

Statistical Modeling of a Ligand Knowledge Base

Ralph A. Mansson* and Alan H. Welsh†

School of Mathematics, University of Southampton, Highfield, Southampton SO17 1BJ, U.K.

Natalie Fey* and A. Guy Orpen

School of Chemistry, University of Bristol, Cantock's Close, Bristol BS8 1TS, U.K.

Received May 23, 2006

A range of different statistical models has been fitted to experimental data for the Tolman electronic parameter (TEP) based on a large set of calculated descriptors in a prototype ligand knowledge base (LKB) of phosphorus(III) donor ligands. The models have been fitted by ordinary least squares using subsets of descriptors, principal component regression, and partial least squares which use variables derived from the complete set of descriptors, least angle regression, and the least absolute shrinkage and selection operator. None of these methods is robust against outliers, so we also applied a robust estimation procedure to the linear regression model. Criteria for model evaluation and comparison have been discussed, highlighting the importance of resampling methods for assessing the robustness of models and the scope for making predictions in chemically intuitive models. For the ligands covered by this LKB, ordinary least squares models of descriptor subsets provide a good representation of the data, while partial least squares, principal component regression, and least angle regression models are less suitable for our dual aims of prediction and interpretation. A linear regression model with robustly fitted parameters achieves the best model performance over all classes of models fitted to TEP data, and the weightings assigned to ligands during the robust estimation procedure are chemically intuitive. The increased model complexity when compared to the ordinary least squares linear model is justified by the reduced influence of individual ligands on the model parameters and predictions of new ligands. Robust linear regression models therefore represent the best compromise for achieving statistical robustness in simple, chemically meaningful models.

1. INTRODUCTION

Networked computing resources have made an increasing volume of information available to the academic community. Attitudes to the provision and access of information have changed to favor instantaneous, (nonexpert) electronic access where activities such as database mining previously required specialist skills.³ In addition to these developments, there have been improvements in the methods of access to specialized databases, e.g. the Cambridge Structural Database (CSD)⁴ and the RCSB Protein Database (PDB),⁵ where automatic data retrieval (*ConQuest*)⁶ and knowledge base derivation tools have been developed (*Mogul*,⁷ *IsoStar*).^{3,8} Statistical analysis methods are often needed to process the information in these databases, and it is important to identify an appropriate method of analysis for a given application. Recent developments in e-science and the Grid architecture⁹ address the changing requirements for database access, management, and analysis, in addition to establishing protocols for the deposition and exploitation of original experimental data¹⁰ and the utilization of high performance computing resources.

We have recently reported the development of a prototype ligand knowledge base (LKB) of calculated descriptors for

phosphorus(III) donor ligands and some of their complexes.¹ In this database the descriptors were calculated using density functional theory (DFT), and the collation and validation steps were facilitated by the original knowledge base design. Our initial work has been focused on the design of this LKB, and statistical analyses have primarily been used to illustrate potential applications. It has been shown that these data can be used to derive variables that map the chemical space and also to fit linear regression models to experimental results. A brief summary of this work is given in section 2 (vide infra).

Since many descriptors in the LKB are correlated, regression models using different subsets of descriptors show similar performance with respect to modeling response variables, and the evaluation and comparison of such models merits further attention. Identification of the *best*, i.e., most suitable, approach from the class of models available depends on whether interpretation or prediction is the overall goal of the model fitting process. In this paper we concentrate on building models for predicting parameters of interest for novel/unknown ligands. Such predictive models are useful in the computational design and evaluation of metal complexes for a range of applications in synthetic and materials chemistry and for the screening of homogeneous catalysts and metallodrugs. We also describe modeling approaches that are suited to small samples of response variables, which is a feature of many experimental data sets in the LKB context. In addition, we aim to provide guidance on the most

* Corresponding authors phone: +44 2380 595 153; fax: +44 2380 595 147; e-mail: ram@soton.ac.uk (R.A.M.) and phone +44 117 95 46398; fax: +44 117 925 1295; e-mail: Natalie.Fey@Bristol.ac.uk (N.F.).

† Present address: Centre for Mathematics and its Applications, Mathematical Sciences Institute, Australian National University, Canberra ACT 0200, Australia.

relevant criteria for the comparison of model performance in the context of this LKB analysis. The suitability and interpretation of the resulting models in a chemical context will also be discussed.

2. PHOSPHORUS(III) DONOR LIGAND KNOWLEDGE BASE

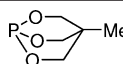
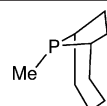
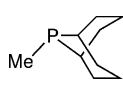
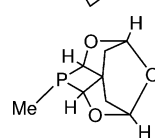
Our initial work has focused on developing a prototype ligand knowledge base (LKB) of DFT-calculated descriptors for phosphorus(III) donor ligands.¹ These ligands are widely used in synthetic organometallic chemistry because changing the substituents on the ligand can modify their steric and electronic characteristics. The properties of transition-metal complexes can thus be tuned to suit specific applications. These complexes have been studied extensively, and a number of experimental^{11–16} and computational^{17–20} approaches have been developed to describe and quantify their steric and electronic properties. The derived stereoelectronic parameters (reviewed, for example, in ref 21) have been used to estimate the quantitative contributions of steric and σ/π -electronic effects to experimentally observed linear free energy relationships,^{12,15,16,22,23} with the aim of comparing and interpreting experimental data. The Quantitative Analysis of Ligand Effects (QALE) approach represents a particularly well-established interpretative application of phosphine ligand descriptors derived from experiment.^{12,16} Stereoelectronic parameters have also been used to visualize ligand relationships and similarities in chemical space (see for example refs 12, 17, 20, 22, and 23), and some of these *maps* have been used to identify candidate ligands for experimental²⁰ and catalyst design.^{17,22,23}

The design of this prototype LKB was chosen to be computationally and chemically robust to allow automated data collection and calculations and for the data to be used for interpretation and predictive modeling.¹ *Computational robustness* implies the use of a reliable, well-established computational approach as well as deriving computationally inexpensive descriptors that are simple to extract from the calculated results, making their calculation suitable for automation. *Chemical robustness* requires comprehensive sampling of chemical (ligand) space and use of transferable descriptors derived from representative chemical environments, in this case mainly transition-metal complexes.

In its prototype version,¹ the phosphorus(III) donor LKB contains data for 61 ligands, which include alkyl- and arylphosphines, phosphites, phosphine halides, and aminophosphines in symmetrical PA_3 and asymmetrical PAB_2 species (see Table 1). Structural and energetic parameters have been calculated for the free ligands (L) as well as a range of their complexes ($[\text{HL}]^+$, $\text{H}_3\text{B}\cdot\text{L}$, $[\text{PdCl}_3\text{L}]^-$, $[\text{Pt}(\text{PH}_3)_3\text{L}]$) with a standard DFT functional (see section 6 and ref 1 for computational details). These descriptors were used in multiple linear regression models to illustrate their versatility for modeling a varied range of experimentally determined parameters such as bond lengths, reaction enthalpies, and bond-stretching frequencies.¹ In addition, the use of resampling methods to assess the predictive performance of models was highlighted.¹ Table 2 lists the descriptors used in the statistical analyses described here.

Table 1. Phosphorus(III) Donor Ligands in LKB

L = PA ₃		L = PAB ₂	
number	A =	number	A = B =
1	H	34	H F
2	Me	35	F H
3	Et	36	H Cl
4	Pr	37	Cl H
5	ⁱ Pr	38	Me F
6	Bu	39	F Me
7	^t Bu	40	Me Cl
8	CF ₃	41	Cl Me
9	Cy(C ₆ H ₁₁)	42	Me CF ₃
10	Bz (CH ₂ Ph)	43	CF ₃ Me
11	F	44	Me ^t Bu
12	Cl	45	^t Bu Me
13	OMe	46	Me Ph
14	OEt	47	Ph Me
15	OPh	48	Et Ph
16	NH ₂	49	Ph Et
17	NMe ₂	50	Cy Ph
18	Pyr (NC ₄ H ₄ , pyrrolyl)	51	Ph Cy
19	NC ₄ H ₈	52	Pyr Ph
20	Pip (NC ₅ H ₁₀ , piperidyl)	53	Ph Pyr
21	CHCH ₂	54	o-Me-Ph Ph
22	Ph	55	Ph o-Me-Ph
23	C ₆ H ₅	56	o-MeO-Ph Ph
24	o-Me-Ph	57	Ph o-MeO-Ph
25	m-Me-Ph		
26	p-Me-Ph		
27	o-MeO-Ph		
28	p-MeO-Ph		
29	3,5-(CF ₃) ₂ -Ph		
30	p-CF ₃ -Ph		
31	p-F-Ph		
32	p-Cl-Ph		
33	p-Me ₂ N-Ph		

number	L	number	L
58		59	
60		61	

3. STATISTICAL METHODOLOGY

Data in the ligand knowledge base can be used to build models that approximate the relationship between experimental results and calculated descriptors. Many of the LKB descriptors are correlated, and models of similar performance can be derived both from different subsets of descriptors and by using different classes of statistical regression models. In comparing these models, we are assessing their performance with respect to both statistical and chemical criteria: 1. The success of approximating the relationship between the response variable of interest and the calculated descriptors (*quality of model fit*). 2. The *complexity* of the model, i.e., whether an appropriate tradeoff between model simplicity/transferability and performance has been achieved. 3. The *statistical robustness* of the model to incomplete sampling of chemical space and changes in the ligand set. This criterion is desirable but problematic to assess without completing the sampling, and in this case we would build a model using the completed data. 4. The suitability for making *predictions*

Table 2. Calculated Descriptors in Prototype LKB¹

descriptor ^a	derivation (unit)
	Free Phosphorus(III) Species (L)
E_{HOMO}	energy of highest occupied molecular orbital (Hartree)
E_{LUMO}	energy of lowest unoccupied molecular orbital (Hartree)
LP s-character	contribution of P s-orbital to lone pair (LP), from NBO analysis (%)
He ₈ 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)
	Protonated Ligand ([HL] ⁺)
PA	proton affinity, calculated as the difference between the energy of the neutral and protonated L (kcal mol ⁻¹)
	Borane Adduct (H ₃ B.L)
Q(B fragm)	NBO charge on BH ₃ fragment
BE(B)	bond energy for dissociation of L from BH ₃ fragment (kcal mol ⁻¹) ^b
$\Delta\text{P-A(B)}$	change in average P-A bond length compared to free ligand (Å)
$\Delta\text{A-P-A(B)}$	change in average A-P-A angle compared to free ligand (°)
P-B	P-B distance (Å)
	Palladium Complexes ([PdCl ₃ L] ⁻)
Q(Pd fragm)	NBO charge on [PdCl ₃] ⁻ fragment
BE(Pd)	bond energy for dissociation of L from [PdCl ₃] ⁻ fragment (kcal mol ⁻¹) ^b
$\Delta\text{P-A(Pd)}$	change in average P-A bond length compared to free ligand (Å)
$\Delta\text{A-P-A(Pd)}$	change in average A-P-A angle compared to free ligand (°)
P-Pd	P-Pd distance (Å)
Pd-Cl trans	Pd-Cl distance, trans to L (Å)
	Platinum Complexes ([Pt(PH ₃) ₃ L])
Q(Pt fragm)	NBO charge on [(PH ₃) ₃ Pt] fragment
BE(Pt)	bond energy for dissociation of L from [Pt(PH ₃) ₃] fragment (kcal mol ⁻¹) ^b
$\Delta\text{P-A(Pt)}$	change in average P-A bond length compared to free ligand (Å)
$\Delta\text{A-P-A(Pt)}$	change in average A-P-A angle compared to free ligand (°)
P-Pt	P-Pt distance (Å)
<(H ₃ P)Pt(PH ₃)	average (H ₃ P)Pt(PH ₃) angle (°)
	Cumulative
S4' calc	($\Sigma < \text{ZPA} - \Sigma < \text{APA}$), where Z = BH ₃ , [PdCl ₃] ⁻ , [Pt(PH ₃) ₃] (°)

^a All calculations were performed on isolated molecules. ^b BE = [$E_{\text{tot}}(\text{fragment}) + E_{\text{tot}}(\text{L})$] - $E_{\text{tot}}(\text{complex})$.

for novel/untested ligands. 5. The scope for interpretation of the model and its underlying chemistry, which we have termed *chemical contextualization*.

This section provides an overview of the statistical methodology applied to the data in the LKB.

3.1. Classes of Models. In situations where there is no prior model assumed for the data, linear regression models can be used to derive empirically useful approximations to the relationship between a response, such as Tolman's electronic parameter (TEP),¹¹ and a set of descriptors. These models are simple and frequently provide an adequate and interpretable description of the relationship. For the prediction of a response Y , using a group of p descriptors X_1, \dots, X_p , the multiple regression model has the form

$$Y_i = \beta_0 + \sum_{j=1}^p \beta_j X_{ij} + \epsilon_i \quad i = 1, \dots, n \quad (1)$$

The coefficients of this model $\beta_0, \beta_1, \dots, \beta_p$ can be estimated from the data using ordinary least squares (OLS) methodology. The residuals ϵ_i ; $i = 1, \dots, n$ are assumed to be independent, identically distributed (normal) random variables with zero mean and common variance σ^2 ; the validity of these assumptions can be investigated using model diagnostics. Regression diagnostics, e.g. plots of residual against fitted values, may indicate that a small part of the data is poorly fitted by the model. These data may have a large influence on the estimates of the model coefficients and their standard errors. In this case, it may be desirable to

use robust methods for estimating model coefficients to reduce the impact of a few statistically discrepant observations. Robust estimation of regression model coefficients aims to get a good fit to the majority of the data while assuming that some of the data does not follow the same relationship.²⁹ The estimation procedure minimizes a function of the residuals

$$\sum_{i=1}^n \rho(Y_i - \beta_0 - \sum_{j=1}^p \beta_j X_{ij}) \quad (2)$$

We have used a high breakdown estimator with one-step of the Tukey estimator for the initial values. Diagnostic residual plots can be used to highlight the effect of weighting the data. It is possible to consider other classes of models with robust regression approaches, e.g. cubic splines as part of a linear model.²⁴ However, we have limited our investigation to multiple linear regression as this approach is most likely to allow chemical interpretation and the numbers of both ligands and descriptors restrict the complexity of potential models.

In cases where many descriptors not only are considered important but also are highly correlated, OLS estimates of model coefficients can become unreliable,²⁵ and models based on derived variables, such as principal component regression (PCR) and partial least squares (PLS), which fit a model to linear combinations of the original descriptors, may be more useful. When fitting PCR models to chemical data, the original descriptor variables are transformed into a

new set of orthogonal variables (i.e. uncorrelated variables) called principal components. This transformation ranks the new orthogonal variables in order of decreasing variability. The model selection proceeds by eliminating some of the principal components in order to effect a reduction in variance and to reduce the complexity of the model. The principal components are found by computing the eigenvalues of the covariance matrix after the original variables have been standardized—mean centered and scaled to unit variance.

The model is $y = \beta_0 + \sum_{i=1}^k \beta_i z_i + \epsilon$ where z_i are the principal components. A standard approach to variable selection for PCR corresponds to removing components with the smallest eigenvalues sequentially, while the total variation explained remains above a prespecified threshold. After elimination of the principal components in order of decreasing variability, a multiple regression analysis of the response variable against the reduced set of principal components is performed using ordinary least squares. Once regression coefficients for the orthogonal variables have been estimated, these are transformed into a new set of coefficients corresponding to the original variables.

Some cautionary notes have been published about the use of principal components in this way.²⁶ In summary, the computation of principal components is based on variances, so the method is sensitive to the presence of outliers, and there are situations where the most variable components are not strongly related to the response. In addition, the interpretation of both principal components and latent variables as generated in PLS regression (see below) may be problematic and not statistically robust whether due to the presence of outliers or the descriptors chosen to form the derived variables.

Partial least squares (PLS) regression is a popular technique in chemometric applications^{20,22,27} and is frequently employed as an alternative to ordinary least squares (OLS) when the matrix of descriptors is poorly conditioned, i.e., for highly correlated sets of descriptors. The aim of the method is to identify a small number of PLS factors, sometimes known as latent variables, to explain a large proportion of the variation in both descriptors and response. The PLS method initially projects the descriptors $X = (X_1, \dots, X_p)$ and response Y onto one or more new axes. The model can be expressed in the form

$$X = TP + E \text{ and } Y = UQ + F \quad (3)$$

where T and U are the X - and Y -scores, P and Q are the X - and Y -loadings, and E and F are the X - and Y -residuals. The PLS approach chooses successive factors to maximize the covariance of each X -score and corresponding Y -score. For a good PLS model there is high correlation between the X - and Y -scores based on using only the first few factors. There are similar concerns to PCR about whether sensible interpretations of the latent variables can be achieved, and although the number of variables may be less than the original number of descriptors, it will still be necessary to calculate all of the descriptors, preventing any savings in terms of computational time.

3.2. Model Diagnostics. When fitting regression models to data, a number of assumptions are made. These typically include the following: 1. form of the expectation function,

which covers the variables and function used to relate the response to the descriptor variables; 2. additivity of the error term in the model; 3. constant variance for the error term in the model; 4. the distribution of the model residual is approximately normal; and 5. independence of model residuals.

These assumptions can be investigated using diagnostic plots of the model residuals. Plotting residuals, $R_i = Y_i - \hat{Y}_i$, against fitted values, \hat{Y}_i , allows investigation of the form of the model, additivity of errors, and constant variance. In general, model assumptions are considered reasonable if the points form an even, horizontal band with no patterns, whereas curvature or increasing variability in the residuals indicates potential problems with model robustness. The assumption of normality for the residuals can be investigated by plotting ordered residuals against expected normal quantiles. This diagnostic allows the detection of discrepant points, i.e., those which deviate from a straight (horizontal) line, again indicative of a nonrobust regression model. The problem associated with outliers and non-normality can be addressed by robust estimation of model parameters (vide supra).

3.3. Variable Selection. Two issues have to be considered simultaneously when identifying suitable subsets of the chemical descriptors: optimization of the number of parameters in a model (*complexity*) and the identification of a best subset given a fixed number of parameters. A tradeoff is often required to obtain models with good predictive power, because an increase in model complexity improves the precision of future predictions but also introduces bias.

Various automatic and manual procedures exist for identifying good subsets of descriptors. Two nested models can be compared using an F-test, which is based on the residual sum of squares for these models. This test considers whether the improvement in the fit of the model is sufficiently large to justify increasing the complexity of the model.²⁹ When the parameters are estimated robustly in the regression model, a robust version of this test is used to investigate whether to include a descriptor in the model.

Automatic subset selection methods²⁹ include best subset regression where, for a given number of parameters, the subset of variables with the smallest residual sum of squares is identified. When there are a large number of descriptors, it is often more feasible to analyze a sequence of models. Forward stepwise selection starts with only a constant in the model and sequentially adds descriptors. At each stage the descriptor that provides the largest improvement in the fit is added to the model. Backward stepwise selection is the reverse process, which starts with the full model (all descriptors included), and sequentially excludes descriptors if the fit of the model is not substantially decreased. Automatic procedures have a tendency to overfit, so it is often prudent to combine this approach with manual intervention to reduce the complexity of the model.

A group of similar methods that are a more recent alternative to subset selection methods include the Least Absolute Shrinkage and Selection Operator (LASSO) and least angle regression (LAR).^{30,31} These methods combine parameter estimation and model selection into one procedure, using a single parameter to vary the complexity of the model. When this parameter is varied, both the number of descriptors in the model and the parameter estimate change. These

methods have two advantages over subset selection models: the shrinkage parameter can be computed by a standard continuous optimization procedures and the estimate varies smoothly with the learning set and with the hyperparameter setting. As a result, the methods are stable with respect to slight changes in data and with respect to errors in the hyperparameter tuning.

3.4. Measures of Model Performance. Standard regression statistics, such as adjusted R^2 , assess the reproduction of data used in the model derivation. [Model quality is always improved by including descriptors, and the number of variables is taken into account using the adjusted version of the statistic.] A more appropriate way to assess the quality of a fitted model is to estimate prediction errors for the experimental data and then to use these values to discriminate between different models. Prediction errors can be calculated by dividing the data into a training and a validation set and using the model fitted to the training set to predict values for the validation set. However, when there are insufficient data to partition the set into two groups, resampling methods, for example cross-validation or bootstrapping, can be considered.

The main challenge in selecting a suitable model arises from the question of model complexity or degrees of freedom, i.e., how many descriptors or derived variables should be included in a model. Ten-fold cross-validation (vide infra) has been used to locate the point where increasing the complexity of the model does not produce a substantial reduction in prediction sum of squares.³¹ Hastie et al.³¹ have compared different classes of models, such as ordinary least squares (OLS) linear regression, principal component regression (PCR), and partial least squares (PLS), using cross-validation. Modern methods, such as the Least Absolute Shrinkage and Selection Operator (LASSO) and least angle regression (LAR), have a tuning parameter that is optimized using 10-fold cross-validation. The least complicated model with a cross-validation estimate close to the minimum is selected. In this application of cross-validation the data are divided randomly into 10 roughly equal groups. The model is then fitted to nine of the 10 groups, and the prediction error is computed for the tenth group that has been excluded from the model fitting. This process is repeated by excluding each of the 10 groups, and the prediction error estimates are averaged; the approach has been discussed in detail by Shao.³²

The bootstrap can also be used to estimate prediction error for a given model. In this case, data sets of size n are randomly drawn from the original data with replacement, and the model is fitted to the bootstrap data. This process is done B times, and for each sample b the conditional expected loss of prediction for model j is calculated using the formula

$$\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_{ij})^2 + \frac{1}{nB} \sum_{b=1}^B \sum_{i=1}^n \{ (y_i - \hat{y}_{bij}^*)^2 - (y_i^{**} - \hat{y}_{bij}^*)^2 \} \quad (4)$$

where \hat{y}_{ij} is the fitted value for the i th observation, y_i , from model j , and \hat{y}_{bij}^* is the fitted value for the i th observation in bootstrap sample b from model j . The use of the bootstrap for model selection has been considered by Shao³³ and for robust regression models by Wisnowski et al.³⁴

Residuals ϵ_i have been computed for the full model and then adjusted based on the models and fitting methods considered. In the nonrobust case the raw residuals are adjusted using the formula

$$r_i = \frac{\epsilon_i - \bar{\epsilon}}{\sqrt{1 - P/n}}, \quad (5)$$

where $\bar{\epsilon}$ is the mean of the raw residuals, P is the number of parameters in the model, and n is the number of observations. In the case of robust regression models there may be problems with the convergence of the fitting procedure for some of the bootstrap samples, and the residuals are leverage corrected as discussed by Davison and Hinkley.³⁵ In this case the formula is

$$r_i = \frac{\epsilon_i}{\sqrt{1 - dh_i}},$$

where $d = \frac{2 \sum (\epsilon_i/s) \psi(\epsilon_i/s)}{\sum \dot{\psi}(\epsilon_i/s)} - \frac{n \psi^2(\epsilon_i/s)}{\{\sum \dot{\psi}(\epsilon_i/s)\}^2}$ (6)

and h_i is the leverage of the i th observation. Here $\psi = \dot{\rho}$ is the derivative of the criterion function ρ . In some cases, the bootstrap process might generate *bad* samples, and, assuming the proportion of bad bootstrap samples is not substantial, the results from these problematic samples are simply ignored. [Bootstrap samples are chosen by sampling with replacement, so there is the danger of large residuals being selected repeatedly leading to situations where the robust estimation procedure does not converge to a solution.]

To compare the various models for a given response the conditional expected loss of prediction is used for the bootstrap and the prediction sum of squares for the cross-validation. There are B bootstrap samples and N predictions for the 10-fold cross-validation. The mean value for these 2 quantities was estimated both robustly and nonrobustly for comparison, and *standard errors* were calculated for these quantities using the approach as described by Hastie et al.³¹ For the nonrobust case the computation is more straightforward, and the variance is simply divided by the number of values, but in the robust situation, using the Huber proposal 2 estimator, the variance of the robust estimator is

$$\frac{1}{B} \frac{1/B \sum \psi(r_i/s)^2}{\{1/B \sum \dot{\psi}(r_i/s)/s\}^2} \quad (7)$$

where s is the robust scale estimate for the data, and r_i are the differences between the data and the robust estimate of the mean.

4. LIGAND KNOWLEDGE BASE MODEL BUILDING EXAMPLES

A general feature of the prototype ligand knowledge base (LKB) is the large number of descriptors in comparison to the number of ligands. The database does not sample ligand space evenly, e.g. there are only a few halides but many alkyl and aryl phosphines (Table 1). While calculated descriptors are available for all ligands, there is typically experimental data for only 15–25 of the 61 ligands in the LKB. These features of the data make robust methods

Table 3. Models Fitted to LKB Data To Describe the Relationship between TEP and Descriptors

model ^a	variables	parameters ^b
OLS	PA, LP s-character, Q(Pt fragm.), $\Delta A-P-A(Pd)$, P-B, P-Pt	7
RLM1	PA, LP s-character, Q(Pt fragm.), BE(Pd), $\Delta A-P-A(Pd)$, P-Pd, P-Pt, He ₈ steric, Pd-Cl trans	10
RLM2	E_{HOMO} , Q(Pt fragm.), $\Delta P-A(B)$, $\Delta A-P-A(Pt)$, P-Pt, He ₈ steric	7
LASSO	PA, E_{HOMO} , LP s-character, BE(B), BE(Pd), $\Delta P-A(B)$, $\Delta P-A(Pt)$, $\Delta A-P-A(B)$, $\Delta A-P-A(Pd)$, $\Delta A-P-A(Pt)$, P-B, P-Pt, He ₈ steric, $<(H_3P)Pt(PH_3)$, $s = 0.434^c$	16
LAR	PA, E_{HOMO} , $\Delta P-A(B)$, $\Delta A-P-A(B)$, $\Delta A-P-A(Pd)$, P-Pt, Pd-Cl trans, $s = 0.242^c$	8
PCR	19 principal components	20
PLS	12 latent variables	13

^a OLS – ordinary least squares, RLM – robust linear model, LASSO – least absolute shrinkage and selection operator, LAR – least angle regression, PCR – principal component regression, PLS – partial least squares. ^b A constant is included in the models. ^c Tuning parameter, estimated using 10-fold cross-validation.

Table 4. Comparison of the Different Classes of Models Considered for Describing the Relationship between TEP and the Calculated Descriptors^c

model	10-fold cross-validation ^a		bootstrap ^b	
	robust	nonrobust	robust	nonrobust
OLS	3.758 (0.291)	5.659 (1.218)	5.179 (0.100)	5.202 (0.076)
RLM1	N/A		3.375 (0.181)	4.978 (1.176)
RLM2			5.964 (0.190)	5.851 (0.256)
LASSO			7.325 (0.259)	8.415 (0.757)
LAR	10.213 (0.491)	16.172 (4.486)	7.452 (0.280)	8.339 (0.737)
PCR	7.799 (0.450)	10.546 (2.098)	4.360 (0.109)	4.364 (0.093)
PCR (10 param)	11.263 (0.525)	17.500 (4.882)	16.382 (0.099)	16.369 (0.070)
PLS	5.541 (0.355)	9.546 (2.508)	3.425 (0.099)	3.417 (0.065)
PLS (10 param)	9.079 (0.470)	17.995 (5.610)	4.962 (0.115)	4.975 (0.090)

^a Prediction sum of squares. ^b Conditional expected loss of prediction. ^c See Table 3 for model abbreviations. Standard errors for estimates are given in parentheses.

suitable for modeling the relationships between experimental data and the calculated descriptors.

The general statistical process of model building for variables in the LKB is as follows: 1. Consider regression models with linear terms and, where appropriate, estimate the model parameters robustly to address problems of non-normality. 2. Fit models to derived variables using principal component regression and partial least squares. 3. Investigate identifying model complexity alternatively via shrinkage methods, using cross-validation to select a sensible number of variables for the model. 4. Compare models using 10-fold cross-validation and bootstrapping to assess their performance.

Once the statistical performance of a range of models is established (i.e. *reproduction* of experimental data and quality of *prediction*), the chemical context of any outliers is investigated, and the models are contextualized in terms of chemical concepts and understanding of experimental results (*interpretation*).

The Tolman Electronic Parameter (TEP)¹¹ was considered as a response variable. As reported in our previous work,¹ an ordinary least squares (OLS) model of the TEP showed worse performance than models derived for other experimental data sets, making this a useful test case for other regression approaches. Furthermore, the overlap between ligands in Tolman's original work and those in the LKB is good, and the TEP has been measured directly for 31 ligands. There are an additional 18 TEP values that have been estimated from individual substituent contributions (vide infra, see Table S1 in the Supporting Information for data).

The TEP can be determined experimentally and corresponds to the vibrational frequency of the A₁ carbonyl stretching mode of [Ni(CO)₃L] complexes measured in dichloromethane by infrared spectroscopy. The nickel complex can be readily prepared for many phosphorus(III) donor ligands, and Tolman estimated that this frequency could be measured with an accuracy of $\pm 0.3 \text{ cm}^{-1}$.¹¹ From measurements on mixed PAB₂ complexes Tolman derived individual substituent contributions χ_i , and, by considering these contributions as additive, the TEP could be estimated as $TEP = 2056.1 + \sum_{i=1}^3 \chi_i$.¹¹ The validity of this assumption has been discussed;^{14,36} however, the TEP and related vibrational frequencies of transition-metal carbonyl complexes are frequently used as measures of net ligand electronic effects (see ref 21 for a recent review). Measurements have been extended to other ligands (see for example refs 19, 37, and 38 and references cited), and a number of computational approaches have been reported.^{17,19,37}

A series of predictive models of the types discussed in earlier sections have been derived using the 23 calculated descriptors shown in Table 2. These statistical models have been summarized in Table 3, and the measures of model performance are compared in Table 4. The descriptor coefficients for all models have been summarized in the Supporting Information (Table S2).

Best subsets ordinary least squares (OLS) regression suggests a model with six descriptors (PA, LP s-character, Q(Pt fragm.), $\Delta A-P-A(Pd)$, P-B, P-Pt), and the relationship between the response and descriptors is described reasonably well by a linear relationship (adj. $R^2 = 0.981$), but the

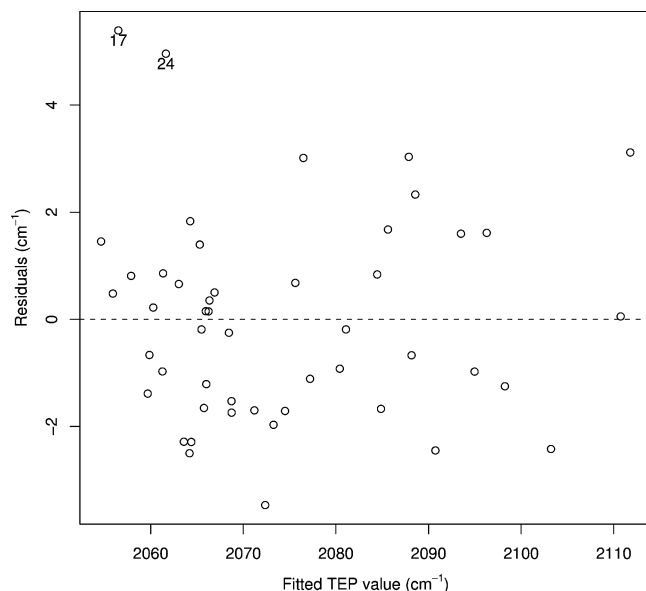


Figure 1. Residual versus fitted (TEP) value diagnostic plot for the ordinary least squares (OLS) model. (See the Supporting Information for the plot with all ligands identified, Figure S1).

diagnostic plot of residuals vs predicted values (Figure 1) shows two ligands that are not well described by the relationship—ligands **17** ($\text{P}(\text{NMe}_2)_3$) and **24** ($\text{P}(\text{o-tolyl})_3$) have large positive residuals. The quality of predictions for the bulky ligand **24** might be affected by both its size and its distinct conformational preferences in different coordination environments;⁴⁰ such steric effects are likely to contribute to fragment-ligand bond lengths (P-B, P-Pt) and perhaps to angle change ($\Delta\text{A-P-A}(\text{Pd})$) descriptors in the model. A chemical explanation for the outlier behavior of aminophosphine **17** is not straightforward; however, it should be noted that there are only 2 TEP values for aminophosphine ligands available (ligands **17** and **20**), and there are only 4 aminophosphine ligands considered in the LKB. The subset of ligands in this region of chemical space is thus not represented well.

There are only a few ligands where the OLS model provides a poor fit, and when using either the cross-validation or the bootstrap statistics the model performance is good in comparison to the other classes of models. It should be noted that the cross-validation and bootstrap estimates are not directly comparable because they investigate prediction error by different approaches. The descriptors in this model could be interpreted in terms of a combination of σ -/ π -electronic effects, with PA, LP s-character and P-B relating to σ -bonding, while the Pd- and Pt-derived descriptors likely include contributions from both σ - and π -effects. As stated above, there may also be implicit steric effects in these descriptors, in particular for the metal fragment-ligand bond lengths (P-B, P-Pt).

Phosphorus(III) donor ligand space is not covered completely by the LKB, so the linear model parameter were estimated robustly (robust linear model, RLM) to investigate possible improvements in the accuracy of predictions, in particular for those areas in the predictor space that are more sparsely sampled by ligands (e.g. aminophosphines, phosphites, phosphine halides). When the model parameters are estimated robustly, a model with 9 descriptors shows the best performance (RLM1, Table 3). Given the increase in

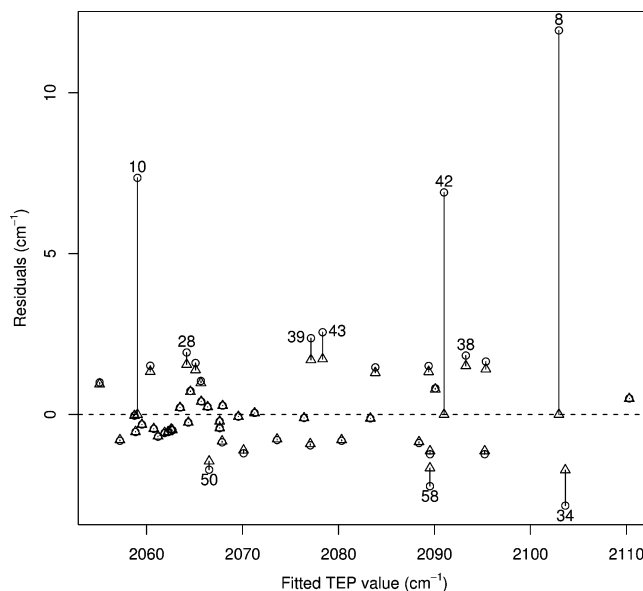


Figure 2. Residual versus fitted (TEP) value diagnostic plot for best robustly estimated linear model (RLM1), including indication of the effect of weighting. (See the Supporting Information for plot with all ligand numbers, Figure S2a).

complexity/dimension when compared to ordinary least squares parameter estimation, a robust linear model with the same number of descriptors as OLS was also considered (model RLM2). The important features of these models are reported in Table 3. The diagnostic plot of fitted values vs residuals for RLM1 is shown in Figure 2 and includes an indication of the weightings used in model derivation (see the Supporting Information for a detailed listing and the corresponding results for RLM2 (Table S1, Figure S2)). The majority of the ligands are not allocated small weights, but the influence of some ligands on the model parameters has been substantially reduced.

The descriptors in RLM1 can be interpreted broadly in terms of σ -/ π -electronic and steric effects contributing to the value of Tolman electronic parameter (TEP). The robust estimation procedure allocates weights to the ligands so that the overall fit is good for the majority of the data. There are some interesting features of the ligands that have reduced weights. Ligands **8**, **10**, and **42** have been allocated zero weights and so do not contribute toward the parameter estimates. In addition, ligands **28**, **34**, **38**, **39**, **43**, **50**, and **58** have been assigned relatively low weights (<0.85) by the robust fitting procedure. The TEP values for most of these ligands (**8**, **34**, **38**, **39**, **42**, and **50**) have actually been calculated by (linear) interpolation from substituent contributions (vide supra) rather than measured experimentally. Previous research has indicated^{14,36} that the contributions of individual substituents are not perfectly additive so that these TEP values are potentially less accurate than those determined experimentally. The low weightings assigned to some of these ligands appear to support this, suggesting that predictions are more difficult in these cases. However, some ligands, for which experimental data have been measured directly, have also been down weighted (**10**, **28**, **43**, and **58**); for these, incomplete sampling of ligand space for more exotic ligands (**58**) and conformational issues (**10** and **28**) might affect the quality of predictions. It should be noted that, while very reliable and robust for most of ligand space,

this model would be less successful for making predictions for these down-weighted ligands and potentially also for chemically similar ligands. The weightings assigned by this approach should thus be considered in the chemical context of the derived model.

Comparison of the estimated prediction errors (Table 4) shows that RLM1 is more accurate than the OLS model discussed above, while RLM2 is less precise when making predictions. The large discrepancy between the robust and nonrobust bootstrap statistics for RLM as well as the large standard error for the nonrobust bootstrap estimate indicates that this model performs consistently better for most ligands but gives poor predictions for some of the ligands. RLM1 has 9 descriptors compared to 6 for the *best* OLS model, so the apparent improvement in performance achieved for the robust model is partially due to increased model complexity. However, the RLM model parameters have been influenced less by the subset of ligands discussed above.

Two models based on the least angle regression fitting procedure (LAR and LASSO) were investigated for modeling TEP and are summarized in Table 3, with bootstrap and cross-validation prediction errors shown in Table 4. The standard LAR plots of the model coefficients against the relevant tuning parameter are shown in the Supporting Information (Figure S3). Both models derived have more descriptors than the linear model (OLS), and the LASSO model is more complex than the best robust linear model (RLM1) as well (Table 3). In addition, their predictive performance is inferior, and the model derivation procedure is more complicated than for least-squares regression.

We also fitted principal component regression (PCR) and partial least squares (PLS) models to the Tolman electronic parameter data to assess the performance of these different classes of models. The resulting models are also summarized in Table 3, and the estimated cross-validation and bootstrap prediction errors are shown in Table 4. The relevant diagnostic plots have been included in the Supporting Information (Figures S4 and S5). The bootstrap estimates of prediction error indicate that these models perform well when compared to the other models fitted to the TEP data. However, both models are considerably more complex than the OLS model and the robust linear model RLM1 discussed above: PCR requires 19 principal components, while the best number of latent variables for the PLS model is 12. Table 4 also shows the estimated prediction errors for PCR and PLS models with 10 parameters, i.e., with complexity comparable to RLM1. The predictive performance of these models is worse than that obtained for RLM1 in both cases, with the 10-parameter PCR giving the worst performance of all models considered here.

One of the attractions of using projection techniques such as principal component regression and partial least squares techniques for situations where there are highly correlated descriptors is to reduce the dimensionality of the model to fewer derived variables, so the PCR model in particular is unsatisfactory in this respect. In addition, all 23 descriptors would need to be calculated when making predictions for new ligands with these models, compared to OLS and RLM models that would utilize a smaller subset of descriptors. In PCR and PLS models, most descriptors contribute to several derived variables, and chemically intuitive interpretations of these variables are likely to be both vague and sensitive to

changes in the training set thus complicating chemical *contextualization*. While these models thus appear suitable for making reliable predictions, they are less appropriate for our dual aims of prediction and interpretation.

5. SUMMARY AND CONCLUSIONS

We have investigated fitting a range of different statistical models to experimental data based on a large set of calculated descriptors in the ligand knowledge base (LKB). These models are primarily aimed at the reliable prediction of data for future ligands using chemically intuitive models. The classes of model considered include ordinary least squares (OLS) linear models using subsets of descriptors as well as principal component regression (PCR) and partial least squares (PLS) which use variables derived from descriptors for a variety of chemical environments. We have also considered robust estimation procedures within the linear regression model (RLM) to reduce the impact of a small subset of the ligands on the overall results. Least angle regression (LAR and lasso), a more recent approach to model selection and parameter estimation from the statistical literature, has also been applied to data from the LKB.

Model building for this LKB is rendered difficult because available experimental data sets are not extensive and sampling of ligand space is both uneven and likely to be incomplete. Standard statistical approaches assume that any data sample modeled is representative of a whole population. In addition, the number of descriptors in the LKB is large compared to the total number of ligands, mainly because the LKB has been designed to be chemically robust by deriving descriptors from a variety of chemical environments. This also causes descriptors to be quite highly correlated, potentially resulting in multiple models of similar performance, with individual descriptor contributions difficult to interpret. These issues are likely to be encountered for other chemical applications of statistical models as well, and in this paper we have examined criteria for model evaluation and comparison, highlighting the importance of resampling methods for assessing the scope for making predictions and the robustness of models to outliers and changes in ligand and descriptor sets. Where possible, we have also sought chemical explanations for model parameters and outliers, highlighting the knowledge value of chemically intuitive results.

The main outcome of fitting models to the LKB data has been that OLS models of descriptor subsets provide a good representation of the data, on a par with more complicated linear regression approaches such as PLS and PCR. However, in OLS regression models outliers may influence the model parameters unduly, which can impair accurate predictions for future ligands of interest as indicated by estimated prediction errors. Fitting the model parameters robustly (RLM) gives less weight to some of the ligands to produce models that should be more accurate for future predictions overall, as they are not unduly influenced by ligands that occupy a sparse area of knowledge space. However, if strongly down-weighted ligands are of particular interest, explanations for these weights should be sought, and other statistical approaches may be more useful. For the Tolman electronic parameter at least, the improved accuracy of prediction achieved with robust models comes at the cost of

a greater number of model parameters. On the other hand, many of the ligands assigned low or zero weightings by the robust regression were actually estimated from individual substituent contributions rather than measured experimentally, suggesting that this approach can give results of chemical value.

PCR and PLS models allow retention of the full set of descriptors. Although their performance with the optimal number of principal components or latent variables is similar to OLS and RLM models, the number of variables in these models is larger. The benefits of dimension reduction compared to OLS and RLM are thus not observed in this particular application, and the resulting models lose both robustness and knowledge value, as the derived variables are sensitive to outliers and changes in the data set and difficult to interpret. Least angle regression models have also been considered for parameter estimation and automatic selection of model complexity, but these models show very poor performance in comparison to the other techniques considered here. The results of the modeling indicate that RLM models represent a good compromise for achieving statistical robustness and produce simple, chemically meaningful models.

The results of the statistical modeling indicate that robust linear regression models represent a good compromise for achieving statistical robustness even when sampling of chemical space is incomplete and uneven, enabling the reliable predictions of response data for novel or unknown ligands. The robust linear regression approach is suitable for the modeling of experimental data sets with small samples of response variables. These models could thus be particularly useful for the computational design and evaluation of metal complexes, especially for the screening of expensive homogeneous catalysts and metallodrugs. Unlike approaches using derived variables, such models are simple and chemically meaningful, allowing the interpretation of descriptor contributions in terms of familiar steric and electronic effects.

6. TECHNICAL DETAILS

The model fitting was performed in the R statistical software system³⁹ using some of the standard libraries of routines and other additional contributed packages. For the different classes of models these include the following: (i) PCR and PLS models fitted using *pls.pcr* which works out the optimal number of derived variables. This library also has routines for 10-fold cross-validation which are used for all types of models. (ii) RLM fits with robust estimation from the routines included in the Modern Applied Statistics in S (MASS) library. The robustly estimated regression models have been fitted using an MM-estimator,⁴¹ which is a high breakdown estimator. (iii) The *boot* library is the basis for custom functions produced to estimate the conditional expected loss of prediction via the bootstrap.

Descriptors were calculated using the BP86⁴² DFT functional with a standard basis set combination (6-31G* on all atoms except Pd and Pt, which used LACV3P) in the Jaguar package⁴³ (see ref 1 for details of these calculations).

ACKNOWLEDGMENT

The authors would like to thank Dr. J. N. Harvey for helpful discussions. Financial support of the Engineering and

Physical Sciences Research Council is gratefully acknowledged. This paper is part 3 in the series Development of a Ligand Knowledge Base. See refs 1 and 2 for parts 1 and 2.

Supporting Information Available: Tables of experimental and predicted values for TEP, descriptor coefficients in all models, and individual ligand weightings in robust models as well as diagnostic plots for models. This material is available free of charge via the Internet at <http://pubs.acs.org>.

REFERENCES AND NOTES

- (1) Fey, N.; Tsipis, A.; 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.
- (2) Fey, N.; Harris, S. E.; Harvey, J. N.; Orpen, A. G. Adding Value to Crystallographically-Derived Ligand Knowledge Bases. *J. Chem. Inf. Model.* **2006**, *46* (2), 912–929.
- (3) Allen, F. H.; Taylor, R. Research applications of the Cambridge Structural Database (CSD). *Chem. Soc. Rev.* **2004**, *33*, 463–475, and references cited.
- (4) Allen, F. H. The CSD: a quarter of a million structures and rising. *Acta Crystallogr.* **2002**, *B58*, 380–388. Orpen, A. G. Applications of the Cambridge Structural Database to molecular inorganic chemistry. *Acta Crystallogr.* **2002**, *B58*, 398–406, and references cited.
- (5) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; T. N. Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–242. Berman, H. M.; Battistuz, T.; Bhat, T. N.; Bluhm, W. F.; Bourne, P. E.; Burkhardt, K.; Feng, Z.; Gilliland, G.; Iype, L.; Jain, S.; Fagan, P.; Marvin, J.; Padilla, D.; Varichandran, V.; Schneider, B.; Thanki, N.; Weissig, H.; Westbrook, J. D.; Zardecki, C. The Protein Data Bank. *Acta Crystallogr.* **2002**, *D58*, 899–907. RSCB, Protein Data Bank. <http://www.pdb.org/> (accessed Feb 22, 2005).
- (6) Bruno, I. J.; Cole, J. C.; Edgington, P. R.; Kessler, M.; Macrae, C. F.; McCabe, P.; Pearson, J.; Taylor, R. New software for searching the Cambridge Structural Database and visualizing crystal structures. *Acta Crystallogr.* **2002**, *B58*, 389–397.
- (7) Bruno, I. J.; Cole, J. C.; Kessler, M.; Luo, J.; Motherwell, W. D. S.; Purkis, L. H.; Smith, B. R.; Taylor, R.; Cooper, R. I.; Harris, S. E.; Orpen, A. G. Retrieval of Crystallographically-Derived Molecular Geometry Information. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 2133–2144. Harris, S. E.; Orpen, A. G.; Bruno, I. J.; Taylor, R. Factors Affecting d-Block Metal–Ligand Bond Lengths: Towards An Automated Library of Molecular Geometry For Metal Complexes. *J. Chem. Inf. Model.* **2005**, *45*, 1727–1748.
- (8) Bruno, I. J.; Cole, J. C.; Lommerse, J. P. M.; Rowland, R. S.; Taylor, R.; Verdonk, M. J. *Comput.-Aided Mol. Des.* **1997**, *11*, 525–537.
- (9) Defining e-Science. <http://www.nesc.ac.uk/nesc/define.html> (accessed Jun 4, 2004). Hey, T.; Trefethen, A. e-Science and its implications. *Philos. Trans. R. Soc. London A* **2003**, *361*, 1809–1825.
- (10) Martin, R. ePrints UK: Developing a national e-prints archive. <http://www.ariadne.ac.uk/issue35/martin> (accessed Feb 22, 2005). Day, M. EBank UK project scenarios and user requirements. <http://www.ukoln.ac.uk/projects/ebank-uk/requirements/scenarios.html> (accessed Feb 22, 2005).
- (11) Tolman, C. A. Steric Effects of Phosphorus Ligands in Organometallic Chemistry and Homogeneous Catalysis. *Chem. Rev.* **1977**, *77*, 313.
- (12) Fernandez, A. L.; Prock, A.; Giering, W. P. The QALE Web Site. <http://www.bu.edu/qale/> (accessed Jun 10, 2004). Wilson, M. R.; Prock, A.; Giering, W. P.; Fernandez, A. L.; Haar, C. M.; Nolan, S. P.; Foxman, B. M. π Effects Involving Rh-PZ₃ Compounds. The Quantitative Analysis of Ligand Effects (QALE). *Organometallics* **2002**, *21*, 2758–2763, and references cited.
- (13) Babji, C.; Poë, A. J. Deconstruction of Taft's σ^* parameter: QSAR meets QALE. *J. Phys. Org. Chem.* **2004**, *17*, 162–167. Henderson, W. A., Jr.; Streuli, C. A. The Basicity of Phosphines. *J. Am. Chem. Soc.* **1960**, *82*, 5791–5794. Bodner, G. M.; May, M. P.; McKinney, L. E. A Fourier Transform C-13 NMR—Study of the Electronic Effects of Phosphorus, Arsenic and Antimony Ligands in Transition-Metal Carbonyl Complexes. *Inorg. Chem.* **1980**, *19*, 1951–1958. Orpen, A. G.; Connelly, N. G. Structural Evidence for the Participation of P–X σ^* Orbitals in Metal-PX₃ Bonding. *J. Chem. Soc., Chem. Commun.* **1985**, 1310–1311. Lever, A. B. P. Electrochemical Parametrization of Metal Complex Redox Potentials, Using the Ru(III)/Ru(II) Couple to Generate a Ligand Electrochemical Series. *Inorg. Chem.* **1990**, *29*, 1271–1285. Orpen, A. G.; Connelly, N. G. Structural Systematics: Role of P–A σ^* Orbitals in M–P π -Bonding in Redox-Related Pairs of M-PA₃ Complexes. *Organometallics* **1990**, *9*, 1206–1210. Chen, L.; Poë, A. J. Associative reactions of metal carbonyl clusters:

- systematic kinetic studies of some ruthenium and other clusters. *Coord. Chem. Rev.* **1995**, *143*, 265–295. Fielder, S. S.; Osborne, M. C.; Lever, A. B. P.; Pietro, W. J. First-Principles Interpretation of Ligand Electrochemical Parameters. *J. Am. Chem. Soc.* **1995**, *117*, 6990–6993. Serron, S.; Nolan, S. P.; Moloy, K. G. Solution Thermochemical Study of Tertiary Phosphine Ligand Substitution Reactions in the $\text{RhCl}(\text{CO})(\text{PR}_3)_2$ System. *Organometallics* **1996**, *15*, 4301–4306. Smith, J. M.; Coville, N. J.; Cook, L. M.; Boeyens, J. C. A. Steric Parameters of Conformationally Flexible Ligands from X-ray Structural Data. 1. $\text{P}(\text{OR})_3$ Ligands in Equivalent Ligand Environments. *Organometallics* **2000**, *19*, 5273–5280. Smith, J. M.; Coville, N. J. Steric Parameters of Conformationally Flexible Ligands from X-ray Structural Data. 2. $\text{P}(\text{OR})_3$ Ligands in Multiple Ligand Environments. *Organometallics* **2001**, *20*, 1210–1215.
- (14) Bartik, T.; Himmler, T.; Schulte, H.-G.; Seevogel, K. Substituenteneinflüsse auf die Basizität von Phosphorliganden in $\text{R}_3\text{P}-\text{Ni}(\text{CO})_3$ -Komplexen. *J. Organomet. Chem.* **1984**, *272*, 29–41.
 - (15) Joerg, S.; Drago, R. S.; Sales, J. Reactivity of Phosphorus Donors. *Organometallics* **1998**, *17*, 589–599. Drago, R. S.; Joerg, S. Phosphine E_B and C_B Values. *J. Am. Chem. Soc.* **1996**, *118*, 2654–2663.
 - (16) Buntin, K. A.; Farrar, D. H.; Poß, A. J. Potential Energy Surfaces in Transition States for Associative Reactions of Metal Carbonyl Clusters: Reactions of $\text{Rh}_4(\text{CO})_{12}$ with P-Donor Nucleophiles. *Organometallics* **2003**, *22*, 3448–3454, and references cited.
 - (17) Cooney, K. D.; Cundari, T. R.; Hoffman, N. W.; Pittard, K. A.; Temple, M. D.; Zhao, Y. A Priori Assessment of the Stereoelectronic Profile of Phosphines and Phosphites. *J. Am. Chem. Soc.* **2003**, *125*, 4318–4324.
 - (18) Brown, T. L. A Molecular Mechanics Model of Ligand Effects. 3. A New Measure of Ligand Steric Effects. *Inorg. Chem.* **1992**, *31*, 1286–1294. Brown, T. L.; Lee, K. J. Ligand steric properties. *Coord. Chem. Rev.* **1993**, *128*, 89–116. Howard, S. T.; Platts, J. A. Relationship between Phosphine Proton Affinities and Lone Pair Density Properties. *J. Phys. Chem.* **1995**, *99*, 9027–9033. Howard, S. T.; Foreman, J. P.; Edwards, P. G. Electronic Structure of Aryl- and Alkylphosphines. *Inorg. Chem.* **1996**, *35*, 5805–5812. Steinmetz, W. E. A CoMFA Model of Steric and Electronic Effects of Phosphorus Ligands. *Quant. Struct.-Act. Relat.* **1996**, *15*, 1–6. Bubel, R. J.; Douglass, W.; White, D. P. Molecular Mechanics-Based Measures of Steric Effects: Customized Code to Compute Ligand Repulsive Energies. *J. Comput. Chem.* **2000**, *21*, 239–246. Landis, C. R.; Feldgus, S.; Uddin, J.; Wozniak, C. E.; Moloy, K. G. Computational Assessment of the Effect of $\sigma-\pi$ Bonding Synergy and Reorganisation Energies on Experimental Trends in Rh–P Bond Enthalpies. *Organometallics* **2000**, *19*, 4878–4886. Senn, H. M.; Deubel, D. V.; Bloechl, P. E.; Togni, A.; Frenking, G. Phosphane lone-pair energies as a measure of ligand donor strengths and relation to activation energies. *J. Mol. Str. (THEOCHEM)* **2000**, *506*, 233–242. Frenking, G.; Wichmann, K.; Fröhlich, N.; Grobe, J.; Golla, W.; Van, D. L.; Krebs, B.; Läge, M. Nature of the Metal–Ligand Bond in $\text{M}(\text{CO})_5\text{PX}_3$ Complexes (M = Cr, Mo, W; X = H, Me, F, Cl): Synthesis, molecular structure, and quantum-chemical calculations. *Organometallics* **2002**, *21*, 2921–2930.
 - (19) Perrin, L.; Clot, E.; Eisenstein, O.; Loch, J.; Crabtree, R. H. Computed Ligand Electronic Parameters from Quantum Chemistry and Their Relation to Tolman Parameters, Lever Parameters, and Hammett Constants. *Inorg. Chem.* **2001**, *40*, 5806–5811.
 - (20) Björsvik, H.-R.; Hansen, U. M.; Carlson, R. Principal Properties of Monodentate Phosphorus Ligands. Predictive Model for the Carbonyl Absorption Frequencies in $\text{Ni}(\text{CO})_3\text{L}$ Complexes. *Acta Chem. Scand.* **1997**, *51*, 733–741.
 - (21) Köhl, O. Predicting the net donating ability of phosphines – do we need sophisticated theoretical methods? *Coord. Chem. Rev.* **2005**, *249*, 693–704, and references cited.
 - (22) Burello, E.; Rothenberg, G. Optimal Heck Cross-Coupling Catalysis: A Pseudo-Pharmaceutical Approach. *Adv. Synth. Catal.* **2003**, *345*, 1334–1340. Burello, E.; Marion, P.; Galland, J.-C.; Chamard, A.; Rothenberg, G. Ligand Descriptor Analysis in Nickel-Catalysed Hydrocyanation: A Combined Experimental and Theoretical Study. *Adv. Synth. Catal.* **2005**, *347*, 803–810.
 - (23) Burello, E.; Farrusseng, D.; Rothenberg, G. Combinatorial Explosion in Homogeneous Catalysis: Screening 60,000 Cross-Coupling Reactions. *Adv. Synth. Catal.* **2004**, *346*, 1844–1853.
 - (24) Mansson, R. A.; Frey, J. G.; Essex, J. W.; Welsh, A. H. Prediction of Properties from Simulations: A Re-examination with Modern Statistical Methods. *J. Chem. Inf. Model.* **2005**, *45*, 1791–1803.
 - (25) Filzmoser, P.; Croux, C. A Projection Algorithm for Regression with Collinearity. In *Classification, Clustering and Data Analysis*; Jajuga, K., Sokolowski, A., Bock, H.-H., Eds.; Springer-Verlag: Berlin, 2002; pp 227–234. Geladi, P. Some recent trends in the calibration literature. *Chemom. Intell. Lab. Syst.* **2002**, *60*, 211–224.
 - (26) Ramsey, F. L. A Fable of PCA. *Am. Statistician* **1986**, *40*, 323–324. Hadi, A. S.; Ling, R. F. Some Cautionary Notes on the Use of Principal Components Regression. *Am. Statistician* **1998**, *52*, 15–19.
 - (27) Eriksson, L.; Anti, H.; Holmes, E.; Johansson, E.; Londstedt, T.; Schockcor, J.; Wold, S. Partial Least Squares (PLS) in Cheminformatics. In *Handbook of Chemoinformatics*; Gasteiger, J., Ed.; Wiley-VCH GmbH & Co. KGaA: Weinheim, 2003; Vol. 4, pp 1134–1166, and references cited.
 - (28) Hawkins, D. M. The Problem of Overfitting. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1–12.
 - (29) Draper, N. R.; Smith, H. *Applied Regression Analysis*; John Wiley & Sons: 1998.
 - (30) Tibshirani, R. Regression Shrinkage and Selection via the Lasso. *J. R. Statist. Soc. B* **1996**, *58*, 267–288.
 - (31) Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*; Chapman & Hall: 2001.
 - (32) Shao, J. Linear Model Selection by Cross-Validation. *J. Am. Stat. Assoc.* **1993**, *88*, 486–494.
 - (33) Shao, J. Bootstrap Model Selection. *J. Am. Stat. Assoc.* **1996**, *91*, 655–665.
 - (34) Wisnowski, J. W.; Simpson, J. R.; Montgomery, D. C.; Runger, G. C. Resampling methods for variable selection in robust regression. *Comput. Stat. Data Anal.* **2003**, *43*, 341–355.
 - (35) Davison, A. C.; Hinkley, D. V. *Bootstrap Methods and their Application*; Cambridge University Press: 1997.
 - (36) Dias, P. B.; Piedade, M. E. M. d.; Simes, J. A. M. Bonding and energetics of P(III) ligands in TM complexes. *Coord. Chem. Rev.* **1994**, *135/136*, 737–807.
 - (37) Gillespie, A. M.; Pittard, K. A.; Cundari, T. R.; White, D. P. Semiempirical Quantum Mechanics and the Quantification of Ligand Electronic Parameters. *Internet Elec. J. Mol. Des.* **2002**, *1*, 242–251.
 - (38) Dorta, R.; Stevens, E. D.; Scott, N. M.; Costabile, C.; Cavallo, L.; Hoff, C. D.; Nolan, S. P. Steric and Electronic Properties of N-Heterocyclic Carbenes (NHC): A Detailed Study of Their Interaction with $\text{Ni}(\text{CO})_4$. *J. Am. Chem. Soc.* **2005**, *127*, 2485–2495.
 - (39) R Development Core Team, R: *A language and environment for statistical computing*; R Foundation for Statistical Computing: Vienna, Austria, 2004. <http://www.R-project.org>.
 - (40) Baber, R. A.; Orpen, A. G.; Pringle, P. G.; Wilkinson, M. J.; Wingad, R. L. Special effects of ortho-isopropylphenyl groups. Diastereoisomerism in platinum(II) and palladium(II) complexes of helically chiral PAr_3 ligands. *Dalton Trans.* **2005**, 659–667.
 - (41) Yohai, V. J. High breakdown-point and high efficiency robust estimates for regression. *Annals Statistics* **1987**, *15*, 642–656.
 - (42) Slater, J. C. *Quantum Theory of Molecules and Solids, Vol. 4: The Self-Consistent Field for Molecules and Solids*; McGraw-Hill: New York, 1974. Becke, A. D. Density-Functional Exchange-Energy Approximation with Correct Asymptotic-Behaviour. *Phys. Rev. A* **1988**, *38*, 3098–3100. Perdew, J. P.; Zunger, A. Self-interaction correction to density functional approximations for many-electron systems. *Phys. Rev. B* **1981**, *23*, 5048–5079. Perdew, J. P. Density-functional approximation for the correlation-energy of the inhomogeneous electron gas. *Phys. Rev. B* **1986**, *33*, 8822–8824. Perdew, J. P. Correction. *Phys. Rev. B* **1986**, *34*, 7406.
 - (43) Jaguar, version 4.0; Schrödinger Inc.: Portland, Oregon, 2000. Jaguar, version 5.0; Schrödinger LLC: Portland, Oregon, 2002.

CI600212T