. J.; Kemp, C. M.; Gilbert, P. J., Molecular Mechanics for Coordination Complexes: The Impact of Adding d-Electron Stabilization Energies. *J. Am. Chem. Soc.* **1995**, *117*, 8407-8415.
17. Deeth, R. J.; Anastasi, A.; Diedrich, C.; Randell, K., Molecular Modelling for Transition Metal Complexes: Dealing with d-Electron Effects. *Coord. Chem. Rev.* **2009**, *253*, 795-816.
18. Comba, P.; Zimmer, M., Molecular Mechanics and the Jahn-Teller Effect. *Inorg. Chem.* **1994**, *33*, 5368-5369.
19. Comba, P.; Remenyi, R., Inorganic and Bioinorganic Molecular Mechanics Modeling—the Problem of the Force Field Parameterization. *Coord. Chem. Rev.* **2003**, *238-239*, 9-20.
20. Rappe, A. K.; Colwell, K. S.; Casewit, C. J., Application of a Universal Force Field to Metal Complexes. *Inorg. Chem.* **1993**, *32*, 3438-3450.
21. Zerner, M.; Gouterman, M.; Kobayashi, H., Porphyrins. *Theor. Chim. Acta* **1966**, *6*, 363-400.
22. Zerner, M. C.; Loew, G. H.; Kirchner, R. F.; Mueller-Westerhoff, U. T., An Intermediate Neglect of Differential Overlap Technique for Spectroscopy of Transition-Metal Complexes. Ferrocene. *J. Am. Chem. Soc.* **1980**, *102*, 589-599.
23. Bacon, A. D.; Zerner, M. C., An Intermediate Neglect of Differential Overlap Theory for Transition Metal Complexes: Fe, Co and Cu Chlorides. *Theor. Chim. Acta* **1979**, *53*, 21-54.
24. Gaus, M.; Cui, Q.; Elstner, M., DFTB3: Extension of the Self-Consistent-Charge Density-Functional Tight-Binding Method (SCC-DFTB). *J. Chem. Theory Comput.* **2011**, *7*, 931-948.
25. Grimme, S.; Bannwarth, C.; Shushkov, P., 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.
26. Rappe, A. K.; Casewit, C. J.; Colwell, K. S.; Goddard, W. A.; Skiff, W. M., UFF, a Full Periodic Table Force Field for Molecular Mechanics and Molecular Dynamics Simulations. *J. Am. Chem. Soc.* **1992**, *114*, 10024-10035.
27. Root, D. M.; Landis, C. R.; Cleveland, T., Valence Bond Concepts Applied to the Molecular Mechanics Description of Molecular Shapes. 1. Application to Nonhypervalent Molecules of the P-Block. *J. Am. Chem. Soc.* **1993**, *115*, 4201-4209.
28. Tubert-Brohman, I.; Schmid, M.; Meuwly, M., Molecular Mechanics Force Field for Octahedral Organometallic Compounds with Inclusion of the Trans Influence. *J. Chem. Theory Comput.* **2009**, *5*, 530-539.
29. Coupry, D. E.; Addicoat, M. A.; Heine, T., Extension of the Universal Force Field for Metal–Organic Frameworks. *J. Chem. Theory Comput.* **2016**, *12*, 5215-5225.
30. Bureekaew, S.; Amirjalayer, S.; Tafipolsky, M.; Spickermann, C.; Roy, T. K.; Schmid, R., Mof-Ff – a Flexible First-Principles Derived Force Field for Metal-Organic Frameworks. *Phys. Status Solidi B* **2013**, *250*, 1128-1141.

31. Bol, J. E.; Buning, C.; Comba, P.; Reedijk, J.; Ströhle, M., Molecular Mechanics Modeling of Organic Backbone of Metal-Free and Coordinated Ligands. *J. Comput. Chem.* **1998**, *19*, 512-523.
32. Mayo, S. L.; Olafson, B. D.; Goddard, W. A., Dreiding: A Generic Force Field for Molecular Simulations. *J. Phys. Chem.* **1990**, *94*, 8897-8909.
33. Shi, S.; Yan, L.; Yang, Y.; Fisher-Shaulsky, J.; Thacher, T., An Extensible and Systematic Force Field, Esff, for Molecular Modeling of Organic, Inorganic, and Organometallic Systems. *J. Comput. Chem.* **2003**, *24*, 1059-1076.
34. Janet, J. P.; Liu, F.; Nandy, A.; Duan, C.; Yang, T.; Lin, S.; Kulik, H. J., Designing in the Face of Uncertainty: Exploiting Electronic Structure and Machine Learning Models for Discovery in Inorganic Chemistry. *Inorg. Chem.* **2019**, *58*, 10592-10606.
35. Kulik, H. J., Making Machine Learning a Useful Tool in the Accelerated Discovery of Transition Metal Complexes. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2020**, *10*, e1439.
36. Allinger, N. L.; Yuh, Y. H.; Lii, J. H., Molecular Mechanics. The Mm3 Force Field for Hydrocarbons. 1. *J. Am. Chem. Soc.* **1989**, *111*, 8551-8566.
37. Hunger, J.; Beyreuther, S.; Huttner, G.; Allinger, K.; Radelof, U.; Zsolnai, L., How to Derive Force Field Parameters by Genetic Algorithms: Modelling Tripod-Mo(CO)3 Compounds as an Example. *Eur. J. Inorg. Chem.* **1998**, *1998*, 693-702.
38. Allured, V. S.; Kelly, C. M.; Landis, C. R., Shapes Empirical Force Field: New Treatment of Angular Potentials and Its Application to Square-Planar Transition-Metal Complexes. *J. Am. Chem. Soc.* **1991**, *113*, 1-12.
39. Landis, C. R.; Cleveland, T.; Firman, T. K., Making Sense of the Shapes of Simple Metal Hydrides. *J. Am. Chem. Soc.* **1995**, *117*, 1859-1860.
40. Landis, C. R.; Cleveland, T.; Firman, T. K., Valence Bond Concepts Applied to the Molecular Mechanics Description of Molecular Shapes. 3. Applications to Transition Metal Alkyls and Hydrides. *J. Am. Chem. Soc.* **1998**, *120*, 2641-2649.
41. Deeth, R. J.; Hitchman, M. A., Factors Influencing Jahn-Teller Distortions in Six-Coordinate Copper(II) and Low-Spin Nickel(II) Complexes. *Inorg. Chem.* **1986**, *25*, 1225-1233.
42. Coe, B. J.; Glenwright, S. J., Trans-Effects in Octahedral Transition Metal Complexes. *Coord. Chem. Rev.* **2000**, *203*, 5-80.
43. Kanal, I. Y.; Keith, J. A.; Hutchison, G. R., A Sobering Assessment of Small-Molecule Force Field Methods for Low Energy Conformer Predictions. *Int. J. Quantum Chem.* **2018**, *118*, e25512.
44. Besora, M.; Braga, A. A. C.; Ujaque, G.; Maseras, F.; Lledós, A., The Importance of Conformational Search: A Test Case on the Catalytic Cycle of the Suzuki–Miyaura Cross-Coupling. *Theor. Chem. Acc.* **2011**, *128*, 639-646.
45. Vitek, A. K.; Jugovic, T. M.; Zimmerman, P. M., Revealing the Strong Relationships between Ligand Conformers and Activation Barriers: A Case Study of Bisphosphine Reductive Elimination. *ACS Catal.* **2020**, *10*, 7136-7145.
46. Bartol, J.; Comba, P.; Melter, M.; Zimmer, M., Conformational Searching of Transition Metal Compounds. *J. Comput. Chem.* **1999**, *20*, 1549-1558.
47. Balcells, D.; Drudis-Solé, G.; Besora, M.; Dölker, N.; Ujaque, G.; Maseras, F.; Lledós, A., Some Critical Issues in the Application of Quantum Mechanics/Molecular Mechanics Methods to the Study of Transition Metal Complexes. *Faraday Discuss.* **2003**, *124*, 429-441.
48. Gillespie, A. M.; Morello, G. R.; White, D. P., De Novo Ligand Design: Understanding Stereoselective Olefin Binding to [(H5-C5h5)Re(NO)(PPh3)]+ with Molecular Mechanics,

Semiempirical Quantum Mechanics, and Density Functional Theory. *Organometallics* **2002**, *21*, 3913-3921.
49. Buda, C.; Burt, S. K.; Cundari, T. R.; Shenkin, P. S., De Novo Structural Prediction of Transition Metal Complexes: Application to Technetium. *Inorg. Chem.* **2002**, *41*, 2060-2069.
50. Foscato, M.; Occhipinti, G.; Venkatraman, V.; Alsberg, B. K.; Jensen, V. R., Automated Design of Realistic Organometallic Molecules from Fragments. *J. Chem. Inf. Model.* **2014**, *54*, 767-780.
51. Foscato, M.; Venkatraman, V.; Occhipinti, G.; Alsberg, B. K.; Jensen, V. R., Automated Building of Organometallic Complexes from 3D Fragments. *J. Chem. Inf. Model.* **2014**, *54*, 1919-1931.
52. Janet, J. P.; Gani, T. Z. H.; Steeves, A. H.; Ioannidis, E. I.; Kulik, H. J., Leveraging Cheminformatics Strategies for Inorganic Discovery: Application to Redox Potential Design. *Ind. Eng. Chem. Res.* **2017**, *56*, 4898-4910.
53. Ioannidis, E. I.; Gani, T. Z. H.; Kulik, H. J., molSimplify: A Toolkit for Automating Discovery in Inorganic Chemistry. *J. Comput. Chem.* **2016**, *37*, 2106-2117.
54. Janet, J. P.; Kulik, H. J., Predicting Electronic Structure Properties of Transition Metal Complexes with Neural Networks. *Chem. Sci.* **2017**, *8*, 5137-5152.
55. Nandy, A.; Duan, C.; Janet, J. P.; Gugler, S.; Kulik, H. J., Strategies and Software for Machine Learning Accelerated Discovery in Transition Metal Chemistry. *Ind. Eng. Chem. Res.* **2018**, *57*, 13973-13986.
56. Schäffer, C. E.; Jørgensen, C. K., The Angular Overlap Model, an Attempt to Revive the Ligand Field Approaches. *Mol. Phys.* **1965**, *9*, 401-412.
57. Bernhardt, P. V.; Comba, P., Prediction and Interpretation of Electronic Spectra of Transition Metal Complexes via the Combination of Molecular Mechanics and Angular Overlap Model Calculations. *Inorg. Chem.* **1993**, *32*, 2798-2803.
58. Bernhardt, P. V.; Comba, P., Molecular Mechanics Calculations of Transition Metal Complexes. *Inorg. Chem.* **1992**, *31*, 2638-2644.
59. Comba, P.; Ströhle, M.; Hambley, T. W., The Directionality of d-Orbitals and Molecular-Mechanics Calculations of Octahedral Transition-Metal Compounds. *Helv. Chim. Acta* **1995**, *78*, 2042-2047.
60. Glerup, J.; Moensted, O.; Schaeffer, C. E., Nonadditive and Additive Ligand Fields and Spectrochemical Series Arising from Ligand Field Parameterization Schemes. Pyridine as a Nonlinearly Ligating .Pi.-Back-Bonding Ligand toward Chromium(III). *Inorg. Chem.* **1976**, *15*, 1399-1407.
61. Handley, C. M.; Deeth, R. J., A Multi-Objective Approach to Force Field Optimization: Structures and Spin State Energetics of d6 Fe(II) Complexes. *J. Chem. Theory Comput.* **2012**, *8*, 194-202.
62. Deeth, R. J.; Foulis, D. L.; Williams-Hubbard, B. J., Molecular Mechanics for Multiple Spin States of Transition Metal Complexes. *Dalton Trans.* **2003**, 3949-3955.
63. Deeth, R. J.; Fey, N.; Williams–Hubbard, B., Dommimoe: An Implementation of Ligand Field Molecular Mechanics in the Molecular Operating Environment. *J. Comput. Chem.* **2005**, *26*, 123-130.
64. Deeth, R. J.; Randell, K., Ligand Field Stabilization and Activation Energies Revisited: Molecular Modeling of the Thermodynamic and Kinetic Properties of Divalent, First-Row Aqua Complexes. *Inorg. Chem.* **2008**, *47*, 7377-7388.

65. Piquemal, J.-P.; Williams-Hubbard, B.; Fey, N.; Deeth, R. J.; Gresh, N.; Giessner-Prettre, C., Inclusion of the Ligand Field Contribution in a Polarizable Molecular Mechanics: Sibfa-Lf. *J. Comput. Chem.* **2003**, *24*, 1963-1970.
66. Foscato, M.; Deeth, R. J.; Jensen, V. R., Integration of Ligand Field Molecular Mechanics in Tinker. *J. Chem. Inf. Model.* **2015**, *55*, 1282-1290.
67. Pham, C. H.; Cirera, J.; Paesani, F., Molecular Mechanisms of Spin Crossover in the {Fe(Pz)[Pt(Cn)4]} Metal–Organic Framework Upon Water Adsorption. *J. Am. Chem. Soc.* **2016**, *138*, 6123-6126.
68. Dajnowicz, S.; Ghoreishi, D.; Modugula, K.; Damm, W.; Harder, E. D.; Abel, R.; Wang, L.; Yu, H. S., Advancing Free-Energy Calculations of Metalloenzymes in Drug Discovery via Implementation of LFMM Potentials. *J. Chem. Theory Comput.* **2020**, *16*, 6926-6937.
69. Eksterowicz, J. E.; Houk, K., Transition-State Modeling with Empirical Force Fields. *Chem. Rev.* **1993**, *93*, 2439-2461.
70. Norrby, P.-O., Selectivity in Asymmetric Synthesis from QM-Guided Molecular Mechanics. *J. Mol. Struct.: THEOCHEM* **2000**, *506*, 9-16.
71. Hansen, E.; Rosales, A. R.; Tutkowski, B.; Norrby, P.-O.; Wiest, O., Prediction of Stereochemistry Using Q2mm. *Acc. Chem. Res.* **2016**, *49*, 996-1005.
72. Jensen, F.; Norrby, P. O., Transition States from Empirical Force Fields. *Theor. Chem. Acc.* **2003**, *109*, 1-7.
73. Norrby, P.-O.; Liljefors, T., Automated Molecular Mechanics Parameterization with Simultaneous Utilization of Experimental and Quantum Mechanical Data. *J. Comput. Chem.* **1998**, *19*, 1146-1166.
74. Rosales, A. R.; Ross, S. P.; Helquist, P.; Norrby, P.-O.; Sigman, M. S.; Wiest, O., Transition State Force Field for the Asymmetric Redox-Relay Heck Reaction. *J. Am. Chem. Soc.* **2020**, *142*, 9700-9707.
75. 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.* **2021**, *11*, e1493.
76. Husch, T.; Vaucher, A. C.; Reiher, M., Semiempirical Molecular Orbital Models Based on the Neglect of Diatomic Differential Overlap Approximation. *Int. J. Quantum Chem.* **2018**, *118*, e25799.
77. Thiel, W., Semiempirical Quantum–Chemical Methods. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2014**, *4*, 145-157.
78. Filatov, M. J.; Gritsenko, O. V.; Zhidomirov, G. M., CNDO-S2—a Semiempirical SCF MO Method for Transition Metal Organometallics. *Theor. Chim. Acta* **1987**, *72*, 211-222.
79. Filatov, M. J.; Gritsenko, O. V.; Zhidomirov, G. M., CNDO-S2: A New Semiempirical SCF MO Method for Transition Metal Organometallics. *J. Mol. Catal.* **1989**, *54*, 462-477.
80. Sakaki, S.; Nishikawa, M.; Tsuru, N.; Yamashita, T.; Ohyoshi, A., A Modified Indo-MO Study of Electronic Spectra of Open-Shell Transition Metal Complexes. *J. Inorg. Nucl. Chem.* **1979**, *41*, 673-680.
81. Blyholder, G.; Head, J.; Ruette, F., Semi-Empirical Calculation Method for Transition Metals. *Theor. Chim. Acta* **1982**, *60*, 429-444.
82. Ivanova, N. V.; Sizova, O. V.; Baranovskii, V. I., Application of Indo Method to Calculation of Electronic Structure and Electronic Spectra of Transition-Metal Complexes. *J. Struct. Chem.* **1983**, *23*, 838-844.

83. Anderson, W. P.; Cundari, T. R.; Zerner, M. C., An Intermediate Neglect of Differential Overlap Model for Second-Row Transition Metal Species. *Int. J. Quantum Chem.* **1991**, *39*, 31-45.
84. Li, J.; De Mello, P. C.; Jug, K., Extension of Sindo1 to Transition Metal Compounds. *J. Comput. Chem.* **1992**, *13*, 85-92.
85. Filatov, M. J.; Zilberberg, I. L.; Zhidomirov, G. M., NDDO/Mc: A New Semiempirical Scfmo Method for Transition Metal Complexes. *Int. J. Quantum Chem.* **1992**, *44*, 565-585.
86. Voityuk, A. A., Intermediate Neglect of Differential Overlap for Spectroscopy. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2013**, *3*, 515-527.
87. Harris, D.; Loew, G. H.; Komornicki, A., Structure and Relative Spin-State Energetics of [Fe(H2O)6]3+: A Comparison of Uhf, Møller−Plesset, Nonlocal DFT, and Semiempircal Indo/S Calculations. *J. Phys. Chem. A* **1997**, *101*, 3959-3965.
88. Stewart, J. J. P., Optimization of Parameters for Semiempirical Methods V: Modification of NDDO Approximations and Application to 70 Elements. *J. Mol. Model.* **2007**, *13*, 1173-1213.
89. Stewart, J. J. P., Optimization of Parameters for Semiempirical Methods VI: More Modifications to the NDDO Approximations and Re-Optimization of Parameters. *J. Mol. Model.* **2013**, *19*, 1-32.
90. Dewar, M. J.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J., Development and Use of Quantum Mechanical Molecular Models. 76. Am1: A New General Purpose Quantum Mechanical Molecular Model. *J. Am. Chem. Soc.* **1985**, *107*, 3902-3909.
91. Voityuk, A. A.; Rösch, N., Am1/d Parameters for Molybdenum. *J. Phys. Chem. A* **2000**, *104*, 4089-4094.
92. Ignatov, S. K.; Razuvaev, A. G.; Kokorev, V. N.; Alexandrov, Y. A., Extension of the Pm3 Method on S,P,d Basis. Test Calculations on Organochromium Compounds. *J. Phys. Chem.* **1996**, *100*, 6354-6358.
93. Minenkov, Y.; Sharapa, D. I.; Cavallo, L., Application of Semiempirical Methods to Transition Metal Complexes: Fast Results but Hard-to-Predict Accuracy. *J. Chem. Theory Comput.* **2018**, *14*, 3428-3439.
94. Bosque, R.; Maseras, F., Performance of the Semiempirical Pm3 (Tm) Method in the Geometry Optimization of Transition Metal Complexes. *J. Comput. Chem.* **2000**, *21*, 562-571.
95. Buda, C.; Cundari, T. R., De Novo Prediction of Ground State Multiplicity and Structural Isomerism for Transition Metal Complexes. *J. Mol. Struct.: THEOCHEM* **2004**, *686*, 137-145.
96. Daniel, C. R. A.; Rodrigues, N. M.; da Costa, N. B.; Freire, R. O., Are Quantum Chemistry Semiempirical Methods Effective to Predict Solid State Structure and Adsorption in Metal Organic Frameworks? *J. Phys. Chem. C* **2015**, *119*, 23398-23406.
97. Sizova, O. V.; Baranovski, V. I., Indo Parameters for the Elements of the I and II Transition Rows. *J. Comput. Chem.* **1995**, *16*, 586-594.
98. Bredow, T.; Geudtner, G.; Jug, K., Msindo Parameterization for Third-Row Transition Metals. *J. Comput. Chem.* **2001**, *22*, 861-887.
99. Clark, T., Quo Vadis Semiempirical MO-Theory? *J. Mol. Struct.: THEOCHEM* **2000**, *530*, 1-10.
100. O'Brien, T. A.; Davidson, E. R., Semiempirical Local Spin: Theory and Implementation of the Zilsh Method for Predicting Heisenberg Exchange Constants of Polynuclear Transition Metal Complexes. *Int. J. Quantum Chem.* **2003**, *92*, 294-325.
101. Cundari, T. R.; Deng, J.; Fu, W., Pm3 (Tm) Parameterization Using Genetic Algorithms. *Int. J. Quantum Chem.* **2000**, *77*, 421-432.

102. Fredin, L. A.; Allison, T. C., Predicting Structures of Ru-Centered Dyes: A Computational Screening Tool. *J. Phys. Chem. A* **2016**, *120*, 2135-2143.
103. Dral, P. O.; Wu, X.; Spörkel, L.; Koslowski, A.; Thiel, W., Semiempirical Quantum-Chemical Orthogonalization-Corrected Methods: Benchmarks for Ground-State Properties. *J. Chem. Theory Comput.* **2016**, *12*, 1097-1120.
104. Bursch, M.; Neugebauer, H.; Grimme, S., Structure Optimisation of Large Transition-Metal Complexes with Extended Tight-Binding Methods. *Angew. Chem., Int. Ed.* **2019**, *58*, 11078-11087.
105. Sinha, V.; Laan, J. J.; Pidko, E. A., Accurate and Rapid Prediction of pKa of Transition Metal Complexes: Semiempirical Quantum Chemistry with a Data-Augmented Approach. *Phys. Chem. Chem. Phys.* **2021**, *23*, 2557-2567.
106. Neugebauer, H.; Bohle, F.; Bursch, M.; Hansen, A.; Grimme, S., Benchmark Study of Electrochemical Redox Potentials Calculated with Semiempirical and DFT Methods. *J. Phys. Chem. A* **2020**, *124*, 7166-7176.
107. Dohm, S.; Bursch, M.; Hansen, A.; Grimme, S., Semiautomated Transition State Localization for Organometallic Complexes with Semiempirical Quantum Chemical Methods. *J. Chem. Theory Comput.* **2020**, *16*, 2002-2012.
108. Elstner, M.; Porezag, D.; Jungnickel, G.; Elsner, J.; Haugk, M.; Frauenheim, T.; Suhai, S.; Seifert, G., Self-Consistent-Charge Density-Functional Tight-Binding Method for Simulations of Complex Materials Properties. *Phys. Rev. B* **1998**, *58*, 7260.
109. Gaus, M.; Cui, Q.; Elstner, M., Density Functional Tight Binding: Application to Organic and Biological Molecules. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2014**, *4*, 49-61.
110. Vujović, M.; Huynh, M.; Steiner, S.; Garcia-Fernandez, P.; Elstner, M.; Cui, Q.; Gruden, M., Exploring the Applicability of Density Functional Tight Binding to Transition Metal Ions. Parameterization for Nickel with the Spin-Polarized DFTB3 Model. *J. Comput. Chem.* **2019**, *40*, 400-413.
111. Gaus, M.; Jin, H.; Demapan, D.; Christensen, A. S.; Goyal, P.; Elstner, M.; Cui, Q., DFTB3 Parametrization for Copper: The Importance of Orbital Angular Momentum Dependence of Hubbard Parameters. *J. Chem. Theory Comput.* **2015**, *11*, 4205-4219.
112. Zheng, G.; Witek, H. A.; Bobadova-Parvanova, P.; Irle, S.; Musaev, D. G.; Prabhakar, R.; Morokuma, K.; Lundberg, M.; Elstner, M.; Köhler, C., et al., Parameter Calibration of Transition-Metal Elements for the Spin-Polarized Self-Consistent-Charge Density-Functional Tight-Binding (Dftb) Method: Sc, Ti, Fe, Co, and Ni. *J. Chem. Theory Comput.* **2007**, *3*, 1349-1367.
113. Mori-Sánchez, P.; Cohen, A. J.; Yang, W., Many-Electron Self-Interaction Error in Approximate Density Functionals. *J. Chem. Phys.* **2006**, *125*, 201102.
114. Ruzsinszky, A.; Perdew, J. P.; Csonka, G. I.; Vydrov, O. A.; Scuseria, G. E., Density Functionals That Are One- and Two- Are Not Always Many-Electron Self-Interaction-Free, as Shown for H2+, He2+, LiH+, and Ne2+. *J. Chem. Phys.* **2007**, *126*, 104102.
115. Haunschild, R.; Henderson, T. M.; Jiménez-Hoyos, C. A.; Scuseria, G. E., Many-Electron Self-Interaction and Spin Polarization Errors in Local Hybrid Density Functionals. *J. Chem. Phys.* **2010**, *133*, 134116.
116. Cohen, A. J.; Mori-Sánchez, P.; Yang, W., Insights into Current Limitations of Density Functional Theory. *Science* **2008**, *321*, 792-794.
117. Schmidt, T.; Kümmel, S., One- and Many-Electron Self-Interaction Error in Local and Global Hybrid Functionals. *Phys. Rev. B* **2016**, *93*, 165120.

118. Perdew, J. P.; Parr, R. G.; Levy, M.; Balduz, J. L., Density-Functional Theory for Fractional Particle Number: Derivative Discontinuities of the Energy. *Phys. Rev. Lett.* **1982**, *49*, 1691-1694.
119. Perdew, J. P.; Levy, M., Physical Content of the Exact Kohn-Sham Orbital Energies: Band Gaps and Derivative Discontinuities. *Phys. Rev. Lett.* **1983**, *51*, 1884-1887.
120. Sham, L. J.; Schlüter, M., Density-Functional Theory of the Energy Gap. *Phys. Rev. Lett.* **1983**, *51*, 1888-1891.
121. Becke, A. D., Perspective: Fifty Years of Density-Functional Theory in Chemical Physics. *J. Chem. Phys.* **2014**, *140*, 18A301.
122. Burke, K., Perspective on Density Functional Theory. *J. Chem. Phys.* **2012**, *136*, 150901.
123. Mori-Sánchez, P.; Cohen, A. J.; Yang, W., Localization and Delocalization Errors in Density Functional Theory and Implications for Band-Gap Prediction. *Phys. Rev. Lett.* **2008**, *100*, 146401.
124. Cohen, A.; Mori-Sanchez, P.; Yang, W., Fractional Spins and Static Correlation Error in Density Functional Theory. *J. Chem. Phys.* **2008**, *129*, 121104.
125. Duignan, T. J.; Autschbach, J., Impact of the Kohn–Sham Delocalization Error on the 4f Shell Localization and Population in Lanthanide Complexes. *J. Chem. Theory Comput.* **2016**, *12*, 3109-3121.
126. Bajaj, A.; Janet, J. P.; Kulik, H. J., Communication: Recovering the Flat-Plane Condition in Electronic Structure Theory at Semi-Local DFT Cost. *J. Chem. Phys.* **2017**, *147*, 191101.
127. Becke, A. D., Density Functionals for Static, Dynamical, and Strong Correlation. *J. Chem. Phys.* **2013**, *138*, 074109.
128. Bajaj, A.; Liu, F.; Kulik, H. J., Non-Empirical, Low-Cost Recovery of Exact Conditions with Model-Hamiltonian Inspired Expressions in jmDFT. *J. Chem. Phys.* **2019**, *150*, 154115.
129. Janesko, B. G.; Proynov, E.; Kong, J.; Scalmani, G.; Frisch, M. J., Practical Density Functionals Beyond the Overdelocalization–Underbinding Zero-Sum Game. *J. Phys. Chem. Lett.* **2017**, *8*, 4314-4318.
130. Su, N. Q.; Li, C.; Yang, W., Describing Strong Correlation with Fractional-Spin Correction in Density Functional Theory. *Proc. Natl. Acad. Sci.* **2018**, *115*, 9678-9683.
131. Bühl, M.; Reimann, C.; Pantazis, D. A.; Bredow, T.; Neese, F., Geometries of Third-Row Transition-Metal Complexes from Density-Functional Theory. *J. Chem. Theory Comput.* **2008**, *4*, 1449-1459.
132. Jensen, K. P.; Cirera, J., Accurate Computed Enthalpies of Spin Crossover in Iron and Cobalt Complexes. *J. Phys. Chem. A* **2009**, *113*, 10033-10039.
133. Swart, M.; Groenhof, A. R.; Ehlers, A. W.; Lammertsma, K., Validation of Exchange−Correlation Functionals for Spin States of Iron Complexes. *J. Phys. Chem. A* **2004**, *108*, 5479-5483.
134. Paulsen, H.; Duelund, L.; Winkler, H.; Toftlund, H.; Trautwein, A. X., Free Energy of Spin-Crossover Complexes Calculated with Density Functional Methods. *Inorg. Chem.* **2001**, *40*, 2201-2203.
135. Ganzenmüller, G.; Berkaïne, N.; Fouqueau, A.; Casida, M. E.; Reiher, M., Comparison of Density Functionals for Differences between the High- (T2g5) and Low- (A1g1) Spin States of Iron(II) Compounds. IV. Results for the Ferrous Complexes [Fe(L)('NHS4')]. *J. Chem. Phys.* **2005**, *122*, 234321.
136. Droghetti, A.; Alfè, D.; Sanvito, S., Assessment of Density Functional Theory for Iron (II) Molecules across the Spin-Crossover Transition. *J. Chem. Phys.* **2012**, *137*, 124303.

137. Mortensen, S. R.; Kepp, K. P., Spin Propensities of Octahedral Complexes from Density Functional Theory. *J. Phys. Chem. A* **2015**, *119*, 4041-4050.
138. Ioannidis, E. I.; Kulik, H. J., Towards Quantifying the Role of Exact Exchange in Predictions of Transition Metal Complex Properties. *J. Chem. Phys.* **2015**, *143*, 034104.
139. Ioannidis, E. I.; Kulik, H. J., Ligand-Field-Dependent Behavior of Meta-GGA Exchange in Transition-Metal Complex Spin-State Ordering. *J. Phys. Chem. A* **2017**, *121*, 874-884.
140. Zhao, Y.; Truhlar, D. G., Comparative Assessment of Density Functional Methods for 3 d Transition-Metal Chemistry. *J. Chem. Phys.* **2006**, *124*, 224105.
141. Iron, M. A.; Janes, T., Evaluating Transition Metal Barrier Heights with the Latest Density Functional Theory Exchange–Correlation Functionals: The Mobh35 Benchmark Database. *J. Phys. Chem. A* **2019**, *123*, 3761-3781.
142. Brumboiu, I. E.; Haldar, S.; Luder, J.; Eriksson, O.; Herper, H. C.; Brena, B.; Sanyal, B., Influence of Electron Correlation on the Electronic Structure and Magnetism of Transition-Metal Phthalocyanines. *J. Chem. Theory Comput.* **2016**, *12*, 1772-1785.
143. Herper, H. C.; Brena, B.; Puglia, C.; Bhandary, S.; Wende, H.; Eriksson, O.; Sanyal, B., Electronic Structure of Isolated Molecules. In *Molecular Nanomagnets*, Springer: 2020; pp 25-34.
144. Brumboiu, I. E.; Prokopiou, G.; Kronik, L.; Brena, B., Valence Electronic Structure of Cobalt Phthalocyanine from an Optimally Tuned Range-Separated Hybrid Functional. *J. Chem. Phys.* **2017**, *147*, 044301.
145. Huang, W.; Xing, D.-H.; Lu, J.-B.; Long, B.; Schwarz, W. E.; Li, J., How Much Can Density Functional Approximations (DFA) Fail? The Extreme Case of the FeO4 Species. *J. Chem. Theory Comput.* **2016**, *12*, 1525-1533.
146. Liu, F.; Kulik, H. J., Impact of Approximate DFT Density Delocalization Error on Potential Energy Surfaces in Transition Metal Chemistry. *J. Chem. Theory Comput.* **2020**, *16*, 264-277.
147. Weymuth, T.; Couzijn, E. P. A.; Chen, P.; Reiher, M., New Benchmark Set of Transition-Metal Coordination Reactions for the Assessment of Density Functionals. *J. Chem. Theory Comput.* **2014**, *10*, 3092-3103.
148. Gani, T. Z. H.; Kulik, H. J., Unifying Exchange Sensitivity in Transition Metal Spin-State Ordering and Catalysis through Bond Valence Metrics *J. Chem. Theory Comput.* **2017**, *13*, 5443-5457.
149. Jonas, V.; Thiel, W., Theoretical Study of the Vibrational Spectra of the Transition Metal Carbonyls M (CO) 6 [M= Cr, Mo, W], M (CO) 5 [M= Fe, Ru, Os], and M (CO) 4 [M= Ni, Pd, Pt]. *J. Chem. Phys.* **1995**, *102*, 8474-8484.
150. Waller, M. P.; Braun, H.; Hojdis, N.; Bühl, M., Geometries of Second-Row Transition-Metal Complexes from Density-Functional Theory. *J. Chem. Theory Comput.* **2007**, *3*, 2234-2242.
151. Laury, M. L.; Wilson, A. K., Performance of Density Functional Theory for Second Row (4 D) Transition Metal Thermochemistry. *J. Chem. Theory Comput.* **2013**, *9*, 3939-3946.
152. Nandy, A.; Chu, D. B. K.; Harper, D. R.; Duan, C.; Arunachalam, N.; Cytter, Y.; Kulik, H. J., Large-Scale Comparison of 3d and 4d Transition Metal Complexes Illuminates the Reduced Effect of Exchange on Second-Row Spin-State Energetics. *Phys. Chem. Chem. Phys.* **2020**, *22*, 19326-19341.
153. Husch, T.; Freitag, L.; Reiher, M., Calculation of Ligand Dissociation Energies in Large Transition-Metal Complexes. *J. Chem. Theory Comput.* **2018**, *14*, 2456-2468.

154. Riplinger, C.; Neese, F., An Efficient and near Linear Scaling Pair Natural Orbital Based Local Coupled Cluster Method. *J. Chem. Phys.* **2013**, *138*, 034106.
155. Chan, B.; Gill, P. M.; Kimura, M., Assessment of DFT Methods for Transition Metals with the Tmc151 Compilation of Data Sets and Comparison with Accuracies for Main-Group Chemistry. *J. Chem. Theory Comput.* **2019**, *15*, 3610-3622.
156. Dohm, S.; Hansen, A.; Steinmetz, M.; Grimme, S.; Checinski, M. P., Comprehensive Thermochemical Benchmark Set of Realistic Closed-Shell Metal Organic Reactions. *J. Chem. Theory Comput.* **2018**, *14*, 2596-2608.
157. Cheng, L.; Gauss, J. r.; Ruscic, B.; Armentrout, P. B.; Stanton, J. F., Bond Dissociation Energies for Diatomic Molecules Containing 3d Transition Metals: Benchmark Scalar-Relativistic Coupled-Cluster Calculations for 20 Molecules. *J. Chem. Theory Comput.* **2017**, *13*, 1044-1056.
158. Moltved, K. A.; Kepp, K. P., Chemical Bond Energies of 3d Transition Metals Studied by Density Functional Theory. *J. Chem. Theory Comput.* **2018**, *14*, 3479-3492.
159. Kulik, H. J.; Marzari, N., Transition-Metal Dioxides: A Case for the Intersite Term in Hubbard-Model Functionals. *J. Chem. Phys.* **2011**, *134*, 094103.
160. Groom, C. R.; Bruno, I. J.; Lightfoot, M. P.; Ward, S. C., The Cambridge Structural Database. *Acta Crystallogr., Sect. B: Struct. Sci., Cryst. Eng. Mater.* **2016**, *72*, 171-179.
161. Hocking, R. K.; Deeth, R. J.; Hambley, T. W., DFT Study of the Systematic Variations in Metal−Ligand Bond Lengths of Coordination Complexes: The Crucial Role of the Condensed Phase. *Inorg. Chem.* **2007**, *46*, 8238-8244.
162. Bühl, M.; Kabrede, H., Geometries of Transition-Metal Complexes from Density-Functional Theory. *J. Chem. Theory Comput.* **2006**, *2*, 1282-1290.
163. Anisimov, V. I.; Zaanen, J.; Andersen, O. K., Band Theory and Mott Insulators: Hubbard U Instead of Stoner I. *Phys. Rev. B* **1991**, *44*, 943.
164. Liechtenstein, A.; Anisimov, V.; Zaanen, J., Density-Functional Theory and Strong Interactions: Orbital Ordering in Mott-Hubbard Insulators. *Phys. Rev. B* **1995**, *52*, R5467.
165. Dudarev, S. L.; Botton, G. A.; Savrasov, S. Y.; Humphreys, C. J.; Sutton, A. P., Electron-Energy-Loss Spectra and the Structural Stability of Nickel Oxide: An LSDA+U Study. *Phys. Rev. B* **1998**, *57*, 1505-1509.
166. Himmetoglu, B.; Floris, A.; De Gironcoli, S.; Cococcioni, M., Hubbard-Corrected DFT Energy Functionals: The LDA+ U Description of Correlated Systems. *Int. J. Quantum Chem.* **2014**, *114*, 14-49.
167. Kulik, H. J.; Cococcioni, M.; Scherlis, D. A.; Marzari, N., Density Functional Theory in Transition-Metal Chemistry: A Self-Consistent Hubbard U Approach. *Phys. Rev. Lett.* **2006**, *97*, 103001.
168. Cococcioni, M.; De Gironcoli, S., Linear Response Approach to the Calculation of the Effective Interaction Parameters in the LDA+ U Method. *Phys. Rev. B* **2005**, *71*, 035105.
169. Zhao, Q.; Ioannidis, E. I.; Kulik, H. J., Global and Local Curvature in Density Functional Theory. *J. Chem. Phys.* **2016**, *145*, 054109.
170. Mosey, N.; Carter, E., Ab Initio Evaluation of Coulomb and Exchange Parameters for DFT+U Calculations. *Phys. Rev. B* **2007**, *76*, 155123.
171. Huang, M.; Fabris, S., CO Adsorption and Oxidation on Ceria Surfaces from DFT+U Calculations. *J. Phys. Chem. C* **2008**, *112*, 8643-8648.

172. Loschen, C.; Carrasco, J.; Neyman, K.; Illas, F., First-Principles LDA+U and GGA+U Study of Cerium Oxides: Dependence on the Effective U Parameter. *Phys. Rev. B* **2007**, *75*, 035115.
173. Wang, L.; Maxisch, T.; Ceder, G., Oxidation Energies of Transition Metal Oxides within the GGA+U Framework. *Phys. Rev. B* **2006**, *73*, 195107.
174. Friesner, R. A.; Knoll, E. H.; Cao, Y., A Localized Orbital Analysis of the Thermochemical Errors in Hybrid Density Functional Theory: Achieving Chemical Accuracy via a Simple Empirical Correction Scheme. *J. Chem. Phys.* **2006**, *125*, 124107.
175. Rinaldo, D.; Tian, L.; Harvey, J. N.; Friesner, R. A., Density Functional Localized Orbital Corrections for Transition Metals. *J. Chem. Phys.* **2008**, *129*, 164108.
176. Friesner, R. A.; Jerome, S. V., Localized Orbital Corrections for Density Functional Calculations on Transition Metal Containing Systems. *Coord. Chem. Rev.* **2017**, *344*, 205-213.
177. Hughes, T. F.; Friesner, R. A., Correcting Systematic Errors in DFT Spin-Splitting Energetics for Transition Metal Complexes. *J. Chem. Theory Comput.* **2011**, *7*, 19-32.
178. Hughes, T. F.; Friesner, R. A., Development of Accurate DFT Methods for Computing Redox Potentials of Transition Metal Complexes: Results for Model Complexes and Application to Cytochrome P450. *J. Chem. Theory Comput.* **2012**, *8*, 442-459.
179. Hughes, T. F.; Harvey, J. N.; Friesner, R. A., A B3LYP-DBLOC Empirical Correction Scheme for Ligand Removal Enthalpies of Transition Metal Complexes: Parameterization against Experimental and CCSD(T)-F12 Heats of Formation. *Phys. Chem. Chem. Phys.* **2012**, *14*, 7724-7738.
180. Coskun, D.; Jerome, S. V.; Friesner, R. A., Evaluation of the Performance of the B3LYP, PBE0, and M06 DFT Functionals, and DBLOC-Corrected Versions, in the Calculation of Redox Potentials and Spin Splittings for Transition Metal Containing Systems. *J. Chem. Theory Comput.* **2016**, *12*, 1121-1128.
181. Galstyan, A.; Knapp, E. W., Accurate Redox Potentials of Mononuclear Iron, Manganese, and Nickel Model Complexes. *J. Comput. Chem.* **2009**, *30*, 203-211.
182. Reiher, M.; Salomon, O.; Hess, B. A., Reparameterization of Hybrid Functionals Based on Energy Differences of States of Different Multiplicity. *Theor. Chem. Acc.* **2001**, *107*, 48-55.
183. Reiher, M., Theoretical Study of the Fe (Phen) 2 (NCS) 2 Spin-Crossover Complex with Reparametrized Density Functionals. *Inorg. Chem.* **2002**, *41*, 6928-6935.
184. Fouqueau, A.; Mer, S.; Casida, M. E.; Lawson Daku, L. M.; Hauser, A.; Mineva, T.; Neese, F., Comparison of Density Functionals for Energy and Structural Differences between the High- [5t2g:(T2g)4(Eg)2] and Low- [1a1g:(T2g)6(Eg)0] Spin States of the Hexaquoferrous Cation [Fe(H2O)6]2+. *J. Chem. Phys.* **2004**, *120*, 9473-9486.
185. Fouqueau, A.; Casida, M. E.; Lawson Daku, L. M.; Hauser, A.; Neese, F., Comparison of Density Functionals for Energy and Structural Differences between the High-[5t2g:(T2g) 4 (Eg) 2] and Low-[1a1g:(T2g) 6 (Eg) 0] Spin States of Iron (II) Coordination Compounds. II. More Functionals and the Hexaminoferrous Cation,[Fe (NH3) 6] 2+. *J. Chem. Phys.* **2005**, *122*, 044110.
186. Liu, F.; Yang, T.; Yang, J.; Xu, E.; Bajaj, A.; Kulik, H. J., Bridging the Homogeneous-Heterogeneous Divide: Modeling Spin and Reactivity in Single Atom Catalysis. *Front. Chem.* **2019**, *7*, 219.
187. Mahler, A.; Janesko, B. G.; Moncho, S.; Brothers, E. N., When Hartree-Fock Exchange Admixture Lowers DFT-Predicted Barrier Heights: Natural Bond Orbital Analyses and Implications for Catalysis. *J. Chem. Phys.* **2018**, *148*, 244106.

188. Kronik, L.; Stein, T.; Refaely-Abramson, S.; Baer, R., Excitation Gaps of Finite-Sized Systems from Optimally Tuned Range-Separated Hybrid Functionals. *J. Chem. Theory Comput.* **2012**, *8*, 1515-1531.
189. Stein, T.; Autschbach, J.; Govind, N.; Kronik, L.; Baer, R., Curvature and Frontier Orbital Energies in Density Functional Theory. *J. Phys. Chem. Lett.* **2012**, *3*, 3740-3744.
190. Autschbach, J.; Srebro, M., Delocalization Error and "Functional Tuning" in Kohn–Sham Calculations of Molecular Properties. *Acc. Chem. Res.* **2014**, *47*, 2592-2602.
191. Srebro, M.; Autschbach, J., Does a Molecule-Specific Density Functional Give an Accurate Electron Density? The Challenging Case of the Cucl Electric Field Gradient. *J. Phys. Chem. Lett.* **2012**, *3*, 576-581.
192. Duignan, T. J.; Autschbach, J.; Batista, E.; Yang, P., Assessment of Tuned Range Separated Exchange Functionals for Spectroscopies and Properties of Uranium Complexes. *J. Chem. Theory Comput.* **2017**, *13*, 3614-3625.
193. Gani, T. Z. H.; Kulik, H. J., Where Does the Density Localize? Convergent Behavior for Global Hybrids, Range Separation, and DFT+U *J. Chem. Theory Comput.* **2016**, *12*, 5931-5945.
194. Konezny, S. J.; Doherty, M. D.; Luca, O. R.; Crabtree, R. H.; Soloveichik, G. L.; Batista, V. S., Reduction of Systematic Uncertainty in DFT Redox Potentials of Transition-Metal Complexes. *J. Phys. Chem. C* **2012**, *116*, 6349-6356.
195. Roy, L. E.; Jakubikova, E.; Guthrie, M. G.; Batista, E. R., Calculation of One-Electron Redox Potentials Revisited. Is It Possible to Calculate Accurate Potentials with Density Functional Methods? *J. Phys. Chem. A* **2009**, *113*, 6745-6750.
196. Casasnovas, R.; Ortega-Castro, J.; Frau, J.; Donoso, J.; Munoz, F., Theoretical pKa Calculations with Continuum Model Solvents, Alternative Protocols to Thermodynamic Cycles. *Int. J. Quantum Chem.* **2014**, *114*, 1350-1363.
197. Man, I. C.; Su, H. Y.; Calle-Vallejo, F.; Hansen, H. A.; Martinez, J. I.; Inoglu, N. G.; Kitchin, J.; Jaramillo, T. F.; Nørskov, J. K.; Rossmeisl, J., Universality in Oxygen Evolution Electrocatalysis on Oxide Surfaces. *ChemCatChem* **2011**, *3*, 1159-1165.
198. Eichkorn, K.; Treutler, O.; Öhm, H.; Häser, M.; Ahlrichs, R., Auxiliary Basis Sets to Approximate Coulomb Potentials. *Chem. Phys. Lett.* **1995**, *240*, 283-290.
199. Eichkorn, K.; Weigend, F.; Treutler, O.; Ahlrichs, R., Auxiliary Basis Sets for Main Row Atoms and Transition Metals and Their Use to Approximate Coulomb Potentials. *Theor. Chem. Acc.* **1997**, *97*, 119-124.
200. Ochsenfeld, C.; Kussmann, J.; Lambrecht, D. S., Linear-Scaling Methods in Quantum Chemistry. *Rev. Comput. Chem.* **2007**, *23*, 1.
201. Ufimtsev, I. S.; Martínez, T. J., Quantum Chemistry on Graphical Processing Units. 3. Analytical Energy Gradients, Geometry Optimization, and First Principles Molecular Dynamics. *J. Chem. Theory Comput.* **2009**, *5*, 2619-2628.
202. Riplinger, C.; Sandhoefer, B.; Hansen, A.; Neese, F., Natural Triple Excitations in Local Coupled Cluster Calculations with Pair Natural Orbitals. *J. Chem. Phys.* **2013**, *139*, 134101.
203. Xu, X.; Zhang, W.; Tang, M.; Truhlar, D. G., Do Practical Standard Coupled Cluster Calculations Agree Better Than Kohn–Sham Calculations with Currently Available Functionals When Compared to the Best Available Experimental Data for Dissociation Energies of Bonds to 3 d Transition Metals? *J. Chem. Theory Comput.* **2015**, *11*, 2036-2052.
204. Stein, C. J.; von Burg, V.; Reiher, M., The Delicate Balance of Static and Dynamic Electron Correlation. *J. Chem. Theory Comput.* **2016**, *12*, 3764-3773.

205. Flöser, B. M.; Guo, Y.; Riplinger, C.; Tuczek, F.; Neese, F., Detailed Pair Natural Orbital-Based Coupled Cluster Studies of Spin Crossover Energetics. *J. Chem. Theory Comput.* **2020**, *16*, 2224-2235.
206. Feldt, M.; Phung, Q. M.; Pierloot, K.; Mata, R. A.; Harvey, J. N., Limits of Coupled-Cluster Calculations for Non-Heme Iron Complexes. *J. Chem. Theory Comput.* **2019**, *15*, 922-937.
207. Fang, Z.; Vasiliu, M.; Peterson, K. A.; Dixon, D. A., Prediction of Bond Dissociation Energies/Heats of Formation for Diatomic Transition Metal Compounds: CCSD (T) Works. *J. Chem. Theory Comput.* **2017**, *13*, 1057-1066.
208. Veryazov, V.; Malmqvist, P. Å.; Roos, B. O., How to Select Active Space for Multiconfigurational Quantum Chemistry? *Int. J. Quantum Chem.* **2011**, *111*, 3329-3338.
209. Stein, C. J.; Reiher, M., Automated Selection of Active Orbital Spaces. *J. Chem. Theory Comput.* **2016**, *12*, 1760-1771.
210. Sayfutyarova, E. R.; Sun, Q.; Chan, G. K.-L.; Knizia, G., Automated Construction of Molecular Active Spaces from Atomic Valence Orbitals. *J. Chem. Theory Comput.* **2017**, *13*, 4063-4078.
211. Jeong, W.; Stoneburner, S. J.; King, D.; Li, R.; Walker, A.; Lindh, R.; Gagliardi, L., Automation of Active Space Selection for Multireference Methods via Machine Learning on Chemical Bond Dissociation. *J. Chem. Theory Comput.* **2020**, *16*, 2389-2399.
212. Rissler, J.; Noack, R. M.; White, S. R., Measuring Orbital Interaction Using Quantum Information Theory. *Chem. Phys.* **2006**, *323*, 519-531.
213. Legeza, Ö.; Sólyom, J., Optimizing the Density-Matrix Renormalization Group Method Using Quantum Information Entropy. *Phys. Rev. B* **2003**, *68*, 195116.
214. Boguslawski, K.; Tecmer, P.; Legeza, O.; Reiher, M., Entanglement Measures for Single- and Multireference Correlation Effects. *J. Phys. Chem. Lett.* **2012**, *3*, 3129-3135.
215. Ghigo, G.; Roos, B. O.; Malmqvist, P.-Å., A Modified Definition of the Zeroth-Order Hamiltonian in Multiconfigurational Perturbation Theory (CASPT2). *Chem. Phys. Lett.* **2004**, *396*, 142-149.
216. Pierloot, K.; Phung, Q. M.; Domingo, A., Spin State Energetics in First-Row Transition Metal Complexes: Contribution of (3s3p) Correlation and Its Description by Second-Order Perturbation Theory. *J. Chem. Theory Comput.* **2017**, *13*, 537-553.
217. Phung, Q. M.; Feldt, M.; Harvey, J. N.; Pierloot, K., Toward Highly Accurate Spin State Energetics in First-Row Transition Metal Complexes: A Combined CASPT2/Cc Approach. *J. Chem. Theory Comput.* **2018**, *14*, 2446-2455.
218. Lee, T. J.; Taylor, P. R., A Diagnostic for Determining the Quality of Single-Reference Electron Correlation Methods. *Int. J. Quantum Chem.* **1989**, *36*, 199-207.
219. Sears, J. S.; Sherrill, C. D., Assessing the Performance of Density Functional Theory for the Electronic Structure of Metal-Salens: The d(2)-Metals. *J. Phys. Chem. A* **2008**, *112*, 6741-6752.
220. Sears, J. S.; Sherrill, C. D., Assessing the Performance of Density Functional Theory for the Electronic Structure of Metal-Salens: The 3d(0)-Metals. *J. Phys. Chem. A* **2008**, *112*, 3466-3477.
221. Langhoff, S. R.; Davidson, E. R., Configuration Interaction Calculations on the Nitrogen Molecule. *Int. J. Quantum Chem.* **1974**, *8*, 61-72.
222. Janssen, C. L.; Nielsen, I. M. B., New Diagnostics for Coupled-Cluster and Moller-Plesset Perturbation Theory. *Chem. Phys. Lett.* **1998**, *290*, 423-430.

223. Nielsen, I. M. B.; Janssen, C. L., Double-Substitution-Based Diagnostics for Coupled-Cluster and Moller-Plesset Perturbation Theory. *Chem. Phys. Lett.* **1999**, *310*, 568-576.
224. Fogueri, U. R.; Kozuch, S.; Karton, A.; Martin, J. M. L., A Simple DFT-Based Diagnostic for Nondynamical Correlation. *Theor. Chem. Acc.* **2013**, *132*, 1291.
225. Tishchenko, O.; Zheng, J. J.; Truhlar, D. G., Multireference Model Chemistries for Thermochemical Kinetics. *J. Chem. Theory Comput.* **2008**, *4*, 1208-1219.
226. Schultz, N. E.; Zhao, Y.; Truhlar, D. G., Density Functionals for Inorganometallic and Organometallic Chemistry. *J. Phys. Chem. A* **2005**, *109*, 11127-11143.
227. Ramos-Cordoba, E.; Salvador, P.; Matito, E., Separation of Dynamic and Nondynamic Correlation. *Phys. Chem. Chem. Phys.* **2016**, *18*, 24015-24023.
228. Ramos-Cordoba, E.; Matito, E., Local Descriptors of Dynamic and Nondynamic Correlation. *J. Chem. Theory Comput.* **2017**, *13*, 2705-2711.
229. Grimme, S.; Hansen, A., A Practicable Real-Space Measure and Visualization of Static Electron-Correlation Effects. *Angew. Chem., Int. Ed.* **2015**, *54*, 12308-12313.
230. Kesharwani, M. K.; Sylvetsky, N.; Kohn, A.; Tew, D. P.; Martin, J. M. L., Do CCSD and Approximate CCSD-F12 Variants Converge to the Same Basis Set Limits? The Case of Atomization Energies. *J. Chem. Phys.* **2018**, *149*, 154109.
231. Jensen, H. J. A.; Jorgensen, P.; Ågren, H.; Olsen, J., Second-Order Moller–Plesset Perturbation Theory as a Configuration and Orbital Generator in Multiconfiguration Self-Consistent Field Calculations. *J. Chem. Phys.* **1988**, *88*, 3834-3839.
232. Jiang, W.; DeYonker, N. J.; Wilson, A. K., Multireference Character for 3d Transition-Metal-Containing Molecules. *J. Chem. Theory Comput.* **2012**, *8*, 460-468.
233. Wang, J.; Manivasagam, S.; Wilson, A. K., Multireference Character for 4d Transition Metal-Containing Molecules. *J. Chem. Theory Comput.* **2015**, *11*, 5865-5872.
234. Marques, H. M.; Cukrowski, I., Molecular Mechanics Modelling of Porphyrins. Using Artificial Neural Networks to Develop Metal Parameters for Four-Coordinate Metalloporphyrins. *Phys. Chem. Chem. Phys.* **2002**, *4*, 5878-5887.
235. Li, H.; Collins, C.; Tanha, M.; Gordon, G. J.; Yaron, D. J., A Density Functional Tight Binding Layer for Deep Learning of Chemical Hamiltonians. *J. Chem. Theory Comput.* **2018**, *14*, 5764-5776.
236. Ramakrishnan, R.; Dral, P. O.; Rupp, M.; von Lilienfeld, O. A., Big Data Meets Quantum Chemistry Approximations: The Δ-Machine Learning Approach. *J. Chem. Theory Comput.* **2015**, *11*, 2087-2096.
237. Grambow, C. A.; Li, Y.-P.; Green, W. H., Accurate Thermochemistry with Small Data Sets: A Bond Additivity Correction and Transfer Learning Approach. *J. Phys. Chem. A* **2019**, *123*, 5826-5835.
238. Zhao, Y.; Truhlar, D. G., A New Local Density Functional for Main-Group Thermochemistry, Transition Metal Bonding, Thermochemical Kinetics, and Noncovalent Interactions. *J. Chem. Phys.* **2006**, *125*, 194101.
239. Zhao, Y.; Truhlar, D. G., The M06 Suite of Density Functionals for Main Group Thermochemistry, Thermochemical Kinetics, Noncovalent Interactions, Excited States, and Transition Elements: Two New Functionals and Systematic Testing of Four M06-Class Functionals and 12 Other Functionals. *Theor. Chem. Acc.* **2008**, *120*, 215-241.
240. Yu, H. S.; Li, S. L.; Truhlar, D. G., Perspective: Kohn-Sham Density Functional Theory Descending a Staircase. *J. Chem. Phys.* **2016**, *145*, 130901.

241. Jensen, F., Method Calibration or Data Fitting? *J. Chem. Theory Comput.* **2018**, *14*, 4651-4661.
242. Mardirossian, N.; Head-Gordon, M., ωB97X-V: A 10-Parameter, Range-Separated Hybrid, Generalized Gradient Approximation Density Functional with Nonlocal Correlation, Designed by a Survival-of-the-Fittest Strategy. *Phys. Chem. Chem. Phys.* **2014**, *16*, 9904-9924.
243. Walker, E.; Ammal, S. C.; Terejanu, G. A.; Heyden, A., Uncertainty Quantification Framework Applied to the Water–Gas Shift Reaction over Pt-Based Catalysts. *J. Phys. Chem. C* **2016**, *120*, 10328-10339.
244. Wellendorff, J.; Lundgaard, K. T.; Møgelhøj, A.; Petzold, V.; Landis, D. D.; Nørskov, J. K.; Bligaard, T.; Jacobsen, K. W., Density Functionals for Surface Science: Exchange-Correlation Model Development with Bayesian Error Estimation. *Phys. Rev. B* **2012**, *85*, 235149.
245. Mortensen, J. J.; Kaasbjerg, K.; Frederiksen, S. L.; Norskov, J. K.; Sethna, J. P.; Jacobsen, K. W., Bayesian Error Estimation in Density-Functional Theory. *Phys. Rev. Lett.* **2005**, *95*, 216401.
246. Sutton, J. E.; Guo, W.; Katsoulakis, M. A.; Vlachos, D. G., Effects of Correlated Parameters and Uncertainty in Electronic-Structure-Based Chemical Kinetic Modelling. *Nat. Chem.* **2016**, *8*, 331-337.
247. Simm, G. N.; Reiher, M., Systematic Error Estimation for Chemical Reaction Energies. *J. Chem. Theory Comput.* **2016**, *12*, 2762-2773.
248. Evans, D.; Osborn, J.; Wilkinson, G., Hydroformylation of Alkenes by Use of Rhodium Complex Catalysts. *J. Chem. Soc. A* **1968**, 3133-3142.
249. Collman, J. P., Patterns of Organometallic Reactions Related to Homogeneous Catalysis. *Acc. Chem. Res.* **1968**, *1*, 136-143.
250. Beer, P. D.; Hayes, E. J., Transition Metal and Organometallic Anion Complexation Agents. *Coord. Chem. Rev.* **2003**, *240*, 167-189.
251. Bousseksou, A.; Molnár, G.; Salmon, L.; Nicolazzi, W., Molecular Spin Crossover Phenomenon: Recent Achievements and Prospects. *Chem. Soc. Rev.* **2011**, *40*, 3313-3335.
252. Champin, B.; Mobian, P.; Sauvage, J.-P., Transition Metal Complexes as Molecular Machine Prototypes. *Chem. Soc. Rev.* **2007**, *36*, 358-366.
253. Gomez-Coca, S.; Cremades, E.; Aliaga-Alcalde, N.; Ruiz, E., Mononuclear Single-Molecule Magnets: Tailoring the Magnetic Anisotropy of First-Row Transition-Metal Complexes. *J. Am. Chem. Soc.* **2013**, *135*, 7010-7018.
254. Gatteschi, D.; Sessoli, R., Quantum Tunneling of Magnetization and Related Phenomena in Molecular Materials. *Angew. Chem., Int. Ed.* **2003**, *42*, 268-297.
255. Noack, J.; Roznyatovskaya, N.; Herr, T.; Fischer, P., The Chemistry of Redox-Flow Batteries. *Angew. Chem., Int. Ed.* **2015**, *54*, 9776-9809.
256. Yeoh, J. S.; Armer, C. F.; Lowe, A., Transition Metal Oxalates as Energy Storage Materials. A Review. *Mater. Today Energy* **2018**, *9*, 198-222.
257. Kalyanasundaram, K.; Grätzel, M., Applications of Functionalized Transition Metal Complexes in Photonic and Optoelectronic Devices. *Coord. Chem. Rev.* **1998**, *177*, 347-414.
258. Hagfeldt, A.; Grätzel, M., Molecular Photovoltaics. *Acc. Chem. Res.* **2000**, *33*, 269-277.
259. de Vries, J. G.; Jackson, S. D., Homogeneous and Heterogeneous Catalysis in Industry. *Catal. Sci. Technol.* **2012**, *2*, 2009-2009.
260. Slaugh, L. H.; Mullineaux, R. D., Novel Hydroformylation Catalysts. *J. Organomet. Chem.* **1968**, *13*, 469-477.

261. Trnka, T. M.; Grubbs, R. H., The Development of L2x2ru Chr Olefin Metathesis Catalysts: An Organometallic Success Story. *Acc. Chem. Res.* **2001**, *34*, 18-29.
262. Topczewski, J. J.; Sanford, M. S., Carbon–Hydrogen (C–H) Bond Activation at Pd IV: A Frontier in C–H Functionalization Catalysis. *Chem. Sci.* **2015**, *6*, 70-76.
263. Twilton, J.; Zhang, P.; Shaw, M. H.; Evans, R. W.; MacMillan, D. W., The Merger of Transition Metal and Photocatalysis. *Nat. Rev. Chem.* **2017**, *1*, 1-19.
264. Cheung, K.-C.; Wong, W.-L.; Ma, D.-L.; Lai, T.-S.; Wong, K.-Y., Transition Metal Complexes as Electrocatalysts—Development and Applications in Electro-Oxidation Reactions. *Coord. Chem. Rev.* **2007**, *251*, 2367-2385.
265. Anantharaj, S.; Aravindan, V., Developments and Perspectives in 3d Transition-Metal-Based Electrocatalysts for Neutral and near-Neutral Water Electrolysis. *Adv. Energy Mater.* **2020**, *10*, 1902666.
266. Crabtree, R. H., Deactivation in Homogeneous Transition Metal Catalysis: Causes, Avoidance, and Cure. *Chem. Rev.* **2015**, *115*, 127-150.
267. Jacobsen, E. N.; Zhang, W.; Guler, M. L., Electronic Tuning of Asymmetric Catalysts. *J. Am. Chem. Soc.* **1991**, *113*, 6703-6704.
268. Johnson, C., Linear Free Energy Relations and the Reactivity-Selectivity Principle. *Chem. Rev.* **1975**, *75*, 755-765.
269. Jencks, W. P., A Primer for the Bema Hapothle. An Empirical Approach to the Characterization of Changing Transition-State Structures. *Chem. Rev.* **1985**, *85*, 511-527.
270. Gassman, P. G.; Macomber, D. W.; Willging, S. M., Isolation and Characterization of Reactive Intermediates and Active Catalysts in Homogeneous Catalysis. *J. Am. Chem. Soc.* **1985**, *107*, 2380-2388.
271. Widegren, J. A.; Finke, R. G., A Review of the Problem of Distinguishing True Homogeneous Catalysis from Soluble or Other Metal-Particle Heterogeneous Catalysis under Reducing Conditions. *J. Mol. Catal. A: Chem.* **2003**, *198*, 317-341.
272. Raynal, M.; Ballester, P.; Vidal-Ferran, A.; van Leeuwen, P. W., Supramolecular Catalysis. Part 1: Non-Covalent Interactions as a Tool for Building and Modifying Homogeneous Catalysts. *Chem. Soc. Rev.* **2014**, *43*, 1660-1733.
273. Bartlett, S. A.; Wells, P. P.; Nachtegaal, M.; Dent, A. J.; Cibin, G.; Reid, G.; Evans, J.; Tromp, M., Insights in the Mechanism of Selective Olefin Oligomerisation Catalysis Using Stopped-Flow Freeze-Quench Techniques: A Mo K-Edge Qexafs Study. *J. Catal.* **2011**, *284*, 247-258.
274. Kleifeld, O.; Frenkel, A.; Martin, J. M.; Sagi, I., Active Site Electronic Structure and Dynamics During Metalloenzyme Catalysis. *Nat. Struct. Biol.* **2003**, *10*, 98-103.
275. Cherepanov, A. V.; De Vries, S., Microsecond Freeze-Hyperquenching: Development of a New Ultrafast Micro-Mixing and Sampling Technology and Application to Enzyme Catalysis. *Biochim. Biophys. Acta, Bioenerg.* **2004**, *1656*, 1-31.
276. Kiefer, L. M.; King, J. T.; Kubarych, K. J., Dynamics of Rhenium Photocatalysts Revealed through Ultrafast Multidimensional Spectroscopy. *Acc. Chem. Res.* **2015**, *48*, 1123-1130.
277. Rountree, E. S.; McCarthy, B. D.; Eisenhart, T. T.; Dempsey, J. L., Evaluation of Homogeneous Electrocatalysts by Cyclic Voltammetry. *Inorg. Chem.* **2014**, *53*, 9983-10002.
278. Simmons, E. M.; Hartwig, J. F., On the Interpretation of Deuterium Kinetic Isotope Effects in C–H Bond Functionalizations by Transition-Metal Complexes. *Angew. Chem., Int. Ed.* **2012**, *51*, 3066-3072.

279. Chen, P., Electrospray Ionization Tandem Mass Spectrometry in High-Throughput Screening of Homogeneous Catalysts. *Angew. Chem., Int. Ed.* **2003**, *42*, 2832-2847.
280. Vikse, K. L.; Ahmadi, Z.; McIndoe, J. S., The Application of Electrospray Ionization Mass Spectrometry to Homogeneous Catalysis. *Coord. Chem. Rev.* **2014**, *279*, 96-114.
281. Mao, Z.; Campbell, C. T., Kinetic Isotope Effects: Interpretation and Prediction Using Degrees of Rate Control. *ACS Catal.* **2020**, *10*, 4181-4192.
282. Besora, M.; Maseras, F., Microkinetic Modeling in Homogeneous Catalysis. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2018**, *8*, e1372.
283. Stegelmann, C.; Andreasen, A.; Campbell, C. T., Degree of Rate Control: How Much the Energies of Intermediates and Transition States Control Rates. *J. Am. Chem. Soc.* **2009**, *131*, 8077-8082.
284. Kozuch, S.; Shaik, S., How to Conceptualize Catalytic Cycles? The Energetic Span Model. *Acc. Chem. Res.* **2011**, *44*, 101-110.
285. Kozuch, S.; Martin, J. M. L., What Makes for a Bad Catalytic Cycle? A Theoretical Study on the Suzuki−Miyaura Reaction within the Energetic Span Model. *ACS Catal.* **2011**, *1*, 246-253.
286. Mandal, D.; Mallick, D.; Shaik, S., Kinetic Isotope Effect Determination Probes the Spin of the Transition State, Its Stereochemistry, and Its Ligand Sphere in Hydrogen Abstraction Reactions of Oxoiron (IV) Complexes. *Acc. Chem. Res.* **2018**, *51*, 107-117.
287. Kupper, C.; Mondal, B.; Serrano-Plana, J.; Klawitter, I.; Neese, F.; Costas, M.; Ye, S.; Meyer, F., Nonclassical Single-State Reactivity of an Oxo-Iron (IV) Complex Confined to Triplet Pathways. *J. Am. Chem. Soc.* **2017**, *139*, 8939-8949.
288. Poli, R.; Harvey, J. N., Spin Forbidden Chemical Reactions of Transition Metal Compounds. New Ideas and New Computational Challenges. *Chem. Soc. Rev.* **2003**, *32*, 1-8.
289. Kwon, Y. H.; Mai, B. K.; Lee, Y.-M.; Dhuri, S. N.; Mandal, D.; Cho, K.-B.; Kim, Y.; Shaik, S.; Nam, W., Determination of Spin Inversion Probability, H-Tunneling Correction, and Regioselectivity in the Two-State Reactivity of Nonheme Iron(IV)-Oxo Complexes. *J. Phys. Chem. Lett.* **2015**, *6*, 1472-1476.
290. Dhuri, S. N.; Seo, M. S.; Lee, Y. M.; Hirao, H.; Wang, Y.; Nam, W.; Shaik, S., Experiment and Theory Reveal the Fundamental Difference between Two-State and Single-State Reactivity Patterns in Nonheme Feiv O Versus Ruiv O Oxidants. *Angew. Chem., Int. Ed.* **2008**, *47*, 3356-3359.
291. Nørskov, J. K.; Bligaard, T.; Logadottir, A.; Bahn, S.; Hansen, L. B.; Bollinger, M.; Bengaard, H.; Hammer, B.; Sljivancanin, Z.; Mavrikakis, M., Universality in Heterogeneous Catalysis. *J. Catal.* **2002**, *209*, 275-278.
292. Warren, J. J.; Tronic, T. A.; Mayer, J. M., Thermochemistry of Proton-Coupled Electron Transfer Reagents and Its Implications. *Chem. Rev.* **2010**, *110*, 6961-7001.
293. Tsay, C.; Livesay, B. N.; Ruelas, S.; Yang, J. Y., Solvation Effects on Transition Metal Hydricity. *J. Am. Chem. Soc.* **2015**, *137*, 14114-14121.
294. Ceballos, B. M.; Yang, J. Y., Directing the Reactivity of Metal Hydrides for Selective CO2 Reduction. *Proc. Natl. Acad. Sci.* **2018**, *115*, 12686-12691.
295. Dyson, P. J.; Jessop, P. G., Solvent Effects in Catalysis: Rational Improvements of Catalysts via Manipulation of Solvent Interactions. *Catal. Sci. Technol.* **2016**, *6*, 3302-3316.
296. Rappé, A. K.; Skiff, W. M.; Casewit, C. J., Modeling Metal-Catalyzed Olefin Polymerization. *Chem. Rev.* **2000**, *100*, 1435-1456.

297. Ho, J.; Ertem, M. Z., Calculating Free Energy Changes in Continuum Solvation Models. *J. Phys. Chem. B* **2016**, *120*, 1319-1329.
298. Kuliaev, P. O.; Pidko, E. A., Operando Modeling of Multicomponent Reactive Solutions in Homogeneous Catalysis: From Non-Standard Free Energies to Reaction Network Control. *ChemCatChem* **2020**, *12*, 795.
299. Marenich, A. V.; Ho, J.; Coote, M. L.; Cramer, C. J.; Truhlar, D. G., Computational Electrochemistry: Prediction of Liquid-Phase Reduction Potentials. *Phys. Chem. Chem. Phys.* **2014**, *16*, 15068-15106.
300. Mavros, M. G.; Tsuchimochi, T.; Kowalczyk, T.; McIsaac, A.; Wang, L.-P.; Voorhis, T. V., What Can Density Functional Theory Tell Us About Artificial Catalytic Water Splitting? *Inorg. Chem.* **2014**, *53*, 6386-6397.
301. Gagne, R. R.; Koval, C. A.; Lisensky, G. C., Ferrocene as an Internal Standard for Electrochemical Measurements. *Inorg. Chem.* **1980**, *19*, 2854-2855.
302. Pavlishchuk, V. V.; Addison, A. W., Conversion Constants for Redox Potentials Measured Versus Different Reference Electrodes in Acetonitrile Solutions at 25 C. *Inorg. Chim. Acta* **2000**, *298*, 97-102.
303. Roth, H.; Romero, N.; Nicewicz, D., Experimental and Calculated Electrochemical Potentials of Common Organic Molecules for Applications to Single-Electron Redox Chemistry. *Synlett* **2016**, *27*, 714-723.
304. Bottomley, L.; Kadish, K., Counterion and Solvent Effects on the Electrode Reactions of Iron Porphyrins. *Inorg. Chem.* **1981**, *20*, 1348-1357.
305. Eisenhart, R. J.; Clouston, L. J.; Lu, C. C., Configuring Bonds between First-Row Transition Metals. *Acc. Chem. Res.* **2015**, *48*, 2885-2894.
306. Patra, S.; Sarkar, B.; Mobin, S. M.; Kaim, W.; Lahiri, G. K., Separating Innocence and Non-Innocence of Ligands and Metals in Complexes [(L)Ru(Acac)2]N (N = −1, 0, +1; L = O-Iminoquinone or O-Iminothioquinone). *Inorg. Chem.* **2003**, *42*, 6469-6473.
307. Chirik, P. J., Preface: Forum on Redox-Active Ligands. *Inorg. Chem.* **2011**, *50*, 9737-9740.
308. Jørgensen, C. K., Differences between the Four Halide Ligands, and Discussion Remarks on Trigonal-Bipyramidal Complexes, on Oxidation States, and on Diagonal Elements of One-Electron Energy. *Coord. Chem. Rev.* **1966**, *1*, 164-178.
309. Pierpont, C. G., Studies on Charge Distribution and Valence Tautomerism in Transition Metal Complexes of Catecholate and Semiquinonate Ligands. *Coord. Chem. Rev.* **2001**, *216*, 99-125.
310. Matsui, T.; Kitagawa, Y.; Shigeta, Y.; Okumura, M., A Density Functional Theory Based Protocol to Compute the Redox Potential of Transition Metal Complex with the Correction of Pseudo-Counterion: General Theory and Applications. *J. Chem. Theory Comput.* **2013**, *9*, 2974-2980.
311. Bím, D.; Rulíšek, L. r.; Srnec, M., Accurate Prediction of One-Electron Reduction Potentials in Aqueous Solution by Variable-Temperature H-Atom Addition/Abstraction Methodology. *J. Phys. Chem. Lett.* **2016**, *7*, 7-13.
312. Li, J.; Fisher, C. L.; Chen, J. L.; Bashford, D.; Noodleman, L., Calculation of Redox Potentials and P K a Values of Hydrated Transition Metal Cations by a Combined Density Functional and Continuum Dielectric Theory. *Inorg. Chem.* **1996**, *35*, 4694-4702.

313. Zeng, X.; Hu, H.; Hu, X.; Cohen, A. J.; Yang, W., Ab Initio Quantum Mechanical/Molecular Mechanical Simulation of Electron Transfer Process: Fractional Electron Approach. *J. Chem. Phys.* **2008**, *128*, 124510.
314. Wang, L.-P.; Van Voorhis, T., A Polarizable QM/MM Explicit Solvent Model for Computational Electrochemistry in Water. *J. Chem. Theory Comput.* **2012**, *8*, 610-617.
315. Orgel, L., Spectra of Transition-Metal Complexes. *J. Chem. Phys.* **1955**, *23*, 1004-1014.
316. Crosby, G. A., Spectroscopic Investigations of Excited States of Transition-Metal Complexes. *Acc. Chem. Res.* **1975**, *8*, 231-238.
317. Fabian, J.; Nakazumi, H.; Matsuoka, M., Near-Infrared Absorbing Dyes. *Chem. Rev.* **1992**, *92*, 1197-1226.
318. Tao, Y.; Yuan, K.; Chen, T.; Xu, P.; Li, H.; Chen, R.; Zheng, C.; Zhang, L.; Huang, W., Thermally Activated Delayed Fluorescence Materials Towards the Breakthrough of Organoelectronics. *Adv. Mater.* **2014**, *26*, 7931-7958.
319. Tsuchida, R., Absorption Spectra of Co-Ordination Compounds. I. *Bull. Chem. Soc. Jpn.* **1938**, *13*, 388-400.
320. Warzecha, E.; Berto, T. C.; Wilkinson, C. C.; Berry, J. F., Rhodium Rainbow: A Colorful Laboratory Experiment Highlighting Ligand Field Effects of Dirhodium Tetraacetate. *J. Chem. Educ.* **2019**, *96*, 571-576.
321. Zhao, J.; Ji, S.; Wu, W.; Wu, W.; Guo, H.; Sun, J.; Sun, H.; Liu, Y.; Li, Q.; Huang, L., Transition Metal Complexes with Strong Absorption of Visible Light and Long-Lived Triplet Excited States: From Molecular Design to Applications. *RSC Adv.* **2012**, *2*, 1712-1728.
322. Evans, D. R.; Reed, C. A., Reversal of H2O and Oh-Ligand Field Strength on the Magnetochemical Series Relative to the Spectrochemical Series. Novel 1-Equiv Water Chemistry of Iron (III) Tetraphenylporphyrin Complexes. *J. Am. Chem. Soc.* **2000**, *122*, 4660-4667.
323. Ward, M. D., Near-Infrared Electrochromic Materials for Optical Attenuation Based on Transition-Metal Coordination Complexes. *J. Solid State Electrochem.* **2005**, *9*, 778-787.
324. Chia, Y.; Tay, M., An Insight into Fluorescent Transition Metal Complexes. *Dalton Trans.* **2014**, *43*, 13159-13168.
325. Schinabeck, A.; Chen, J.; Kang, L.; Teng, T.; Homeier, H. H.; Suleymanova, A. F.; Shafikov, M. Z.; Yu, R.; Lu, C.-Z.; Yersin, H., Symmetry-Based Design Strategy for Unprecedentedly Fast Decaying Thermally Activated Delayed Fluorescence (TADF). Application to Dinuclear Cu (I) Compounds. *Chem. Mater.* **2019**, *31*, 4392-4404.
326. Linfoot, C. L.; Leitl, M. J.; Richardson, P.; Rausch, A. F.; Chepelin, O.; White, F. J.; Yersin, H.; Robertson, N., Thermally Activated Delayed Fluorescence (TADF) and Enhancing Photoluminescence Quantum Yields of [Cui (Diimine)(Diphosphine)]+ Complexes Photophysical, Structural, and Computational Studies. *Inorg. Chem.* **2014**, *53*, 10854-10861.
327. Nair, R. B.; Cullum, B. M.; Murphy, C. J., Optical Properties of [Ru(Phen)2dppz]2+ as a Function of Nonaqueous Environment. *Inorg. Chem.* **1997**, *36*, 962-965.
328. Green, M. L.; Marder, S. R.; Thompson, M. E.; Bandy, J. A.; Bloor, D.; Kolinsky, P.; Jones, R., Synthesis and Structure of (Cis)-[1-Ferrocenyl-2-(4-Nitrophenyl) Ethylene], an Organotransition Metal Compound with a Large Second-Order Optical Nonlinearity. *Nature* **1987**, *330*, 360-362.
329. Bella, S. D., Second-Order Nonlinear Optical Properties of Transition Metal Complexes. *Chem. Soc. Rev.* **2001**, *30*, 355-366.

330. Maury, O.; Le Bozec, H., Molecular Engineering of Octupolar Nlo Molecules and Materials Based on Bipyridyl Metal Complexes. *Acc. Chem. Res.* **2005**, *38*, 691-704.
331. Powell, C. E.; Humphrey, M. G., Nonlinear Optical Properties of Transition Metal Acetylides and Their Derivatives. *Coord. Chem. Rev.* **2004**, *248*, 725-756.
332. Halcrow, M. A., Structure: Function Relationships in Molecular Spin-Crossover Complexes. *Chem. Soc. Rev.* **2011**, *40*, 4119-4142.
333. Gütlich, P., Spin Crossover – Quo Vadis? *Eur. J. Inorg. Chem.* **2013**, *2013*, 581-591.
334. Coronado, E.; Espallargas, G. M., Dynamic Magnetic Mofs. *Chem. Soc. Rev.* **2013**, *42*, 1525-1539.
335. Phonsri, W.; Harding, P.; Liu, L.; Telfer, S. G.; Murray, K. S.; Moubaraki, B.; Ross, T. M.; Jameson, G. N.; Harding, D. J., Solvent Modified Spin Crossover in an Iron (III) Complex: Phase Changes and an Exceptionally Wide Hysteresis. *Chem. Sci.* **2017**, *8*, 3949-3959.
336. Halder, G. J.; Kepert, C. J.; Moubaraki, B.; Murray, K. S.; Cashion, J. D., Guest-Dependent Spin Crossover in a Nanoporous Molecular Framework Material. *Science* **2002**, *298*, 1762-1765.
337. Barrett, S. A.; Kilner, C. A.; Halcrow, M. A., Spin-Crossover in [Fe (3-Bpp) 2][Bf 4] 2 in Different Solvents–a Dramatic Stabilisation of the Low-Spin State in Water. *Dalton Trans.* **2011**, *40*, 12021-12024.
338. Fumanal, M.; Jiménez-Grávalos, F.; Ribas-Arino, J.; Vela, S., Lattice-Solvent Effects in the Spin-Crossover of an Fe(II)-Based Material. The Key Role of Intermolecular Interactions between Solvent Molecules. *Inorg. Chem.* **2017**, *56*, 4474-4483.
339. Decurtins, S.; Gütlich, P.; Köhler, C.; Spiering, H.; Hauser, A., Light-Induced Excited Spin State Trapping in a Transition-Metal Complex: The Hexa-1-Propyltetrazole-Iron (II) Tetrafluoroborate Spin-Crossover System. *Chem. Phys. Lett.* **1984**, *105*, 1-4.
340. Krivokapic, I.; Zerara, M.; Daku, M. L.; Vargas, A.; Enachescu, C.; Ambrus, C.; Tregenna-Piggott, P.; Amstutz, N.; Krausz, E.; Hauser, A., Spin-Crossover in Cobalt (II) Imine Complexes. *Coord. Chem. Rev.* **2007**, *251*, 364-378.
341. Milocco, F.; de Vries, F.; Bartels, I. M.; Havenith, R. W.; Cirera, J.; Demeshko, S.; Meyer, F.; Otten, E., Electronic Control of Spin-Crossover Properties in Four-Coordinate Bis (Formazanate) Iron (II) Complexes. *J. Am. Chem. Soc.* **2020**, *142*, 20170-20181.
342. Creutz, S. E.; Peters, J. C., Spin-State Tuning at Pseudo-Tetrahedral d6 Ions: Spin Crossover in [Bp3] FeII–X Complexes. *Inorg. Chem.* **2016**, *55*, 3894-3906.
343. Gütlich, P.; Bill, E.; Trautwein, A. X., *Mössbauer Spectroscopy and Transition Metal Chemistry: Fundamentals and Applications*; Springer Science & Business Media, 2010.
344. Mabbs, F. E.; Collison, D., *Electron Paramagnetic Resonance of d Transition Metal Compounds*; Elsevier, 2013.
345. Fataftah, M. S.; Krzyaniak, M. D.; Vlaisavljevich, B.; Wasielewski, M. R.; Zadrozny, J. M.; Freedman, D. E., Metal–Ligand Covalency Enables Room Temperature Molecular Qubit Candidates. *Chem. Sci.* **2019**, *10*, 6707-6714.
346. Bader, K.; Dengler, D.; Lenz, S.; Endeward, B.; Jiang, S.-D.; Neugebauer, P.; Van Slageren, J., Room Temperature Quantum Coherence in a Potential Molecular Qubit. *Nat. Commun.* **2014**, *5*, 1-5.
347. Zadrozny, J. M.; Niklas, J.; Poluektov, O. G.; Freedman, D. E., Millisecond Coherence Time in a Tunable Molecular Electronic Spin Qubit. *ACS Cent. Sci.* **2015**, *1*, 488-492.

348. Bogani, L.; Wernsdorfer, W., Molecular Spintronics Using Single-Molecule Magnets. In *Nanoscience and Technology: A Collection of Reviews from Nature Journals*, World Scientific: 2010; pp 194-201.
349. Layfield, R. A., Organometallic Single-Molecule Magnets. *Organometallics* **2014**, *33*, 1084-1099.
350. Murrie, M., Cobalt (II) Single-Molecule Magnets. *Chem. Soc. Rev.* **2010**, *39*, 1986-1995.
351. Burgi, H. B., Chemical Reaction Coordinates from Crystal Structure Data. I. *Inorg. Chem.* **1973**, *12*, 2321-2325.
352. Taylor, R.; Allen, F. H., Statistical and Numerical Methods of Data Analysis

Structure Correlation. In *Structure Correlation*, 1994; pp 111-161.
353. Bruno, I. J.; Cole, J. C.; Edgington, P. R.; Kessler, M.; Macrae, C. F.; McCabe, P.; Pearson, J.; Taylor, R., New Software for Searching the Cambridge Structural Database and Visualizing Crystal Structures. *Acta Crystallogr., Sect. B: Struct. Sci.* **2002**, *58*, 389-397.
354. Reeves, M. G.; Wood, P. A.; Parsons, S., Automated Oxidation-State Assignment for Metal Sites in Coordination Complexes in the Cambridge Structural Database. *Acta Crystallogr., Sect. B: Struct. Sci., Cryst. Eng. Mater.* **2019**, *75*, 1096-1105.
355. Orpen, A., Applications of the Cambridge Structural Database to Molecular Inorganic Chemistry. *Acta Crystallogr., Sect. B: Struct. Sci.* **2002**, *58*, 398-406.
356. Dudev, M.; Wang, J.; Dudev, T.; Lim, C., Factors Governing the Metal Coordination Number in Metal Complexes from Cambridge Structural Database Analyses. *J. Phys. Chem. B* **2006**, *110*, 1889-1895.
357. Orpen, A. G.; Connelly, N. G., Structural Systematics: The Role of P-A .Sigma.* Orbitals in Metal-Phosphorus .Pi.-Bonding in Redox-Related Pairs of M-Pa3 Complexes (a = R, Ar, or; R = Alkyl). *Organometallics* **1990**, *9*, 1206-1210.
358. Dunne, B. J.; Morris, R. B.; Orpen, A. G., Structural Systematics. Part 3. Geometry Deformations in Triphenylphosphine Fragments: A Test of Bonding Theories in Phosphine Complexes. *J. Chem. Soc., Dalton Trans.* **1991**, 653-661.
359. Morton, D. A. V.; Orpen, A. G., Structural Systematics. Part 4. Conformations of the Diphosphine Ligands in M2(M-Ph2pch2pph2) and M(Ph2pch2ch2pph2) Complexes. *J. Chem. Soc., Dalton Trans.* **1992**, 641-653.
360. Garner, S. E.; Orpen, A. G., Structural Systematics. Part 5. Conformation and Bonding in the Chiral Metal Complexes [M(H5-C5r5)(Xo)Z(PPh3)]. *J. Chem. Soc., Dalton Trans.* **1993**, 533-541.
361. Martín, A.; Orpen, A. G., Structural Systematics. 6.1 Apparent Flexibility of Metal Complexes in Crystals. *J. Am. Chem. Soc.* **1996**, *118*, 1464-1470.
362. Alvarez, S.; Llunell, M., Continuous Symmetry Measures of Penta-Coordinate Molecules: Berry and Non-Berry Distortions of the Trigonal Bipyramid. *J. Chem. Soc., Dalton Trans.* **2000**, 3288-3303.
363. Keinan, S.; Avnir, D., Continuous Symmetry Analysis of Tetrahedral/Planar Distortions. Copper Chlorides and Other Ab4 Species. *Inorg. Chem.* **2001**, *40*, 318-323.
364. Alvarez, S.; Avnir, D.; Llunell, M.; Pinsky, M., Continuous Symmetry Maps and Shape Classification. The Case of Six-Coordinated Metal Compounds. *New J. Chem.* **2002**, *26*, 996-1009.
365. Alvarez, S.; Alemany, P.; Casanova, D.; Cirera, J.; Llunell, M.; Avnir, D., Shape Maps and Polyhedral Interconversion Paths in Transition Metal Chemistry. *Coord. Chem. Rev.* **2005**, *249*, 1693-1708.

366. Canton, S.; Zhang, X.; Lawson Daku, M.; Liu, Y.; Zhang, J.; Alvarez, S., Mapping the Ultrafast Changes of Continuous Shape Measures in Photoexcited Spin Crossover Complexes without Long-Range Order. *J. Phys. Chem. C* **2015**, *119*, 3322-3330.
367. Shevchenko, A. P.; Blatov, I. A.; Kitaeva, E. V.; Blatov, V. A., Local Coordination Versus Overall Topology in Crystal Structures: Deriving Knowledge from Crystallographic Databases. *Cryst. Growth Des.* **2017**, *17*, 774-785.
368. Cordero, B.; Gómez, V.; Platero-Prats, A. E.; Revés, M.; Echeverría, J.; Cremades, E.; Barragán, F.; Alvarez, S., Covalent Radii Revisited. *Dalton Trans.* **2008**, 2832-2838.
369. Hocking, R. K.; Hambley, T. W., Structural Measure of Metal−Ligand Covalency from the Bonding in Carboxylate Ligands. *Inorg. Chem.* **2003**, *42*, 2833-2835.
370. Aravena, D.; Ruiz, E., Shedding Light on the Single-Molecule Magnet Behavior of Mononuclear Dyiii Complexes. *Inorg. Chem.* **2013**, *52*, 13770-13778.
371. Nimmermark, A.; Öhrström, L.; Reedijk, J., Metal-Ligand Bond Lengths and Strengths: Are They Correlated? A Detailed Csd Analysis. *Z. Kristallogr. Cryst. Mater.* **2013**, *228*, 311-317.
372. Kuppuraj, G.; Dudev, M.; Lim, C., Factors Governing Metal−Ligand Distances and Coordination Geometries of Metal Complexes. *J. Phys. Chem. B* **2009**, *113*, 2952-2960.
373. Anderson, K. M.; Orpen, A. G., On the Relative Magnitudes of Cis and Trans Influences in Metal Complexes. *Chem. Commun.* **2001**, 2682-2683.
374. Donnelly, M. A.; Zimmer, M., Structural Analysis of All the Nickel 14-Membered Tetraaza Macrocycles in the Cambridge Structural Database. *Inorg. Chem.* **1999**, *38*, 1650-1658.
375. Peigne, B.; Aullon, G., Structural Analysis of the Coordination of Dinitrogen to Transition Metal Complexes. *Acta Crystallogr., Sect. B: Struct. Sci., Cryst. Eng. Mater.* **2015**, *71*, 369-386.
376. Hocking, R. K.; Hambley, T. W., Database Analysis of Transition Metal Carbonyl Bond Lengths: Insight into the Periodicity of π Back-Bonding, Σ Donation, and the Factors Affecting the Electronic Structure of the TM−C:O Moiety. *Organometallics* **2007**, *26*, 2815-2823.
377. Bowman, S. E.; Bridwell-Rabb, J.; Drennan, C. L., Metalloprotein Crystallography: More Than a Structure. *Acc. Chem. Res.* **2016**, *49*, 695-702.
378. Fey, N.; Harris, S. E.; Harvey, J. N.; Orpen, A. G., Adding Value to Crystallographically-Derived Knowledge Bases. *J. Chem. Inf. Model.* **2006**, *46*, 912-929.
379. Chung, Y. G.; Camp, J.; Haranczyk, M.; Sikora, B. J.; Bury, W.; Krungleviciute, V.; Yildirim, T.; Farha, O. K.; Sholl, D. S.; Snurr, R. Q., Computation-Ready, Experimental Metal–Organic Frameworks: A Tool to Enable High-Throughput Screening of Nanoporous Crystals. *Chem. Mater.* **2014**, *26*, 6185-6192.
380. Chung, Y. G.; Haldoupis, E.; Bucior, B. J.; Haranczyk, M.; Lee, S.; Zhang, H.; Vogiatzis, K. D.; Milisavljevic, M.; Ling, S.; Camp, J. S., et al., Advances, Updates, and Analytics for the Computation-Ready, Experimental Metal–Organic Framework Database: Core Mof 2019. *J. Chem. Eng. Data* **2019**, *64*, 5985-5998.
381. Rosen, A. S.; Notestein, J. M.; Snurr, R. Q., Structure–Activity Relationships That Identify Metal–Organic Framework Catalysts for Methane Activation. *ACS Catal.* **2019**, *9*, 3576-3587.
382. Datar, A.; Chung, Y. G.; Lin, L.-C., Beyond the Bet Analysis: The Surface Area Prediction of Nanoporous Materials Using a Machine Learning Method. *J. Phys. Chem. Lett.* **2020**, *11*, 5412-5417.

383. Anderson, R.; Biong, A.; Gómez-Gualdrón, D. A., Adsorption Isotherm Predictions for Multiple Molecules in Mofs Using the Same Deep Learning Model. *J. Chem. Theory Comput.* **2020**, *16*, 1271-1283.
384. Moosavi, S. M.; Nandy, A.; Jablonka, K. M.; Ongari, D.; Janet, J. P.; Boyd, P. G.; Lee, Y.; Smit, B.; Kulik, H. J., Understanding the Diversity of the Metal-Organic Frameworks Ecosystems. *Nat. Commun.* **2020**, *11*, 4068.
385. Kim, E.; Huang, K.; Saunders, A.; McCallum, A.; Ceder, G.; Olivetti, E., Materials Synthesis Insights from Scientific Literature via Text Extraction and Machine Learning. *Chem. Mater.* **2017**, *29*, 9436-9444.
386. Park, S.; Kim, B.; Choi, S.; Boyd, P. G.; Smit, B.; Kim, J., Text Mining Metal–Organic Framework Papers. *J. Chem. Inf. Model.* **2018**, *58*, 244-251.
387. Moghadam, P. Z.; Li, A.; Liu, X.-W.; Bueno-Perez, R.; Wang, S.-D.; Wiggin, S. B.; Wood, P. A.; Fairen-Jimenez, D., Targeted Classification of Metal–Organic Frameworks in the Cambridge Structural Database (Csd). *Chem. Sci.* **2020**, *11*, 8373-8387.
388. Balcells, D.; Skjelstad, B. B., Tmqm Dataset—Quantum Geometries and Properties of 86k Transition Metal Complexes. *J. Chem. Inf. Model.* **2020**, *60*, 6135-6146.
389. Curtarolo, S.; Setyawan, W.; Wang, S.; Xue, J.; Yang, K.; Taylor, R. H.; Nelson, L. J.; Hart, G. L.; Sanvito, S.; Buongiorno-Nardelli, M., AFLOWLIB. Org: A Distributed Materials Properties Repository from High-Throughput Ab Initio Calculations. *Comput. Mater. Sci.* **2012**, *58*, 227-235.
390. Jain, A.; Ong, S. P.; Hautier, G.; Chen, W.; Richards, W. D.; Dacek, S.; Cholia, S.; Gunter, D.; Skinner, D.; Ceder, G., et al., Commentary: The Materials Project: A Materials Genome Approach to Accelerating Materials Innovation. *APL Mater.* **2013**, *1*, 011002.
391. Saal, J. E.; Kirklin, S.; Aykol, M.; Meredig, B.; Wolverton, C., Materials Design and Discovery with High-Throughput Density Functional Theory: The Open Quantum Materials Database (OQMD). *JOM* **2013**, *65*, 1501-1509.
392. Fey, N.; Tsipis, A. C.; Harris, S. E.; Harvey, J. N.; Orpen, A. G.; Mansson, R. A., Development of a Ligand Knowledge Base, Part 1: Computational Descriptors for Phosphorus Donor Ligands. *Chem. - Eur. J.* **2006**, *12*, 291-302.
393. Mansson, R. A.; Welsh, A. H.; Fey, N.; Orpen, A. G., Statistical Modeling of a Ligand Knowledge Base. *J. Chem. Inf. Model.* **2006**, *46*, 2591-2600.
394. Ahn, S.; Hong, M.; Sundararajan, M.; Ess, D. H.; Baik, M.-H., Design and Optimization of Catalysts Based on Mechanistic Insights Derived from Quantum Chemical Reaction Modeling. *Chem. Rev.* **2019**, *119*, 6509-6560.
395. Zahrt, A. F.; Athavale, S. V.; Denmark, S. E., Quantitative Structure–Selectivity Relationships in Enantioselective Catalysis: Past, Present, and Future. *Chem. Rev.* **2020**, *120*, 1620-1689.
396. Falivene, L.; Cao, Z.; Petta, A.; Serra, L.; Poater, A.; Oliva, R.; Scarano, V.; Cavallo, L., Towards the Online Computer-Aided Design of Catalytic Pockets. *Nat. Chem.* **2019**, *11*, 872-879.
397. Tolman, C. A., Phosphorus Ligand Exchange Equilibriums on Zerovalent Nickel. Dominant Role for Steric Effects. *J. Am. Chem. Soc.* **1970**, *92*, 2956-2965.
398. Tolman, C. A., Steric Effects of Phosphorus Ligands in Organometallic Chemistry and Homogeneous Catalysis. *Chem. Rev.* **1977**, *77*, 313-348.
399. Hillier, A. C.; Sommer, W. J.; Yong, B. S.; Petersen, J. L.; Cavallo, L.; Nolan, S. P., A Combined Experimental and Theoretical Study Examining the Binding of N-Heterocyclic

Carbenes (Nhc) to the Cp*Rucl (Cp* = H^5-C_5me_5) Moiety: Insight into Stereoelectronic Differences between Unsaturated and Saturated Nhc Ligands. *Organometallics* **2003**, *22*, 4322-4326.
400. Falivene, L.; Credendino, R.; Poater, A.; Petta, A.; Serra, L.; Oliva, R.; Scarano, V.; Cavallo, L., Sambvca 2. A Web Tool for Analyzing Catalytic Pockets with Topographic Steric Maps. *Organometallics* **2016**, *35*, 2286-2293.
401. 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.
402. Brown, T. L., A Molecular Mechanics Model of Ligand Effects. 3. A New Measure of Ligand Steric Effects. *Inorg. Chem.* **1992**, *31*, 1286-1294.
403. Cundari, T. R.; Deng, J.; Pop, H. F.; Sârbu, C., Structural Analysis of Transition Metal β-X Substituent Interactions. Toward the Use of Soft Computing Methods for Catalyst Modeling. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1052-1061.
404. Dierkes, P.; van Leeuwen, P. W. N. M., The Bite Angle Makes the Difference: A Practical Ligand Parameter for Diphosphine Ligands. *J. Chem. Soc., Dalton Trans.* **1999**, 1519-1530.
405. Phan, H.; Hrudka, J. J.; Igimbayeva, D.; Lawson Daku, L. v. M.; Shatruk, M., A Simple Approach for Predicting the Spin State of Homoleptic Fe (II) Tris-Diimine Complexes. *J. Am. Chem. Soc.* **2017**, *139*, 6437-6447.
406. Fey, N.; Harvey, J. N.; Lloyd-Jones, G. C.; Murray, P.; Orpen, A. G.; Osborne, R.; Purdie, M., Computational Descriptors for Chelating P,P- and P,N-Donor Ligands. *Organometallics* **2008**, *27*, 1372-1383.
407. Lipkowitz, K. B.; Schefzick, S.; Avnir, D., Enhancement of Enantiomeric Excess by Ligand Distortion. *J. Am. Chem. Soc.* **2001**, *123*, 6710-6711.
408. Lipkowitz, K. B.; Schefzick, S., Ligand Distortion Modes Leading to Increased Chirality Content of Katsuki-Jacobsen Catalysts. *Chirality* **2002**, *14*, 677-682.
409. Cramer, R. D.; Patterson, D. E.; Bunce, J. D., Comparative Molecular Field Analysis (Comfa). 1. Effect of Shape on Binding of Steroids to Carrier Proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959-5967.
410. Burello, E.; Rothenberg, G., In Silico Design in Homogeneous Catalysis Using Descriptor Modelling. *Int. J. Mol. Sci.* **2006**, *7*, 375-404.
411. Martínez, S.; Cruz, V. L.; Ramos, J.; Martínez-Salazar, J., Polymerization Activity Prediction of Zirconocene Single-Site Catalysts Using 3D Quantitative Structure–Activity Relationship Modeling. *Organometallics* **2012**, *31*, 1673-1679.
412. Urbano-Cuadrado, M.; Carbó, J. J.; Maldonado, A. G.; Bo, C., New Quantum Mechanics-Based Three-Dimensional Molecular Descriptors for Use in Qssr Approaches: Application to Asymmetric Catalysis. *J. Chem. Inf. Model.* **2007**, *47*, 2228-2234.
413. Cruz, V. L.; Martinez, S.; Ramos, J.; Martinez-Salazar, J., 3D-QSAR as a Tool for Understanding and Improving Single-Site Polymerization Catalysts. A Review. *Organometallics* **2014**, *33*, 2944-2959.
414. Lipkowitz, K. B.; D'Hue, C. A.; Sakamoto, T.; Stack, J. N., Stereocartography: A Computational Mapping Technique That Can Locate Regions of Maximum Stereoinduction around Chiral Catalysts. *J. Am. Chem. Soc.* **2002**, *124*, 14255-14267.
415. Lipkowitz, K. B.; Kozlowski, M. C., Understanding Stereoinduction in Catalysis via Computer: New Tools for Asymmetric Synthesis. *Synlett* **2003**, *2003*, 1547-1565.

416. Hansch, C.; Leo, A.; Taft, R., A Survey of Hammett Substituent Constants and Resonance and Field Parameters. *Chem. Rev.* **1991**, *91*, 165-195.
417. Hammett, L. P., The Effect of Structure Upon the Reactions of Organic Compounds. Benzene Derivatives. *J. Am. Chem. Soc.* **1937**, *59*, 96-103.
418. Perrin, L.; Clot, E.; Eisenstein, O.; Loch, J.; Crabtree, R. H., Computed Ligand Electronic Parameters from Quantum Chemistry and Their Relation to Tolman Parameters, Lever Parameters, and Hammett Constants. *Inorg. Chem.* **2001**, *40*, 5806-5811.
419. Yang, B.; Truhlar, D. G., Computational Design of an Iron Catalyst for Olefin Metathesis. *Organometallics* **2018**, *37*, 3917-3927.
420. Cruz, V. L.; Ramos, J.; Martinez, S.; Muñoz-Escalona, A.; Martinez-Salazar, J., Structure−Activity Relationship Study of the Metallocene Catalyst Activity in Ethylene Polymerization. *Organometallics* **2005**, *24*, 5095-5102.
421. Schilling, M.; Patzke, G. R.; Hutter, J.; Luber, S., Computational Investigation and Design of Cobalt Aqua Complexes for Homogeneous Water Oxidation. *J. Phys. Chem. C* **2016**, *120*, 7966-7975.
422. Kim, J. Y.; Kulik, H. J., When Is Ligand pKa a Good Descriptor for Catalyst Energetics? In Search of Optimal CO2 Hydration Catalysts. *J. Phys. Chem. A* **2018**, *122*, 4579-4590.
423. Lipkowitz, K. B.; Pradhan, M., Computational Studies of Chiral Catalysts: A Comparative Molecular Field Analysis of an Asymmetric Diels−Alder Reaction with Catalysts Containing Bisoxazoline or Phosphinooxazoline Ligands. *J. Org. Chem.* **2003**, *68*, 4648-4656.
424. Baik, M.-H.; Friesner, R. A., Computing Redox Potentials in Solution: Density Functional Theory as a Tool for Rational Design of Redox Agents. *J. Phys. Chem. A* **2002**, *106*, 7407-7412.
425. Comba, P.; Sickmüller, A. F., Modeling of the Redox Properties of (Hexaamine)Cobalt(III/II) Couples. *Inorg. Chem.* **1997**, *36*, 4500-4507.
426. Hambley, T. W., Steric Contributions to the Thermodynamics of Electron Transfer in Cobalt(III) Hexaamine Complexes. *Inorg. Chem.* **1988**, *27*, 2496-2501.
427. Haines, D. E.; O'Hanlon, D. C.; Manna, J.; Jones, M. K.; Shaner, S. E.; Sun, J.; Hopkins, M. D., Oxidation-Potential Tuning of Tungsten–Alkylidyne Complexes over a 2 V Range. *Inorg. Chem.* **2013**, *52*, 9650-9658.
428. Bowman, D. N.; Bondarev, A.; Mukherjee, S.; Jakubikova, E., Tuning the Electronic Structure of Fe(II) Polypyridines via Donor Atom and Ligand Scaffold Modifications: A Computational Study. *Inorg. Chem.* **2015**, *54*, 8786-8793.
429. Dam, H. C.; Pham, T. L.; Ho, T. B.; Nguyen, A. T.; Nguyen, V. C., Data Mining for Materials Design: A Computational Study of Single Molecule Magnet. *J. Chem. Phys.* **2014**, *140*, 044101.
430. Santiago, C. B.; Guo, J.-Y.; Sigman, M. S., Predictive and Mechanistic Multivariate Linear Regression Models for Reaction Development. *Chem. Sci.* **2018**, *9*, 2398-2412.
431. Harper, K. C.; Sigman, M. S., Predicting and Optimizing Asymmetric Catalyst Performance Using the Principles of Experimental Design and Steric Parameters. *Proc. Natl. Acad. Sci.* **2011**, *108*, 2179-2183.
432. Orlandi, M.; Hilton, M. J.; Yamamoto, E.; Toste, F. D.; Sigman, M. S., Mechanistic Investigations of the Pd(0)-Catalyzed Enantioselective 1,1-Diarylation of Benzyl Acrylates. *J. Am. Chem. Soc.* **2017**, *139*, 12688-12695.

433. Sandford, C.; Fries, L. R.; Ball, T. E.; Minteer, S. D.; Sigman, M. S., Mechanistic Studies into the Oxidative Addition of Co(I) Complexes: Combining Electroanalytical Techniques with Parameterization. *J. Am. Chem. Soc.* **2019**, *141*, 18877-18889.
434. Brethome, 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.
435. Parveen, R.; Cundari, T. R.; Younker, J. M.; Rodriguez, G.; McCullough, L., DFT and QSAR Studies of Ethylene Polymerization by Zirconocene Catalysts. *ACS Catal.* **2019**, *9*, 9339-9349.
436. Occhipinti, G.; Bjørsvik, H.-R.; Jensen, V. R., Quantitative Structure−Activity Relationships of Ruthenium Catalysts for Olefin Metathesis. *J. Am. Chem. Soc.* **2006**, *128*, 6952-6964.
437. Nazemi, A.; Cundari, T. R., Control of C–H Bond Activation by Mo-Oxo Complexes: pKa or Bond Dissociation Free Energy (Bdfe)? *Inorg. Chem.* **2017**, *56*, 12319-12327.
438. Kim, J. Y.; Steeves, A. H.; Kulik, H. J., Harnessing Organic Ligand Libraries for First-Principles Inorganic Discovery: Indium Phosphide Quantum Dot Precursor Design Strategies. *Chem. Mater.* **2017**, *29*, 3632-3643.
439. Lakuntza, O.; Besora, M.; Maseras, F., Searching for Hidden Descriptors in the Metal-Ligand Bond through Statistical Analysis of Density Functional Theory (DFT) Results. *Inorg. Chem.* **2018**, *57*, 14660-14670.
440. Fu, R.; Nielsen, R. J.; Goddard, W. A.; Fortman, G. C.; Gunnoe, T. B., DFT Virtual Screening Identifies Rhodium–Amidinate Complexes as Potential Homogeneous Catalysts for Methane-to-Methanol Oxidation. *ACS Catal.* **2014**, *4*, 4455-4465.
441. Andrikopoulos, P. C.; Michel, C.; Chouzier, S.; Sautet, P., In Silico Screening of Iron-Oxo Catalysts for CH Bond Cleavage. *ACS Catal.* **2015**, *5*, 2490-2499.
442. Cooney, K. D.; Cundari, T. R.; Hoffman, N. W.; Pittard, K. A.; Temple, M. D.; Zhao, Y., A Priori Assessment of the Stereoelectronic Profile of Phosphines and Phosphites. *J. Am. Chem. Soc.* **2003**, *125*, 4318-4324.
443. Occhipinti, G.; Koudriavtsev, V.; Törnroos, K. W.; Jensen, V. R., Theory-Assisted Development of a Robust and Z-Selective Olefin Metathesis Catalyst. *Dalton Trans.* **2014**, *43*, 11106-11117.
444. Gillespie, A. M.; White, D. P., Understanding the Steric Control of Stereoselective Olefin Binding in Cyclopentadienyl Complexes of Rhenium: An Application of De Novo Ligand Design. *Organometallics* **2001**, *20*, 5149-5155.
445. Nørskov, J. K.; Bligaard, T.; Rossmeisl, J.; Christensen, C. H., Towards the Computational Design of Solid Catalysts. *Nat. Chem.* **2009**, *1*, 37-46.
446. Falivene, L.; Kozlov, S. M.; Cavallo, L., Constructing Bridges between Computational Tools in Heterogeneous and Homogeneous Catalysis. *ACS Catal.* **2018**, *8*, 5637-5656.
447. Latimer, A. A.; Kulkarni, A. R.; Aljama, H.; Montoya, J. H.; Yoo, J. S.; Tsai, C.; Abild-Pedersen, F.; Studt, F.; Nørskov, J. K., Understanding Trends in C-H Bond Activation in Heterogeneous Catalysis. *Nat. Mater.* **2017**, *16*, 225-229.
448. Liao, P.; Getman, R. B.; Snurr, R. Q., Optimizing Open Iron Sites in Metal–Organic Frameworks for Ethane Oxidation: A First-Principles Study. *ACS Appl. Mater. Interfaces* **2017**, *9*, 33484-33492.
449. Yang, X.-F.; Wang, A.; Qiao, B.; Li, J.; Liu, J.; Zhang, T., Single-Atom Catalysts: A New Frontier in Heterogeneous Catalysis. *Acc. Chem. Res.* **2013**, *46*, 1740-1748.

450. Flytzani-Stephanopoulos, M., Gold Atoms Stabilized on Various Supports Catalyze the Water–Gas Shift Reaction. *Acc. Chem. Res.* **2014**, *47*, 783-792.
451. Xu, H.; Cheng, D.; Cao, D.; Zeng, X. C., A Universal Principle for a Rational Design of Single-Atom Electrocatalysts. *Nat. Catal.* **2018**, *1*, 339-348.
452. Wodrich, M. D.; Sawatlon, B.; Solel, E.; Kozuch, S.; Corminboeuf, C., Activity-Based Screening of Homogeneous Catalysts through the Rapid Assessment of Theoretically Derived Turnover Frequencies. *ACS Catal.* **2019**, *9*, 5716-5725.
453. Wodrich, M. D.; Busch, M.; Corminboeuf, C., Accessing and Predicting the Kinetic Profiles of Homogeneous Catalysts from Volcano Plots. *Chem. Sci.* **2016**, *7*, 5723-5735.
454. Busch, M.; Wodrich, M. D.; Corminboeuf, C., A Generalized Picture of C–C Cross-Coupling. *ACS Catal.* **2017**, *7*, 5643-5653.
455. Busch, M.; Wodrich, M. D.; Corminboeuf, C., Linear Scaling Relationships and Volcano Plots in Homogeneous Catalysis – Revisiting the Suzuki Reaction. *Chem. Sci.* **2015**, *6*, 6754-6761.
456. Gani, T. Z. H.; Kulik, H. J., Understanding and Breaking Scaling Relations in Single-Site Catalysis: Methane-to-Methanol Conversion by Fe(IV)=O. *ACS Catal.* **2018**, *8*, 975-986.
457. Nandy, A.; Kulik, H. J., Why Conventional Design Rules for C-H Activation Fail for Open Shell Transition Metal Catalysts. *ACS Catal.* **2020**, *10*, 15033-15047.
458. Anand, M.; Nørskov, J. K., Scaling Relations in Homogeneous Catalysis: Analyzing the Buchwald–Hartwig Amination Reaction. *ACS Catal.* **2019**, *10*, 336-345.
459. Busico, V.; Cipullo, R.; Pellecchia, R.; Ronca, S.; Roviello, G.; Talarico, G., Design of Stereoselective Ziegler-Natta Propene Polymerization Catalysts. *Proc. Natl. Acad. Sci.* **2006**, *103*, 15321-15326.
460. Herbert, M. B.; Suslick, B. A.; Liu, P.; Zou, L.; Dornan, P. K.; Houk, K. N.; Grubbs, R. H., Cyclometalated Z-Selective Ruthenium Metathesis Catalysts with Modified N-Chelating Groups. *Organometallics* **2015**, *34*, 2858-2869.
461. Xu, Y.; Wong, J. J.; Samkian, A. E.; Ko, J. H.; Chen, S.; Houk, K. N.; Grubbs, R. H., Efficient Z-Selective Olefin-Acrylamide Cross-Metathesis Enabled by Sterically Demanding Cyclometalated Ruthenium Catalysts. *J. Am. Chem. Soc.* **2020**, *142*, 20987-20993.
462. Kwon, D.-H.; Fuller, J. T.; Kilgore, U. J.; Sydora, O. L.; Bischof, S. M.; Ess, D. H., Computational Transition-State Design Provides Experimentally Verified Cr(P,N) Catalysts for Control of Ethylene Trimerization and Tetramerization. *ACS Catal.* **2018**, *8*, 1138-1142.
463. Sperger, T.; Sanhueza, I. A.; Schoenebeck, F., Computation and Experiment: A Powerful Combination to Understand and Predict Reactivities. *Acc. Chem. Res.* **2016**, *49*, 1311-1319.
464. Mondal, B.; Neese, F.; Ye, S., Toward Rational Design of 3d Transition Metal Catalysts for CO_2 Hydrogenation Based on Insights into Hydricity-Controlled Rate-Determining Steps. *Inorg. Chem.* **2016**, *55*, 5438-5444.
465. Ramakrishnan, S.; Moretti, R. A.; Chidsey, C. E. D., Mapping Free Energy Regimes in Electrocatalytic Reductions to Screen Transition Metal-Based Catalysts. *Chem. Sci.* **2019**, *10*, 7649-7658.
466. Uhe, A.; Hölscher, M.; Leitner, W., Carboxylation of Arene C-H Bonds with CO_2: A DFT-Based Approach to Catalyst Design. *Chem. - Eur. J.* **2012**, *18*, 170-177.
467. Orian, L.; Wolters, L. P.; Bickelhaupt, F. M., In Silico Design of Heteroaromatic Half-Sandwich Rh^i Catalysts for Acetylene [2+2+2] Cyclotrimerization: Evidence of a Reverse Indenyl Effect. *Chem. - Eur. J.* **2013**, *19*, 13337-13347.

468. Nandy, A.; Zhu, J.; Janet, J. P.; Duan, C.; Getman, R. B.; Kulik, H. J., Machine Learning Accelerates the Discovery of Design Rules and Exceptions in Stable Metal- Oxo Intermediate Formation. *ACS Catal.* **2019**, *9*, 8243-8255.
469. Cremades, E.; Ruiz, E., Mononuclear Feiisingle-Molecule Magnets: A Theoretical Approach. *Inorg. Chem.* **2011**, *50*, 4016-4020.
470. Pickup, O. J. S.; Khazal, I.; Smith, E. J.; Whitwood, A. C.; Lynam, J. M.; Bolaky, K.; King, T. C.; Rawe, B. W.; Fey, N., Computational Discovery of Stable Transition-Metal Vinylidene Complexes. *Organometallics* **2014**, *33*, 1751-1761.
471. Shiekh, B. A.; Kaur, D.; Kumar, S., Bio-Mimetic Self-Assembled Computationally Designed Catalysts of Mo and W for Hydrogenation of CO_2/Dehydrogenation of Hcooh Inspired by the Active Site of Formate Dehydrogenase. *Phys. Chem. Chem. Phys.* **2019**, *21*, 21370-21380.
472. Minkin, V. I.; Starikova, A. A.; Minyaev, R. M., Computational Design of Valence Tautomeric Adducts of Co^{ii} Diketonates with Redox-Active O-Benzoquinone Ligands. *Dalton Trans.* **2013**, *42*, 1726-1734.
473. Nazemi, A.; Cundari, T. R., Computational Analysis of Proton-Coupled Electron Transfer in Hydrotris(Triazolyl)Borate Mid–Late 3d and 4d Transition Metal Complexes. *Organometallics* **2019**, *38*, 3521-3531.
474. Rolfes, J. D.; van Gastel, M.; Neese, F., Where Is the Fluoro Wall?: A Quantum Chemical Investigation. *Inorg. Chem.* **2020**, *59*, 1556-1565.
475. Mavros, M. G.; Shepherd, J. J.; Tsuchimochi, T.; McIsaac, A. R.; Van Voorhis, T., Computational Design Principles of Two-Center First-Row Transition Metal Oxide Oxygen Evolution Catalysts. *J. Phys. Chem. C* **2017**, *121*, 15665-15674.
476. Nielsen, M. C.; Bonney, K. J.; Schoenebeck, F., Computational Ligand Design for the Reductive Elimination of $Arcf_3$ from a Small Bite Angle Pd^{ii} Complex: Remarkable Effect of a Perfluoroalkyl Phosphine. *Angew. Chem., Int. Ed.* **2014**, *53*, 5903-5906.
477. Chen, X.; Yang, X., Bioinspired Design and Computational Prediction of Iron Complexes with Pendant Amines for the Production of Methanol from CO_2 and H_2. *J. Phys. Chem. Lett.* **2016**, *7*, 1035-1041.
478. Burrows, L. C.; Jesikiewicz, L. T.; Lu, G.; Geib, S. J.; Liu, P.; Brummond, K. M., Computationally Guided Catalyst Design in the Type I Dynamic Kinetic Asymmetric Pauson–Khand Reaction of Allenyl Acetates. *J. Am. Chem. Soc.* **2017**, *139*, 15022-15032.
479. Farha, O. K.; Özgür Yazaydın, A.; Eryazici, I.; Malliakas, C. D.; Hauser, B. G.; Kanatzidis, M. G.; Nguyen, S. T.; Snurr, R. Q.; Hupp, J. T., De Novo Synthesis of a Metal–Organic Framework Material Featuring Ultrahigh Surface Area and Gas Storage Capacities. *Nat. Chem.* **2010**, *2*, 944-948.
480. Gomez-Gualdron, D. A.; Gutov, O. V.; Krungleviciute, V.; Borah, B.; Mondloch, J. E.; Hupp, J. T.; Yildirim, T.; Farha, O. K.; Snurr, R. Q., Computational Design of Metal–Organic Frameworks Based on Stable Zirconium Building Units for Storage and Delivery of Methane. *Chem. Mater.* **2014**, *26*, 5632-5639.
481. Avci, G.; Velioglu, S.; Keskin, S., In Silico Design of Metal Organic Frameworks with Enhanced CO_2 Separation Performances: Role of Metal Sites. *J. Phys. Chem. C* **2019**, *123*, 28255-28265.
482. Cundari, T. R.; Sârbu, C.; Pop, H. F., Robust Fuzzy Principal Component Analysis (Fpca). A Comparative Study Concerning Interaction of Carbon−Hydrogen Bonds with Molybdenum−Oxo Bonds. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1363-1369.

483. Jover, J.; Fey, N., The Computational Road to Better Catalysts. *Chem. - Asian J.* **2014**, *9*, 1714-1723.
484. 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.
485. 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 Chelating P,P-Donor Ligands (Lkb-Pp). *Organometallics* **2012**, *31*, 5302-5306.
486. Fey, N.; Haddow, M. F.; Harvey, J. N.; McMullin, C. L.; Orpen, A. G., A Ligand Knowledge Base for Carbenes (Lkb-C): Maps of Ligand Space. *Dalton Trans.* **2009**, 8183-8196.
487. Beyreuther, S.; Hunger, J.; Huttner, G.; Mann, S.; Zsolnai, L., Conformation of Tripod Metal Templates in Ch3c(Ch2pph2)3mln (N = 2, 3): Neural Networks in Conformational Analysis. *Chem. Ber.* **1996**, *129*, 745-757.
488. Saadun, A. J.; Pablo-García, S.; Paunović, V.; Li, Q.; Sabadell-Rendón, A.; Kleemann, K.; Krumeich, F.; López, N.; Pérez-Ramírez, J., Performance of Metal-Catalyzed Hydrodebromination of Dibromomethane Analyzed by Descriptors Derived from Statistical Learning. *ACS Catal.* **2020**, *10*, 6129-6143.
489. Maldonado, A. G.; Rothenberg, G., Predictive Modeling in Homogeneous Catalysis: A Tutorial. *Chem. Soc. Rev.* **2010**, *39*, 1891.
490. Maggiora, G.; Medina-Franco, J. L.; Iqbal, J.; Vogt, M.; Bajorath, J., From Qualitative to Quantitative Analysis of Activity and Property Landscapes. *J. Chem. Inf. Model.* **2020**, *60*, 5873-5880.
491. Burello, E.; Farrusseng, D.; Rothenberg, G., Combinatorial Explosion in Homogeneous Catalysis: Screening 60,000 Cross-Coupling Reactions. *Adv. Synth. Catal.* **2004**, *346*, 1844-1853.
492. See, X. Y.; Wen, X.; Wheeler, T. A.; Klein, C. K.; Goodpaster, J. D.; Reiner, B. R.; Tonks, I. A., Iterative Supervised Principal Component Analysis Driven Ligand Design for Regioselective Ti-Catalyzed Pyrrole Synthesis. *ACS Catal.* **2020**, *10*, 13504--13517.
493. van der Maaten, L.; Hinton, G., Visualizing Data Using T-Sne. *J. Mach. Learn. Res.* **2008**, *9*, 2579-2605.
494. McInnes, L. Umap: Uniform Manifold Approximation and Projection for Dimension Reduction Software. https://umap-learn.readthedocs.io/en/latest/ (accessed May 23, 2020).
495. McInnes, L.; Healy, J.; Melville, J., Umap: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv:1802.03426* **2018**.
496. Chang, A. M.; Freeze, J. G.; Batista, V. S., Hammett Neural Networks: Prediction of Frontier Orbital Energies of Tungsten–Benzylidyne Photoredox Complexes. *Chem. Sci.* **2019**, *10*, 6844-6854.
497. Kier, L. B., A Shape Index from Molecular Graphs. *Quant. Struct.-Act. Relat.* **1985**, *4*, 109-116.
498. Moreau, G.; Broto, P., The Autocorrelation of a Topological Structure: A New Molecular Descriptor. *Nouv. J. Chim.* **1980**, *4*, 359.
499. Burello, E.; Rothenberg, G., Topological Mapping of Bidentate Ligands: A Fast Approach for Screening Homogeneous Catalysts. *Adv. Synth. Catal.* **2005**, *347*, 1969-1977.
500. Janet, J. P.; Kulik, H. J., Resolving Transition Metal Chemical Space: Feature Selection for Machine Learning and Structure-Property Relationships. *J. Phys. Chem. A* **2017**, *121*, 8939-8954.

501. Virshup, A. M.; Contreras-García, J.; Wipf, P.; Yang, W.; Beratan, D. N., Stochastic Voyages into Uncharted Chemical Space Produce a Representative Library of All Possible Drug-Like Compounds. *J. Am. Chem. Soc.* **2013**, *135*, 7296-7303.
502. Allahyari, Z.; Oganov, A. R., Nonempirical Definition of the Mendeleev Numbers: Organizing the Chemical Space. *J. Phys. Chem. C* **2020**, *124*, 23867-23878.
503. Ramakrishnan, R.; Dral, P. O.; Rupp, M.; Von Lilienfeld, O. A., Quantum Chemistry Structures and Properties of 134 Kilo Molecules. *Sci. Data* **2014**, *1*, 140022.
504. Rupp, M.; Tkatchenko, A.; Muller, K. R.; von Lilienfeld, O. A., Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning. *Phys. Rev. Lett.* **2012**, *108*, 058301.
505. Hansen, K.; Biegler, F.; Ramakrishnan, R.; Pronobis, W.; Von Lilienfeld, O. A.; Müller, K.-R.; Tkatchenko, A., Machine Learning Predictions of Molecular Properties: Accurate Many-Body Potentials and Nonlocality in Chemical Space. *J. Phys. Chem. Lett.* **2015**, *6*, 2326-2331.
506. Bartók, A. P.; Kondor, R.; Csányi, G., On Representing Chemical Environments. *Phys. Rev. B* **2013**, *87*, 184115.
507. Huang, B.; von Lilienfeld, O. A., Quantum Machine Learning Using Atom-in-Molecule-Based Fragments Selected on the Fly. *Nat. Chem.* **2020**, *12*, 945-951.
508. Huo, H.; Rupp, M., Unified Representation of Molecules and Crystals for Machine Learning. *arXiv preprint arXiv:1704.06439* **2017**.
509. Duan, C.; Liu, F.; Nandy, A.; Kulik, H. J., Data-Driven Approaches Can Overcome the Cost-Accuracy Tradeoff in Multireference Diagnostics. *J. Chem. Theory Comput.* **2020**, *16*, 4373-4387.
510. Anderson, R.; Rodgers, J.; Argueta, E.; Biong, A.; Gómez-Gualdrón, D. A., Role of Pore Chemistry and Topology in the CO_2 Capture Capabilities of Mofs: From Molecular Simulation to Machine Learning. *Chem. Mater.* **2018**, *30*, 6325-6337.
511. Martin, R. L.; Smit, B.; Haranczyk, M., Addressing Challenges of Identifying Geometrically Diverse Sets of Crystalline Porous Materials. *J. Chem. Inf. Model.* **2011**, *52*, 308-318.
512. Willems, T. F.; Rycroft, C. H.; Kazi, M.; Meza, J. C.; Haranczyk, M., Algorithms and Tools for High-Throughput Geometry-Based Analysis of Crystalline Porous Materials. *Microporous Mesoporous Mater.* **2012**, *149*, 134-141.
513. Bucior, B. J.; Bobbitt, N. S.; Islamoglu, T.; Goswami, S.; Gopalan, A.; Yildirim, T.; Farha, O. K.; Bagheri, N.; Snurr, R. Q., Energy-Based Descriptors to Rapidly Predict Hydrogen Storage in Metal–Organic Frameworks. *Mol. Syst. Des. Eng.* **2019**, *4*, 162-174.
514. Bucior, B. J.; Rosen, A. S.; Haranczyk, M.; Yao, Z.; Ziebel, M. E.; Farha, O. K.; Hupp, J. T.; Siepmann, J. I.; Aspuru-Guzik, A.; Snurr, R. Q., Identification Schemes for Metal–Organic Frameworks to Enable Rapid Search and Cheminformatics Analysis. *Cryst. Growth Des.* **2019**, *19*, 6682-6697.
515. Pardakhti, M.; Moharreri, E.; Wanik, D.; Suib, S. L.; Srivastava, R., Machine Learning Using Combined Structural and Chemical Descriptors for Prediction of Methane Adsorption Performance of Metal Organic Frameworks (Mofs). *ACS Comb. Sci.* **2017**, *19*, 640-645.
516. Fanourgakis, G. S.; Gkagkas, K.; Tylianakis, E.; Froudakis, G. E., A Universal Machine Learning Algorithm for Large-Scale Screening of Materials. *J. Am. Chem. Soc.* **2020**, *142*, 3814-3822.

517. Borboudakis, G.; Stergiannakos, T.; Frysali, M.; Klontzas, E.; Tsamardinos, I.; Froudakis, G. E., Chemically Intuited, Large-Scale Screening of Mofs by Machine Learning Techniques. *npj Comput. Mater.* **2017**, *3*, 40.
518. Fernandez, M.; Trefiak, N. R.; Woo, T. K., Atomic Property Weighted Radial Distribution Functions Descriptors of Metal–Organic Frameworks for the Prediction of Gas Uptake Capacity. *J. Phys. Chem. C* **2013**, *117*, 14095-14105.
519. Duan, C.; Janet, J. P.; Liu, F.; Nandy, A.; Kulik, H. J., Learning from Failure: Predicting Electronic Structure Calculation Outcomes with Machine Learning Models. *J. Chem. Theory Comput.* **2019**, *15*, 2331-2345.
520. Landrum, G. A.; Penzotti, J. E.; Putta, S., Machine-Learning Models for Combinatorial Catalyst Discovery. *Meas. Sci. Technol.* **2005**, *16*, 270-277.
521. Yada, A.; Nagata, K.; Ando, Y.; Matsumura, T.; Ichinoseki, S.; Sato, K., Machine Learning Approach for Prediction of Reaction Yield with Simulated Catalyst Parameters. *Chem. Lett.* **2018**, *47*, 284-287.
522. Welborn, M.; Cheng, L.; Miller III, T. F., Transferability in Machine Learning for Electronic Structure via the Molecular Orbital Basis. *J. Chem. Theory Comput.* **2018**, *14*, 4772-4779.
523. Husch, T.; Sun, J.; Cheng, L.; Lee, S. J. R.; Miller, T. F., Improved Accuracy and Transferability of Molecular-Orbital-Based Machine Learning: Organics, Transition-Metal Complexes, Non-Covalent Interactions, and Transition States. *J. Chem. Phys.* **2021**, *154*, 064108.
524. Mikami, K., Interactive-Quantum-Chemical-Descriptors Enabling Accurate Prediction of an Activation Energy through Machine Learning. *Polymer* **2020**, *203*, 122738.
525. Burello, E.; Rothenberg, G., Optimal Heck Cross-Coupling Catalysis: A Pseudo-Pharmaceutical Approach. *Adv. Synth. Catal.* **2003**, *345*, 1334-1340.
526. Gal, Y.; Ghahramani, Z. In *Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning*, international conference on machine learning, 2016; pp 1050-1059.
527. Janet, J. P.; Duan, C.; Yang, T.; Nandy, A.; Kulik, H. J., A Quantitative Uncertainty Metric Controls Error in Neural Network-Driven Chemical Discovery. *Chem. Sci.* **2019**, *10*, 7913-7922.
528. Lunghi, A.; Sanvito, S., Surfing Multiple Conformation-Property Landscapes via Machine Learning: Designing Single-Ion Magnetic Anisotropy. *J. Phys. Chem. C* **2020**, *124*, 5802-5806.
529. Holleis, L.; Shivaram, B. S.; Balachandran, P. V., Machine Learning Guided Design of Single-Molecule Magnets for Magnetocaloric Applications. *Appl. Phys. Lett.* **2019**, *114*, 222404.
530. Bahlke, M. P.; Mogos, N.; Proppe, J.; Herrmann, C., Exchange Spin Coupling from Gaussian Process Regression. *J. Phys. Chem. A* **2020**, *124*, 8708-8723.
531. Duan, C.; Janet, J. P.; Nandy, A.; Liu, F.; Ioannidis, E. I.; Arunachalam, N.; Taylor, M. G.; Harper, D. R.; Kulik, H. J. molSimplify Github. http://github.com/hjkgrp/molsimplify (accessed March 31, 2021).
532. Arunachalam, N.; Taylor, M. G.; Nandy, A.; Duan, C.; Janet, J. P.; Kulik, H. J. molSimplify Lite. https://molsimplify.mit.edu (accessed March 31, 2021).
533. Eckhoff, M.; Lausch, K. N.; Blöchl, P. E.; Behler, J., Predicting Oxidation and Spin States by High-Dimensional Neural Networks: Applications to Lithium Manganese Oxide Spinels. *J. Chem. Phys.* **2020**, *153*, 164107.

534. Marenich, A. V.; Kelly, C. P.; Thompson, J. D.; Hawkins, G. D.; Chambers, C. C.; Giesen, D. J.; Winget, P.; Cramer, C. J.; Truhlar, D. G., Minnesota Solvation Database. *Minnesota Solvation Database version* **2012**, *20*.
535. Hutchinson, S. T.; Kobayashi, R., Solvent-Specific Featurization for Predicting Free Energies of Solvation through Machine Learning. *J. Chem. Inf. Model.* **2019**, *59*, 1338-1346.
536. Lim, H.; Jung, Y., Delfos: Deep Learning Model for Prediction of Solvation Free Energies in Generic Organic Solvents. *Chem. Sci.* **2019**, *10*, 8306-8315.
537. Jinich, A.; Sanchez-Lengeling, B.; Ren, H.; Harman, R.; Aspuru-Guzik, A., A Mixed Quantum Chemistry/Machine Learning Approach for the Fast and Accurate Prediction of Biochemical Redox Potentials and Its Large-Scale Application to 315 000 Redox Reactions. *ACS Cent. Sci.* **2019**, *5*, 1199-1210.
538. Liu, F.; Duan, C.; Kulik, H. J., Rapid Detection of Strong Correlation with Machine Learning for Transition-Metal Complex High-Throughput Screening. *J. Phys. Chem. Lett.* **2020**, *11*, 8067-8076.
539. Friederich, P.; dos Passos Gomes, G.; 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.
540. Meyer, B.; Sawatlon, B.; Heinen, S.; von Lilienfeld, O. A.; Corminboeuf, C., Machine Learning Meets Volcano Plots: Computational Discovery of Cross-Coupling Catalysts. *Chem. Sci.* **2018**, *9*, 7069-7077.
541. Cordova, M.; Wodrich, M. D.; Meyer, B.; Sawatlon, B.; Corminboeuf, C., Data-Driven Advancement of Homogeneous Nickel Catalyst Activity for Aryl Ether Cleavage. *ACS Catal.* **2020**, *10*, 7021-7031.
542. Zahrt, A. F.; Henle, J. J.; Rose, B. T.; Wang, Y.; Darrow, W. T.; Denmark, S. E., Prediction of Higher-Selectivity Catalysts by Computer-Driven Workflow and Machine Learning. *Science* **2019**, *363*, eaau5631.
543. Henle, J. J.; Zahrt, A. F.; Rose, B. T.; Darrow, W. T.; Wang, Y.; Denmark, S. E., Development of a Computer-Guided Workflow for Catalyst Optimization. Descriptor Validation, Subset Selection, and Training Set Analysis. *J. Am. Chem. Soc.* **2020**, *142*, 11578-11592.
544. Yang, W. H.; Fidelis, T. T.; Sun, W. H., Prediction of Catalytic Activities of Bis(Imino)Pyridine Metal Complexes by Machine Learning. *J. Comput. Chem.* **2020**, *41*, 1064-1067.
545. Ahneman, D. T.; Estrada, J. G.; Lin, S.; Dreher, S. D.; Doyle, A. G., Predicting Reaction Performance in C–N Cross-Coupling Using Machine Learning. *Science* **2018**, *360*, 186-190.
546. Nielsen, M. K.; Ahneman, D. T.; Riera, O.; Doyle, A. G., Deoxyfluorination with Sulfonyl Fluorides: Navigating Reaction Space with Machine Learning. *J. Am. Chem. Soc.* **2018**, *140*, 5004-5008.
547. Chuang, K. V.; Keiser, M. J., Comment on ``Predicting Reaction Performance in C--N Cross-Coupling Using Machine Learning" *Science* **2018**, *362*, eaat8603.
548. Estrada, J. G.; Ahneman, D. T.; Sheridan, R. P.; Dreher, S. D.; Doyle, A. G., Response to Comment on ``Predicting Reaction Performance in C–N Cross-Coupling Using Machine Learning" *Science* **2018**, *362*, eaat8763.
549. Maley, Steven M.; Kwon, D.-H.; Rollins, N.; Stanley, J. C.; Sydora, O. L.; Bischof, S. M.; Ess, D. H., Quantum-Mechanical Transition-State Model Combined with Machine Learning Provides Catalyst Design Features for Selective Cr Olefin Oligomerization. *Chem. Sci.* **2020**, *11*, 9665-9674.

550. Rizkin, B. A.; Hartman, R. L., Supervised Machine Learning for Prediction of Zirconocene-Catalyzed Alpha-Olefin Polymerization. *Chem. Eng. Sci.* **2019**, *210*, 115224.
551. Siebert, M.; Krennrich, G.; Seibicke, M.; Siegle, A. F.; Trapp, O., Identifying High-Performance Catalytic Conditions for Carbon Dioxide Reduction to Dimethoxymethane by Multivariate Modelling. *Chem. Sci.* **2019**, *10*, 10466-10474.
552. Meredig, B.; Agrawal, A.; Kirklin, S.; Saal, J. E.; Doak, J. W.; Thompson, A.; Zhang, K.; Choudhary, A.; Wolverton, C., Combinatorial Screening for New Materials in Unconstrained Composition Space with Machine Learning. *Phys. Rev. B* **2014**, *89*, 094104.
553. Isayev, O.; Oses, C.; Toher, C.; Gossett, E.; Curtarolo, S.; Tropsha, A., Universal Fragment Descriptors for Predicting Properties of Inorganic Crystals. *Nat. Commun.* **2017**, *8*, 15679.
554. Xie, T.; Grossman, J. C., Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties. *Phys. Rev. Lett.* **2018**, *120*, 145301.
555. Chen, C.; Ye, W. K.; Zuo, Y. X.; Zheng, C.; Ong, S. P., Graph Networks as a Universal Machine Learning Framework for Molecules and Crystals. *Chem. Mater.* **2019**, *31*, 3564-3572.
556. Wilmer, C. E.; Leaf, M.; Lee, C. Y.; Farha, O. K.; Hauser, B. G.; Hupp, J. T.; Snurr, R. Q., Large-Scale Screening of Hypothetical Metal–Organic Frameworks. *Nat. Chem.* **2012**, *4*, 83.
557. Janet, J. P.; Chan, L.; Kulik, H. J., Accelerating Chemical Discovery with Machine Learning: Simulated Evolution of Spin Crossover Complexes with an Artificial Neural Network. *J. Phys. Chem. Lett.* **2018**, *9*, 1064-1071.
558. Musil, F.; Willatt, M. J.; Langovoy, M. A.; Ceriotti, M., Fast and Accurate Uncertainty Estimation in Chemical Machine Learning. *J. Chem. Theory Comput.* **2019**, *15*, 906-915.
559. Peterson, A. A.; Christensen, R.; Khorshidi, A., Addressing Uncertainty in Atomistic Machine Learning. *Phys. Chem. Chem. Phys.* **2017**, *19*, 10978-10985.
560. Cortés-Ciriano, I.; Bender, A., Deep Confidence: A Computationally Efficient Framework for Calculating Reliable Prediction Errors for Deep Neural Networks. *J. Chem. Inf. Model.* **2019**, *59*, 1269-1281.
561. Gomez-Bombarelli, R.; Wei, J. N.; Duvenaud, D.; Hernandez-Lobato, J. M.; Sanchez-Lengeling, B.; Sheberla, D.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Adams, R. P.; Aspuru-Guzik, A., Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Cent. Sci.* **2018**, *4*, 268-276.
562. Zahrt, A. F.; Rose, B. T.; Darrow, W. T.; Henle, J. J.; Denmark, S. E., Computational Methods for Training Set Selection and Error Assessment Applied to Catalyst Design: Guidelines for Deciding Which Reactions to Run First and Which to Run Next. *React. Chem. Eng.* **2021**, *6*, 694-708.
563. Janet, J. P.; Ramesh, S.; Duan, C.; Kulik, H. J., Accurate Multiobjective Design in a Space of Millions of Transition Metal Complexes with Neural-Network-Driven Efficient Global Optimization. *ACS Cent. Sci.* **2020**, *6*, 513-524.
564. Iovanac, N. C.; Savoie, B. M., Improved Chemical Prediction from Scarce Data Sets via Latent Space Enrichment. *J. Phys. Chem. A* **2019**, *123*, 4295-4302.
565. Duan, C.; Liu, F.; Nandy, A.; Kulik, H. J., Semi-Supervised Machine Learning Enables the Robust Detection of Multireference Character at Low Cost. *J. Phys. Chem. Lett.* **2020**, *11*, 6640-6648.
566. Babor, M.; Gerzon, S.; Raveh, B.; Sobolev, V.; Edelman, M., Prediction of Transition Metal-Binding Sites from Apo Protein Structures. *Proteins: Struct., Funct., Bioinf.* **2008**, *70*, 208-217.

567. Hauser, A.; Enachescu, C.; Daku, M. L.; Vargas, A.; Amstutz, N., Low-Temperature Lifetimes of Metastable High-Spin States in Spin-Crossover and in Low-Spin Iron (II) Compounds: The Rule and Exceptions to the Rule. *Coord. Chem. Rev.* **2006**, *250*, 1642-1652.
568. Taylor, M. G.; Yang, T.; Lin, S.; Nandy, A.; Janet, J. P.; Duan, C.; Kulik, H. J., Seeing Is Believing: Experimental Spin States from Machine Learning Model Structure Predictions. *J. Phys. Chem. A* **2020**, *124*, 3286-3299.
569. Cundari, T. R.; Russo, M., Database Mining Using Soft Computing Techniques. An Integrated Neural Network−Fuzzy Logic−Genetic Algorithm Approach. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 281-287.
570. Fernandez, M.; Boyd, P. G.; Daff, T. D.; Aghaji, M. Z.; Woo, T. K., Rapid and Accurate Machine Learning Recognition of High Performing Metal Organic Frameworks for CO_2 Capture. *J. Phys. Chem. Lett.* **2014**, *5*, 3056-3060.
571. Van Engelen, J. E.; Hoos, H. H., A Survey on Semi-Supervised Learning. *Mach. Learn.* **2020**, *109*, 373-440.
572. Bunn, J. K.; Hu, J.; Hattrick-Simpers, J. R., Semi-Supervised Approach to Phase Identification from Combinatorial Sample Diffraction Patterns. *JOM* **2016**, *68*, 2116-2125.
573. Court, C. J.; Cole, J. M., Auto-Generated Materials Database of Curie and Néel Temperatures via Semi-Supervised Relationship Extraction. *Sci. Data* **2018**, *5*, 180111.
574. Huo, H.; Rong, Z.; Kononova, O.; Sun, W.; Botari, T.; He, T.; Tshitoyan, V.; Ceder, G., Semi-Supervised Machine-Learning Classification of Materials Synthesis Procedures. *npj Comput. Mater.* **2019**, *5*, 1-7.
575. Düren, T.; Sarkisov, L.; Yaghi, O. M.; Snurr, R. Q., Design of New Materials for Methane Storage. *Langmuir* **2004**, *20*, 2683-2689.
576. Boyd, P. G.; Chidambaram, A.; García-Díez, E.; Ireland, C. P.; Daff, T. D.; Bounds, R.; Gładysiak, A.; Schouwink, P.; Moosavi, S. M.; Maroto-Valer, M. M., et al., Data-Driven Design of Metal–Organic Frameworks for Wet Flue Gas CO_2 Capture. *Nature* **2019**, *576*, 253-256.
577. Boyd, P. G.; Woo, T. K., A Generalized Method for Constructing Hypothetical Nanoporous Materials of Any Net Topology from Graph Theory. *CrystEngComm* **2016**, *18*, 3777-3792.
578. Coudert, F.-X.; Fuchs, A. H., Computational Characterization and Prediction of Metal–Organic Framework Properties. *Coord. Chem. Rev.* **2016**, *307*, 211-236.
579. Gómez-Gualdrón, D. A.; Colón, Y. J.; Zhang, X.; Wang, T. C.; Chen, Y.-S.; Hupp, J. T.; Yildirim, T.; Farha, O. K.; Zhang, J.; Snurr, R. Q., Evaluating Topologically Diverse Metal–Organic Frameworks for Cryo-Adsorbed Hydrogen Storage. *Energy Environ. Sci.* **2016**, *9*, 3279-3289.
580. Li, S.; Chung, Y. G.; Simon, C. M.; Snurr, R. Q., High-Throughput Computational Screening of Multivariate Metal–Organic Frameworks (Mtv-Mofs) for CO_2 Capture. *J. Phys. Chem. Lett.* **2017**, *8*, 6135-6141.
581. Martin, R. L.; Haranczyk, M., Construction and Characterization of Structure Models of Crystalline Porous Polymers. *Cryst. Growth Des.* **2014**, *14*, 2431-2440.
582. Colón, Y. J.; Gómez-Gualdrón, D. A.; Snurr, R. Q., Topologically Guided, Automated Construction of Metal–Organic Frameworks and Their Evaluation for Energy-Related Applications. *Cryst. Growth Des.* **2017**, *17*, 5801-5810.
583. Martin, R. L.; Simon, C. M.; Smit, B.; Haranczyk, M., In Silico Design of Porous Polymer Networks: High-Throughput Screening for Methane Storage Materials. *J. Am. Chem. Soc.* **2014**, *136*, 5006-5022.

584. Earl, D. J.; Deem, M. W., Toward a Database of Hypothetical Zeolite Structures. *Ind. Eng. Chem. Res.* **2006**, *45*, 5449-5454.
585. Deem, M. W.; Pophale, R.; Cheeseman, P. A.; Earl, D. J., Computational Discovery of New Zeolite-Like Materials. *J. Phys. Chem. C* **2009**, *113*, 21353-21360.
586. Ruddigkeit, L.; van Deursen, R.; Blum, L. C.; Reymond, J.-L., Enumeration of 166 Billion Organic Small Molecules in the Chemical Universe Database GDB-17. *J. Chem. Inf. Model.* **2012**, *52*, 2864-2875.
587. Smith, J. S.; Isayev, O.; Roitberg, A. E., ANI-1, a Data Set of 20 Million Calculated Off-Equilibrium Conformations for Organic Molecules. *Sci. Data* **2017**, *4*, 170193.
588. Kotrle, K.; Herchel, R., Are Inorganic Single-Molecule Magnets a Possibility? A Theoretical Insight into Dysprosium Double-Deckers with Inorganic Ring Systems. *Inorg. Chem.* **2019**, *58*, 14046-14057.
589. Gugler, S.; Janet, J. P.; Kulik, H. J., Enumeration of De Novo Inorganic Complexes for Chemical Discovery and Machine Learning. *Mol. Syst. Des. Eng.* **2020**, *5*, 139-152.
590. Hageman, J. A.; Westerhuis, J. A.; Frühauf, H.-W.; Rothenberg, G., Design and Assembly of Virtual Homogeneous Catalyst Libraries –Towards in Silico Catalyst Optimisation. *Adv. Synth. Catal.* **2006**, *348*, 361-369.
591. Freeze, J. G.; Kelly, H. R.; Batista, V. S., Search for Catalysts by Inverse Design: Artificial Intelligence, Mountain Climbers, and Alchemists. *Chem. Rev.* **2019**, *119*, 6595-6612.
592. Le, T. C.; Winkler, D. A., Discovery and Optimization of Materials Using Evolutionary Approaches. *Chem. Rev.* **2016**, *116*, 6107-6132.
593. Leardi, R., Genetic Algorithms in Chemometrics and Chemistry: A Review. *J. Chemom.* **2001**, *15*, 559-569.
594. Leardi, R., Genetic Algorithms in Chemistry. *J. Chromatogr. A* **2007**, *1158*, 226-233.
595. Boussie, T. R.; Diamond, G. M.; Goh, C.; Hall, K. A.; LaPointe, A. M.; Leclerc, M.; Lund, C.; Murphy, V.; Shoemaker, J. A. W.; Tracht, U., et al., A Fully Integrated High-Throughput Screening Methodology for the Discovery of New Polyolefin Catalysts: Discovery of a New Class of High Temperature Single-Site Group (IV) Copolymerization Catalysts. *J. Am. Chem. Soc.* **2003**, *125*, 4306-4317.
596. Kreutz, J. E.; Shukhaev, A.; Du, W.; Druskin, S.; Daugulis, O.; Ismagilov, R. F., Evolution of Catalysts Directed by Genetic Algorithms in a Plug-Based Microfluidic Device Tested with Oxidation of Methane by Oxygen. *J. Am. Chem. Soc.* **2010**, *132*, 3128-3132.
597. Kim, K.; Graf, P. A.; Jones, W. B., A Genetic Algorithm Based Inverse Band Structure Method for Semiconductor Alloys. *J. Comput. Phys.* **2005**, *208*, 735-760.
598. Froemming, N. S.; Henkelman, G., Optimizing Core-Shell Nanoparticle Catalysts with a Genetic Algorithm. *J. Chem. Phys.* **2009**, *131*, 234103.
599. Chu, Y.; Heyndrickx, W.; Occhipinti, G.; Jensen, V. R.; Alsberg, B. K., An Evolutionary Algorithm for De Novo Optimization of Functional Transition Metal Compounds. *J. Am. Chem. Soc.* **2012**, *134*, 8885-8895.
600. Foscato, M.; Houghton, B. J.; Occhipinti, G.; Deeth, R. J.; Jensen, V. R., Ring Closure to Form Metal Chelates in 3D Fragment-Based De Novo Design. *J. Chem. Inf. Model.* **2015**, *55*, 1844-1856.
601. Bernhardt, P. V.; Bilyj, J. K.; Brosius, V.; Chernyshov, D.; Deeth, R. J.; Foscato, M.; Jensen, V. R.; Mertes, N.; Riley, M. J.; Törnroos, K. W., Spin Crossover in a Hexaamineiron(II) Complex: Experimental Confirmation of a Computational Prediction. *Chem. Eur. J.* **2018**, *24*, 5082--5085.

602. Bao, Y.; Martin, R. L.; Simon, C. M.; Haranczyk, M.; Smit, B.; Deem, M. W., In Silico Discovery of High Deliverable Capacity Metal–Organic Frameworks. *J. Phys. Chem. C* **2015**, *119*, 186-195.
603. Collins, S. P.; Daff, T. D.; Piotrkowski, S. S.; Woo, T. K., Materials Design by Evolutionary Optimization of Functional Groups in Metal-Organic Frameworks. *Sci. Adv.* **2016**, *2*, e1600954.
604. Chung, Y. G.; Gómez-Gualdrón, D. A.; Li, P.; Leperi, K. T.; Deria, P.; Zhang, H.; Vermeulen, N. A.; Stoddart, J. F.; You, F.; Hupp, J. T., et al., In Silico Discovery of Metal-Organic Frameworks for Precombustion CO_2 Capture Using a Genetic Algorithm. *Sci. Adv.* **2016**, *2*, e1600909.
605. Gustafson, J. A.; Wilmer, C. E., Intelligent Selection of Metal–Organic Framework Arrays for Methane Sensing via Genetic Algorithms. *ACS Sens.* **2019**, *4*, 1586-1593.
606. Andersson, M.; Bligaard, T.; Kustov, A.; Larsen, K.; Greeley, J.; Johannessen, T.; Christensen, C.; Norskov, J., Toward Computational Screening in Heterogeneous Catalysis: Pareto Optimal Methanation Catalysts. *J. Catal.* **2006**, *239*, 501-506.
607. Chen, H.-Z.; Zhang, Y.-Y.; Gong, X.; Xiang, H., Predicting New TiO_2 Phases with Low Band Gaps by a Multiobjective Global Optimization Approach. *J. Phys. Chem. C* **2014**, *118*, 2333-2337.
608. Martin, R. L.; Haranczyk, M., Insights into Multi-Objective Design of Metal–Organic Frameworks. *Cryst. Growth Des.* **2013**, *13*, 4208-4212.
609. Llamas-Galilea, J.; Gobin, O. C.; Schüth, F., Comparison of Single- and Multiobjective Design of Experiment in Combinatorial Chemistry for the Selective Dehydrogenation of Propane. *J. Comb. Chem.* **2009**, *11*, 907-913.
610. Häse, F.; Roch, L. M.; Kreisbeck, C.; Aspuru-Guzik, A., Phoenics: A Bayesian Optimizer for Chemistry. *ACS Cent. Sci.* **2018**, *4*, 1134-1145.
611. Häse, F.; Roch, L. M.; Aspuru-Guzik, A., Chimera: Enabling Hierarchy Based Multi-Objective Optimization for Self-Driving Laboratories. *Chem. Sci.* **2018**, *9*, 7642-7655.
612. Huang, K.; Chen, F.-Q.; Lü, D.-W., Artificial Neural Network-Aided Design of a Multi-Component Catalyst for Methane Oxidative Coupling. *Appl. Catal., A* **2001**, *219*, 61-68.
613. Corma, A.; Serra, J.; Serna, P.; Valero, S.; Argente, E.; Botti, V., Optimisation of Olefin Epoxidation Catalysts with the Application of High-Throughput and Genetic Algorithms Assisted by Artificial Neural Networks (Softcomputing Techniques). *J. Catal.* **2005**, *229*, 513-524.
614. Huang, K.; Zhan, X.-L.; Chen, F.-Q.; Lü, D.-W., Catalyst Design for Methane Oxidative Coupling by Using Artificial Neural Network and Hybrid Genetic Algorithm. *Chem. Eng. Sci.* **2003**, *58*, 81-87.
615. Rodemerck, U.; Baerns, M.; Holena, M.; Wolf, D., Application of a Genetic Algorithm and a Neural Network for the Discovery and Optimization of New Solid Catalytic Materials. *Appl. Surf. Sci.* **2004**, *223*, 168-174.
616. Umegaki, T.; Watanabe, Y.; Nukui, N.; Omata, K.; Yamada, M., Optimization of Catalyst for Methanol Synthesis by a Combinatorial Approach Using a Parallel Activity Test and Genetic Algorithm Assisted by a Neural Network. *Energy Fuels* **2003**, *17*, 850-856.
617. Watanabe, Y.; Umegaki, T.; Hashimoto, M.; Omata, K.; Yamada, M., Optimization of Cu Oxide Catalysts for Methanol Synthesis by Combinatorial Tools Using 96 Well Microplates, Artificial Neural Network and Genetic Algorithm. *Catal. Today* **2004**, *89*, 455-464.

618. Scott, D. J.; Manos, S.; Coveney, P. V., Design of Electroceramic Materials Using Artificial Neural Networks and Multiobjective Evolutionary Algorithms. *J. Chem. Inf. Model.* **2008**, *48*, 262-273.
619. Deb, K.; Agrawal, S.; Pratap, A.; Meyarivan, T. In *A Fast Elitist Non-Dominated Sorting Genetic Algorithm for Multi-Objective Optimization: NSGA-II*, International conference on parallel problem solving from nature, Springer: 2000; pp 849-858.
620. Cuéllar, M. P.; Lapresta-Fernández, A.; Herrera, J. M.; Salinas-Castillo, A.; Pegalajar, M. d. C.; Titos-Padilla, S.; Colacio, E.; Capitán-Vallvey, L. F., Thermochromic Sensor Design Based on Fe(II) Spin Crossover/Polymers Hybrid Materials and Artificial Neural Networks as a Tool in Modelling. *Sens. Actuators, B* **2015**, *208*, 180-187.
621. Jennings, P. C.; Lysgaard, S.; Hummelshøj, J. S.; Vegge, T.; Bligaard, T., Genetic Algorithms for Computational Materials Discovery Accelerated by Machine Learning. *npj Comput. Mater.* **2019**, *5*, 46.
622. del Rosario, Z.; Rupp, M.; Kim, Y.; Antono, E.; Ling, J., Assessing the Frontier: Active Learning, Model Accuracy, and Multi-Objective Candidate Discovery and Optimization. *J. Chem. Phys.* **2020**, *153*, 024112.
623. Gopakumar, A. M.; Balachandran, P. V.; Xue, D.; Gubernatis, J. E.; Lookman, T., Multi-Objective Optimization for Materials Discovery via Adaptive Design. *Sci. Rep.* **2018**, *8*, 3738.
624. Herbol, H. C.; Hu, W.; Frazier, P.; Clancy, P.; Poloczek, M., Efficient Search of Compositional Space for Hybrid Organic–Inorganic Perovskites via Bayesian Optimization. *npj Comput. Mater.* **2018**, *4*, 51.
625. Xue, D.; Balachandran, P. V.; Hogden, J.; Theiler, J.; Xue, D.; Lookman, T., Accelerated Search for Materials with Targeted Properties by Adaptive Design. *Nat. Commun.* **2016**, *7*, 11241.
626. Yuan, R.; Liu, Z.; Balachandran, P. V.; Xue, D.; Zhou, Y.; Ding, X.; Sun, J.; Xue, D.; Lookman, T., Accelerated Discovery of Large Electrostrains in BaTiO3-Based Piezoelectrics Using Active Learning. *Adv. Mater.* **2018**, *30*, 1702884.
627. Schweidtmann, A. M.; Clayton, A. D.; Holmes, N.; Bradford, E.; Bourne, R. A.; Lapkin, A. A., Machine Learning Meets Continuous Flow Chemistry: Automated Optimization Towards the Pareto Front of Multiple Objectives. *Chem. Eng. J.* **2018**, *352*, 277-282.
628. Zhang, Y.; Lee, A. A., Bayesian Semi-Supervised Learning for Uncertainty-Calibrated Prediction of Molecular Properties and Active Learning. *Chem. Sci.* **2019**, *10*, 8154-8163.
629. Forrester, A. I. J.; Keane, A. J., Recent Advances in Surrogate-Based Optimization. *Prog. Aeronaut. Sci.* **2009**, *45*, 50-79.
630. Keane, A. J., Statistical Improvement Criteria for Use in Multiobjective Design Optimization. *AIAA J.* **2006**, *44*, 879-891.
631. Shields, B. J.; Stevens, J.; Li, J.; Parasram, M.; Damani, F.; Alvarado, J. I. M.; Janey, J. M.; Adams, R. P.; Doyle, A. G., Bayesian Reaction Optimization as a Tool for Chemical Synthesis. *Nature* **2021**, *590*, 89-96.
632. Amar, Y.; Schweidtmann, A.; Deutsch, P.; Cao, L. W.; Lapkin, A., Machine Learning and Molecular Descriptors Enable Rational Solvent Selection in Asymmetric Catalysis. *Chem. Sci.* **2019**, *10*, 6697-6706.
633. Montoya, J. H.; Winther, K. T.; Flores, R. A.; Bligaard, T.; Hummelshøj, J. S.; Aykol, M., Autonomous Intelligent Agents for Accelerated Materials Discovery. *Chem. Sci.* **2020**, *11*, 8517-8532.

634. Dan, Y.; Zhao, Y.; Li, X.; Li, S.; Hu, M.; Hu, J., Generative Adversarial Networks (Gan) Based Efficient Sampling of Chemical Composition Space for Inverse Design of Inorganic Materials. *npj Comput. Mater.* **2020**, *6*, 84.
635. Kim, B.; Lee, S.; Kim, J., Inverse Design of Porous Materials Using Artificial Neural Networks. *Sci. Adv.* **2020**, *6*, eaax9324.
636. Yao, Z.; Sánchez-Lengeling, B.; Bobbitt, N. S.; Bucior, B. J.; Kumar, S. G. H.; Collins, S. P.; Burns, T.; Woo, T. K.; Farha, O. K.; Snurr, R. Q., et al., Inverse Design of Nanoporous Crystalline Reticular Materials with Deep Generative Models. *Nat. Mach. Intell.* **2021**, *3*, 76-86.
637. Curtarolo, S.; Hart, G. L. W.; Nardelli, M. B.; Mingo, N.; Sanvito, S.; Levy, O., The High-Throughput Highway to Computational Materials Design. *Nat. Mater.* **2013**, *12*, 191-201.
638. Agrawal, A.; Choudhary, A., Perspective: Materials Informatics and Big Data: Realization of the "Fourth Paradigm" of Science in Materials Science. *APL Mater.* **2016**, *4*, 053208.
639. Gromski, P. S.; Henson, A. B.; Granda, J. M.; Cronin, L., How to Explore Chemical Space Using Algorithms and Automation. *Nat. Rev. Chem.* **2019**, *3*, 119-128.
640. Larsen, A. H.; Mortensen, J. J.; Blomqvist, J.; Castelli, I. E.; Christensen, R.; Dulak, M.; Friis, J.; Groves, M. N.; Hammer, B.; Hargus, C., et al., The Atomic Simulation Environment-a Python Library for Working with Atoms. *J. Phys.: Condens. Matter* **2017**, *29*, 273002.
641. Bahn, S. R.; Jacobsen, K. W., An Object-Oriented Scripting Interface to a Legacy Electronic Structure Code. *Comput. Sci. Eng.* **2002**, *4*, 56-66.
642. Curtarolo, S.; Setyawan, W.; Hart, G. L. W.; Jahnatek, M.; Chepulskii, R. V.; Taylor, R. H.; Wanga, S. D.; Xue, J. K.; Yang, K. S.; Levy, O., et al., AFLOW: An Automatic Framework for High-Throughput Materials Discovery. *Comput. Mater. Sci.* **2012**, *58*, 218-226.
643. Curtarolo, S.; Morgan, D.; Persson, K.; Rodgers, J.; Ceder, G., Predicting Crystal Structures with Data Mining of Quantum Calculations. *Phys. Rev. Lett.* **2003**, *91*, 135503.
644. Ong, S. P.; Richards, W. D.; Jain, A.; Hautier, G.; Kocher, M.; Cholia, S.; Gunter, D.; Chevrier, V. L.; Persson, K. A.; Ceder, G., Python Materials Genomics (Pymatgen): A Robust, Open-Source Python Library for Materials Analysis. *Comput. Mater. Sci.* **2013**, *68*, 314-319.
645. Jain, A.; Ong, S. P.; Chen, W.; Medasani, B.; Qu, X. H.; Kocher, M.; Brafman, M.; Petretto, G.; Rignanese, G. M.; Hautier, G., et al., Fireworks: A Dynamic Workflow System Designed for High-Throughput Applications. *Concurr. Comp.-Pract. E.* **2015**, *27*, 5037-5059.
646. Pizzi, G.; Cepellotti, A.; Sabatini, R.; Marzari, N.; Kozinsky, B., Aiida: Automated Interactive Infrastructure and Database for Computational Science. *Comput. Mater. Sci.* **2016**, *111*, 218-230.
647. O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R., Open Babel: An Open Chemical Toolbox. *J Cheminformatics* **2011**, *3*, 33.
648. Weininger, D., Smiles, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31-36.
649. Bento, A. P.; Gaulton, A.; Hersey, A.; Bellis, L. J.; Chambers, J.; Davies, M.; Krüger, F. A.; Light, Y.; Mak, L.; McGlinchey, S., The ChEMBL Bioactivity Database: An Update. *Nucleic Acids Res.* **2014**, *42*, D1083-D1090.
650. Hay, B. P.; Firman, T. K., Hostdesigner: A Program for the De Novo Structure-Based Design of Molecular Receptors with Binding Sites That Complement Metal Ion Guests. *Inorg. Chem.* **2002**, *41*, 5502-5512.

651. Sobez, J.-G.; Reiher, M., Molassembler: Molecular Graph Construction, Modification, and Conformer Generation for Inorganic and Organic Molecules. *J. Chem. Inf. Model.* **2020**, *60*, 2884-3900.
652. Guan, Y. F.; Ingman, V. M.; Rooks, B. J.; Wheeler, S. E., Aaron: An Automated Reaction Optimizer for New Catalysts. *J. Chem. Theory Comput.* **2018**, *14*, 5249-5261.
653. Guan, Y. F.; Wheeler, S. E., Automated Quantum Mechanical Predictions of Enantioselectivity in a Rhodium-Catalyzed Asymmetric Hydrogenation. *Angew. Chem., Int. Ed.* **2017**, *56*, 9101-9105.
654. Rosales, A. R.; Wahlers, J.; Lime, E.; Meadows, R. E.; Leslie, K. W.; Savin, R.; Bella, F.; Hansen, E.; Helquistl, P.; Munday, R. H., et al., Rapid Virtual Screening of Enantioselective Catalysts Using Catvs. *Nat. Catal.* **2019**, *2*, 41-45.
655. Young, T. A.; Silcock, J.; Sterling, A.; Duarte, F., Autode: Automated Calculation of Reaction Energy Profiles— Application to Organic and Organometallic Reactions. *Angew. Chem., Int. Ed.* **2020**, *60*, 4266-4274.
656. Foscato, M.; Venkatraman, V.; Jensen, V. R., Denoptim: Software for Computational De Novo Design of Organic and Inorganic Molecules. *J. Chem. Inf. Model.* **2019**, *59*, 4077-4082.
657. Dittner, M.; Hartke, B., Globally Optimal Catalytic Fields - Inverse Design of Abstract Embeddings for Maximum Reaction Rate Acceleration. *J. Chem. Theory Comput.* **2018**, *14*, 3547-3564.
658. Heinen, S.; Schwilk, M.; von Rudorff, G. F.; von Lilienfeld, O. A., Machine Learning the Computational Cost of Quantum Chemistry. *Mach. Learn.: Sci. Tech.* **2020**, *1*, 025002.
659. Seritan, S.; Bannwarth, C.; Fales, B. S.; Hohenstein, E. G.; Isborn, C. M.; Kokkila-Schumacher, S. I.; Li, X.; Liu, F.; Luehr, N.; Snyder Jr, J. W., et al., Terachem: A Graphical Processing Unit-Accelerated Electronic Structure Package for Large-Scale Ab Initio Molecular Dynamics. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2021**, *11*, e1494.
660. Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. In *Grad-Cam: Visual Explanations from Deep Networks via Gradient-Based Localization*, Proceedings of the IEEE international conference on computer vision, 2017; pp 618-626.
661. Duan, C.; Liu, F.; Nandy, A.; Kulik, H. J., Putting Density Functional Theory to the Test in Machine-Learning-Accelerated Materials Discovery. *J. Phys. Chem. Lett.* **2021**, *12*, 4628--4637.
662. Mitrofanov, A.; Korolev, V.; Andreadi, N.; Petrov, V.; Kalmykov, S., Simple Automatized Tool for Exchange-Correlation Functional Fitting. *J. Phys. Chem. A* **2020**, *124*, 2700-2707.
663. McAnanama-Brereton, S.; Waller, M. P., Rational Density Functional Selection Using Game Theory. *J. Chem. Inf. Model.* **2018**, *58*, 61-67.
664. Karton, A.; Daon, S.; Martin, J. M. L., W4-11: A High-Confidence Benchmark Dataset for Computational Thermochemistry Derived from First-Principles W4 Data. *Chem. Phys. Lett.* **2011**, *510*, 165-178.
665. Sprague, M. K.; Irikura, K. K., Quantitative Estimation of Uncertainties from Wavefunction Diagnostics. In *Thom H. Dunning, Jr.*, Springer: 2015; pp 307-318.
666. Townsend, J.; Vogiatzis, K. D., Data-Driven Acceleration of the Coupled-Cluster Singles and Doubles Iterative Solver. *J. Phys. Chem. Lett.* **2019**, *10*, 4129-4135.
667. Smith, J. S.; Nebgen, B. T.; Zubatyuk, R.; Lubbers, N.; Devereux, C.; Barros, K.; Tretiak, S.; Isayev, O.; Roitberg, A. E., Approaching Coupled Cluster Accuracy with a General-Purpose Neural Network Potential through Transfer Learning. *Nat. Commun.* **2019**, *10*, 1-8.

668. Coe, J. P., Machine Learning Configuration Interaction for Ab Initio Potential Energy Curves. *J. Chem. Theory Comput.* **2019**, *15*, 6179-6189.
669. Sameera, W. M. C.; Maeda, S.; Morokuma, K., Computational Catalysis Using the Artificial Force Induced Reaction Method. *Acc. Chem. Res.* **2016**, *49*, 763--773.
670. Bergeler, M.; Simm, G. N.; Proppe, J.; Reiher, M., Heuristics-Guided Exploration of Reaction Mechanisms. *J. Chem. Theory Comput.* **2015**, *11*, 5712-5722.
671. Wang, L.-P.; Titov, A.; McGibbon, R.; Liu, F.; Pande, V. S.; Martínez, T. J., Discovering Chemistry with an Ab Initio Nanoreactor. *Nat. Chem.* **2014**, *6*, 1044-1048.
672. Wang, L.-P.; McGibbon, R. T.; Pande, V. S.; Martinez, T. J., Automated Discovery and Refinement of Reactive Molecular Dynamics Pathways. *J. Chem. Theory Comput.* **2016**, *12*, 638-649.
673. Rodriguez, A.; Rodriguez-Fernandez, R.; Vazquez, S. A.; Barnes, G. L.; Stewart, J. J. P.; Martinez-Nunez, E., Tsscds2018: A Code for Automated Discovery of Chemical Reaction Mechanisms and Solving the Kinetics. *J. Comput. Chem.* **2018**, *39*, 1922-1930.
674. Pracht, P.; Bohle, F.; Grimme, S., Automated Exploration of the Low-Energy Chemical Space with Fast Quantum Chemical Methods. *Phys. Chem. Chem. Phys.* **2020**, *22*, 7169-7192.
675. Peterson, A. A., Acceleration of Saddle-Point Searches with Machine Learning. *J. Chem. Phys.* **2016**, *145*, 074106.
676. Denzel, A.; Kastner, J., Gaussian Process Regression for Geometry Optimization. *J. Chem. Phys.* **2018**, *148*, 094114.
677. Denzel, A.; Kastner, J., Gaussian Process Regression for Transition State Search. *J. Chem. Theory Comput.* **2018**, *14*, 5777-5786.
678. Koistinen, O. P.; Asgeirsson, V.; Vehtari, A.; Jonsson, H., Nudged Elastic Band Calculations Accelerated with Gaussian Process Regression Based on Inverse Interatomic Distances. *J. Chem. Theory Comput.* **2019**, *15*, 6738-6751.
679. Meyer, R.; Schmuck, K. S.; Hauser, A. W., Machine Learning in Computational Chemistry: An Evaluation of Method Performance for Nudged Elastic Band Calculations. *J. Chem. Theory Comput.* **2019**, *15*, 6513-6523.
680. Torres, J. A. G.; Jennings, P. C.; Hansen, M. H.; Boes, J. R.; Bligaard, T., Low-Scaling Algorithm for Nudged Elastic Band Calculations Using a Surrogate Machine Learning Model. *Phys. Rev. Lett.* **2019**, *122*, 156001.
681. Carr, S.; Garnett, R.; Lo, C., Accelerating the Search for Global Minima on Potential Energy Surfaces Using Machine Learning. *J. Chem. Phys.* **2016**, *145*, 154106.
682. Simm, G. N.; Reiher, M., Error-Controlled Exploration of Chemical Reaction Networks with Gaussian Processes. *J. Chem. Theory Comput.* **2018**, *14*, 5238-5248.
683. Ulissi, Z. W.; Medford, A. J.; Bligaard, T.; Nørskov, J. K., To Address Surface Reaction Network Complexity Using Scaling Relations Machine Learning and DFT Calculations. *Nat. Commun.* **2017**, *8*, 14621.
684. Bergstra, J.; Yamins, D.; Cox, D. D. In *Hyperopt: A Python Library for Optimizing the Hyperparameters of Machine Learning Algorithms*, Proceedings of the 12th Python in science conference, Citeseer: 2013; p 20.
685. Liu, Y.; Sid-Lakhdar, W. M.; Marques, O.; Zhu, X.; Meng, C.; Demmel, J. W.; Li, X. S. In *Gptune: Multitask Learning for Autotuning Exascale Applications*, Proceedings of the 26th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming, 2021; pp 234-246.

686. Ueno, T.; Rhone, T. D.; Hou, Z.; Mizoguchi, T.; Tsuda, K., Combo: An Efficient Bayesian Optimization Library for Materials Science. *Mater. Discovery* **2016**, *4*, 18-21.
687. Ward, L.; Dunn, A.; Faghaninia, A.; Zimmermann, N. E. R.; Bajaj, S.; Wang, Q.; Montoya, J.; Chen, J. M.; Bystrom, K.; Dylla, M., et al., Matminer: An Open Source Toolkit for Materials Data Mining. *Comput. Mater. Sci.* **2018**, *152*, 60-69.
688. Ward, L.; Agrawal, A.; Choudhary, A.; Wolverton, C., A General-Purpose Machine Learning Framework for Predicting Properties of Inorganic Materials. *npj Comput. Mater.* **2016**, *2*, 16028.
689. Gossett, E.; Toher, C.; Oses, C.; Isayev, O.; Legrain, F.; Rose, F.; Zurek, E.; Carrete, J.; Mingo, N.; Tropsha, A., AFLOW-ML: A RESTful API for Machine-Learning Predictions of Materials Properties. *Comput. Mater. Sci.* **2018**, *152*, 134-145.
690. Wu, Z. Q.; Ramsundar, B.; Feinberg, E. N.; Gomes, J.; Geniesse, C.; Pappu, A. S.; Leswing, K.; Pande, V., Moleculenet: A Benchmark for Molecular Machine Learning. *Chem. Sci.* **2018**, *9*, 513-530.
691. El Khatib, M.; de Jong, W., Ml4chem: A Machine Learning Package for Chemistry and Materials Science. *ChemRxiv* **2020**.
692. Haghighatlari, M.; Vishwakarma, G.; Altarawy, D.; Subramanian, R.; Kota, B. U.; Sonpal, A.; Setlur, S.; Hachmann, J., Chemml: A Machine Learning and Informatics Program Package for the Analysis, Mining, and Modeling of Chemical and Materials Data. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2020**, *10*, e1458.
693. Kausar, S.; Falcao, A. O., An Automated Framework for QSAR Model Building. *J Cheminformatics* **2018**, *10*, 1.
694. Minnich, A. J.; McLoughlin, K.; Tse, M.; Deng, J.; Weber, A.; Murad, N.; Madej, B. D.; Ramsundar, B.; Rush, T.; Calad-Thomson, S., et al., Ampl: A Data-Driven Modeling Pipeline for Drug Discovery. *J. Chem. Inf. Model.* **2020**, *60*, 1955-1968.
695. Obrezanova, O.; Gola, J. M. R.; Champness, E. J.; Segall, M. D., Automatic QSAR Modeling of Adme Properties: Blood-Brain Barrier Penetration and Aqueous Solubility. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 431-440.
696. Lindsey, J. S., A Retrospective on the Automation of Laboratory Synthetic Chemistry. *Chemom. Intell. Lab. Syst.* **1992**, *17*, 15-45.
697. Dimitrov, T.; Kreisbeck, C.; Becker, J. S.; Aspuru-Guzik, A.; Saikin, S. K., Autonomous Molecular Design: Then and Now. *ACS Appl. Mater. Interfaces* **2019**, *11*, 24825-24836.
698. Coley, C. W.; Eyke, N. S.; Jensen, K. F., Autonomous Discovery in the Chemical Sciences Part II: Outlook. *Angew. Chem., Int. Ed.* **2020**, *59*, 23414-23436.
699. Coley, C. W.; Eyke, N. S.; Jensen, K. F., Autonomous Discovery in the Chemical Sciences Part I: Progress. *Angew. Chem., Int. Ed.* **2020**, *59*, 22858-22893.
700. Steiner, S.; Wolf, J.; Glatzel, S.; Andreou, A.; Granda, J. M.; Keenan, G.; Hinkley, T.; Aragon-Camarasa, G.; Kitson, P. J.; Angelone, D., et al., Organic Synthesis in a Modular Robotic System Driven by a Chemical Programming Language. *Science* **2019**, *363*, 144-144.
701. Segler, M. H. S.; Preuss, M.; Waller, M. P., Planning Chemical Syntheses with Deep Neural Networks and Symbolic Ai. *Nature* **2018**, *555*, 604-610.
702. Bedard, A. C.; Adamo, A.; Aroh, K. C.; Russell, M. G.; Bedermann, A. A.; Torosian, J.; Yue, B.; Jensen, K. F.; Jamison, T. F., Reconfigurable System for Automated Optimization of Diverse Chemical Reactions. *Science* **2018**, *361*, 1220-1225.

703. Coley, C. W.; Thomas, D. A.; Lummiss, J. A. M.; Jaworski, J. N.; Breen, C. P.; Schultz, V.; Hart, T.; Fishman, J. S.; Rogers, L.; Gao, H. Y., et al., A Robotic Platform for Flow Synthesis of Organic Compounds Informed by Ai Planning. *Science* **2019**, *365*, eaax1566.
704. Sans, V.; Porwol, L.; Dragone, V.; Cronin, L., A Self Optimizing Synthetic Organic Reactor System Using Real-Time in-Line NMR Spectroscopy. *Chem. Sci.* **2015**, *6*, 1258-1264.
705. Wei, J. N.; Duvenaud, D.; Aspuru-Guzik, A., Neural Networks for the Prediction of Organic Chemistry Reactions. *ACS Cent. Sci.* **2016**, *2*, 725-732.
706. Coley, C. W.; Barzilay, R.; Jaakkola, T. S.; Green, W. H.; Jensen, K. F., Prediction of Organic Reaction Outcomes Using Machine Learning. *ACS Cent. Sci.* **2017**, *3*, 434-443.
707. Zhou, Z. P.; Li, X. C.; Zare, R. N., Optimizing Chemical Reactions with Deep Reinforcement Learning. *ACS Cent. Sci.* **2017**, *3*, 1337-1344.
708. Reizman, B. J.; Jensen, K. F., Feedback in Flow for Accelerated Reaction Development. *Acc. Chem. Res.* **2016**, *49*, 1786-1796.
709. Epps, R. W.; Bowen, M. S.; Volk, A. A.; Abdel-Latif, K.; Han, S. Y.; Reyes, K. G.; Amassian, A.; Abolhasani, M., Artificial Chemist: An Autonomous Quantum Dot Synthesis Bot. *Adv. Mater.* **2020**, *32*, 2001626.
710. Gomez-Bombarelli, R.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Duvenaud, D.; Maclaurin, D.; Blood-Forsythe, M. A.; Chae, H. S.; Einzinger, M.; Ha, D. G.; Wu, T., et al., Design of Efficient Molecular Organic Light-Emitting Diodes by a High-Throughput Virtual Screening and Experimental Approach. *Nat. Mater.* **2016**, *15*, 1120-1127.
711. Kim, E.; Huang, K.; Jegelka, S.; Olivetti, E., Virtual Screening of Inorganic Materials Synthesis Parameters with Deep Learning. *npj Comput. Mater.* **2017**, *3*, 53.
712. Swain, M. C.; Cole, J. M., Chemdataextractor: A Toolkit for Automated Extraction of Chemical Information from the Scientific Literature. *J. Chem. Inf. Model.* **2016**, *56*, 1894-1904.
713. Kononova, O.; He, T.; Huo, H.; Trewartha, A.; Olivetti, E. A.; Ceder, G., Opportunities and Challenges of Text Mining in Aterials Research. *iScience* **2021**, *24*, 102155.
714. Jensen, Z.; Kim, E.; Kwon, S.; Gani, T. Z. H.; Roman-Leshkov, Y.; Moliner, M.; Corma, A.; Olivetti, E., A Machine Learning Approach to Zeolite Synthesis Enabled by Automatic Literature Data Extraction. *ACS Cent. Sci.* **2019**, *5*, 892-899.
715. Tshitoyan, V.; Dagdelen, J.; Weston, L.; Dunn, A.; Rong, Z.; Kononova, O.; Persson, K. A.; Ceder, G.; Jain, A., Unsupervised Word Embeddings Capture Latent Knowledge from Materials Science Literature. *Nature* **2019**, *571*, 95-98.
716. Raccuglia, P.; Elbert, K. C.; Adler, P. D.; Falk, C.; Wenny, M. B.; Mollo, A.; Zeller, M.; Friedler, S. A.; Schrier, J.; Norquist, A. J., Machine-Learning-Assisted Materials Discovery Using Failed Experiments. *Nature* **2016**, *533*, 73-76.
717. Jia, X.; Lynch, A.; Huang, Y.; Danielson, M.; Lang'at, I.; Milder, A.; Ruby, A. E.; Wang, H.; Friedler, S. A.; Norquist, A. J., Anthropogenic Biases in Chemical Reaction Data Hinder Exploratory Inorganic Synthesis. *Nature* **2019**, *573*, 251-255.
718. Cáceres, E. L.; Mew, N. C.; Keiser, M. J., Adding Stochastic Negative Examples into Machine Learning Improves Molecular Bioactivity Prediction. *J. Chem. Inf. Model.* **2020**, *60*, 5957-5970.
719. Moosavi, S. M.; Chidambaram, A.; Talirz, L.; Haranczyk, M.; Stylianou, K. C.; Smit, B., Capturing Chemical Intuition in Synthesis of Metal-Organic Frameworks. *Nat. Commun.* **2019**, *10*, 539.

For Table of Contents Only