

June 2009

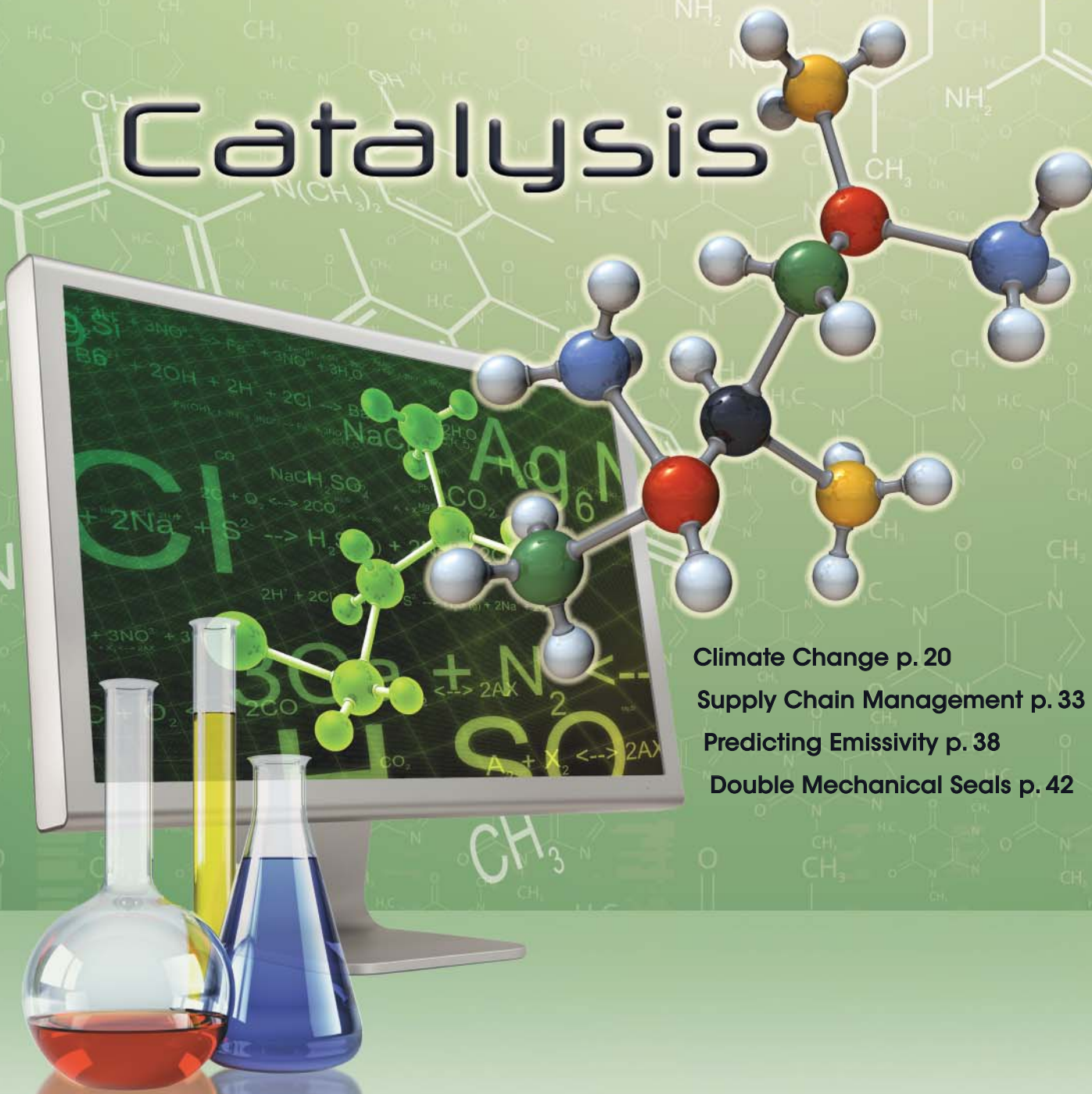
**Chemical
Engineering
Progress**

An AIChE Publication

CEP

www.aiche.org/cep

Catalysis



Climate Change p. 20

Supply Chain Management p. 33

Predicting Emissivity p. 38

Double Mechanical Seals p. 42



On the Horizon

Predictive Modeling in Catalysis — From Dream to Reality

ANA G. MALDONADO
GADI ROTHENBERG
UNIV. OF AMSTERDAM

In silico catalyst optimization is the ultimate application of computers in catalysis. This article provides an overview of the basic concepts of predictive modeling and describes how this technique can be used in catalyst and reaction design.

Imagine a computer program that is fed data for a particular reaction and then outputs the structure of the optimal catalyst for that reaction. Although such a program does not yet exist, much progress has been made in this direction in the last decade, due in large part to advances in laboratory automation and data analysis methods.

Just like any other new idea, using computers to design and optimize catalysts *in silico* is accepted by some researchers and met with skepticism by others. Nevertheless, it is essential for realizing the potential of high-throughput screening and combinatorial chemistry in catalysis research. Computers will not replace chemists, and data mining methods will not replace mechanistic studies — but these techniques will be essential tools for the 21st century chemist.

This article provides a brief overview of predictive modeling in catalysis. It is directed at engineers and chemists who are working on or are interested in catalysts and catalytic processes, and it focuses on the basic concepts and the possibilities of predictive modeling, rather than on specific mathematical techniques. A more detailed overview is published in Ref. 1.

Predictive modeling takes available catalyst and reaction data and uses a computer model to predict the outcome for catalysts and reactions that are unavailable or have not yet been tested. Process engineers who are familiar with response-surface methodology will recognize the similarities.

Still, predictive modeling in catalysis has some challenges. The key to success is developing a simple yet

realistic model that combines the chemistry with the process. Until recently, this capability was out of reach for everyday computers. Technological advances have made today's computers much cheaper and faster. Moreover, experimental validation of such models is now easier with high-throughput experimentation (HTE), which gives access to large amounts of reproducible data.

Computer models are similar to laboratory experiments in several ways. Unless they are trivially simple, you cannot predict their results by looking at the code. If they are planned badly, or if they are programmed badly, they may either crash or yield meaningless numbers. Like experiments, computer modeling is hardware-dependent, and yet, as in experiments, surprisingly nice results can sometimes be obtained using simple equipment. Importantly, computer models offer us no understanding, only numbers. We must examine the meaning and the significance of these numbers, always also considering the statistical errors, such as noise in measurements and sampling influence, involved. (For a detailed discussion on filtering out worthless data, see Refs. 2 and 3.)

Unfortunately, too many scientists tend to accept the results of “successful” computer models at face value. Just because a program did not crash does not mean that the results are meaningful. Understanding each model's limitations and setting realistic expectations will increase the chances of finding good catalysts.

A structured approach to predictive modeling includes four steps (Figure 1):

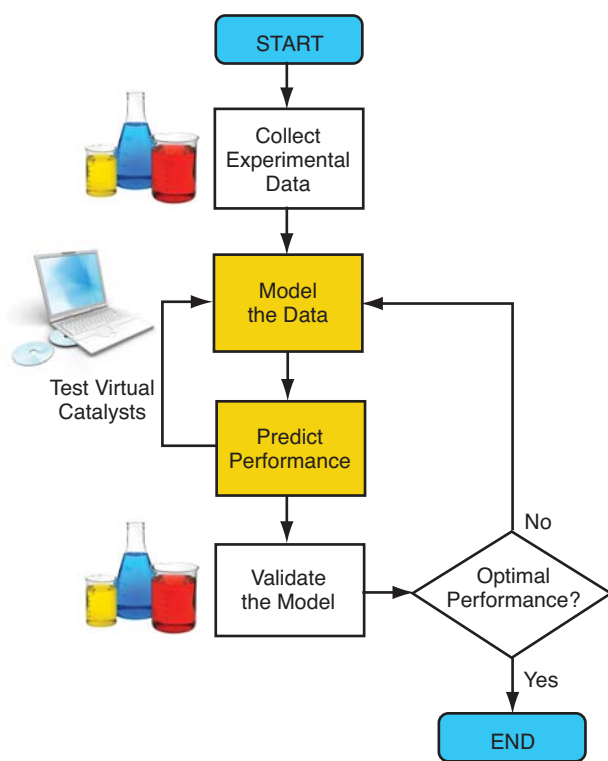
1. collecting experimental data
2. modeling these data
3. predicting the performance for new catalysts/reactions
4. validating (part of) these predictions experimentally.

The advantage of such an approach is that the dataset can be altered *in silico* to create and model virtual libraries of catalysts and reaction conditions. This increases the scope of the search space and thereby the chance of finding a better, cheaper, or more eco-friendly catalyst or process.

As a rule of thumb, predictive modeling requires experimental data that cover at least 5–15% of the search space. For example, synthesis and testing of 100 catalysts can permit the modeling of 1,500–5,000 catalysts or reaction conditions, by “mixing” the data on the computer. In this way, the model can point you in the direction of “good” experiments or, equally importantly, direct you away from potentially “bad” ones.

Catalysts, descriptors, and figures of merit

Before making any predictions, we must understand the problem of catalyst optimization, and define the space in which we want to optimize our data. To do this, the catalyst system is redefined in three multidimensional spaces (Figure 2).



▲ **Figure 1.** Predictive modeling involves a series of experimental and computational steps.

Space A is a grid containing all the catalyst structures — *e.g.*, if the catalyst in question is a bidentate transition-metal complex, Space A will contain all of the combinations of transition-metal atoms and bidentate ligands, with each point representing a different catalyst.

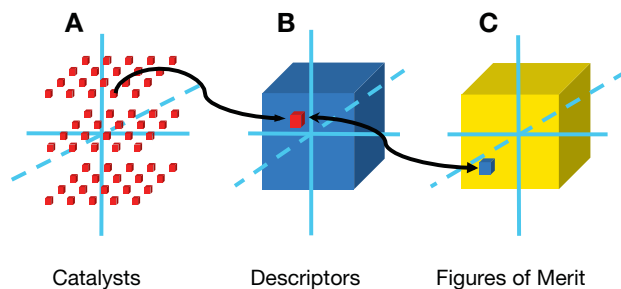
Space B, the descriptor space, contains the values of the catalysts’ descriptors (internal parameters such as backbone flexibility, partial charge on the metal atom, polarity, lipophilicity, surface area, or crystallite size) as well as the reaction conditions (external parameters such as temperature, pressure, and solvent type). All of these parameters may influence the reaction outcome, and a model may include both internal and external descriptors.

Space C is the space of the figures of merit (FOM), *e.g.*, the turnover number (TON), turnover frequency (TOF), product selectivity, price, and so forth. Note that while Space A is a discrete grid, Spaces B and C are continuous and are arranged such that each dimension represents one property.

Redefining the system thus translates the abstract catalysis problem to the (still abstract) problem of relating one multidimensional space to another. The advantage is that now the relationship between Spaces B and C can be quantified using quantitative structure-activity relationship (QSAR) and quantitative structure-property relationship (QSPR) models (4, 5).

The key to the model lies in finding the right descriptors for Space B. This is relatively easy in homogeneous catalysis, where the catalyst is usually a well-defined molecule or organometallic complex (6). There are several levels of complexity for descriptors.

Three-dimensional (3D) descriptors, which are based on optimized geometries, can be derived from molecular-mechanics force-field and quantum-mechanics calculations. They offer a realistic representation of chemical systems. However, they are computationally expensive, with the cost depending on the system’s size and number of degrees of freedom. When optimizing large numbers of catalysts (*e.g.*, large virtual libraries for combinatorial

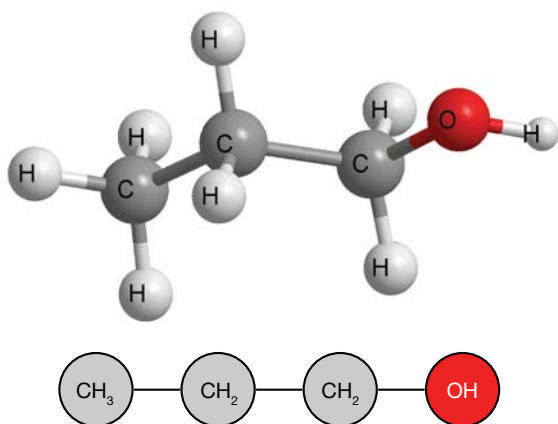


▲ **Figure 2.** A catalyst system is represented by three multidimensional spaces containing the catalysts (Space A), descriptor values (Space B), and figures of merit (Space C).

optimization studies), 3D descriptors are simply too costly.

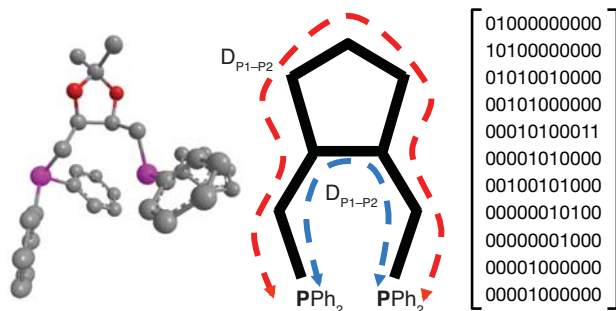
In such cases, the simpler 2D descriptors (also called topological descriptors) provide a viable alternative. Topological descriptors are derived directly from molecular connectivity tables, without using any 3D atom coordinates. They are easily calculated using graph theory (7, 8). Figures 3 and 4 show two approaches to reducing the amount of information for a given molecule, thereby creating a simplified structure and reducing computational costs.

Topological descriptors give information on the molecular size, flexibility, electron distribution, and various other physicochemical properties. Their calculation is typically three to five orders of magnitude faster than 3D descriptors, depending on the geometry optimization method used for the latter (Figure 5) (8). Unfortunately, this lower cost is offset by several limitations. First, although 2D descriptors account for specific physicochemical properties, they have no direct



▲ **Figure 3.** A 3D atomistic model of 1-propanol (top) can be simplified to a coarse-grained bead model, in which each group of atoms is represented by one bead (bottom). In the latter, all beads typically have the same size, which simplifies the computations required.

▼ **Figure 4.** Kagan's DIOP ligand, (-)-2,2-dimethyl-4,5-bis(diphenylphosphinomethyl)-1,3-dioxolane-PP', can be represented as a 3D molecular model (left), which is more accurate but requires expensive geometry optimization. The 2D molecular graph (center) is much simpler, requiring only counting of the atoms, and can be easily translated into the so-called adjacency matrix (right).

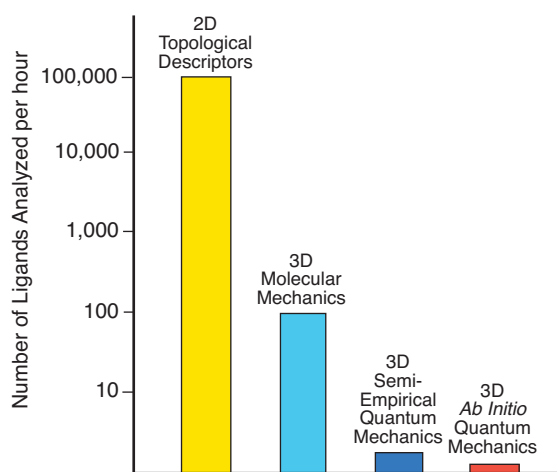


heuristic interpretation, because they are far from our chemical intuition. Second, 2D descriptors neglect conformational information, and because they are two-dimensional, they cannot be used for modeling chirality.

Finding good descriptors for heterogeneous catalysts is much more difficult. Unlike molecular catalysts and organometallic complexes, the activity of solid catalysts depends on a multitude of parameters: different types of active sites, synthesis conditions, thermal treatments, and aging. Moreover, the properties of many solids can change discontinuously. New phases may form at different compositions, temperatures and pressures, and even the catalyst particle size can influence the reaction. Nano-sized gold particles, for example, are very different catalysts from bulk gold, and two supported gold catalysts can have different activities even if they contain identical amounts of gold and support (9). Solid surfaces are anything but uniform, and solid catalysts have a variety of sites. To complicate things further, sometimes the real active sites are not those observed in characterization studies, but rather metastable defects that are difficult to characterize.

Because of this, relying solely on catalyst composition parameters is impractical (compositional descriptors are applicable when the catalyst is a crystalline material, and where changes in composition do not cause phase changes; see, for example, Ref. 10). Instead, a descriptor toolbox that can account for the discontinuities and nonlinear dependencies is needed. A few promising starts have been made in this direction (11–14).

Klanner, *et al.*, combined compositional descriptors and tabulated physicochemical data, collecting more than



Note: All calculations were performed on a desktop PC with a single 2.5-GHz processor. The analytical capacity will undoubtedly improve over time with better computers and software, but the 2D:3D ratio will remain essentially the same.

▲ **Figure 5.** The number of bidentate ligands that can be analyzed per hour depends on the type of descriptors and how they were determined.



3,000 descriptors for 467 catalysts, which were then tested in a high-throughput reactor in propene autoxidation. Such a dataset is over-determined — *i.e.*, there are many more descriptors than data points, so that one can always find good correlations, but these correlations are often meaningless. To obtain meaningful correlations, the researchers focused on a subset of 75 descriptors that they deemed to be chemically relevant. Prediction models were then constructed using artificial neural networks (ANNs) and classification trees (15). Significantly, both methods could predict good or bad propene oxidation catalysts. The prediction rate of the ANNs was typically 0.5–0.7, much higher than that of random models (typically 0.2–0.3).

Artyushkova and coworkers used predictive modeling to correlate the structure and electrochemical performance of various non-platinum (and thus cheaper) porphyrin electrocatalysts for oxygen reduction, with the aim of increasing catalytic activity (16). Other researchers combined genetic algorithms (GAs) and ANNs for predicting the performance of virtual catalyst libraries for oxidative dehydrogenation of ethane, using the catalyst composition as input parameters (17, 18). The virtual screening was again combined with high-throughput experimentation, using the predictions of the ANNs as a theoretical prescreening to avoid the testing of poorly performing materials. Although the subsets tested were very small compared to the catalyst space, a significant improvement was obtained after seven generations.

Predictive modeling also plays an increasingly important role in the search for new biocatalysts, despite the fact that enzymes are far too complex for detailed descriptor models at the molecular level (19). The experimental techniques for designing enzymes that can operate under harsh process conditions (high temperatures, acidic/basic pH, and/or organic solvents) rely heavily on genetic engineering and combinatorial chemistry. These efforts are complemented by a variety of computational screening tools (20). The main composition variable is the primary structure (*i.e.*, the amino acid sequence of the protein). Computer algorithms screen the sequence space, eliminating those sequences that are incompatible with the protein folding model and thereby reducing the number of useless experiments (21). Protein modeling algorithms can even insert potentially active catalytic residues into virtual proteins and search for candidates with an improved binding affinity to high-energy reaction intermediates (22, 23). Examples include the redesigning of active sites for improved catalytic activity (24) and the designing of thermostable enzymes (25).

Building and testing virtual catalyst libraries

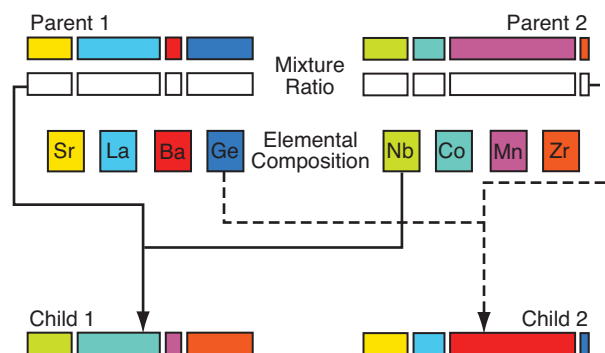
The next step after data collection is to choose the modeling approach. Several options, all based on well-known statistical methods, are available. To those who are not statis-

ticians, the many acronyms may look daunting, but all these methods can be applied by chemists using commercially available user-friendly software interfaces. One option is classification models, which can discriminate between good and bad catalysts. Typical methods that apply well to classification models are classification trees (26, 27), supported vector machines (SVMs) (28, 29), and ANNs (30, 31).

Another option is linear or nonlinear regression models, such as multiple linear regression (MLR) (32) or partial least-squares (PLS) regression (33). These models connect the structural data about the catalysts to their figure of merit via QSAR/QSPR. The results of the model indicate how important each descriptor is, as well as how much of the variance in the data the model predicts. Typically, a small number of descriptors should explain a large portion of the variance in the data. These important descriptors can be discerned using principal component analysis (PCA) (34) and variable importance (VIP) (35) techniques.

Next, new catalyst sets are created *in silico* and the model is used to predict their figures of merit. These new sets are called virtual catalyst libraries, as they have not been tested experimentally. Indeed, most of them will never be prepared and tested. Nevertheless, using such virtual libraries of catalysts and reaction conditions allows a larger part of the search space to be explored. This is essential, even for laboratories with high-throughput experimentation equipment, because the search space is always much larger than the lab's experimental capacity. In fact, HTE labs can benefit most from modeling virtual libraries, as their experiments are typically more reproducible and model validation is easier.

There are several approaches for constructing such virtual libraries. With solid catalysts, one can mix catalyst precursors virtually or simulate the result of sputtering and masking experiments and obtain a good analogy with the actual experiments (36, 37). When working with well-



▲ **Figure 6.** A simplified “evolutionary step” for crystalline mixed-perovskite oxides (which have the general formula $AA'BB'O_3$) illustrates how in each new generation, a “child perovskite” can inherit from each “parent perovskite” either its elemental composition (represented by the colors of the rectangles) or its elemental mixture ratio (the sizes of the rectangles).

defined crystalline materials, such as mixed oxides, genetic algorithms can be used to evolve new generations of catalysts *in silico*.

Figure 6 shows such a case for mixed perovskite oxides. These have the general formula $AA'BB'O_3$, where A and A' are the main large cation and the dopant large cation, and B and B' are the main small cation and the dopant small cation, respectively. A simple yet effective description of these materials uses two descriptors: the elemental composition and the relative concentration or mixture ratio. Here, each pair of "parent perovskites" can "mate," resulting in two "child perovskites." Each child inherits from each of its parents either the elemental composition or the mixture ratio.

Model validation: separating knowledge from garbage

As noted previously, one lamentable problem with computer models — or with the scientists using them — is that when the model does not crash, the scientist tends to believe the result. Without proper validation, however, deducing anything from any model is risky business. The model may be over-fitting (finding trends in noise), or predictions may be out of range (extrapolation), either of which can lead to ridiculous results. Model validation is like a control experiment in the laboratory. It may be tedious and time-consuming, but it is essential.

Two useful validation techniques are cross-validation and *y*-randomizing, which are described here. A detailed treatment of model validation is published elsewhere (38).

Regardless of the model used, the validation process will depend on the amount and quality of the data. When there are enough data, they should be divided into a training set, a test set, and a validation set. The model is constructed using the training set, tested with the test set, and perhaps improved and retested. When you are satisfied with the model, you can test its performance on the validation set. After this, do not tinker with the model again, because it has already been exposed to the validation set. If there are not enough data for three subsets, divide the data into a training set and a test set, construct the model using the training set, validate it using cross-validation, and then test its performance with the test set.

Cross-validation. Assume that we have measured the TON and TOF of 200 catalytic reactions and developed a QSAR/QSPR regression model that relates the catalyst descriptors to the figures of merit. To validate the model, the dataset is divided into a training set (also known as the calibration set), which is used to develop the model, and a test set (also known as the prediction set) (33). The figures of merit for the test set are known, but they are not used to generate the model. Instead, the regression equation for the

training set (say for $n = 150$ reactions) is calculated and used to predict the TON and the TOF for the test set (the remaining 50 reactions). In this way, the performance of different models can be compared because they are all trained with the same training set. This is known as cross-validation.

The model's predictive performance is measured by q^2 , the cross-validation squared correlation coefficient:

$$q^2 = 1 - \frac{\sum_{i=1}^n (y_i - y_{i,p})^2}{\sum_{i=1}^n (y_i - y_{i,m})^2} \quad (1)$$

where y_i , $y_{i,p}$, and $y_{i,m}$ are the measured, predicted, and mean values of y , respectively.

Note, however, that although a high q^2 value is a necessary condition for good predictions, it is not a sufficient condition (39). Cross-validation can be complemented by other methods, *e.g.* bootstrapping (40), to ensure the model's robustness and prediction accuracy.

The choice of training and test sets may bias the results (41). To avoid problems, the original set can be partitioned in several different ways (*i.e.*, choosing different combinations of training sets and test sets) and an average score over the different partitions computed. This is known as leave- n -out cross-validation.

For example, results for 100 experiments can be divided into a training set of 80 and a test set of 20. This process can be repeated several times with different 80:20 combinations ($n = 20$). An extreme variant of this is splitting the 200 reactions into a training set of 199 and a test set of 1, and repeating the validation 200 times, each time with a different test experiment. This is called leave-*one*-out cross-validation.

The advantage of leave- n -out and leave-*one*-out cross-validation is that all the data are used for training.

Mixing the dependent variables, or *y*-randomizing. Randomizing of the y variables (also known as a permutation test) is a common method for testing a model's robustness. The vector containing the figures of merit is shuffled randomly so that the figures of merit are no longer matched to the original descriptor values. A new model is then generated using the original descriptor values, and the process is repeated several times. In principle, models developed in this way should have very low correlation coefficients (both R^2 , which predicts the results that have already been tested, and q^2 , which predicts new, untested results). Models that fail this negative test should be discarded, because any random collection of values for the figures of merit would do just as well.

A simpler, yet effective, variant of this method is testing the model on completely random data. Generate a random series of numbers for the figure of merit (y) and run



LITERATURE CITED

1. **Rothenberg, G.**, "Catalysis: Concepts and Green Applications," Wiley-VCH, Weinheim, Germany (2008).
2. **Massart, D. L., et al.**, "Handbook of Chemometrics and Qualimetrics," Elsevier, Amsterdam, The Netherlands (1998).
3. **Tranter, R. L.**, "Design and Analysis in Chemical Research," CRC Press, Sheffield, U.K. (2000).
4. **Bönnemann, H.**, "Organocobalt Compounds in Pyridine Syntheses — An Example for Structure-Activity Relations in Homogeneous Catalysis," *Angew. Chem. Int. Ed. Engl.*, **24**, pp. 248–262 (1985).
5. **Cooney, K. D., et al.**, "A Priori Assessment of the Stereoelectronic Profile of Phosphines and Phosphites," *J. Am. Chem. Soc.*, **125**, pp. 4318–4324 (2003).
6. **Burello, E., and G. Rothenberg**, "In Silico Design in Homogeneous Catalysis Using Descriptor Modeling," *Int. J. Mol. Sci.*, **7**, pp. 375–404 (2006).
7. **Diestel, R.**, "Graph Theory," Springer Verlag, New York, NY (2000).
8. **Burello, E., and G. Rothenberg**, "Topological Mapping of Bidentate Ligands: A Fast Approach for Screening Homogeneous Catalysts," *Adv. Synth. Catal.*, **347**, pp. 1969–1977 (2005).
9. **Haruta, M., et al.**, "Gold Catalysts Prepared by Coprecipitation for Low-Temperature Oxidation of Hydrogen and of Carbon Monoxide," *J. Catal.*, **115**, pp. 301–309 (1989).
10. **Beckers, J., et al.**, "Clean Diesel Power Via Microwave Susceptible Oxidation Catalysts," *ChemPhysChem*, **7**, pp. 747–755 (2006).
11. **Klanner, C., et al.**, "How to Design Diverse Libraries of Solid Catalysts," *QSAR Comb. Sci.*, **22**, pp. 729–736 (2003).
12. **Farrusseng, D., et al.**, "Design of Discovery Libraries for Solids Based on QSAR Models," *QSAR Comb. Sci.*, **24**, pp. 78–93 (2005).
13. **Holena, M., and M. Baerns**, "Feedforward Neural Networks in Catalysis. A Tool for the Approximation of the Dependency of Yield on Catalyst Composition, and for Knowledge Extraction," *Catalysis Today*, **81**, pp. 485–494 (2003).
14. **Baumes, L., et al.**, "Using Artificial Neural Networks to Boost High-Throughput Discovery in Heterogeneous Catalysis," *QSAR Comb. Sci.*, **23**, pp. 767–778 (2004).
15. **Klanner, C., et al.**, "The Development of Descriptors for Solids: Teaching "Catalytic Intuition" To a Computer," *Angew. Chem. Int. Ed.*, **43**, pp. 5347–5349 (2004).
16. **Artyushkova, K., et al.**, "Predictive Modeling of Electrocatalyst Structure Based on Structure-to-Property Correlations of X-Ray Photoelectron Spectroscopic and Electrochemical Measurements," *Langmuir*, **24**, pp. 9082–9088 (2008).
17. **Corma, A., et al.**, "Application of Artificial Neural Networks to Combinatorial Catalysis: Modeling and Predicting ODHE Catalysts," *ChemPhysChem*, **3**, pp. 939–945 (2002).
18. **Serra, J. M., et al.**, "Soft Computing Techniques Applied to Combinatorial Catalysis: A New Approach for the Discovery and Optimization of Catalytic Materials," *QSAR Comb. Sci.*, **26**, pp. 11–26 (2007).
19. **Arnold, F. H.**, "Combinatorial and Computational Challenges for Biocatalyst Design," *Nature*, **409**, pp. 253–257 (2001).
20. **Edward, G. H., and A. D. Paul**, "Directed Evolution Strategies for Improved Enzymatic Performance," *Microb. Cell Fact.*, **4**, p. 29 (2005).
21. **Hayes, R. J., et al.**, "Combining Computational and Experimental Screening for Rapid Optimization of Protein Properties," *Proc. Natl. Acad. Sci. USA*, **99**, pp. 15926–15931 (2002).
22. **Bolon, D. N., and S. L. Mayo**, "Enzyme-Like Proteins by Computational Design," *Proc. Natl. Acad. Sci. USA*, **98**, pp. 14274–14279 (2001).
23. **Dwyer, M. A., et al.**, "Computational Design of a Biologically Active Enzyme," *Science*, **304**, pp. 1967–1971 (2004).
24. **Lassila, J. K., et al.**, "Computationally Designed Variants of Escherichia Coli Chorismate Mutase Show Altered Catalytic Activity," *PEDS*, **18**, pp. 161–163 (2005).
25. **Korkegian, A., et al.**, "Computational Thermostabilization of an Enzyme," *Science*, **308**, pp. 857–860 (2005).
26. **De'ath, G., and K. E. Fabricius**, "Classification and Regression Trees: A Powerful yet Simple Technique for Ecological Data Analysis," *Ecology*, **81**, pp. 3178–3192 (2000).
27. **Buntine, W.**, "Learning Classification Trees," *Stat. Comp.*, **2**, pp. 63–73 (1992).
28. **Smola, A. J., and B. Schölkopf**, "A Tutorial on Support Vector Regression," *Stat. Comp.*, **14**, pp. 199–222 (2004).
29. **Schölkopf, B., and A. J. Smola**, "Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond," MIT Press, Cambridge, MA (2002).
30. **Serra, J. M., et al.**, "Can Artificial Neural Networks Help the Experimentation in Catalysis?," *Catalysis Today*, **81**, pp. 393–403 (2003).
31. **Basheer, I. A., and M. Hajmeer**, "Artificial Neural Networks: Fundamentals, Computing, Design, and Application," *J. Microbiol. Meth.*, **43**, pp. 3–31 (2000).
32. **Montgomery, D. C., et al.**, "Introduction to Linear Regression Analysis," Wiley, Hoboken, NJ (2001).
33. **Geladi, P., and B. R. Kowalski**, "Partial Least-Squares Regression: A Tutorial," *Anal. Chim. Acta*, **185**, pp. 1–17 (1986).
34. **Wold, S., et al.**, "Principal Component Analysis," *Chemom. Intell. Lab. Syst.*, **2**, pp. 37–52 (1987).
35. **Burello, E., et al.**, "Ligand Descriptor Analysis in Nickel-Catalysed Hydrocyanation: A Combined Experimental and Theoretical Study," *Adv. Synth. Catal.*, **347**, pp. 803–810 (2005).
36. **Beckers, J., et al.**, "Selective Hydrogen Oxidation Catalysts Via Genetic Algorithms," *Adv. Synth. Catal.*, **350**, pp. 2237–2249 (2008).
37. **Maier, W. R., et al.**, "Combinatorial and High-Throughput Materials Science," *Angew. Chem. Int. Ed.*, **46**, pp. 6016–6067 (2007).
38. **Tropsha, A., et al.**, "The Importance of Being Earnest: Validation Is the Absolute Essential for Successful Application and Interpretation of QSPR Models," *QSAR Comb. Sci.*, **22**, pp. 69–77 (2003).
39. **Golbraikh, A., and A. Tropsha**, "Beware of q^2 !," *J. Molec. Graph. Model.*, **20**, pp. 269–276 (2002).
40. **Wehrens, R., et al.**, "The Bootstrap: A Tutorial," *Chemom. Intell. Lab. Syst.*, **54**, pp. 35–52 (2000).
41. **Golbraikh, A., et al.**, "Rational Selection of Training and Test Sets for the Development of Validated QSAR Models," *J. Computer-Aided Molec. Design*, **17**, pp. 241–253 (2003).
42. **Hageman, J. A., et al.**, "Design and Assembly of Virtual Homogeneous Catalyst Libraries — Towards in Silico Catalyst Optimization," *Adv. Synth. Catal.*, **348**, pp. 361–369 (2006).
43. **Maldonado, A. G., et al.**, "Backbone Diversity Analysis in Catalyst Design," *Adv. Synth. Catal.*, **351**, pp. 387–396 (2009).

On the Horizon

the model. You should get only noise. If you get meaningful results (*i.e.*, high R^2 and q^2 values) using random data, then there is something seriously wrong with the model.

Looking ahead

Predictive modeling in catalysis is here to stay. Convincing traditional chemists and engineers of its usefulness may take another decade, but the ultimate driver for using it will be, simply, the overall economic benefits. Case studies that show the advantages of combining predictive modeling with catalyst synthesis and testing, both in academia (42) and in industry (43), are accumulating fast. And today's computer hardware and software put us on the brink of true *in silico* catalyst design.

Nevertheless, keep in mind that:

- Predictive modeling will not solve all problems.
- It's a research tool that works well especially when complemented by high-throughput experiments, good thinking, and good planning.
- Garbage in, garbage out — start with reliable data.
- Data in, garbage out — proper model validation

is essential for getting meaningful results.



ANA G. MALDONADO is a researcher in the Van 't Hoff Institute for Molecular Sciences at the Univ. of Amsterdam, where she is working on a joint project with Rhodia Chemicals, applying predictive modeling and rational design for finding new hydrocyanation catalysts. She has worked extensively with chemoinformatic and chemometrics tools in industry and academia over the past seven years. Her research interests include molecular diversity, QSAR/QSPR predictive models, and data mining. She is the architect of the MoDiA software, an XML-based high throughput screening tool for chemical databases. She earned a BS in chemistry and an MS in theoretical chemistry at the Univ. Simon Bolivar (Caracas, Venezuela) and a PhD in computational and theoretical chemistry from the Univ. Denis Diderot (Paris, France).

GADI ROTHENBERG is a professor and Chair of Heterogeneous Catalysis and Sustainable Chemistry at the Van 't Hoff Institute for Molecular Sciences at the Univ. of Amsterdam (Nieuwe Achtergracht 166, 1018 WV Amsterdam, The Netherlands; Fax: +31 20 525 5604; E-mail: g.rothenberg@uva.nl; Website: www.science.uva.nl/~gadi), where he teaches courses on catalysis, thermodynamics, and scientific writing. He has published more than 100 papers in peer-reviewed journals and discovered two catalysts, for which he received the Marie Curie Excellence Award in 2004 and the Paul N. Rylander Award in 2006. He also invented a method for monitoring pollutants in water, and co-founded the companies Sorbisense and Yellow Diesel. In 2007, he was voted Teacher of the Year by the chemistry students, and his book "Catalysis: Concepts and Green Applications," was published by Wiley-VCH in 2008. He holds a BSc in chemistry, and an MSc and a PhD in applied chemistry, all from the Hebrew Univ. of Jerusalem.