

Development of a Ligand Knowledge Base, Part 1: Computational Descriptors for Phosphorus Donor Ligands

Natalie Fey,^[a] Athanassios C. Tsipis,^[a] Stephanie E. Harris,^[a] Jeremy N. Harvey,^{*,[a]}
A. Guy Orpen,^{*,[a]} and Ralph A. Mansson^[b]

Abstract: A prototype collection of knowledge on ligands in metal complexes, termed a ligand knowledge base (LKB), has been developed. This contribution describes the design of DFT-calculated descriptors for monodentate phosphorus(III) donor ligands in a range of representative complexes. Using the resulting data, a ligand space is mapped and predictive models are derived for metal complexes. Important characteristics, including chemical, computational and statistical robustness for the generation and exploitation of such an LKB are described. Chemical robustness ensures transferability of the descriptors, as well as comprehensive sampling of ligand space.

To make the calculations amenable to automation in an e-science setting, a reliable, well-defined computational approach has been sought from which the descriptors can be readily extracted. The LKB has been explored with multivariate statistical methods. Principal component analysis (PCA) is used for the mapping of chemical space, projecting multiple descriptors into scatter plots which illustrate the clustering of chemically similar ligands. Interpreta-

tion of the resulting principal components in terms of established steric and electronic properties and the importance of its statistical robustness to variations in the ligand set are discussed. Multiple linear regression (MLR) models have been derived, demonstrating the versatility of the descriptors for modeling varied experimentally determined parameters (bond lengths, reaction enthalpies and bond-stretching frequencies). The importance of re-sampling methods for testing the robustness of predictions is highlighted. A strategy for the construction of a robust LKB suitable for the modeling of ligand and complex behavior is outlined based on these observations.

Keywords: density functional calculations • ligand effects • P ligands • statistical analysis • stereoelectronic parameters

Introduction

Ligands play a central role in coordination and organometallic chemistry, and their steric and electronic properties are crucial for determining and controlling the structures, properties and functions of metal complexes. Considerable efforts have been directed at quantifying these properties (see, e.g., references [1–5]) to place ligand chemistry on a rational footing. Such work has a range of potential applications—for example to facilitate the screening of metallo-drugs and homogeneous catalysts, as well as the design of new metal complexes tailored to specific applications in synthesis and materials chemistry.

More generally, ligands are chemical entities that may be described by an interesting mix of essentially intrinsic or inherent properties (connectivity, geometry, energy and electronic structure) of the “free” or pro-ligand and those that emerge or are perturbed on complexation to the metal (including conformation, geometry, charge distribution etc).

[a] Dr. N. Fey, Dr. A. C. Tsipis,⁺ Dr. S. E. Harris, Dr. J. N. Harvey,
Prof. Dr. A. G. Orpen
School of Chemistry
University of Bristol
Cantock's Close, Bristol BS8 1TS (UK)
Fax: (+44) 117-925-1295
E-mail: jeremy.harvey@bristol.ac.uk
guy.orpen@bristol.ac.uk

[b] Dr. R. A. Mansson
School of Mathematics
University of Southampton
Highfield, Southampton SO17 1BJ (UK)

[⁺] Current Address:
Laboratory of Inorganic and General Chemistry
Department of Chemistry
University of Ioannina
Ioannina 451 10 (Greece)

Supporting information for this article is available on the WWW under <http://www.chemeurj.org/> or from the author: Details of LKB data, CSD searches, principal component analysis results and detailed linear regression models.

Furthermore, when bound they confer properties on the new entity formed, that is, the metal complex (e.g. solubility, reactivity, structure). For example: the behavior, including reactivity of the trifluoromethyl group, CF_3 , is markedly different when coordinated to a metal ($\text{M}-\text{CF}_3$) than when part of a fluorinated alkyl group ($\text{C}-\text{CF}_3$); the conformation of ethylenediamine (en, $\text{NH}_2(\text{CH}_2)_2\text{NH}_2$) is very much affected by chelate coordination; the geometry of triphenylphosphine (PPh_3) is dependent on the Lewis acid to which it is bound.^[6] It is clearly more challenging to provide a succinct and robust quantitative description of ligands and their properties (and those of their complexes), which is transferable and chemically useful, than it is for other chemical systems that do not show such strongly context dependent behavior.

Here we consider the characteristics of a collection of structured and validated knowledge on ligands and their properties, both inherent and dependent—what might be termed a ligand knowledge base (LKB). We also explore how an LKB might be constructed so as to span the full range of chemical (ligand) space, hence covering greater diversity in ligand structure and type than can be readily obtained by experimental studies. We seek to develop a protocol for the robust characterization of ligand properties, based on a trial with an important class of ligands (phosphorus(III) derivatives, see Figure 1). With this knowledge in hand we then develop statistical models that are able, amongst other things, to reproduce experimentally derived data on their (emergent or dependent) behavior in metal complexes.

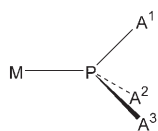


Figure 1. Metal-bound phosphorus(III) ligand, $\text{M}-\text{PA}_3$.

to reproduce experimentally derived data on their (emergent or dependent) behavior in metal complexes.

First let us consider what such an LKB might consist of and what it would allow its user to do.

- 1) A computationally derived LKB would contain robust descriptors of ligand properties. The robustness of the values of these descriptors should ideally mean both that they can be reliably obtained, that is, they are *computationally* robust, and that they can be transferred to a variety of circumstances, that is, they are *chemically* robust.
- 2) The ligand descriptors in the LKB should afford the chemist insight into the part of chemical space in which a given ligand is located—perhaps better termed its location in *ligand space*. A sufficient range, robustness and diversity of descriptors are required so that all the important properties, both inherent and otherwise, of the ligands are captured in the LKB.
- 3) The principal dimensions of this ligand space might ideally be related to concepts established in the literature of coordination chemistry (steric bulk, σ -donor ability, π -acidity) or related fields (e.g. physical organic chemistry).
- 4) The LKB would ideally allow the robust statistical modeling and prediction of a range of behavior of metal com-

plexes of the ligands in question, that is, give models that are *statistically* robust.

One way forward in studies of ligand chemistry is to conduct systematic, large-scale experimental studies of ligands, both as free (pro-)ligands and in a range of metal complexes, as well as their behavior in reactions. This generates empirical data, such as thermochemical, electrochemical or reactivity, which allows description and analysis of their properties. However, such studies are rarely undertaken (perhaps for reasons of expense), and the impact of varying experimental conditions can preclude the reliable interpretation of experimental data from different (literature) sources. Chemical diversity in experimental datasets is therefore typically rather limited.

Given the practical difficulties and expense of assembling an LKB from empirical data, alternative approaches are of interest. We are investigating an approach in which information from structural (crystallographic) and computational sources are combined to provide descriptors of ligand properties which can form a large-scale structured collection of knowledge about transition metal complexes. In developing and using such a knowledge base we seek eventually to lay the foundations for the exploitation of the growing availability of networked computational and data storage resources (so-called e-science and the Grid architecture).^[7] In its mature form, this LKB might combine structural and experimental information “mined” from available databases, such as the Cambridge Structural Database (CSD)^[8] with descriptors calculated using appropriate computational approaches, for example, density functional theory (DFT), for a wide range of ligands.

As a first step in developing an LKB, we have generated a collection of calculated structural and energetic descriptors for monodentate phosphorus(III) ligands. These ligands were chosen for their ubiquity in organometallic and coordination chemistry. The properties of the metal–phosphorus bond and the resulting complexes can be fine-tuned by modifying the substituents on the phosphorus to control the steric and electronic profile of each ligand. A range of experimental^[1–4,9–12] and computational^[5,13–19] approaches have been reported to describe and quantify these steric and electronic properties; some of these approaches have recently been reviewed.^[20] The resulting stereoelectronic parameters have been used to describe phosphorus donor ligands and to project their properties and hence map relationships in multidimensional chemical space. Independently, Bjørsvik^[19] and Cundari^[5] have reported how these maps may be used to identify ligand targets for screening experiments and catalyst design, with the former work by using PCA-derived variables based on calculated ligand descriptors while the latter is based on direct plots of descriptors calculated for rhodium complexes. In addition, a number of groups have sought to estimate the quantitative contributions of steric and σ -/ π -electronic effects by fitting regression models to various experimentally observed linear free energy relation-

ships and to explore their use in quantitative structure–property relationships.^[2,3,11,21,22]

The availability of a range of experimental data makes phosphorus donor ligands attractive for a proof-of-concept study such as this. We report below the design of calculated phosphorus(III) ligand descriptors and their statistical analysis, and describe potential applications in mapping chemical space as well as the interpretation and prediction of experimental data. On the basis of these observations we offer some conclusions describing a way forward for the design and development of a mature LKB.

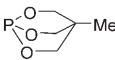

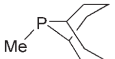
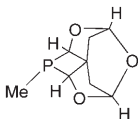
Results and Discussion

Knowledge base design

Ligands: Structural and electronic descriptors were calculated for 61 monodentate phosphorus(III) ligands (Table 1) with a range of substituents. The dataset contains symmetrical PA₃ species **1–33** (A = alkyl, aryl, halide, alkoxy, aryloxy, amino), and so includes substituted tertiary phosphines and phosphites, phosphine halides and aminophosphines. In addition, a range of simple asymmetric PA₂B species **34–57** (A, B = alkyl, aryl, halide, amino) and some more unusual examples where the phosphorus is incorporated into cage or bicyclic systems (**58–61**) were included. This ligand selection was designed to achieve optimal overlap with available experimental data and to sample chemical space for phosphorus(III) ligands widely, thereby improving the robustness of statistical analyses (as discussed below) and the chance of modeling a wider range of ligand and complex behavior.

Complexes: Structural and energetic parameters were calculated for the free phosphorus(III) species, L, and a range of their complexes. Protonated ligands ([HL]⁺) and borane adducts ([H₃B·L]) were chosen to investigate σ-electronic effects. Two metal complexes, square-planar [PdCl₃L][−] and tetrahedral [Pt(PH₃)₃L], were included as such species may involve contributions from both metal–L σ- and π-bonding

Table 1. Phosphorus(III) donor ligands in prototype ligand knowledge base.

No.	L = PA ₃	No.	L = PA ₂ B	
	A		A	B
1	H	34	F	H
2	Me	35	H	F
3	Et	36	Cl	H
4	Pr	37	H	Cl
5	<i>i</i> Pr	38	F	Me
6	Bu	39	Me	F
7	<i>t</i> Bu	40	Cl	Me
8	CF ₃	41	Me	Cl
9	Cy (C ₆ H ₁₁)	42	CF ₃	Me
10	Bz (CH ₂ Ph)	43	Me	CF ₃
11	F	44	<i>t</i> Bu	Me
12	Cl	45	Me	<i>t</i> Bu
13	OMe	46	Ph	Me
14	OEt	47	Me	Ph
15	OPh	48	Ph	Et
16	NH ₂	49	Et	Ph
17	NMe ₂	50	Ph	Cy
18	pyr (NC ₄ H ₄ , pyrrolyl)	51	Cy	Ph
19	NC ₄ H ₈	52	Ph	Pyr
20	pip (NC ₅ H ₁₀ , piperidyl)	53	Pyr	Ph
21	CHCH ₂	54	Ph	<i>o</i> -Me-Ph
22	Ph	55	<i>o</i> -Me-Ph	Ph
23	C ₆ F ₅	56	Ph	<i>o</i> -MeO-Ph
24	<i>o</i> -Me-Ph	57	<i>o</i> -MeO-Ph	Ph
25	<i>m</i> -Me-Ph	no.	L	
26	<i>p</i> -Me-Ph	58		
27	<i>o</i> -MeO-Ph	59		
28	<i>p</i> -MeO-Ph	60		
29	3,5-(F ₃ C) ₂ -Ph	61		
30	<i>p</i> -F ₃ C-Ph			
31	<i>p</i> -F-Ph			
32	<i>p</i> -Cl-Ph			
33	<i>p</i> -Me ₂ N-Ph			

interactions. In addition, structural *cis* and *trans* influences^[28] can be monitored for the palladium complexes.¹ Table 2 summarizes the calculated descriptors and the complete LKB is included in the Supporting Information (Table S1).

¹ In preliminary work, the set of LKB descriptors also included those from octahedral molybdenum complexes [Mo(PH₃)₃L], where the electron-rich Mo⁰ d⁶ metal centre was expected to donate π-electrons to appropriate ligands, L. However, the descriptors arising from these complexes were found to correlate very highly with steric parameters and the optimized geometries displayed considerable steric hindrance due to the square-pyramidal {Mo(PH₃)₃} fragment. These steric interactions seem to dominate the structure of the complexes, and (weaker) electronic effects are apparently masked and indeed a number of complexes {Mo(PH₃)₃L} are not bound. Given this non-robust behavior all data derived from these complexes were therefore excluded from the analysis reported below.

Table 2. Calculated descriptors in prototype ligand knowledge base (LKB).

Descriptor ^[a]	Derivation (Unit)	Mean	σ	Range	Used?
free phosphorus(III) species (L)					
E_{HOMO}	energy of highest occupied molecular orbital [Hartree]	-0.1996	0.0310	-0.2819--0.1417	yes
E_{LUMO}	energy of lowest unoccupied molecular orbital [Hartree]	-0.0296	0.0396	-0.1087--0.0327	yes
$Q(\text{P})$	NBO charge on P in L	0.94	0.30	-0.03--1.63	no ^[b]
LP s-character	contribution of P s-orbital to lone pair (LP), from NBO analysis (%)	57.5	7.5	49.0--80.3	yes
$\text{He}_8\text{-steric}$	interaction energy between L in ground state conformation and ring of 8 Helium atoms, $E_{\text{ster.}} = E_{\text{tot}}(\text{system}) - [E_{\text{tot}}(\text{He}_8) + E_{\text{tot}}(\text{L})]$ [kcal mol ⁻¹] (Figure 2)	7.2	5.7	0.8--29.8	yes
protonated ligand ([HL] ⁺)					
PA	proton affinity, $\text{PA} = E(\text{L}) - E([\text{HL}]^+)$ [kcal mol ⁻¹]	227.8	24.2	163.6--264.8	yes
$Q(\text{P,prot})$	NBO charge on P in [HL] ⁺	1.43	0.36	0.45--2.28	no ^[b]
$Q(\text{H,prot})$	NBO charge on H in [HL] ⁺	0.06	0.03	-0.02--0.14	no ^[c,d]
$\Delta\text{P-A}(\text{H})$	change in av. $r(\text{P-A})$ compared with free ligand, L [Å]	-0.057	0.024	-0.115--0.023	no ^[d]
$\Delta\text{A-P-A}(\text{H})$	change in av. $\angle(\text{A-P-A})$ cf. L [°]	11.0	1.8	6.5--17.0	no ^[d]
P-H	$r(\text{L-H})$ in [HL] ⁺ [Å]	1.416	0.005	1.399--1.427	no ^[c,d]
borane adduct (H ₃ B·L)					
Q(B fragm.)	NBO charge on BH ₃ fragment	-0.65	0.04	-0.72--0.52	yes
BE(B)	bond energy for dissociation of P-ligand from BH ₃ fragment [kcal mol ⁻¹] ^[e]	35.5	4.3	23.4--41.7	yes
$\Delta\text{P-A}(\text{B})$	change in av. $r(\text{P-A})$ cf. L [Å]	-0.022	0.009	-0.037--0.009	yes
$\Delta\text{A-P-A}(\text{B})$	change in av. $\angle(\text{A-P-A})$ cf. L [°]	3.4	1.2	0.1--6.3	yes
P-B	$r(\text{P-B})$ [Å]	1.927	0.026	1.861--1.977	yes
B-H	av. $r(\text{B-H})$ [Å]	1.221	0.002	1.216--1.224	no ^[c]
H-B-H	av. $\angle(\text{H-B-H})$ [°]	113.6	0.7	112.3--115.6	no ^[c]
palladium complexes ([PdCl ₃ L] ⁻)					
Q(Pd fragm.)	NBO charge on [PdCl ₃] ⁻ fragment	-1.24	0.06	-1.37--1.06	yes
BE(Pd)	bond energy for dissociation of L from [PdCl ₃] ⁻ fragment [kcal mol ⁻¹] ^[e]	34.6	5.1	22.5--48.3	yes
$\Delta\text{P-A}(\text{Pd})$	change in av. $r(\text{P-A})$ cf. L [Å]	-0.006	0.011	-0.026--0.023	yes
$\Delta\text{A-P-A}(\text{Pd})$	change in av. $\angle(\text{A-P-A})$ cf. L [°]	0.8	1.7	-3.2--5.4	yes
P-Pd	$r(\text{P-Pd})$ [Å]	2.277	0.039	2.188--2.418	yes
Pd-Cl <i>cis</i>	$r(\text{Pd-Cl})$, <i>cis</i> to L [Å]	2.385	0.006	2.372--2.404	no ^[c]
Pd-Cl <i>trans</i>	$r(\text{Pd-Cl})$, <i>trans</i> to L [Å]	2.366	0.013	2.334--2.392	yes
platinum complexes ([Pt(PH ₃) ₃ L])					
Q(Pt fragm.)	NBO charge on [(PH ₃) ₃ Pt] fragment	0.01	0.07	-0.10--0.26	yes
BE(Pt)	Bond energy for dissociation of P ligand from [Pt(PH ₃) ₃] fragment [kcal mol ⁻¹] ^[e]	16.4	4.1	6.2--24.2	yes
$\Delta\text{P-A}(\text{Pt})$	change in av. $r(\text{P-A})$ cf. L [Å]	0.002	0.011	-0.017--0.043	yes
$\Delta\text{A-P-A}(\text{Pt})$	change in av. $\angle(\text{A-P-A})$ cf. L [°]	-0.04	1.7	-6.0--3.4	yes
P-Pt	$r(\text{P-Pt})$ distance [Å]	2.320	0.038	2.234--2.390	yes
H ₃ P-Pt	av. $r(\text{H}_3\text{P-Pt})$ [Å]	2.325	0.006	2.312--2.339	no ^[c]
$\angle(\text{H}_3\text{P})\text{Pt}(\text{PH}_3)$	av. $\angle(\text{H}_3\text{P})\text{-Pt}(\text{-PH}_3)$ [°]	108.0	1.0	105.4--110.6	yes
cumulative					
S4' calcd	$(\Sigma \angle \text{APA} - \Sigma \angle \text{ZPA})$, where Z = BH ₃ , [PdCl ₃] ⁻ , [Pt(PH ₃) ₃] [°]	38.4	11.1	7.8--65.9	yes

[a] All calculations were performed on isolated molecules. [b] Large range of values, no clear trend in data, see text for discussion. [c] Small range of values, see text for discussion. [d] Highly correlated with H₃B·L descriptors, see text for discussion. [e] $\text{BE} = [E_{\text{tot}}(\text{fragment}) + E_{\text{tot}}(\text{L})] - E_{\text{tot}}(\text{complex})$.

Computational requirements: From a technical point of view, the computational method chosen to calculate descriptors for any LKB must fulfill three main criteria: good reliability, relatively low computational expense and scope for automation of calculations.

The performance of any chosen computational approach should be reliable, in order to prevent the occurrence of random errors in the data set produced. Methods that yield data disagreeing significantly with relevant experimental data in some but not all cases should thus be avoided if no chemical explanation can be given. We have used the BP86 density functional, because it has been shown to give reasonably good structural and energetic performance in the description of organometallic complexes^[29] and it is unlikely

that significant errors would occur for individual structures within a relatively homogenous class of metal-phosphine complexes. Other gradient-corrected or hybrid density functionals may have been equally suitable, because *systematic* deviations of structural or energetic parameters, as often observed for these functionals when compared to experimental data, do not affect the results of the statistical analyses and are thus negligible in a large data set.

Given the large number of species to be studied, the calculation of descriptors should be relatively computationally inexpensive (at least by present standards). More importantly, they must be easily performed and analyzed, that is, the required geometric or energetic information readily extracted, in a standard way with zero (or minimal) human inter-

vention. This will aid in future data generation and analysis in an e-science/Grid framework, where multiple calculations can be distributed across a network and information extraction must be automated.

For now, more complex computational data, for example frequency calculations, have been avoided, even though established (experimental) electronic parameters for phosphorus donor ligands are often based on carbonyl stretching frequencies (see for example references [1–4, 10, 16]). Frequency calculations are significantly more computationally expensive than geometry optimizations and may require inspection of the output to identify the correct result, especially if little or no molecular symmetry is found and so vibrational modes are highly mixed. We have also avoided more complex properties such as the minimum in the electrostatic potential proposed by Koga et al.^[18] and parameters derived from energy decomposition analysis.^[17] In addition, we have not performed extensive conformational searches, as these are too expensive to be attempted at DFT level. Furthermore, the energetic ranking derived from cheaper approaches such as molecular mechanics (MM) is potentially unreliable^[30] and the analysis of multiple conformers would be difficult to automate. Some or all of these constraints might be removed in a later version of a LKB, most likely in response to a lowering of computational cost, that is, due to improvements of data management software as well as hardware, both targeted by the development of the Grid architecture. As shown in Table 2, all descriptors in this prototype LKB can either be extracted directly from geometry optimizations or can be derived by simple mathematical operations, thus making the calculations amenable to automation.

Descriptors: The calculated descriptors (Table 2) include:

- 1) frontier molecular orbital energies of the free ligands,
- 2) ligand proton affinities,
- 3) adduct binding energies,
- 4) ligand and metal fragment charges,
- 5) a range of structural parameters describing geometry changes of both ligand and metal fragments upon complexation,
- 6) two measures of ligand steric bulk, the $S4'$ parameter^[5,6] and an energetic measure of steric bulk, He_8 _steric (see below).

To facilitate the development of linear regression models for experimental data, these descriptors are linearly related to energy, either directly (orbital energies, binding energies, proton affinity) or indirectly. For example, structural changes on complexation can be expressed as a perturbation from the ideal geometry of a free ligand. We have not included Tolman's cone angle (θ)^[1] of the ligands as a descriptor, both because it is difficult to compute automatically^[5] and because its relationship to energy cannot be readily established (see below). Instead, we have developed a new steric parameter (termed He_8 _steric), calculated as the inter-

action energy between the phosphorus(III) ligand and a ring of eight helium atoms. The helium atoms are held in regular, fixed positions on a circle of radius 2.5 Å. The phosphine geometry is re-optimized in the presence of this He_8 ring, starting from an optimized conformation of the free ligand, with the phosphorus atom constrained to lie exactly 2.28 Å^[1] above the ring centroid (Figure 2) along the perpendicular

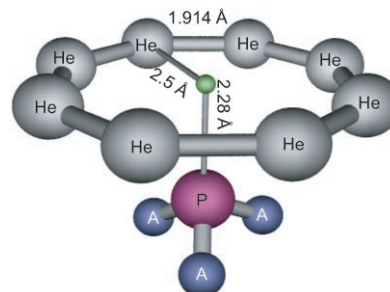


Figure 2. Geometry used for computation of the He_8 _steric parameter: the interaction energy between the phosphorus ligand (PA_3) and a ring of eight helium atoms.

to its plane. This arrangement seeks to mimic the non-bonded, closed-shell/closed-shell, interactions of a phosphine with, for example, the *cis* ligands in an octahedral complex. Given the van der Waals radii of He and P (1.4 and 1.8 Å)^[31] and the He...P distance in this model (3.383 Å) the lone pair of the phosphorus does not interact strongly with the He_8 ring, so that only substituent steric effects contribute significantly to the interaction energy. This is confirmed by the small values of this parameter calculated for the smallest phosphine **1** (2.3 kcal mol⁻¹) and its halide derivatives, for example, ligands **34–37** (range 1.6–1.9 kcal mol⁻¹), as well as the parent ligands trifluorophosphine **11** (1.5 kcal mol⁻¹) and trichlorophosphine **12** (2.2 kcal mol⁻¹, see Tables 2 and S1 for value range and full dataset respectively).

Descriptor reduction: In the first instance, it is appealing simply to compute and include a wide range of chemically varied parameters to describe ligand properties. However, the use of a substantial number of highly correlated descriptors can be problematic in applying multivariate analysis techniques, one of the intended uses of the LKB. For example, in regression models based on such data the estimates of coefficients may be unreliable and hence not robust.^[32] Descriptive statistics (mean, standard deviation, range, frequency distributions) were therefore used to assess the suitability of individual descriptors for inclusion in the LKB. In particular, those descriptors with either very large or very small value ranges were excluded, because of their poor robustness and unresponsiveness respectively. Several variables associated with structural changes in the metal fragment (e.g. H_3P-Pt , $H-B-H$, see Tables 2 and S1) were excluded from further analysis, because their range of values was so small (0.027 Å and 3.3°, respectively) as to make them insensitive to variation of ligand. Indeed it is likely that for these descriptors computational “noise” might be

significant on the scale of ligand effects. Thus comparison with analogous ligand-centered quantities, for example Pt–P distance and A–P–A angle in BH₃·L adducts, which have ranges of 0.156 Å and 10.9°, respectively, suggests that some of these Lewis acid fragments do not show a significant structural response to variations in the phosphorus substituents.

Pearson correlation coefficients were computed to quantify the (linear) relationship between pairs of descriptors and with a view to reducing the number of descriptors required in the LKB. Structural variables of the same kind, for example, changes in bond length and angles on complexation, were found to be highly correlated for chemically similar complexes. For example, reasonably high correlation coefficients were observed for NBO fragment charges (Q , $r=0.804$) and bond angle changes on complexation ($\Delta A-P-A$, $r=0.729$) for the σ -bound adducts BH₃·L and [HL]⁺. This suggests that σ -bonding could be well represented by the BH₃·L adduct data alone and, with the exception of proton affinity, the descriptors for the protonated ligand [HL]⁺ were therefore excluded from further analysis (without substantial loss of information content).

For the NBO charges calculated for the phosphorus atom in both free and protonated [HL]⁺ ligands ($Q(P)$, $Q(P,prot)$), no significant linear relationship with other measures of ligand electronic properties, such as E_{HOMO} , E_{LUMO} and PA, could be identified. These two sets of NBO charges on P seem unable to capture the electronic effects of substituent variations, notably in response to changes in the *para*-substituent of arylphosphines (**26**, **28**, **30–33**, Table S1), on the electronic properties of the ligands and were excluded from further analysis, because they are apparently not computationally robust.

Contextualization: Exploratory graphical data analysis and regression modeling can be used to relate the new descriptors to more familiar ligand parameters from the literature (a process one might term contextualization). The relationship between the values of the original Tolman cone angle $\theta^{[1]}$ and the He₈_steric parameter is shown in Figure 3. While the Pearson linear correlation coefficient is quite high ($r=0.861$), fitting a simple linear regression equation ($R^2=0.742$) gives rise to physically unrealistic negative He₈_steric values for phosphine **1** and the phosphite cage ligand **58**, {P(OCH₂)₃CMe}. The data are better described by a cubic function ($R^2=0.869$) as indicated in Figure 3. It is worth noting that exclusion of the most hindered ligands, tris(*tert*-butyl)phosphine (**7**) and tris(*ortho*-tolyl)phosphine (**24**) from a linear regression fit gives only slightly improved

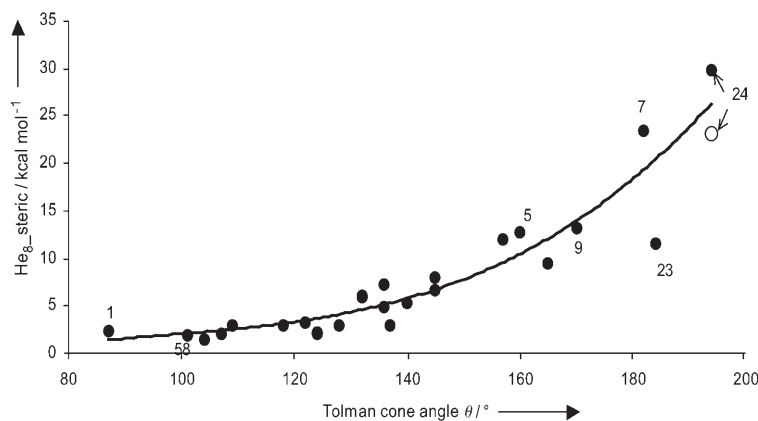


Figure 3. Plot of Tolman cone angle versus He₈_steric descriptor. Open circle refers to alternative ligand conformation; see text for discussion and Figure S1 for plot showing all ligand numbers. Nonlinear relationship illustrated by cubic function, $y=0.00002x^3-0.0061x^2+0.6347x-21.623$, $R^2=0.869$.

agreement ($R^2=0.782$), but poorer fits for the bulkier ligands tris(isopropyl)phosphine (**5**) and tris(cyclohexyl)phosphine (**9**) than observed for the cubic fit.

As shown in Figure 3, the sterically demanding ligands **7** (PtBu₃) and **24** (P(*o*-tolyl)₃) have higher He₈_steric parameters than would be predicted from the corresponding cone angles, whereas that of tris(pentafluorophenyl)phosphine (**23**, P(C₆F₅)₃) is lower. The deviation observed for ligand **24** may be explained by conformational differences. Tolman minimized the cone angle by assuming the least bulky conformation, that is, folding all substituents away from the metal.^[1] Later work (discussed for example in references [12,14 and 33]) has aimed to improve agreement between Tolman's cone angles and those measured from crystal structures, where alternative ligand conformations with larger cone angles are observed. As stated previously, we have used an optimized conformation of the free ligand as the input geometry for the calculation of the He₈_steric parameter. Even though the ligand geometry in the He₈·L species was re-optimized, significant conformational relaxation did not always occur, substituent rotation often being restricted. A survey of the *o*-tolyl ring conformations in crystal structures involving ligand **24**^[34] confirms that two conformer types (ggg/exo3 and gga/exo2, see reference [34] and references therein) are adopted in different coordination environments (see Table S2). The He₈_steric parameter was calculated for both conformers. The value for the conformer most commonly observed in sterically hindered complexes (gga/exo2, open circle, Figure 3, see Table S2 and reference [34]), is substantially lower.² A summary of relevant CSD reference codes and torsion angles for crystal structures of

² The relationship between the He₈_steric parameter and the MM-calculated ligand repulsive energy parameter developed by Brown^[13,14] can be described reasonably well by a linear function ($R^2=0.873$), if the He₈_steric value for tris(*o*-tolyl)phosphine ligand **24** is treated as an outlier ($R^2=0.742$, if included). This observation, as well as the effect of different conformational preferences in different coordination environments, will be discussed in detail elsewhere.

free ligands and representative complexes of phosphine **24** may be found in Table S2 (Supporting Information).

Complexes of the conformationally rigid tris(*tert*-butyl)-phosphine ligand **7** often have longer metal–phosphorus bond lengths than those observed for smaller ligands (see Table S3 for comparison of representative M–L distances observed in the CSD for ligands **2** (PMe₃) and **7** (PtBu₃)), presumably in response to the steric hindrance encountered in this ligand. This structural relaxation is not possible in the He₈-steric model, because the ring centroid–phosphorus distance is fixed (as it is also in the Tolman cone angle protocol). It seems likely that in this case the calculated value of the He₈-steric parameter exceeds that expected from the cone angle because the energetic measure takes better account of highly unfavorable interactions with other ligands. The apparently low value of He₈-steric for the conformationally flexible ligand tris(pentafluorophenyl)phosphine (**23**, P(C₆F₅)₃) is puzzling but may have its origin in an artificially high cone angle value (184°, cf. 145° for PPh₃).

While the effect of conformational changes is considerable for the bulky *o*-tolyl ligand **24**, amounting to an energy difference of 6.7 kcal mol⁻¹, most ligands in this version of the LKB are sterically less hindered and we estimate that this “conformational noise” is unlikely to exceed, on average, 2 or 3 kcal mol⁻¹. Work is currently under way to explore the effect of conformational variability on the calculated descriptors and how this may be captured by suitable descriptors in future versions of the LKB.

Multivariate data analysis

The collection of ligands and descriptors in this phosphorus(III) donor LKB can be investigated using a variety of multivariate statistical methods.^[35,36] Our choice of methodology has been determined by the characteristics of the data, consisting of relatively few ligands and a rather wide range of descriptors, which are often quite highly correlated. Although future extensions of the LKB will increase the diversity of ligands investigated, they are likely to fall into chemically distinct subsets, so these statistical analysis protocols will continue to be relevant. We have identified two distinct aims: mapping of chemical (ligand) space and property interpretation/prediction. The following section is structured accordingly.

Mapping chemical space: Principal component analysis (PCA) typically simplifies a multidimensional set of descriptors to a few derived variables (the principal components or PCs) that capture a large proportion of the variation in the data set. These PCs are orthogonal and are linear combinations of the original descriptors. The distribution of ligands in multidimensional descriptor space can be projected to fewer dimensions in pairwise plots of their values on the resulting PCs. Such plots can be used to identify clustering of subsets of ligands in the knowledge base. Since estimates of regression coefficients may be unreliable if highly correlated descriptors are used (see above),^[32] regression on the or-

thogonal PCs arising from PCA may be preferable when many descriptor variables are considered to be important. In such principal component regression (PCR), the PCs are used to build regression models with the aim of providing a good approximation of the relationship between a response variable and a small number of these PCs (which carry information from a larger number of the original descriptors). However, in both the projection and regression applications, interpretation and contextualization of PCs may be problematic when they are composed of many of the original descriptors (see below), and following PCA with PC rotation may be useful in interpreting important contributions.

The correlation matrix of the descriptor variables detailed in Table 2 was used in PCA followed by Varimax rotation. This technique reduces the number of descriptors in each PC in favor of those with large contributions, giving the resulting PCs a simpler structure while leaving them orthogonal.^[36] A score plot of the first two PCs is shown in Figure 4 and a more detailed summary of the PCA results is reported in the Supporting Information (Tables S4–S6).

The first two PCs describe some 67% of the variation in the descriptor dataset and the third PC a further 13% (see Table S4). The twodimensional map of phosphorus ligand space derived from PCs 1 and 2 (Figure 4) shows clusters of ligands corresponding to chemically familiar subsets and offers an appealing insight into ligand classification and similarity. Thus the arylphosphines appear grouped around PC1=1, PC2=0.5; the trialkylphosphines around PC1=0, PC2=-1 etc. This map is useful in locating unusual trialkylphosphine ligands (such as the adamantane-derived ligand **61**) as being similar to more familiar ligands (notably alkyl and aryl phosphites) in neighboring regions of ligand space. Most interestingly, the rhodium complex of ligand **61** is an active hydroformylation catalyst^[37]—not a characteristic of orthodox trialkylphosphines, but commonplace for trialkylphosphites. In addition, the consequences of systematic ligand variations can be visualized (e.g. P(Pyr)₃ (**18**), P(Pyr)₂Ph (**53**), P(Pyr)Ph₂ (**52**)). A plot of the first two PCs can also be used to identify where in chemical space new ligands would occur, even if experimental data are not available, and to determine their proximity to current ligands.

In contrast to previous stereoelectronic maps,^[1,5] which explicitly plot a steric versus an electronic parameter and do not consider properties of ligand complexes, the PC maps shown in Figure 4 are based on the projection of a large number of variables to pair-wise plots of PCs, that is, of linear combinations of multiple descriptors (see reference [22] for a similar approach to that used here but using free ligand-only properties). Our approach allows for a more objective capture of the properties of ligands in a range of chemically different coordination environments. The reasons for a ligand adopting a particular location in ligand space can be investigated by further analysis of the individual descriptors contributing to each PC.

The use of PCA for reducing multidimensional descriptor space to a small set of linear combinations of descriptors is appealing, but it is also interesting to try to associate the re-

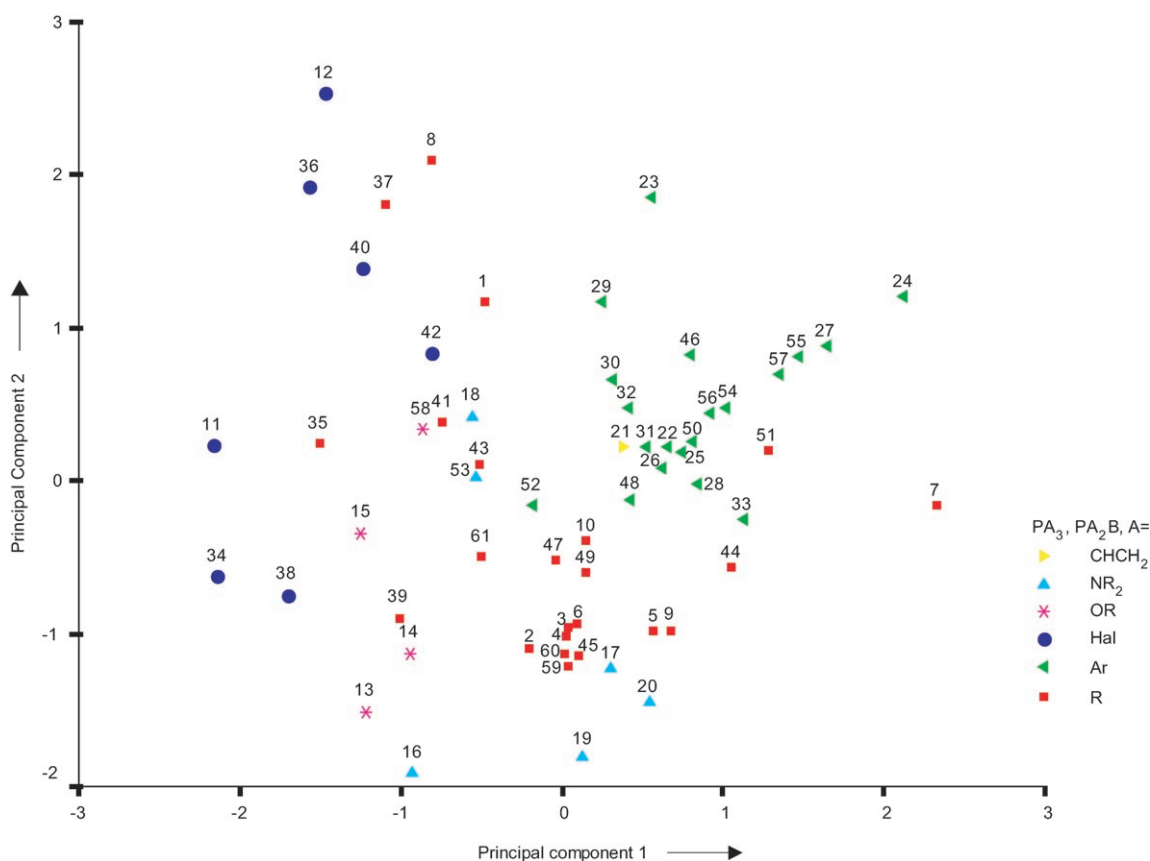


Figure 4. Principal component score plot (PC1 vs PC2) for all ligands in LKB. (Mixed ligands PA₂B are denoted by the same symbol as the analogous PA₃ species. See Supporting Information for larger plot (Figure S2).

sulting PCs with familiar steric and electronic properties. Inspection of the descriptor loadings on the rotated PCs (Tables 3 and S6) suggests that the first PC consists mainly of descriptors associated with steric and σ -electronic effects and the second PC can perhaps be interpreted in terms of π -electronic effects. The first two PCs therefore appear to correspond to some extent to established stereoelectronic parameters and to explain about two thirds of the variation in the present dataset (Table S4). Interpretation of the third PC is less obvious, but it is notable that it is mainly derived from the changes in A-P-A angle on complexation to B, Pd or Pt in adducts. It apparently records an aspect of the sensitivity of a ligand to the stimulus provided by binding to a Lewis acid, and so perhaps a facet of their behavior that emerges on complexation.

In general, interpretation of PCs is problematic and not statistically robust, primarily because PCA is based on the variance of descriptors and so is in turn sensitive to outlier values.^[38] For example, if we change the subset of ligands that is used, it is desirable that the PC composition remains reasonably consistent. This is important when developing PCR models for the interpretation of experimental data, particularly if these data are not available for all ligands in the LKB.

The correlation matrix on which PCA is based is derived from the descriptor covariance matrix by standardization of

Table 3. Principal component loadings (Varimax rotation), analysis for all ligands in LKB. (Descriptors with contributions $< |0.3|$ are not displayed, see Table S6 for full results.)

Descriptor	PC1	PC2	PC3
E_{HOMO}	0.718	-0.552	
E_{LUMO}		-0.800	
PA	0.811	-0.502	
LP s-character	-0.863		
Q(B fragm.)		0.894	
Q(Pd fragm.)		0.871	-0.313
Q(Pt fragm.)	-0.399	0.811	
BE(B)		-0.856	
BE(Pd)	-0.536	-0.554	
BE(Pt)	-0.755		-0.356
$\Delta\text{P-A(B)}$	0.501	0.529	
$\Delta\text{P-A(Pd)}$	0.417	0.762	
$\Delta\text{P-A(Pt)}$		0.780	
$\Delta\text{A-P-A(B)}$			0.809
$\Delta\text{A-P-A(Pd)}$			0.926
$\Delta\text{A-P-A(Pt)}$	0.432		0.794
P-B	0.925		
P-Pd	0.935		
P-Pt	0.924		
He_8steric	0.841		
S4' calcd	-0.822	0.300	
Pd-Cl <i>trans</i>	0.599	-0.745	
$\chi(\text{H}_3\text{P})\text{Pt}(\text{PH}_3)$		0.737	

the data. Standardization is useful in order to remove the effects of scale and unit (e.g. angles vs degrees vs kcal mol^{-1}) from the data set. It does, however, leave outliers and their potentially distorting effects in the data set. In addition, standardization can accentuate the noise content of a dataset if essentially invariant descriptors are retained (i.e., those termed unresponsive above). As noted above, such descriptors were eliminated from the dataset, for exactly this reason.

We have therefore tested the statistical robustness of our PCA results by repeating the analysis for i) randomly selected subsets of ligands and ii) some chemically defined subsets, for example, arylphosphines. The order and composition of the PCs does indeed change for these subsets, and this variation becomes more pronounced when Varimax rotation is applied. It was this lack of robustness that led to the inclusion of a group of mixed ligands of the form PA_2B in order to better sample chemical space. Their inclusion has to some extent improved the robustness of the PCA results (both chemically and statistically). However, *quantitative* interpretations of the principal components remain dubious and use of a statistically more robust version of PCA, where the impact of outliers has been reduced by robust estimation of the correlation matrix, may be beneficial.^[38,39]

Property interpretation and prediction: Given Figure 4, the descriptors seem to capture chemically intuitive ligand similarities and so provide a useful qualitative approach for visualizing chemical space. The LKB data were explored in greater detail by considering multiple linear regression (MLR) models for experimentally measured quantities using the calculated descriptor values. These MLR models were used to investigate whether a range of rather different experimental variables can be described by suitable linear functions of ligand descriptors. In addition, the application of these models to the prediction of experimental data for new ligands was assessed.

Many of the descriptors in the LKB are correlated (see above) and thus different regression models, using subsets of descriptors, can be derived. These models might have similar performance when assessed according to the successful description of the relationship between a response variable and a set of descriptors, that is, all models have similar regression coefficients R^2 , with values close to 1. The complexity (i.e., dimensionality) of the model will depend on the variable selection procedure used, for which various manual and automatic procedures are known. These are generally based on evaluating regression diagnostics for different models derived from a given set of descriptors so as to balance model complexity (having fewer descriptors typically leads to improved transferability and simplifies interpretation) and performance (having more descriptors allows better approximation of response variable).

The identification of a “best” model further depends on whether this model will be used in the interpretation of observed data or for the prediction of unknowns. In the former case, the regression coefficient, R^2 , and the adjusted R^2

value indicate how well the experimental data is described by the model. The adjusted R^2 statistic gives a better account of the balance between model complexity and quality, unlike the standard R^2 which usually increases upon using more descriptors. The quality of a model fit can further be assessed by estimating the prediction error as the mean squared residual and by diagnostic plots. While plotting observed versus predicted data can be used to confirm the successful description of the experimental data by an MLR model, if data points are clustered around the diagonal individual deviations become much clearer in a plot of predicted versus residual error values.

These diagnostics give no indication of the predictive capabilities of the models, or indeed of their robustness to variations in the ligand set. To investigate whether an MLR model is useful for estimating experimental parameters for new ligands, additional diagnostic statistics have to be considered. In a large database, the response data can be split into representative training and test sets to establish the quality and robustness of model predictions. However, when there is a limited amount of data, models may be extrapolating, so re-sampling methods such as cross-validation should be used.^[40,41] Prediction errors were estimated here using 10-fold cross-validation^[40,42] and bootstrapping.^[43] When using cross-validation to assess a model, we fit to a subset of the original data and make predictions for the cases (ligands) that were excluded.^[40,42] In the bootstrap re-sampling method,^[43] a random sample is drawn from the original data to generate a new data set of the same size as the original. This is achieved by replacement, that is, by allowing multiple occurrences of the same sample. A regression model is then fitted to this bootstrap sample, but the model is validated by making predictions for the original sample. This approach can be used to mimic variation in the original data to obtain a further measure of the predictive power of a given model.

To illustrate this application of a LKB, three examples of MLR models for experimental predictors are shown (Tables 4, S7 and S8), with diagnostic plots (fits and residuals) shown in Figure 5. These examples were selected to illustrate the application of this prototype LKB for the MLR-based interpretation and prediction of a range of experimentally determined parameters. It is notable that these parameters span a considerable range of information types, being geometric (describing molecular structure), energetic (describing reaction thermodynamics) and “electronic” (actually reporting vibrational behavior), respectively. More advanced variable selection and model evaluation methods, as well as protocols for robust parameter estimation and determining appropriate model complexity, will be evaluated and discussed in detail in due course.^[44]

Table 4 summarizes the experimental data and lists the descriptors used in the regression models as well as giving representative diagnostic data. The full models are included in the Supporting Information (Tables S7 and S8). Overall, these models provide a good description of the relationship between the experimental data and the calculated descrip-

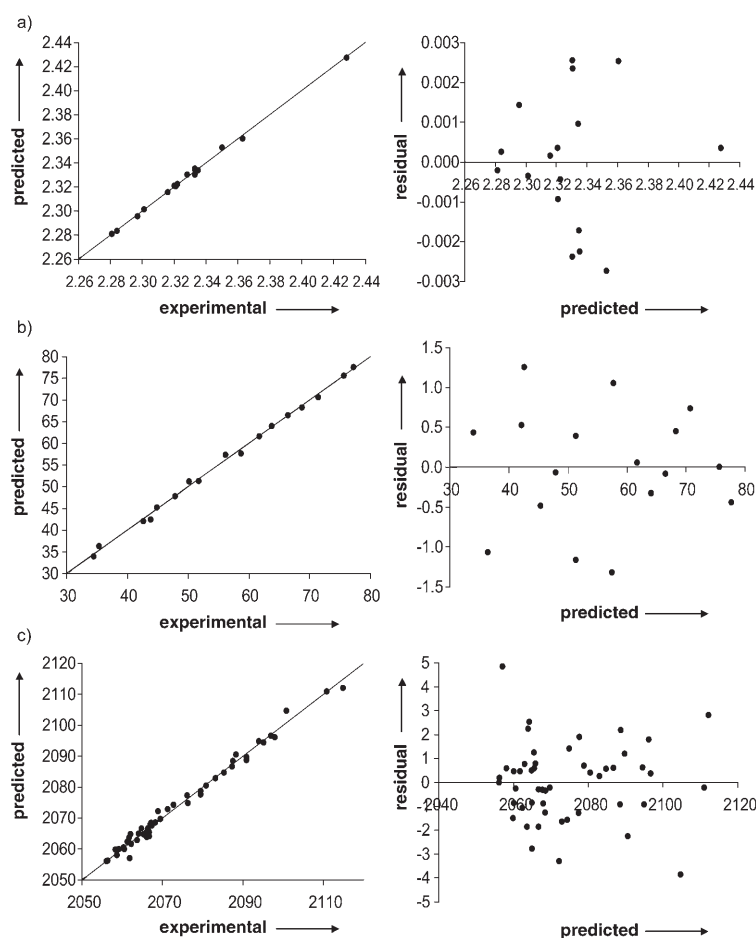


Figure 5. Diagnostic (fit and residual) plots for multiple linear regression models; a) Rh–P distance [Å] in four-coordinate, square-planar Rh^I complexes with a phosphorus ligand *trans* (from CSD survey, Table S9), b) ΔH_{rxn} [kcal mol⁻¹], for $[\text{Rh}(\text{CO})_2\text{Cl}]_2 + 4\text{PX}_3 \rightarrow 2\text{Rh}(\text{CO})(\text{Cl})(\text{PX}_3)_2 + 2\text{CO}$,^[2] c) Tolman electronic parameter, $A_1 \nu(\text{CO})$ in $[(\text{CO})_3\text{NiL}]^{\text{I}}$ [cm⁻¹]. (Full models are summarized in the Supporting Information (Tables S7 and S8).)

tors, with regression coefficients close to 1. This is further illustrated by small estimated prediction errors (Table 4) and the scatter of residuals in the diagnostic plots (Figure 5). The models for P–Rh and ΔH_{rxn} in particular demonstrate that the LKB descriptors can be used to derive linear models, which reproduce the experimental data closely, with residuals at about the level of experimental noise. While it might be argued that these models are over-fitted with six descriptors and only 17 experimental observations (Table S8), they have been chosen to illustrate this potential application of a LKB and as yet optimizing model complexity has been of secondary importance. For both P–Rh and ΔH_{rxn} , the descriptors used in the models (Table 4 and S7) could be interpreted as measures of steric (S_4' calcd) and σ -/ π -electronic effects, with PA and LP s-character indicative of σ -bonding and E_{LUMO} related to π -bonding, whereas the Pd- and Pt-derived descriptors ($\Delta\text{P-A}(\text{Pd})$, $\Delta\text{A-P-A}(\text{Pd})$, P–Pd, BE(Pd), Q(Pt fragm.), $\chi(\text{H}_3\text{P})\text{Pt}(\text{PH}_3)$) seem to in-

Table 4. Multiple linear regression models using LKB data.^[a]

Experimental variable	$N^{\text{[b]}}$	Mean	STD	Descriptors in model	R^2 (adj. R^2) ^[c]	Estimated prediction errors ^[d]		
						MLR	10-fold cross-validation	Bootstrap ^[e]
P–Rh [Å], in four-coordinate, square planar Rh ^I complexes with L <i>trans</i> , (CSD survey, Table S9)	17	2.279	0.026	PA, $\Delta\text{P-A}(\text{Pd})$, $\Delta\text{A-P-A}(\text{Pd})$, P–Pd, S_4' calcd, $\chi(\text{H}_3\text{P})\text{Pt}(\text{PH}_3)$	0.998 (0.996)	1.612×10^{-3}	2.917×10^{-3}	2.274×10^{-3}
ΔH_{rxn} [kcal mol ⁻¹], for $[\text{Rh}(\text{CO})_2\text{Cl}]_2 + 4\text{PX}_3 \rightarrow 2\text{Rh}(\text{CO})(\text{Cl})(\text{PX}_3)_2 + 2\text{CO}$ ^[2]	17	55.9	13.5	E_{LUMO} , LP s-character, Q(Pt fragm.), BE(Pd), P–Pd, S_4' calcd	0.997 (0.995)	0.72	1.40	0.93
Tolman electronic parameter (TEP), $A_1 \nu(\text{CO})$ in $[(\text{CO})_3\text{NiL}]^{\text{I}}$ [cm ⁻¹]	49	2074.4	14.8	PA, LP s-character, Q(Pt fragm.), $\Delta\text{A-P-A}(\text{Pd})$, P–Pt, P–B, He_8 -steric	0.988 (0.986)	1.60	1.99	1.80

[a] See Table 2 for variable names, all models include a constant. Descriptor coefficients, which are sensitive to the subset of ligands used, have not been listed, as we are mainly interested in the descriptors contributing to each model. Full models are summarized in the Supporting Information (Tables S7 and S8). [b] Number of ligands in sample. [c] Model quality is always improved by additional descriptors, so adjusted R^2 takes the number of variables in the model into account when computing the regression coefficient. [d] Mean absolute residual. [e] Conditional loss of prediction.

clude contributions from both σ - and π -bonding effects. This interpretation broadly corresponds to conventional understanding of bonding in transition metal–phosphorus ligand complexes. However, the increase in prediction errors estimated using 10-fold cross-validation and bootstrapping compared with the MLR mean absolute residuals suggests caution in using these models for making predictions for future ligands given the limited number of experimental observations available for P–Rh and ΔH_{rxn} (17 in each case).

Modeling the Tolman electronic parameter (TEP)^[1] leads to a lower regression coefficient (Table 4), but the fitted model again provides a good approximation to the relationship with the LKB descriptors. Steric ($\text{He}_{\text{s}}^{\text{steric}}$) as well as σ - and π -electronic contributions to the model can be identified (Tables 4 and S7), with PA, LP s-character and P–B related to σ -bonding and various Pd- and Pt-derived descriptors ($Q(\text{Pt fragm.})$, $\Delta A\text{-P-A}(\text{Pd})$, P–Pt) again including contributions from both σ - and π -bonding. The combination of σ - and π -electronic effects is in good agreement with previous interpretations of both this parameter and carbonyl stretching frequencies in related complexes, which have been used as measures of the electronic properties of phosphorus ligands and are thought to include both σ - and π -electronic effects.^[31,42] It should be noted that for the majority of mixed phosphorus ligands the TEP has not been measured experimentally, but was instead derived by Tolman by assuming that substituent contributions are additive,^[1] which may not adequately describe their interactions in asymmetric ligands.^[10,33]

Conclusions

A prototype ligand knowledge base of DFT-calculated descriptors for phosphorus(III) donor ligands has been developed. The descriptors have been designed to achieve chemical and computational robustness by sampling the properties of a range of ligands and their representative complexes using a standard DFT approach and parameters amenable to automated calculation. The resulting knowledge base can be used to map chemical space and visualize clustering of ligands in chemically meaningful subsets. In addition, linear regression models can be developed that describe the relationship between the descriptors and a range of experimental parameters. The good performance of the models for TEP, P–Rh and ΔH_{rxn} discussed in this work demonstrates that the LKB descriptors as presently constructed are competent to predict a substantial range of ligand behavior. This allows us to attempt the interpretation of experimental data and the evaluation of novel or untested ligands by predicting their properties.

The problems associated with building models of ligand behavior that emerge on complexation seem not to be insuperable, given the satisfactory performance of the prototype LKB in these MLR studies. This has been achieved, because the LKB explicitly includes data on a range of complexes in which such behavior can in principle be (and presumably

has been) recorded. The nature of PC3 is perhaps the best indication that this is indeed happening. Therefore, to capture the full extent of ligand behavior, a wider range of robust descriptors for ligands in complexes should be sought to record the changes in their properties on binding to metals.

Key steps to the construction and use of a mature LKB include the following:

- 1) Identification of computationally robust, responsive descriptors, therefore identifying and discarding non-robust and unresponsive descriptors.
- 2) Design and computation of a range of descriptors sufficiently diverse to sample the inherent and complexation dependent properties of ligands, both in the “free” state and in a range of coordination environments.
- 3) Inclusion of a range of ligands sufficiently diverse and numerous to sample comprehensively the chemical space they span.
- 4) Development of robust statistical protocols for the modeling and prediction of the behavior of ligands in complexes. The models derived map the chemical (ligand) space in Figure 4 above on to the behavior space of the complexes in which the ligands are employed.

While this LKB approach shows promise as a tool for the design of transition-metal complexes and their properties, the protocols for statistical analysis need to be refined to improve model robustness to outliers and variations in the subset of ligands. Similarly, there is a need to establish criteria for comparison and evaluation of competing models. The effect of changes in conformational preferences in response to different coordination environments should be explored and incorporated in relevant descriptors. An extension to bidentate phosphorus ligands as well as to a more chemically diverse set of ligands will require development of a more extended set of descriptors and representative species. In addition, the data generation process should be fully automated and the calculated LKB should eventually be interfaced with databases of structural and experimental data. Work is currently under way in all these areas to develop and implement the conceptual framework outlined in this paper.

Computational Details

All calculations used the Jaguar package^[23] and the standard Becke-Perdew (BP86) density functional.^[24] The Jaguar triple-zeta form of the standard Los Alamos ECP basis set (LACV3P) was used on Pd and Pt, employing the 6-31G* basis for all other atoms. “Loose” convergence (five times larger than default criteria) was used for all geometry optimizations. Test calculations using the more stringent default convergence criteria did not lead to significant changes in energies, bond lengths, or angles, but were much more time-consuming. Calculations were performed on isolated molecules and NBO atomic charges were calculated.^[25] Vibrational frequencies were not computed, and so the energetic data do not include a correction for zero-point energy, although we note that this would be expected to be quite small. In the absence of frequency calculations, stationary points have not been verified as minima. How-

ever, most ligands and complexes are large and optimization to transition states seems unlikely for these carefully built low symmetry starting geometries. Although multiple conformations are viable for some of the ligands and complexes, conformational searching was not attempted, but rather the choice of input geometry was guided by those observed in crystal structures. The impact of potentially resulting "conformational noise" (i.e., variations in descriptor values between alternative conformers) on the data is discussed below. Initial statistical analyses were performed in SPSS for Windows,^[26] and linear regression models evaluated in R.^[27]

Acknowledgement

The authors would like to thank A. H. Welsh for helpful discussions of robust statistical techniques and the Cambridge Crystallographic Data Centre (CCDC) for providing access to the Cambridge Structural Database (CSD). Financial support of the Engineering and Physical Sciences Research Council, including an Advanced Research Fellowship (J.N.H.), is gratefully acknowledged.

- [1] C. A. Tolman, *Chem. Rev.* **1977**, *77*, 313–348.
- [2] A. L. Fernandez, A. Prock, W. P. Giering, The QALE Web Site, <http://www.bu.edu/qale/> (accessed June 10, 2004).
- [3] M. R. Wilson, A. Prock, W. P. Giering, A. L. Fernandez, C. M. Haar, S. P. Nolan, B. M. Foxman, *Organometallics* **2002**, *21*, 2758–2763, and references therein.
- [4] C. Babji, A. J. Poë, *J. Phys. Org. Chem.* **2004**, *17*, 162–167.
- [5] K. D. Cooney, T. R. Cundari, N. W. Hoffman, K. A. Pittard, M. D. Temple, Y. Zhao, *J. Am. Chem. Soc.* **2003**, *125*, 4318–4324.
- [6] B. J. Dunne, R. B. Morris, A. G. Orpen, *J. Chem. Soc. Dalton Trans.* **1991**, 653–661.
- [7] National e-Science Centre, Defining e-Science, <http://www.nesc.a-c.uk/nesc/define.html> (accessed June 4, 2004).
- [8] F. H. Allen, *Acta Crystallogr. Sect. B* **2002**, *58*, 380–388; A. G. Orpen, *Acta Crystallogr. Sect. B* **2002**, *58*, 398–406.
- [9] W. A. Henderson, Jr., C. A. Streuli, *J. Am. Chem. Soc.* **1960**, *82*, 5791–5794; G. M. Bodner, M. P. May, L. E. McKinney, *Inorg. Chem.* **1980**, *19*, 1951–1958; A. G. Orpen, N. G. Connelly, *J. Chem. Soc. Chem. Commun.* **1985**, 1310–1311; A. B. P. Lever, *Inorg. Chem.* **1990**, *29*, 1271–1285; A. G. Orpen, N. G. Connelly, *Organometallics* **1990**, *9*, 1206–1210; L. Chen, A. J. Poë, *Coord. Chem. Rev.* **1995**, *143*, 265–295; S. S. Fielder, M. C. Osborne, A. B. P. Lever, W. J. Pietro, *J. Am. Chem. Soc.* **1995**, *117*, 6990–6993; S. Serron, S. P. Nolan, K. G. Moloy, *Organometallics* **1996**, *15*, 4301–4306; J. M. Smith, B. C. Taverner, N. J. Coville, *J. Organomet. Chem.* **1997**, *530*, 131–140; J. M. Smith, N. J. Coville, L. M. Cook, J. C. A. Boeyens, *Organometallics* **2000**, *19*, 5273–5280; J. M. Smith, N. J. Coville, *Organometallics* **2001**, *20*, 1210–1215.
- [10] T. Bartik, T. Himmler, H.-G. Schulte, K. Seevogel, *J. Organomet. Chem.* **1984**, *272*, 29–41.
- [11] S. Joerg, R. S. Drago, J. Sales, *Organometallics* **1998**, *17*, 589–599; R. S. Drago, S. Joerg, *J. Am. Chem. Soc.* **1996**, *118*, 2654–2663.
- [12] K. A. Bunten, L. Chen, A. L. Fernandez, A. J. Poë, *Coord. Chem. Rev.* **2002**, *233–234*, 41–51.
- [13] T. L. Brown, *Inorg. Chem.* **1992**, *31*, 1286–1294.
- [14] T. L. Brown, K. J. Lee, *Coord. Chem. Rev.* **1993**, *128*, 89–116.
- [15] S. T. Howard, J. P. Foreman, P. G. Edwards, *Inorg. Chem.* **1996**, *35*, 5805–5812; S. T. Howard, J. A. Platts, *J. Phys. Chem.* **1995**, *99*, 9027–9033; W. E. Steinmetz, *Quant. Struct. Act. Relat.* **1996**, *15*, 1–6; R. J. Bubel, W. Douglass, D. P. White, *J. Comput. Chem.* **2000**, *21*, 239–246; C. R. Landis, S. Feldgus, J. Uddin, C. E. Wozniak, K. G. Moloy, *Organometallics* **2000**, *19*, 4878–4886; H. M. Senn, D. V. Deubel, P. E. Blöchl, A. Togni, G. Frenking, *J. Mol. Str. (Theochem)* **2000**, *506*, 233–242; A. M. Gillespie, K. A. Pittard, T. R. Cundari, D. P. White, *Internet Electron. J. Mol. Des.* **2002**, *1*, 242–251.
- [16] L. Perrin, E. Clot, O. Eisenstein, J. Loch, R. H. Crabtree, *Inorg. Chem.* **2001**, *40*, 5806–5811.
- [17] G. Frenking, K. Wichmann, N. Fröhlich, J. Grobe, W. Golla, D. L. Van, B. Krebs, M. Läge, *Organometallics* **2002**, *21*, 2921–2930.
- [18] C. H. Suresh, N. Koga, *Inorg. Chem.* **2002**, *41*, 1573–1578.
- [19] H.-R. Bjørsvik, U. M. Hansen, R. Carlson, *Acta Chem. Scand.* **1997**, *51*, 733–741.
- [20] O. Köhl, *Coord. Chem. Rev.* **2005**, *249*, 693–704.
- [21] K. A. Bunten, D. H. Farrar, A. J. Poë, *Organometallics* **2003**, *22*, 3448–3454.
- [22] E. Burello, G. Rothenberg, *Adv. Synth. Catal.* **2003**, *345*, 1334–1340; E. Burello, D. Farrusseng, G. Rothenberg, *Adv. Synth. Catal.* **2004**, *346*, 1844–1853; E. Burello, P. Marion, J.-C. Galland, A. Chamard, G. Rothenberg, *Adv. Synth. Catal.* **2005**, *347*, 803–810.
- [23] Schrödinger Inc., Jaguar 4.0, Portland, Oregon, **2000**; Schrödinger LLC, Jaguar 5.0, Portland, OR, **2002**.
- [24] J. C. Slater, *Quantum Theory of Molecules and Solids, Vol. 4, The Self-Consistent Field for Molecules and Solids*, McGraw-Hill, New York, **1974**; A. D. Becke, *Phys. Rev. A* **1988**, *38*, 3098–3100; J. P. Perdew, A. Zunger, *Phys. Rev. B* **1981**, *23*, 5048–5079; J. P. Perdew, *Phys. Rev. B* **1986**, *34*, 7406; J. P. Perdew, *Phys. Rev. B* **1986**, *33*, 8822–8824.
- [25] E. D. Glendening, J. K. Badenhop, A. E. Reed, J. E. Carpenter, J. A. Bohmann, C. M. Morales, F. Weinhold, NBO 5.0, Madison, **2001**.
- [26] SPSS Inc., SPSS for Windows Release 11.5, 233 S. Wacker Drive, Chicago, Illinois 60606, **2002**.
- [27] R Development Core Team, R: A language and environment for statistical computing, Vienna, Austria, **2004**.
- [28] J. N. Harvey, K. M. Heslop, A. G. Orpen, P. G. Pringle, *Chem. Commun.* **2003**, 278–279; K. M. Anderson, A. G. Orpen, *Chem. Commun.* **2001**, 2682–2683.
- [29] M. Torrent, M. Sola, G. Frenking, *Chem. Rev.* **2000**, *100*, 439–493.
- [30] A. M. Gillespie, G. R. Morello, D. P. White, *Organometallics* **2002**, *21*, 3913–3921; P. T. Olsen, F. Jensen, *J. Chem. Phys.* **2003**, *118*, 3523–3531; D. Balcells, G. Drudis-Solé, M. Besora, N. Dölker, G. Ujaque, F. Maseras, A. Lledós, *Faraday Discuss.* **2003**, *124*, 429–441.
- [31] A. Bondi, *J. Phys. Chem.* **1964**, *68*, 441–451.
- [32] P. Filzmoser, C. Croux, in *Classification, Clustering and Data Analysis* (Eds.: K. Jajuga, A. Sokolowski, H.-H. Bock), Springer, Berlin, **2002**, pp. 227–234; P. Geladi, *Chemom. Intell. Lab. Syst.* **2002**, *60*, 211–224.
- [33] P. B. Dias, M. E. M. d. Piedade, J. A. M. Simões, *Coord. Chem. Rev.* **1994**, *135/136*, 737–807.
- [34] R. A. Baber, A. G. Orpen, P. G. Pringle, M. J. Wilkinson, R. L. Wingad, *Dalton Trans.* **2005**, 659–667.
- [35] C. Chatfield, A. J. Collins, *Introduction to Multivariate Analysis*, Chapman and Hall, London, **1980**; J. Townend, *Practical Statistics for Environmental and Biological Scientists*, Wiley, Chichester, **2002**.
- [36] D. Livingstone, *Data Analysis for Chemists*, Oxford University Press, Oxford, **1995**.
- [37] R. A. Baber, M. L. Clarke, K. M. Heslop, A. C. Marr, A. G. Orpen, P. G. Pringle, A. Ward, D. E. Zambrano-Williams, *Dalton Trans.* **2005**, 1079–1085.
- [38] M. Hubert, P. J. Rousseeuw, S. Varboven, *Chemom. Intell. Lab. Syst.* **2002**, *60*, 101–111; M. Hubert, S. Engelen, *Bioinformatics* **2004**, *20*, 1728–1736.
- [39] T. R. Cundari, C. Sârbu, H. F. Pop, *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1363–1369.
- [40] D. M. Hawkins, S. C. Basak, D. Mills, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 579–586.
- [41] D. M. Hawkins, *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1–12.
- [42] J. Shao, *J. Am. Stat. Assoc.* **1993**, *88*, 486–494.
- [43] J. Shao, *J. Am. Stat. Assoc.* **1996**, *91*, 655–665.
- [44] R. A. Mansson, A. H. Welsh, N. Fey, A. G. Orpen, unpublished results.

Received: July 27, 2005
Published online: November 9, 2005