doi:10.2533/chimia.2013.71 Chimia 67 (2013) 71–77 © Schweizerische Chemische Gesellschaft

Optimal Compounds Discovery by Design of Experiments and Algorithmic Evolution of Linear Models

Raphaël Dumeunier* and Florent A. Hascher

Abstract: Based on the premise that, for a given class of related chemical compounds, there exists a relationship between their structure and their properties (*i.e.* activity), it is demonstrated herein that an elementary algorithm can readily identify, with simplistic models and without recourse to molecular descriptors, the most active compounds of a categorical, pre-defined space of molecules. In an actual case study using public experimental data on two thousand related molecules, D-optimal design of experiments initially identified the best subset of compounds considered for the construction of simple models. Subsequently, predictions of a first generation of best candidates, their preparation and inclusion into a new data set, allowed the exploration of the most active region within the space of interest. Survival of the algorithm by iterative generations ensured that most of the best (active) compounds had been prepared. A certain partial survival condition, followed by a complete termination criterion, helped to minimize the total amount of compounds to prepare while identifying the *n* best individuals of the matrix.

Keywords: Algorithm · Design of experiments · Regression analysis · Statistic models · Structure–activity relationship

Introduction

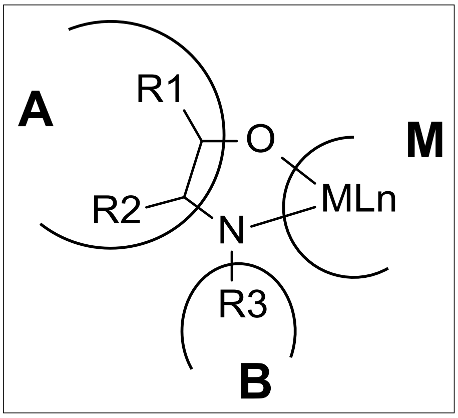
Research in industrial settings requires using minimal time and resources to achieve a specific goal. In chemical companies – pharmaceutical, agrochemical or others – the research chemist is usually asked to find the best compounds of a class without having to prepare many. A common answer to this problem is based on the premise that, within a class of chemical compounds, there is a relationship between structure and activity (SAR, for Structure–Activity Relationship). By far the most favoured method to visualize this function is by changing one parameter at a time (*i.e.* replacing one fragment of the molecule by another), and to observe its effect on the response. But comparing pairs of compounds in which a single point (*i.e.* substituent) has been modified is sensitive to the random variability of the responses, especially when they are quite close. More fundamentally, it makes the severe assumption that the change of substituent accounts completely for the variation observed in the response *independently of the rest of the molecule*.[1] In other words, unless double mutant variations are specifically conceived and prepared,[2] such a method implies that interactions are irrelevant.

The requirement to find the best compound of a class with minimal time and resources is particularly suited for the application of heuristic methods in problem solving. As a matter of fact, more optimal approaches were already invented decades ago,[3] and ever since, a large body of researchers has endeavoured to model structure–activity functions using a fraction only of a given chemical class.[4] These models are particularly useful as they can help both to understand and predict the activity.

However, converting a chemical structure into a usable input for statistical analyses usually requires its 'translation' into continuous descriptors[5] (initial step in quantitative SAR[6] techniques). Each contribution of a descriptor to the activity is an unknown to be estimated, and in a typical QSAR study the relationship between number of unknowns and number of compounds has to be minimally one, ideally six to seven. Once a model is then built and checked for adequacy, the optimal values of every descriptor can be predicted, leaving the chemist free to conceive a specific structure that would fit their combination. A great advantage of QSAR is that the class of compounds is not limited to a finite, given set of structures, as for every combination of continuous descriptors, there should exist an approaching specific structure. However, QSAR has also the disadvantage that information is inevitably lost when the structures used for building the model are converted into calculated descriptors, and when the less significant descriptors are left out for ease of calculation. The alternative choice of keeping a large number of diverse descriptors would instead bring the problem of intercorrelation, or collinearity of independent variables. Finally, there will never be a single molecule able to fit *perfectly* a combination of optimal predicted descriptors, and in the end, a compromise will have to be reached.

More recently, methods circumventing the use of conventional descriptors have begun to emerge.[7] Alongside these sometimes sophisticated algorithms, a simple alternative to descriptors may easily be found when the exploration is restricted to a pre-defined categorical molecular space,[8] wherein all possible compounds can be inventoried and numbered. In order to construct such a space of molecules, expanding from a prototypical compound of interest (for instance, in Scheme 1, hydrogenation catalysts), the chemist can divide the latter into parts (in Scheme 1, A, B and M), and define arbitrarily, as categorical levels for every variable, the nature and number of substituents variations; that is, different specific A_1, A_2... B_1, B_2... M_1, M_2... also of promising potential.

*Correspondence: Dr. R. Dumeunier
Syngenta Crop Protection
Münchwilen AG, Schaffhauserstrasse
CH-4332 Stein
Tel.: +41 62 8660271
E-mail: raphael.dumeunier@syngenta.com

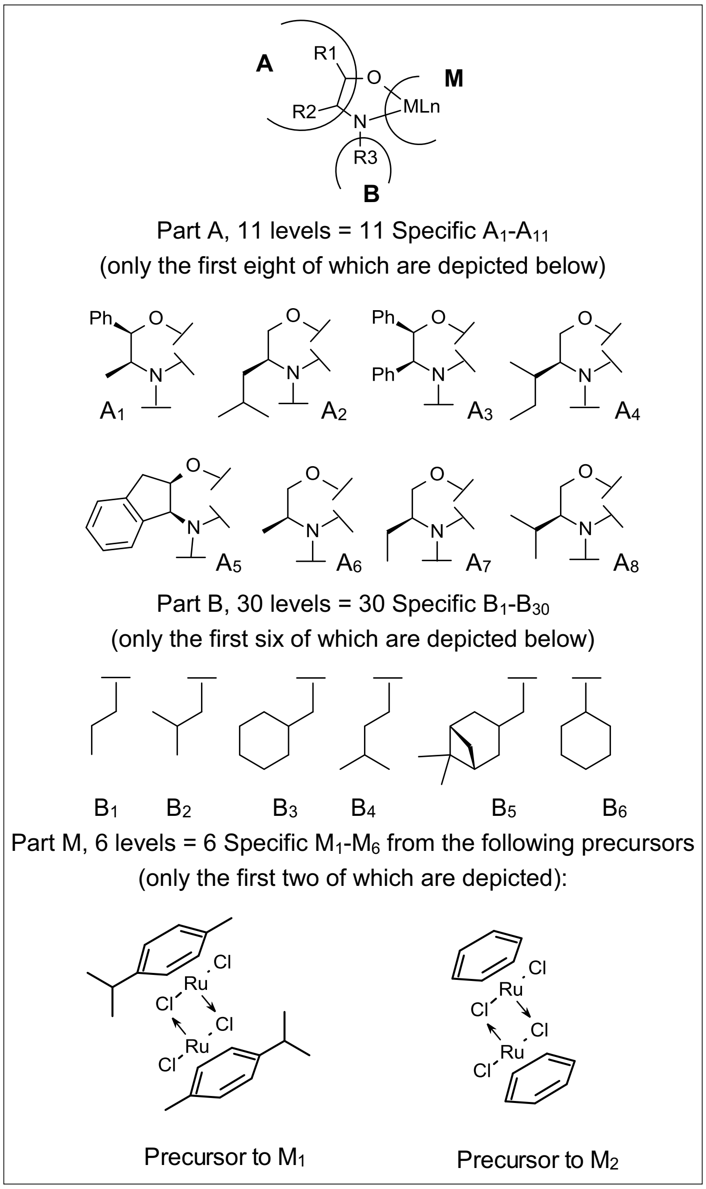


Scheme 1.

Compounds are then simply described by the combination of their fragments (for example, compound $A_1B_3M_2$), before being converted into linear equations by the use of meaningless, *dummy variables*, as will be seen later.

A key question is then, how to find the most active molecule(s) of this categorical space, by making a minimal number of compounds. Like in classical QSAR,[9] a good starting point is an optimal Design of Experiment (DoE).[10] This brings the advantage of taking into account, if desired and *a priori*, possible interactions between fragments, while minimizing the variance of prediction with a most diverse and balanced subset of structures. Once all compounds of the data set, defined by this prior DoE operation, would have been prepared and their experimental activity measured, a most simple linear model can be fitted and used to predict the rest of the molecular space. Preparing the best *n* of the remaining compounds and measuring their activity should lead the chemist into the most active area of the complete space. But as with naïve linear models, only a fraction of the variation in measured activity will be attributed to the nature of substituents, an evolution of the algorithm ought to be envisaged. For example, after a first generation of potentially active compounds has been prepared, one can envisage combining the initial DoE data set with this generation in a new data set. The very same model can be built once again, this time from the augmented set, predictions can be made, and the best predicted compounds can be prepared and evaluated. This cycle is reiterated thus as long as a certain survival condition stays fulfilled, to ensure that a large proportion of the best compounds have been found before meeting a termination criterion.

The present paper reports the application of this algorithm to the exploration of a pre-defined, categorical population of almost two thousand transfer hydrogenation catalysts, and to the discovery of the best ten specimens. The influence of arbitrary factors such as survival and termination conditions, or the nature of the model, has also been investigated.


A
R1
O
M
R2
MLn
N
R3
B
Part A, 11 levels = 11 Specific A1-A11
(only the first eight of which are depicted below)
Ph
O
N
A1
A2
Ph
Ph
A3
A4
A5
A6
A7
A8
Part B, 30 levels = 30 Specific B1-B30
(only the first six of which are depicted below)
B1
B2
B3
B4
B5
B6
Part M, 6 levels = 6 Specific M1-M6 from the following precursors
(only the first two of which are depicted):
Cl
Ru
Cl
Cl
Ru
Cl
Precursor to M1
Precursor to M2

Scheme 2.

Discussion and Results

In 2009, Riant and Vriamont reported the preparation of a *complete* library of transfer hydrogenation catalysts.[11] Completeness was effected in the sense that a generic catalyst had been divided into three parts (A, B and M; Scheme 2), their levels were defined (respectively 11, 30 and 6) and, in a herculean effort, all 1980 possible molecules were successfully prepared and evaluated.

The catalysts were evaluated both for conversion and enantiomeric excess in the asymmetric reduction of acetophenone, and their performance was normalised in a single number (0-1) against the best catalyst (NPF; Normalised Performance Factor).[12] As can be seen from Fig. 1, a very large proportion performed very poorly. To find the ten best candidates, for example, without having to make them all, might quickly appear as a daunting task, as more than half of the population has no activity and therefore, gives no information as to where is the most active area of the space.

In exactly such an exercise, Riant and coworkers demonstrated the use of a genetic algorithm[13] for simulated evolu-

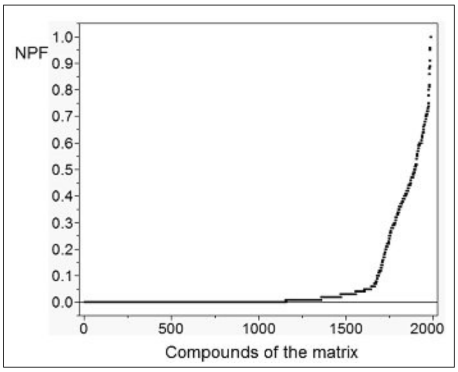


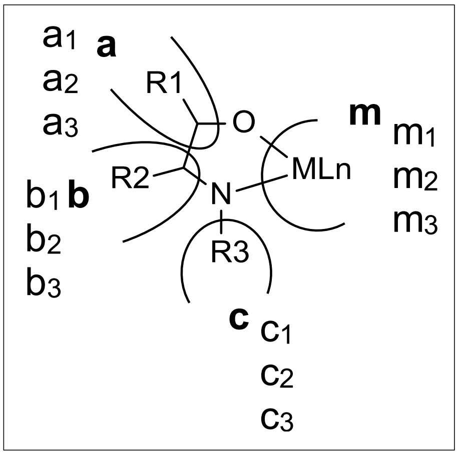
Fig. 1.

tion. They indeed endeavoured to find a maximum of catalysts among the best ten of the whole space, without having to prepare more than two hundred of them. This endeavour is in some respect quite similar to many of research scenarios in life-sciences companies, with the difference that the performance (responses) of interesting molecules is being measured in terms of *in vitro* and *in vivo* biological potencies, stability, physico-chemical properties and many others variables, up to the complete absence of toxicity at the required doses.

Our simple heuristic model, unfolded below, developed exactly for these scenarios and already in use within our company, may actually outperform the genetic algorithm proposed by Riant and coworkers. Even though success was undeniably met in the search of the best candidates with a genetic algorithm (around six out of the ten best were found on average), we realised indeed that an iterative evolution of simple linear models should allow the preparation of fewer compounds before finding a better proportion of the best ten catalysts. The first model would be fitted from a most diverse subset of compounds (D-optimal[14]), and taking into account, *a priori*, interaction between fragments.

Models Rationale

Because of the length of the models equations, the mathematical mechanics of our algorithm will be exemplified with a different, much smaller imaginary library composed of 3×3×3×3 specific combinations (Scheme 3).



Scheme 3.

Without Interactions

A standard[15] *main effects model* can be expressed in the following equation (Eqn. (1)).

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \\ + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8 + \varepsilon$$

where y_i = measured response variable for the i^{th} compound of the data set, β_0 to β_8 are constants with values that would have to be estimated from the data set. β_0, called the intercept, needs to be associated (arbitrarily) to one specific compound of the data set (here, say $\mathbf{a_1b_1c_1m_1}$), then:

x_1 = 1 if substituent in **a** is $\mathbf{a}_2$, 0 if not
x_2 = 1 if substituent in **a** is $\mathbf{a}_3$, 0 if not
x_3 = 1 if substituent in **b** is $\mathbf{b}_2$, 0 if not
x_4 = 1 if substituent in **b** is $\mathbf{b}_3$, 0 if not
x_5 = 1 if substituent in **c** is $\mathbf{c}_2$, 0 if not
x_6 = 1 if substituent in **c** is $\mathbf{c}_3$, 0 if not
x_7 = 1 if substituent in **m** is $\mathbf{m}_2$, 0 if not
x_8 = 1 if substituent in **m** is $\mathbf{m}_3$, 0 if not
ε = unexplainable, or random, error

The variables x_1 to x_8 are not meaningful independent variables, but *dummy indicator variables*. Note that even if there are a total of twelve levels in our space ($\mathbf{a}_1$ to $\mathbf{m}_3$), it is possible to describe all of them with only eight dummy variables, because the four base levels $\mathbf{a}_1$, $\mathbf{b}_1$, $\mathbf{c}_1$ and $\mathbf{m}_1$ are accounted for by the intercept β_0. On the choice of the base levels associated to the intercept will thus depend the attributions, or coding, of the independent variables x_1 to x_8.[16]

As we make the implicit assumption, for one,[17] that the mean of the probability distribution of the random errors is zero ($E(\varepsilon) = 0$), our best estimate of ε is zero. The predicted variable response (activity), of any compound ***j*** of the complete categorical space, thus becomes

$$\hat{y}_j = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + \hat{\beta}_4 x_4 + \hat{\beta}_5 x_5 + \\ \hat{\beta}_6 x_6 + \hat{\beta}_7 x_7 + \hat{\beta}_8 x_8$$

with $\hat{y}_j$ = predicted variable response of the ***j***$^{\mathbf{th}}$ compound of the complete space, x_1 to x_8 are defined as above and their coefficients are the least squares estimates of the model 'true' parameters β_0 to β_8.

In order to numerically solve this prediction equation for any compound, the values of the coefficient predictors have first to be found. And in order to find solutions for these nine unknowns, an equal number (at least) of equations is to be assembled in a system. As *(i)* each different compound can be associated to a distinct equation, and *(ii)* the intercept β_0 has arbitrarily been chosen to account for the four base levels $\mathbf{a_1b_1c_1m_1}$ eight additional compounds remain to be prepared. The key question is which eight exactly? An optimal design would require, for example, a global set of nine compounds containing all possible substituents an equal and maximal number of times, with all possible combinations of any two substituents being present. Such a design ensures a maximum diversity in the structures, and as we assume a structure–activity function, a maximum diversity in the responses is also expected (from inactive to active).[18] An example may be the following set of nine catalysts (Set 1).

$\mathbf{a_1\ b_1\ c_1\ m_1}$
$\mathbf{a_1\ b_2\ c_2\ m_2}$
$\mathbf{a_1\ b_3\ c_3\ m_3}$
$\mathbf{a_2\ b_1\ c_2\ m_3}$
$\mathbf{a_2\ b_2\ c_3\ m_1}$
$\mathbf{a_2\ b_3\ c_1\ m_2}$
$\mathbf{a_3\ b_1\ c_3\ m_2}$
$\mathbf{a_3\ b_2\ c_1\ m_3}$
$\mathbf{a_3\ b_3\ c_2\ m_1}$
Set 1

As can be seen in Set 1, all fragments ($\mathbf{a}_1$ to $\mathbf{m}_3$) do indeed appear three times, and all possible combinations of any two fragments are represented (*i.e.* $\mathbf{c}_1$ with $\mathbf{a}_1$, but also with $\mathbf{a}_2$, $\mathbf{a}_3$, $\mathbf{b}_1$, $\mathbf{b}_2$, $\mathbf{b}_3$, $\mathbf{m}_1$, $\mathbf{m}_2$, $\mathbf{m}_3$). There are seven other sets of nine compounds including $\mathbf{a_1b_1c_1m_1}$ that would fit these two conditions, *and be equally good to start with*. This data set is converted into a system of nine equations, using the definitions of Eqn. (1), once their experimental activity would have been measured ($\boldsymbol{y_1}$ to $\boldsymbol{y_9}$; if we number the compounds of Set 1 from one to nine), written hereafter under matrix algebraic form:

$$\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{Y}$$

or, under developed matrix form:

$$\begin{vmatrix} 1\,0\,0\,0\,0\,0\,0\,0\,0 \\ 1\,0\,0\,1\,0\,1\,0\,1\,0 \\ 1\,0\,0\,0\,1\,0\,1\,0\,1 \\ 1\,1\,0\,0\,0\,1\,0\,0\,1 \\ 1\,1\,0\,1\,0\,0\,1\,0\,0 \\ 1\,1\,0\,0\,1\,0\,0\,1\,0 \\ 1\,0\,1\,0\,0\,0\,1\,1\,0 \\ 1\,0\,1\,1\,0\,0\,0\,0\,1 \\ 1\,0\,1\,0\,1\,1\,0\,0\,0 \end{vmatrix} \begin{vmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \\ \hat{\beta}_4 \\ \hat{\beta}_5 \\ \hat{\beta}_6 \\ \hat{\beta}_7 \\ \hat{\beta}_8 \end{vmatrix} = \begin{vmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \\ y_8 \\ y_9 \end{vmatrix}$$

In this case, **X** is a *square* data matrix, and a simple Gaussian elimination would give the values of the estimators. But we rather recommend the more general *least squares* solutions: if indeed, as will be seen later in the actual case study, the model uses more equations than unknowns, the least squares matrix solution can be obtained from the following equation (Eqn. (4), where $\mathbf{X}^{\mathbf{T}}$ is the transpose of **X** and $(\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}$ is the inverse of the product $\mathbf{X}^{\mathbf{T}}\mathbf{X}$):

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}(\mathbf{X}^{\mathrm{T}}\mathbf{Y})$$

Solutions of the normal equations will deliver the least squares estimators of β_0 to β_8 and from Eqn. (2), the activity of all compounds of the space of interest can be predicted.

This *main effect model* gives in essence the same predictions as a Free-Wilson analysis would, and its mechanics mirror almost exactly those of the Fujita-Ban analysis.[19] However, in our case, the algorithm that started with an optimal subset will evolve into generations, and survive to identify the best n% of the full space before its termination. Moreover, such a simple regression analysis can easily be extended to account for interactions between the categorical, qualitative variables, as will be demonstrated in the next case.

With Interactions

If interaction between two specific fragments would have been foreseen as likely (or even interactions between all fragments of any two parts of the prototypical molecule), then another linear model can be used in place of Eqn. (1). For example, secondary interactions between all possible specific fragments of parts **a** and **b** are accounted for in the new model written in Eqn. (5).

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8 + \beta_9 (x_1 x_3) + \beta_{10}(x_1 x_4) + \beta_{11}(x_2 x_3) + \beta_{12}(x_2 x_4) + \varepsilon$$

with the same definitions as (Eqn. (1))

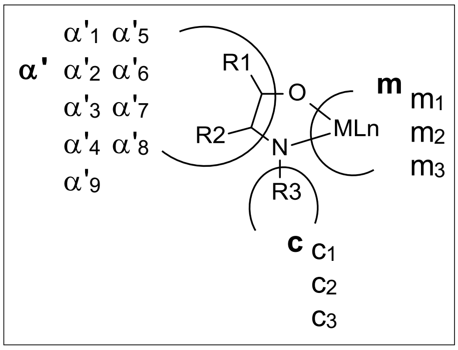
As can be seen from Eqn. (5), thirteen unknowns are now present (β_0 to β_{12}), and therefore thirteen compounds at least are required to find their estimators. An optimal design might lead this time, for example, to the following set of compounds (Set 2).

$\mathbf{a_1\, b_1\, c_1\, m_1}$
$\mathbf{a_1\, b_2\, c_2\, m_2}$
$\mathbf{a_1\, b_3\, c_3\, m_3}$
$\mathbf{a_2\, b_1\, c_2\, m_3}$
$\mathbf{a_2\, b_2\, c_3\, m_1}$
$\mathbf{a_2\, b_3\, c_1\, m_2}$
$\mathbf{a_3\, b_1\, c_3\, m_2}$
$\mathbf{a_3\, b_2\, c_1\, m_3}$
$\mathbf{a_3\, b_3\, c_2\, m_1}$
$\boldsymbol{a_1\, b_1\, c_2\, m_2}$
$\boldsymbol{a_2\, b_2\, c_1\, m_3}$
$\boldsymbol{a_3\, b_3\, c_3\, m_3}$
$\boldsymbol{a_2\, b_1\, c_1\, m_2}$

Set 2: in which the first nine individuals are identical to those of Set 1. All combinations of any two levels of parts being indeed already present in Set 1, the other four extra compounds of Set 2 (*ital.*) are thus 'only' required to equate the number of compounds with the number of unknowns, in a new system allowing unique solutions. This is an optimal starting point composed of a strict minimum of compounds.[20] The algorithm now takes into account all interactions between every fragments of parts **a** and **b**.

In a similar way, we could have decided to introduce more interactions terms in our model, not only between parts **a&b**, but also between **b&m**, **b&c**, or any combinations of them.

In the exemplified case where only interactions terms between **a&b** had been taken into account, it is noteworthy that the exact same predictions would have been found, if instead of dividing the molecule into four parts, we would have considered three different ones (Scheme 4), *and ignore any interactions!*



Scheme 4.

Indeed, if $\alpha'_1=[\mathbf{a}_1\mathbf{b}_1]$, $\alpha'_2=[\mathbf{a}_1\mathbf{b}_2]$ *etc.*, the categorical space of eighty-one compounds is identical to the one depicted in Scheme 3. A linear model without interactions would require 13 compounds, and from the same, *it would lead exactly to the predictions of the interaction model of Eqn. (5).* This highlights the importance of how to divide the molecule into parts. Less of them implies larger data set for predictions, but also inherently accounts for 'interactions' *within each individual part*, like part α does for former parts **a** and **b**.

Actual Case Study

These simplified cases exemplify succinctly possible models behind the actual object of interest, that is, Olivier Riant's library of 1980 different catalysts (Scheme 2). With a combination of 11×30×6 fragments (A, B and M), a linear model without interactions would contain 45 unknowns (1+10+29+5), and require minimally as many compounds for predictions.

Obviously, we assumed no prior knowledge of *any* single experimental activity of the catalyst library, like in a real case scenario. The difference being that instead of having to actually prepare and measure the activity of the compounds identified in our algorithm, we would simply extract the information from the work of Riant and coworkers.

Using the statistical software JMP®9,[21] we initially planned to perform a design of experiment to identify which data set (or generation zero, G^0) to prepare. To maximize our chance to find the most active catalysts in the subsequent steps, we would have liked, *a priori,* to take into account interactions. However, interaction of part B (30 levels) with part A creates 290 (10×29) new unknowns; and with part M, 145 (29×5) unknowns. As we wanted to keep the number of catalysts in our initial data set G^0 relatively low (<5% of the whole), we could only take into account interactions between parts A&M, creating only 50 (10×5) new unknowns.

With Interactions

We therefore opted for an optimal design over A, B, and M, with secondary interactions between A&M, that is, 45 unknowns for main effects + 50 new unknowns for interactions. From a chemical viewpoint, it is equivalent to having bisected the prototypical catalyst in parts [AM] and B, and considering to fit a linear model without interactions.

JMP®9 defined readily the list of the 95 compounds required, and after we had read their experimental activity (in this case, the normalized performance factor, NPF, being a function of the measured *ee* and conversion, ranging from 0-1),[22] we built a linear model from G^0 by implementing the required matrix operations in Java programing language.[23]

The system of 95 equations was thus solved, providing a first formula used to calculate the predicted NPF for all 1885 remaining compounds. Those predicted NPF were ranked, and the ten best of the 1885 candidates were 'prepared' and their performance 'evaluated'. This new, first generation G^1, had a significantly higher average performance than G^0 (0.34 *vs* 0.07) but no individual catalyst was outstanding. So we decided to iterate the model from there, but using G^0 and G^1 combined as a new data set. By doing so, we were voluntarily creating an unbalance in the data set; this bias towards the more active catalysts being supposed to help us to explore better the most active region of space.

From these 105 catalysts (G^0 and G^1), we therefore re-built the same linear model. In this case, there are more equations than unknowns (over-determined system), and the Java program returned true *least squares* estimators from Eqn. (4). Again, the predicted activity (NPF) of the 1875 remaining catalysts was calculated using the new solutions, and the best ten predicted were 'prepared' and evaluated (second generation G^2). This time the average NPF was quite good (0.66), and three catalysts were performing very well (>0.8).

A question that appeared naturally was when to stop building new generations of catalysts. As the former genetic algorithm had been evaluated by its score on the overall top 10, we had also decided to find most of the top 10 compounds with as few generations as possible. And had we already found, in the three generations G^0 to G^2, the 10 bests catalysts of all? This fact can only be proven true by looking up all the solutions, but it may possibly be *proven wrong* otherwise, without having to. For this, it is only required to make a new generation G^3, and check that in G^3, no catalyst is better than the *tenth best of the 115 belonging to G^0, $G^1$and G^2*.

If, indeed, at least one catalyst of G^3 is found to be better than the best 10th of the previous combined generations, it proves unambiguously that the top ten catalysts of the whole space had not been found yet. That is part of the survival condition, allowing our algorithm to enter a new cycle.

If however no better compound than the 10th best of the samples already made is found, it obviously does not prove that the global top 10 had been identified. To compensate for this uncertainty, we decided to always stop our algorithm after *another* of such an unproductive generation. Having discovered *two* unproductive generations, consecutive or not, is then our arbitrary termination criterion.

As we had found in G^1, four catalysts being more active than the 95 others, we moved to G^2, and in this generation, eight new catalysts were better than the 10th best of the 105 others. The third generation G^3, found by modelling the data set [G^0+G^1+G^2], allowed the identification of four new catalysts entering the new top 10; G^4 identified three new of such catalysts again; G^5, one and G^6, two. At last, G^7 did not identify any new better than the latest top tenth compound, G^8 led to a similar observation, and the condition to stop evolution after the second of such an unproductive generation was fulfilled.

In the end, we had to prepare 175 catalysts, in a total of nine generations, including G^0. As our algorithm was ended, we could finally look up in the reference file of Riant and coworkers and count how many of the global top ten (present in Table 1) had been identified.

The full algorithm (from matrix building to survival and termination criteria) was implemented in our Java program and ran thirty-nine more times, from *randomly chosen, different but equally optimal*, generations zero. The average score of the 39 attempts is given in Table 2, along with the average number of catalysts and generations required before termination.

On average, 8 catalysts out of the best 10 were found, by having to 'prepare' only 180 compounds. Individually, exactly one third (13 out of the 39 different generations zero) did actually lead to a perfect 10/10 score.

Without Interactions

In the previous case, we had decided arbitrarily to account for interactions in our model and we succeeded in finding most of the best catalysts. Without taking interactions into account, G^0 would be composed of only 45 catalysts. The average scores on forty runs, from as many different, equally optimal *generations zero* of 45 compounds only is given in Table 3.

Table 1.

Entry	Catalyst	NPF	Found in
1	$A_1B_{19}M_2$	1	G^5
2	$A_1B_{17}M_2$	0.96	G^4
3	$A_1B_4M_1$	0.95	
4	$A_7B_4M_1$	0.91	
5	$A_7B_{19}M_1$	0.89	G^6
6	$A_1B_{20}M_2$	0.88	G^2
7	$A_1B_{11}M_2$	0.86	G^2
8	$A_1B_{22}M_2$	0.82	G^2
9	$A_1B_{15}M_2$	0.81	G^4
10	$A_1B_9M_1$	0.8	G^6

Table 2.

Run	Score /10	Cat.	Gen.
1 - above	8	175	9
2 to 40	7.97	180.4	9.54

Table 3.

Runs	Average		
	Score /10	Cat.	Gen.
1 to 40	7.25	144	10.9

The number of catalysts to prepare is significantly lower than in the previous model, but this is exclusively due to the smaller size of G^0. More details as to the performance of both algorithms, averaged on forty runs, can be read from Fig. 2.[24]

Judging *a posteriori* if the initial choice to take into account interactions was best is here a vain enterprise, but in a real case it would depend on many factors, *i.e.* the actual ease of synthesis of the catalysts, the extra amount of work required to prepare 95 catalysts (G^0 with interactions) compared to 45 (G^0 without interactions), or the importance to find most of the top ten *versus* the total number of compounds to be prepared. A rigorous comparison of means leads here, with more than 95% certainty ($\alpha = 0.05$), to the conclusion that *more compounds of the top ten were found* per generation *with the interaction model*, *whereas more compounds of the top ten were found* per total compounds to prepare *with the main effects model*.

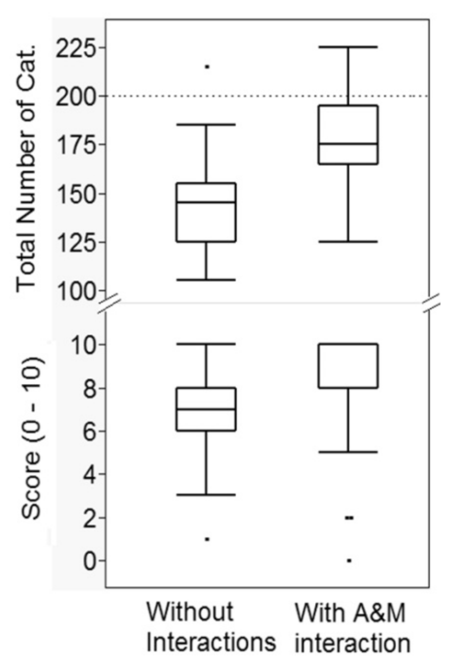


Fig. 2.

In other words, a higher predictive power of the interactive model was reflected in the need for fewer generations to find more active compounds, but the main effects model is in the end more *economic*, with an optimal score per total compounds to prepare.

Regression Analysis of the Complete Library

With the categorical space being filled experimentally, it is both possible and informative to check *a posteriori* which of the two models (in bold, Table 4) used here above was the most correct. The same question may be asked about the other interactions models which we had deliber-

Table 4[25]

Model	R^2	R^2 Adj	RMSE	PRESS
Main Effects (ME)	**0.38**	**0.37**	**0.131**	**34.7**
ME+A*B	0.49	0.39	0.129	39.6
ME+A*M	**0.55**	**0.53**	**0.113**	**26.5**
ME+B*M	0.46	0.41	0.127	35.3
ME+A*B+A*M	0.66	0.58	0.107	28.0
ME+A*B+B*M	0.57	0.43	0.124	40.0
ME+A*M+B*M	0.63	0.58	0.106	25.4
ME+A*M+B*M+A*B	0.74	0.65	0.098	25.7

Table 5.

Entry	G^1 to G^n	Unprod.	Average		
	Size	gen.	Score /10	Cat.	Gen.
1	10	2	7.25	144	10.9
2	15	2	8.88	201	11.4
3	10	3	7.95	163	12.8

ately left out at the onset of our study. Their evaluations are summed up below.

F-tests (α = 0.05) for interactions have been performed on all models, and there was sufficient evidence to conclude that not only A and M, but *all parts* of the generic catalyst do interact two-by-two. And it is striking that the most spartan Main Effects (ME) model, even with a regression equation explaining only 38% of the total variation in catalyst performance, gave out very decent scores.

Influence of the Survival and Termination Criteria

Undeniably, the excellent start of the algorithm with a prior DoE is crucial to this relative success. Two other elements are also key to this performance: the algorithm survival and the evolution of generations. When the model is poor, which for the linear ones it was certainly safe to assume, it is essential for the algorithm to survive long enough, so that every little success cropped in each cycle adds up to a final, befitting score. But what if we would have given the algorithm an extra chance with three unproductive generations allowed instead of two? The influence of these arbitrarily pre-defined conditions was investigated by running forty times the Main Effects algorithm with 15 compounds per generation (entry 2, Table 5), or with three unproductive generations allowed (entry 3), compared to our standard algorithm (entry 1).

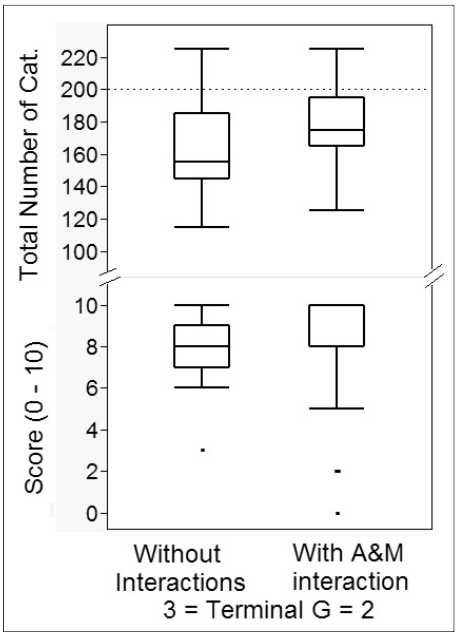


Fig. 3.

If the average scores did improve significantly in entry 2, the total number of compounds to prepare seems to deviate from optimality. But it is worth noting with entry 3 that opting for a model without interactions, *even when they are relevant and the model is poorer*, is slightly better than choosing a model with interactions, *at the condition to compensate by adopting a less severe termination criterion*. This can be seen from comparing entry 3 with the results of Table 2, and eventually visualised with more details in the box-and-whisker diagrams of Fig. 3.

A final, legitimate question is about the size (10 compounds) of generations. It would only be optimal *for certain* with a perfectly predictive model, as a single generation after G^0 would identify the overall top ten compounds. Even in our case, even with a large deviation from a perfect predictive model, choosing more than 10 compounds (entry 2, Table 5) is sub-optimal.

Conclusions

We have shown how iterative generations of simple linear models can be used to find efficiently the most active compounds within pre-defined molecular spaces, *even when a large fraction of the space is completely inactive*. A good initial design of experiment, the knowledge transfer from one generation to the data set of the next, and a not-so-severe termination criterion certainly make up for the simplicity of the predictive models. All of the algorithms presented above, including models with or without interactions, using standard or larger generations, or even with different termination criteria, lead to the identification of most of the overall best ten compounds, by preparing less than 10% of the whole matrix.

The algorithm may easily be adapted to specific situations: for example, if it costs time to measure the response, it may be better to work with *a priori* more precise interactive models, as fewer generations are required. And if it is desired instead to find most, if not all, of the top n, the size of generations and the termination condition can be tuned to maximize the score.

It is also crucial, if the most active compound has been found but is not up to the expectations, to perform a post-algorithm interpretation.[26] Due to its simplicity, the models presented in this paper are particularly well suited for this essential operation. A qualitative analysis of the data, in order to try to understand why such are active where others are not, and to generate hypotheses, is required to build new spaces, or expand the existing one. The algorithm may well be useful for optimising search time, but it is only the first step in a more global strategy. A proper interpretation remains imperative for breaking the boundaries of the arbitrarily predefined space of molecules.

Assuming that for a given property, there exists a structure–response relationship, such a method may be used for all types of properties. Within our company, Syngenta, chemical structures are linked to *in vivo* biological activity on insects, fungi, plants and nematodes. They are also used to collect experimental data on toxicity, metabolism, uptake, stability, and many more properties. The relationships between the structures and all these responses are obviously different, but if they indeed exist, a single optimal data set, obtained from DoE techniques, remains valid for assessing different models for each response. A compromise can then be looked for in the predictions. The algorithm described in this paper makes here a first step into the labyrinthine paths of multi-optimisation,[27] in the hope to lead more efficiently to the best chemical candidates for development.

Acknowledgments

We thank Dr. Guillaume Berthon and Dr. Sebastian Wendeborn for active support and encouragements, Prof. Bernadette Govaerts for insightful discussions, and Prof. Olivier Riant.

Received: October 18, 2012

[1] Y. Patel, V. J. Gillet, T. Howe, J. Pastor, J. Oyarzabal, P. Willett, *J. Med. Chem.* **2008**, *51*, 7552.

[2] F. J. Carver, C. A. Hunter, P. S. Jones, D. J. Livingstone, J. F. McCabe, E. M. Seward, P. Tiger, *Chem. Eur. J.* **2001**, *7*, 4854.

[3] a) C. Hansch, P. P. Maloney, T. Fujita, R. M. Muir, *Nature* **1962**, *194*, 178; b) S. M. Free, J. W. Wilson, *J. Med. Chem.* **1964**, *7*, 395; c) C. Hansch, *Acc. Chem. Res.* **1969**, *2*, 232.

[4] a) For a review on chemometrics, see B. Lavine, J. Workman, *Anal. Chem.* **2008**, *80*, 4519; b) For a review on chem(o)informatics, see M. Vogt, J. Bajorath, *Bioorg. Med. Chem.* **2012**, *20*, 5317; c) For a review on machine learning methods in chemistry, see A. Varnek, I. Baskin, *J. Chem. Inf. Model.* **2012**, *52*, 1413.

[5] R. Todeschini, V. Consonni, 'Molecular descriptors for Chemoinformatics', Wiley-VCH, Berlin, **2009**.

[6] For a comprehensive description of QSAR methods, see E. X. Esposito, A. J. Hopfinger, J. D. Madura, *Methods Mol. Biol.* **2004**, *275*, 131.

[7] a) I. Baskin, V. A. Palyulin, N. S. Zefirov, *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 715; b) A. Goulon, T. Picot, A. Duprat, G. Dreyfus, *SAR QSAR Environ. Res.* **2007**, *18*, 141; c) P. Mahe, L. Ralaivola, V. Stoven, J.-P. Vert, *J. Chem. Inf. Model.* **2006**, *46*, 2003; d) T. W. Heritage, D. R. Lowis, *Rational Drug Design* **1999**, 212; e) W. K. Yeo, K. L. Tan, S. B. Koh, M. Khan, S. Nilar, M. L. Go, *ChemMedChem* **2012**, *7*, 977.

[8] For a discussion on the use of the terms 'molecular space' or 'space of molecules' instead of 'chemical space', see R. Carbo-Dorca, *J. Math. Chem.* **2012**, in press, *http://dx.doi.org/10.1007/s10910-012-0091-x*

[9] a) M. L. Tosato, S. M. Marchini, L. Passerini, A. Pino, L. Eriksson, F. Lindgren, S. Hellberg, J. Jonsson, M. Sjöström, B. Skagerberg, S. Wold, *Environ. Toxicol. Chem.* **1990**, *9*, 265; b) P. M. Mager, *Med. Res. Rev.* **1997**, *17*, 453; c) H. Tye, *Drug Discovery Today* **2004**, *9*, 485.

[10] T. Lundstedt, E. Seifert, L. Abramo, B. Thelin, A. Nyström, J. Pettersen, R. Bergman, *Chemomet. Intell. Lab. Sys.* **1998**, *42*, 3.

[11] N. Vriamont, B. Govaerts, P. Grenouillet, C. de Bellefon, O. Riant, *Chem. Eur. J.* **2009**, *15*, 6267.

[12] As mentioned judiciously by a referee, the use of NPFs is an excellent choice for parameters with well-defined upper and lower limits, like here with yields (0-100) and *ee* (0-100). This choice would also be required, but more difficult to set in place, when parameters to optimize have no upper or lower bounds (especially when multiple parameters are to be optimized, to insure properly balanced metrics).

[13] For a review on genetic algorithms in chemometrics, see: A. Niazi, R. Leardi, *J. Chemometrics* **2012**, *26*, 345.

[14] a) K. Smith, *Biometrika* **1918**, *1*, 1; b) P. F. de Aguiar, B. Bourguignon, M. S. Khots, D. L. Massart, R. Phan-Than-Luu, *Chem. Intell. Lab.* **1995**, *30*, 199.

[15] (a) N. R. Draper, H. Smith, 'Applied Regression Analysis', John Wiley & Sons, Somerset, **1981**; b) W. Mendenhall, T. Sincich, 'A Second Course in Statistics: Regression Analysis', Prentice Hall, Boston USA, **2012**.

[16] A comment on the special role and future influence of the arbitrary choice of $\mathbf{a_1b_1c_1m_1}$ in this set has to be made. Compound $\mathbf{a_1b_1c_1m_1}$ has, when included in this optimal design, nothing more special than the eight others of the set, as one could choose to stay with this exact set of compounds and use any eight of the nine other compounds to incorporate in the intercept β_0. The coding of the variables x_1-x_8 would change accordingly, and the predictions would stay exactly the same.

[17] Three other assumptions on ε are (i) $V(\varepsilon)=\mathbf{I}\sigma^2$, (ii) ε has a normal probability distribution (required for F-tests) and (iii) εs associated with any two different observations are independent.

[18] This is actually an orthogonal array (OA), built easily or found in public OA tables. A nice property of an OA is that the mean of the responses of the OA subset is equal to the mean of the responses of the whole space investigated, if a linear model is valid within.

[19] T. Fujita, T. Ban, *J. Med. Chem.* **1971**, *14*, 148.

[20] To retain 100% D-optimality, 18 compounds would be required (a combination of two orthogonal arrays).

[21] JMP®, Statistical Software from SAS Institute, *http://www.jmp.com*

[22] It is unsure whether converting the performance to a log scale would improve the algorithm. See: E. Gaebler, R. Franke, P. Oehme, *Pharmazie* **1976**, *31*, 1.

[23] Written and developed in an Eclipse Indigo environment.

[24] The bottom and top of the box is the lower and upper quartile (25th and 75th percentile); the line in the box is the median (50th percentile); whiskers contain the lowest and highest data (still within a maximum distance of 1.5 the interquartile range) and any data not included in the boxplot is represented as an outlier with a dot. In Figs 2 and 3, the lower right box may appear to have no median in the box, but it actually coincides with the lower edge of the box, at a value of 8.

[25] Adj R^2 is the Adjusted Coefficient of Multiple Determination; PRESS stands for Prediction Sum of Square; RMSE is Root Mean Square Error. As criteria for selecting the best model, Adj R2 has to be maximum (closest to 1) and both RMSE and PRESS have to be as low as possible.

[26] R. Guha, *J. Comp. Aided Mol. Des.* **2008**, *22*, 857.

[27] I. Nilsson, N. Polla, *J. Comput. Aided Mol. Des.* **2012**, in press, *http://dx.doi.org/10.1007/s10822-012-9605-7.*