DOI: 10.1002/adsc.200505299

Design and Assembly of Virtual Homogeneous Catalyst Libraries – Towards *in silico* Catalyst Optimisation

Jos A. Hageman,[a] Johan A. Westerhuis,[a] Hans-Werner Frühauf,[b]
Gadi Rothenberg[b,*]

[a] Biosystems Data Analysis, University of Amsterdam, Nieuwe Achtergracht 166, 1018 WV Amsterdam, The Netherlands
[b] van 't Hoff Institute of Molecular Sciences, University of Amsterdam, Nieuwe Achtergracht 166, 1018 WV Amsterdam, The Netherlands
E-mail: gadi@science.uva.nl

Received: July 27, 2005; Accepted: November 28, 2005

Abstract: An alternative top-down concept for searching for homogeneous catalysts is introduced. In this approach, three multi-dimensional spaces are considered. These represent the catalysts, the descriptor values (e.g., cone angle, lipophilicity indices), and the figures of merit (e.g., turnover frequency, enantiomeric excess, or product selectivity), respectively. By generating and connecting these spaces, it is possible to screen virtual catalyst libraries and indicate regions in the catalyst space where good catalysts are likely to be found. The generation of the catalyst space from simple building blocks is presented and the application of this approach is demonstrated for two cases: predicting the bite angle in a library of 600 ligand-Rh complexes, and predicting the linear : branched aldehyde product ratio that these 600 catalysts would give in the hydroformylation of 1-octene. In the latter case, the model is first trained on a set of 39 octene hydroformylation reactions. The limitations and applications of this concept are discussed.

Keywords: catalyst discovery; hydroformylation; ligand design; molecular descriptors; QSAR; virtual library

Introduction

Designing new catalysts and exploring novel catalytic reactions is one of the key challenges of homogeneous catalysis in the 21st century.[1,2] The last two decades have seen an enormous development in discovery and optimisation tools, especially in the area of high-throughput experimentation (HTE) and process optimisation.[3] However, the concepts for exploring the catalyst space in homogeneous catalysis have changed little over the past fifty years. Basically, once an active catalyst complex is discovered, small modifications are made on the structure to screen the activity of neighbouring complexes, covering the space like an ink drop spreads on a sheet of paper. This is by no means a bad method, and until recently this was also the only possible approach.[4]

With the commercialisation of HTE systems and in particular synthesis robots in the 1990s, a different catalyst screening approach was made possible, with numerous reactions carried out in parallel. However, although many thousands of catalysts were tested, very few good catalysts were discovered.[5,6] The reason for this is that in most of the published reports the robots performed a systematic 'grid search' of the catalyst space. As shown by Chen and Deem, this approach is unsuitable, due to the sheer scale of the problem, plus the fact that many of the structure-activity relationships in catalysis are non-linear.[7] New search strategies are therefore needed to complement the new robotic systems.[8,9]

In this paper, we introduce an alternative concept for optimising homogeneous catalysts. Using a 'virtual synthesis' platform, we assemble large catalyst libraries (10^6–10^9 candidates) *in silico*. We then extract subsets from these libraries and predict their catalytic performance using statistical models, molecular descriptors, and QSAR models. First, we present the theoretical basis of this approach. Then, we construct a dataset of 600 bidentate ligand-Rh complexes and test the model's ability to predict the bite angle values for these catalysts. Finally, we generate a quantitative structure-activity relationship (QSAR) model for a set of 39 Rh-catalysed hydroformylation reactions using the *n*:*iso* product ratio as the figure of merit, project the 600 structures on this model and analyse which descriptors are important and which catalysts are promising.

Theory

Defining the Catalyst Space

As an example, assume that we are looking for a homogeneous catalyst comprised of a transition metal atom bound to a bidentate organic ligand. Now let us consider three multi-dimensional spaces, **A**, **B**, and **C** (Figure 1). Space **A** is a grid containing all the catalyst structures (i.e., all of the combinations of transition metal atoms and bidentate ligands, where each point in space **A** pertains to a different catalyst); space **B** contains the values of the catalyst and the reaction descriptors (backbone flexibility, partial charge on the metal atom, lipophilicity and temperature, pressure, solvent type, and so on); and space **C** contains the catalysts' figures of merit (i.e., the TON, TOF, ee, price and so forth). By dividing our problem in this way, we translate it from an abstract problem in catalysis to a (still abstract) problem of relating one multi-dimensional space to another. The advantage is that the relationship between spaces **B** and **C** can be quantified using QSAR and quantitative structure-property relationship (QSPR) models.[10,11] Note that space **B** contains molecular descriptor values, rather than structures, but these are related directly to the structures, as we showed recently for the cases of monodentate complexes in Pd-catalysed Heck reactions,[12,13] and bidentate complexes in Ni-catalysed hydrocyanation.[14]

Therefore, if one can generate a sufficiently large number of sufficiently diverse structures in space **A**, and link the spaces **A** and **B**, one should be able to predict the relevant figures of merit in space **C** using QSAR/QSPR descriptor models. In the following subsection, we will present a building block assembly approach for generating space **A**, and show that it is relatively easy to obtain a large number of structures (the question of structure diversity is much more complex, and a full treatment of it is out of the scope of this paper[15]).

Assembling the Virtual Library (Space A)

To follow the above example, let us assume that each catalyst contains one metal atom M and one bidentate ligand, which includes two ligating groups L_1 and L_2, a backbone group B, and three residue groups R_1, R_2, and R_3. To make things simple, we will limit the R groups to one per ligating or backbone group. There is no restriction on similarity between the groups, i.e., it is possible that $L_1 = L_2$, and so forth. Each ligand has a unique identifier, in the form $L_1(R_1)$-$B(R_2)$-$L_2(R_3)$. Figure 2 shows a schematic representation of the catalyst parts and assembly possibilities. The connection points for the R groups and between the L and B groups are predefined for each building block (for example, the pyridine ligating group **6** can connect to a backbone on positions 2, 3, or 4, and an R group can subsequently be attached to it on one of the remaining 'free' positions).

Here one already faces a barrier: The number of possible structures, even using just twenty ligating groups, ten bridging groups, ten residue groups, and this very simplified catalyst representation, is over 1.7 billion!

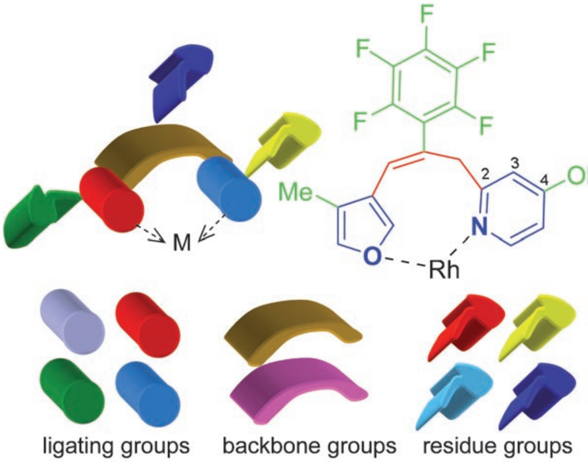


Figure 2. Cartoon showing the construction of a bidentate ligand (*top left*) and an example structure (*top right*) from building blocks comprised of ligating, backbone, and residue groups (*bottom*).

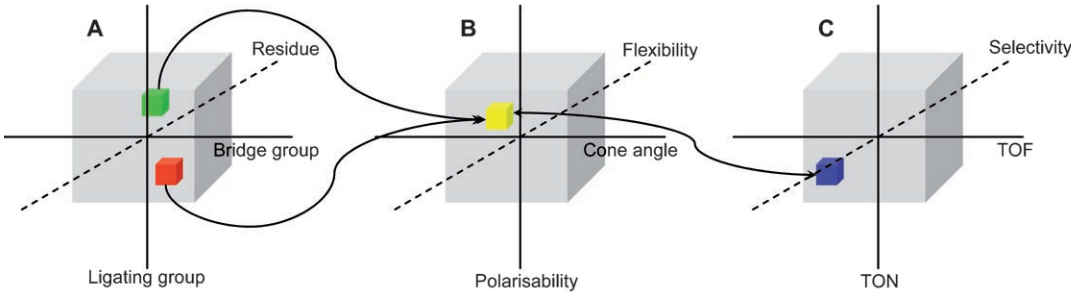


Figure 1. Simplified projection of the three multi-dimensional spaces representing the catalysts (space **A**, *left*), the descriptor values (space **B**, *middle*) and the figures of merit (space **C**, *right*). Reaction conditions, e.g., temperature, pH, or solvent type can also in principle be incorporated in space **B**.

Computing any parameters for such a large library is not realistic, and even storing it is problematic (if each structure requires just 300 bytes, space **A** would take up 510 gigabytes). Therefore, we will keep space **A** as a well defined but virtual space, that exists in thought only. From this space, we will sample subspaces small enough to be analysed (typically 500–1000 catalyst structures). These subspaces, or virtual libraries, are assembled by generating sets of different $L_1(R_1)$-$B(R_2)$-$L_2(R_3)$ identifiers and subsequently optimising the corresponding structures. The entire process is fully automated and a record is kept of the sampled structures (see the Computational Methods Section for details).

Choosing the subsets for the virtual libraries is not trivial. For the demonstration and validation purposes in this paper, we used random subsets. In practice, one may insert into the selection process also prior knowledge about the reaction at hand, and apply an iterative approach using global optimisation methods (*vide infra*).[16]

Results and Discussion

Predicting Bite Angles of Bidentate Ligand-Rh Complexes

As a first example, we will demonstrate the possibilities of the above approach in categorising and predicting the properties of virtual catalyst libraries.[17] We will use the ligand bite angle as a representative property (this descriptor was shown by van Leeuwen and co-workers to be a key parameter in determining the activity of bidentate ligands[18,19]). First, we 'synthesised' a library of Rh chelate complexes as a specific test case, using as building blocks 20 ligating groups, 10 backbone groups, and 10 residue groups (Figure 3).[1,20] These building blocks encompass a variety of functions (among them the degree of lipophilicity, polarity, steric hindrance and the ability to form π-bonds and hydrogen bonds).

The program then selects at random two L groups, three R groups and one B group, and generates a L_1 (R_1)-$B(R_2)$-$L_2(R_3)$ string that is translated into a structure. In cases where a group has several connection points, the program selects one of these points at random. To avoid 'bond entanglement' (e.g., the creation of six-membered catenanes), we connect the blocks us-

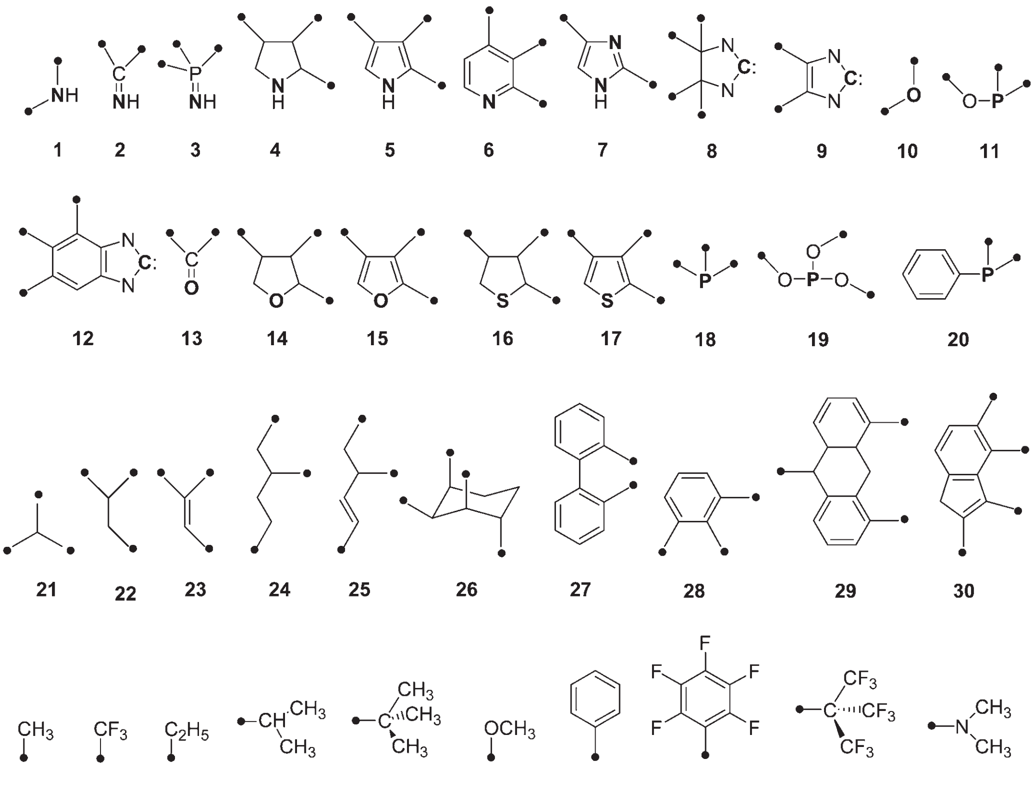


Figure 3. Building blocks used in the assembly of the virtual library. Ligating groups (structures **1**–**20**, boldface type indicates the ligating atoms), bridging groups (**21**–**30**) and residue groups (**31**–**40**). The • symbols denote the connection points. After assembling the ligand structure, the program automatically assigns H atoms to any unused connection points.

ing very long (20 Å) bonds, and subsequently minimise the structures using molecular mechanics (MM) force fields. This simplified approach gives good results and is computationally inexpensive. Out of the 1.7×10^9 possible Rh ligand complexes (space **A**) we chose 600 structures at random. We then calculated for each of these 600 structures 138 QSAR parameters, resulting in a 138×600 matrix of descriptor values (space **B**). This matrix was then divided in two subsets: A training set of 500 structures and a validation set of 100 structures.

To create a figure of merit (space **C)** for this simulated dataset, we calculated the actual bite angles of all 600 structures (note that the bite angle is in fact a descriptor, used here for demonstration purposes as a figure of merit[21]). The relation between spaces **B** and **C** was modelled using principal least squares (PLS) in combination with the training set. The number of latent variables was determined using leave-one-out cross-validation. To test the model's predictive capabilities towards new structures, we applied the validation set to the model. Figure 4 shows the predicted *vs.* the calculated bite angles of the validation set. The correlation coefficient of the predicted and calculated values is $R^2 = 0.73$. This correlation can be improved by using more sophisticated models, but it can already give us a rough indication of the desired (and undesired) regions in the catalyst space.

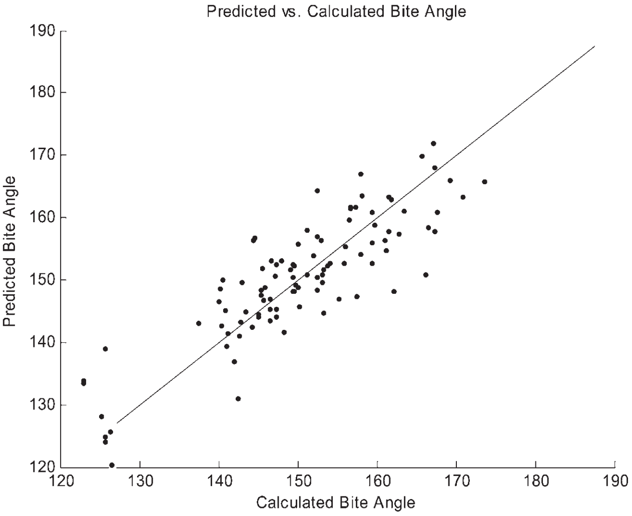
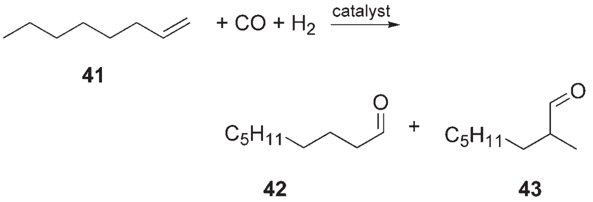


Figure 4. Predicted *vs.* calculated bite angles.

It is important to emphasise that the existence of structure-activity relationships in homogeneous catalysis means that space **A** is not simply a random grid of structures. So-called 'good catalysts' should share some important structural attributes. Were this not the case, then any search for a good homogeneous catalyst would be a hopeless task. In the following section, we show Rh-catalysed hydroformylation as an example. We chose this example expressly because this reaction shows a clear dependency on the ligand descriptors. In other cases, the relationship exists but it is not clear, and descriptor selection is indeed the primary task.[14,22]

Application to Rh-Catalysed Hydroformylation

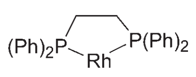
The chain extension of alkenes by CO and hydrogen insertion, commonly known as hydroformylation, is a key synthetic protocol in the chemical industry. Furthermore, it is one of the few examples where homogeneous catalysis is practiced on a large industrial scale. The Rh-catalysed variation has been extensively studied under a variety of conditions,[23,24] and many results point to strong ligand effects on activity and product selectivity.[25] For these reasons, we chose this reaction as a case study. We examined the possibilities of using the above approach to predict the product selectivity of bidentate Rh-complexes in the hydroformylation of 1-octene **41** [Eq. (1)]. For this, we collected 39 Rh-catalysed hydroformylation reactions from the literature,[26–32] using the *n:iso* product ratio as the figure of merit. Table 1 shows a partial representation of this dataset, together with catalyst structures **44–53**. Since the number of structures is relatively small, the division into training and test set requires special care. We based this division on Euclidian distances, with 30 samples forming the training set, and the remaining nine samples forming a representative test set. The relation between spaces **B** and **C** was then modelled with PLS, determining the number of latent variables with leave-one-out cross-validation on the training set. Figure 5 shows the predicted *vs.* calculated values for the test set. The correlation coefficient of the predicted and calculated values is $R^2 = 0.84$. This is not an excellent correlation, and this is also one of the limitations of this approach. However, this correlation is sufficient for avoiding 'inactive regions' of the catalyst space.



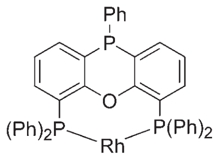
(1)

Subsequently, we used this PLS model to screen the above library of 600 Rh-ligand complexes. Figure 6 shows a histogram of the resulting 600 *n:iso* ratios, divided in 30 groups (30 bins). The *n:iso* ratios show a median at ~20. Four of the structures have predicted *n:iso* ratios of >85. As Figure 7 shows, all four complexes contain two P-ligating groups. This is not so surprising, as the training set contains only P-ligating groups. Note, however, that catalysts **54** and **55** are more bulky, while **56** and **57** contain more electron-withdrawing groups. Catalyst **57** is especially interesting, as it seems

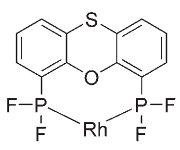
Table 1. Partial representation of the Rh-catalysed 1-octene hydroformylation dataset.



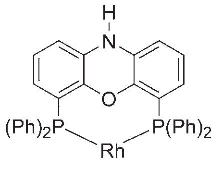
44



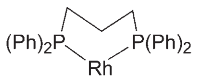
45



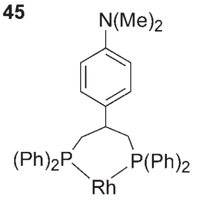
46



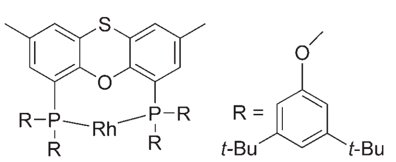
47



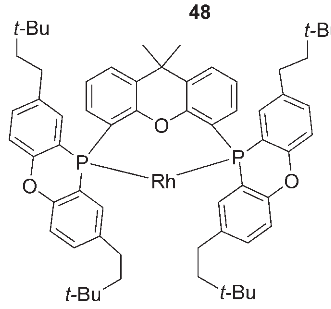
48



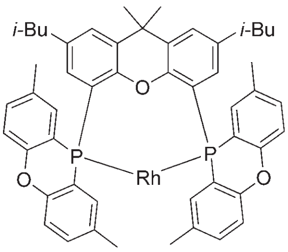
49



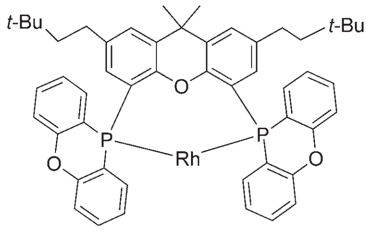
50



51



52



53

Entry	Reaction conditions[a]							Figure of merit
	Catalyst	T [°C]	Time [h]	Catalyst precursor	Ligand : Rh ratio	Octene : Rh ratio	$P_H = P_{CO}$ [bar]	*n* : *iso* product ratio
1	**44**	34	18	Rh(acac)(CO)$_2$	1.00	645	3	2.10
2	**45**	60	18	Rh(acac)(CO)$_2$	5.00	637	10	14.60
3	**46**	60	18	Rh(acac)(CO)$_2$	5.00	637	10	51.50
4	**47**	60	18	Rh(acac)(CO)$_2$	5.00	637	10	69.40
5	**48**	85	18	RhH	1.00	150	0.5	1.04
6	**49**	60	18	Rh(cod)BF$_4$	1.20	400	50	1.87
7	**50**	80	1.5	Rh(acac)(CO)$_2$	6.00	2500	10	7.70
8	**51**	80	n/a	Rh(acac)(CO)$_2$	5.00	637	10	11.50
9	**52**	80	n/a	Rh(acac)(CO)$_2$	5.00	637	10	32.33
10	**53**	80	n/a	Rh(acac)(CO)$_2$	5.00	637	10	99.00

[a] Benzene is the solvent for entry 1; toluene for entries 2–10.

very different from the usual 'bulky ligand' approach. This shows that in principle it is possible to find ligands with very different structures that exhibit similar behaviour (*vide infra*). For our demonstration purposes, however, the important part is not so much the performance of individual structures, but the fact that the model can identify trends in the catalysts' performance.

Variable Importance (VIP) Studies

Not all QSAR parameters are equally important for the model's performance. To rank the parameters, we conducted a VIP analysis. VIP helps us to discard unimportant parameters, and make the model easier to understand.[33,34] Figure 8 shows the key variables according to the VIP analyses for determining the complex bite angles (*top*) and the *n*:*iso* ratios (*bottom*). In the case of the bite angles, the Randić index,[35] the molecular surface area, the Wiener index[36] and the Kier shape index[37] are the most influential parameters. This is not so surprising, as these are all indeed topological parameters. Conversely, in the case of the *n*:*iso* ratios, the two most significant parameters are electrostatic: Zefirov's empirical atomic partial charges for the O and H atom, respectively.[38,39] This may reflect the participation of specific oxygen atoms in the competition between two chemical pathways to form the *n* and the *iso* aldehydes, respectively. Although in this case the experimental dataset of 39 reactions is too small for drawing

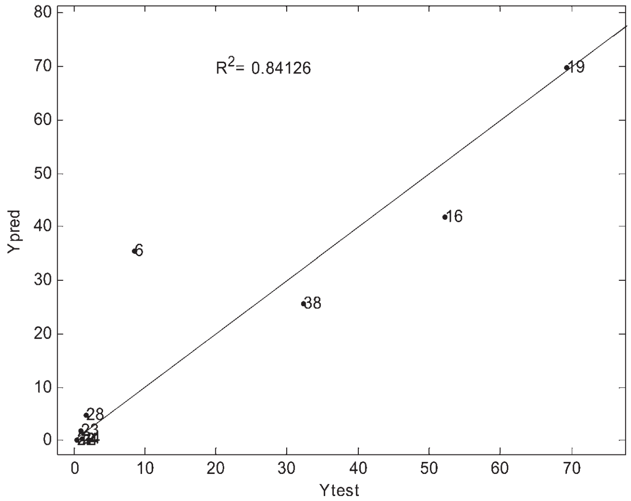


Figure 5. Predicted *vs.* calculated values of *n:iso* product ratios for 1-octene hydroformylation.

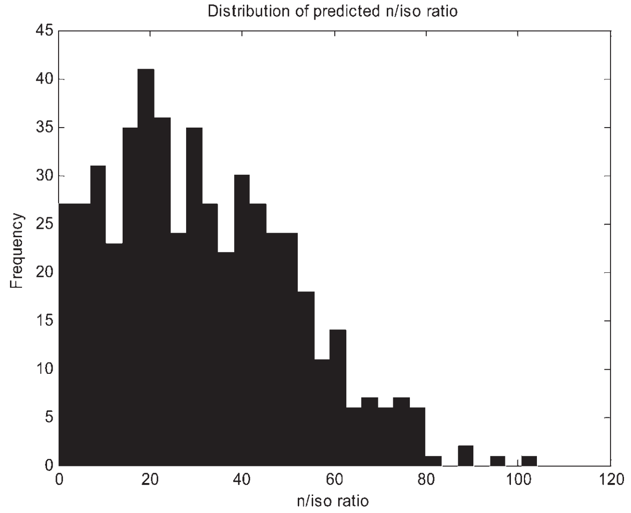


Figure 6. Histogram of the resulting 600 *n:iso* product ratios for 1-octene hydroformylation.

specific mechanistic conclusions, it can give us some insight with regard to the parameters involved in the reaction mechanism. With larger datasets, more accurate mechanistic information can be inferred.[13]

Limitations and Future Prospects

If a direct relationship is found between the catalysts (space **A**), the molecular descriptors (space **B**), and the figures of merit (space **C**), then in principle it should be possible to 'backtrack' from space **C** to **A**. Thus, if one knows the structure of a 'good catalyst' (i.e., knows the corresponding positions in spaces **B** and **C**), it would be possible to select structures in space **A** that correspond to these points. It may well be, however, that

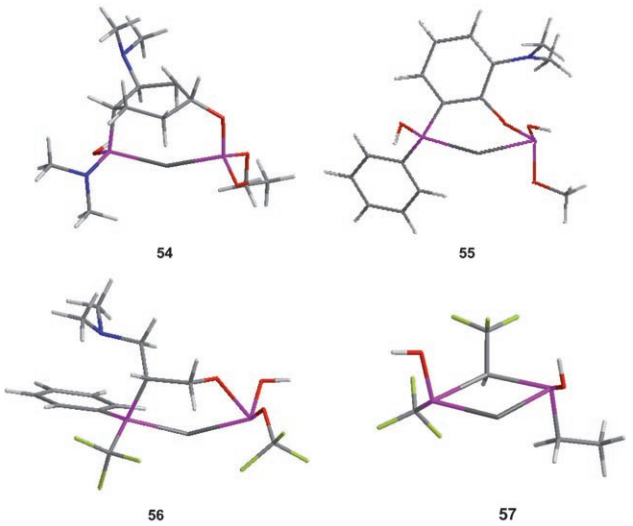


Figure 7. Four catalyst structures generated by the model, with predicted *n:iso* ratios >85 (P atoms shown in magenta and F atoms in yellow).

the corresponding catalyst structures would be different, aqua bonding and functional groups, from the original structure. Since patents in catalysis relate mainly to structures, and not to molecular descriptors or figures of merit (it would be difficult to claim all "good catalysts"), this may lead to changes in intellectual property definitions, as far as homogeneous catalysis is concerned. Backtracking from space **B** to **A** is not trivial, however, and further work is needed to realise the potential of this concept. We plan using global optimisation methods such as genetic algorithms,[40] simulated annealing and tabu search[41] methods to connect the desired regions in space **C** to the corresponding regions in spaces **A** and **B**. Further work is in progress to realise the potential of this concept.

Conclusions

Assembling large homogeneous catalyst libraries *in silico* from a small number of simple building blocks is possible. Molecular descriptors may then be combined with statistical tools for extracting meaningful structure/activity and structure/property trends from these libraries. We believe that coupling *in silico* screening to high-throughput experimentation workflows may enable the study of large catalyst spaces in the future. We emphasise that this approach does not negate or replace 'chemical intuition' or mechanistic research. Rather, it is a tool that can help catalysis chemists pinpoint good regions in the catalyst space. Further work, namely implementing combined computational/experimental workflows, examining the issues of catalyst diversity and optimal subset selection, and, most importantly,

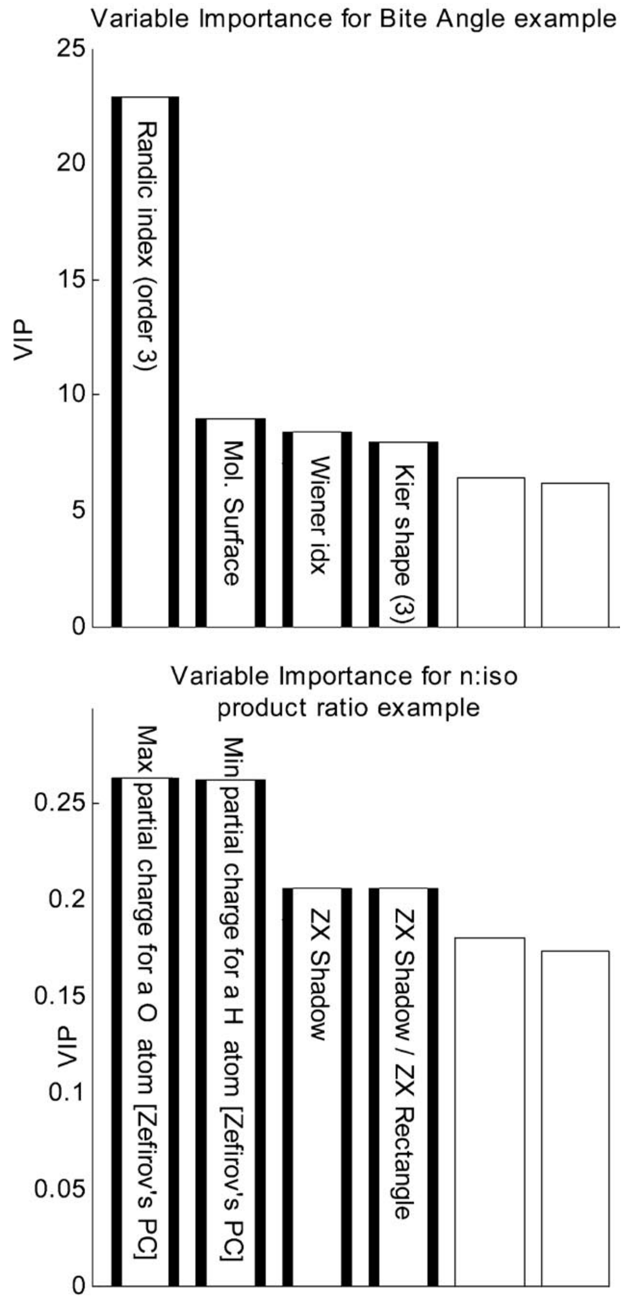


Figure 8. Variable importance (VIP) analysis of the QSAR parameters calculated for 500 ligand-Rh complex bite angles; parameters with VIP > 8 are considered as influential on the model (*top*). QSAR parameters ranked according to VIP analysis for the *n:iso* product ratio example (*bottom*).

synthesising the new catalysts, will be the subject of future research in our laboratory.

Computational Methods

String generation, assembly routines and matrix analysis were coded in MATLAB,[42] generating the unoptimised catalyst structures as Brookhaven protein database (PDB) files. Geometry optimisation was done in Hyperchem,[43] using the MM+ force field in combination with a conjugate gradient optimisation method (Polak-Ribiere). The optimised structures were then analysed using the Codessa software package. The entire process was automated using a combination of batch and MATLAB code. All computations were performed on a single-processor 2.0 GHz laptop computer. To give an idea of the computational costs, the automated selection, geometry optimisation, and analysis of the 600 Rh complex structures took 8 h.

Assembling the Virtual Library

Since the number of possible catalysts in space **A** is gigantic and cannot be stored in a 'normal' manner, this space is defined in thought only. The program chooses ligating groups L, bridging groups B, and residue groups R, and generates sets of $L_1(R_1)$-$B(R_2)$-$L_2(R_3)$ strings (typically 200–5000 per library) that are then assembled and optimised. Each structure has a unique combination of building blocks and connection points.

Example: Assembling a catalyst from Rh and blocks **6** and **14** (ligating groups), **22** (bridging group) and **33**, **34** and **34** (residue groups). The algorithm chooses at random the connection points for each building block, depending on the molecule (for example, for pyridine the possibilities are 2, 3, and 4). Two connection points are selected for each ligating group, to connect the bridge and the residue group. Three points are selected for the bridging group, to connect the two ligating groups and one residue group. The program connects the building blocks with very long (~20 Å) bonds to avoid bond entanglement (see Figure 9, *left*). The geometries of these structures are subsequently optimised (Figure 9, *right*). QSAR parameters are then calculated for each optimised structure, and parameters with null or near-null values are deleted from the dataset.

Regression Analysis

With regression one aims to find the relationship $\mathbf{y}=\mathbf{Xb}+\mathbf{e}$ between the regressors (**X**) (in this case space **B**, the QSAR parameters) and the dependent variable (**y**) (in this case space **C**, the bite angles and the *n:iso* product ratios). The vector **b** contains the regression coefficients and **e** the residuals. Here we used partial least squares (PLS) regression. In PLS, the data matrix **X** is replaced by factors (latent variables) extracted from the $\mathbf{X'yy'X}$ matrix. These factors are predictive of **y**, and also utilize **X** efficiently. The correct number of latent variables is unknown *a priori* and must be determined by (cross)-validating the model. Too many latent variables will lead to overfitting, while too few latent variables will result in poor model performance, because not all characteristics of the data are taken into account.

The complete dataset of 600 catalysts was divided into a training set of (500 structures) and a test set of (100 structures). All structures were randomly divided over both sets. The training set was used for constructing and validating the model. The test set was used for assessing the final predictive abilities of the model (this testing is meaningful only with structures which have not been used for constructing the model). The number of latent variables is determined by leave-one-out cross-validation. In this procedure, for a limited number of latent variables

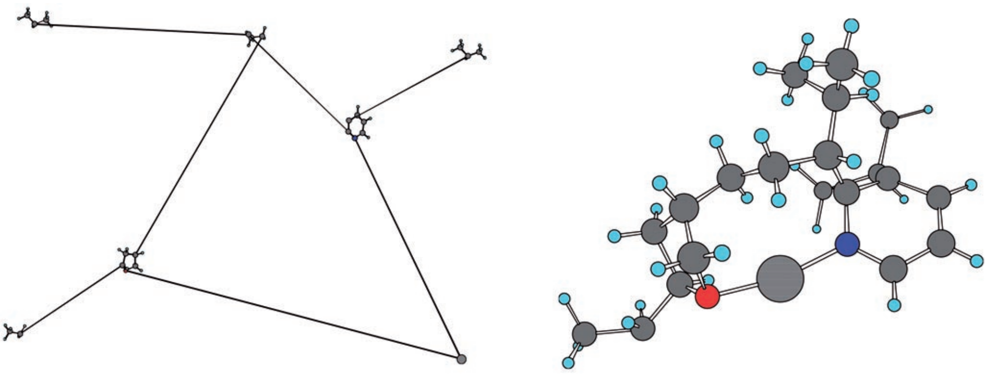


Figure 9. Catalyst structure directly after assembly (*left*), and following MM structure minimisation using Hyperchem (*right*).

(e.g., 1 to 10) each of the 500 structures in the training set is left out once and predicted with a model based on the other 499 structures. The number of latent variables in the model with the lowest average cross validation error is chosen.[44]

Variable Importance Analysis

Variable importance is a tool to examine the relative importance of each predictor variable in a regression model.[14] The influence is calculated as the variable importance parameter (VIP), calculated using Eq. (2) for a parameter k:

$$VIP_k = \frac{\sum_{a=1}^{lv} b_{ak}^2 \cdot SSQ_a}{n \cdot \sum_{a=1}^{lv} SSQ_a} \tag{2}$$

In this equation, b_{ak} is the regression weight for variable k and factor a, SSQ_a is the percentage variance captured by latent variable a, n is the total number of variables and lv is the number of latent variables used in the regression model. The VIP magnitude depends on the number of latent variables, so that the absolute VIP values are less meaningful than the relative values in a given dataset.

References and Notes

[1] For a recent and excellent general monograph on homogeneous catalysis see P. W. N. M. van Leeuwen, *Homogeneous Catalysis: Understanding the Art*, Kluwer, Dordrecht, **2004**.

[2] M. Thommen, H.-U. Blaser, *Chim. Oggi* **2003**, *21*, 6.

[3] For a recent collection of essays on the principles and practice of THE, see: a) *High-Throughput Synthesis*, (Ed.: I. Sucholeiki), Marcel Dekker, New York, **2001**; for reviews on applying HTE in homogeneous catalysis, see: b) M. T. Reetz, *Angew. Chem. Int. Ed.* **2001**, *40*, 284, and references cited therein; for applications in heterogeneous catalysis, see: c) S. Senkan, *Angew. Chem. Int. Ed.* **2001**, *40*, 312; d) A. Holzwarth, P. Denton, H. Zanthoff, C. Mirodatos, *Catalysis Today* **2001**, *67*, 309.

[4] For a modular approach to constructing chiral ligands, see: F. Spindler, C. Malan, M. Lotz, M. Kesselgruber, U. Pittelkow, A. Rivas-Nass, O. Briel, H.-U. Blaser, *Tetrahedron: Asymmetry* **2004**, *15*, 2299.

[5] There are many reports on applications of HTE methods in catalysis, but relatively few commercial successes. An exception is the Pd-phosphinous acid systems discovered by Li and colleagues from DuPont's Combiphos, see: G. Y. Li, *Chemical Industries (CRC Press)*, **2005**, *104*, (Catal. Org. React.) 177.

[6] G. Y. Li, *J. Org. Chem.* **2002**, *67*, 3643.

[7] L. Chen, M. W. Deem, *J. Chem. Inform. Comp. Sci.* **2001**, *41*, 950.

[8] This premise applies just as well to heterogeneous catalysis, where the situation is further complicated by the plethora of parameters that must be taken into account in HTE catalyst preparation and treatment steps. See, for example a) C. Klanner, D. Farrusseng, L. Baumes, M. Lengliz, C. Mirodatos, F. Schüth, *Angew. Chem. Int. Ed.* **2004**, *43*, 5347; b) C. Klanner, D. Farrusseng, L. Baumes, C. Mirodatos, F. Schüth, *QSAR Comb. Sci.* **2003**, *22*, 729.

[9] For a discussion on combining selection algorithms and THE, see: J. A. Westerhuis, H. F. M. Boelens, D. Iron, G. Rothenberg, *Anal. Chem.* **2004**, *76*, 3171.

[10] H. Bönnemann, *Angew. Chem. Int. Ed. Engl.* **1985**, *24*, 248.

[11] K. D. Cooney, T. R. Cundari, N. W. Hoffman, K. A. Pittard, M. D. Temple, Y. Zhao, *J. Am. Chem. Soc.* **2003**, *125*, 4318.

[12] E. Burello, G. Rothenberg, *Adv. Synth. Catal.* **2003**, *345*, 1334.

[13] E. Burello, D. Farrusseng, G. Rothenberg, *Adv. Synth. Catal.* **2004**, *346*, 1844.

[14] E. Burello, P. Marion, J. C. Galland, A. Chamard, G. Rothenberg, *Adv. Synth. Catal.* **2005**, *347*, 803.

[15] For an excellent review on measuring chemical similarity/diversity, see: N. Nikolova, J. Jaworska, *QSAR Comb. Sci.* **2003**, *22*, 1006.

[16] For an application of such global search methods, see: J. A. Hageman, R. Wehrens, H. A. Van Sprang, L. M. C. Buydens, *Anal. Chim. Acta* **2003**, *490*, 211.

[17] For a discussion on choosing molecular descriptors for virtual screening, see: L. Xue, J. Bajorath, *Comb. Chem. High Throughput Scr.* **2000**, 3.

[18] P. W. N. M. van Leeuwen, P. C. J. Kamer, J. N. H. Reek, *Pure Appl. Chem.* **1999**, *71*, 1443.

[19] P. Dierkes, P. W. N. M. Van Leeuwen, *J. Chem. Soc. Dalton Trans.* **1999**, 1519.

[20] For a review on the properties of non-phosphine chelate ligands, see: K. J. Cavell, *Aust. J. Chem.* **1994**, *47*, 769.

[21] For a discussion on the natural bite angle of diphosphines, see: C. P. Casey, G. T. Whiteker, *Isr. J. Chem.* **1990**, *30*, 299.

[22] For a comparison between 2D and 3D descriptors in homogeneous catalysis, see: E. Burello, G. Rothenberg, *Adv. Synth. Catal.* **2005**, *347*, 1969–1977.

[23] Even using state-of-the-art hydroformylation catalysts, optimisation of process conditions is still a challenge; see: P. B. Webb, T. E. Kunene, D. J. Cole-Hamilton, *Green Chem.* **2005**, *7*, 373.

[24] D. F. Foster, D. Gudmunsen, D. J. Adams, A. M. Stuart, E. G. Hope, D. J. Cole-Hamilton, G. P. Schwarz, P. Pogorzelec, *Tetrahedron* **2002**, *58*, 3901.

[25] P. Cheliatsidou, D. F. S. White, D. J. Cole-Hamilton, *Dalton Trans.* **2004**, 3425.

[26] Z. Freixa, P. W. N. M. Van Leeuwen, *Dalton Trans.* **2003**, 1890.

[27] Z. Freixa, M. S. Beentjes, G. D. Batema, C. B. Dieleman, G. P. F. van Strijdonck, J. N. H. Reek, P. C. J. Kamer, J. Fraanje, K. Goubitz, P. W. N. M. van Leeuwen, *Angew. Chem. Int. Ed.* **2003**, *42*, 1284.

[28] M. Matsumoto, M. Tamura, *J. Mol. Catal. 1982*, *16*, 209.

[29] M. T. Reetz, S. R. Waldvogel, R. Goddard, *Tetrahedron Lett.* **1997**, *38*, 5967.

[30] J. I. van der Vlugt, R. Sablong, P. C. M. M. Magusin, A. M. Mills, A. L. Spek, D. Vogt, *Organometallics* **2004**, *23*, 3177.

[31] R. P. J. Bronger, J. P. Bermon, J. Herwig, P. C. J. Kamer, P. W. N. M. Van Leeuwen, *Adv. Synth. Catal.* **2004**, *346*, 789.

[32] C. P. Casey, G. T. Whiteker, M. G. Melville, L. M. Petrovich, J. A. Gavney, Jr., D. R. Powell, *J. Am. Chem. Soc.* **1992**, *114*, 5535.

[33] For a general discussion on regression variable selection, see: A. J. Burnham, J. F. MacGregor, R. Viveros, *J. Chemom.* **2001**, *15*, 265.

[34] J. Niu, G. Yu, *Environ. Toxicol. Pharmacol.* **2004**, *18*, 39.

[35] M. Randic, *J. Am. Chem. Soc.* **1975**, *97*, 6609.

[36] H. Wiener, *J. Am. Chem. Soc.* **1947**, *69*, 17.

[37] L. B. Kier, L. H. Hall, *Eur. J. Med. Chem.* **1977**, #20#*12*, 307.

[38] N. S. Zefirov, M. A. Kirpichenok, F. F. Izmailov, M. I. Trofimov, *Dokl. Akad. Nauk SSSR* **1987**, *296*, 883.

[39] M. A. Kirpichenok, N. S. Zefirov, *Zh. Org. Khim.* **1987**, 23.

[40] J. A. Hageman, R. Wehrens, R. de Gelder, W. Leo Meerts, L. M. C. Buydens, *J. Mol. Phys.* **2000**, *113*, 7955.

[41] J. A. Hageman, M. Streppel, R. Wehrens, L. M. C. Buydens, *J. Chemom.* **2003**, *17*, 427.

[42] MATLAB^{TM} is commercially available from MathWorks, Natick USA, version 6.1, 2001.

[43] HyperChem^{TM} Professional 7.51, Hypercube, Inc., 1115 NW 4th Street, Gainesville, Florida 32601, USA.

[44] For a detailed description of regression methods, see: a) B. G. M. Vandeginste, D. L. Massart, L. M. C. Buydens, S. d. Jong, P. J. Lewi, J. Smeyers-Verbeke, *Handbook of chemometrics*, Vol. 20B, Elsevier, Amsterdam, **1998**; for a specific tutorial on PLS regression, see: b) P. Geladi, B. R. Kowalski, *Anal. Chim. Acta* **1986**, *185*, 1.