Building a toolbox for the analysis and prediction of ligand and catalyst effects in organometallic catalysis

Derek J. Durand, Natalie Fey*

School of Chemistry, University of Bristol, Cantock's Close, Bristol BS8 1TS, UK

*Corresponding author: Natalie.Fey@Bristol.ac.uk

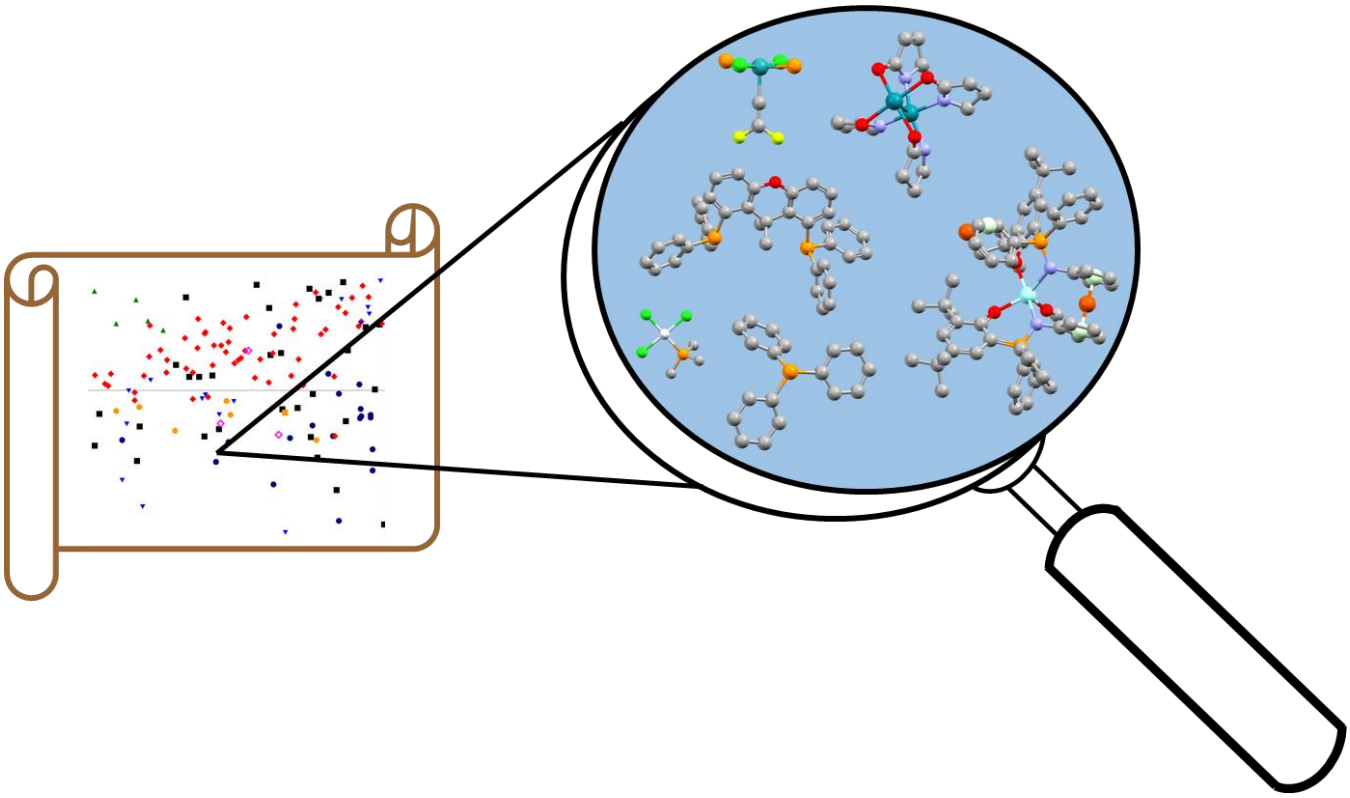
Conspectus

Computers have become closely involved with most aspects of modern life and these developments are tracked in the chemical sciences. Recent years have seen the integration of computing across chemical research, made possible by investment in equipment, software development, improved networking between researchers and rapid growth in the application of predictive approaches to chemistry, but also a change of attitude rooted in the successes of computational chemistry – it is now entirely possible to complete research projects where computation and synthesis are cooperative, integrated, and work in synergy to achieve better insights and so improved results. It remains our ambition to put computational prediction before experiment, and we have been working towards developing the key ingredients and workflows to achieve this.

The ability to precisely tune selectivity along with high catalyst activity make organometallic catalysts using transition metal (TM) centres ideal for high value-added transformations, and this can make them appealing for industrial applications. However, mechanistic variations of TM-catalysed reactions across the vast chemical space of different catalysts and substrates are not fully explored, nor is such an exploration feasible with current resources. This can lead to complete synthetic failures when new substrates are used, but more commonly we see outcomes that require further optimisation, such as incomplete conversion, insufficient selectivity, or the appearance of unwanted side products. These processes consume time and resources, but the insights and data generated are usually not tied to a broader predictive workflow where experiments test hypotheses quantitatively, reducing their impact.

These failures suggest at least a partial deviation of the reaction pathway from that hypothesised, hinting at quite complex mechanistic manifolds for organometallic catalysts which are affected by the combination of input variables. Mechanistic deviation is most likely when challenging, multifunctional substrates are being used, and the quest for so-called privileged catalysts is quickly replaced by a need to screen catalysts libraries until a new "best" match between catalyst and substrate can be identified and reaction conditions optimised. As a community we remain confined to broad interpretations of the substrate scope of new catalysts and focus on small changes based on idealised catalytic cycles, rather than working towards a "big data" view of organometallic homogeneous catalysis, with routine use of predictive models and transparent data sharing.

Databases of DFT-calculated steric and electronic descriptors can be built for such catalysts, and we summarise here how these can be used in the mapping, interpretation, and prediction of catalyst properties and reactivities. Our motivation is to make these databases useful as tools for synthetic chemists, so they challenge and validate quantitative computational approaches. In this account we demonstrate their application to different aspects of catalyst design and discovery, their integration with computational mechanistic studies, and so describe the progress of our journey towards truly predictive models in homogeneous organometallic catalysis.



Key References

- N. Fey, A. Koumi, A. V. Malkov, J. D. Moseley, B. N. Nguyen, S. N. G. Tyler, C. E. Willans. Mapping the properties of bidentate ligands with calculated descriptors (LKB-bid*). Dalton Trans.*, **2020**, *49*, 8169-8178.[1] *Extension of the Ligand Knowledge Base (LKB) approach to produce a new descriptor dataset for bidentate ligands capturing P, N, O & C donor atom combinations. Principal component analysis (PCA) of data reveals relationships between donor atoms and bridge lengths in donor properties, as well as showcasing the ability to map multiple ligand classes in one dataset.*
- A. Lai, J. Clifton, P. L. Diaconescu, N. Fey. Computational mapping of redox-switchable metal complexes based on ferrocene derivatives. *Chem. Commun.*, **2019**, *55*, 7021-7024.[2] *DFT calculations used to capture properties of 64 systematically chosen redox-switchable metal complexes relevant to ring-opening polymerisation of cyclic esters. Mapping this space using PCA to aid future synthetic choices highlights that ligand and metal changes have a more significant effect on catalyst properties than changing oxidation state of redox switch.*
- A. L. Green, C. P. Tinworth, S. Warriner, A. Nelson, N. Fey. Computational Mapping of Dirhodium (II) Catalysts. *Chem Eur. J.*, **2021**, *27*, DOI: 10.1002/chem.202003801.[3] *Development of a database of DFT-calculated descriptors which can be used in conjunction with PCA analysis to map dirhodium catalyst space, allowing exploration of reactivity landscape for any dirhodium catalysed reaction. PCA map shown to correlate with the results of an androgen receptor agonist molecular discovery project.*

1. Introduction

While computational results are now routinely used to contribute to the analysis of experimental data and confirmation of mechanistic postulates, computational predictions made before experiments remain the exception[4] in organometallic catalysis. More commonly, computational data feed into predictions in iterative feedback cycles, and so the lines between interpretation and prediction become blurred.[5-8] Better use of what we know about chemical reactions, whether from experiment or calculation, will facilitate *ab initio*, structure-

based prediction.[7] Here the adoption of statistics and data science into the molecular sciences are proving crucial, and this account will discuss some of the contributions the Fey group and their collaborators have made.

At this stage we want to be clear: the application of statistical methods to organometallic chemistry has deep roots,[4,6,9-12] and our work builds on the success of quantitative structure-property and linear free energy relationships. Other groups utilising different approaches have published successful examples, predicting the outcome of reactions such as Buchwald-Hartwig cross-coupling, nucleophilic aromatic substitution, and alkene ethenolysis.[13-15] These quantitative models, which combine computed descriptors with experimental data and statistical techniques such as multivariate linear regression (MLR) and machine learning algorithms, have been shown to be successful in predicting the outcomes of reactions within a well-defined region of catalyst space. They are reliant on large-scale training data and a high level of user confidence when it comes to the selection and evaluation of models, presenting barriers to non-expert users.

Our own efforts have approached the problem of prediction by providing chemists with tools designed to be widely applicable to organometallic chemistry, as well as responsive to subtle changes in catalyst properties, providing a platform for large-scale mapping and predictive modelling in this field. This is a shared endeavour which has databases of calculated catalyst descriptors developed in Bristol at its heart. Data-led approaches rely on access to high quality responses, and experimental and computational mechanistic studies represent considerable efforts, so while this account is focussed on the work of the Fey group, we give full credit to many discussions with our collaborators who are also co-authors of the references cited. Our focus will be on the descriptor databases and their analysis, assessing our journey towards prediction. It also is worth bearing in mind that "prediction" can range from a data-led design of experiments to a structure-based computational guide to the "best" catalyst for a given substrate; the former can be achieved now, but the latter remains a formidable challenge, since few catalytic routes are mechanistically robust, nor are they fully mapped.

1.1 A formidable challenge – organometallic catalysis

The multivariate nature of organometallic catalysis makes it time consuming to explore, and we often remain confined to broad interpretations of the substrate scope of new catalysts based on idealised catalytic cycles, rather than working towards a "big data" view of organometallic homogeneous catalysis. Efficient exploration of new reaction species is desirable, and transferability can be achieved by capturing variable effects in representative environments. There is a tension between a useable, linear models of few dimensions which have informed linear free energy relationships[9,12,16] and the exploitation of investments in calculations and experiments to generate a complete picture of higher dimensionality, which will give exquisite results for a small area of catalyst space but with more limited transferability.

Our efforts to build databases of DFT-calculated steric and electronic descriptors are thus focussed on their utility to chemical synthesis. The so-called "maps" of ligand space are intuitively appealing to synthetic chemists, and relatively simple regression models enhance our understanding of the mechanism of catalysis. Despite our preference for simplicity, more sophisticated models can also be developed when data for training and validation are available.

1.2 Computational matters

While a common concern for computational mechanistic studies is the selection of an appropriate level of theory, calculations contributing to descriptor database are more concerned with the trends observed rather than whether calculated energies match a "gold standard" calculation or experimental benchmark. The choice of DFT approach is thus less crucial.

For each database, the computational approach was chosen to produce results which capture trends reliably and robustly. (Computational details for all databases described here have been reproduced in the supporting information.) While the BP86 functional used in most cases generally produces good agreement with experimental geometries, it is known to overestimate binding energies, [17,18] but this and many of the other

concerns associated with density functionals are of less importance here. These errors are systematic and entirely benign for comparing trends across chemical space. When we started this work, dispersion corrections were still in their infancy, and optimisations with hybrid density functional approaches were prone to dissociate sterically demanding ligands than the non-local functionals. As we demonstrated later,[19] this choice will slightly shift the balance between descriptors in models (overstating steric effects).

More significant than the choice of density functional is the exploration of conformational space, and we have favoured using the lowest energy conformer for each coordination environment, adding computational noise to the data.[20] An alternative would be to introduce an energetic weighting to descriptor data, trading a different type of noise for breaking the structural link with the favoured conformation in different coordination environments. When possible, exploring conformational space is clearly sensible and actually facilitates the generation of input geometries for DFT calculations. However, developing bespoke force fields for challenging metal complexes is likely to lie outside of the scope of most studies. In the work described here we have relied on general force fields where possible, and available crystal structure geometries combined with our expertise in structural chemistry where these failed, evaluating multiple conformers to identify low energy structures. As noted, we have also analysed the likely level of conformational noise from such a strategy and found it manageable; this would not be true if we attempted enantioselectivity prediction.[8,21-24]

2 Exploring organometallic ligand properties

Structural and computational chemistry can elucidate the coordination and reactivity of organometallic catalysts, and the Ligand Knowledge Base (LKB) approach, reported by a consortium of authors (Orpen, Harvey, Fey)[25,26] is rooted in this tradition, challenged by novel ligands (Pringle)[27] and mechanistic insights (Lloyd-Jones).[28-30] Published contributions from the Fey group have focussed on the application of such databases to the analysis of ligand effects, as well as the derivation of statistical models for interpretation and prediction. As the LKBs have been described and reviewed repeatedly in past publications,[10,12,31] we will focus here on data-led predictions.

2.1 Visualisation of chemical space

Generating plots of a single steric descriptor *vs.* a net electronic descriptor is an appealing way of visualising similarities and differences between ligands, and in our work (see Table 1) we have used a statistical projection approach, Principal Component Analysis (PCA, discussed in greater detail below) to accommodate multivariate databases. While the dimensions become harder to assign cleanly to steric and electronic contributions, the resolution of properties is improved and our use of DFT calculations provides exquisite resolution of electronic effects. In addition, linear relationships are intuitively appealing, and establishing the relationships between structures and energies is a powerful application of computational methods.

Table 1: Summary of published Knowledge Bases for organometallic catalysis. An expanded version with references and timelines for each LKB is available in the ESI.

Knowledge Base	Chemical Space Covered	No. of Descriptors	No. of Entries
LKB-P	Monodentate P ligands	28	366
LKB-PP	Bidentate P,P and P,N ligands	28	324
LKB-PP_{screen}	Bidentate P,P ligands, accelerated computational screen	28	275
LKB-C	Monodentate C ligands	26	113
LKB-Bid	Bidentate P, C, N & O ligands	20	224
Alkyne/Vinylidene	Alkyne/Vinylidene coordination	35	60 pairs
Metallocene catalysts	Redox-switchable catalysts	19	64 pairs
Dirhodium catalysts	Dirhodium (II) catalysts	14	48

LKB-P

Our initial report, detailing a prototype ligand knowledge base for 61 monodentate phosphines (LKB-P), explored descriptors that could be obtained from the coordination of a range of representative complexes and adducts to the ligand, as well as optimisation of the free ligand.[25] Geometry optimisation with a standard DFT approach (BP86/6-31G*) of the free ligand, as well as the protonated ligand ($[HL]^+$) and a borane adduct ($[H_3B•L]$) can be used to explore the σ-donating properties of ligands, while representative metal complexes, in this case a tetrahedral $[Pt(PH_3)_3L]$ and square planar $[Pd(Cl)_3L]^-$, offer descriptors covering both σ- and π-bonding properties, as well as capturing the *trans* influence in the palladium complex, alongside the steric interactions with the ligand.[9,25] A final representative adduct, the He_{8_ring} fragment, was used to calculate the He_{8_steric} descriptor, which measures the interaction energy between a ligand and a ring of eight helium atoms aligned to resemble the *cis* ligands in an octahedral complex. This descriptor was found to correlate well with the Tolman cone angle (Fig. 1) while also affording additional information about the steric bulk of the ligand from the perspective of a hindered metal fragment. Together with the so-called Orpen S4' descriptor, this captures a purely steric effect.[32]

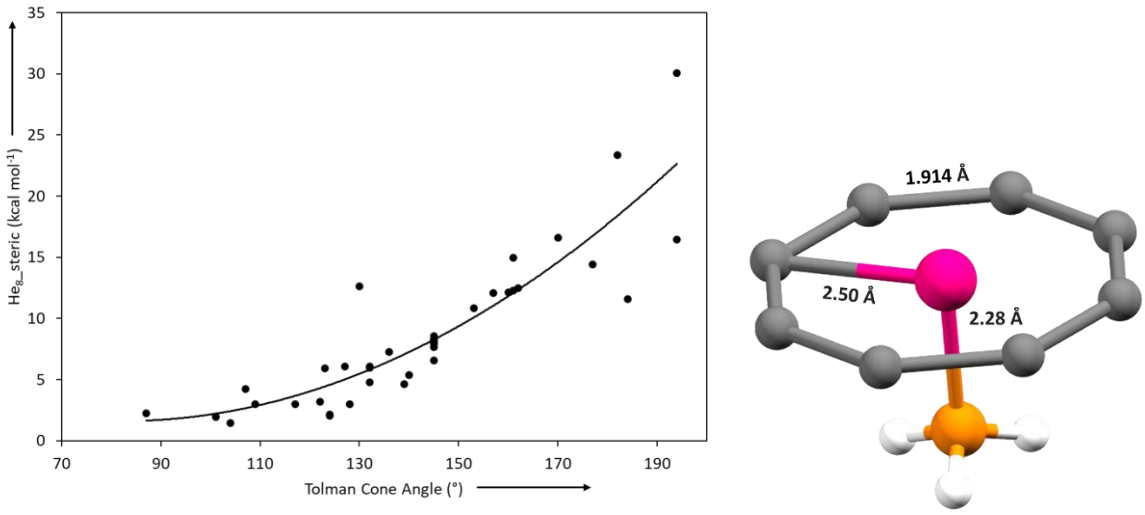


Figure 1: (left) Plot of Tolman cone angle versus He_{8_steric} descriptor for monodentate P ligands, nonlinear relationship described by a cubic function, $R^2 = 0.796$. Adapted with permission from ref. 25. Copyright 2006, John Wiley & Sons. (right) Geometry used for computation of the He_{8_steric} descriptor: the interaction energy between the phosphorus ligand (PA_3) and a ring of eight helium atoms.

These calculations gave rise to 28 descriptors, curated to be responsive to changes in ligand, straightforward to calculate and easy to extract. These descriptors are highly correlated, making Principal Component Analysis (PCA) useful for visualisation of the datasets.[33,34] We note that other approaches for dimensionality reduction may be of interest, too, but linear models appear to be adequate in these cases and are perhaps easier to interpret. This projection method seeks to capture most of the variation in the data in as few dimensions as possible, giving rise to new orthogonal descriptors (principal components, PCs) which are linear combinations of the original descriptors. For ligand data, the first two PCs routinely capture between 50-65 % of the variation, with all descriptors loading highly, capturing the variation across ligand space. While 28 descriptors lead to 28 PCs, their importance declines and the first few PCs are usually most important. PCA is not statistically robust, such that PC compositions and orientations change as ligands and descriptors are changed, but chemically-intuitive clustering of ligands is observed and sustained when ligands are added, facilitating the use of these PC score plots as "maps" of the ligand space. The maps can be enhanced by projecting additional information, either about the ligands (such as the type of substituent) or about their chemical behaviour, facilitating the observation of stereoelectronic trends. We generally counsel against interpretations of PC composition and PCA is not statistically robust, meaning that as the composition of individual PCs, as well as their order will change in

response to changes to the dataset. PCA identifies the largest sources of variation in the dataset, so a new ligand class can cause changes to these derived variables. For monodentate P-donor ligands, however, this can be attempted as most of accessible ligand space appears to have been sampled in recent published iterations of the LKB-P and these show the same general split between loadings, with PC1 capturing steric and σ-donor properties, while PC2 loads descriptors associated with π-backdonation and steric effects.[29] Higher PCs may be of interest, and for LKB-P, PC3 captures the ligand response to coordination. The maps are not directly predictive but there is a relationship between ligand structure and catalyst properties, supporting catalyst selection.

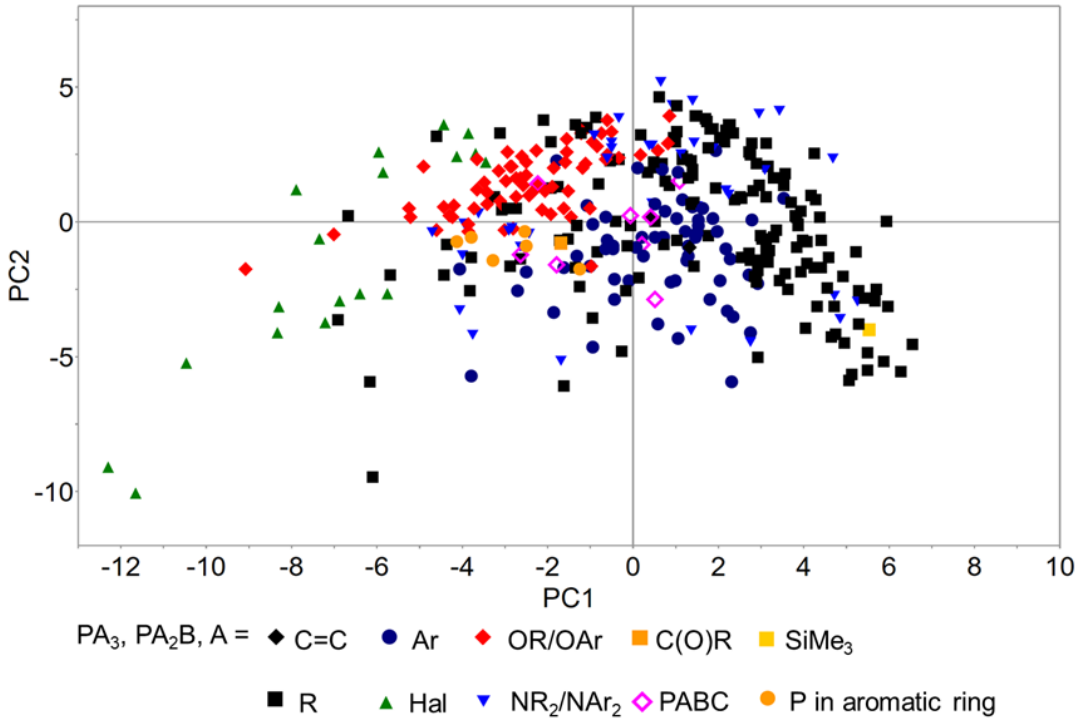


Figure 2: Principal component score plot (PC1 vs PC2) for ligands in the LKB-P.[35] PABC ligands indicate three different substituent groups. 366 ligands, *ca.* 62% of variation in data. Adapted with permission from ref. 35. Copyright 2015, Taylor & Francis.

The LKB-P database has been updated, substantially improving the sampling of P-donor ligand space while further additions focussed on fluorophosphines and unusual ligand designs (Fig. 2).[27,29,35]

These maps can be useful in their own right, as demonstrated in a collaboration with Paul Pringle's group concerned with the design of novel ligand structures, with a particular focus on monodentate P-donor ligands substituted by fluorine for applications in hydroformylation and hydrocyanation.[27] In this work, we were interested in assessing ligand designs before synthesis and three fluorophosphine ligands (a phospha-adamantane cage ligand (**1**) and two phosphabicyclic ligands (**2**, **3**), Fig. 3) were found to be located in a similar sector of P-ligand space to R_2PF ligands. The phospha-adamantane cage ligand **1** was found to be completely stable to thermal disproportionation and hydrolysis, leading to the synthesis of **2** and **3** as similar ligands for screening purposes. These were not as stable as **1**, but still far exceeded the stability of $(tBu)_2PF$. In addition, all three of these ligands occupy P-ligand space more similar to bulky aryl phosphite ($P(O\text{-}Ar)_3$) ligands, which perform well for range of catalytic reactions, than known R_2PF ligands.

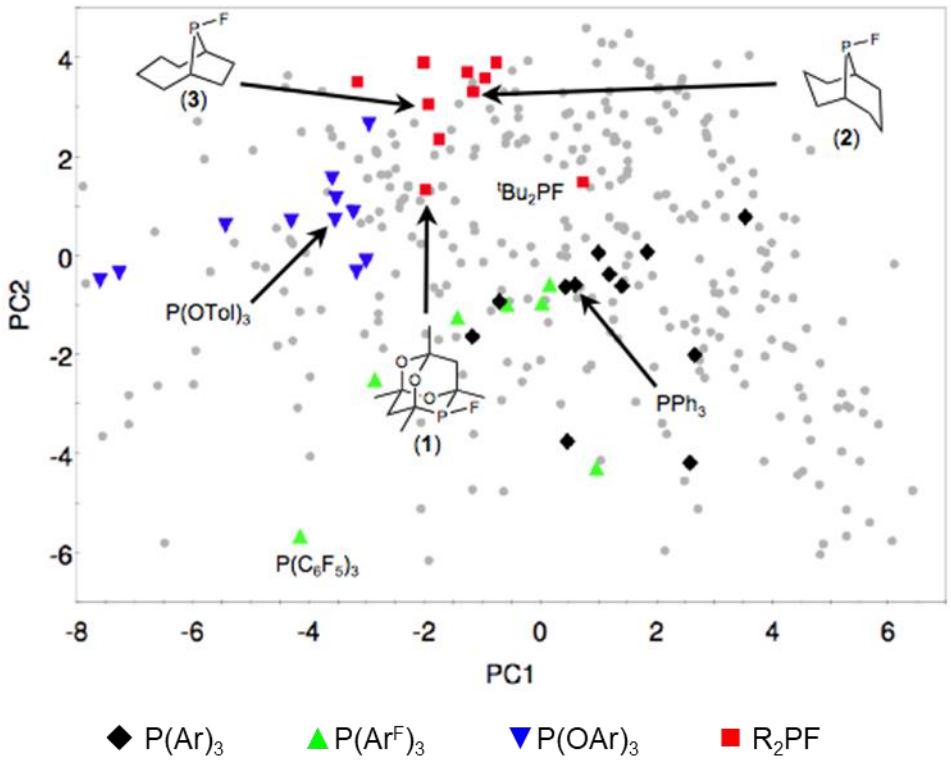


Figure 3: Principal component score plot (PC1 vs PC2), showing PAr_3 (black diamonds), $P(Ar^F)_3$ (green triangles), $P(OAr)_3$ (blue triangles), and R_2PF (red squares). Other ligand classes are represented as grey dots, Ar^F=fluoroaryl.[27] Reproduced with permission from ref. 27. Copyright 2012 John Wiley & Sons.

Experimentally, these new ligands proved to be promising for catalysis, achieving activities and selectivity comparable to ligands used in industry for both the hydroformylation of 1-heptene (PPh_3) and the hydrocyanation of 3-pentenenitrile ($P(O\text{-}o\text{-}Tol)_3$). This observation corresponded well with their location on the map, and we further demonstrated that proximity in ligand space can correspond to similar chemical behaviour[35] and that projection of experimental observations onto the map can highlight clusters of ligands that give rise to active catalysts.[29,36] These applications of LKB-P data are most useful for the design and analysis of experimental screening campaigns, but hold limited predictive power alone.

However, the utility of the LKB-P database goes beyond this mapping and we have used it in the fitting of multivariate regression models (linear and potentially non-linear (LASSO, PLSR)), trained on structural and calculated data,[29] Tolman electronic parameters,[26] and experimental measurements of catalyst activity (Fig. 4).[29,36] Model evaluation suggested that these would be of limited usefulness for prediction since experimental data were clustered rather than capturing all of chemical space. In such situations, global models may be difficult to achieve; not all ligands will give active catalysts, hampering wide-ranging screening.

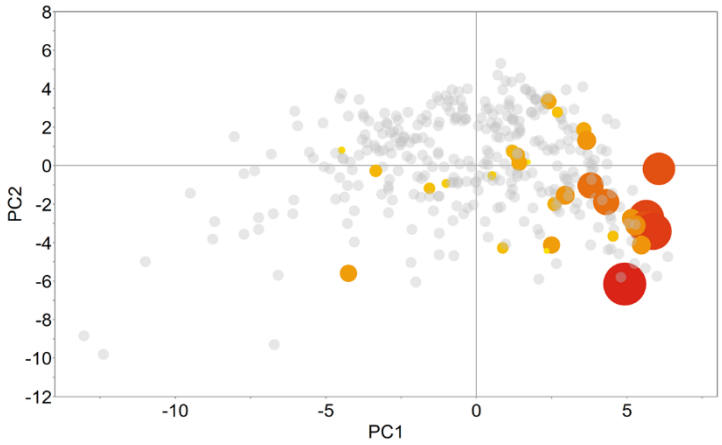


Figure 4: Projection of FRET yields for palladium-catalysed amination of aryl bromide onto LKB-P map at the time of LKB-P expansion paper.[29] Size and colouring relate to yield, with dark red, large spots corresponding to the highest yields and small yellow spots corresponding to lowest yields (range 8-80%).[29] Adapted with permission from ref. 29. Copyright 2010 American Chemical Society.

We have also demonstrated the use of LKB-P descriptors for the analysis of ligand effects on calculated barriers.[30]

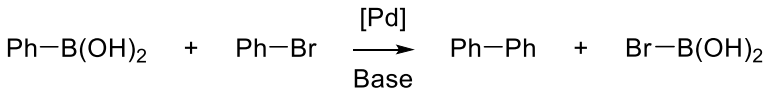


Figure 5: Overall reaction for Pd catalysed cross-coupling of phenyl bromide and $PhB(OH)_2$, which under reaction conditions forms $PhB(OH)_3^-$.

Early initial work focused on a computational exploration of the mechanism for the Suzuki coupling of phenyl bromide and $PhB(OH)_3^-$ (Fig. 5) using four different $Pd(PR_3)_2$ catalysts (R= Me, CF_3, Ph and ^tBu).[30] The barriers to fundamental steps in the catalytic cycle were calculated for four model catalysts and used to build simple multi-variate linear regression models, utilising two of the LKB-P descriptors. All four models (two models were created for the transmetallation step) showed high regression coefficients (R^2= 0.820 – 0.996), indicating that the relationship between our descriptors and the individual barriers is described well by a linear equation.

In later mechanistic studies,[37,38] consideration of competing pathways has highlighted the challenges for using computational data to train such models. For the oxidative addition step common to most cross-coupling reactions, even relatively small changes to aryl halide substrates[37] or ligands[38] lead into different pathways, with predicted differences at the very edge of computational reliability.

Such an approach could be extended to a wider range of ligands and mechanistic variations, either for mechanistically more robust cases such as metathesis, as demonstrated by Jensen's team,[39] or by more extensive computational exploration. These studies show that using computational data as a response in predictive model is viable, provided enough of the variable effects on the manifold of competing reaction pathways can be explored. We are currently working on descriptors which can directly encode mechanistic information.

LKB-C

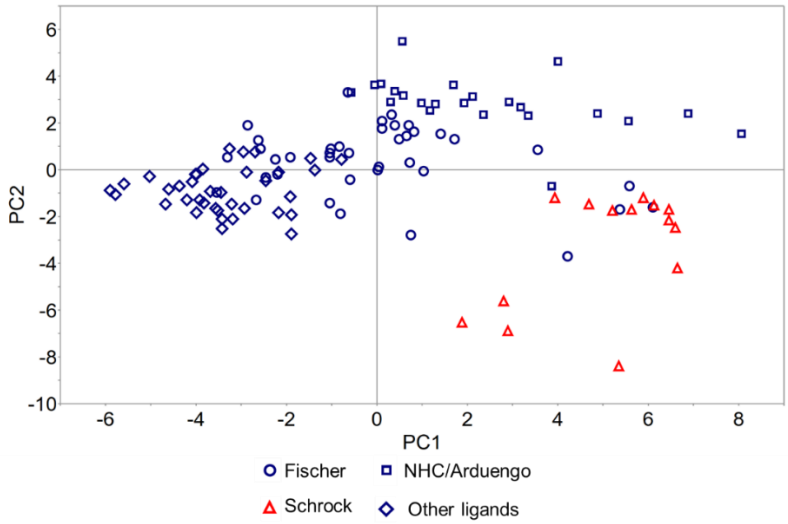


Figure 6: Principal component score plot (PC1 vs PC2) for the LKB-C. 113 ligands, *ca.* 57% of variation in data shown.[40] Adapted with permission from ref. 40. Copyright 2009 The Royal Society of Chemistry.

The LKB-C is designed to capture the properties of carbene ligands and related C-donors (Fig. 6).[40] Carbene-metal bonding is potentially more difficult to quantify than that of P or N donor ligands, as the substituents of carbene ligands determine whether they are more stable in their singlet or triplet state, which in turn affects their bonding bond as either Fischer or Schrock (Fig. 7).

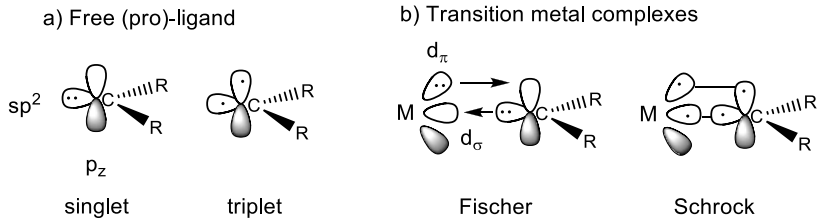


Figure 7: Carbene electronic configurations and binding.

The electronic configuration affects the behaviour of carbene ligands, and the LKB-C assesses this by calculation of the free ligand in both singlet and triplet configurations. The $He_{8\text{-steric}}$ descriptor was adapted for carbenes (D-X was changed from 2.28 Å to 1.88 Å to represent the shorter bond distances between metals and carbenes) and new complexes were introduced to improve characterisation of carbene ligands. A linear gold(I) chloride complex, [AuClL], was introduced to assess ligand-metal interactions largely free of steric hindrance, and this fragment was later also added to the LKB-P database.[29] *Trans*-$[RuCl_2(PH_3)_2L]$, with the carbene ligand bound axially, was added to capture ligand behaviour in a more sterically-demanding environment than other fragments.

Few published datasets sample both singlet and triplet carbenes, except when modelling the steric descriptor %VBur. We modelled the calculated binding energy of the trans CO ligand in $[Cr(CO)_5L]$, where L is the carbene in either singlet or triplet state, with LKB-C descriptors. While reasonable model performance and prediction errors could be achieved in these cases, it would be desirable to test this database on experimental results.

LKB-PP

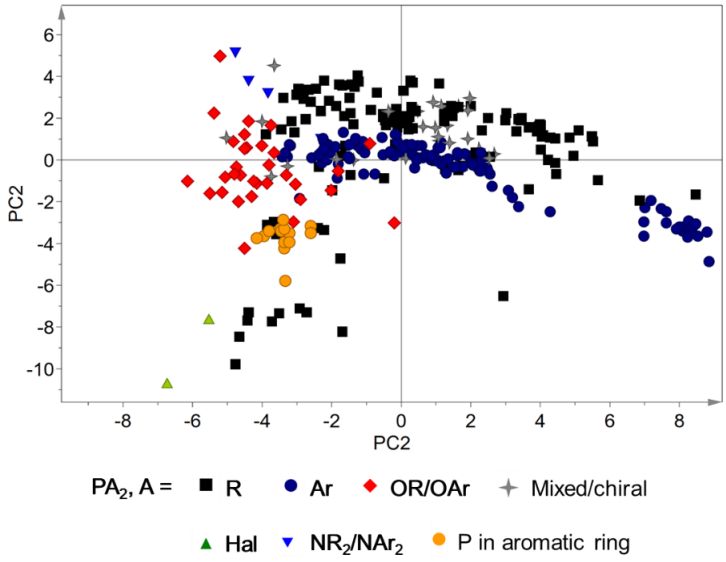


Figure 8: Principal component score plot (PC1 vs PC2) for LKB-PP. 324 ligands, *ca.* 58% of variation in data shown.[41] Adapted with permission from ref. 41. Copyright 2018 American Chemical Society.

We applied a similar philosophy to P,P (and P,N) ligand bidentate space leading to a LKB-PP database (Fig. 8).[20] Here, descriptors were adapted and modified to accommodate two donor atoms, capturing ligands across a range of coordination environments. A new steric descriptor, the He_{8_wedge} fragment, consists of He atoms positioned to have a comparable steric profile to the closest approach of ligands to a chelating ligand in an octahedral complex,[20] with a second version of this wedge (nHe_8) allowing for structural compression of the ligands in response to steric "pressure"(Fig. 9). The palladium fragment used in LKB-P was modified to form the analogous *cis*-ligated bidentate square-planar complex $[Pd(Cl)_2\{LL\}]$, and a tetrahedral zinc chloride $[(Zn(Cl)_2\{LL\}]$ was used to represent a typical Lewis acid complex, focusing on σ-donation.

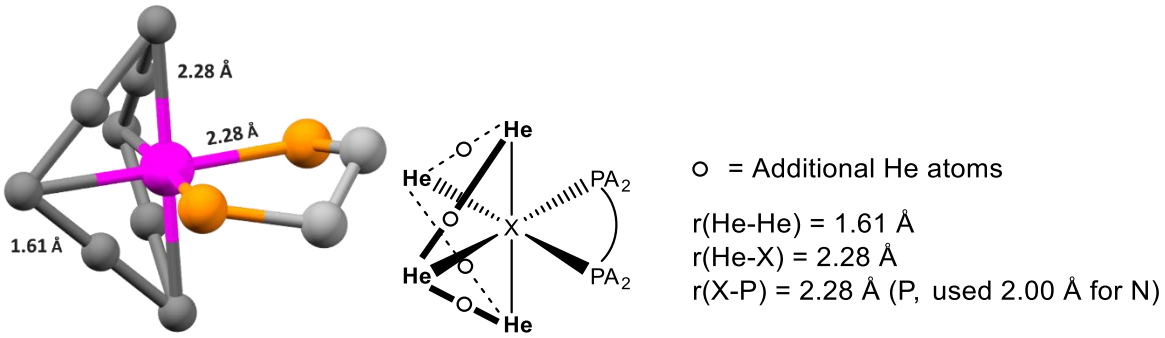


Figure 9: Representations of the $[He_8 \cdot \{LL\}]$ system used for the calculation of the steric descriptors He_{8_wedge} and nHe_8.[42] Adapted with permission from ref. 42. Copyright 2012 American Chemical Society.

Non-symmetrical chelating ligands, as well as those with differing donor atoms, mean that calculation of the frontier molecular orbitals (to obtain HOMO and LUMO descriptors for the donor atom) and proton affinities are harder to locate and analyse due to orbital mixing. To accommodate this, a simple set of splitting rules (truncation) for bidentate ligands along the bridging backbone yields these descriptors for each donor atom separately. This truncation became unwieldy when considering multiple chelating ligand classes for the recently reported LKB-Bid, so this descriptor was not included there (discussed below).[1] Similar to the LKB-P, we demonstrated that LKB-PP descriptors can be used to train regression models on available experimental and calculated data.[20]

The version of the LKB-PP map in Fig. 8 indicates that P,P ligand chemical space within the LKB-PP contains many alkyl, aryl and phosphite substituted ligands, but sampling of other substituents (halogenated and amines), is

relatively uncommon. Literature reports of bidentate ligands show a stronger focus on backbones than substituents. [11,43]

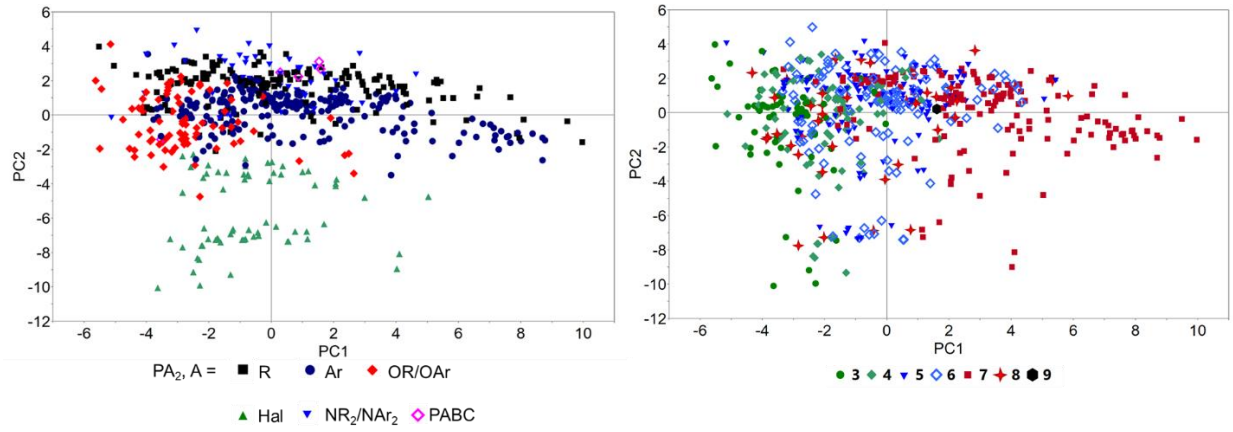


Figure 10: Pairwise principal component plots (PC1 vs PC2) for LKB-PP$_{screen}$.[44] Coloured by (left) substituent type and (right) backbone length. 275 ligands, *ca.* 56% of variation in data shown. Reproduced with permission from ref. 44. Copyright 2013 The Royal Society of Chemistry.

We were interested in exploring a wider range of substituents and used a streamlined computational approach to screen substituent and backbone effects in a new version of this database, LKB-PP$_{screen}$, with a view to predicting where experimental efforts could broaden ligand properties most effectively.[44] Utilising a representative subset of 11 substituents and 25 backbones, a computational evaluation of 275 bidentate *cis*-chelating P,P ligands was carried out, and the resulting pairwise PC plots from the extracted dataset are shown above (Fig. 10).

By colouring the plots by variation of both substituent groups (left) and backbone length (right), bands can be observed for substituent groups (in the PC2 direction) and backbone length (in the PC1 direction). While these bands are not fully separated (due to the combined influence of steric and electronic effects) the clustering of different substituent types and backbone structures is becoming a little more apparent. Halogenated substituents occupy a region of ligand space which was previously not captured by our LKB-PP database, and this highlights the potential of an *in silico* approach for assessing the properties of novel ligands with desirable catalyst activities.

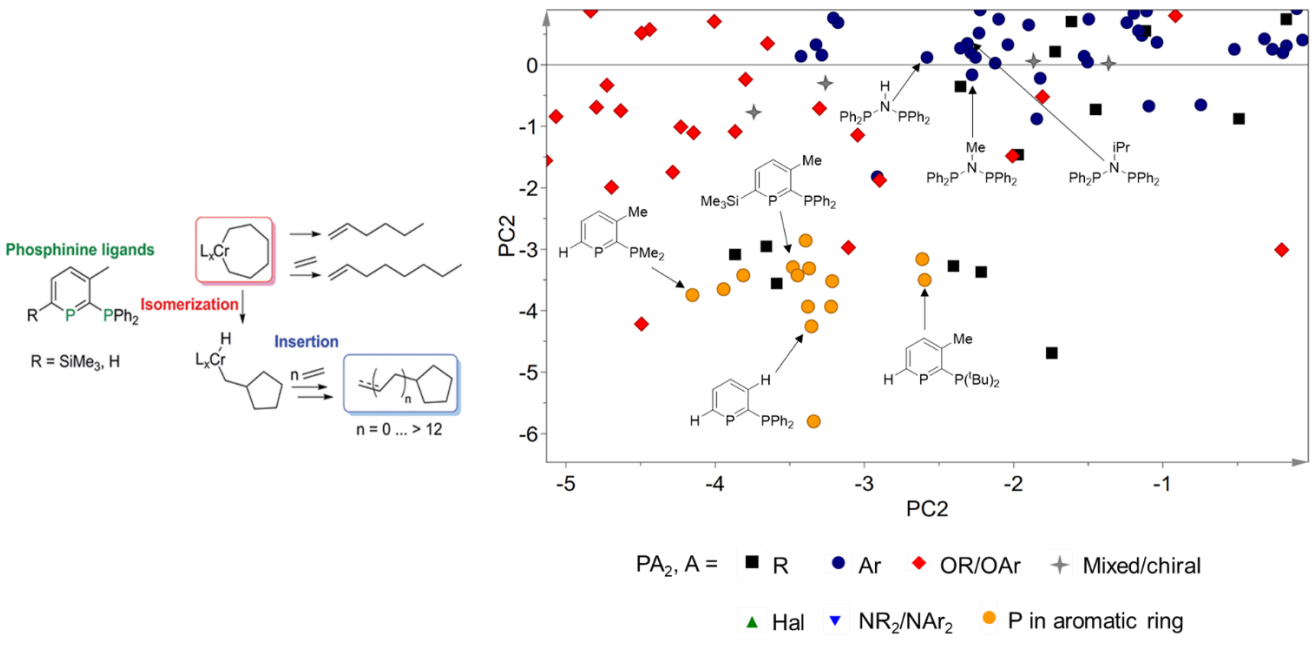


Figure 11: Summary of phosphinine ligands explored and the unexpected insertion chemistry observed for these species when Cr-catalysed oligomerisation of ethene was carried out (left, adapted with permission from ref. 41. Copyright 2018 American Chemical Society). Subset of Fig. 8, highlighting phosphinine ligands and nearby aryl-substituted PNP ligands.[41]

Computational mapping was used to guide possible catalytic applications of novel chelating ligands in work published in collaboration with colleagues at Heriot-Watt University and Sasol Technology Ltd.[41] The two novel 2-phosphinophosphinine ligands shown in Fig. 11, with remarkably small bite angles (65.1−68.3°), were synthesised and then added to LKB-PP along with 12 phosphinine ligands. Analyses highlighted the tunability of these ligands and allowed us to compare them (shown as orange circles in Fig. 8) with other ligand architectures. They were found to occupy a region of chemical space adjacent to aryl-substituted PNP ligands, which are used in chromium-catalysed oligomerisation of ethene to produce 1-hexene and 1-octene, prompting their experimental testing. They actually promote the production of isomerised products, alkyl- and alkenylcyclopentanes, highlighting the sensitivity of catalytic pathways to ligand modifications.

LKB-bid

We have recently presented the LKB-bid database, visualising ligand space for a wider range of chelating ligand, capturing combinations of P, N, O & C donor atoms, along with both neutral and charged ligands.[1] The descriptors used in this database are closely related to those in LKB-PP, with only minor tweaks, such as the removal of split ligand descriptors from consideration. As for P,P bidentates, grouping according to substituent and backbone variations can be observed (Fig. 12) and similarities of ligand classes become apparent, with the larger/softer P-donor ligands clearly separate and ligands involving carbenes also standing out (these bind more strongly to the metals than other donors). The mixing of donor atoms leads to interesting overlaps of ligand clusters, in particular for O,O, N,N and O,N donor ligands, and this prediction aligns well with the usage of these ligands in copper-catalysed Ullman cross-coupling reactions.

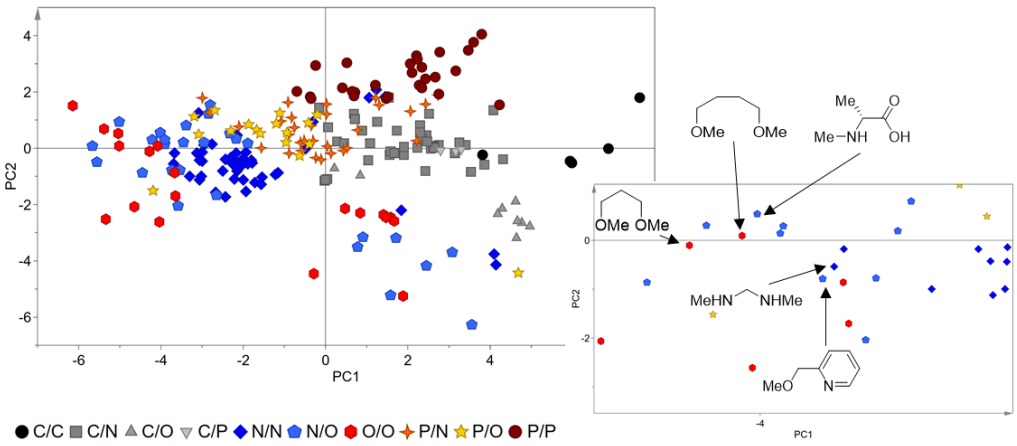


Figure 12: Principal component score plot (PC1 vs PC2) for LKB-Bid (left) and zoomed version for O/O, O/N and N/N region (right). Colouring by donor type, 224 ligands, *ca.* 52% of variation in data shown.[1] Adapted with permission from ref. 1. Copyright 2020 Royal Society of Chemistry.

2.2 Understanding and controlling substrates

Having built a foundation to model ligand effects through calculated descriptor sets, we have expanded our approach to explore substrate effects. Working with Jason Lynam and collaborators at the University of York, we have captured a range of variables which affect tautomer preferences on alkyne coordination to metals.[45]

Alkynes can either be η^2-coordinated substrates or tautomerise to form a metal-vinylidene complex. This tautomerisation affects the reactivity of alkyne substrates, as the formation of the vinylidene leads to a change in selectivity for nucleophilic attacks, promoting an Anti-Markovnikov addition whereas alkynes undergo Markovnikov additions (Fig. 13).

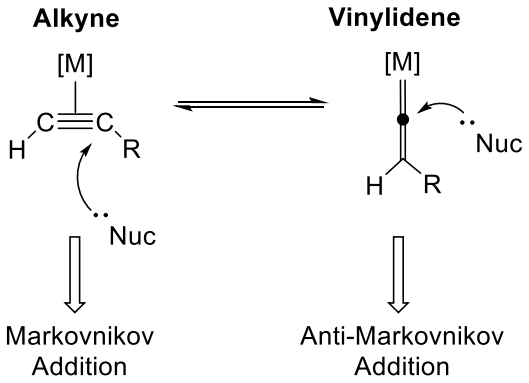


Figure 13: Alkyne/Vinylidene tautomerisation in metal complexes, and their differing reactivity in nucleophilic attacks.[45] Reproduced with permission from ref. 45. Copyright 2014 American Chemical Society.

Initial experimental studies indicated that these systems were sensitive to variation in substituents, making them amenable to computational study along similar lines as the LKBs described above. We considered three representative metal complexes ((V/A)-Au-Cl, (V/A)-Rh$(PMe_3)_2Cl$ and (V/A)-Ru$(PH_3)_2(Cl)_2$, where V=vinylidene and A=alkyne), in part related to work on monodentate ligands, giving 35 descriptors, capturing the properties and responses of these complexes to the introduction of tautomeric pairs of monosubstituted (32) and disubstituted (28) alkyne/vinylidenes. Of these descriptors, three ($\Delta E_{v\text{-}a}$ for each of the complexes) directly measure the stability of the two tautomers, giving rise to the PC plot shown below, mapping the substrate space effectively and in an interpretable manner. In addition, calculations showed that the energy difference between the tautomers varies systematically according to both substituents and metal (Fig 14D).

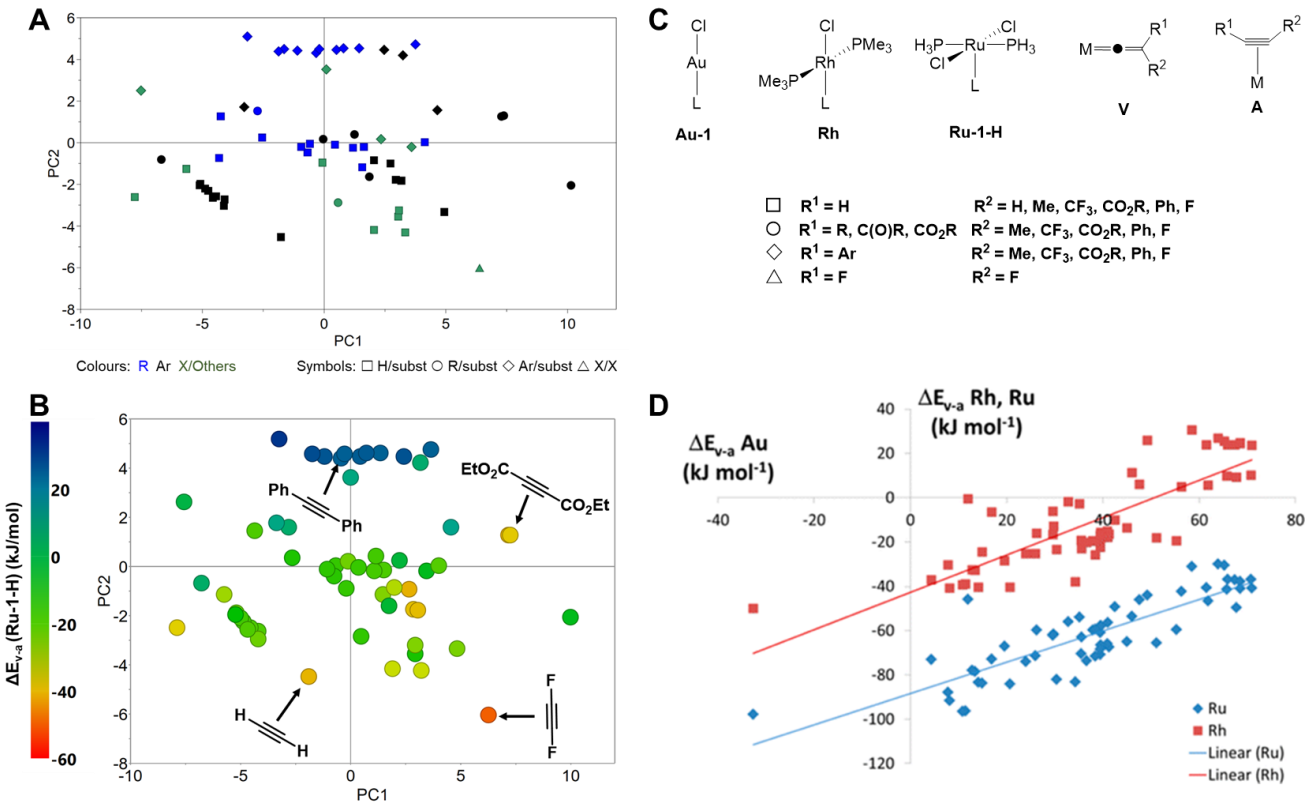


Figure 14A-D: A+B Principal component score plots (PC1 vs PC2) coloured by substituent group (A) and $\Delta E_{v\text{-}a}$ (**Ru-1-H**(B), 60 datapoints, ca. 65% of data variation shown. **C** Representative complexes for tautomeric pairs and tautomeric pairs for computational investigation, totalling 60 pairs. **D** Scatter plot of tautomer energy differences ($\Delta E_{v\text{-}a}$, kJ mol^{-1}) for metal complexes considered in the substituent effect branch. R^2 = 0.709 (**Ru-1-H**), 0.679 (**Rh**).[45] (Adapted with permission from ref. 45. Copyright 2014, American Chemical Society.)

In addition to the visualisation of substrate space, colouring the map by the $\Delta E_{v\text{-}a}$ (**Ru-1-H**) descriptor allowed us to interpret and make predictions about which types of substrates are likely to form either tautomer, with the colour scheme clearly showing that bulky disubstituted alkynes are less likely to form the vinylidene on complexation than electron-poor substituted alkynes, particularly for halogen-substituted substrates such as 1,2 difluoro-ethyne.

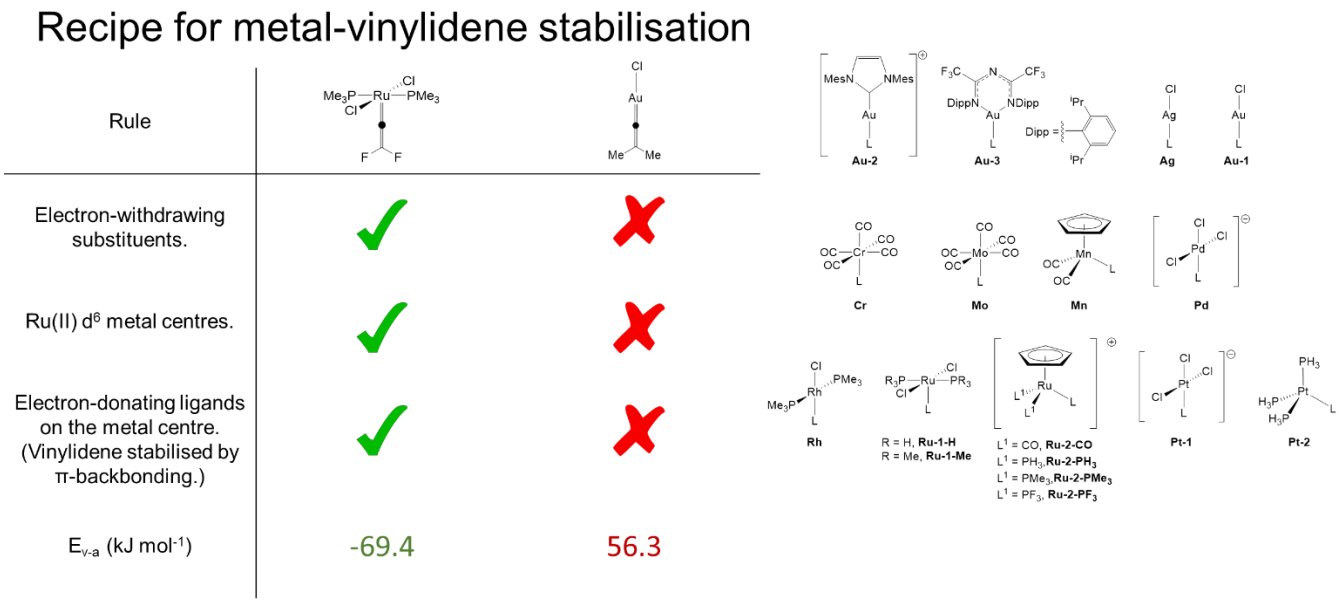


Figure 15: General rules for stabilising metal-vinylidene complexes and the seventeen complexes used to explore metal and ligand effects on $\Delta E_{v\text{-}a}$.[45] (Adapted with permission from ref. 45. Copyright 2014, American Chemical Society.)

We explored our ability to manipulate the tautomeric preference by modifying the metal centre and calculating $\Delta E_{v\text{-}a}$ for a subset of six tautomers chosen to sample substituent effects. We also considered ligand effects on some of the metal complexes. We were able to formulate design criteria for the stabilisation of vinylidene complexes (Fig. 15). We then tested whether these insights could be used to stabilise gold vinylidenes;[46] while those remained elusive, demonstrating the perils of predictions in chemistry, we discovered new intermediates and reaction pathways.

2.3 Catalyst characterisation

In some cases, a focus on ligand descriptors is insufficient for the characterisation of complex catalytic manifolds, especially where the metal centre substantially defines the coordination sphere, and we have developed broader catalyst databases to identify key variables for modification and fine-tuning in two cases, redox-switchable polymerisation catalysts and dirhodium catalysts.

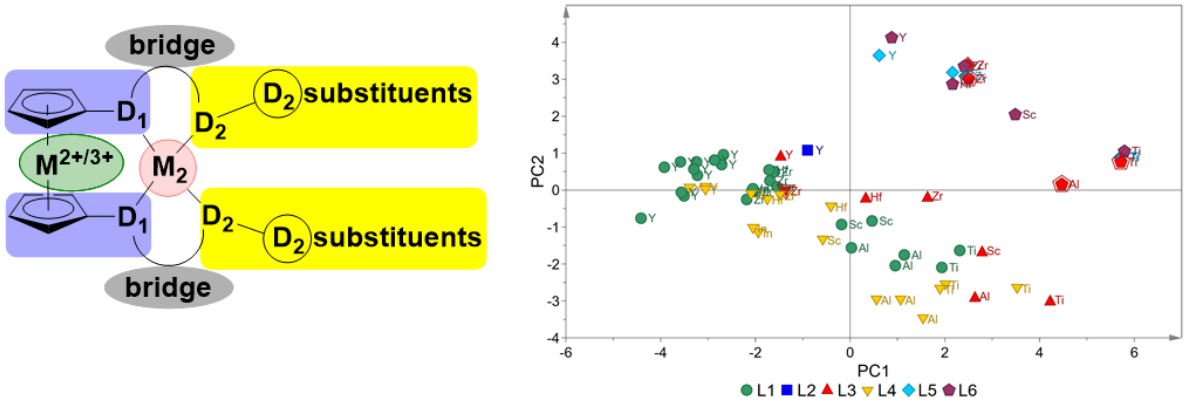


Figure 16: Variables explored for redox-switchable catalysts (left, reproduced with permission from Ref. 2, Copyright 2019 The Royal Society of Chemistry) and principal component score plot (right, PC1 vs PC2, adapted with permission from ref. 2. Copyright 2019 Royal Society of Chemistry) showing 64 redox pairs of metal complexes, coloured by ligand type and labelled by M_2, ca. 56% of variation in data shown. Red border rings highlight species known to be active for ROMP catalysis.[2]

The first of these examples arose from a collaboration with the Diaconescu group at UCLA, focusing on their work with redox-switchable metallocene-based catalysts.[2] These complexes are difficult to synthesise and have a range of targets for modification (Fig 16). Computational assessment of novel catalyst designs would be desirable, both to understand how modifying each variable affects properties, and to prioritise targets for synthesis. Mechanistic studies are challenging as reaction pathways are complicated by the redox-switch of the metallocene backbone and would not be feasible within a reasonable timeframe for catalyst development.

By combining seven metals (Sc(III), Y(III), Ti(IV), Zr(IV), Hf(IV), Al(III) and In(III)) with six supporting ligand architectures, we calculated descriptors for 128 metal complexes, capturing their properties for both Fe(II) and Fe(III) oxidation states of the backbone. We found that changing the oxidation state of the metallocene had a comparatively small effect on the descriptors considered, while modification of the active metal and ligand architecture were more important. Highlighting species known to be active catalysts for ring-opening metathesis polymerisation (ROMP) showed that our map was in broad agreement with experimental observations that there is no single activity hotspot for these catalysts, as well as providing a framework for further exploration of catalyst space.

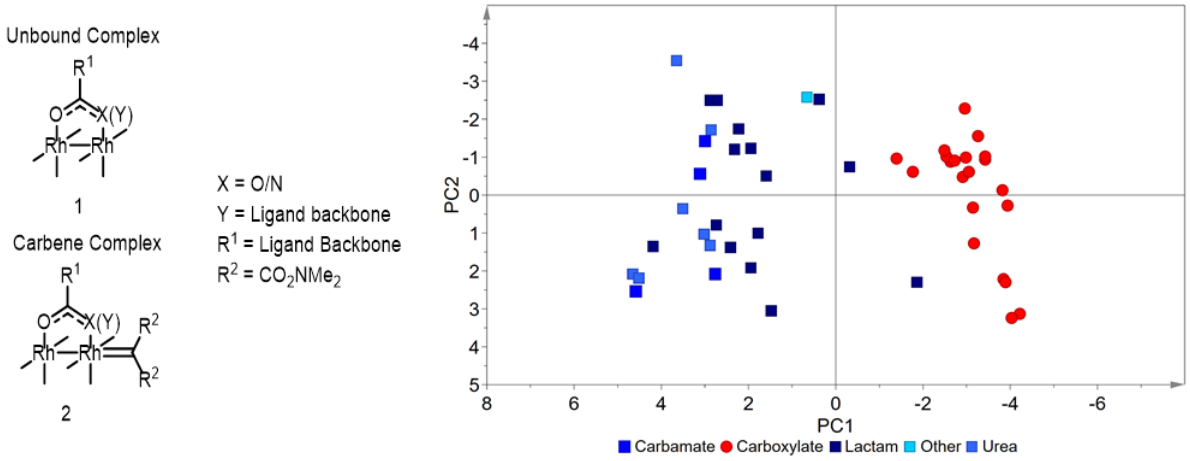


Figure 17: Structures calculated for dirhodium descriptors (left) and principle component score plot (right, PC1 vs PC2) for 48 catalysts, coloured by ligand donor type (red for O,O and blues for N,O). ca. 60% of data variation shown. Adapted from ref. 3, which is published under a CC BY Open Access license by John Wiley & Sons.

The second example of catalyst descriptors was developed in collaboration with the Nelson group at the University of Leeds and industrial partners at GlaxoSmithKline, and explores dirhodium(II) catalysts, which are known to be active for a wide range of transformations.[3] Mechanistic study of these catalysts is challenging as they catalyse a range of different transformations and a robust mechanistic hypothesis is unlikely.[47] The Nelson group are interested in this catalytic promiscuity to achieve diverse outcomes and we were keen to avoid biasing models towards a specific reaction pathway. We selected 14 descriptors from two metal complexes (the unbound Rh-Rh complex and the carbene bound complex) (Fig. 17, left) for 48 bidentate ligands from a broader set of calculated descriptors. The descriptors capture key features of the dirhodium core (Table S10), as well as the frontier molecular orbitals of complex 1, ligand and metal charges, steric descriptors for the cavity generated by the ligands and the coordination energy for the formation of complex 2 from a diazo precursor. Ligand descriptors alone did not capture the geometry of the dirhodium core well. These descriptors were used to map the catalyst space, showing clustering based on ligand donor atom, as well as the steric environment produced by the ligands. This approach was validated by showing that the resulting catalyst map/database could be used to support the discovery of androgen receptor agonists, through the visualisation of reaction yields which identified reactivity hotspots. While this is promising for experimental design, true prediction was hampered by the lack of training data.

3. Towards data-led prediction

Our databases provide chemically-intuitive tools for the design of screening experiments, which can support the interpretation of observations and lead to the prediction of catalyst behaviour. We have described several databases focussed on ligand properties, as well as investigating the coordination of alkyne/vinylidene tautomeric pairs to different metal centres. We have also considered two groups of organometallic catalysts where the coordination sphere is more complicated and there we have sacrificed transferability for a more focussed exploration of the effects of variables on the catalyst.

While we have demonstrated how such data can be used to build multivariate models throughout, this relies on the availability of suitable data for the training and validation of models. It is here where the experimental literature and computational capabilities fail us, and this is often for good reason – the mechanism of reaction may well change, or fail, as we traverse chemical space.[38,46] Along with investigating catalyst failure as a matter of routine, standard catalyst screening conditions and benchmarking sets will help to address this over time, as will large-scale, systematic computational studies. But there is still a long way to go before we reach *ab initio* structure-based prediction in organometallic catalysis, and we look forward to more exciting adventures and collaborations along the journey.

Biographies

Derek J. Durand obtained his undergraduate M.Chem. degree from the University of Southampton (2017). He joined the EPSRC Centre for Doctoral Training (CDT) in Catalysis and in 2018 joined the Fey group at the University of Bristol for his Ph.D., focusing on new workflows to computationally predict catalysts for homogeneous organometallic processes from calculated descriptors and mechanistic studies.

Natalie Fey is a senior lecturer at the University of Bristol. After completing a B.Sc. (chemistry and economics, 1997) and a Ph.D. (2001) at Keele University, she worked with Rob Deeth at Warwick University, then joined the University of Bristol in 2003, to work with Guy Orpen, Jeremy Harvey and later also Guy Lloyd-Jones. She was awarded an EPSRC Advanced Research Fellowship in 2007, appointed as lecturer in 2015 and promoted to in 2018. Her group use computational approaches as a driver for scientific discovery.

Acknowledgments

Ph.D. project funding (D.J.D.) from the EPSRC Centre for Doctoral Training in Catalysis (EP/L016443/1) is gratefully acknowledged. Aspects of this work have been funded by EPSRC (EP/E059376/1), AstraZeneca, CatSci, the Welsh Government, as well as seedcorn funding from Bristol's Jean Golding Institute, the University of York and the Dial-a-Molecule network. We thank our collaborators, most of them cited as co-authors in the references below, for fruitful discussions, and recognise Bristol's Centre of Computational Chemistry as our "work home" and source of computing power, still free at the point of access.

Supporting Information

Overview of descriptors, timelines and publications, as well as reproduction of all computational details. This material is available free of charge via the Internet at http://pubs.acs.org.

References

1. Fey, N.; Koumi, A.; Malkov, A. V.; Moseley, J. D.; Nguyen, B. N.; Tyler, S. N. G.; Willans, C. E., Mapping the properties of bidentate ligands with calculated descriptors (LKB-bid). *Dalton Trans.* **2020,** *49*, 8169-8178.
2. Lai, A.; Clifton, J.; Diaconescu, P. L.; Fey, N., Computational mapping of redox-switchable metal complexes based on ferrocene derivatives. *Chem. Commun.* **2019,** *55*, 7021-7024.
3. Fey, N.; Green, A. I.; Tinworth, C. P.; Warriner, S.; Nelson, A., Computational Mapping of Dirhodium(II) Catalysts. *Chem. Eur. J.* **2021,** *27*, 1-9. DOI: 10.1002/chem.202003801.
4. Jover, J.; Fey, N., The Computational Road to Better Catalysts. *Chem. Asian J.* **2014,** *9*, 1714-1723.
5. 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.
6. 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.
7. Foscato, M.; Jensen, V. R., Automated in Silico Design of Homogeneous Catalysts. *ACS Catal.* **2020,** *10*, 2354-2377.
8. Peng, Q.; Duarte, F.; Paton, R. S., Computing organic stereoselectivity – from concepts to quantitative calculations and predictions. *Chem. Soc. Rev.* **2016,** *45*, 6093-6107.
9. Fey, N.; Orpen, A. G.; Harvey, J. N., Building ligand knowledge bases for organometallic chemistry: Computational description of phosphorus(III)-donor ligands and the metal–phosphorus bond. *Coord. Chem. Rev.* **2009,** *253*, 704-722.
10. Fey, N., The contribution of computational studies to organometallic catalysis: descriptors, mechanisms and models. *Dalton Trans.* **2010,** *39*, 296-310.
11. Gillespie, J. A.; Dodds, D. L.; Kamer, P. C. J., Rational design of diphosphorus ligands – a route to superior catalysts. *Dalton Trans.* **2010,** *39*, 2751-2764.
12. Durand, D. J.; Fey, N., Computational Ligand Descriptors for Catalyst Design. *Chem. Rev.* **2019,** *119*, 6561–6594.
13. Engl, P. S.; Santiago, C. B.; Gordon, C. P.; Liao, W.-C.; Fedorov, A.; Copéret, C.; Sigman, M. S.; Togni, A., Exploiting and Understanding the Selectivity of Ru-N-Heterocyclic Carbene Metathesis Catalysts for the Ethenolysis of Cyclic Olefins to α,ω-Dienes. *J. Am. Chem. Soc.* **2017,** *139*, 13117-13125.
14. 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.
15. Jorner, K.; Brinck, T.; Norrby, P.-O.; Buttar, D., Machine learning meets mechanistic modelling for accurate prediction of experimental activation energies. *Chem. Sci* **2021**. DOI: 10.1039/d0sc04896h
16. Tolman, C. A., Steric effects of phosphorus ligands in organometallic chemistry and homogeneous catalysis. *Chem. Rev.* **1977,** *77*, 313-348.
17. Mardirossian, N.; Head-Gordon, M., Thirty years of density functional theory in computational chemistry: an overview and extensive assessment of 200 density functionals. *Mol. Phys.* **2017,** *115*, 2315-2372.
18. Jensen, K. P.; Roos, B. O.; Ryde, U., Performance of density functionals for first row transition metal systems. *The Journal of Chemical Physics* **2007,** *126*, 014103.

19. Fey, N.; Ridgway, B. M.; Jover, J.; McMullin, C. L.; Harvey, J. N., Organometallic reactivity: the role of metal–ligand bond energies from a computational perspective. *Dalton Trans.* **2011,** *40*, 11184-11191.
20. 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.
21. 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.
22. Rosales, A. R.; Wahlers, J.; Limé, E.; Meadows, R. E.; Leslie, K. W.; Savin, R.; Bell, F.; Hansen, E.; Helquist, P.; Munday, R. H.; Wiest, O.; Norrby, P.-O., Rapid virtual screening of enantioselective catalysts using CatVS. *Nature Catalysis* **2019,** *2*, 41-45.
23. Guan, Y.; Ingman, V. M.; Rooks, B. J.; Wheeler, S. E., AARON: An Automated Reaction Optimizer for New Catalysts. *Journal of Chemical Theory and Computation* **2018,** *14*, 5249-5261.
24. Hansen, E.; Rosales, A. R.; Tutkowski, B.; Norrby, P.-O.; Wiest, O., Prediction of Stereochemistry using Q2MM. *Acc. Chem. Res.* **2016,** *49*, 996-1005.
25. 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.
26. 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.
27. Fey, N.; Garland, M.; Hopewell, J. P.; McMullin, C. L.; Mastroianni, S.; Orpen, A. G.; Pringle, P. G., Stable fluorophosphines: predicted and realized ligands for catalysis. *Angew. Chem. Int. Ed. Engl.* **2012,** *51*, 118-22.
28. Evans, L. A.; Fey, N.; Harvey, J. N.; Hose, D.; Lloyd-Jones, G. C.; Murray, P.; Orpen, A. G.; Osborne, R.; Owen-Smith, G. J. J.; Purdie, M., Counterintuitive Kinetics in Tsuji-Trost Allylation: Ion-Pair Partitioning and Implications for Asymmetric Catalysis. *J. Am. Chem. Soc.* **2008,** *130*, 14471-14473.
29. 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.
30. Jover, J.; Fey, N.; Purdie, M.; Lloyd-Jones, G. C.; Harvey, J. N., A computational study of phosphine ligand effects in Suzuki–Miyaura coupling. *J. Mol. Catal. A: Chem.* **2010,** *324*, 39-47.
31. Fey, N., Lost in chemical space? Maps to support organometallic catalysis. *Chem. Cent. J.* **2015,** *9*, 38.
32. 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.
33. Townend, J., *Practical Statistics for Environmental and Biological Scientists*. John Wiley & Sons Ltd.: Chichester, 2002.
34. Livingstone, D., *A Practical Guide to Scientific Data Analysis*. Wiley & Sons Ltd.: Chichester, UK, 2009.
35. Fey, N.; Papadouli, S.; Pringle, P. G.; Ficks, A.; Fleming, J. T.; Higham, L. J.; Wallis, J. F.; Carmichael, D.; Mézailles, N.; Müller, C., Setting P-Donor Ligands into Context: An Application of the Ligand Knowledge Base (LKB) Approach. *Phosphorus, Sulfur, Silicon Relat. Elem.* **2015,** *190*, 706-714.
36. Stauffer, S. R.; Hartwig, J. F., Fluorescence Resonance Energy Transfer (FRET) as a High-Throughput Assay for Coupling Reactions. Arylation of Amines as a Case Study. *J. Am. Chem. Soc.* **2003,** *125*, 6977-6985.
37. McMullin, C. L.; Jover, J.; Harvey, J. N.; Fey, N., Accurate modelling of Pd(0) + PhX oxidative addition kinetics. *Dalton Trans.* **2010,** *39*, 10833-10836.
38. McMullin, C. L.; Fey, N.; Harvey, J. N., Computed ligand effects on the oxidative addition of phenyl halides to phosphine supported palladium(0) catalysts. *Dalton Trans.* **2014,** *43*, 13545-13556.
39. 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.
40. 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.
41. Newland, R. J.; Smith, A.; Smith, D. M.; Fey, N.; Hanton, M. J.; Mansell, S. M., Accessing Alkyl- and Alkenylcyclopentanes from Cr-Catalyzed Ethylene Oligomerization Using 2-Phosphinophosphinine Ligands. *Organometallics* **2018,** *37*, 1062-1073.

42. 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.
43. van Leeuwen, P. W. N. M.; Kamer, P. C. J.; Reek, J. N. H.; Dierkes, P., Ligand Bite Angle Effects in Metal-catalyzed C–C Bond Formation. *Chem. Rev.* **2000,** *100*, 2741-2770.
44. Jover, J.; Fey, N., Screening substituent and backbone effects on the properties of bidentate P,P-donor ligands (LKB-PPscreen). *Dalton Trans.* **2013,** *42*, 172-181.
45. 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.
46. Ciano, L.; Fey, N.; Halliday, C. J. V.; Lynam, J. M.; Milner, L. M.; Mistry, N.; Pridmore, N. E.; Townsend, N. S.; Whitwood, A. C., Dispersion, solvent and metal effects in the binding of gold cations to alkynyl ligands: implications for Au(i) catalysis. *Chem. Commun.* **2015,** *51*, 9702-9705.
47. Hare, S. R.; Tantillo, D. J., Cryptic post-transition state bifurcations that reduce the efficiency of lactone-forming Rh-carbenoid C–H insertions. *Chem. Sci* **2017,** *8*, 1442-1449.