Acta Chemica Scandinavica, 1997: 51: 733–741
Printed in UK – all rights reserved

Copyright © Acta Chemica Scandinavica 1997
ACTA
CHEMICA SCANDINAVICA
ISSN 0904–213X

Principal Properties of Monodentate Phosphorus Ligands. Predictive Model for the Carbonyl Absorption Frequencies in $Ni(CO)_3L$ Complexes

Hans-René Bjørsvik,[a,*] Unni Marie Hansen[b] and Rolf Carlson[c]

[a]Borregaard Synthesis, PO Box 162, N-1701 Sarpsborg, Norway, [b]Department of Organic Chemistry, Umeå University, S 90 187 Umeå, Sweden and [c]Department of Chemistry, Institute of Mathematical and Physical Sciences, University of Tromsø, N-9037 Tromsø, Norway

Bjørsvik, H.-R., Hansen, U. M. and Carlson, R., 1997. Principal Properties of Monodentate Phosphorus Ligands. Predictive Model for the Carbonyl Absorption Frequencies in $Ni(CO)_3L$ Complexes. – Acta Chem. Scand. 51: 733–741. © Acta Chemica Scandinavica 1997.

By use of the principal component analysis method, the principal properties of 45 different monodentate phosphorus ligands, each described by fifteen molecular descriptors, have been calculated. A principal component model composed of five principal components was determined according to cross-validation and accounted for 90% of the variance in the original data table. This model divided the P ligands into two classes, a small one which contained six polyhalogenated P ligands, and another which contained 39 different P ligands. Another principal component analysis was carried out on the large class of P ligands. This model described 90.5% of the total variance by using four principal components. Among these a subset of compounds were used for validating the relevance of the molecular descriptor variables. Two experimentally measured IR-frequencies, the carbonyl frequencies ν_{A1} and ν_E were used as independent variables and correlated to the molecular descriptors using the partial least-squares regression (PLSR) method. The PLSR models showed that the calculated molecular descriptors contained relevant information concerning the investigated compounds to be used for deriving the principal properties of the monodentate phosphorus ligands. The PLSR models for the two different IR frequencies were used to predict the CO frequencies for P ligands where the IR frequencies have not been reported in the literature.

An expanding and important field of synthetic organic chemistry is application of transition metals as catalysts.[1] A general feature of such catalysts is that the transition metals are stabilised by certain ligands of which phosphines are the most important single class. In order to achieve a desired synthetic transformation several conditions must be considered, such as which transition metal will be the best; which ligand will be the most suitable and which solvent to employ.

For newly discovered reactions, details of the reaction mechanisms are often not known with certainty. Hence, deduction from theoretical considerations based upon reaction mechanisms cannot be used for determining which ligand would be the best choice. Any conclusions in these directions must be inferred from experimental observations. To this end, screening experiments with different types of ligand are often carried out. In the catalytic process, the substrate to be converted interacts with the transition metal, whereupon bond breaking and bond formation occurs. In these steps it is reasonable to assume that the donating properties of the P ligand, as well as back-donation from the metal, will play roles for fine-tuning the catalytic process. It is also reasonable to assume that effects of steric congestion intervene.

This paper present a multivariate characterisation of 45 different monodentate phosphorus ligands. It is shown how such a characterisation furnishes methods for the systematic selection of test ligands for screening experiments when the objective is to find a suitable ligand. The descriptors used to characterise the ligands were also used to predict the vibrational carbonyl frequencies of phosphine-coordinated nickel carbonyl complexes. The method is based on multivariate characterisation in principal component analysis of molecular descriptors of the ligands. The latent variables thus determined are called principal properties.[2,3]

Description of the ligands

Data. The different monodentate phosphorus ligands, in total 45, were characterised by a set of molecular

[*] To whom correspondence should be addressed.

descriptors which describe a physical or chemical property. A general problem in the present type of investigation is the lack of consistency in measured data from literature sources. To overcome these difficulties, the present work is based on descriptors determined using computational chemistry. The following descriptors were obtained for phosphorus ligands and used in the multivariate data analysis: H_f/kcal mol^{-1}, the heat of formation; μ/debye, the dipole moment; χ/eV, the hardness; η/eV, the absolute electronegativity; $\varepsilon_{\text{HOMO}}$/eV, the energy of the highest occupied molecular orbital; $\varepsilon_{\text{LUMO}}$/eV, the energy of the lowest unoccupied molecular orbital; $\varepsilon_{\text{LUMO}+1}$/eV, the energy of the level next to $\varepsilon_{\text{LUMO}}$; $\varepsilon_{\text{LUMO}+2}$/eV, the energy of the level two up from $\varepsilon_{\text{LUMO}}$; $\varepsilon_{\text{LUMO}+3}$/eV, the energy of the level three up from $\varepsilon_{\text{LUMO}}$; δP [charge] the partial charge on phosphorus; M_W/g mol^{-1} the molecular weight; n_C, the number of carbon atoms; n_H, the number of hydrogen atoms; psA/Å^2, the polar surface area; vdWA/Å^2, the van der Waals surface; ν_{A1}/cm^{-1}, the IR A1 stretching frequencies of the carbonyl in $Ni(CO)_3L$ in CH_2Cl_2, where L is a P ligand; ν_E/cm^{-1}, the IR carbonyl E stretching frequencies of $Ni(CO)_3L$ in CH_2Cl_2, where L is a P ligand.

The carbonyl frequencies ν_{A1} and ν_E were experimentally determined and compiled from the literature: Grim and McFarlane,[4] Moedritzer *et al.*,[5] Bemi *et al.*,[6] Grim *et al.*,[7] and Tolman.[8–10] The other descriptors were calculated mainly using semiempirical quantum chemistry which furnishes versatile methods for describing molecular properties. Such methods[11,12] were used for calculating the molecular descriptors: H_f, μ, $\varepsilon_{\text{HOMO}}$, $\varepsilon_{\text{LUMO}}$, $\varepsilon_{\text{LUMO}+1}$, $\varepsilon_{\text{LUMO}+2}$, $\varepsilon_{\text{LUMO}+3}$, δP, psA, and vdWA. These calculated molecular descriptors can be considered as point measurements of various molecular properties which portray different aspects of the energies and electronic properties of the characterised molecule. The hardness and absolute electronegativity were calculated according to Pearson.[13] The molecular descriptor data for the P ligands studied in the present work are summarised in Table 1. However, some numerical values are not listed, the IR frequencies ν_{A1} and ν_E, due to lack of literature data. Those ligands where there are no data are still of interest, since within the ranges of substitution there is a certain similarity with the ligands which also have numerical values for the IR frequencies. Thus, it is reasonable to expect that predictions carried out for these P ligands will give satisfactory results. Furthermore, it is reasonable to assume that molecular descriptors which depend on the *same* intrinsic properties of the molecule will be correlated over the set of compounds, while other molecular descriptors which depend on *different* intrinsic properties will be uncorrelated or only weakly correlated over the whole set of compounds. By using the principal component analysis (PCA) method, these features will be taken into account. Hence, the PCA modelling will reveal which different underlying intrinsic properties may influence the experimental results.

Methods and results

Principal component analysis (PCA). When selecting test compounds for screening in experimental studies, it is desirable that the set of selected items span a sufficiently large range of variation of the properties of the test compounds. If this is not fulfilled, potentially useful new procedures run the risk of being overlooked due to too narrow a choice. If a series of potential test compounds is characterised and sufficiently described by a single property descriptor, the selection is a simple task. However, if the compounds are characterised by several property descriptors, the selection becomes difficult. In such cases, a good selection should ensure sufficient variation in all properties considered. Any data table, like Table 1, displays two types of variation: horizontally, the within-compound variation of the molecular descriptors, and vertically, the between-compound variation of the molecular descriptors.

A data-analytic method which permits the separate analysis of these features is principal component analysis. The essence of the PCA method is that the systematic variation can be portrayed by fewer variables, the principal components, than the number of descriptor variables present in the original data table. Such a procedure makes it easy to obtain an overview of the data and furnishes a tool for experimental design. An example is given below.

Mathematically, this involves a factorisation of the original data matrix X, into means ($\bar{x}_k$), the principal component scores (t_{ia}), which displays the between-compound variation, the principal component loading (p_{ak}), which describes the within-compound variation of the descriptors, and residuals (ε_{ik}) mathematically described by eqn. (1),

$$x_{ik} = \bar{x}_k + \sum_{a=1}^{A} t_{ia} p_{ak} + \varepsilon_{ik} \qquad (1)$$

where A denotes the number of significant principal components determined according to, e.g., cross-validation.

The absolute value of the loading, p_{ak}, tells how much a certain descriptor variable contributes to the ath principal component, whereas the signs give information as to whether the descriptor variables are negatively or positively correlated with the principal component. Detailed accounts are given in Ref. 14.

Cross validation (CV). The essence of the cross-validation method is to determine the model complexity of the principal component analysis (PCA) and partial least-squares regression (PLS) as well as to estimate the expected prediction error level. The cross-validation method is composed of a series of calibrations and predictions. For each of the calibrations in the PLS, some objects (samples), N, are kept outside the original

Table 1.

No.	P-Ligand[a]	H_f/ kcal mol^{-1}	μ/ D	η/ eV	χ/ eV	ε_{HOMO}/ eV	ε_{LUMO}/ eV	ε_{LUMO+1}/ eV	ε_{LUMO+2}/ eV	ε_{LUMO+3}/ eV	δP (Charge)	M_W/ g mol^{-1}	n_C	n_H	psA/ Å^2	vdWA/ Å^2	ν_{A1}/ cm^{-1}	ν_E cm^{-1}
1	PMe_2CF_3	−160.23	3.09	5.0259	4.9209	−9.9468	0.1049	1.2850	1.6970	3.2840	0.4635	130.0	3	6	4.9	144	2081	2004
2	$P(OEt)_3$	−204.90	2.36	5.6420	3.9113	−9.5533	1.7307	1.9810	1.9570	3.8450	0.8345	166.2	6	15	25.9	244	2076	1996
3	PEt_3	−28.21	1.45	5.2274	3.5669	−8.7943	1.6606	2.5900	2.6020	3.8560	0.3510	118.2	6	15	3.5	192	2062	1978
4	PBu_3	−71.46	1.25	5.3100	3.6280	−8.9380	1.6820	2.5480	2.5650	3.7590	0.3570	202.3	12	27	2.7	318	2060	1976
5	P(cy-hx)$_3$	−61.05	1.32	4.7989	3.6583	−8.4572	1.1405	2.6030	2.6680	3.5830	0.3086	280.4	18	33	1.9	349	2056	1973
6	$PHPh_2$	67.75	1.45	4.4958	4.2178	−8.7136	0.2780	0.2940	0.4140	0.4980	0.5415	186.2	12	11	8.1	234	2073	1995
7	$PPhEt_2$	13.47	1.19	4.5777	4.2583	−8.8359	0.3194	0.4820	1.7430	2.5690	0.4981	166.2	10	15	3.2	234	2064	1982
8	PPh_2Me	70.59	1.21	4.3237	4.0438	−8.3675	0.2798	0.3470	0.4680	0.5600	0.6926	200.2	13	13	5.2	250	2067	1987
9	PPh_2vin	101.45	0.92	4.3435	4.0356	−8.3791	0.3080	0.3530	0.5220	0.5610	0.7555	212.2	14	13	4.9	268	2069	1990
10	PPh_3	99.22	1.07	4.2540	3.9940	−8.2480	0.2600	0.2660	0.4160	0.5360	0.8011	262.3	18	15	4.6	313	2069	1990
11	PBz_3	67.07	1.47	4.4440	4.2970	−8.7410	0.1470	0.3260	0.3800	0.4820	0.4309	304.4	21	21	4.0	365	2066	1986
12	PPh_2Bz	91.15	1.11	4.1816	4.0281	−8.2096	0.1535	0.3310	0.4770	0.5170	0.6890	276.3	19	17	4.4	322	2068	1989
13	P(*p*-tol)$_3$	85.99	1.29	4.1326	3.8988	−8.0314	0.2338	0.3050	0.3990	0.5700	0.8181	304.4	21	21	4.6	377	2067	1987
14	P(*o*-$C_6H_4OMe)_3$	6.91	0.90	3.9205	3.4725	−7.3930	0.4480	0.4510	0.6220	0.7510	1.0070	352.4	21	21	18.1	409	2058	1974
15	$P(OPh)_3$	−67.31	3.04	4.7228	4.5381	−9.2608	0.1847	0.2130	0.2660	0.3100	0.8953	310.3	18	15	30.7	356	2085	2012
16	P(O-*p*-tol)$_3$	−88.31	3.52	4.6201	4.4064	−9.0265	0.2137	0.2270	0.2760	0.3510	0.8921	352.4	21	21	30.9	425	2084	2009
17	P(*p*-$C_6H_4OMe)_3$	−15.47	1.58	4.1701	3.8599	−8.0300	0.3102	0.3300	0.4420	0.4890	0.8200	352.4	21	21	28.0	414	2066	1987
18	P(*o*-tol)$_3$	93.34	1.35	4.2171	3.9360	−8.1530	0.2811	0.2860	0.4360	0.5400	0.7989	304.4	21	21	3.7	344	2067	1986
19	P(*m*-tol)$_3$	76.36	1.37	4.2319	3.9411	−8.1730	0.2908	0.3270	0.4510	0.5220	0.8012	304.4	21	21	5.1	381	2067	1989
20	PMe_3	−22.00	1.52	5.4019	3.5310	−8.9330	1.8709	2.6430	2.6440	4.2970	0.4104	76.1	3	9	4.6	135	2064	1982
21	$PPhMe_2$	18.44	1.34	4.3683	4.0294	−8.3977	0.3388	0.5180	1.8510	2.6200	0.5702	138.1	8	11	4.6	187	2065	1982
22	PPh_2Et	57.07	1.18	4.3075	4.0423	−8.3498	0.2652	0.3660	0.4530	0.5750	0.6766	214.2	14	15	4.0	267	2067	1986
23	P(*i*-Pr)$_3$	−25.12	1.32	5.0858	3.5949	−8.6806	1.4909	2.4970	2.5870	3.8090	0.3047	160.2	9	21	2.9	230	2059	1977
24	P(*t*-Bu)$_3$	0.02	1.27	4.6871	3.6284	−8.3155	1.0587	2.5140	2.5570	3.8990	0.2280	202.3	12	27	1.5	260	2056	1971
25	$PPhBz_2$	78.58	1.35	4.3405	4.1890	−8.5295	0.1515	0.3340	0.4300	0.5490	0.5851	290.3	20	19	4.5	338	2068	1988
26	$PPh_2(OEt)$	−1.00	0.33	4.3692	4.0878	−8.4570	0.2814	0.3460	0.5200	0.5800	0.8268	230.2	14	15	7.6	291	2072	1994
27	$PPh(OEt)_2$	−101.81	3.65	4.7853	4.5809	−9.3662	0.2044	0.5250	1.7040	2.1970	0.8070	198.2	10	15	17.0	268	2074	1997
28	P(O-*o*-tol)$_3$	−89.07	1.59	4.6049	4.4345	−9.0394	0.1705	0.2270	0.2600	0.3040	0.9458	352.4	21	21	22.7	410	2084	2011
29	$P(OMe)_3$	−187.15	2.58	5.6765	3.9418	−9.6184	1.7347	1.8490	1.8830	3.9550	0.8279	124.1	3	9	27.7	177	2080	2000
30	P(*n*-Pr)$_3$	−48.48	1.45	5.2488	3.5619	−8.8107	1.6869	2.5620	2.6150	3.7190	0.3495	160.2	9	21	3.5	251	—	—
31	PPh(*i*-Pr)$_2$	15.57	1.15	4.4885	4.1268	−8.6153	0.3617	0.5150	1.5790	2.5550	0.4740	194.3	12	19	3.1	256	—	—
32	PPh_2(*i*-Pr)	64.23	1.21	4.2530	3.9724	−8.2254	0.2806	0.3360	0.4650	0.5630	0.6694	228.3	15	17	3.9	276	—	—
33	$PPhBu_2$	−11.83	1.26	4.3683	4.0294	−8.3977	0.3388	0.5290	1.6980	2.5500	0.5380	222.3	14	23	4.1	312	—	—
34	PPh_2Bu	46.62	1.15	4.3077	4.0554	−8.3631	0.2523	0.3590	0.4480	0.5600	0.6764	242.3	16	19	4.3	312	—	—
35	P(*i*-Bu)$_3$	−62.80	1.41	5.2434	3.5337	−8.7772	1.7097	2.5720	2.6230	3.7640	0.3454	202.3	12	27	3.5	302	—	—
36	PPh(cy-hx)$_2$	−8.55	1.19	4.4109	4.0782	−8.4892	0.3327	0.4950	1.5560	2.5780	0.4843	274.4	18	27	3.7	332	—	—
37	PPh_2(cy-hx)	45.91	1.07	4.3738	4.1181	−8.4919	0.2557	0.4100	0.5040	0.5280	0.6379	268.3	18	21	4.5	324	—	—
38	PBz_2Et	35.27	1.38	4.4865	4.2762	−8.7627	0.2102	0.4060	0.4340	0.6180	0.3941	242.3	16	19	3.2	308	—	—
39	$PBzEt_2$	4.22	1.49	4.5841	4.1855	−8.7696	0.3987	0.6290	1.7070	2.4590	0.3838	180.2	11	17	3.3	250	—	—
40	$PPh_2C_6F_5$	−100.78	2.86	0.0729	0.0186	−0.0915	0.0543	0.2340	0.3370	1.1710	0.8712	352.2	18	10	3.7	334	2075	1998
41	P(*p*-$C_6C_4F)_3$	−27.06	0.30	0.0727	0.0456	−0.1183	0.0271	0.0540	0.0910	1.5250	0.8338	316.3	18	12	4.6	323	2071	1995
42	P(*p*-$C_6H_4Cl)_3$	77.84	0.18	0.0993	0.0199	−0.1192	0.0793	0.1030	0.1590	1.1890	0.8309	365.6	18	12	4.6	358	2073	1996
43	$PPh(C_6F_5)_2$	−298.96	2.81	0.1823	0.1218	−0.3041	0.0605	0.6020	0.9080	0.9720	0.9127	442.2	18	5	2.1	355	—	—
44	P(*o*-$C_6H_4F)_3$	−14.73	0.46	0.1145	−0.0715	−0.0430	0.1860	0.2230	0.2520	0.2620	0.8370	316.3	18	12	1.9	328	—	—
45	P(*o*-$C_6H_4Cl)_3$	162.25	2.23	0.0751	−0.0286	−0.0465	0.1037	0.1440	0.1620	0.3300	0.8807	365.6	18	12	4.6	348	—	—

[a] Me, methyl; Et, ethyl; Pr, propyl; Bu, butyl; vin, vinyl; cy-hx; cyclohexyl; Bz, benzyl; Ph, phenyl; tol, tolyl; *o-*, *ortho* (e.g. *o-tol* : *ortho*-tolyl); *m-*, *meta*; *p-*, *para*; *n-*, normal (e.g. *n*-Pr: normal propyl); *i-*, *iso-* (e.g. *i*-Pr: isopropyl); *t*, *tert* (e.g. *t*-Bu: *tert*-butyl).

calibration data set which containing I objects. The reduced calibration data set is then composed of $(I-N)$ objects and the prediction data set of N quantities, denoted as a *cross-validation segment.* This is repeated until all samples have been omitted from the calibration set once. By using this technique, each of the objects in the data set is used once as a test objects and different models are made based on different calibration data sets. The squared prediction sums of the deleted objects will then give an estimate of the significance of a PLS dimension. In PCA the cross-validation is usually done slightly differently. Here individual observations are deleted in a pseudo-random way. A PCA based on the reduced set is then used to predict the deleted observations. Detailed accounts of one CV method are given in Ref. 15.

The principal properties. In the principal component analysis, all quantities and all of the molecular descriptor variables described above, except ν_{A1}, and ν_E were used. A model with five principal components was significant according to cross-validation and accounted for 90% of the total variance. A score projection of the two first principal components (Fig. 1) shows that the set of ligands can be divided into two separate homogeneous classes. One class accommodating the majority of the ligands and another, small class containing the polyhalogenated phosphorus ligands are, in many aspects, dissimilar to the other ones. For this reason the polyhalogenated compounds: $PPh(C_6F_5)_2$, $PPh_2C_6F_5$, $P(p\text{-}C_6H_4F)_3$, $P(o\text{-}C_6H_4F)_3$, $P(p\text{-}C_6H_4Cl)_3$, and $P(o\text{-}C_6H_4Cl)_3$ were excluded when the final the principal component model and the predictive models discussed below were determined. This class of compounds is not further investigated here, but one phosphine from this class should also be included in an experimental design in order also to explore the behaviour of polyhalogenated P-ligands.

A new principal components model was determined after the exclusion of the polyhalogenated compounds. A four-component model was significant according to cross-validation and accounted for 90.5% (41.0%+21.5%+18.0%+10.0%) of the total variance. Projections of score and loadings are shown in Figs. 2 and 3. The computed values of the scores and loadings are summarised in Table 2 and 3.

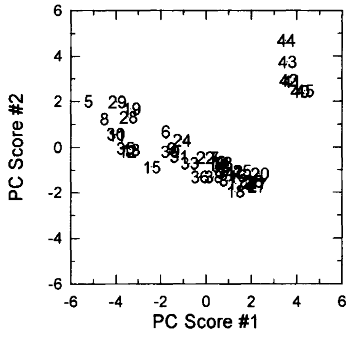

Fig. 1. Two way score plot for principal component (PC) #1 versus PC#2 of the whole data set. The principal component analysis 'splits' the phosphines into two different classes.

The score plots in Fig. 2 portray the principal properties of the compounds as projections of the original descriptors into the space spanned by the principal components. The loading plots (Fig. 3) show how the original molecular descriptors contribute to the principal components. These plots make it easy to discern how the descriptors are correlated. The first principal component is largely composed of the descriptors ε_{HOMO}, H_f, δP, M_W, n_C and vdWA which are correlated to each other, and the descriptors ε_{LUMO}, ε_{LUMO+1}, ε_{LUMO+2} and ε_{LUMO+3} which are correlated to each other and inversely correlated to descriptors ε_{HOMO}, δP, M_W, n_C and vdWA. The second principal component is composed of the following descriptors H_f, n_H and ε_{HOMO}, which are correlated, and descriptors μ, χ, δP and psA which are correlated to each other and inversely correlated to the descriptors H_f, n_H and ε_{HOMO}. In principal component #3, the molecular descriptors: H_f, χ, n_H, vdWA, M_W, psA and n_C are the most important. H_f and χ are correlated to each other and inversely correlated to the others. The numerical values for the principal component loading, $p_1, \ldots, p_4$, are given in Table 3.

Validity of the molecular descriptors as predictors for the IR frequencies. Tolman has shown that the nature of the phosphine ligand (L) influences the carbonyl stretching frequencies ν_{A1} and ν_E in $Ni(CO)_3L$ complexes. Thus, if the molecular descriptor variables, H_f , μ, χ, η, ε_{HOMO}, $\varepsilon_{LUMO}, \ldots$, account for metal–substrate interactions, this would be detected through a multivariate model such as a *partial least-squares regression* model. Two PLSR models (using PLS1), i.e., one response for each (ν_{A1} and ν_E) were derived for prediction purposes. These models relate the experimentally determined carbonyl IR stretching frequencies ν_{A1} in eqn. (2) and ν_E in eqn. (3) to the molecular descriptors,

$$\nu_{A1} = \alpha_0 + \alpha_1 \cdot H_f + \alpha_2 \cdot \mu + \alpha_3 \cdot \eta + \alpha_4 \cdot \chi + \cdots + \alpha_{15} \cdot \text{vdWA} \qquad (2)$$

$$\nu_E = \beta_0 + \beta_1 \cdot H_f + \beta_2 \cdot \mu + \beta_3 \cdot \eta + \beta_4 \cdot \chi + \cdots + \beta_{15} \cdot \text{vdWA} \qquad (3)$$

where $\alpha_1, \alpha_2, \ldots, \alpha_{15}$ are the regression coefficients in the predictive model for the IR carbonyl frequency ν_{A1}, and $\beta_1, \beta_2, \ldots, \beta_{15}$ are the regression coefficients for the IR carbonyl frequency ν_E, respectively. The numerical data for these equations are given in the last rows of Table 3.

The PLSR modelling is accomplished through a set of partial least-squares (PLS) components. A PLS-component represents systematic regression found in the data. Each P-ligand is represented in the PLS components by its scalar values called *PLS scores.* Each variable in the data matrices, X (which contain the calculated molecular

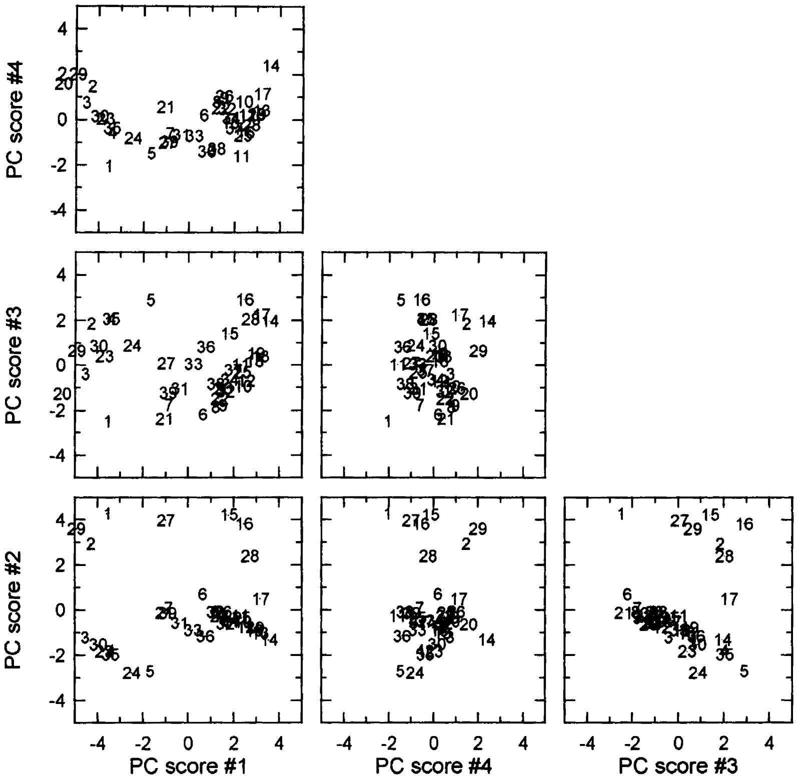

Fig. 2. Multivariate principal component score plot for a principal component analysis performed on the large class of phosphines (39). The principal component scores for the total principal component model are plotted: PC#1 versus PC#2, PC#1 versus PC#3, PC#1 versus PC#4, PC#2 versus PC#3, PC#2 versus PC#4, and PC#3 versus PC#4. The numbers in the plots are entries (objects) in Tables 1 and 2.

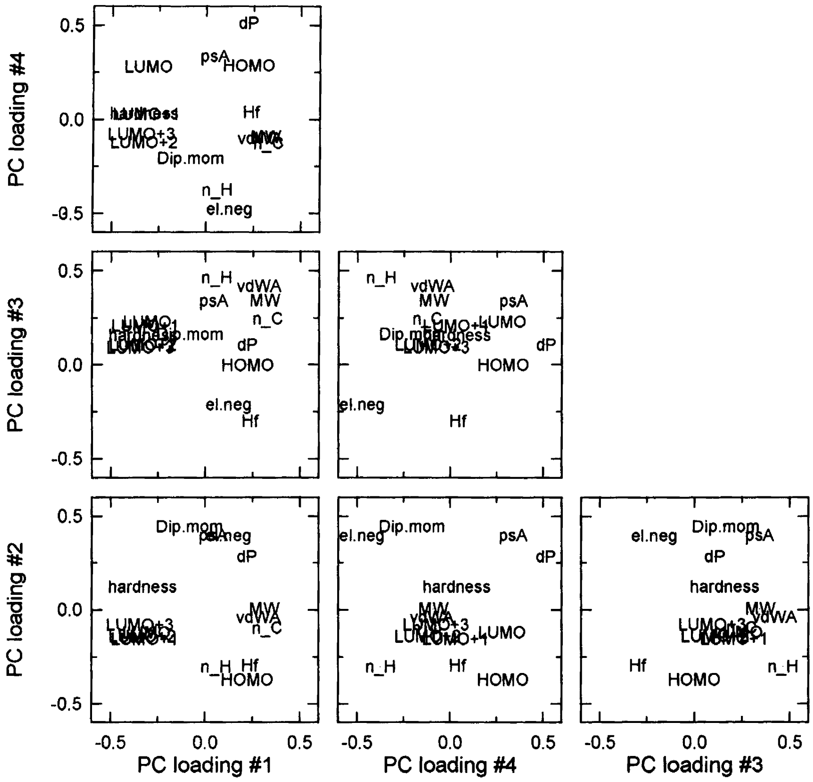

Fig. 3. Multivariate principal component loading plot for a principal component analysis performed on the larger class of phosphines (39). The principal component loadings for the total principal component model are plotted: PC#1 versus PC#2, PC#1 versus PC#3, PC#1 versus PC#4, PC#2 versus PC#3, PC#2 versus PC#4, and PC#3 versus PC#4.

Table 2.

		Estimated principal properties				Predicted CO frequencies	
No.	P-Ligand[a]	t_1	t_2	t_3	t_4	ν_{A1}	ν_E
1	PMe_2CF_3	−3.5701	4.2572	−2.4977	−2.0569	2084±3.0	2008±3.4
2	$P(OEt)_3$	−4.3058	2.9175	1.8380	1.4480	2076±2.2	1996±3.2
3	PEt_3	−4.5605	−1.2396	−0.4057	0.7471	2062±1.9	1977±2.3
4	PBu_3	−3.4094	−1.8188	2.0476	−0.5956	2059±1.5	1974±2.2
5	$P(cy\text{-}hx)_3$	−1.6884	−2.7323	2.8849	−1.4912	2056±2.3	1971±3.2
6	$PHPh_2$	0.6450	0.6766	−2.2022	0.1964	2071±1.9	1993±2.5
7	$PPhEt_2$	−0.8779	0.0684	−1.7713	−0.6244	2068±2.1	1988±2.5
8	PPh_2Me	1.1956	−0.0913	−1.8608	0.7672	2068±1.7	1989±2.3
9	PPh_2vin	1.5045	−0.3562	−1.7900	0.9578	2067±1.6	1987±2.3
10	PPh_3	2.3810	−0.5394	−0.9517	0.7850	2067±1.1	1987±1.6
11	PBz_3	2.2776	−0.2645	0.0191	−1.6211	2068±1.8	1989±2.4
12	PPh_2Bz	2.5094	−0.7767	−0.7151	0.2171	2066±0.9	1986±1.4
13	$P(p\text{-tol})_3$	3.1502	−1.0533	0.3734	0.3952	2065±1.1	1984±1.3
14	$P(o\text{-}C_6H_4OMe)_3$	3.5553	−1.3473	1.9604	2.3780	2062±2.4	1980±2.8
15	$P(OPh)_3$	1.7889	4.2146	1.3966	−0.1510	2084±1.9	2009±2.3
16	$P(O\text{-}p\text{-tol})_3$	2.4521	3.8068	2.8876	−0.5816	2082±2.9	2007±3.5
17	$P(p\text{-}C_6H_4OMe)_3$	3.1838	0.4721	2.2347	1.1248	2069±2.2	1990±2.5
18	$P(o\text{-tol})_3$	2.8846	−0.9392	0.1483	0.2450	2065±1.1	1985±1.4
19	$P(m\text{-tol})_3$	2.9431	−0.8024	0.5002	0.2245	2065±1.1	1985±1.3
20	PMe_3	−5.5761	−0.6602	−1.2883	1.5693	2064±2.8	1980±3.3
21	$PPhMe_2$	−1.0595	−0.1296	−2.3847	0.5448	2067±2.7	1987±3.2
22	PPh_2Et	1.3642	−0.2853	−1.5023	0.4952	2068±1.4	1988±1.9
23	$P(i\text{-Pr})_3$	−3.7163	−1.8501	0.3945	0.0164	2059±1.4	1975±1.7
24	$P(t\text{-Bu})_3$	−2.4986	−2.7934	0.8729	−0.8233	2056±1.7	1971±2.2
25	$PPhBz_2$	2.3404	−0.3537	−0.3422	−0.7108	2068±1.2	1988±1.6
26	$PPh_2(OEt)$	1.5194	−0.1042	−1.0489	1.0292	2067±1.7	1988±2.3
27	$PPh(OEt)_2$	−0.9970	3.9482	0.0661	−1.0183	2082±1.7	2006±2.3
28	$P(O\text{-}o\text{-tol})_3$	2.6962	2.3760	2.0278	−0.2451	2076±2.4	1999±2.9
29	$P(OMe)_3$	−4.9524	3.5912	0.6337	1.9783	2079±2.5	2000±3.4
30	$P(n\text{-Pr})_3$	−3.9803	−1.5517	0.8472	0.1480	2060±1.3	1976±1.7
31	$PPh(i\text{-Pr})_2$	−0.3795	−0.6038	−1.0667	−0.6911	2066±1.7	1985±2.1
32	$PPh_2(i\text{-Pr})$	1.6212	−0.6320	−1.1755	0.4603	2066±1.2	1986±1.7
33	$PPhBu_2$	0.2086	−0.9162	0.0362	−0.7320	2064±1.5	1983±1.9
34	PPh_2Bu	1.8024	−0.4922	−0.6625	0.0687	2067±1.0	1987±1.4
35	$P(i\text{-Bu})_3$	−3.4285	−1.9994	2.0450	−0.3797	2058±1.5	1974±2.2
36	$PPh(cy\text{-}hx)_2$	0.7503	−1.1842	0.7965	−1.4140	2063±1.7	1982±2.3
37	$PPh_2(cy\text{-}hx)$	1.9487	−0.5515	−0.2424	−0.3876	2066±1.0	1986±1.4
38	PBz_2Et	1.1907	−0.1020	−0.8542	−1.2876	2069±1.6	1989±2.2
39	$PBzEt_2$	−0.9128	−0.1577	−1.2484	−0.9851	2067±1.9	1987±2.2

[a] Ligand notation as given in the footnote to Table 1.

descriptors) and *Y* (which contain the IR carbonyl frequencies), is represented by *PLS loadings,* similar to the PC-loadings in the principal component analysis. The PLSR analysis also includes determination of some statistics and the optimal number of PLS components to be used in the model. Detailed accounts concerning multivariate calibration and validation are given in Refs. 16 and 17.

Experimental design. The two-way score plots, Fig. 2, were used for the selection of a small subset of 17 compounds which spans a range of properties (if the selected molecular descriptors do). This subset was used as a calibration data set for the PLSR modelling. The following 17 P ligands with available IR stretching frequencies were selected: PMe_2CF_3, $P(OEt)_3$, PEt_3, PBu_3, $P(cy\text{-}hx)_3$, $PHPh_2$, $PPhEt_2$, PPh_2Me, PPh_2vin, PPh_3, PBz_3, PPh_2Bz, $P(p\text{-tol})_3$, $P(o\text{-PhOMe})_3$, $P(OPh)_3$, $P(O\text{-}p\text{-tol})_3$, $P(p\text{-}C_6H_4OMe)_3$. The selection was based upon a uniform spread in the principal properties. The data of these ligands were used for PLSR modelling, to determine the α and β values in eqns. (2) and (3), respectively.

For ν_{A1} and ν_E two-component PLSR models were determined as significant according to cross-validation. The model for ν_{A1} was 89% (83.0%+6.0%) and the model for ν_E was 88.5% (81.0%+ 7.5%) of the total variance. In order to evaluate the predictive ability of these models, they were (*i*) used to predict the IR carbonyl frequencies of an independent test set for which IR data were available and (*ii*) evaluated by using the cross-validation method (see above). The independent test data set contained the following ligands: $P(o\text{-tol})_3$, $P(m\text{-tol})_3$, PMe_3 , $PPhMe_2$, PPh_2Et, $P(i\text{-Pr})_3$, $P(t\text{-Bu})_3$,

Table 3. Loadings for the principal component model and regression coefficients for PLSR models for the carbonyl frequencies ν_{A1} and ν_E.

a	p_{a1}	p_{a2}	p_{a3}	p_{a4}	p_{a5}	p_{a6}	p_{a7}	p_{a8}	p_{a9}	p_{a10}	p_{a11}	p_{a12}	p_{a13}	p_{a14}	p_{a15}
	H_f	μ	η	χ	ε_{HOMO}	ε_{LUMO}	ε_{LUMO+1}	ε_{LUMO+2}	ε_{LUMO+3}	δP	M_W	n_C	n_H	psA	vdWA
1	0.2330	−0.0843	−0.3334	0.1208	0.2225	−0.3100	−0.3209	−0.3327	−0.3467	0.2190	0.3096	0.3271	0.0569	0.0409	0.2787
2	−0.2968	0.4430	0.1220	0.3950	−0.3719	−0.1222	−0.1573	−0.1428	−0.0812	0.2825	0.0041	−0.0973	−0.3022	0.3969	−0.0407
3	−0.3003	0.1636	0.1569	−0.2127	−0.0016	0.2295	0.2083	0.1055	0.0924	0.1049	0.3437	0.2472	0.4650	0.3406	0.4184
4	0.0391	−0.2057	0.0331	−0.4749	0.2842	0.2791	0.0287	−0.1215	−0.0767	0.5123	−0.0917	−0.1255	−0.3740	0.3360	−0.0999
ν_{A1}															
α_0	α_1	α_2	α_3	α_4	α_5	α_6	α_7	α_8	α_9	α_{10}	α_{11}	α_{12}	α_{13}	α_{14}	α_{15}
2018	−0.0098	2.6637	1.4151	5.1179	−2.7213	−1.1586	−0.8005	−1.0105	−0.3928	4.2790	−0.0005	−0.0586	−0.2020	0.1284	−0.0032
ν_E															
β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}	β_{11}	β_{12}	β_{13}	β_{14}	β_{15}
1924	−0.0115	3.4531	1.4371	7.0603	−3.4515	−1.9296	−1.2261	−1.4555	−0.6124	5.4432	0.0006	−0.0574	−0.2671	0.1655	−0.0036

Table 4. Statistical parameters – evaluating the PLSR models predictive ability.

Statistical parameters[a]	Prediction using independent test data set	Prediction using cross-validation segments
For ν_{A1}		
RMSEP	3.60	3.09
k	0.87	0.87
R^2	0.89	0.93
For ν_E		
RMSEP	5.09	4.18
k	0.83	0.88
R^2	0.88	0.93

[a] RMSEP: root mean square of error of prediction; k, slope of correlation line ($y_{predicted}$ versus $y_{measured}$); R^2, multiple correlation coefficient.

$PPhBz_2$ $PPh_2(OEt)$, $PPh(OEt)_2$, $P(O\text{-}o\text{-}tol)_3$, and $P(OMe)_3$. The statistics of these predictions are summarised in Table 4. The poorest fit was observed for the ligands $PPh(OEt)_2$ and $P(O\text{-}o\text{-}tol)_3$. These P ligands are classified by the PCA model as 'extreme' ligands, i.e., that they have extreme values of the principal properties and thus are found in the outer region of this class of ligands.

It is interesting to note that the derived PLSR models for the IR carbonyl frequencies can be used to predict frequencies for P-ligands for which such measurements not have been carried out. We have used the PLSR models to determine the IR carbonyl frequencies of all ligands which belong to the largest class (all except ligands Nos. 40–45 in Table 1). These results are summarised in Table 2. By studying the sign and numerical value of α and β for the IR carbonyl frequencies ν_{A1} and ν_E, respectively, it can be seen that the molecular parameters μ, χ, δP, and psA contribute to an increase in the IR frequency, whereas H_f, ε_{HOMO}, ε_{LUMO}, ε_{LUMO+1}, ε_{LUMO+2} and n_H contribute to a lowering of the frequencies.

Discussion and conclusion

When many molecular properties are calculated and/or measured for a series of compounds, it is often found that some of the property descriptors are correlated to each other, which means that they depend on the same underlying intrinsic property. Independent properties will be described by different principal component vectors. This is a consequence of the fact that principal component vectors are mutually orthogonal. The PCA model derived for the phosphorus ligands showed that all of the molecular descriptors contained some information used to obtain the principal properties. Since these principal properties could be used as predictors for describing the influence on carbonyl frequencies in

$Ni(CO)_3L$ complexes, one may conclude that the original molecular descriptors orthogonalised by projections upon principal components may well be used as *real* principal properties for the class(es) of phosphorus ligands. The result of this PLSR modelling was considered satisfactory, and our conclusion is that the descriptors obtained through semiempirical quantum chemistry methods can be used satisfactorily in multivariate modelling and deriving of the principal properties.

An important conclusion follows from the PLSR modelling: the position of the carbonyl frequencies reflects the energy of the C–O bond. Coordination of a phosphine ligand to the central Ni atom perturbs the Ni–CO interaction (back donation) which is observed as a shift in the carbonyl absorption frequencies. The PLSR model shows that the properties of the P-ligand *determine* the properties of the entire $Ni(CO)_3L$ complex, i.e., the principal properties of the P-ligand *can explain* the variation of carbonyl shift. The implication of this result can be far-reaching since the PCA score plots allows for experimental design for the systematic exploration of phosphorus ligand transition metal complexes for use in new synthetic procedures. Thorough accounts and examples of the use of principal properties for experimental design are given in Refs. 18–23.

From the interpretation of the PCA model and the results from the PLSR model, a chemical interpretation of the model describing the P-ligand–metal interaction can be deduced: the dipole moment gives information about the electronic distribution over the whole molecule and the partial charge on the phosphorus gives information about the electron density on the atom that coordinates directly to the metal atom of the catalyst. Both these parameters give information about the electron-donating properties of the P-ligands. The energies of the unoccupied molecular orbital tells how *good* the P-ligand is at accepting back-donation of electrons from the metal atom. The number of hydrogen atoms contributes to the size of the compound (a high value of n_H indicates a small molecule). Steric effects alter the bond angles and thus the hybridisation of the phosphorus atom. Increasing the bond angle between substituents will decrease the percentage *s* character in the phosphorus lone pair, and hence steric effects have important electronic consequences.

Experimental

Molecular calculations. Energy-optimised structures obtained by molecular mechanics, the MM2 routine implemented in the PCMODEL version 4.0 software, were subjected to the AM1 (Austin Model 1) routine[11,12] implemented in the MOPAC software version 6.0 (A General Molecular Orbital Package). The AM1 routine was used to calculated the heat of formation, dipole moment, energies of the occupied and unoccupied molecular orbitals, and partial charge on phosphorus.

The final energy-optimised structures from the AM1 computations were used for the calculations of the van der Waals surfaces ($Å^2$) and the polar surface areas ($Å^2$) using PCMODEL version 4.0 software. The calculations were done on an Arche 486DX, 50 MHz microcomputer under DOS 6.0.

Multivariate computation. Prior to computing the principal component models (for both the PLSR and the PCA methods), each descriptor variable was scaled to unit variance. This was done to avoid the situation where different units of measurement of the descriptors distort the variance structure and thereby bias the projections. UNSCRAMBLER version 5.0 or The Unscrambler® version 5.5 software[24] under DOS 5.0 on a COMPAQ 486/50DX microcomputer or on an IBM ThinkPad 755Cs under DOS 6.3, was used for the multivariate data analysis on the scaled data. To avoid overfitting in deriving the principal component models, the cross-validation method[15] with the maximum number of segments, was used to determine how many principal components were significant.

Acknowledgements. The Swedish Natural Science Research Council (R.C., U.M.H) is gratefully acknowledged for financial support .

References

1. Collman, J. P., Hegedus, L. S., Norton, J. R. and Finke, R. G. *Principles and Applications of Organotransition Metal Chemistry*, University Science Books, Mill Valley CA 1987, p. 1.
2. Hellberg, S., Sjöström, M., Skagerberg, B. and Wold, S. *J. Med. Chem. 30* (1987), 1126.
3. Hellberg, S., Sjöström, M. and Wold, S. *Acta Chem. Scand. Ser. B 40* (1986) 135.
4. Grim, S. O. and McFarlane, W. *Nature* (1965) 995.
5. Moedritzer, K., Maier, L. and Groenweghe, L. C. D. *J. Chem. Eng. Ref. Data, 7* (1962).
6. Bemi, L., Clark, H. C., Davies, J. A., Fyfe, C. A. and Wasylishen, R. E. *J. Am. Chem. Soc. 104* (1982) 438.
7. Grim, S. O., McFarlane, W. and Davidoff, E. F. *J. Am. Chem. Soc. 32* (1967) 781.
8. Tolman, C. A. *Chem. Rev. 77* (1977) 313.
9. Tolman, C. A. *J. Am. Chem. Soc. 92* (1970) 2953.
10. Tolman, C. A. *J. Am. Chem. Soc. 92* (1970) 2956.
11. Stewart, J.J.P. MOPAC: *A General Molecular Orbital Package*, Frank J. Seiler Research Laboratory, US Air Force Academy, Colorado Spring, CO 80840, USA.
12. Stewart, J. J. P. *J. Computer-Aided Mol. Des. 4* (1990) 1.
13. Pearson, R. G. *J. Org. Chem. 54* (1989) 1423.
14. Jolliffe, I. T. *Principal Component Analysis*, Springer-Verlag, New York 1986, pp 1–114.
15. Wold, S. *Technometrics 20* (1978) 397.
16. Martens, H. and Næs, T. *Multivariate Calibration*, Wiley, New York 1991, pp. 116–165.
17. Jackson, J. E. *A User's Guide to Principal Components*, Wiley, New York 1991, pp. 282–290.
18. Carlson, R. *Design and Optimization in Organic Synthesis*, Vol. 8 in *Data Handling in Science and Technology*, Elsevier, Amsterdam, 1992, pp. 337–387.
19. Carlson, R., Lundsted, T. and Albano, C. *Acta Chem. Scand., Ser. B 39* (1985) 79.

20. Carlson, R., Prochazka, M. P. and Lundsted, T. *Acta Chem. Scand., Ser. B 42* (1988) 145.
21. Carlson, R., Prochazka, M. P. and Lundsted, T. *Acta Chem. Scand., Ser. B 42* (1988) 157.
22. Carlson, R., Lundsted, T., Nordahl, Å. and Prochazka, M. P. *Acta Chem. Scand., Ser. B 40* (1986) 522.
23. Bjørsvik, H.-R. and Priebe, H. *Acta Chem. Scand. 49* (1995) 446.
24. UNSCRAMBLER, version 5.0. Extended memory version. User's Guide 1993 and The Unscrambler® version 5.5. User's Guide 1994. Camo AS, Trondheim, Norway.