DOI: 10.1002/anie.201208487

The Catalyst Genome

Jens K. Nørskov and Thomas Bligaard**

Jens K. Nørskov
Professor, Stanford University
and SLAC National
Accelerator Laboratory

Thomas Bligaard
Senior Staff Scientist
SLAC National
Accelerator Laboratory

The quest for the materials genome—the properties of a material that define its functional properties—has started. This signifies a transition to a new era of materials research where large amounts of materials data become available. The expectation is that this will significantly speed up the discovery of new materials. This is particularly true in the area of catalytic materials, where there is an urgent need for new catalysts and processes to enable the sustainable production of fuels and chemicals.

Catalysis is used everywhere to control chemical processes: In nature, enzymes regulate the chemical factories in cells, and catalyst are the cornerstone of the chemical industry today. Essentially all chemicals, including fuels, have seen one and often many catalysts during production. Most catalysts used in large-scale processes in industry are solids and the catalysis takes place at the surface of the catalyst material, but there are also areas where the superior selectivity of low-temperature processes makes molecular catalysts preferable.

What is it that makes one material a good catalyst and another a bad one? The quest for the catalyst genome has been going on for over 100 years. Progress is very rapid these days, and the demands for new approaches to catalyst design are daunting. Not only would more efficient catalysts reduce energy consumption and waste products in today's chemical industry, they are also essential for building a completely new, sustainable chemical sector. Unless we want to continue relying on fossil resources, we need to find new, sustainable routes to make both fuels and chemicals. Most scenarios rely on energy input from the sun, and new catalytic processes are of central importance for making this possible (Figure 1).

What would the catalyst genome look like? It could be a map linking all possible catalyst structures to rates of all possible elementary reactions at all possible reaction conditions coupled with electronic structure and spectroscopic data characterizing the different intermediates. Imagine that we had all this data and efficient methods to mine them. It would then be possible to construct catalysts for any given catalytic reaction by first considering all possible reaction paths and then search for the material that would best catalyze the selected process.

We are very far from this dream scenario. The amount of data would be enormous and the experimental work needed to obtain the data would be unfeasible. A major step forward in this direction has been the development of electronic-structure methods to calculate catalyst properties. Computational methods lend themselves very well to providing systematic data for a large body of materials and reactions. An essential part of this process is to benchmark the calculations (often at the density functional theory level) against detailed experiments or more accurate calculations for selected systems in order to establish limits of reliability. The "CatApp" is one such example of a database of reaction and activation energies for elementary processes on transition metal surfaces: http://suncat.slac.stanford.edu/catapp/ and *Angew. Chem. Int. Ed.* **2012**, *51*, 272–274.

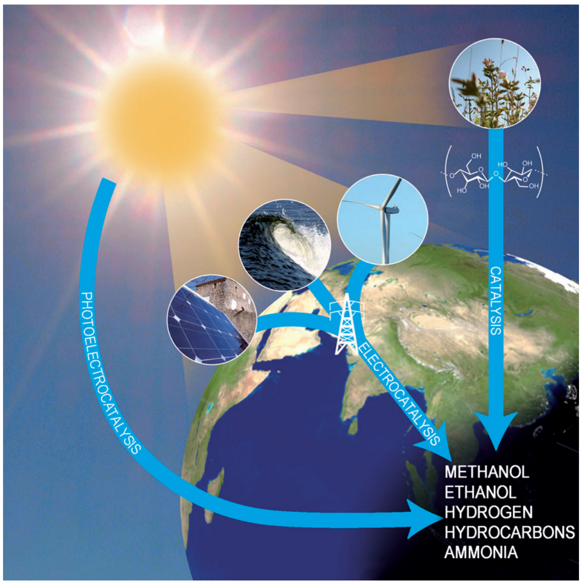

Figure 1. Different ways of arriving at fuels and base chemicals from sunlight. Whether the energy flux from sunlight is harvested through biomass, through intermediate electricity production from for example, photovoltaic cells or wind turbines, or directly through photoelectrochemical reactions, the process always requires an efficient catalyst preferably made of earth-abundant materials.

[*] Prof. J. K. Nørskov, Dr. T. Bligaard
SUNCAT Center for Interface Science
and Catalysis
SLAC National Accelerator Laboratory
and
Department of Chemical Engineering
Stanford University
Stanford (USA)
E-mail: norskov@stanford.edu
bligaard@stanford.edu

There are large areas (transition-metal oxides, for instance) where the conventional electronic structure methods are not accurate enough and where considerably more work needs to go into development of methods that are accurate and efficient enough to be useful in mapping out large amounts of data. There is also the question of dealing with the complexity of real catalysts. For molecular catalysts in solution, the challenges are associated with solvation effects, and for heterogeneous catalysts the nature of the active site and the existence of several solid phases need further understanding. It is clear that rather than describing each system in complete detail, it will be essential to extract the most important properties that determine the catalytic properties. Once such descriptors have been established it is possible to do searches in the much smaller space of parameters spanned by the descriptors.

How does one identify descriptors? One way is to invoke an understanding of the reactivity in terms of the underlying electronic structure of the catalyst. Another is to look for correlations in existing data. The catalyst activity or selectivity that is measured in costly, carefully designed experiments can for example be correlated against a calculated descriptor. The easily accessible calculated descriptor can thus be used to aim new experiments at compounds that the descriptor suggests have good catalytic properties. Likewise, more expensive calculations can sometimes be correlated with much less expensive calculations, thus defining less accurate but more economically obtained descriptors for advanced simulations. This could, for example, be more advanced electronic structure calculations versus a more basic level of theory, more complex sampled free energy calculations versus ground state energy ones, or simply more difficultly obtainable reaction barrier heights versus adsorption energies. If the correlations are accurate enough, such simplifications can increase the number of systems that can be computationally addressed by many orders of magnitudes, thus providing a boost to the data richness and allowing a search to be carried out. Continuously improving theoretical methods have also resulted in the accuracy of the calculated catalyst properties increasing over time. As the analyzed catalytic processes become more complex, and the amount of computationally obtained data grows rapidly, it appears likely that machine-learning approaches will start playing a more prominent role in shaping the catalyst genome. Machine learning is loosely defined a set of approaches which attempt to determine (less obvious) patterns in large amounts of data or to make (reasonable) predictions based on incomplete data sets.

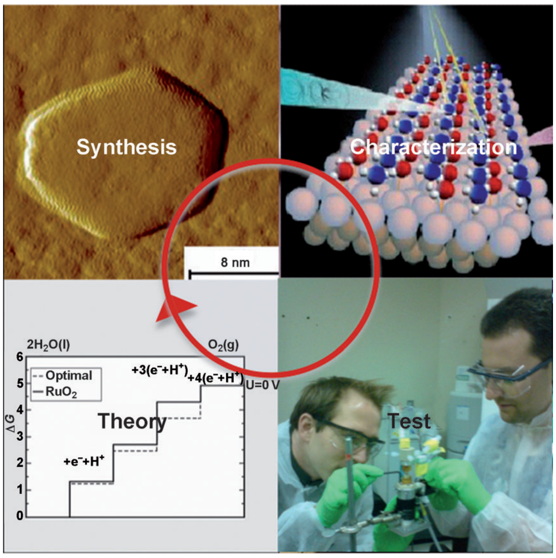

Figure 2. Illustration of an integrated catalyst genomics approach to the discovery of new catalytic materials. Experimental synthesis, characterization, and testing are rapidly becoming interwoven with theoretical calculations. The interplay not only pushes theory towards higher accuracy and allows validation and evaluation of calculational uncertainties, it also works the other way, with theory pushing the experiments toward the limits of atomic-scale control.

Since all the approaches aimed at accelerating catalyst discovery are centered around the availability of large amounts of data, the catalyst genome is likely in the near future to reveal its form primarily as a database of calculated properties augmented with key experimental data for benchmarking and for establishing correlations. The catalyst genome, however, should be considered as much more than just the underlying data. Since the knowledge relevant for catalysis is based on understanding the correlations and interrelationships between catalytically relevant data sets that quite often have still to be discovered, the catalyst genome is also a collection of relevant concepts, analysis tools, search methods, and learning algorithms to create data where none is yet present.

All the approaches aimed at accelerating the discovery are centered around harvesting integration, both between many research groups and between different methodologies. Thus, a central issue to address is the platform for interaction between research groups and between different methodologies (Figure 2). It will be essential for progress that the whole community is working together. It is likely that there will be several databases or platform technologies. The key to success is to make sure that they are formatted in a transparent way so that software can access and utilize all data available. Since no group is likely to have the total overview of or full vision for what the catalyst genome will develop into, it seems that an international scientific open forum should be established with the specific task of defining some "standards" for how catalysis data and method integration should occur. An important issue will be the treatment of private data. For the catalyst genome to become important as a vehicle for the discovery of novel technical catalysts, industry for example has to have access to the full "public genome" and be enabled to combine it seamlessly with their own specialized "proprietary genome" for specific processes or catalysts, without fear that their data become public.