Organopalladium Catalysis as a Proving Ground for Data-Rich Approaches to Reaction Development and Quantitative Predictions

Jingru Lu and David C. Leitch

Department of Chemistry, University of Victoria, 3800 Finnerty Rd., Victoria, BC V8P 5C2, Canada. *dcleitch@uvic.ca

KEYWORDS: palladium catalysis; data science; high-throughput experimentation; molecular descriptors; machine learning; multivariate analysis

ABSTRACT. With the advent of high-throughput methods for both computation and experimentation, data-rich approaches to discovering and understanding chemical reactions are becoming ever more central to catalysis research. Organopalladium catalysis is at the forefront of these new approaches, providing a rich proving ground for method development and validation. This critical Perspective discusses a number of recent case studies from academic and industrial laboratories that illustrate how to generate, analyze, and correlate large data sets for quantitative predictions of reactivity and selectivity. Both the power and potential pitfalls of these approaches are discussed, as are the opportunities for both practical predictions and fundamental mechanistic insights.
INTRODUCTION

Organopalladium catalysis, and in particular palladium-catalyzed coupling, continues to be indispensable for fragment coupling in organic synthesis. Its success as a catalytic method for forming all manner of carbon-element bonds is unparalleled, especially in pursuit of complex molecule targets. Pd-catalyzed reactions are (and continue to be) central to the synthesis of myriad natural products, active pharmaceutical ingredients, agrochemicals, and organic materials. While the data was collected nearly 10 years ago, Brown and Boström’s analysis of reaction classes most often used in medicinal chemistry has 3 of the top 20 as Pd-catalyzed: Suzuki-Miyaura, Sonogashira, and Buchwald-Hartwig. No other examples of homogeneous organometallic catalysis are in this set. Certainly, the plethora of excellent research into non-precious metal catalysis for cross-coupling and related transformations means that these systems are emerging as viable alternatives. However, the sheer number and variety of documented examples of Pd catalysis in complex molecule synthesis attests to its enduring importance as a reliable and effective means to access new chemical matter.

As a direct result of its fairly unique combination of wide applicability, highly variable reaction conditions, and massive amounts of published data, organopalladium catalysis is also one of the most frequently studied reaction classes in emerging data-rich methods for reaction discovery, optimization, and prediction. A 2018 survey of practitioners reveals catalytic C–C and C–N coupling as 2 of the top 3 most frequently screened reaction types by high-throughput experimentation (HTE) groups within (or affiliated with) pharmaceutical R&D. And a recent comprehensive review of HTE in organometallic chemistry and catalysis from 2006-2020 reveals cross-coupling (sum of all varieties) as the most frequently reported reaction type undergoing high-throughput screening. This is a stark contrast to the prior 10 years, where HTE
for asymmetric hydrogenation and alkene polymerization were overwhelmingly more common.\(^{16}\) Finally, due to its clear dominance as a preferred method in medicinal chemistry, the patent literature contains millions of unique examples of Pd-catalyzed transformations across a wide swath of chemical space.

Thus, organopalladium catalysis is an ideal proving ground for emerging data-rich approaches to solving chemical synthesis problems. The recent ultrafast growth of computing and data analysis power combined with significant advancements in laboratory automation, miniaturization, and rapid chemical analysis have paved new ways to study and understand chemical systems. In organic synthesis specifically, data-driven methods have made a significant impact on accelerating the process of reaction optimization\(^{17–19}\) and expanding the applicability of multivariate quantitative structure-reactivity/selectivity prediction models for synthesis planning.\(^{20–25}\) A number of literature reviews published in recent years have covered a wide range of applications of the data-rich approaches in the field of organopalladium catalysis.\(^{26–30}\)

Large, accurate, consistent, and comprehensive datasets from reliable sources are the foundation of any data-driven methodology, and often the determining factor for overall success. The data foundation for studies of organopalladium catalysis is built from a combination of three major sources (Figure 1):

1) Previously reported reaction conditions/outcomes obtained directly from the academic/patent literature, and/or from open-source/proprietary reaction databases;

2) Calculated reaction-based parameters, including structurally and/or mechanistically-relevant molecular descriptors and computed reaction energy barriers;
3) “Do-it-yourself” experimentally determined reaction data, ideally collected via automated and/or HTE approaches.

![Diagram showing data sources and their characteristics]

**Figure 1.** Data foundations for building quantitative models in catalyst reactivity and selectivity predictions.

This critical Perspective will cover key aspects of these three approaches to data collection and analysis, and the resulting outcomes. Illustrative examples from organopalladium catalysis show how these approaches are continuing to evolve. There are, of course, advantages and challenges specific to each approach; it is clear from the success stories thus far that a combination of approaches is the ideal strategy. In addition to accelerating catalyst and/or reaction discovery and optimization, taking a data-rich approach to studying chemical synthesis has the potential to significantly expand our understanding of the interplay between molecular structure, experimental conditions, and reaction mechanisms.
DATASETS FROM EXISTING SOURCES

The power (and pitfalls) of training on the literature. Due to its major role in pharmaceutical synthesis for both discovery and manufacturing activities, there is an abundance of reported examples for Pd-catalyzed cross-coupling across an extremely broad range of chemical space. The details of these reactions are spread throughout the academic and (especially) patent literature, and are readily accessible from major chemical databases and platforms. From a data science and statistical modeling perspective, this appears to be an ideal situation: a large, diverse, and available dataset containing “real-world” reaction systems to feed data-hungry but powerful machine learning algorithms.

Unfortunately, as noted by several experts in the field, machine learning predictive models trained exclusively by literature/patent-derived data can suffer from several issues, including low prediction accuracy and lack of generalizability. Reaction outcome data from the literature, despite its abundance, is skewed by the objectives of each individual practitioner. A medicinal chemist making structural analogues during a lead-optimization campaign is not concerned with achieving comprehensive reaction condition coverage or collecting highly accurate yield/rate data; their job is to rapidly prepare target compounds by any means necessary. Likewise, an academic chemist developing new synthetic methods is not interested in reporting a large number of “failed” reactions; their job is to demonstrate the utility and scope of the new method. In other words, the experiments that produced the existing data were not designed or executed with an eye toward predictive statistical modeling, and thus great care must be taken before using these inputs. Glorious and co-workers have nicely addressed the common types of errors and biases in literature-based reaction datasets, which include experimental noise due to reporting errors, selection biases caused by researchers’ preference for specific reagents over
others, and reporting biases toward successful results. A comparison of reported yield frequency between the *Reaxys* database and data obtained from HTE clearly reveals this latter point, with literature-derived data skewed heavily toward higher yields (illustrated in Figure 2 for Buchwald-Hartwig couplings).32

![Bar chart showing frequency of reported yield values for Buchwald-Hartwig coupling reactions taken from the literature (*Reaxys* database) or from an HTE-based dataset. Values obtained from ref. 27.]

Figure 2. Comparison of frequency of reported yield values for Buchwald-Hartwig coupling reactions taken from the literature (*Reaxys* database) or from an HTE-based dataset. Values obtained from ref. 27.

Selection bias is particularly problematic in synthetic chemistry data. This is because synthetic chemists have a high tendency to select specific reagents/catalysts/conditions based on familiarity, past success on similar systems, and ease of implementation. Such practices are very successful in pursuit of new molecules, but gives rise to a highly unbalanced distribution of reaction conditions in the reported datasets. Specifically for organopalladium catalysis, the
reported cross-coupling reaction conditions reveal a strong selection bias toward specific catalysts. Burke, Grzybowski and co-workers highlighted this issue when attempting to generate machine learning based models for Suzuki-Miyaura reactions.\textsuperscript{33} They trained the models using >10,000 literature reported Suzuki couplings with an aim to predict conditions that would lead to success in a hypothetical reaction. Instead of making any chemically-meaningful predictions with respect to catalyst, solvent, or base choice, the models simply captured popularity trends\textsuperscript{35,36} in the literature data. For example, >50\% of reported Suzuki couplings use Pd(PPh\textsubscript{3})\textsubscript{4} as the palladium source, and therefore Pd(PPh\textsubscript{3})\textsubscript{4} was the top recommendation in >80\% of the test cases.

Selection bias also extends to the specific target reactions being studied. Pd-catalyzed coupling reactions benefit from the large number of commercially available substrates: organohalides, boronic acids, amines, etc. However, even with myriad readily accessible and relatively complex substrates, reported chemical space coverage of the resulting products is actually rather limited. Krška, Dreher, and co-workers at Merck revealed a large disconnect in molecular properties between the reported products of Pd-catalyzed cross-couplings and many approved small molecule drugs. The latter class of molecules generally possess more atoms, more H-bond acceptors, lower lipophilicities, and an overall higher degree of molecular complexity. Thus, prediction models trained exclusively on literature data may not be generalizable to late-stage coupling reactions on advanced pharmaceutical intermediates, where additional functional groups and molecular topologies can dramatically affect catalytic performance.

Reporting bias is another prevalent issue in the synthetic chemistry literature. Methodology studies frequently report univariate optimization within a relatively limited range
of conditions, and/or a series of “deviations from standard” conditions as control studies. Furthermore, the highest achievable yields of the target products are the key outcome reported, with little/no discussion of unsuccessful reactions (though this is changing for the better more recently). In addition, other information that is equally important to guide reaction prediction is infrequently disclosed. This includes aspects such as extent of reaction, rate of reaction, mass balance, and specific byproducts observed. For example, formation of homocoupling-derived byproducts is a common concern in Suzuki cross-couplings, but the presence and/or quantity of these byproducts often goes unreported. This reporting bias results in literature-based datasets being skewed and/or incomplete, which is not ideal when applying them as training/test sets for statistical modeling. Glorius and co-workers investigated the impact of insufficient low yield data on prediction accuracy for machine learning models. In a case study on Buchwald-Hartwig reaction modeling, they discovered that the literature dataset was heavily skewed toward high yielding examples (>50% of data points with >70% yield); the resulting ML model performed relatively poorly, with a mean absolute error (MAE) of ±15%. If an HTE-generated dataset was instead used, which has a more realistic distribution of yields (>25% of data points with <10% yield), the resulting model was significantly more accurate (MAE = ±10%). Similarly, the authors simulated the effect of introducing additional experimental data to cover the low yielding reaction space, which again led to improved model performance.

Fitzner, Wuitschik, and co-workers have conducted large meta-analyses on more than 62,000 Buchwald-Hartwig coupling reactions reported in three databases (CAS, Reaxys, and USPTO). Their aims were to identify and address the common issues in using literature data to guide reaction condition optimization, and to generate useful and broadly generalizable predictions. They pointed out that many reactions are reported with very limited reaction
information. For example, only 75% of the C–N couplings analyzed have reaction yield reported. In addition, essential reaction conditions such as temperature, reaction time, reagent scale and catalyst loading are significantly underreported. As already discussed in the context of Glorius’s study, yields that are reported are skewed toward higher values in both the patent and non-patent literature (median yield >60% in recent years).

Importantly, this meta-analysis revealed reporting biases extend beyond (high) yield and (popular) catalyst/ligand to include solvent – nearly 80% of reactions use toluene or dioxane – and base – nearly 80% of reactions use NaOtBu or Cs₂CO₃. Reporting bias can also result from specific target applications becoming more prominent: the authors noted a large increase in use of P(tBu)₃ and diarylamine nucleophiles beginning around 2014, due to an explosion of patent claims around OLED-relevant materials. Thus, while 62,000 reactions may seem like an ideal “big data” set, there are huge swaths of homogeneity in reaction conditions used, as well as overrepresented substrate/product classes based on commercial need. Thus, simply increasing the number of reactions in a dataset provides very limited effect in expanding the chemical/reaction space coverage.
Figure 3. Diversity analysis of reported reaction conditions for C–N couplings from a combination of three databases (CAS, Reaxys, and USPTO). Plots represent % of reported conditions (y-axis) that contain the top N settings (x-axis). For example, ~80% of reactions in the complete database use just the top two solvents (top N = 2). Reproduced with permission from ref. 29. Copyright 2023, American Chemical Society.

These data issues have an obvious impact on model accuracy and applicability. Fitzner and Wuitschik’s initial meta-analysis did result in a set of qualitative predictive tools to aid condition selection for Buchwald-Hartwig couplings based on substrate type.40 Subsequent work from this group on quantitative machine learning models trained on this large dataset revealed how such models can be misleading when applied to new synthetic cases beyond the initial training set.34 Although their model performs well in training/test splits within the literature data, it fails to extend its predictive power to a new reaction dataset obtained by experiment (external predictions). They attribute this lack of generality to all of the aforementioned biases in the literature data. These hidden biases retained in the literature data make it impossible to account for all the important reaction parameters that must be considered and explored when discovering
or optimizing a new reaction. By presenting their failed attempt to achieve a general-purpose predictive model trained solely on literature data, they have identified a potentially general problem with such models: evaluating performance using test data taken from the overall literature data set may lead to overestimation of accuracy/generality. Thus, the authors suggest that it is important to use lab-generated reaction data (e.g. obtained by HTE) for reaction predictive model training and assessment.

Importantly, expert curation can be a powerful approach in leveraging literature data for predictive models. A study by Schleinitz et al. on predicting reaction yield highlights the importance of the quality of the data, rather than the quantity, in assembling a robust training dataset. While this is not specifically organopalladium catalysis, the approach and lessons from this work are highly relevant. Among an initial dataset of >2000 Ni-catalyzed C–O couplings from the literature, the authors carefully selected a small subset of reactions (~200) that best represent the reaction information and chemical space covered by the original dataset. They then trained machine learning models using the large and curated datasets for comparison. The model trained with the small subset achieved the same predictive performance compared to the model trained on the original large dataset. Their study also reveals the importance of including failed experiments and low yield reactions extracted from optimization tables, such as those contained in Supporting Information sections. This will help to overcome reporting biases and lead to reaction datasets with a greater potential for generalization.

**Perspectives on the Available Data.** The above examples highlight the importance of knowing the limitations and biases of literature databases as a source of training data for predictive models. Using expert chemical knowledge in dataset assembly and curation prior to model training is critical to ensure accurate and useful predictions. The success of any statistical
modeling approach to reactivity predictions will rely on the quality and diversity of the training dataset. Even a relatively small dataset can be suitable if it covers a significant chemical/reaction space. The above examples also highlight how synthetic chemists can help improve the applicability of literature data toward predictive modeling. For any new method, evaluating a diverse and balanced chemical space when exploring reaction scope is enormously useful, such as with an informer library. And evaluating/reporting diverse reaction conditions that result in a range of reaction yields is more useful than reporting only the “champion” conditions and yield. Recent initiatives such as the Open Reaction Database seek to standardize and streamline dataset curation to ensure each entry is complete and in machine-readable format, which should greatly improve the quality and accuracy of resulting models.

One final perspective on existing literature datasets is whether the synthetic chemists’ main metric (yield) is the best response variable for predictive modeling. Whether in academic papers or patent examples, the isolated yield is the universal measure of reactivity in organic synthesis. Accordingly, all of the aforementioned models are trained to build connections between molecular structure, reaction conditions, and yield values, with the key predicted outcome being the yield. However, the reliability, consistency, and reproducibility of reported yields data from different literature sources give rise to concerns for its use in predictive model training.

Due to lack of consistency in our reporting practices, a literature yield could refer to several different measurements, such as solution or assay yield by NMR, GC, or HPLC analysis (common during optimization), or isolated yield after workup/purification (for reaction scope exploration). Isolated yields also tend to suffer from reproducibility issues (as any practicing synthetic chemist is painfully aware). Importantly, isolated yields are not only a measure of
reactivity, but the efficiency of the isolation procedure. Thus, there must be systematic errors in isolated yield values due to product loss during workup and purification. This would be especially prevalent in cases where high purity is desired over maximum yield – such as in the pharmaceutical syntheses that populate the USPTO database! Thus, it is critical for us in the organic synthetic community to more frequently report solution yields / assay yields for reaction outcomes in addition to isolated yields, and for reaction databases (CAS, Reaxys) to include and tag these data accordingly, as opposed to only including information from reaction scope studies. Assay yields obtained using quantitative HPLC analysis are standard practice in process chemistry research. More consistent reporting of this measure of reaction outcome would definitely improve the quality of literature data for the purposes of predictive modeling.

Another concern with reaction yield as the major response variable is the small range of values it can take. Yields are bounded between 0-100%, and nearly always reported to the nearest percent. The small product masses (10-100 mg) isolated from many reactions also mean that isolated yields have fairly significant measurement errors. The narrow reporting range and low precision of isolated yields makes it difficult to capture and model significant reactivity differences, especially in lower yielding reactions. For example, the difference between 85% and 95% yield and between 0% and 10% yield are both 10%; however, from a reactivity standpoint the former is barely significant, while the latter is very significant. Other measures of reaction outcome, such as rate or selectivity, can span multiple orders of magnitude and have distinct advantages in reactivity prediction; unfortunately, they are also much less frequently reported.
DATASETS FROM COMPUTATIONAL APPROACHES

To go beyond the existing literature/patent database sources, alternative methods of data generation and collection are clearly required. Computational approaches are a powerful complement to experimental methods in validating mechanistic hypotheses and accelerating reaction optimization in synthetic chemistry.\(^45\) They are also invaluable for dataset generation, especially when limited (or no) experimental data is available, and/or when collecting experimental data would be difficult (or impossible). In organopalladium chemistry and the corresponding synthetic transformations it enables, there are two types of data that are generated using computational approaches: calculated reaction coordinates and corresponding energy profiles; and calculated molecular descriptors that can be linked to reaction outcome data (whether computational or experimental) to build quantitative structure-reactivity/selectivity relationships. These are two distinct applications of computational methods, with the former having closer links to what is typically categorized as computational/theoretical chemistry, and the latter more in the realm of cheminformatics.

**High-Throughput Computation of Reaction Energy Profiles.** In the absence of large experimental data sets of catalyst activity and/or selectivity for a given transformation (or set of transformations), computational studies of catalytic reactions are invaluable as source of data. Comparing relative energies of putative intermediates and/or transition states between different systems gives a molecular-level view of catalyst reactivity that is highly tunable. This is especially powerful where obtaining sufficient experimental data is cost and/or time prohibitive. It also provides a method to interrogate aspects of the reaction mechanism that are impossible to study using purely experimental means. While computational chemistry is often used to study
specific mechanistic questions in palladium catalysis, data-rich studies require high-throughput computational methods to interrogate a large number of potential catalysts and/or substrates.

Corminboeuf and coworkers have successfully applied high-throughput computational tools to construct linear free energy scaling relationships for several catalytic reactions, and represent these using “volcano plots”\(^\text{46,47}\). Such plots are commonly used in heterogeneous catalysis and electrochemistry, and originate from Sabatier’s principle that an ideal catalyst-reactant (or catalyst-product) interaction should be neither too strong nor too weak.\(^\text{48,49}\) By extending this concept to homogeneous catalysis, Corminboeuf’s group is able to gain new insights and make predictions of catalyst performance for transition metal-catalyzed cross-coupling reactions.

They first validated the concept of molecular volcano plots using Suzuki coupling, perhaps the prototypical organopalladium-catalyzed reaction, as a key case study (Figure 5).\(^\text{50}\) For this initial analysis, catalytic behavior was modeled using thermodynamic aspects of the catalytic cycle: namely, the free energies of calculated catalytic intermediates for a variety of transition metals, and the linear relationship between the free energies of those intermediates. This analysis effectively reproduced the experimentally validated fact that Pd-based catalysts are optimal for Suzuki cross-coupling.
Figure 5. Molecular volcano plot of the calculated thermodynamics of oxidative addition (Rxn A), transmetallation (Rxn B) and reductive elimination (Rxn C) for a series of Ni, Pd, Pt, Cu, Ag, and Au catalysts Optimal performance is achieved by Pd-based catalysts, which have intermediate exergonicities of oxidative addition as compared to other transition metals Reproduced with permission from ref. 46. Copyright 2021, American Chemical Society.

In follow-up work, Corminboeuf and co-workers significantly expanded the catalyst pool under investigation to include more than 25,000 potential complexes. Using DFT-obtained thermodynamic data on the free energy of oxidative addition for 7,000 of these as a training set, they were able to use machine learning analysis to predict those values for the remaining 18,000 potential catalysts. Constructing a molecular volcano plot using these thermodynamic parameters and filtering candidate catalysts by cost led to identification of 37 ideal candidates based on Pd or Cu. The authors also exploited this large dataset for further analysis through dimensionality
reduction via clustering analysis, extracting ligand effect trends across multiple cross-coupling reaction classes.\textsuperscript{52} Finally, in addition to studying catalyst effects on reaction performance, they have also applied the volcano plot analysis to studying how substrate structures and substituents affect the outcome.\textsuperscript{53}

The use of calculated transition state energies is another powerful indicator of catalyst performance, though collecting sufficient data can be challenging. A recent example by Yu, Fu, and coworkers showcases this approach to study the activating effect of Brønsted acids on the Pd-catalyzed C–O bond cleavage of allyl alcohols.\textsuperscript{54} They created a computationally derived reaction database containing 393 DFT calculated activation barriers, followed by multivariate linear regression (MLR) model training using calculated molecular descriptors. The prediction accuracy was confirmed by a small experimental test set (taken from the literature\textsuperscript{55}), with good agreement between predicted $\Delta G^\ddagger$ and the experimental product yield.

Currently, accurate and fast mechanistic modeling of organometallic transformations remains challenging. Schoenebeck has highlighted that computational studies of large, complex systems with myriad mechanistic possibilities lead to exponentially growing search directions (and thus exponentially growing computational resources).\textsuperscript{45} An additional challenge is that calculation outcomes often depend on the theoretical method employed,\textsuperscript{56} leading to another layer of complexity in designing high-throughput computational studies. Therefore, amassing computational data for a large array of Pd-catalyzed reactions needs to be performed with caution, balancing both calculation speed with accuracy.

A crucial aspect of any quantitative structure-reactivity dataset is the collection of descriptors used to define molecular structures and reaction conditions. Thanks to tremendous
developments in computer performance and theoretical methods, computational chemistry has made impressive progress in performing fast and reliable calculations for a wide variety of chemical systems. Especially germane to reactivity prediction, computational chemistry is extensively used to create molecular structural representations and corresponding numerical descriptors, and to compute reaction coordinates, transition states, and reaction barriers. It is

Whether by univariate regression or unsupervised machine learning, data-driven prediction models are developed by establishing quantitative relationships between experimentally obtained or computed molecular descriptors and measured reactivity outcome values. The application of powerful computational tools and mechanistic insights has vastly increased the number of relevant physicochemical descriptors to quantify molecular properties. Mechanistically relevant descriptors that directly impact chemical reactivity have found great value in reactivity prediction, such as the computed notations representing the electronic, steric, and vibrational properties of a substrate or catalyst. Activation barriers determined from calculated transition state energies have also contributed to quantitatively accurate reaction outcome predictions. Finally, computational methods that combine statistical algorithms are used to develop automated workflows for tasks ranging from training data selection to error assessment of the output numbers for improved efficiency and reliability in reaction predictions.

**Mapping Catalyst Chemical Space.** A key aspect of organopalladium catalysis where computationally-derived datasets are prevalent is in ligand parameterization. Understanding and predicting the effect that a given ligand will have on reactivity and selectivity is central to Pd-catalyzed coupling. Ligand choice is often the most consequential factor for catalyst performance (both rate and productivity), especially with less reactive but more abundant substrates (e.g.
unactivated Ar–Cl). In addition, ligand identity is known to be a crucial factor to control site selectivity for Pd-catalyzed couplings of multihalogenated substrates. Accordingly, ligand screening from large and diverse ligand libraries is a frequent activity in industrial HTE labs. However, to adequately cover the available chemical space even within a subclass of supporting ligand is time and resource intensive. Computational tools have been invaluable in mapping this chemical space with large sets of relevant numerical descriptors.

Fey and co-workers have pioneered this approach through curation of the Ligand Knowledge Bases, a series of focused ligand descriptor datasets for specific ligand classes relevant to catalysis. These sets include mono and bidentate ligands with P (Figure 4A), C (carbene), N, and O donor atoms, with more than 1,300 entries containing at least 20 descriptors each. This focused approach allows each ligand class to be described by a relevant set of descriptors; for example, descriptors for LKB-PP (bidentate P,P and P,N ligands) were adapted to account for multiple donor atoms and specific steric aspects of chelate bite angle. Likewise, LKB-C (carbene ligands) was designed to take into account differences in electronic configuration between Schrock, Fischer, and NHC type carbene ligands. To highlight the LKB’s relevance to organopalladium catalysis, Fey and co-workers mapped HTE-derived Pd-catalyzed C–N coupling data against the principal component analysis (PCA) map of LKB-P (monodentate P). While a quantitative model was not reported, the data visualization revealed a clear structure-reactivity trend where large, electron-rich phosphines are clearly optimal. In fact, adjacent to the most reactive ligands in the PCA map are several Buchwald-type biaryl ligands (not tested in the experimental data set) that are known to be very effective in this chemistry. Thus, LKB descriptors can potentially help to guide subsequent ligand screening.
Figure 4. A comparison of two monodentate phosphine databases: A) LKB-P; and B) Kraken, with representative phosphines shown. Each database is represented by two-dimensional principal component analysis (PCA) plot. In the LKB-P, PC1 is predominantly electronic in nature, while PC2 is predominantly steric. In the Kraken visualization, PC1-PC4 have the indicated contributors. Figure 4A reproduced with permission from ref. 64, copyright 2010, American Chemical Society. Figure 4B reproduced with permission from ref. 66, copyright 2022, American Chemical Society.
Schoenebeck and co-workers recently reported using the LKB descriptors as a basis for ligand clustering analysis.76 Their group’s interest in dimeric Pd(I) species as precatalysts led them to investigate the ligand properties that stabilize these dimers versus those that fail to generate stable dimers. The existing knowledge of the structure features promoting Pd(I) dimer formation is extremely limited, as there are only four known dimer-inducing ligands from the LKB-P library. Furthermore, these ligands are proximal to non-dimer forming ligands (e.g. PCy3) in the initial principal component analysis (PCA) plot of the LKB-P library. Thus, simple visual clustering of ligands with similar properties fails to reveal the key features for Pd(I) dimer formation. An alternative means of grouping ligands based on their descriptors – k-means clustering (k = 8) – led to dimer-forming and non-dimer forming ligands being grouped in different clusters. Strikingly, these clusters overlap in the PCA plot, showing that the k-means algorithm is capable of distinguishing apparently similar ligands; however, the sizes of the two candidate clusters were too large to be useful in predicting dimer formation for a given ligand.

To refine the ligand classification analysis, the authors introduced a group of new, “problem-specific” descriptors obtained by DFT calculations. These include some of the original LKB-P descriptors, but also many steric, electronic, and thermodynamic descriptors specific to Pd-based intermediates along the mechanism for Pd(I) dimer formation. Further clustering of the two initial candidate clusters using k-means (k = 6) and this new set of descriptors led to clear differentiation and identification of two smaller clusters that contained all of the known dimer-forming ligands. Importantly, eight predicted dimer-inducing phosphine ligands belonging to the same clusters were verified experimentally to support Pd(I) dimer formation, leading to the isolation of eight previously unknown Pd(I) dimers. This example shows how effective data analysis and descriptor generation is just as important as efficient collection and collation of...
computational data. While large descriptor databases like LKB provide an excellent starting point, problem-specific descriptor sets are often critical for more granular analyses and predictions.

More recently, Gensch, Sigman, Aspuru-Guzik and co-workers created an online descriptor database for monodentate phosphines (Figure 4B). This database significantly expands the chemical and descriptor space applied to this important class of ligand, and incorporates conformational effects. For an initial set of 1,558 ligands (including the 200 most commonly cited phosphines in the literature), the authors calculated 190 individual descriptors using DFT methods. From this set, they then were able to train machine learning models to predict properties for a further 300,000 potential phosphate ligands. Thus, combining DFT and ML methods is a powerful approach to rapidly generate “big data” for molecular descriptor sets.

In a follow-up study, Gensch, Sigman, and co-workers applied this phosphine descriptor database to guide the development of a diverse monodentate phosphine screening set to maximize chemical space coverage. Unlike continuous reaction variables, such as temperature and reaction time, the electronic and steric properties of the PR₃ ligands are generally treated as categorical/discrete variables in experimental screening. In a typical HTE-based optimization approach, researchers will design an array using a combination of literature precedent and chemical intuition. They will then determine which ligands to test next by analogy/perceived similarity to the most successful candidates in an iterative approach. Inevitably, this leads to selection biases when the search space for the discrete parameters is not well defined.

To explore the discrete ligand space systematically, and reduce these potential biases, the authors sampled all of the commercially available PR₃ ligands from the virtual library and
generated four principal components space from the descriptor set. Using a computational data-driven workflow, they then winnowed down to a diverse set of 32 monophosphines intended to maximally and uniformly cover the 4D chemical space. Application of this screening set to several Pd-catalyzed C–C and C–N coupling reactions revealed the expected spread of catalytic results, with a clear positive hit (or several) observed in each case. Furthermore, given the logical design of the screening set, follow-up investigations can focus on ligands in close PC space proximity to “hit” ligands, reminiscent of the LKB-P example described previously.

The application of cheminformatics tools to organopalladium chemistry can positively impact one of the biggest challenges in using Pd-catalysis in target-based synthesis: how does one pick the right catalyst? Practitioners, especially in pharmaceutical R&D, are moving steadily away from empirically-based screening designs into those resembling the deliberate chemical space coverage described above. Not only does this increase the likelihood of success in identifying a suitable catalyst in a given amount of time – critical in discovery chemistry – it also enables cheminformatics to combine with high-throughput experimentation to generate large, consistent, and quantitative datasets as a foundation for quantitative predictions. Finally, it also makes hypothesis-driven screening much more exact, since specific ligand features can be evaluated quantitatively rather than (at worst) superficially (e.g. investigating “similar” ligands that are not actually that similar).
DATASETS FROM HIGH-THROUGHPUT EXPERIMENTATION

Big data on demand via miniaturized array-based experiments. High throughput experimentation (HTE) is an increasingly common approach in both academic and industrial labs. HTE is especially well-suited to studying catalytic reactions, where the number of factors affecting reaction outcome are generally greater than in non-catalytic systems. While HTE is often used with the aim of either discovery (does this transformation work?) or optimization (what is the best set of conditions to maximize yield/selectivity?), it is also an invaluable approach to generating large internally-consistent datasets on demand. This Perspective will focus only on recent, illustrative applications of HTE in organopalladium chemistry; readers interested in HTE more generally are directed to several recent reviews on the topic.14,15,79–85

As HTE enables a data-rich approach to the discovery of chemical reactions – many more factors and their interactions can be studied simultaneously – it is particularly well-suited to the study of Pd-catalyzed processes. A typical cross-coupling reaction system includes numerous categorical factors (Pd source, ligand, base, solvent) and continuous factors (time, temperature, concentration, stoichiometries) that are impossible to fully assess “one factor at a time” (OFAT). Led by industrial R&D groups, particularly in pharmaceutical discovery and process chemistry, as well as academic HTE centers, data-rich HTE approaches to multifactor exploration and optimization have been demonstrated on many Pd-catalyzed reactions.15 In addition to the direct applicability of these studies to active pharmaceutical ingredient (API) synthesis, organopalladium catalysis is again a general proving ground for data-rich techniques.

Developing new data-rich capabilities often goes hand-in-hand with applications in organopalladium chemistry. One set of examples is the development of new methods for high-
throughput reaction set-up and analysis for dense reaction arrays. Researchers at Merck used a combination of advanced liquid handling automation with rapid UPLC-MS or MALDI-TOF MS analysis to interrogate thousands of Pd-catalyzed C–N coupling reactions in plate-based formats.86,87 Researchers at Pfizer similarly demonstrated a new flow-based HTE platform using Pd-catalyzed Suzuki-Miyaura couplings to run and analyze thousands of examples.88 In academia, the Cernak group is continuing to push the envelope of ultra-high throughput methods in array design and execution.89–91 In the area of laboratory automation, the Hein group, in collaboration with Merck and the Aspuru-Guzik and Sigman groups, demonstrated the power of data-dense experimentation in an autonomous optimization of a Pd-catalyzed Suzuki-Miyaura reaction.18 The Newman group published a user-centered “how-to-HTE” guide on multiwell screening, using a Pd-catalyzed C–N coupling as the prototype reaction.92 And finally, even undergraduate teaching lab experiments on microscale HTE employ Pd-catalyzed reactions as the exemplars.93

Our research group focuses not only using HTE and data-rich techniques to develop new reactions,94–98 but also to better understand how reactivity and mechanism change as a function of the reaction system.58,99,100 To enable these efforts in the realm of organopalladium chemistry, we identified a need for Pd precursor compounds that are specifically suited to HTE studies. In 2021 we reported an easily prepared and air-stable Pd(0) precursor, DMPDB−Pd−MAH, which was designed specifically for HTE applications.101

While HTE has been widely used to conduct catalyst screening exercises as well as multifactor optimization studies, the identities of the substrates themselves are increasingly being explored via HTE. Sather and Martinot used a data-rich approach to find the best reaction conditions for fusing piperidine-based nucleophiles with five-membered heteroaromatic
bromides, a difficult C–N coupling type that has been rarely studied before. Extensive HTE revealed Pd-PEPPSI IHeptCl – an advanced precatalyst from the Organ group as the best catalyst in their model pyrazole-based system, as well as specific solvent/base combinations that prevent substrate decomposition. They subsequently tested the generality of this catalyst system in a panel of 48 unique heteroaryl bromides, with a “hit-rate” of ~50% (products observable by LCMS). This evaluation revealed several important features of the reaction, including incompatibility with certain ester functional groups, as well as protic functional groups including alcohols and amides.

In 2019 we, along with collaborators at GSK and Temple University, conducted a similar study of a challenging C–N coupling involving sulfonamide arylation. Multi-substrate HTE revealed the BippyPhos class of ligands as superior, with AdBippyPhos as the optimal candidate. To evaluate the generality of this reaction, we tested an array of 288 substrate combinations: 24 heteroaryl halides and 12 primary and secondary sulfonamides. In this case, our “hit-rate” was ~25% (where coupled products were observed by LCMS in >10% area). Several substrates were ineffective when paired with any coupling partner, while others were effective with nearly all coupling partners. These kinds of substrate-focused HTE studies not only provide a map of reaction scope and highlight gaps in applicability, they are crucial as training sets for statistical analysis and reaction prediction.

Combining high-throughput experimental and cheminformatics approaches. HTE screening in chemistry is often aimed at achieving the ideal solution to a synthetic task. Beyond the optimization goal, the ability to generate hundreds-to-thousands of data points in a deliberate way means HTE is ideally suited as a means to build training datasets for reaction prediction. Multivariate HTE designs create reaction data libraries that cover a broader range of chemical
structure and reaction space, and have more representative outcomes than literature data (vide supra, Figure 2). Such high-density coverage of reaction space is ideal when building quantitative statistical models for reactivity and/or selectivity, providing the experimental link to large descriptor datasets that are increasingly accessible (vide supra, Figure 4).

Catalytic reactions are known to be acutely sensitive to seemingly minor changes, where subtle alterations to catalyst structure, solvent identity, or even reaction concentration can have a dramatic impact on reaction outcomes. Understanding the underlying connections between reaction conditions, molecular structure, and chemical reactivity is both a fundamental goal of catalysis research, and is crucial to making accurate predictions. Because of the flexibility in experimental design and efficiency in data collection, HTE can be used to study how specific aspects of a reaction system impact reactivity, providing sufficient experimental data to train predictive models. Notably, the following three illustrative examples are all academic/industry collaborations!

A seminal example of combining high-throughput experimentation and computation for predictive modeling is the work of Dreher, Doyle, and co-workers on Pd-catalyzed C–N coupling.\textsuperscript{104} Their study was focused on the effect of a specific type of inhibitory additive, isoxazole heterocycles, on the performance of various C–N couplings. Using Merck’s nanoscale high-throughput platform, they collected a multidimensional experimental dataset of \(~4000\) reactions by combining 23 isoxazoles with different aryl halides, solvents and bases. From the computational side, a large set of molecular descriptors was generated without biasing selection with (possibly incorrect) mechanistic hypotheses. A variety of modeling approaches were evaluated, from multivariate linear regression (MLR) to machine learning (ML) algorithms, with a random forest algorithm providing the most accurate model. Even though ML-based models
can be difficult to interpret mechanistically, the authors used a sensitivity analysis for key descriptors (i.e. how much is the model error increased by randomizing a given descriptor) to derive mechanistic insight. The two most sensitive descriptors are isoxazole-based, and are linked to electrophilicity (calculated $^{13}$C chemical shift of the C$_3$ site, and LUMO energy), suggesting the most potent isoxazole inhibitors undergo competitive oxidative addition to Pd(0). This was verified experimentally by spectroscopic observation of the competitive N–O oxidative addition to Pd(PPh$_3$)$_4$. Notably, this work did generate commentary about the use of different feature sets,\textsuperscript{105,106} whether chemical-based (as in Dreher and Doyle’s approach\textsuperscript{104}) or random-valued (as discussed by Chuang and Keiser\textsuperscript{105}). While the chemical insights and out-of-sample prediction accuracy of the initial random forest model validate the use of chemical-based descriptors in this case,\textsuperscript{106} incorporating control procedures and best practices into ML data analysis is crucial for those looking to use this powerful technique across the chemical sciences.\textsuperscript{107}

In another academic/industry collaboration, Sigman, Doyle, and researchers at Merck reported a quantitative workflow to identify and rationalize reactivity cliffs in Ni- and Pd-catalyzed Suzuki-Miyaura coupling (SMC) datasets using monodentate phosphine ligands.\textsuperscript{109} Using the aforementioned Kraken organophosphorus descriptor library and SMC reaction datasets collected by high-throughput experimentation, they developed a quantitative workflow to evaluate the performance of a single ligand descriptor in identifying reactivity thresholds. Among the 3 steric ligand descriptors evaluated (cone angle, $\%V_{bur}$ (Boltz) and $\%V_{bur}$ (min)), $\%V_{bur}$ (min) offers the best performance in dividing the SMC datasets into active and inactive regions around a threshold value. The success of this univariate classification is based on the ligation state (formation of ML versus ML$_2$ species), which is intimately tied to the size of the
phosphine ligand. In addition, this work highlighted an intriguing mechanistic dichotomy: in the examined Ni- and Pd-catalyzed SMC datasets, the opposite $%V_{bur}$ (min) reactivity threshold directionalities indicate that the Ni-catalyzed SMCs require bisligated (NiL$_2$) species while Pd-catalyzed SMCs require monoligated (PdL) species. This study highlights the power of a conceptually simple univariate threshold approach to reactivity analysis and prediction.

Making predictions for multiple outcomes – such as high activity and selectivity – poses additional challenges for reaction modelling. This is particularly the case for regio- and/or stereoselective catalysis, where both high chemical yield and high selectivities are requirements for a successful process; thus, catalyst/ligand optimization requires a multi-objective approach. Mack, Sigman, and co-workers reported a data-driven workflow using high-throughput computation and experimentation for multi-objective ligand optimization. This approach was demonstrated for two key enantioselective steps toward an investigational API from Genentech. These steps are a regio- and enantioselective Pd-catalyzed Hayashi-Heck coupling, following by regio- and enantioselective Rh-catalyzed hydroformylation; the discussion here will focus on the first step (Figure 6A).

In this case, the data foundations for this study are a computational database consisting of >550 bisphosphine ligands with DFT derived descriptors, and an experimental HTE screening database using a group of selected ligands based on previous research findings. These then informed the multi-objective analysis (Figure 6B). The ligands screened were classified based on a reactivity threshold analysis akin to that described above, where the phosphorus lone pair occupancy was an effective single descriptor. This revealed that phosphines below a threshold value were generally active, enabling a more focused ligand set to be taken forward. Then, regioselectivity as a function of ligand structure was assessed by MLR, revealing only two
descriptors – anisotropic $^{31}$P NMR shielding value and P–C $\sigma^*$ orbital occupancy – were sufficient to build a robust correlation. Finally, a virtual screen of the entire bisphosphine database identified ligands predicted to be even more regioselective than those in the HTE dataset, leading to experimental validation of (S)-HexaMeO-BIPHEP as optimal from a combined activity, regioselectivity, and enantioselectivity standpoint (Figure 6C).

**Figure 6.** A) Hayashi-Heck coupling required in high yield, enantioselectivity, and regioselectivity. B) Step-wise multi-objective optimization via a data-rich computational / HTE approach, with univariate threshold analysis for activity, followed by MLR analysis for regioselectivity, and finally virtual screening for extrapolation to superior catalyst. C) Improved outcome with new catalyst system. Adapted with permission from ref. 108. Copyright 2023, American Chemical Society.
In another example from the Sigman/Genentech collaboration, Xu et al. reported the use of MLR modeling to optimize an atroposelective Pd-catalyzed Negishi reaction using chiral bisphosphine ligands. This step is key for the synthesis of a potent KRAS G12C covalent inhibitor, GDC-6036 (1).\textsuperscript{110} Using Genentech’s HTE capabilities, the authors rapidly collected a focused set of data for a variety of chiral ligands, with only 3 out of 24 giving yields >20%, and none with e.r. >70:30. One Walphos-type ligand was identified as superior in this initial set, leading to a more focused HTE campaign to generate a Walphos training set for MLR analysis. The resulting linear model facilitated virtual ligand screening and experimental verification, akin to that described above, leading to high stereoselectivity for the desired Negishi reaction using W057-2 as the optimal ligand.

Notably in this case, the excellent selectivity exhibited by W057-2 is not transferable to other coupling partners in model Negishi reactions. Even small structural changes to the substrates result in significantly diminished yield and selectivity. This lack of generality is likely due to the focused HTE-based training set, where substrate variations were not taken into account. This is obviously not an issue for the Genentech researchers, who are primarily concerned with synthesis of GDC-6036 (1) rather than simpler biaryls; however, it does reveal the broader challenges and complexities in discovering general catalytic systems, where substantially more data is required to account for interactions between substrate and catalyst.

**Quantitative structure-rate relationships (QSRRs).** Overall, combining HTE with MLR/ML approaches to link experimental outcomes to calculated descriptors is an excellent approach to building robust, accurate, and potentially generalizable predictive models. However, as discussed previously, using reaction yields as the outcome variable is not optimal for quantitative modeling. Kinetic parameters (rates, rate constants, $\Delta G^r$ values) are superior to
yields as an outcome variable, especially for mechanism-based approaches. Many of the best-performing models – such as the two Sigman/Genentech examples above – use selectivity measurements as the outcome metric. If under kinetic control, selectivity represents a ratio of reaction rates for the formation of one product or the other, and therefore can be used to construct linear free energy relationships (LFERs). This is why selectivity values are often represented as $\Delta\Delta G^\ddagger$ – i.e. free energies – in the resulting models.

A classic example of LFERs in physical organic chemistry is the Hammett equation,\textsuperscript{111} which has been widely used to study mechanistic aspects of organic and organometallic reactions.\textsuperscript{112–114} A relevant example in organopalladium chemistry was reported by Maes, Jutand, and co-workers, who studied the oxidative addition of 2-halopyridines to Pd(PPh\textsubscript{3})\textsubscript{4} using Hammett analyses of experimentally measured rates.\textsuperscript{115} The corresponding Hammett plots and DFT calculations of transition states reveals that the oxidative addition mechanism for 2-halopyridines changes depending on the identity of the halide. A polarized nucleophilic displacement mechanism was proposed for 2-chloro and 2-bromopyridines, whereas a classic 3-centered mechanism was proposed for 2-iodopyridines.

Kinetic values can outperform yields in providing chemically-meaningful predictions because the major objective that most models are trained to predict – whether explicitly or implicitly – is kinetic in nature. Higher reaction yields are often correlated to favorable reaction kinetics. However, unlike selectivity or yield measurements – which are easily performed in HTE-type formats – determining large numbers of kinetic parameters \textit{via} rate analysis is much more challenging. Rate data is also not as widely reported in the literature, limiting the usefulness of reaction database mining. A recent successful example of using literature rate data to train an ML QSRR model was reported by Jorner, Brinck, Norrby, and Buttar (yet another
Using a combination of >400 experimental S<sub>N</sub>Ar rate constants from literature, the authors trained an ML model demonstrating excellent prediction accuracy across a wide range of reported S<sub>N</sub>Ar reactions. While this study focused on S<sub>N</sub>Ar chemistry and not organopalladium chemistry, their approach is likely generally applicable (e.g. to organopalladium reactions) provided sufficient data is available; however, as described above, rate data is much less reported in the literature, and therefore must be collected experimentally.

Traditional approaches to obtaining rate constants via reaction progress analysis, even using initial rates, is extremely time and resource intensive, and incompatible with high-throughput experimental approaches. To efficiently assemble rate-based datasets as a foundation for predictive modeling, alternative techniques are therefore required. The concept of “one-pot multi-substrate screening” was introduced by Kagan for its application in fast optimization of asymmetric catalysts for various enantioselective reactions. In 2007-2008, Plenio and co-workers extended this concept to realize high-throughput kinetic analysis of Pd-catalyzed Sonogashira couplings. The authors of these two studies were able to simultaneously collect initial rates for up to 25 substrates in a single reaction flask. This significantly reduces the number of measurements needed to generate large kinetic datasets. Their QSRR studies combined 29 aryl bromide substrates with 17 monophosphine ligands for a total of 410 individual reactions. This pioneering study demonstrates the power of creative experimental design in high-throughput experimentation, where more individual experiments is not necessarily required to amass sufficient data. While the quantitative analysis provided by the authors does not attempt to create a unified multivariate model, this dataset was invaluable to our group as an external test set for catalytic predictions (vide infra).
Plenio and co-workers do note some limitations of the one-pot, multi-substrate kinetic analysis approach. One challenge is measuring initial rates for very fast reactions, due to the need for offline analysis by (in their case) GC. Another challenge specific to this method is to accurately measure rates for the slowest reactions in a multi-substrate set, as the effective catalyst concentration per substrate will increase once the faster reactions finish. Thus, care must be taken to validate the results from these multi-substrate studies. Nevertheless, this approach is likely applicable to many reaction types, and should prove valuable for high-throughput kinetic analysis.

Our group has also considered the challenges of collecting large kinetic parameter datasets, and have taken a different approach to reducing the experimental burden. Rather than collect rate data from multiple measurements of concentration over time, we opted to measure relative rates via competition experiments. This effectively converts rate measurements to selectivity measurements. Using this approach, we recently reported two QSRR models for accurate prediction of S_NAr^99 and Pd-catalyzed cross-coupling reactivity. By correlating calculated molecular descriptors to relative free energies of activation (\(\Delta\Delta G^\ddagger\)), we obtained MLR models that have excellent performance to predict rate and selectivity for many external datasets. A discussion of the Pd-based work is illustrative of our approach.

Rather than study a specific class of Pd-catalyzed reaction, we focused our QSRR study on a key step in the catalytic mechanism: oxidative addition. This fundamental transformation is common to myriad catalytic reactions, and is often turnover and/or selectivity determining in Pd-catalyzed transformations (Figure 7A). We therefore hypothesized that a QSRR model for Ar–X oxidative addition to Pd(0) would be applicable to many different Pd-catalyzed cross-coupling reactions. To collect the required rate data, we assembled a diverse set of (hetero)aryl halides

https://doi.org/10.26434/chemrxiv-2023-2078b-v2 ORCID: https://orcid.org/0000-0002-8726-3318 Content not peer-reviewed by ChemRxiv. License: CC BY-NC-ND 4.0
with a variety of substitution patterns, and ensuring heterocyclic substrates were well-represented. Using a model Pd(0) complex – Pd(PCy₃)₂ – we then performed a series of competition experiments under pseudo first-order conditions by having two Ar–X electrophiles in excess but equal amount compete for oxidative addition (Figure 7B). The product ratio, and therefore relative rate, is easily measured by quantitative ³¹P NMR spectroscopy, and control experiments confirmed that the product ratio is kinetically controlled. This technique enables rapid assembly of a reactivity scale for substrates that spans many orders of magnitude in rate (Figure 7C).

**Figure 7.** A) Generic catalytic cycle for Pd-catalyzed cross-coupling, highlighting the importance of oxidative addition. B) Experimental approach to rapid relative rate collection *via* competition experiments, with analysis of L₂Pd(Ar)(X) complexes by ³¹P qNMR spectroscopy. C) Reactivity scale for representative substrates. Figure adapted with permission from ref. 47. Copyright 2022, Royal Society of Chemistry.
During feature selection for our QSRR models, we quickly determined that the average molecular electrostatic potential (ESP) is a particularly useful electronic descriptor for this reaction. These are the average ESP energy across the surface area of an atom in the molecule, and are easily calculated using the wavefunction analysis application Multiwfn. Molecular ESP has been previously used to interpret the chemical properties and reactivity of various molecular systems. Germane to this Perspective, Suresh and Koga correlated minimum molecular ESP values at P to a variety of phosphine molecular properties, demonstrating ESP is an excellent descriptor for the electronic nature of phosphine ligands. Subsequent work from Anjali and Suresh correlated molecular ESP at Pd to computationally-determined activation barriers for oxidative addition of various Ph–X substrates to L–Pd(0) fragments.

In our oxidative addition QSRR model, five descriptors are sufficient to describe the reactivity for Ar–Cl, Ar–Br, and Ar–OTf substrates (Figure 8). The most significant contributors are the average molecular ESP at the reactive carbon (ESP₁) and at the adjacent atom (C or N, ESP₂); combined, these values contribute >60% to the model output. These local descriptors not only indicate the extent of electron-deficiency, but also the polarization of the C=C or C=N bond. Steric effects were initially accounted for using simple tabulated A values, and work is underway to incorporate calculated steric descriptors. Finally, the identity of the leaving group is linked to two additional descriptors: the C–X bond strength, given by the intrinsic bond strength index (IBSI), and the pKₐ of the leaving group conjugate acid. Both of these descriptors are necessary to unify the three classes of electrophile, and subsequent work indicates this set of features enables incorporating Ar–I substrates as a fourth substrate class.
To test our hypothesis that a QSRR model for oxidative addition would be applicable to multiple classes of catalytic reactions, we used predicted $\Delta G^{\ddag}_{OA}$ values from the model to make predictions about site selectivity in multihalogenated electrophiles. There is a plethora of literature available on cross-coupling site selectivity, and our analysis indicated that the MLR model in Figure 8 is able to predict “conventional” site selectivity for a wide range of substrates in both Suzuki and Buchwald-Hartwig reactions. This is the expected site selectivity when using “simple” catalyst systems, such as those based on PPh$_3$, dppf, or other common phosphines. Given we used Pd(PCy$_3$)$_2$ as the Pd(0) source to generate our model, this consistency with simple phosphines makes sense. Neufeldt’s group have assessed factors that lead to unconventional site
selectivity in Pd-catalyzed coupling, many of which are related to ligand and/or solvent identity, and their impacts on Pd speciation.\textsuperscript{70,71,129–133} Future work in QSRR modeling for oxidative addition must therefore endeavor to capture these ligand/solvent/speciation effects.

Finally, we tested our oxidative addition QSRR model for predicting external catalytic rate data. The aforementioned Sonogashira dataset from Plenio and co-workers was an ideal and comprehensive case study, especially since the data could be cleanly separated into two distinct sets (Figure 9). In one paper, the authors focused on electronic effects (\textit{meta} and \textit{para} substitution), with 20 different Ar–Br substrates (Substrate set #1).\textsuperscript{119} In another report, only steric effects were studied (\textit{ortho} substitution), with 9 additional Ar–Br substrates (Substrate set #2, which also contains Ph–Br from Substrate set #1).\textsuperscript{118} Importantly, most of the substrates from both sets are not included in our $\Delta G^\ddagger_{OA}$ training set. The inclusion of 17 different monophosphine ligands also provided an opportunity to test our model’s applicability beyond PCy$_3$-ligated catalysts.

We exclusively used Substrate set #1 as our training and test set to build an initial MLR model. For each substrate, we calculated the predicted $\Delta G^\ddagger_{OA}$ value from the equation in Figure 8; effectively, this is a pre-weighted “super-descriptor” that takes into account the steric and electronic effects of the electrophile. For the phosphine ligands, we used two descriptors – average molecular ESP at P ($\text{ESP}_P$) and \% buried volume ($\% V_{\text{bur}}$)\textsuperscript{59} – to account for electronic and steric effects, respectively. These three descriptors – $\Delta G^\ddagger_{OA}$, $\text{ESP}_P$, and $\% V_{\text{bur}}$ – lead to an accurate linear model for $\ln k$ of the Sonogashira reactions in Substrate set #1, with statistics given in Figure 9. Notably, $\text{ESP}_P$ and the classic Tolman electronic parameter (TEP)\textsuperscript{134} are closely correlated,\textsuperscript{123} enabling replacement of an established and successful experimental descriptor with an easy-to-obtain computational version.
To challenge this QSRR model, we reserved data from Substrate set #2, which contains all ortho-substituted substrates, as an external validation set. Even though the MLR model training data had none of these substrates included, predictions for Substrate set #2 are still excellent, with a mean absolute error of 0.732 (compared to 0.529-0.542 for the training/test sets). This is possible because steric effects are already accounted for in the predicted $\Delta G^{\dagger}_{OA}$ values, since that model was trained using a far more diverse range of (Het)Ar–X substrates. In addition, the predicted $\Delta G^{\dagger}_{OA}$ model appears to function equally well for all 17 phosphine ligands. Only two outlier points are observed, corresponding to 2,4,6-triisopropylphenyl bromide as a substrate with $\text{P}(t\text{Bu})_3$ and $\text{P(Ad)}_2(t\text{Bu})$; in other words, the most sterically hindered substrate with the two largest phosphines. Individual univariate correlations of $\Delta G^{\dagger}_{OA}$ with $\ln k$ for all 17 phosphines confirm this generality. Thus, even though our initial oxidative addition QSRR model was built using only Pd(PCy$_3$)$_2$, it is clearly able to make accurate predictions for many additional catalysts.
Figure 9. Applying predicted $\Delta G^\dagger_{OA}$ values to external catalytic reaction datasets: unified QSRR of Plenio and co-workers’ Sonogashira initial rates for all substrate/catalyst combinations. Figure adapted with permission from ref. 47. Copyright 2022, Royal Society of Chemistry.
CONCLUSIONS AND OUTLOOK

The art and practice of catalyst and reaction development in synthetic chemistry is in a state of change. The advent of rapid tools to collect, visualize, and analyze the plethora of existing data from the academic and patent literatures is enabling chemists to use these data like never before. Increases in computing power and speed are greatly expanding the utility of high-throughput computational chemistry approaches to mapping chemical space with molecular descriptors, as well as rapid and accurate calculation of reaction intermediates and transition states for ever more complex molecules. And advances in laboratory automation and miniaturization are changing the way chemists design and conduct experiments, moving from traditional “one-factor-at-a-time” iteration to more holistic, multifactor approaches. Organopalladium catalysis is central to each of these three aspects, providing large existing datasets, rich catalyst structure space and mechanistic diversity, and a plethora of high-throughput studies.

With these powerful new tools and approaches becoming more mainstream in catalysis research, we must be cognizant of the challenges and drawbacks of each individual approach. Literature/patent data is known to be skewed by selection and reporting biases; computational results are not acquired instantaneously, nor are they error-free; and HTE is still bottlenecked by reagent/catalyst availability, and by analysis speed/accuracy. The ideal data-rich approach to catalysis research must combine all three sources of data and the corresponding analysis methods. Importantly, massively large datasets are not required to make accurate and actionable reactivity/selectivity predictions, especially if mechanistic aspects are taken into account. As an exemplar of this fact, our work on quantitative predictions for Pd oxidative addition reactivity is
broadly applicable across multiple reaction classes and accurate when presented with new data, despite being built from only ~80 relative rate constants.

While this Perspective showed how organopalladium catalysis has been the proving ground of choice for many data-rich methods, there is nothing inherent about palladium or its reactivity that makes it uniquely suited for these new approaches. Many other reaction classes are well-represented in the chemical literature, and high-throughput experimentation methods are increasingly accessible to many reaction types.\(^{90,135}\) The concepts from the case studies presented here, as well as others from the literature, are applicable to any type of synthetic chemistry. Finally, as remarked throughout, many of the seminal works on data-driven catalysis research are the result of academic/industry collaborations. These research partnerships will likely continue to be at the forefront of this emerging area.

**AUTHOR INFORMATION**

**Corresponding Author**

*dcleitch@uvic.ca

**Author Contributions**

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

**Notes**

The authors declare no competing financial interest.
ACKNOWLEDGMENT

We acknowledge with respect the Lekwungen peoples on whose traditional territory the University of Victoria (UVic) stands, and the Songhees, Esquimalt and WSÁNEĆ peoples whose historical relationships with the land continue to this day. We also acknowledge funding from the New Frontiers in Research Fund – Exploration and the NSERC Discovery Grant programs. D.C.L. thanks the Research Corporation for Science Advancement for a Cottrell Scholar Award.

REFERENCES


https://doi.org/10.1038/s42004-021-00550-x


https://doi.org/10.1039/C7SC04679K


(49) Sabatier, P. *La Catalyse en Chimie Organique*, 2nd ed.; Béranger, 1913.


(65) 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:


Reactions from the Medicinal Chemists’ Toolbox for Ultrahigh-Throughput Experimentation.  
*Nat. Synth.* **2023**. https://doi.org/10.1038/s44160-023-00351-1


https://doi.org/10.1021/acs.orglett.9b03380

https://doi.org/10.1002/anie.202006586


(127) Lu, J.; Leitch, D. C. Unpublished Results.


Dombrowski, A. W.; Gesmundo, N. J.; Aguirre, A. L.; Sarris, K. A.; Young, J. M.; Bogdan, A. R.; Martin, M. C.; Gedeon, S.; Wang, Y. Expanding the Medicinal Chemist Toolbox: