Combinatorial Catalysis

The Development of Descriptors for Solids: Teaching "Catalytic Intuition" to a Computer

*Catharina Klanner, David Farrusseng, Laurent Baumes, Mourad Lengliz, Claude Mirodatos, and Ferdi Schüth**

High-throughput experimentation has become an accepted and important strategy in the search for novel catalysts and materials.[1–7] However, one of the major problems is still the design of libraries, especially, if vast numbers of catalysts are to be explored. On the other end of the work flow, after catalysts have been tested, data mining and the search for trends is equally demanding. Several methods based on expert systems[8] have been proposed to support the development of solid catalysts. Also the correlation of performance with catalyst composition, evaluated by neural networks, has been used for the optimization of catalysts.[9–11] For such optimization programs in catalysis, evolutionary algorithms were found to be helpful as well.[12,13] However, in these approaches the scope was usually very limited, and an attempt to include a wide range of properties to describe the solids was not made. A more integrated "knowledge extraction engine" has been proposed by Caruthers et al.[14] for propane aromatization which is, however, focused on the reaction engineering aspects.

There is a great need for software-based methods to plan the design of libraries based on chemical knowledge, in addition to the statistical tools which are implemented in some of the commercial software packages. QSAR (quantitative structure–activity relationship) is one of the most powerful methods used in drug discovery to design libraries and to extract knowledge from tests on possible drug libraries. Such structure–activity relationships are discovered by computer programs, for which molecules need to be represented in computer data bases by so-called descriptors. The descriptors can, for instance, be two-dimensional fingerprints, such as absence or presence of certain chemical functional groups, or can be pharmocophores, which relate to the relative spatial arrangement of three selected chemical functional groups, or physico–chemical properties, or many others.[15] Whole journals are by now devoted to this topic.[16]

However, owing to the different nature of the problem, a transfer of descriptor concepts to solids has not been possible to date. In contrast to molecules, a solid can not easily be represented in a computer, since no structural formula can be given and encoded. If only the composition of a solid would be used, many important factors would be lost, since properties of a catalyst, for instance, are also very much dependent on the synthesis and the conditions of the reaction itself. There is one preceding study in which descriptors have been used to correlate structural features of zeolites with the ring size in the structures.[17] Recently, we suggested a methodology to apply to heterogeneous catalysis which was expected to work in a similar manner to the molecular descriptors.[18] It can to some extent be considered to be a multidimensional version of the volcano principle known in catalysis for decades.[19] Herein, we show that these concepts can indeed be implemented. The descriptors thus developed have predictive power and the concept can therefore be considered as the transfer of "catalytic intuition" to a computer.

The method, in short, consists of the creation of a library of solids, testing of the performance in a catalytic reaction, description of the solids in terms of a multitude of attributes which are available either from the synthesis of the solid or from tabulated physico–chemical data, and, finally, the identification of a set of those attributes which allow discrimination between different catalytic performance.

A highly diverse library was synthesized, consisting of 467 different catalysts. Diversity in this case was judged by chemical intuition, based on the accumulated knowledge in the field. The library included binary oxides, multinary mixed oxides, supported catalysts on different support materials with various supported compounds, zeolites, and many other types. All the catalysts of this library were tested in the oxidation of propene with oxygen (O_2:C_3H_6 = 5:1, that is, slightly above stoichiometric for total oxidation) in a 16-fold parallel reactor which was a more advanced stainless steel version of the system described by Hoffmann et al.[20] Products were analyzed sequentially by GC, which allowed the detection of about 30 products. Each catalyst was measured twice, which also allowed its temporal behavior to be analyzed, at five temperature levels (200, 250, 300, 400, and 500 °C). In this way 120 parameters, that is, conversions, selectivities to 21 products, temporal behavior, and carbon mass balance, all at each temperature, were generated for every catalyst.

This set of data is too vast by far for a meaningful attempt at a correlation. We have thus classified the catalytic performances into distinct groups with respect to an analysis of the 120 output parameters, using principal components analysis and then clustering techniques based on euclidian distance. Figure 1 shows as an example the results of a tree cluster analysis. Each of the classes can be identified with a specific catalytic performance of the solids, as given in the legend to Figure 1.

The other major task was the encoding of the solids. For a virtual screening, only such attributes are useful, which are either derived from a possible synthesis method or are tabulated, so that they do not need to be measured. For each catalyst we have created a set of 3179 attributes, which include the concentrations of 60 elements from the periodic table, 19 attributes which are related to the synthesis method, and 3100 attributes which are taken from tabulated data. These are, for instance, enthalpies of formation of different oxides, possible coordination numbers of the atoms, ionization energies, electronegativities, averages of such values for

[*] Dr. C. Klanner, Prof. Dr. F. Schüth
Max-Planck-Institut für Kohlenforschung
Kaiser-Wilhelm-Platz 1
45470 Mülheim (Germany)
Fax: (+49) 208-306-2995
E-mail: schueth@mpi-muelheim.mpg.de
Dr. D. Farrusseng, L. Baumes, Prof. Dr. C. Mirodatos
Institut de Recherches sur la Catalyse—CNRS
2, avenue A. Einstein, 69626 Villeurbanne Cedex (France)

 DOI: 10.1002/ange.200460731

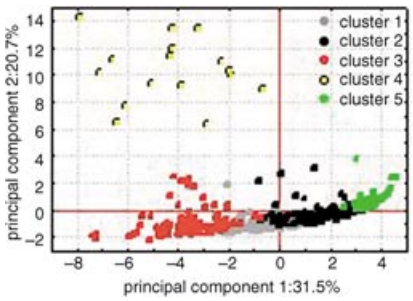

Figure 1. Results of a k-means cluster analysis based on eight principal components (PC) in the projection on the PC1-PC2 plane. The clusters which overlap in this projection are very well separated in the other principal components. With respect to catalytic performance, the catalysts can be described as follows: cluster 1 (gray) low activity, total oxidation, cluster 2 (black) medium activity, total oxidation, cluster 3 (red) low activity, CO and partial oxidation products, cluster 4 (yellow) hydrocarbon formation, cluster 5 (green) high activity, total oxidation.

test	cluster 1	2	3	4	5	sum	ratio	prediction rate	sensibility
1 predicted	16	3	4	0	2	25	0.25	0.64	0.57
2 predicted	5	18	6	0	8	37	0.29	0.49	0.55
3 predicted	5	4	16	0	1	26	0.24	0.62	0.59
4 predicted	1	0	0	3	0	4	0.03	0.75	1.00
5 predicted	1	8	1	0	11	21	0.19	0.52	0.50
sum/ mean	28	33	27	3	22	113		0.60	0.64

Figure 2. Confusion matrix for descriptor-based classification of catalysts using an artificial neural network of the multilayer perceptron type. 45 attributes were selected as relevant by the network out of the 75 initial attributes. In total 113 different catalysts were classified, the colored boxes mark the number of correctly classified catalysts. The columns 1–5 indicate the cluster to which the catalysts belong (see legend to Figure 1), the rows 1 predicted to 5 predicted indicate the cluster (1–5) which was predicted for a given catalyst. The ratio describes the statistical expectation for the fraction of correctly classified catalysts, the prediction rate indicates the fraction of catalysts assigned to a certain cluster which actually belong to this cluster. This value can directly be compared to the ratio.

multicomponent catalysts, variance of such values for multicomponent catalysts, and so on.

With 3179 attributes, a system of 467 catalysts is hopelessly over determined. Prior to a correlation, the number of attributes needed to be reduced. In principle, two different selection approaches can be chosen: The chemist can select those parameters which appear to be most promising, or software-based methods called feature selection can be employed. In our case different feature selection routines were tested, but none allowed discrimination between relevant attributes and attributes which had no correlation to the performance of the catalysts (although we would not claim that such discrimination is impossible). Thus, the number of attributes was reduced to 75 attributes by intuition, including all synthesis-related parameters and a set of parameters related to the properties of the elements, ions, or oxides.

For the correlation, both neural networks and classification trees, were used. In general, neural networks gave a better prediction of the performance class than classification trees. For the neural-network analysis, the catalyst set was divided at random into the training set (50 % of the catalysts), the selection set (25 %), and the test set (25 %), for the classification trees, two groups were formed, the training set (66 %) and the test set (33 %). Neural networks were trained for various different clusters based on a different number of principal components (PC), but in all cases the predictions were of comparable quality and vastly superior over a mere statistical prediction. Figure 2 gives a so-called confusion matrix for the prediction achieved with a neural network of the multilayer perceptron type on a data set with five clusters based on eight PCs. The initial 75 attributes were reduced to 45 relevant ones by the network algorithm. The "ratio" listed in the matrix gives the fraction of cases which would be predicted to belong to the specific class, if the assignment were made at random. This value can be compared directly with the "prediction rate", which accounts for the correctly classified cases in the respective predicted class. As can be seen, the prediction rate far exceeds the ratio in all cases. The prediction is thus substantially better than statistically expected. The "sensibility" is the fraction of cases correctly classified from the respective original class. For this value, no proper statistical benchmark can be given, but the numbers are all rather high, again indicating how good the correlation is. In addition, misclassifications occur predominantly in "catalytically related" classes, that is, catalysts are, for instance, sorted into the medium-activity class instead of the correct high-activity class, but less often into the low-activity class.

Classification tree analysis was also performed for several cases and Figure 3 gives a confusion matrix for such an analysis. The number of attributes was reduced to 23 relevant ones by the algorithm. In general, all classification trees performed far better than a random prediction, but were inferior to neural networks. A major discriminative effect in this type of analysis has the normalized formation free enthalpy of the most stable metal oxide of all the elements in the catalyst. Influence of such a factor in an oxidation reaction would have been expected from heuristic knowledge as well, so that one can say with some justification that chemical intuition has been implemented in an algorithm.

test	cluster 1	2	3	4	5	sum	ratio	prediction rate	sensibility
1 predicted	18	7	14	0	7	46	0.25	0.39	0.45
2 predicted	6	19	10	1	9	45	0.27	0.42	0.44
3 predicted	11	2	12	0	0	25	0.25	0.48	0.31
4 predicted	2	3	2	5	0	12	0.04	0.42	0.83
5 predicted	3	12	1	0	14	30	0.19	0.47	0.47
sum/ mean	40	43	39	6	30	158		0.44	0.50

Figure 3. Confusion matrix for descriptor-based classification of catalysts using a classification tree analysis. 23 attributes out of the initial 75 attributes were selected as relevant by the algorithm. Explanations are as for Figure 2. As can be seen, both the prediction rate and the sensibility are lower than for the neural network based analysis but still substantially better than a random assignment to the clusters.

It is also revealing to inspect those attributes which were selected to be of influence by almost all neural networks and classification trees. These attributes are the maximum difference in the atomic radius of all the elements present in a

catalyst, the mean electron affinity of all the elements in the catalyst, the mean Pauling electronegativity of all the metals and semimetals in the catalyst, the normalized formation free enthalpy of the most stable metal oxide of all the elements in the catalyst, the weighted mean molar mass of all the elements in the catalyst, the difference between the highest and the lowest ionic radius of all the elements (average radius as basis for each element), the synthesis pathway, and the fact whether a base was added in the synthesis. That these eight parameters out of the 75 initially selected show up in most of the correlations suggests that the combination of them has a major influence in determining the catalytic performance, and if one inspects these properties one could indeed expect some predictive power from such parameters.

In summary, based on parameters, which do not have to be measured, sets of attributes for solids were derived which can be used to predict whether a catalyst falls into one out of five performance classes in propene oxidation, with a predictive power substantially exceeding the statistically expected values. Figure 4 summarizes these results. This is the same

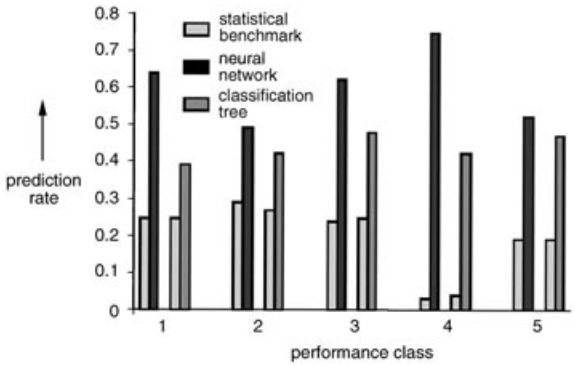

Figure 4. Comparison of the prediction rates for the different catalyst classes based on neural network analysis and classification tree analysis together with the statistical expectation value. Statistical expectation values are not identical, because the test set contained 113 at random selected catalysts for the neural network analysis and 158 at random selected catalysts for the classification tree analysis.

as what an able chemist can do based on his or her experience and chemical knowledge. Each catalysis researcher will, for instance, suggest that Pt/Al_2O_3 should be a good total oxidation catalyst, or that bismuth molybdenum oxide may form partial oxidation products. However, implementing this kind of intuition on a general level into an algorithm is exceedingly difficult. We have implemented a solution to this problem and shown for one test case that the concept works in practice. With a sufficiently broad database, one can expect that the descriptors initially derived for propene oxidation can be generalized to alkene oxidations or even to hydrocarbon oxidation reactions, and the concept will be more and more reliable the broader the database becomes and will thus provide the basis of virtual screening as a first step in a catalyst-discovery program.

The identification of a descriptor vector now opens the pathway to a virtual screening of solids. For such an approach, a multitude of catalysts would be generated theoretically, using, for instance, a randomizer to determine the synthesis pathway and the composition. For each of the catalysts suggested by the algorithm, the descriptor vector would be determined. Then only catalysts for which a desired performance is expected would indeed be synthesized and tested, or, if a highly diverse library is targeted at, several examples would be selected from each predicted performance class. Since the randomizer would suggest a composition, precursors, and a synthesis pathway, there is a high probability, that suggested catalysts can indeed be synthesized.

The concept is not restricted to catalysis. In general, any materials science problem involving complex solids could be tackled by the methodology which we have introduced.

Received: May 20, 2004

Keywords: combinatorial chemistry · heterogeneous catalysis · high-throughput screening · oxidation · virtual screening

[1] X. D. Xiang, X. Sun, G. Briceno, Y. Lou, K.-A. Wang, H. Chang, W. G. Wallace-Freedman, S.-W. Chen, P. G. Schultz, *Science* **1995**, *268*, 1738.
[2] F. C. Moates, M. Somani, J. Annamalai, J. T. Richardson, D. Luss, R. C. Wilson, *Ind. Eng. Chem. Res.* **1996**, *35*, 4801.
[3] S. M. Senkan, *Nature* **1998**, *394*, 350.
[4] E. Reddington, A. Sapienza, B Guraou, R. Viswanathan, S. Sarangapani, E. S. Smotkin, T. E. Mallouk, *Science* **1998**, *280* 1735.
[5] A. Holzwarth, H. W. Schmidt, W. F. Maier, *Angew. Chem.* **1998**, *110*, 2788; *Angew. Chem. Int. Ed.* **1998**, *37*, 2644.
[6] B. Jandeleit, D. J. Schaefer, T. S. Powers, H. W. Turner, W. H. Weinberg, *Angew. Chem.* **1999**, *111*, 2648; *Angew. Chem. Int. Ed.* **1999**, *38*, 2495.
[7] S. Senkan, *Angew. Chem.* **2001**, *113*, 322; *Angew. Chem. Int. Ed.* **2001**, *40*, 312.
[8] A good survey is given in: M. Baerns, E. Körting in *Handbook of Heterogeneous Catalysis* (Eds.: G. Ertl, H. Knözinger, J. Weitkamp), Wiley-VCH, Weinheim, **1997**, pp. 419–426.
[9] T. Hattori, S. Kito, *Catal. Today* **1995**, *23*, 347.
[10] S. Kito, T. Hattori, Y. Murakami, *Ind. Eng. Chem. Res.* **1989**, *31*, 979.
[11] A. Corma, J. M. Serra, E. Argente, V. Botti, S. Valero, *ChemPhysChem* **2002**, *3*, 939.
[12] D. Wolf, O. Buyevskaya, M. Baerns, *Appl. Catal. A* **2000**, *200*, 63.
[13] U. Rodemerck, M. Baerns, M. Holena, D. Wolf, *Appl. Surf. Sci.* **2004**, *223*, 168.
[14] J. M. Caruthers, J. A. Lauterbach, K. T. Thomson, V. Venkatasubramanian, C. M. Snively, A. Bhan, S. Katare, G. Oskarsdottir, *J. Catal.* **2003**, *216*, 98.
[15] R. Todeschini, V. Consonni, *Handbook of Molecular Descriptors*, Wiley-VCH, Weinheim, **2000**.
[16] For instance, the journals *QSAR-Combinatorial Science* and *J. Chem. Inf. Comp. Sci.* carry a majority of papers from this field.
[17] A. Rajagopalan, C. Suh, X. Li, K. Rajan, *Appl. Catal. A* **2003**, *254*, 147.
[18] C. Klanner, D. Farrusseng, L. Baumes, C. Mirodatos, F. Schüth, *QSAR Comb. Sci.* **2003**, *22*, 729.
[19] M. Boudart in *Handbook of Heterogeneous Catalysis* (Eds.: G. Ertl, H. Knözinger, J. Weitkamp), Wiley-VCH, Weinheim, **1997**, pp. 1–13.
[20] C. Hoffmann, A. Wolf, F. Schüth, *Angew. Chem.* **1999**, *111*, 2971; *Angew. Chem. Int. Ed.* **1999**, *38*, 2800.

