UNIVERSITY OF AMSTERDAM

UvA-DARE (Digital Academic Repository)

Descriptors for solid catalysts: 21st century discovery tools

Ras, E.J.

Link to publication

6 Finalizing the workflow

6.1 Using descriptor models in catalyst discovery

Now that we have a toolbox that provides the means of designing experiments, modelling the outcome thereof and even generating novel descriptors to meet our needs, we need to consider how to apply these methods effectively and efficiently. We will demonstrate this on the selection of a subset of bimetallic catalysts from a large set of candidates. This subset selection is important because without a well-balanced design the subsequent generation of a model is difficult.

Figure 1 Selected candidate space for our selection problem. Entries in blue are selected as main metals and entries in purple are selected as dopants.

First, we designate a set of metals as "Main metal" (highlighted in blue in Figure 1). Second, we designate a set of metals as "Dopant" (pink, in Figure 1). These metals, or rather their bimetallic combinations, we characterize with the descriptors based on Slater-type orbitals (see Example on selective hydrogenation in Chapter 5). Please note that as many (or as few) descriptors can be used as are required by the problem at hand. If you know little about a problem, it is generally better to select more descriptors to start with. After performing a first round of experiments the redundant descriptors can be excluded based on data rather than on assumptions. Finally, we select a number of supports (SiO_2, Al_2O_3, TiO_2, ZrO_2, Nb_2O_5, MgO, ZnO). For each of these supports we select a high and a low surface area. Due to the natural difference in practically

accessible surface areas between different supports, the surface area is treated in a relative rather than an absolute manner. Besides surface area, the supports are characterized using their point of zero charge.

Dopants were applied in molar ratios of 0.05, 0.1 and 0.2 relative to the main metal. The first set of candidates consisted of all the combinations of one main metal and dopant at 3 levels, $13 \times 16 \times 3 = 624$ combinations. The second set was composed by combining two main metals at three dopant levels, giving $13 \times 12 \times 3 = 468$. The total candidate set contained therefore $624 + 468 = 1,092$ bimetallic combinations. Note that this assumes the use of a single support and a single loading of the main metal. If support variations are included, one could conceive using seven common supports, each with classified by its isoelectric point. Were we to include surface area of this support, for example in a "high" and "low" fashion, the number of supports available in the candidate set would increase to 14. This increases the size of our candidate set to $14 \times 1,092 = 15,288$. Here we assume that surface area, if important at all, will correlate with performance in a linear fashion (since we only use two levels, low and high, we only have enough degrees of freedom to explain the two coefficients corresponding to a straight line). For metal loading, we will not make this assumption. Instead, we will assume that the effect of metal loading is non-linear and we will use 3 levels. This increases the size of the candidate set to $3 \times 15,288 = 45,864$ catalysts!

Since we cannot synthesize and test over 45.000 catalysts, we must take a stepwise approach. Assuming we can describe the properties of the bimetallic combination by approximately 10 descriptors, we have ten metal parameters + two support parameters + one loading parameter = 13 variables that play a role. As we do not know a priori whether the relationship between variables and performance is linear, we will assume it is nonlinear.

Assuming a second order model, we need to identify an intercept, 13 main effects, $12 \times 11 = 132$ two-variable interactions and 13 quadratic terms. This total of 157 model coefficients is the minimum number of degrees of freedom we need to consider. Adding some replicates (or near neighbours) and some points to determine lack of fit raises this to a number around 200 catalysts that would need to be synthesized and tested. This is a number that is well within reach for most

chemistries using state of the art parallel reactor technology. Note that this initial design only comprises 0.4% of the original search space.
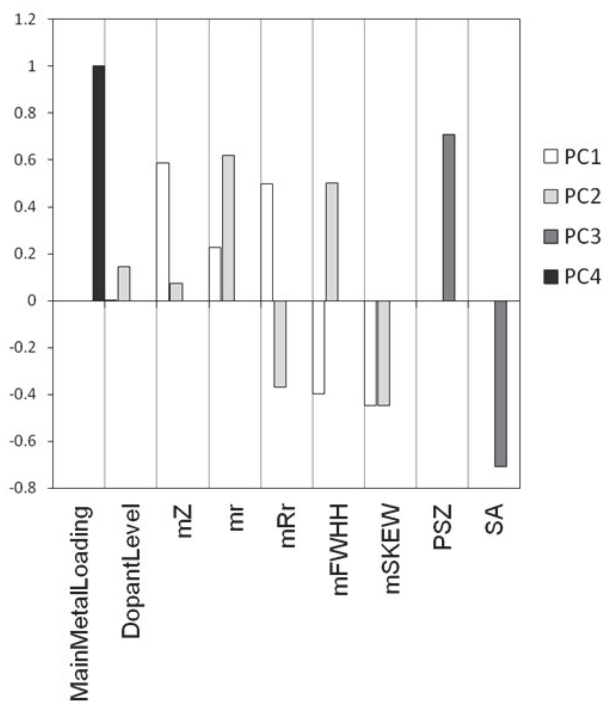

Figure 2 Loadings for the main effects (the raw variables) of our catalyst selection problem for the first four principal components of the PCA model. The vertical axis denotes the magnitude and direction of each effect, the horizontal axis denotes the variable identifiers (mZ = mean atomic number, mr = mean radius, mRr = mean apex value at radius, mFWHH = mean full with at half height of the RDF, mSKEW = mean skewness of RDF, PSZ = point of zero charge of support, SA = surface area of support)

To efficiently select candidates from this search space, we first reduce its dimensionality using PCA (see also the example on selective hydrogenation). In this case, over 98% of the variance in the set of candidates is captured by six principal components. Using experimental design (minimal point designs and distance based designs) we can select an optimal subset of catalysts for a first round of experiments. Note that machine-based selection methods are preferred over human intuition, to avoid any bias. Still, intuition not be ignored, so adding more candidates based on "gut feeling" is certainly recommended.

As we explained above, a PCA model is characterized by two matrices: the scores and the loadings. The loadings matrix gives information about the contribution of each of the original variables to each of the principal components. Figure 2 shows an example using the loadings of the main effects in the first four PCs of our model. We see that PC1 and PC2 mostly contain information about the metals used. This is clear from both the relatively large size of the bars associated with "metal descriptors" as well as the absence of bars for "support descriptors". In contrast, PC3 only contains information about the support used. The scores matrix, in combination with the selected points from the candidate list, shows us the structure of the candidate list in the descriptor space and how well the selected points cover the total space. As an example, Figure 3 shows the scores on PC1 and PC2 for all data points, highlighting those points selected by our algorithm. We see that this selection describes the problem well.
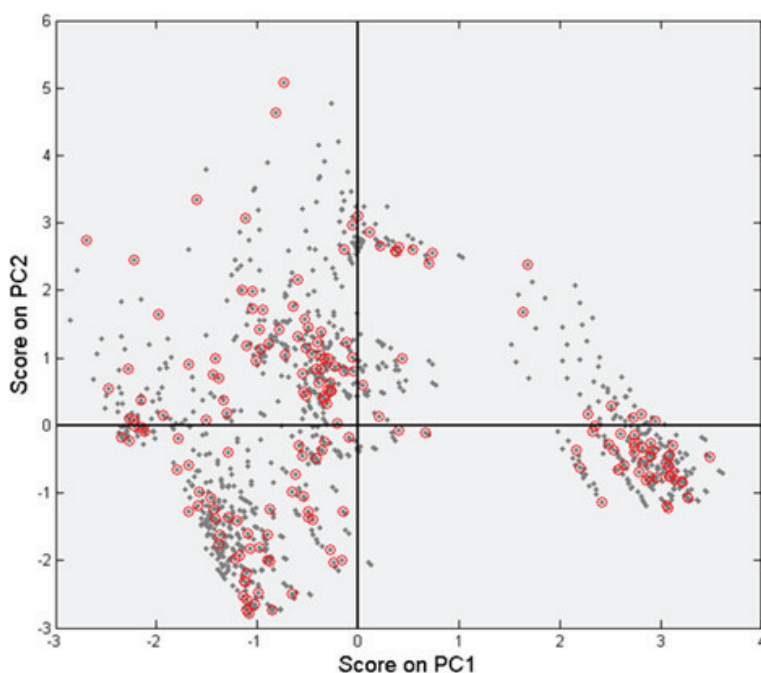

Figure 3 Scores plot of the principal component model describing the candidate set. The 45,864 catalysts in the candidate set are represented as gray dots, the 200 selected candidates are highlighted using red circles.

Note that also the coverage (spread of points) in the other PCs should be evaluated before a decision is made to synthesize and test the selected catalysts. The irregular shape of the scores plot also demonstrates the need for non-classical design methods. Regular experimental design methods are designed to deal with regularly shaped (cubes, spheres, triangles) design spaces. When using this methodology, regularly shaped design spaces are an exception rather than a rule.

The last important concept to consider when using machine-based selection is whether the selected catalysts can actually be synthesized in a meaningful and consistent manner. A chemist will take this into consideration a priori, but for a large set of candidates like the one we consider here this is not a trivial task. Instead of doing this upfront, we need to limit ourselves to carefully reviewing the catalysts once a selection is made. If some materials cannot be synthesized due to, for example, solubility limitations or incompatibility with the support material, a suitable replacement needs to be identified. A concept that can be used for this is similarity. If for some reason a candidate catalyst cannot be synthesized, the "most similar" catalyst that does allow synthesis is selected instead. "Most similar" in this case can be defined as the nearest neighbour of the catalyst that needs to be replaced in descriptor or principal component space. The concept is demonstrated graphically in Figure 4, using the transition metals as an example. For example, we see that Fe and Co are quite similar, but Ir and Ti are not.
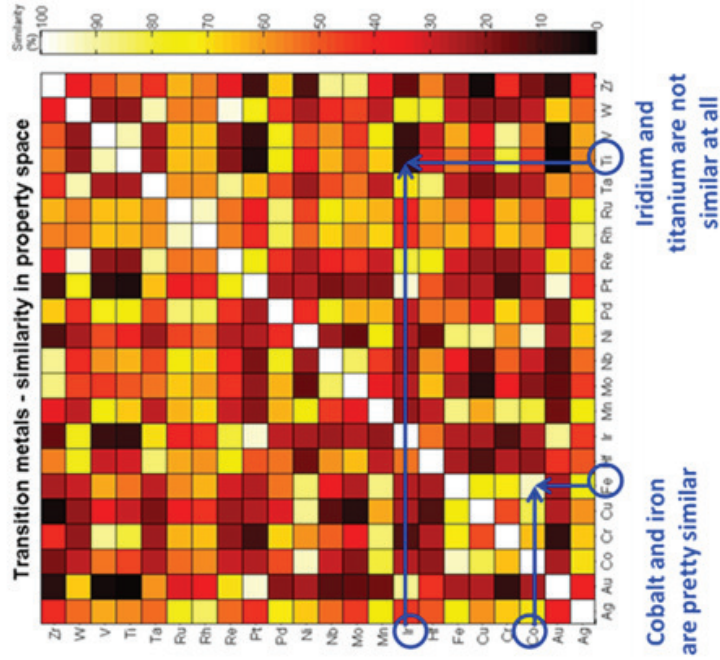
Figure 4 representation and mathematical equations describing the concept of "similarity" in the descriptor space.
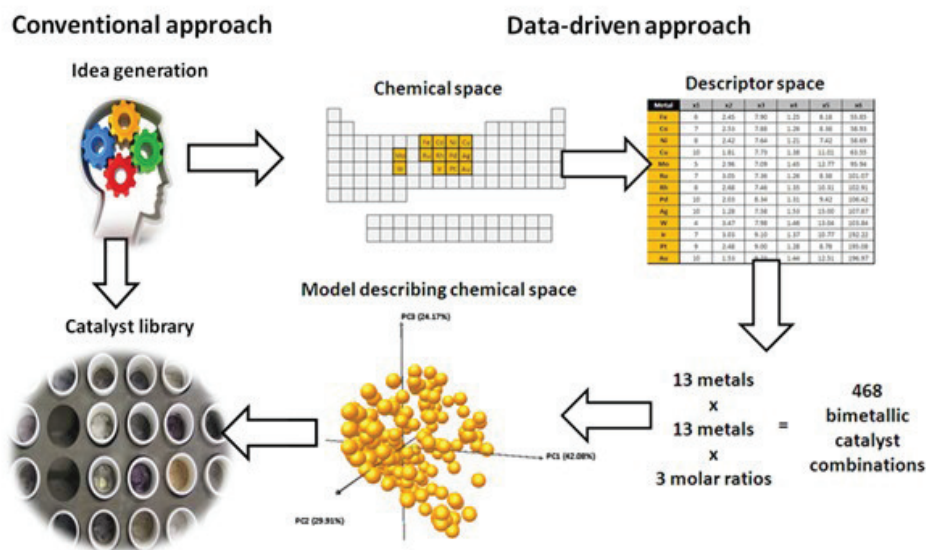
6.1 Conclusions

Figure 5 Workflow for data-driven catalyst development. Note that the workflow is iterative in nature – several cycles will typically have to be completed before reaching the end of a development project.

Discovering and optimising solid catalysts is still largely an empirical business. But, this process can be helped by using the right combination of experimental design, descriptor modelling, high-level modelling, and experimental feedback and validation. A successful optimisation workflow is per definition iterative, and should include all three capacities (statistical design, experimental testing capabilities, and descriptor modelling and validation; see example in Figure 5). Integrating these capacities (and people!) in one team, and realising that multiple iterations are needed, are the keys to success. Furthermore, since much of the research on solid catalysts is done in industrial environment, budgeting multiple iterations of experiments, modelling, and validation will help you create and manage realistic expectations.