Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

Journal of Catalysis 232 (2005) 335–341

JOURNAL OF CATALYSIS

www.elsevier.com/locate/jcat

Integrating high-throughput characterization into combinatorial heterogeneous catalysis: unsupervised construction of quantitative structure/property relationship models

A. Corma [*], J.M. Serra, P. Serna, M. Moliner

Instituto de Tecnología Química, UPV-CSIC, Av. Naranjos s/n, E-46022 Valencia, Spain

Received 23 December 2004; revised 4 March 2005; accepted 13 March 2005

Available online 26 April 2005

Abstract

This paper presents a novel approach in the framework of heterogeneous combinatorial catalysis, which integrates into the global discovery strategy the use of inexpensive high-throughput characterization of libraries of catalysts, as multivariate spectral descriptors for catalytic quantitative structure/property relationship (QSPR) modeling. Moreover, QSPR models can be used to assist the design of new libraries and for extraction of rules and relationships, yielding knowledge about catalysis. This approach can be of special interest when experimental evaluation of catalytic behavior is very expensive or time-consuming, as, for instance, for catalyst deactivation studies, for testing under very severe conditions, or when high amounts of catalyst are demanded. This methodology has been applied to modeling of the behavior of epoxidation catalysts, with the composition vector of the starting synthesis gel and XRD spectra as descriptors. Dimensional reduction was conducted by principal components analysis, clustering, and Kohonen networks, and predictive models were obtained with the use of logistic equations, artificial neural networks, and decision tree techniques. The use of spectral descriptors made it possible to markedly improve the prediction performance obtained with synthesis descriptors alone.

Keywords: Catalyst descriptors; Spectral descriptors; High-throughput experimentation; Characterization; Combinatorial catalysis; Data mining; QSPR

1. Introduction

In the field of heterogeneous catalysis, new experimental tools are available for high-throughput (HT) materials synthesis, catalytic testing, and physicochemical characterization. Such tools make it possible to study simultaneously a large number of variables, like multicomponent catalyst formulation, synthesis procedure, activation conditions, etc. Furthermore, HT experimentation or so-called combinatorial catalysis has become an accepted approach to new catalytic material discovery and development. This approach requires the utilization of complex library design strategies, new data mining techniques, and database technology.

HT experimentation in drug discovery and medical chemistry is a widely applied and mature approach from which combinatorial catalysis has taken and adapted most of its experimental and software tools. However, the proper description of solid extended catalysts and the ab initio calculation [1] of their properties are very complex and are still in their infancy [2,3], especially when compared with the available chemoinformatics software suites for virtual screening of drug candidate molecules [4–6]. Molecular descriptors used in the pharmaceutical chemistry are variables that represent the physiochemical properties of a class of compounds [7–9], and they are commonly classified into four types: (a) constitutional (molecular formula); (b) topological (molecule connectivity matrix); (c) geometrical (3D molecular models); and (d) quantum chemical properties: (semi-empirical or ab initio calculations). In recent years [10–12], a complete descriptor database has been developed for solid

[*] Corresponding author. Fax: +34 96 387 7809.
E-mail address: acorma@itq.upv.es (A. Corma).

0021-9517/$ – see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.jcat.2005.03.019

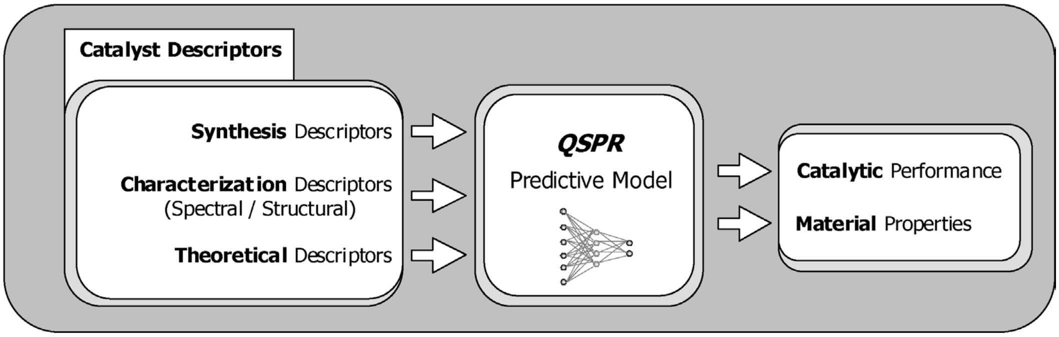

Fig. 1. General concept of QSPR predictive models in catalysis science.

catalysts based on molecular descriptors. Nevertheless, the use of fundamental descriptors is very limited, since heterogeneous catalysts are complex systems, with heterogeneous multicomponent active sites and unpredictable metastable structures. In some cases, the tabulated information corresponding to the starting elements/oxides making up the catalyst can be of help, as was interestingly shown [13,14] for the catalytic oxidation of propylene with oxygen. The authors found that of the more than 3000 tabulated attributes of the catalyst, just six of them could be correlated with certain significance with the experimental performance of the final solid catalyst (e.g., the Pauling electronegativity of all of the metals and semimetals or the normalized formation free enthalpy of the most stable metal oxide of all the elements in the catalyst). Predictive modeling was conducted with the use of neural networks and classification trees. Neural networks gave the best results, and the model output corresponded to the class of catalytic behavior. However, one should expect that tabulated properties would be of help only when an effect of one of the catalyst components is predominant.

Another interesting approach in the field of polymer science is fingerprint technology and high-output screening [15]. In this case, the use of molecular descriptors is also very limited, since a final polymer does not possess an exact chemical formula or structure, and it is better characterized by its average properties, polymerization conditions, and ingredients. Novel fingerprint technology makes use of fast, nondestructive, and inexpensive spectroscopy measurements combined with data mining tools to reduce the need for extensive and tedious polymer testing for rheological and long-term mechanical properties. Therefore, HT characterization of a well-defined model polymers library is used for the rapid establishment of quantitative structure/property relationships (QSPRs) by multivariate analyses. Different characterization techniques [16–18] have been applied for fingerprint polymer and QSPR modeling, such as X-ray fluorescence [19], fluorescence spectroscopy [20], and ATR-FTIR spectroscopy [21], which allow for the rapid analysis of powders and polymeric solids without the need for sample preparation, whereas the data mining tools included principal component analysis, multivariate partial least squares, or decision trees.

Up to now, in the field of heterogeneous catalysis only the development of new HT techniques has been reported; no reports on data analysis and the ulterior integration in the whole combinatorial loop or global discovery strategy have been published. The available HT characterization techniques for solid catalysts include XRD systems [22], acidity determination by TPD-NH_3 [23] and IR-pyridine adsorption [24], parallel TAP reactor studies [25], and photoluminescence [26]. Conversely, data mining has been successfully applied in this field for the analysis and extraction of multifactor relationships (QSPR), with the use of different modeling techniques [13,27] like artificial neural networks, decision trees, principal components regression, etc. This extracted knowledge can subsequently be integrated into library design tools [28–31], making it possible to reduce the experimental effort needed to reach a convergence criterion.

We propose here a novel approach in the framework of heterogeneous combinatorial catalysis, which integrates into the global discovery strategy [32] the use of inexpensive HT characterization of libraries of catalysts, as multivariate spectral descriptors for predictive modeling of the catalytic behavior. Indeed, spectral data can be used together with synthesis and theoretical data as input descriptors for catalytic QSPR modeling (Fig. 1). As summarized in Fig. 2, spectral descriptors can be obtained automatically by processing of the raw characterization data. The QSPR model obtained by different data-mining techniques can be used (i) as a predictive model, assisting the DoE of new catalyst libraries; and (ii) for extraction of rules and relationships between the different variables, and gaining knowledge about catalysis.

As a proof of this principle, we have applied these concepts to a specific case, using experimental data obtained with HT tools, following an evolutionary strategy. First, different unsupervised methods for dimension reduction of the raw spectral HT characterization data, clustering algorithms, principal component analysis and Kohonen neural networks, are used to obtain a series of experimental spectral descriptors. Subsequently, the construction of predictive models for the catalytic activity with the use of synthesis and characterization descriptors is studied, by the application of different modeling techniques. The influence of the dimensionality re-

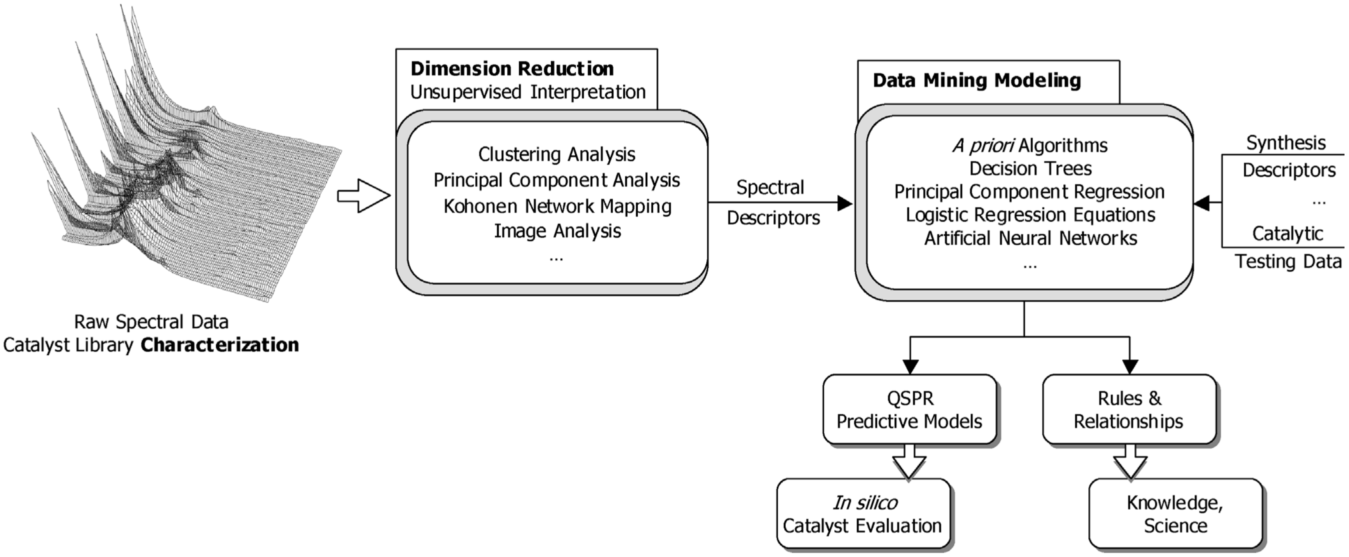

Fig. 2. Workflow of processing of characterization data and the further mining treatment together with synthesis and testing data, aiming to obtain QSPR models and to extract fundamental knowledge.

duction approach on the final prediction performance will be also examined.

2. Experimental

Experimental data were drawn from the HT optimization of epoxidation catalysts based on mesoporous titanium silicate materials [33]. In that work, the experimental design was determined by a hybrid optimizer comprising a genetic algorithm assisted by an ANN [30] and experimental data corresponding to the three evolved catalyst generations. The object was to determine the yield of cyclohexene epoxide for optimal material synthesis parameters, that is, the molar concentrations of the components of starting gel. The distribution of materials in the explored space is complex, because of the evolutionary design strategy applied.

Apart from the optimization process, HT characterization was carried out to gain a fundamental understanding of the catalysis of the process. Therefore, X-ray diffraction analyses were available for most of the screened Ti-silicate materials (extracted and sylilated). Although the interpretation of these data for mesoporous materials is not obvious, low-angle XRD spectra (1–8 2θ) give information about the long-distance order of the materials, that is, the type of "crystallographic system," dimensions of the pores, and particle size. Consequently, it is expected that XRD data could be correlated with the final catalytic activity of the material. Concretely, XRD measurements showed that from a crystallographic point of view, three types of materials are occurring in the explored multivariate space, that is, MCM-41 and MCM-48 with different degrees of structural order, and low-ordered materials. X-ray powder measurements were performed with a Philips X'Pert MPD diffractometer equipped with a PW3050 goniometer, with the use of Cu-K_α radiation and a multisample handler.

The experimental data employed here can therefore be divided into three groups: (i) compositional synthesis data: [TMA], [CTMA], [OH], and [Ti] concentrations; (ii) XRD measurements; and (iii) catalytic performances, that is, epoxide yield, which was classified into five classes, from "very bad" to "very good" yields.

3. Unsupervised analysis of XRD patterns: dimensionality reduction

XRD characterization data consist of arrays of more than 230 data, which could hardly be processed by correlation methods when they were used directly as input variables for predictive modeling (QSPR). For this reason and for reduction of experimental noise, a previous dimension reduction of the abundant raw data is required. In the present case, XRD data are analyzed and projected to a discrete number of dimensions (from 1D, 2D to 5D) with different data-mining techniques: (i) clustering analysis using K-means and two-step algorithms; (ii) principal component analysis (PCA); and (iii) Kohonen neural networks. Furthermore, these data would be used as multivariate spectral descriptors for catalytic activity modeling.

In the present approach, instead of applying direct human interpretation of the XRD spectral data, we decided to employ unsupervised analysis techniques because of the convenience of automating the processing of high amounts of data and in order to avoid the subjectivity introduced by human interpretation.

PCA is a widely applied statistical methodology [34] that allows the dimensionality of information space to be reduced with a minimum loss of information. This method is based

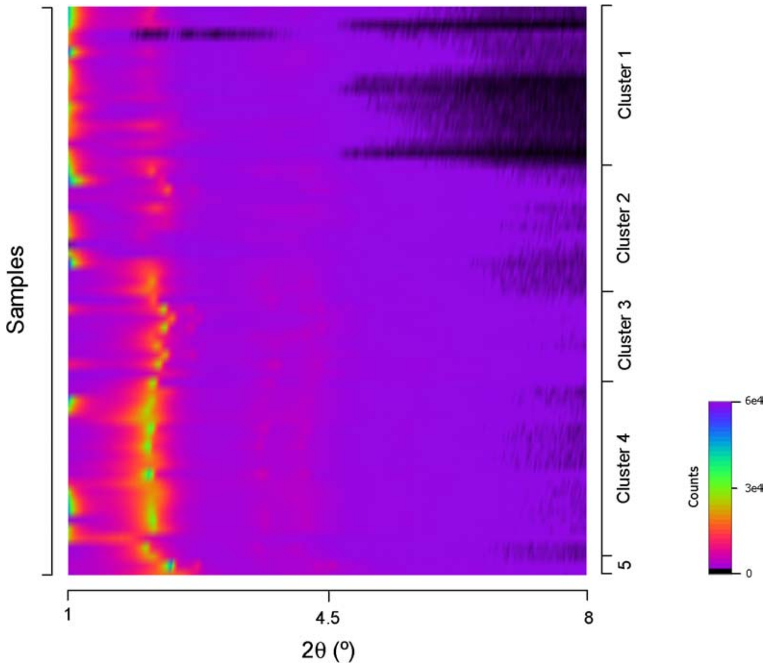

Fig. 3. PXRD measurement data of epoxidation catalysts ordered considering the cluster distribution obtained by Two-Step algorithm.

on the fact that input variables (raw descriptors) are closely correlated; then it is possible to find a set of new uncorrelated variables (descriptors called *principal components*). Therefore, principal components are essentially a linear combination of the original descriptors. On the other hand, clustering [35] can be considered the most important unsupervised learning methods, which find a structure in a collection of unlabeled data by organizing them into groups or *clusters* according to their similarity. Finally, the objective of a Kohonen network [36–38] is to map input vectors (patterns) of arbitrary dimension N onto a discrete map with one or two dimensions. Samples close to one another in the input space should be close to one another in the map Kohonen (topologically order). A Kohonen network is composed of a grid of output units and N input units. The input pattern is fed to each output unit.

Fig. 3 shows the complete spectral data ordered according to which of the different clusters generated by a two-step clustering algorithm they belong to. The five clusters found can be somewhat correlated with knowledge-based observations; that is, the clusters correspond to low-ordered MCM-41, medium-ordered MCM-41, MCM-48, and high-ordered MCM-41. Fig. 4 shows the cluster distribution obtained with K-means and two-step clustering algorithms, when the principal components obtained by PCA computation and the Kohonen map (10×7) are plotted as coordinates. Clustering analyses were performed with these two clustering algorithms, with a minimum of 5 and 10 clusters. The results show that apparently the clusters found by the two-step logarithm are clearly separated from each other, when displayed with the PCA projection, and the most appropriate number of clusters is 5. Furthermore, Figs. 4c–d presents the mapping, obtained by Kohonen network training, of the samples classified according to the cluster distribution obtained by the two-step algorithm. As for the PCA projection, it can also be observed that clusters found with the two-step algorithm are clearly recognizable within the Kohonen mapping (2D projection).

These unsupervised techniques can successfully organize and classify the spectral XRD dataset, reducing the data dimensionality of the redundant raw data. The next step is to study the suitability of these multivariate spectral descriptors for predictive modeling, used alone or in combination with synthesis descriptors.

4. Construction of predictive models with the use of synthesis and multivariate spectral descriptors

Data-mining techniques make it possible to discover hidden patterns or relationships among large amounts of data with multidimensional structure [39,40]. On the basis of this knowledge extracted from experimental or ab initio computed descriptors, data-mining techniques have made it possible to construct predictive models (QSPR) [27,29] in the field of heterogeneous catalysis. Among the different modeling techniques, we can draw attention to decision trees (DT) [41], logistic regression equations (LRE) [42], artificial neural networks (ANN) [43], the support vector machine (SVM) algorithm [44], and principal component regression (PCR) [45]. In this section, we attempt to develop a predictive model that relates the catalytic behavior of solid samples to synthesis descriptors and/or (structural) XRD spectral descriptors. This model will provide a basis for developing a

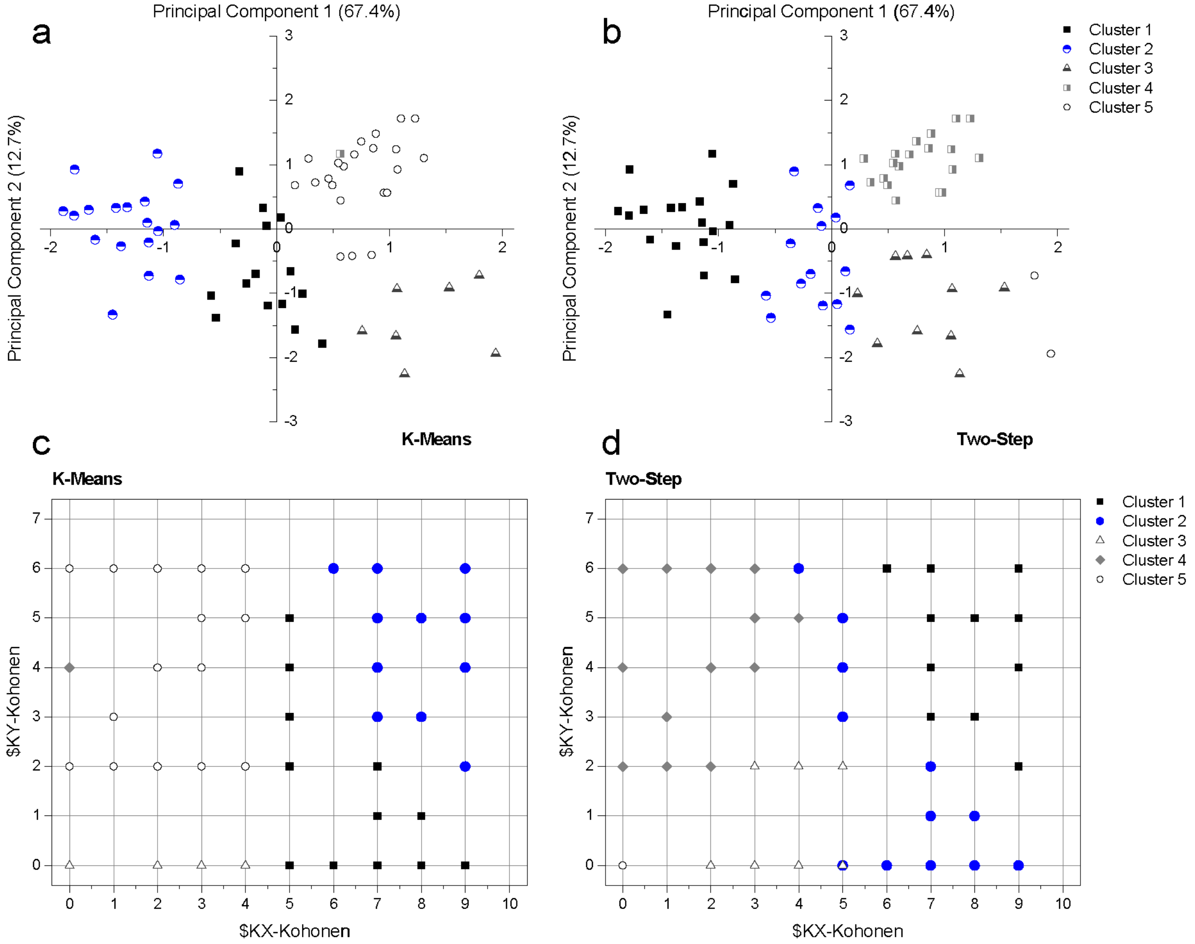

Fig. 4. *K*-Means and Two-Step clustering analysis represented in the PCA projection (a, b) and in the Kohonen map (c, d) obtained for the classification of 63 catalysts. The corresponding percentage of variance for each principal component (PC#) is 67.4, 12.7, 6.0, 5.4 and 2.4%. Kohonen map calculated with a network comprising 234 input neurons and 70 output neurons and using exponential learning rate decay.

QSPR, which can be used in discovery programs when HTS or combinatorial techniques are applied.

The influence of the different descriptors on prediction performance was studied with logistic equations as the model, whereas ANN and DT fitting will be done with the most appropriate descriptors. Table 1 summarizes the modeling results obtained with different catalyst descriptors and modeling techniques. When only the four synthesis variables were used as catalyst descriptors, the LRE fitted model showed a prediction accuracy of 65%. Interestingly, when only spectral descriptors obtained by PCA were used, the LRE prediction accuracy computed with two and five principal components was 50 and 55%, respectively. Although there is no direct information concerning elemental composition or the synthesis procedure, the structural information contained in XRD data permits some correlation of the catalytic activity. When only the Kohonen projections are used as descriptors, the accuracy of the fitted LRE model is 50%, but this value is increased (67%) when Kohonen and two-step clustering data are combined as descriptors. To check the convenience of the XRD data dimension reduction, the LRE model was fitted, with the 263 original attributes used as input variables, yielding a poor accuracy (32%).

The combination of synthesis and spectral descriptor was first studied by LRE modeling with the spectral principal components (2 and 5) and the four compositional descriptors. The model accuracy is higher (73 and 65%) than that obtained with both descriptors sets separately, illustrating the complementarity of the two data sets. Moreover, the introduction of characterization information would also make it possible to correct the experimental deviations introduced during the synthesis process. Although structural properties and synthesis descriptors are correlated, the use of spectral descriptors improves the modeling of highly nonlinear spaces. Furthermore, since only a few spectral descriptors are applied, such correlation between the two types of descriptor should not have any negative effect on the modeling process. Conversely, the accuracy is also increased (71%) when synthesis and Kohonen descriptors are used, and it is significantly increased (88%) when the information of two-step clustering is added. This model makes it possible to correctly predict the outcome of most of the samples, and

Table 1
Modeling results obtained using synthesis and characterization data as catalyst descriptors, processed by PCA, Kohonen networks and Two-Step clustering algorithm. The fitted models were logistic regression equations, artificial neural networks and decision trees. All modeling calculations were carried out using Clementine™ 6.0.2 application (SPSS Inc.)

	Logistic regression equations										Neural networks[a]						D. tree[d]		
	None	Raw data	PCA (2)	PCA (5)	Kohonen	Kohonen + clustering	PCA (2)	PCA (5)	Kohonen	Kohonen + clustering	None	PCA (2)	PCA (5)	Kohonen + clustering	Kohonen + clustering	Kohonen + clustering	None	Kohonen + clustering	Experimental
Synthesis descriptors Nr.	4	0	0	0	0	0	4	4	4	4	4	0	0	0	4[b]	1[c]	4	4	
Catalyst descriptors Nr.	4	234	4	5	2	3	6	9	6	7	4	5	2	3	7	4	4	7	
Very bad	4	10	2	5	3	4	4	4	4	4	0	0	2	7	2	5	4	4	4
Bad	2	2	2	2	3	4	2	2	2	2	0	0	0	0	3	2	2	2	2
Fair	3	3	0	5	1	10	6	11	10	10	0	15	5	9	10	11	11	10	10
Good	34	12	45	33	42	29	31	24	27	27	49	25	35	33	22	25	18	23	23
Very good	6	22	2	4	0	2	6	8	6	6	0	9	7	0	12	6	14	10	10
Prediction[e] (%)	65.3	32.7	49.9	55.1	49.9	67.3	65.3	73.5	71.4	87.8	46.9	67.3	61.2	65.3	93.9	87.8	89.8	100	

[a] Over-training was prevented by using 80% data for training and 20% for testing; fitting was repeated 5 times partitioning randomly the data sets and the obtained prediction deviation was ±4%.

[b] Best ANN model has a topology 11-22-10-5 following a *multiple* training method, as implemented in Clementine™ software (relative weight of input: $KX 0.51, $Clustering 0.50, $KY, 0.50, [CTMA] 0.03, [OH] 0.02, [TMA] 0.02 and [Ti] 0.02).

[c] It was employed only the titanium molar ratio as synthesis descriptor. Best ANN model has a topology 8-7-6-5 following a *multiple* training method.

[d] Fitting done using a classification and regression algorithm (C&R).

[e] Correct prediction of training and testing samples.

it failed in only a few samples when distinguishing between "good" and "very good" classes.

On the other hand, ANN fitting with only synthesis data yielded models with moderate prediction performance ($\sim 45\%$). A better performance ($> 65\%$) was obtained when spectral descriptors, that is, PCA and Kohonen and clustering, were used. The ANN prediction performance is again significantly improved by a combination of synthesis and spectral descriptors; that is, it was possible to reach values of about 90%. The last modeling technique was classification decision trees, with which it is feasible to obtain high prediction performances of about 100% with combined descriptors, improving that obtained with synthesis data alone (90%). The benefit of decision trees is that they make it possible to extract legible rules and conditions from the experimental data.

5. Conclusions

A novel approach integrating HT characterization in combinatorial heterogeneous catalysis has been proposed that describes how spectral characterization descriptors can be used in combination with synthesis descriptors for the unsupervised construction of QSPR models. This makes it possible to increase the prediction capability by introducing information about the *real* catalyst. Furthermore, characterization descriptors could be complemented with other catalyst descriptors, namely those based on tabulated and ab initio computed data [2,13].

This approach is exemplified in the modeling of the catalytic behavior of epoxidation catalyst based on mesoporous Ti-silicate materials. The composition vector of the starting synthesis gel and PXRD spectra were used as catalyst descriptors, whereas the epoxide yield obtained by catalyst testing was used as the outcome of the model. Dimensional reduction was conducted with the use of principal components analysis, clustering, and Kohonen networks, and combinations thereof, permitting extraction of the desired spectral descriptors from the XRD characterization data. Subsequently, predictive models (QSPR) were obtained with the use of logistic equations, artificial neural networks, and decision tree modeling techniques. The use of spectral descriptors made it possible to markedly improve the prediction performance over that obtained with synthesis descriptors alone. Moreover, reactor operating conditions could also be included as model input [43], permitting prediction of the catalyst performance under a wide range of process conditions.

HT characterization permits the incorporation of information about the final solid material, that is, coordination of the elements, types of crystalline structures making up the solid, etc. This complex final state includes the presence of metastable structures or coordination of specific elements, whose a priori prediction is very difficult to make, when only the catalyst composition and tabulated data are considered. However, these specific catalyst properties do have a paramount influence on the catalytic behavior of the material.

This approach can be of special interest when the experimental evaluation of the catalytic behavior is very expensive or time-consuming, as, for instance, for catalyst deactivation studies, testing under very severe conditions, or when high amounts of catalyst are demanded. It should be taken into account that, even though unsupervised QSPR construction is done, the intervention of a scientist to identify the adequate (inexpensive) characterization technique and adjust the experimental and calculation procedures would always be required.

We envisage that future work in HT catalyst development will apply the approach proposed here, complementing the state-of-the-art techniques [13,31]. The resulting QSPR models can be used to *virtually screen* the untested catalysts, providing information useful for guiding the next round of experimentation (reactivity testing). For further implementation of this approach, it would be necessary to increase the diversity of the mapped catalyst space and apply other spectroscopic techniques, with the intention of incorporating information about the coordination state of the elements dispersed over the catalyst surface, that is, UV–visible, Raman, and photoluminescence spectroscopy.

References

[1] D.J. Livingstone, D.T. Manallack, QSAR & Comb. Sci. 22 (2003) 510.

[2] S. Linic, J. Jankowiak, M.A. Barteau, J. Catal. 224 (2004) 489.

[3] P. Strasser, Q. Fan, M. Devenney, W.H. Weinberg, J. Phys. Chem. B 107 (2003) 11013.

[4] J. Bajorath (Ed.), Chemoinformatics: Concepts, Methods and Tools for Drug Discovery, Methods in Molecular Biology, vol. 275, Humana Press, Totowa, US, 2004.

[5] J.F. Blake, Curr. Opin. Chem. Biol. 8 (2004) 407.

[6] http://www.accelrys.com; http://www.inforsense.com.

[7] M. Karelson, Molecular Descriptors in QSAR/QSPR, Wiley, New York, USA, 2000.

[8] B.R. Beno, J.S. Mason, Drug Discovery Today 6 (2001) 251.

[9] B. McKay, M. Hoogenraad, E.W.P. Damen, A.A. Smith, Curr. Opin. Drug Discovery Dev. 6 (2003) 966.

[10] C. Klanner, D. Farrusseng, L. Baumes, C. Mirodatos, F. Schüth, QSAR & Comb. Sci. 22 (7) (2003) 729.

[11] StoCat™ description at http://catalyse.univ-lyon1.fr/gre3b4.htm.

[12] A. Frantzen, D. Sanders, J. Scheidtmann, U. Simon, W.F. Maier, QSAR & Comb. Sci. 24 (2005) 22.

[13] C. Klanner, D. Farrusseng, L. Baumes, C. Mirodatos, F. Schueth, Angew. Chem. Int. Ed. 43 (2004) 5347.

[14] D. Farrusseng, K. Klanner, L. Baumes, M. Lengliz, C. Mirodatos, F. Schüth, QSAR & Comb. Sci. 24 (2005) 78.

[15] A. Tuchbreiter, J. Marquardt, B. Kappler, J. Honerkamp, M.O. Kristen, R. Muelhaupt, Macromol. Rapid Commun. 24 (2003) 47.

[16] R.H. Hoogenboom, M.A.R. Meier, U.S. Schubert, Macromol. Rapid Commun. 24 (2003) 15.

[17] N. Adams, U.S. Schubert, J. Comb. Chem. 6 (2004) 12.

[18] E. Schneiderman, D. Stanton, T. Trinh, W.D. Laidig, M.L. Kramer, E.P. Gosselink, World Patent Application WO02/44686 A2, 2002.

[19] C. Vazquez, S. Boeykens, H. Bonadeo, Talanta 57 (2002) 1113.

[20] R.A. Potyrailo, R.J. Wroczynski, J.P. Lemmon, W.P. Flanagan, O.P. Siclovan, J. Comb. Chem. 5 (2003) 8.

[21] A. Tuchbreiter, J. Marquardt, J. Zimmermann, P. Walter, R. Muelhaupt, B. Kappler, D. Faller, T. Rohts, J. Honnerkaup, J. Comb. Chem. 3 (6) (2001) 598.

[22] J. Klein, C.W. Lehmann, H.W. Schmidt, W.F. Maier, Angew. Chem. Int. Ed. 37 (1998) 3369.

[23] H. Wang, Z. Liu, J. Shen, H. Liu, Catal. Commun. 5 (2004) 55.

[24] O.M. Busch, W. Brijoux, S. Thomson, F. Schuth, J. Catal. 222 (2004) 174.

[25] A.C. van Veen, D. Farrusseng, M. Rebeilleau, T. Decamp, A. Holzwarth, Y. Schuurman, C. Mirodatos, J. Catal. 216 (2003) 135.

[26] P. Atienzar, A. Corma, H. García, J.M. Serra, Chem. Eur. J. 10 (2004) 6043.

[27] A. Corma, J.M. Serra, E. Argente, S. Valero, V. Botti, Chem. Phys. Chem. 3 (2002) 939.

[28] F. Gilardoni, A. Graham, B. McKay, B. Brown, in: Proceedings of the 225th ACS National Meeting, New Orleans (USA), 23–27 March 2003.

[29] L. Baumes, D. Farrusseng, M. Lengliz, C. Mirodatos, QSAR & Comb. Sci. 23 (2004) 767.

[30] S. Valero, E. Argente, V. Botti, J.M. Serra, A. Corma, in: E. Conejo, M. Urretavizcaya, J.L. Perez-de-la-Cruz (Eds.), Soft Computing Techniques Applied to Catalytic Reactions, Current Topics in Artificial Intelligence, vol. 3040, Springer, Heidelberg, 2004, p. 536.

[31] J.M. Serra, A. Corma, S. Valero, E. Argente, V. Botti, QSAR & Comb. Sci., in press.

[32] J.N. Cawse (Ed.), Experimental Design for Combinatorial and High Throughput Materials Development, Wiley, New York, 2003.

[33] A. Corma, J.M. Serra, P. Serna, E. Argente, S. Valero, V. Botti, J. Catal. 229 (2005) 513.

[34] N. Kettaneha, A. Berglundb, S. Wold, Comput. Stat. Data Anal. 48 (2005) 69.

[35] A.K. Jain, M.N. Murty, P.J. Flynn, ACM Comput. Surv. 31 (1999) 264.

[36] L. Chen, J. Gasteiger, J. Am. Chem. Soc. 119 (1997) 4033.

[37] J. Gasteiger, J. Zupan, Angew. Chem. Intl. Ed. Engl. 32 (1993) 503.

[38] B. Hammera, A. Michelib, A. Sperdutic, M. Strickert, Neural Networks 17 (2004) 1061.

[39] A. Bröcker, G. Schneider, A. Teckentrup, QSAR & Comb. Sci. 23 (2004) 207.

[40] D.C. Weaver, Curr. Opin. Chem. Biol. 8 (2004) 264.

[41] T. Mitchell (Ed.), Machine Learning, McGraw–Hill, 1997, p. 52.

[42] D. Hosmer, L. Stanley, Applied Logistic Regression, Wiley, 1989.

[43] J.M. Serra, A. Corma, A. Chica, E. Argente, V. Botti, Catal. Today 81 (2003) 393.

[44] H.-X. Liu, R.-S. Zhang, X.-J. Yao, M.-C. Liu, Z.-D. Hu, B.-T. Fan, Anal. Chim. Acta 525 (2004) 31.

[45] J.T.G. Hwang, D. Nettleton, Technometrics 45 (2003) 70.