DOI: 10.1002/adsc.200505220

Topological Mapping of Bidentate Ligands: A Fast Approach for Screening Homogeneous Catalysts

Enrico Burello, Gadi Rothenberg*

van 't Hoff Institute for Molecular Sciences, University of Amsterdam, Nieuwe Achtergracht 166, 1018 WV Amsterdam, The Netherlands
Fax: (+31)-20-525-5604, e-mail: gadi@science.uva.nl

Received: May 27, 2005; Accepted: September 12, 2005

Supporting Information for this article is available on the WWW under http://asc.wiley-vch.de or from the author.

Abstract: The challenge of predicting the catalytic properties of large libraries of homogeneous catalysts is introduced. A new concept is presented that combines fundamental chemical topology principles with linear and non-linear statistical analysis. These models can predict key properties of bidentate ligand-metal complexes, namely the ligand bite angle and the backbone flexibility, without computing any 3D structural parameters and without using any force fields or any quantum mechanics. The model's performance is demonstrated on a set of 80 biphosphine and biphosphite complex crystal structures. With non-linear methods, the prediction accuracy is 93% for bite angles and 90% for flexibilities. The link between the descriptors and the ligand structures, and the possibilities that this approach opens in the search for new homogeneous catalysts are discussed.

Keywords: catalyst discovery; data mining; graph theory; high-throughput screening; homogeneous catalysis; ligand design

Introduction

Homogeneous catalysis using metal-ligand complexes is one of the most promising routes to sustainable chemistry, which in turn is essential for a sustainable society.[1–4] The problem is that finding the optimal homogeneous catalyst for a given reaction is far from trivial.[5] Robot synthesisers can already perform hundreds of reactions per day, but this is but a drop in the ocean compared to the number of possible molecules that could function as catalyst complexes.[6–9] To succeed in the search for new catalysts, we need to complement the high-throughput experimental set-ups with fast and robust modelling tools.[10–13] These should be able to sift through virtual libraries containing potentially active ligand-metal complexes, singling out the most likely candidates.[14]

In pharmacological studies and drug design, topological modelling and screening of large virtual libraries is an indispensable tool. Quantitative structure-activity and structure-property relationship models (QSAR/QSPR) are used to predict the pharmacological properties of drug candidates. The key advantage of such topological models is their low computational cost. Conversely, in homogeneous catalysis, the majority of the modelling studies use quantum mechanics and/or molecular mechanics models to elucidate mechanistic information. These models are more accurate, but their computational costs are high. In this paper, we will show that by combining chemical principles, topological descriptors, and statistical analysis, it is possible to predict important parameters in very large catalyst libraries with little computational effort. Specifically, we will show that it is possible to predict with good accuracy key properties of bidentate ligand-metal complexes, namely the ligand bite angle and the backbone flexibility, without computing any 3D structural parameters and without using any force fields or any quantum mechanics. These models will be demonstrated on eighty biphosphine and biphosphite complex crystal structures.

Results

The catalytic activity of bidentate ligands can be characterised using two parameters: the angle a ligand forms with the metal centre (the bite angle) and the ligand flexibility profile.[15–17] The bite angle indicates which conformation a ligand may adopt in the catalytic cycle, and can be measured from X-ray experiments or calculated using, e.g., molecular mechanics. The ligand's flexibility denotes the range of bite angles a ligand can adopt, if conformations with energies slightly above that of the minimised structure (to within 3 kcal mol^{-1}) are considered.

These two parameters have been used in several studies to select or modify ligands for catalytic reactions, showing that good results can be achieved by combining simple molecular mechanics force fields and mechanistic information.[18,19] However, when large libraries of ligands ($>10^5$ compounds) are considered, it takes too long to calculate all the structures, even when using molecular mechanics. Here we propose a new method to predict ligand bite angles and flexibilities, using a set of properties derived from the molecular topology. To avoid lengthy energy minimisation calculations that require three-dimensional atom coordinates, we derive a set of two-dimensional descriptors to explain the variance in the bite angles and flexibilities among different bidentate ligands.

Ligand Data Generation and Descriptor Development

First, we assembled a data set of known ligand structures, complete with the figures of merit (in this case, the ligand bite angles and the flexibility profiles), that will serve as a test and validation set. For this, we used the crystal structures of 80 ligand-Pd complexes retrieved from the Cambridge Crystallographic Database.[20] All of these complexes consist of biphosphine or biphosphite ligands that bind to a Pd^{2+} ion, with two Cl^- counterions. The bite angles were measured directly from the complexes' crystal structures. The flexibility profiles were calculated by fixing the ligand's bite angle to standard values (90°, 100°, 110°, 120° and 130°) and minimising the constrained structures using the MMFF molecular mechanics force field implemented in the Spartan software. Figure 1 shows a selection of 20 complexes from the dataset, together with their bite angle and flexibility values (listings for all 80 structures are provided in the Supporting Information).

We then encoded this set of chemical structures numerically. For each ligand, we calculated 25 topological descriptors related to structural and metal-binding features. The ligands were considered as assemblies of two parts: the backbone that connects the P atoms, and the R groups attached to these P atoms.

Figure 2 shows a selection of the most important descriptors (a complete list of the descriptors used in this study is given in the Computational Methods section). Descriptors that pertain to the backbone structure include the number of bonds that make up the minimum P_1-P_2 connectivity path (red), the alternative molecular P_1-P_2 connectivity paths (blue) and the number and position of free-rotating bonds. The most relevant descriptor for the residue R groups is the sum of mass units of the atoms that are less than three bonds distant from the ligating P atom ($SAM_{R<3}$, shown in green).

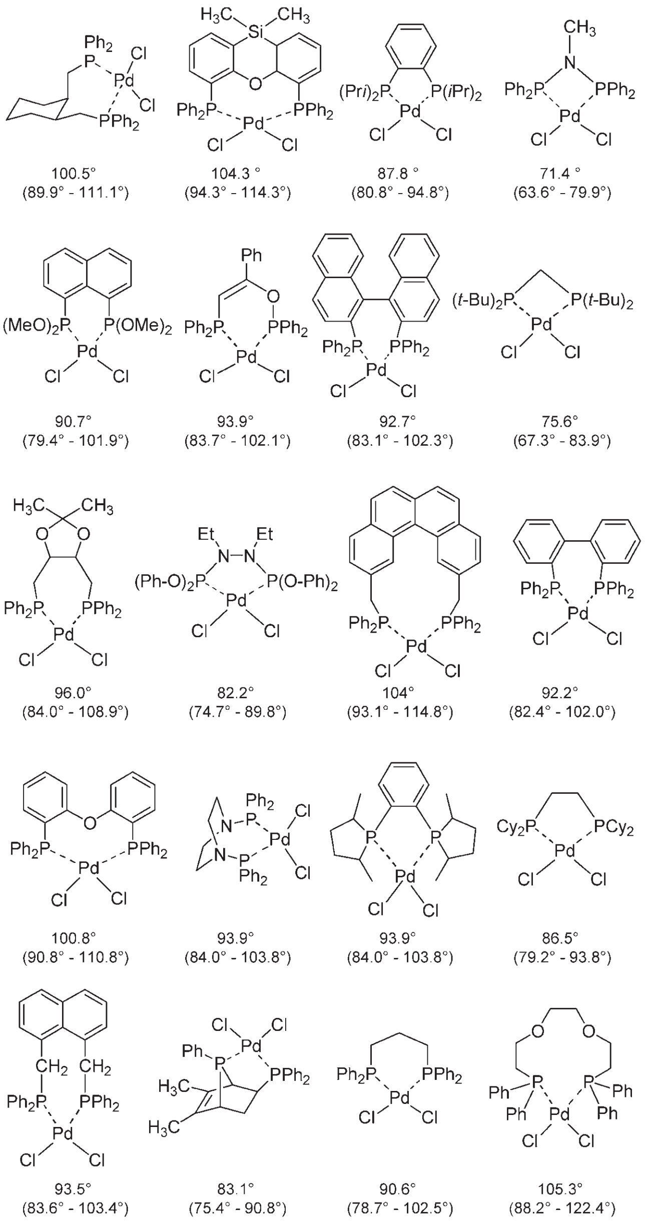


Figure 1. Bite angle and flexibility ranges of twenty selected bidentate ligand structures retrieved from the Cambridge Crystallographic Database. The P–Pd–P bite angles are calculated from the atoms' crystallographic coordinates. The flexibility corresponds to the range of bite angles a ligand can adopt if conformations with energies within 3 kcal mol^{-1} of that of the minimised structure are considered.

In addition to the molecular connectivity data, each pathway connecting the ligating phosphorus atoms P_1 and P_2 (denoted as D_n) is also described by the sum of bond valences (v_{Dn}), and bond distances ($\mathring{A}_{Dn}$), as well as by the sum of mass units of the atoms it comprises (SAM_{Dn}).

The longer P_1-P_2 paths (i.e., $D_2, D_3,...D_n$) represent alternative routes that connect the two ligating P atoms, and contain additional information regarding the back-

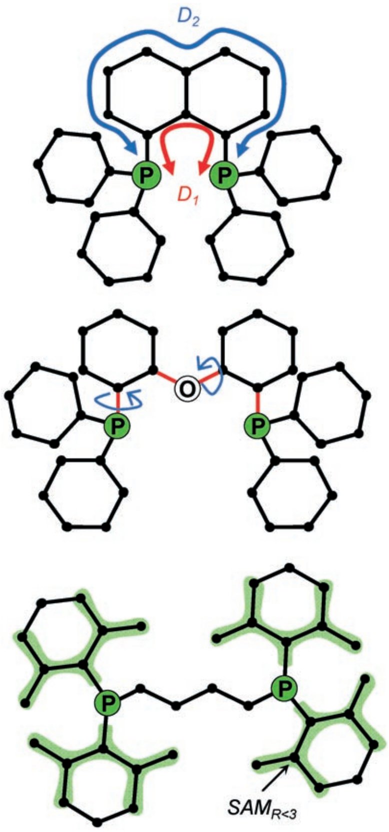


Figure 2. Examples of topological descriptors calculated on backbone and R groups of bidentate ligands. The red fragment indicates the minimum P_1-P_2 connectivity path (D_1). The second P_1-P_2 connectivity (D_2) is shown in blue. Free-rotating bonds are shown in red. The R group descriptor $SAM_{R<3}$ (green) pertains to the sum of mass units of atoms that are connected within three bonds from the P atoms.

bone structure. The ratio between the minimum P_1-P_2 path, D_1, and one of the alternative paths, D_n, can be thought of as a measure of the mutual position of the P atoms in the cyclic fragment formed by D_1+D_n. Here we used the descriptors b_{D1}/b_{D2}, b_{D1}/b_{D3}, v_{D1}/v_{D2} and v_{D1}/v_{D3} to calculate the position of P atoms in the first and second cyclic fragments using the sum of bonds (b) and atom valences (v) as weights (a detailed explanation is given in the Computationl Methods session).

Other backbone descriptors pertain to fragment size and rotational features. Size is encoded in terms of the sum of atomic mass units (SAM_{BB}) and the sum of all P_1-P_2 connectivity paths (ΣD_n). The rotational features of the minimum path D_1, in the backbone and in the P-BB linkers, are described by r_{D1}, r_{BB}, and $r_{(P\text{-}BB)}$, respectively.

The descriptors of the R residue groups include the sum of atomic mass units, (SAM_R), the sum of bonds enclosed in the P_1-P_1 self-returning paths, $S_{(P1,P2)}$, and the sum of mass units of atoms less than three bonds distant from the P atoms ($SAM_{R<3}$).

Descriptor Selection and Cluster Analysis

The initial set of 25 descriptors was reduced to 19 variables using a two-step approach: a Principal Component Analysis (PCA) followed by a hierarchical cluster analysis[21,22] on the loadings of the first thirteen principal components, representing 99% of the total variance in the data set (5 PCs are required to take into account 70% of the variance in the data set). This method clusters the descriptors according to the type of information they pertain to. This is useful, because it is tantamount to clustering according to their "chemical meaning". The resulting dendogram (Figure 3) shows two clusters: a large one (sub-clusters C_I-C_{IV}) and a small one (C_V). Sub-cluster C_I contains descriptors related to the atom connectivity and the size of the R groups. The descriptors in sub-cluster C_{II} pertain to the size (SAM_{BB}) and non-minimal P_1-P_2 connectivities of the backbone (b_{D2}, b_{D3}, v_{D2}, v_{D3} and ΣD_n). The C_{III} sub-cluster consists of four descriptors that relate to the P_1-P_2 minimal connectivity path in terms of sum of bonds (b_{D1}), valences (v_{D1}), bond distances ($\AA_{D1}$) and size (SAM_{D1}). Sub-cluster C_{IV} includes descriptors related to the mutual positions of the P atoms in the cyclic fragments D_1+D_2 and D_1+D_3, as well as the ratio between the minimum path and the backbone sum of atomic mass units (b_{D1}/SAM_{BB}). Sub-cluster C_V contains information regarding the number of P–BB and BB–BB linkers and the ratio b_{D1}/SAM_{D1}.

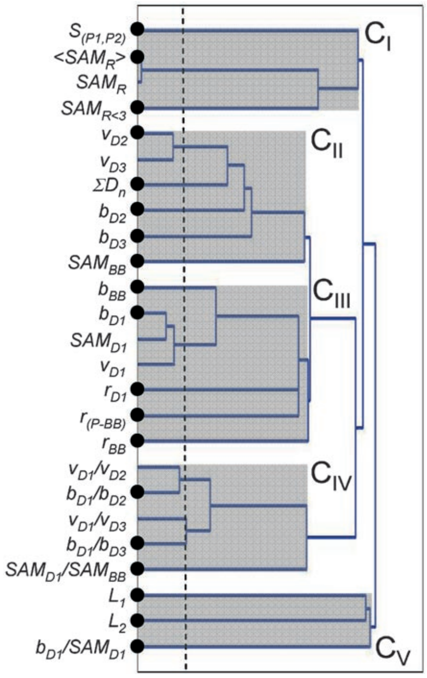


Figure 3. Variable selection dendogram obtained using hierarchical clustering, where the similarity between the variables is evaluated by describing them in terms of their first thirteen principal components. The dashed line indicates a similarity level of 80%; black circles indicate the set of 19 descriptors selected for the models.

Table 1. Prediction quality and accuracy of regression (PLS) and artificial neural networks models.

Model	X[a]	Descriptor classes[b]	Figure of merit	Correct predictions, %[c]		R^2	
				Training set[d]	Validation set[e]	Training set[d]	Validation set[e]
PLS_1	6	C_{III} C_V C_{II} C_{IV}	Bite angle	63	59	0.87	0.83
ANN_1	7	C_{III} C_{II} C_V C_{IV}	Bite angle	70	65	0.93	0.90
PLS_2	7	C_{III} C_V C_I	Flexibility	59	56	0.79	0.74
ANN_2	8	C_{III} C_V C_I	Flexibility	69	67	0.90	0.84

[a] Number of descriptors used in the model. These descriptors were selected using a genetic algorithm (for PLS) and a backward elimination method (for ANN).

[b] Classes of most relevant structure features, see figure 6.

[c] Percentage of compounds with a predicted absolute error (AE) of less than ±2.5° for the bite angle and the flexibility ranges.

[d] The training set contains 65 ligands.

[e] The validation set contains 15 ligands.

The best informative subset of descriptors was then selected by "cutting" the dendogram at a similarity level of 80% (dashed line in Figure 3). This gives a subset of 19 descriptors, i.e., one descriptor for each tree branch crossed by the dashed line.

Predicting Ligand Parameters using Linear and Non-Linear Models

After calculating the descriptors, a statistical procedure was established to relate them to the figures of merit (i.e., the bite angles and flexibilities). We used both linear and non-linear statistical tools, namely Partial Least Squares (PLS) and Artificial Neural Network (ANN) analyses. To reduce the rank of the problem, we discarded descriptors with low or no correlation with the figures of merit. This variable selection process results in more robust and simpler models that reduce the chances of overlearning.[23] The model's prediction performance was validated on a subset of structures not used in the modelling process.

Table 1 shows the performance of the PLS and ANN models, and Figure 4 shows the observed and predicted bite angles and flexibilities for the entire dataset. Both types of models performed well with training and validation sets, the latter showing a decrease in R^2 within accepted values ($0.03 < R^2_{training} - R^2_{validation} < 0.06$). The best results were obtained with ANNs, which gave prediction correlation coefficients of $R^2 = 0.93$ and $R^2 = 0.90$ for the bite angle and the flexibility, respectively. The ANNs outperformed the linear models also when considering the size of the residuals – non-linear models yielded a higher percentage of compounds predicted to within ±2.5° degrees of the experimental value for bite angles and flexibility ranges. Note the overlap between the two methods, observed both for the bite angle predictions and the flexibility profiles. This basically means that the 2D and the 3D models identify the same trends, but the fast 2D model comes with a higher price tag in the form of more noisy results.

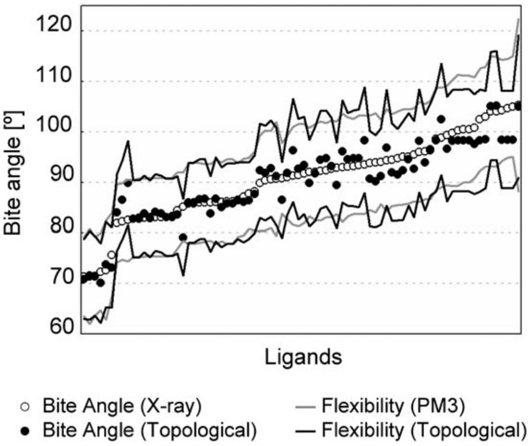


Figure 4. Observed and predicted ligand bite angles (shown as '○' and '●' symbols, respectively) and flexibility ranges using 3D (PM3) and 2D models (grey and black continuous curves, respectively) for the dataset of 80 bidentate ligands.

The most important topological descriptors for the bite angle models PLS_1 and ANN_1 were the sum of bonds in the D_1 and D_2 connectivity paths (b_{D1}, b_{D2} and b_{D1}/SAM_{D1}), the sum of P_1–P_2 connectivities (ΣD_n), and the mutual position of the P atoms in the cyclic fragments $D_1 + D_2$ and $D_1 + D_3$, (b_{D1}/b_{D2} and b_{D1}/b_{D3}). For the PLS models we calculated the influence of every descriptor using the VIP (variable importance in projection) method. The assessment of descriptor importance in ANN models was calculated using a sensitivity analysis that rates a variable according to the deterioration in modelling performance that occurs if that variable is no longer available to the model.

In the flexibility models PLS_2 and ANN_2, the important descriptors were the sum of bonds in the minimum P_1–P_2 path (b_{D1}), the number of free-rotating bonds in the backbone structure (r_{D1} and $r_{(P\text{-}BB)}$), the number of

P-backbone linkers (L_1) and the R-group descriptors ($SAM_{R<3}$ and SAM_R).[24]

Using principal components as inputs for PLS and NN analyses did not produce better models than using pure descriptors.

Discussion

The Link between Descriptors and Ligands

Descriptors found to be important in the bite angle models use only information regarding the backbone connectivities between P_1 and P_2. The sum of bonds enclosed in the minimal P_1-P_2 connectivity path (sub-cluster C_{III}) is especially significant. The easiest way to calculate the bite angle, assuming that the two P–Pd distances are equal, is by knowing the P_1-P_2 distance in three dimensions. However, from the topological point of view, the minimum P_1-P_2 connectivity path is the closest value to the P_1-P_2 distance calculated in three dimensions, so this is the best approximation one can obtain from a 2D representation. The P_1-P_2 minimum connectivity descriptors are included in all models and rank first in order of predictor importance. Indeed, the sum of bonds in the P_1-P_2 minimum path correlates significantly with both the bite angle ($R^2=0.70$) and the flexibility ($R^2=0.63$). This correlation is higher when the bite angle is smaller, but as the number of atoms in the path increases the degrees of freedom of the backbone increase,[25] and the correlation decreases.

Descriptors ΣD_n and b_{D2}, respectively the sum of P_1-P_2 connectivities and the second shortest P_1-P_2 path, rank second and third in terms of predicting importance. These descriptors provide additional information on the backbone structure and size, by considering all of the P_1-P_2 connections. The positions of the P atoms in the cyclic fragments are also involved in the modelling process, as indicated by the importance of the descriptors b_{D1}/b_{D2} and b_{D1}/b_{D3} (sub-cluster C_{IV}). All of the topological descriptors that are important for predicting the bite angle are backbone structure descriptors, indicating that it is the backbone that determines the bite angle, rather than the R groups on the P atoms, for example.

Modelling the ligands' flexibilities requires information on the P_1-P_2 minimum connectivity D_1, the number and position of free-rotating bonds in D_1, and the size of the R groups on the P atoms. Both the PLS_2 and the ANN_2 models rank the b_{D1} descriptor first for predictor importance. This is not surprising, as a ligand's flexibility is known to increase with the backbone size.[16] The number of free-rotating bonds is connected to the backbone's degree of freedom,[25] i.e., to the possibility that a backbone will adopt a stable conformation when the bite angle is varied. The descriptors r_{D1} and $r_{(P\text{-}BB)}$ are calculated by considering also the formation of possible conjugated systems like $C_{sp^{1,2}}-C_{sp^{1,2}}$ and $C_{sp^{1,2}}-N$ (*vide infra*). In this way, the model also takes into consideration the rotational restrictions of BINAP-type structures.

The number of P–BB linkers can also be interpreted in terms of the backbone's degree of freedom. For example, ligands in which a P atom is bound to the backbone *via* two linkers, will display a lower degree of flexibility (because the P atom forms an integral part of the backbone fragment).

The third important class of descriptors includes the R group descriptors (sub-cluster C_I). The $SAM_{R<3}$ descriptor is related to the steric effect of R groups with bulky substituents next to the metal centre (e.g., *ortho*-methyl in tolyl rings). This descriptor plays an increasingly important role when the R groups are bulky, and $R \rightleftharpoons R$ and $R \rightleftharpoons BB$ interactions lower the flexibility values.

Screening Very Large Ligand Libraries *in silico*

Our results show that two-dimensional descriptors can be used to predict bite angles and flexibility ranges avoiding the time-consuming energy minimisation of molecules. The power of this approach lies in the numbers it can handle! Since topological descriptors require little computational time, they allow the screening of large databases of ligands. Figure 5 shows the number of ligands that can be analysed in one hour using a desktop computer with a 2.5 GHz processor, with topological descriptors (2D), molecular mechanics (MM) and semiempirical methods (PM3).[26] The 2D descriptor calculations are three orders of magnitude faster than MM

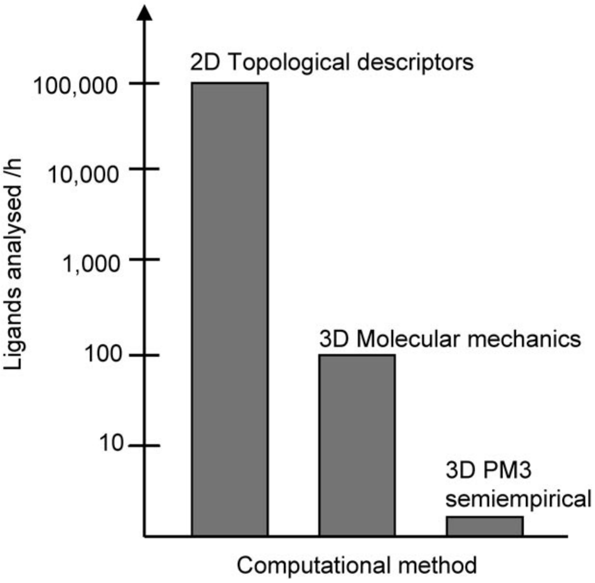


Figure 5. Comparison of the analysis capacity (number of ligands' bite angles and flexibility ranges calculated per hour) using two dimensional descriptors (2D), molecular mechanics (MM) and semi-empirical methods (PM3).

force fields – in ten hours, one can compute the bite angles and flexibilities of 1,000,000 ligands using topological descriptors, compared to only 1,000 with MM methods. The number of possible structural variations in the backbone and R groups of ligands is huge, and therefore can be handled only by calculating 2D descriptors.

The advantages of using 2D descriptors are offset by some limitations. The computational time is shorter, but the results are more scattered. This means that one should combine 2D descriptors to search for likely regions in the catalyst space, followed by more accurate models to pinpoint the best structures in the smaller regions. Moreover, some chemically important attributes, such as chirality, are almost inaccessible to 2D descriptors,[27] because two enantiomers occupy essentially the same point in the 2D descriptors space.

Conclusions

It is possible to model the L–M–L angle and the flexibility of bidentate ligands using 2D descriptors and obtain high quality predictions. By combining molecular connectivity properties with PLS and ANN statistical analyses, one can predict bite angles and flexibility ranges in the dataset, avoiding time-consuming energy minimization of molecules. Since topological descriptors require little computational time they allow the screening of large databases of ligands. Future research in our laboratory will focus on developing descriptor models for ligands bearing N, O and C ligating atoms and binding different metal centres, as well as on the application of these models to high-throughput experimental systems.

Computational Methods

Dataset Construction

The crystallographic coordinates of 80 Pd^{2+} bidentate phosphines and phosphites were downloaded from the Cambridge Crystallographic Database. The metal complexes were obtained in *.cif format and converted to *.mol files to assign aromaticity and atomic valences. The choice of structures was limited to Pd^{2+} complexes with two Cl^- counterions, to minimize the electronic and steric effects of counterions on the P–Pd–P angle, and ensure that the only structural variation in the 80 compounds arises from differences in the ligands. The dataset included bidentate ligands with angles ranging from 70° to 105°. Backbone and R group structures included both rigid and flexible scaffolds, with aromatic and aliphatic moieties (the complete dataset is given in the Supporting Information). For the flexibility calculations, we constrained the Pd–P bond distances to a fixed length, using a value of 2.29 Å, that was the average measured value of the Pd–P distances in the 80 crystal structures.

Calculating Topological Descriptors for Bidentate Ligands

As inputs for the modelling process we developed a series of 25 topological descriptors,[28] focusing on the metal-binding features of bidentate ligands (Table 2). All descriptor values were generated from connectivity tables, without using any 3D atom coordinates. Topological descriptors are derived from graph theory and describe atom connectivity in hydrogen-suppressed molecules.[29,30] A molecular graph G is a structure formed by vertexes (V) and edges (E), where $G=(E+V)$. In this depiction, atoms and bonds of molecules correspond to vertexes and edges of G. A standard representation of a graph is the adjacency matrix $\boldsymbol{A}=(a_{ij})$, defined as:

$$a_{ij}=\begin{cases}1 & if\ i\neq j \quad and\ (i,j)\in E(G)\\ \\ 0 & if\ i=j \quad and\ (i,j)\notin E(G)\end{cases}$$

where i and j are two generic nodes and $E(G)$ represents the set of edges of G. An example of a molecular graph and an adjacency matrix is given below for the ligand (–)-2,2-dimethyl-4,5-bis(diphenylphosphinomethyl)-1,3-dioxolane-P,P′ (DIOP, Figure 6). Vertexes and edges are labelled from 1 to 11 and from a to k, respectively.

Two matrices are particularly important, both of them based on the topological distance between vertexes within a graph: the distance matrix $\boldsymbol{D(G)}$ and the detour matrix $\boldsymbol{\Delta(G)}$. The first contains as values the smallest number of steps from vertex i to vertex j, and the second contains as values the longest paths. For example, for the previous graph corresponding to the DIOP ligand, matrices $\boldsymbol{D}$ and $\boldsymbol{\Delta}$ are:

$$D=\begin{bmatrix}01234434555\\10123323444\\21012212333\\32101223422\\43210123411\\43221012322\\32122101233\\43233210144\\54344321055\\54321234502\\54321234520\end{bmatrix} \qquad \Delta=\begin{bmatrix}01265567866\\10154456755\\21043345644\\65404334555\\54340434511\\54334045655\\65433401244\\76544510155\\87655621066\\65451545602\\65451545620\end{bmatrix}$$

Based on the matrix expression of molecular graphs, we can calculate the length of paths connecting any pair of atoms, i.e., a series of consecutive edges that connect two nodes and do not overlap twice with the same node. In the case of bidentate ligands, the vertexes connections we are interested in are those between P_1 and P_2. For example, the minimum D_{P1-P2} and maximum $\boldsymbol{\Delta}_{P1-P2}$ connectivity paths are calculated from distance and detour matrixes, respectively. In the case of the DIOP ligand, the respective minimum and maximum paths are five and eight bonds long.

Table 2. List of topological descriptors used in this study.

Entry	Symbol	Description
Backbone (BB) descriptors		
1	b_{D1}	Sum of bonds in the minimum P_1-P_2 connectivity path
2	v_{D1}	Sum of atom valences in the min P_1-P_2 connectivity path
3	$\AA_{D1}$	Sum of bond distances in the min P_1-P_2 connectivity path
4	SAM_{D1}	Sum of atomic mass units of the P_1-P_2 min connectivity path
5	r_{D1}	Sum of rotational bonds in the min P_1-P_2 connectivity path
6	b_{BB}	Sum of bonds in the backbone P_1-P_2 min connectivity path
7	r_{BB}	Sum of free-rotating bonds in the backbone P_1-P_2 min connectivity path
8	$r_{(P\text{-}BB)}$	Sum of free-rotating bonds in the linkers (P–BB)
9	b_{D2}	Sum of bonds in the 2^{nd} shortest P_1-P_2 connectivity path
10	b_{D3}	Sum of bonds in the 3^{rd} shortest P_1-P_2 connectivity path
11	v_{D2}	Sum of atoms valences in the 2^{nd} shortest P_1-P_2 connectivity path
12	v_{D3}	Sum of atom valences in the 3^{rd} shortest P_1-P_2 connectivity path
13	ΣD_n	Sum of all P_1-P_2 connectivity paths
14	L_1	Sum of P–BB linkers
15	L_2	Sum of BB–BB linkers
16	SAM_{BB}	Backbone sum of atomic mass units
R group descriptors		
17	$S_{(P1,P2)}$	Sum of P_1-P_1 and P_2-P_2 self-returning paths
18	$SAM_{R<3}$	Sum of mass units of R group atoms enclosed within three bonds from P_1 and P_2.
19	SAM_R	Sum of the atomic mass units of all the R groups
20	SAM_R	Ratio between SAM_R and the total number of R groups
Other descriptors		
21	b_{D1}/b_{D2}	P_1-P_2 mutual position in fragment D_1+D_2 (the ratio between b_{D1} and b_{D2})
22	b_{D1}/b_{D3}	P_1-P_2 mutual position in fragment D_1+D_3 (the ratio between b_{D1} and b_{D3})
23	v_{D1}/v_{D2}	P_1-P_2 mutual position in fragment D_1+D_2 (the ratio between v_{D1} and v_{D2})
24	v_{D1}/v_{D3}	P_1-P_2 mutual position in fragment D_1+D_3 (the ratio between v_{D1} and v_{D3})
25	b_{D1}/SAM_{D1}	Ratio between b_{D1} and SAM_{D1}
26	SAM_{D1}/SAM_{BB}	Ratio between SAM_{D1} and SAM_{BB}

In addition to molecular connectivity data, other types of information can be encoded in each path by assigning specific weights to bonds or atoms encompassed by a connection. Such weights can be, for example, bond valences or distances. The path weight *w(p)*, then, corresponds to the sum of all weights attributed to each bond in the path:

$$w(p) = \sum_{i=1}^{k} w(v_{i-1}, v_i)$$

where *p* is the path, *w* is the weight attributed to each edge and the sum runs over all consecutive nodes v_{i-1} and v_i. In this way each P_1-P_2 connectivity (D_n) is further characterised as a sum of bond valences (v_{Dn}), and bond distances ($\AA_{Dn}$), or mass units (SAM_{Dn}) of atoms enclosed in the path.

Other backbone descriptors pertain to size and rotational features. Size is encoded in terms of sum of atomic mass units (SAM_{BB}) and sum of all P_1-P_2 connectivity paths ($\Sigma\, D_n$). The r_{D1}, r_{BB}, and $r_{(P-BB)}$ descriptors compute the number of free-rotating bonds in the minimum path D_1, in the backbone and in the P–BB linkers. These descriptors are calculated for each ligand using the following four rules defining a free-rotating bond:

The bond must be single and acyclic.

In a bond X–Y, X can be C_{sp^1}, C_{sp^2} (ethylene or aromatic), or C_{sp^3}.

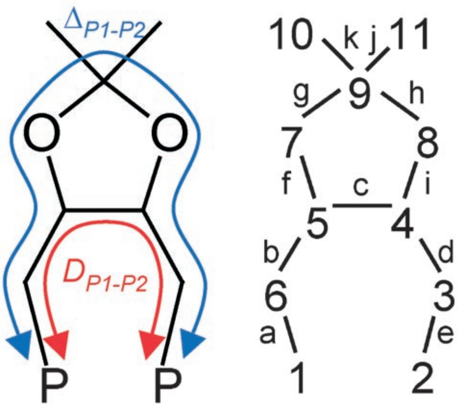
If atom X is C_{sp^3}, then atom Y can be anything.

If atom X is C_{sp^1} or C_{sp^2}, the Y can be only C_{sp^3}, O, S, P or Si. This excludes the most common conjugated systems, $C_{sp^{1,2}} - C_{sp^{1,2}}$ and $C_{sp^{1,2}} - N$, where there are strong restrictions to free rotation.

The R group descriptors include the R group sum of atomic mass units , (SAM_R), the sum of bonds enclosed in the P_1-P_1 self-returning paths, $S_{(P1,P2)}$, and the sum of mass units of atoms less than three bonds distant from the P atoms ($SAM_{R<3}$). The descriptors L_1 and L_2 define the number of linkers between P atoms and the backbone and within the backbone itself. They are calculated by finding the minimum subset of bonds that when deleted cancel all P_1-P_2 connections. L_1 counts for the number of linkers next to the P atoms thus gives information regarding the P–BB connectivity. The L_2 descriptor determines the number of linkers in the backbone structure.

Angle and Flexibility Prediction Models

Variable selection and ANN analysis were performed with STATISTICA 6.1,[31] and PLS analysis was carried out with SIMCA-P 8.0.[32] The original dataset was randomly divided into training and validation sets of 65 and 15 ligands, respectively. The training set is used to generate the predicting models. The validation set is employed to verify the soundness of the model towards new chemical entities, i.e., compounds not used in the model building or in the descriptor selection process. In ANN analysis an additional test set (15% of the training compounds) is used to stop



$$A = \begin{bmatrix} 00000100000 \\ 00100000000 \\ 01010000000 \\ 00101001000 \\ 00010110000 \\ 10001000000 \\ 00001000100 \\ 00010000100 \\ 00000011011 \\ 00000000100 \\ 00000000100 \end{bmatrix}$$

Figure 6. Molecular graph and adjacency matrix of the DIOP ligand. The phenyl rings are omitted for clarity.

the learning process and avoid over-fitting. The model's goodness-of-fit is determined by the R^2 values on both training and validation sets. For each model we report the number (X) of relevant descriptors selected, and the variable classes in order of their predictor importance, i.e., their contribution to the predicting performance. A final descriptor selection for PLS models was performed using a genetic algorithm (GA) resulting in six and seven variables for the bite angle and the flexibility, respectively. For the neural network analysis we used the three layers Perceptron network topology, and the standard backward elimination method to select relevant descriptors.[33] The ANN_1 network has seven nodes in the input layer, two nodes in the hidden layer, and one output (the bite angle). The ANN_2 network has eight descriptors as inputs in the first layer, four nodes in the hidden layer and one output, the ligand flexibility.

Supporting Information

Structures of the eighty ligand-palladium complexes used as the training and test datasets.

References and Notes

[1] M. Misono, *J. Ind. Eng. Chem.* **2004**, *10*, 1126.
[2] S. Z. Luo, Y. Y. Peng, B. L. Zhang, P. G. Wang, J. P. Cheng, *Curr. Org. Synth.* **2004**, *1*, 405.
[3] T. V. RajanBabu, A. L. Casalnuovo, T. A. Ayers, N. Nomura, J. Jin, H. Park, M. Nandi, *Curr. Org. Chem.* **2003**, **7**, 301.
[4] A. F. Littke, G. C. Fu, *Angew. Chem. Int. Ed.* **2002**, **41**, 4176.
[5] For an excellent recent monograph, see: P. W. N. M. van Leeuwen, *Homogeneous Catalysis: Understanding the Art*, Kluwer Academic Press, Amsterdam, **2004**.
[6] For a discussion on the pros and cons of high-throughput experimentation in catalysis research, see: H. F. M. Boelens, D. Iron, J. A. Westerhuis, G. Rothenberg, *Chem. Eur. J.* **2003**, *9*, 3876.
[7] O. Lavastre, *Actual. Chim.* **2000**, 42.
[8] A. Hagemeyer, B. Jandeleit, Y. M. Liu, D. M. Poojary, H. W. Turner, A. F. Volpe, W. H. Weinberg, *Appl. Catal. A: Gen.* **2001**, *221*, 23.
[9] S. Dahmen, S. Brase, *Synthesis* **2001**, 1431.
[10] A. Fernandez, C. Reyes, T. Y. Lee, A. Prock, W. P. Giering, C. M. Haar, S. P. Nolan, *J. Chem. Soc. Perkin Trans. 2* **2000**, 1349.
[11] A. L. Fernandez, C. Reyes, A. Prock, W. P. Giering, *J. Chem. Soc. Perkin Trans. 2* **2000**, 5, 1033.
[12] G. Rothenberg, H. F. M. Boelens, D. Iron, J. A. Westerhuis, *Chim. Oggi* **2003**, *21*, 80.
[13] J. A. Westerhuis, H. F. M. Boelens, D. Iron, G. Rothenberg, *Anal. Chem.* **2004**, *76*, 3171.
[14] E. Burello, P. Marion, J. C. Galland, A. Chamard, G. Rothenberg, *Adv. Synth. Catal.* **2005**, *347*, 803.
[15] C. P. Casey, G. T. Whiteker, *Isr. J. Chem.* **1990**, *30*, 299.
[16] P. Dierkes, P. W. N. M. Van Leeuwen, *J. Chem. Soc. Dalton Trans.* **1999**, 1519.
[17] L. A. van der Veen, P. C. J. Kamer, P. W. N. M. van Leeuwen, *Cattech* **2002**, *6*, 116.
[18] R. P. J. Bronger, P. C. J. Kamer, P. W. N. M. van Leeuwen, *Organometallics* **2003**, *22*, 5358.
[19] K. A. Lenero, M. Kranenburg, Y. Guari, P. C. J. Kamer, P. W. N. M. van Leeuwen, S. Sabo-Etienne, B. Chaudret, *Inorg. Chem.* **2003**, *42*, 2859.
[20] http://www.ccdc.cam.ac.uk.
[21] L. Eriksson, E. Johansson, *Chemom. Intell. Lab. Sys.* **1996**, *34*, 1.
[22] E. Burello, G. Rothenberg, *Adv. Synth. Catal.* **2003**, **345**, 1334.
[23] E. Burello, D. Farrusseng, G. Rothenberg, *Adv. Synth. Catal.* **2004**, *346*, 1845.
[24] We also checked the possibility, as suggested by one referee, of normalising the flexibility by the bite angle. However, as we show here, the flexibility is correlated with the bite angle (the higher the bite angle, the more the ligand). This additional descriptor, therefore, does not explain more variance in the dataset.
[25] Note that the term "degree of freedom" does not mean here "one of 3N degrees of freedom of a molecule" but rather pertains to the flexibility of the backbone.

The word "flexibility" itself is not used to avoid confusion with the flexibility figure of merit.

[26] While CPUs are probably going to improve in the next decade, it is unlikely that the relative analysis time per molecule for different methods is going to change much.

[27] For a study on computing the chirality of steroids using molecular graphs see A. Golbraikh, D. Bonchev, A. Tropsha, *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 147.

[28] R. Todeschini, R. Cazar, E. Collina, *Chemometrics Intell. Lab. Syst.* **1992**, *15*, 51.

[29] For a introductory monograph to graph theory, see: R. Diestel, *Graph Theory*, Vol. 173, Springer Verlag, New York, **2000**.

[30] P. D. Iedema, H. C. J. Hoefsloot, *Macromol. Theor. Simul.* **2001**, *10*, 855.

[31] Statistica is distributed by StatSoft, Inc., 2300 East 14th Street, Tulsa, OK 74104, USA. http://www.statsoft.com.

[32] Simca is distributed by Umetrics AB, Umeå, Sweden, http://www.umetrics.com.

[33] D. L. Massart, B. G. M. Vandeginste, L. M. C. Buydens, S. De Jong, P. J. Lewi, J. Smeyers-Verbeke, *Handbook of Chemometrics and Qualimetrics*, Elsevier, Dordrecht, **1997**.