Combinatorial Explosion in Homogeneous Catalysis: Screening 60,000 Cross-Coupling Reactions

Enrico Burello,[a] David Farrusseng,[b,*] Gadi Rothenberg[a,*]

[a] van' t Hoff Institute of Molecular Sciences, University of Amsterdam, Nieuwe Achtergracht 166, 1018 WV Amsterdam, The Netherlands
Fax: (+31)-20-525-5604, e-mail: gadi@science.uva.nl
[b] Institut de Recherches sur la Catalyse, CNRS, 2 Avenue Albert Einstein, F-69626, Villeurbanne Cedex, France
Fax : (+33)-4-72-44-53-99, e-mail: farrusseng@catalyse.cnrs.fr

Received: June 11, 2004; Accepted: October 15, 2004

Supporting Information for this article is available on the WWW under http://asc.wiley-vch.de or from the author.

Abstract: A new approach to the selection of Heck cross-coupling catalysts and reaction conditions is presented, based on a quantitative structure-activity relationship (QSAR) descriptor set that is coupled to linear and non-linear analysis models. A set of steric and electronic descriptors is defined and calculated. The correlations between ligands, substrates, catalyst precursors and reaction conditions in a dataset of 412 Heck reactions are then analyzed using artificial neural networks, classification tree methods, and linear models. The regression and classification models are powerful, with prediction confidence levels as high as 93%. These models are then used to predict the performance (turnover numbers and turnover frequencies) of 60,000 combinations of virtual catalysts and reaction conditions *in silico*. Furthermore, the models are used to pinpoint mechanistic effects, in this case the important role that small palladium clusters play in Heck cross-coupling. The pros and cons of linear and non-linear models and the application of this approach to the selection of new experiments in high-throughput systems are discussed.

Keywords: artificial neural networks; classification trees; combicat; combinatorial catalysis; Heck reaction; virtual library

Introduction

The end of the 20th century has witnessed a radical change in the value of the basic scientific unit operation – the laboratory experiment. High-throughput experimentation (HTE) and combinatorial chemistry tools have climbed over the science barrier and become an enabling technology.[1,2] Robotic systems can now perform thousands of experiments per day, yielding mind-boggling amounts of experimental results. This fast technological change has also important psychological consequences that must be realized if catalysis chemists are to make the most of these new tools. That is, these tools must be complemented by new modes of operation and, most importantly, new modes of thinking.

In contrast to classical research methods, where scientists decide which experiments to perform based solely on knowledge and 'chemical intuition', HTE techniques enable a search of the catalyst space in a fast and exhaustive manner, but there are two problems: first, the total reaction space is much too large to be fully explored, even with robotic systems, so a choice must be made as to which areas to search.[3,4] Second, today's chemists must face an overwhelming quantity of data, but much of it is perforce 'undesired data', that must be sifted out.[5] This requires both data mining techniques able to extract knowledge from large datasets, and algorithms for optimizing catalysts and process conditions.

One possible approach to minimize the number of samples to be tested while maximizing the chance of finding a 'good catalyst', is to use quantitative structure-activity relationship (QSAR) methods to correlate the input (reaction conditions) and the output data (figures of merit). The model generated can then be used to screen virtual libraries of catalysts in order to select optimized combinations for future experiments.[6–8]

Recently, we examined the application of the QSAR approach to a set of literature data of Heck cross-coupling reactions.[9] Although the reaction mechanism is reasonably well understood, the complexity of the data and the predominance of non-linear effects precluded a full analysis with classical linear methods. Nevertheless, we demonstrated that useful knowledge can be extracted regarding the activity of ligands and solvents.

In this paper, we tackle the problem from a different angle, using non-linear modeling tools. We show that artificial neural networks (ANNs)[10] and classification trees can give very good predictions of the complex da-

  DOI: 10.1002/adsc.200404170

taset. Using these tools, we demonstrate how preselection of catalysts and reaction conditions can be performed from very large libraries. Furthermore, we show that this approach can help pinpoint mechanistic effects by statistical isolation of the key reaction descriptors.

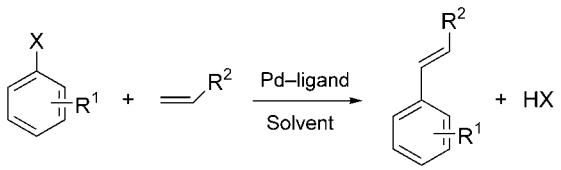
Results

Dataset Construction and Descriptor Selection

A fundamental requirement for meaningful analysis is to have a good dataset. Ideally, comparison should be made between reactions that were performed simultaneously under fully controlled conditions at one location. In practice, such sets are very hard to come by. Here we collected literature data that was in a second step augmented by adding virtual experiments and compounds. Although this set is imperfect, it has two advantages. First, it is accessible (and publishable). Second, the results can lead us to general conclusions on the reaction as the data are collected from different research groups. We compiled experimental data for 412 Heck reactions performed using different reagents, solvents and ligands (Scheme 1).[11–23] The data were arranged in a 412×76 matrix, where each row corresponds to a reaction, the first 74 columns correspond to descriptors, and the last two columns contain the figures of merit, turnover number (TON) and turnover frequency (TOF).

To build a predictive model we need to capture the catalyst patterns of the experiments into numerical data. This transformation involves the selection of steric and electronic descriptors to characterize all the chemical species involved. In other terms, the descriptors define the dimensionality of the space that encompasses all of the 412 Heck reactions analyzed.

The descriptors must fulfill several requirements: They should be simple and fast to calculate, they should capture the structural and electronic variations responsible for catalytic activity, and, finally, they should be chemically meaningful. To be able to employ the same set of stereoelectronic descriptors for all the compounds that bind to the metal atom during the catalytic cycle, we considered all reagents, ligands and solvents as 'ligands' with respect to the metal atom. Among the steric descriptors we calculated are Tolman's cone angle (Θ)[24] and the solid angle (Ω).[25,26] The electronic descriptors included the charge at the ligating atoms, HOMO and LUMO energies, Hammett parameters[27] (for *ortho* and *para* substituents on phenyl rings) and dipole moments. We used 17 descriptors for the aromatic substrate (R^1), 14 for the alkene substrate (R^2), 15 for the ligand, 24 for the solvent, and four additional variables for the temperature, time of reaction, Pd loading, and precursor/Pd ratio. The metal precursors were coded as nominal variables (i.e., variables with a limited range of text values) with 5 modalities {none; $Pd(OAc)_2$; $Pd_2(dba)_3$; CpPd(allyl); $Pd(dba)_2$}. A complete list of the descriptors is given in the Supporting Information.



R^1 = H, OH, CHO, Me, OAc, OBz, NH_2, OMe, NHCOMe, NO_2, CN, COMe, CO_2Me, F, $N(Me)_2$, CF_3

R^2 = CO_2Bu, CO_2Me, Ph, CO_2H, $(CO_2)Et$, $CON(Et)_2$, CN

Scheme 1. The general Heck reaction described by the dataset. Ligands used are monophosphines and monophosphites; solvents are DMF, THF, DMA, dioxane, Et_3N, PhMe, NMP, MeCN, EtCN, PrCN, HMPT, and 1,2-DCE.

The choice of figure of merit depends on the application. Generally, a 'good' catalyst is active, selective, and stable. We used turnover number (TON) and turnover frequency (TOF). These are good indicators of activity and stability, and can be compared across the board as they are calculated per unit catalyst (TON) and per unit time (TOF). The TON and TOF values of the reaction dataset, however, are not normally distributed, and data dispersion is too high to enable good modeling. To obtain more normal distributions for the regression analysis we used log(TON) and log(TOF). For classification, the values were divided into "positive" and "negative" cases, with threshold values of 100 for TON and 10 for TOF (see Figure 1). These thresholds were set *a priori*, as a rough starting point.

To reduce the rank of the problem, we used a two-step process: a relief algorithm[28,29] followed by principal components analysis (PCA). This eliminates descriptors that have low or no correlation with the figures of merit, and discards duplicates (descriptors that are highly correlated with each other). The model becomes robuster and simpler, and the chances of overlearning are reduced. The results were then visualized and a number of uncorrelated descriptors were chosen. Figure 2 shows the correlation between the ligand descriptors. In this radial plot, dots that are close together are correlated. For example, the q_1 and q_2 charge descriptors are strongly correlated, while HOMO is not correlated to S_{occ}.

Regression Analysis

We generated several ANNs with different topologies, and selected two that performed well in predicting the TON and TOF figures of merit (Figures 3**A** and 3**B**, respectively). The TON network had 11 and 3 nodes in the first and second hidden layers. The TOF network had 15

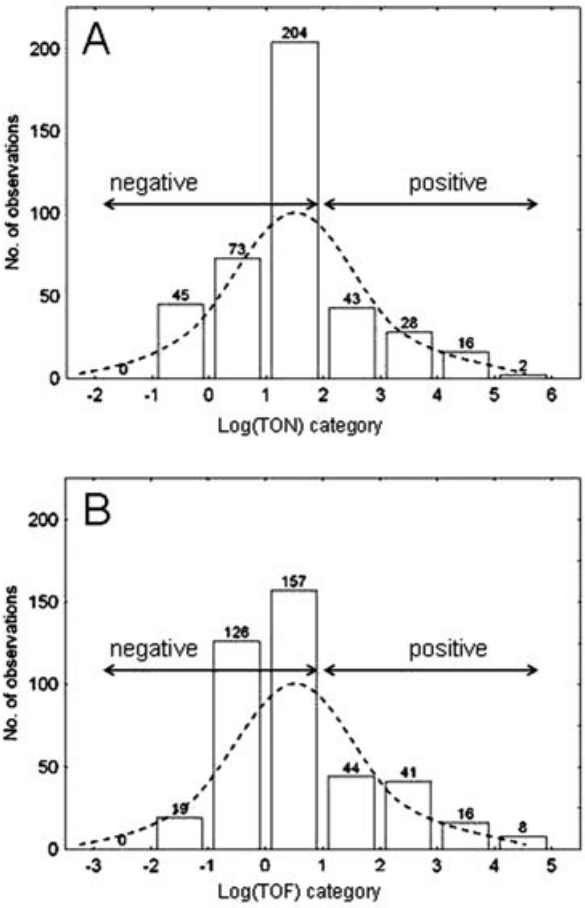


Figure 1. Logarithmic distribution of reactions for figures of merit TON (**A**) and TOF (**B**). The log transformation gives a normal distribution of the cases and is a part of the data pretreatment for the regression analysis. Discretization of samples for classification analysis into positive and negative cases is indicated by arrows. The negative/positive thresholds are TON = 100 and TOF = 10, respectively.

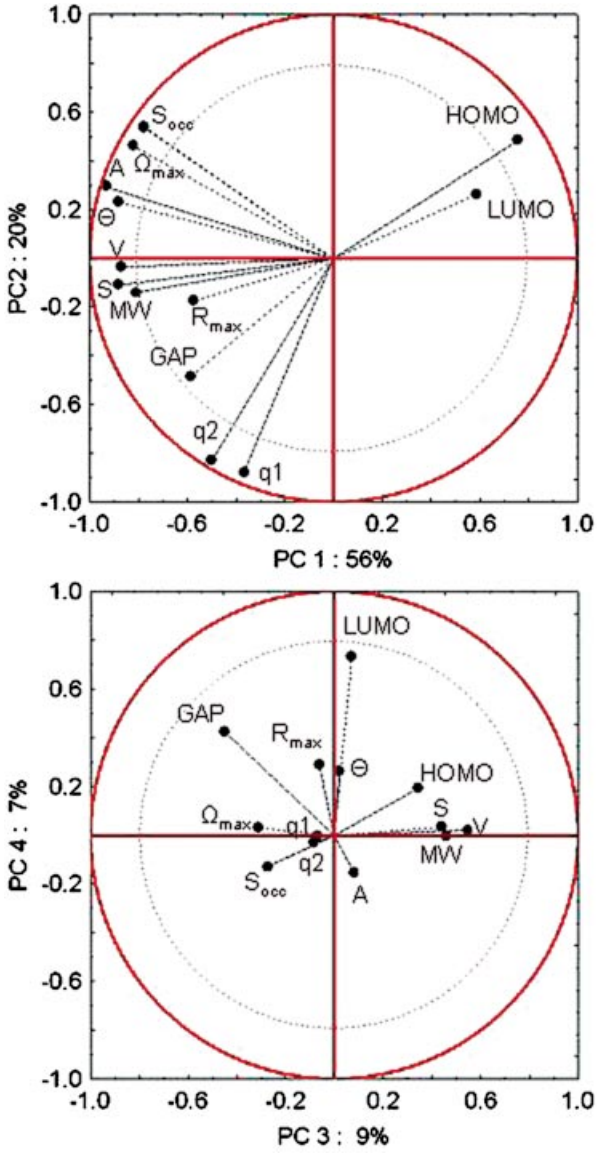


Figure 2. Correlation structures obtained by principal component analysis of steric and electronic ligand descriptors. The figures display the degree of correlation between the descriptors in the first four principal components. The cumulative variation explained by these four PCs is 89%.

and 10 nodes in the first and second hidden layers, respectively. In both cases, the residuals are scattered (see insets), and the errors in the training set ('○' symbols) are comparable to those of the test set ('●' symbols). This shows that there is no overlearning, and the model is sufficiently robust to predict the output of new independent experiments in the reaction space.

The linear regression model, which uses only 21 coefficients (one for each descriptor and one for each modality for the metal precursor) also fits to the TON data (Figure 3**C**). Here, however, the residuals are structured indicating that the model does not account for high-order effects in some of the experiments (*cf.* inset of 3**C** with those of 3**A** and 3**B**). In this respect, the non-linear models are superior. The advantage of the linear model is that it is transparent: you can examine the trends between the descriptors and the TON data by comparing the regression coefficients. Figure 4 shows that only seven of the 21 coefficients are statistically significant, with Pd loading being by far the most significant. The regression coefficient for Pd loading is negative – i.e., the less Pd, the better.[30]

Classification

In classification analysis the variable selection procedure is the same as in regression, but here the reactions are categorized into 'positive' and 'negative' cases according to their catalytic performance (positive cases display TON > 100 and/or TOF > 10, respectively). Here we used ANNs, linear discriminant analysis

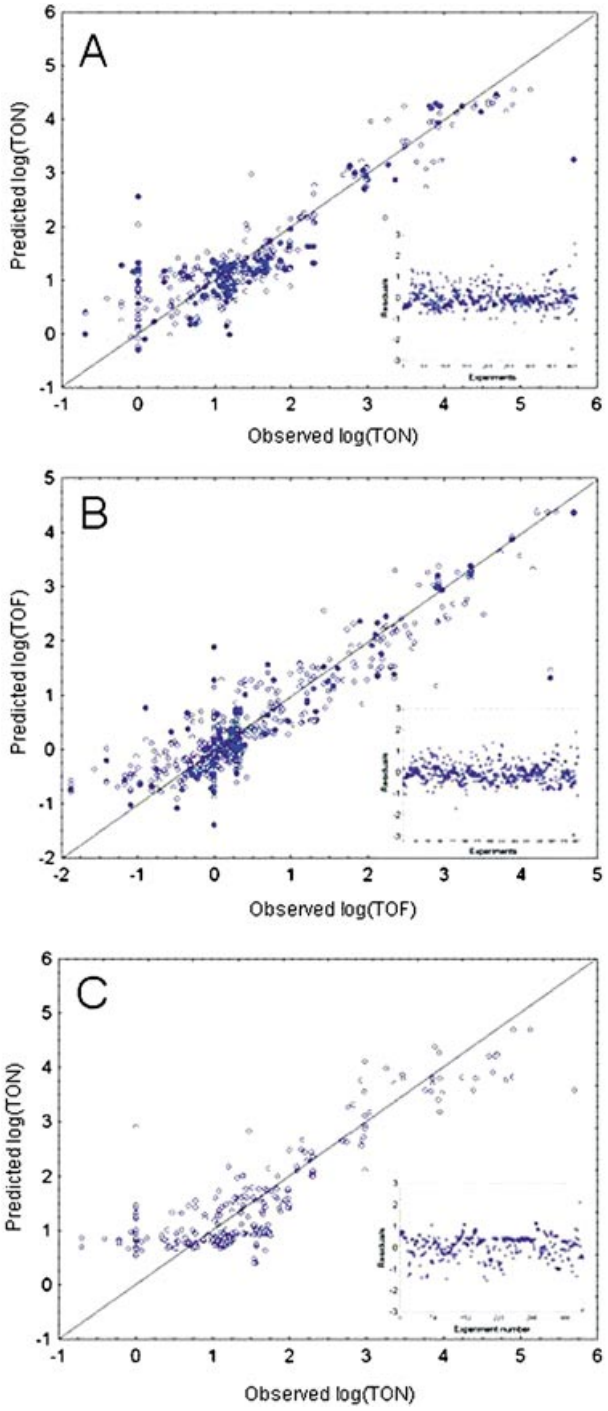


Figure 3. Predicted *vs.* observed TON (**A**) and TOF (**B**) values obtained by NN regression analysis. The inset shows the residuals for each model. Training and test cases are represented by '●' and '○' symbols, respectively. Graph **C** shows the predicted *vs.* observed TON values obtained by multiple linear regression analysis. Note that **A** and **B** show scattered residuals, while **C** shows structured residuals. This indicates that the linear regression model does not account for high-order effects in some of the reactions.

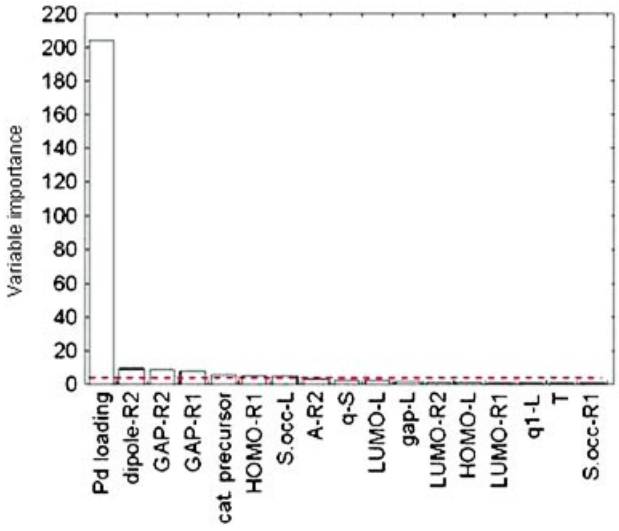


Figure 4. Variable importance (VIP) bar graph obtained by multiple linear regression analysis in the prediction of TON values. Variables above the threshold (red dashed line) are statistically significant. In this case the Pd loading is by far the most significant descriptor.

(LDA) and tree classification methods to model the data.

Figures 5 and 6 show the classification tree structures obtained for prediction of TON and TOF, respectively. The tree structure obtained for classification of experiments according to TON values (Figure 5) has one single split condition which corresponds to a palladium loading threshold of 0.75%. In the whole dataset, there are 104 experiments that fulfill this condition, and 92 of them are positive. On the other hand, 308 experiments have higher Pd loading, and 307 of these are negative. Pd loading can thus be used to distinguish between positive and negative cases. The tree model classifies 104 experiments as positive with 12 false positives, a prediction rate 92/104 (88%). Those values are reported in the confusion matrix (Table 1).

The tree structure obtained for TOF classification (Figure 6) is more complex. It indicates that the time, the palladium loading, and two ligand descriptors (the LUMO energy and the distance between the metal atom and the bulk of the ligand, R_{max}) are the key descriptors in this case. For example, for Pd loading $<$ 0.3% and LUMO $>$ 0.39 eV there are 73 positive cases, of which only 3 are false positives. On the opposite branch, when Pd loading is between 0.3% and 2.15% and the reaction time is $<$ 3.5 h, there are 26 positive cases correctly assigned out of the 27 experiments. All the terminal node numbers are summed and reported in the confusion matrix (Table 1). In total, 102 catalysts are correctly predicted as positive and 295 as negative, with only 7 and 8 cases misclassified as false-positives and false-negatives, respectively. The global prediction rate for positive experiments is 93%.[31]

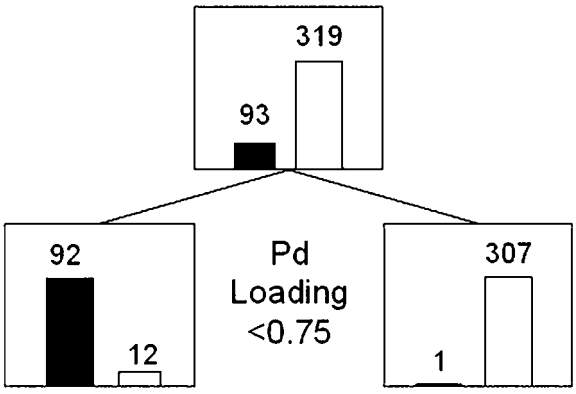


Figure 5. Classification tree structure for TON values. The black and white bars represent positive and negative experiments, respectively. Starting from 93 positive and 319 negative experiments (the threshold is TON = 100) the first node (also called the parental node) is divided into two child nodes according to the most relevant splitting condition (in this case, a Pd loading < 0.75%).

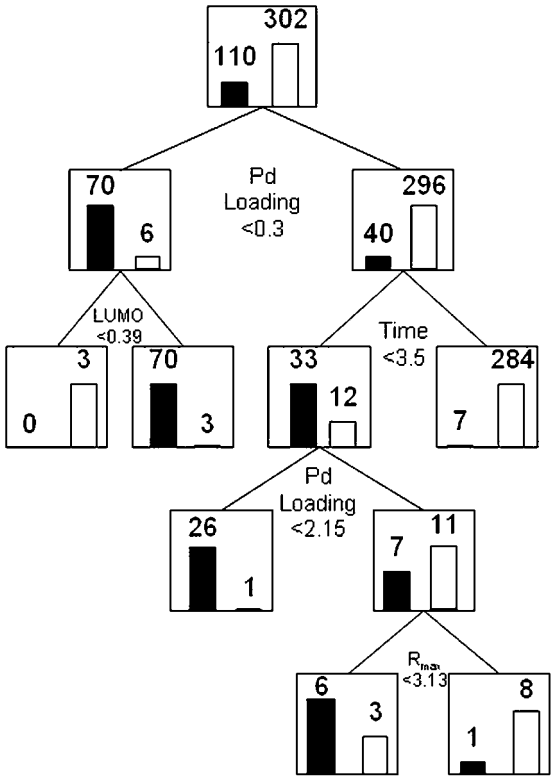


Figure 6. Classification tree structure for TOF values. The first splitting condition is the reaction time, followed by the Pd loading, and the ligand's LUMO energy and R_{max} (the distance between the bulk of the ligand and the metal center).

We then compared the tree results with those obtained from LDA and ANN models. In contrast to the tree method, LDA and ANN are 'black boxes', i.e., they give a final result, but it is difficult to see how the result was obtained. Comparing the values in Table 1, we see that all three methods have similar prediction rates for both TON and TOF. The number of misclassifications in the case of TOF is higher than in the case of TON also using LDA and ANN, in agreement with the more complex tree structure. In each case, the models' robustness was checked using independent tests.

Table 1. Confusion matrix results for classification analyses of TON and TOF values.

		Tree		LDA		ANN	
		true	false	true	false	true	false
TON	positive	92	12	92	13	89	6
	negative	307	1	306	1	313	4
TOF	positive	102	7	92	17	91	29
	negative	295	8	285	18	273	19

A variable importance analysis of the descriptors used in tree and the LDA models showed, once again, that the Pd loading is the most relevant descriptor for both TON and TOF. The time, the metal precursor and various ligand parameters also have a significant effect (Figure 7).

Screening 60,000 Heck Reactions using a Virtual Combinatorial Library

The real power of the ANN regression models lies in their ability to quantitatively predict the figures of merit for virtual libraries. Instead of synthesizing and testing huge numbers of ligands and/or reaction conditions, we can preselect *in silico* likely candidates by generating combinations of catalysts and reaction conditions and using an ANN to predict the TONs and TOFs. First, we designed 61 new monophosphine ligands by using several building blocks available from commercial catalogues.[6] It is important to verify that the descriptor values for the new (virtual) ligands fall within the range of values covered by the original data that was used to train the network. To do this, we performed a PCA on the ligand descriptors and compared the scores for the real and the virtual ligands. The results show that the virtual ligands are within the scope of descriptor values of the real ligands (see Figure 8A, inset). We then compiled a virtual library of 60,000 Heck reactions, containing all possible combinations of these 61 ligands with 4 olefins, 4 aryl halides, 5 catalyst precursors, 4 solvents, and 3 palladium concentrations. The substrates and reaction conditions were chosen from the original matrix shown above.[32]

Figure 8A shows the distribution of the virtual library (red dots) compared to the original 412 reactions (blue dots) in the overall reaction parameter space. The 60,000 new combinations are clustered in two clouds, enabling us to sample many points around the original ex-

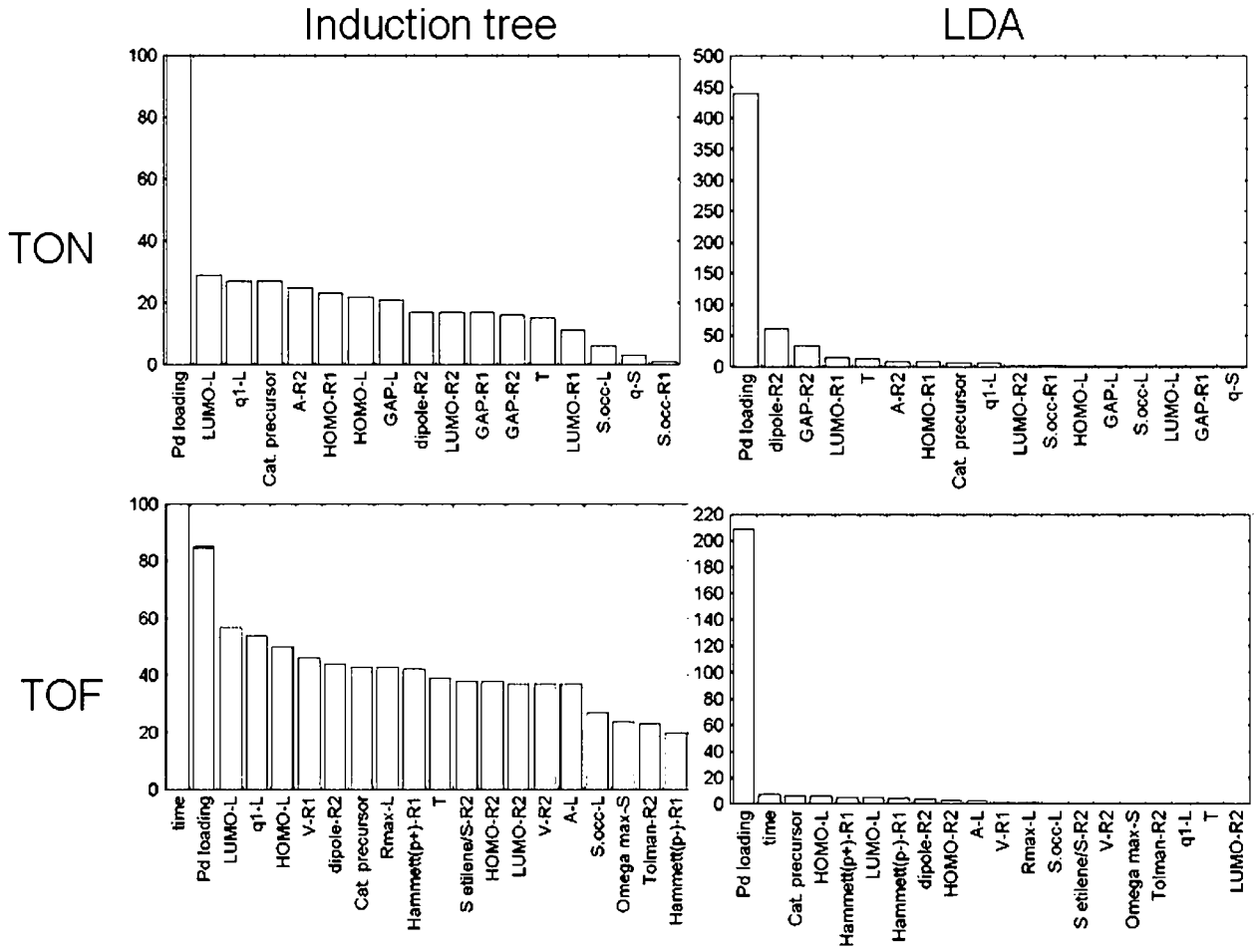


Figure 7. Variable importance (VIP) bar graphs obtained by induction tree methods and linear discriminant analysis (LDA). For the prediction of TON values, the Pd loading is the most significant descriptor while for the prediction of TOF, time and ligand electronic descriptors are also significantly involved in the model.

periments. The sampling of the reaction space is enlarged, while remaining in the same region. The distribution of the TONs in the virtual library is shown in Figure 8B. 4% of the combinations are predicted to have TON > 10,000.

Using these data, it is also possible to visualize directly the relationship between the descriptors and the figure of merit. This is very important in high-throughput experimentation, where the limiting factor is often the choice of reactions to be studied in the next generation, rather than the actual duration of the experiments. Figure 9 shows a contour plot of the predicted TON for the 60,000 reactions in the virtual library *vs.* the two first principal components (the first correlated mainly with the Pd loading and the R^2 electronic descriptors and the second with the ligand's electronic descriptors). In this way, a simple and fast selection of the most promising trends in catalysts and reaction conditions can be made.

Discussion

Linear *vs.* Non-Linear models – Which is Better?

The Heck reaction, like many other catalytic and natural systems, shows a complex non-linear behavior. ANNs describe this type of data better than simple linear regression models. However, this description comes with a heavy price tag: the networks are opaque, and it is difficult to perceive any trends in the underlying chemistry. Linear regression, on the other hand, is transparent, enabling the chemist to compare the model to chemical intuition. In this specific example, because of the primary importance of the Pd loading, the simple linear model is not too bad. However, we expect that in cases where non-linear effects predominate (as in many chemical systems), the prediction ability of the linear models would deteriorate. This is especially true if one studies the reaction system as a whole, rather than isolated effects.

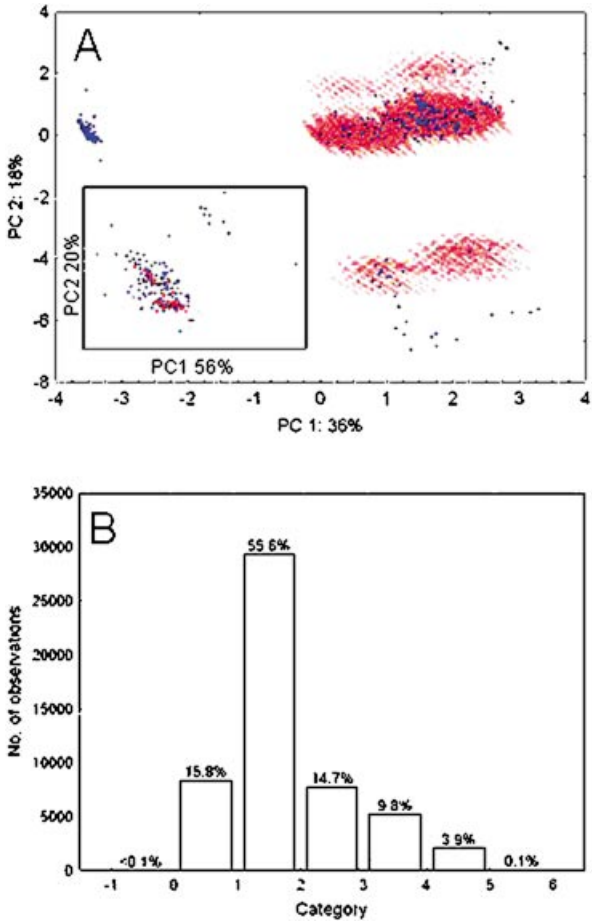


Figure 8. Principal component analysis (**A**) of virtual (red) and real (blue) catalytic Heck reactions (all reaction descriptors have been included). This plot reveals the portion of the space that is explored using the virtual catalysts. The inset shows the PCA of virtual (red) and real (blue) ligands. Compounds are plotted against the first two principal components calculated for the ligand descriptors. This validation is performed to ensure that descriptors of the learning and prediction sets lie in the same ranges. This condition allows intrapolation, thus a correct prediction on the virtual catalysts. Graph **B** shows the TON predicted values for the virtual catalysts (note the similarity between this distribution and Figure 1**A**).

Both types of analysis methods are affected by the quality of the dataset. We used literature data, obtained in different laboratories in the past 20 years. One cannot expect that all 412 reactions were performed under controlled conditions. The advantage of these data, however, is that they are in the public domain.[33] Even with this incomplete and noisy dataset, the non-linear models (both ANNs and classification trees) perform well. The quality of prediction in both cases is $>90\%$, which makes this approach a viable option for optimizing catalytic reactions.

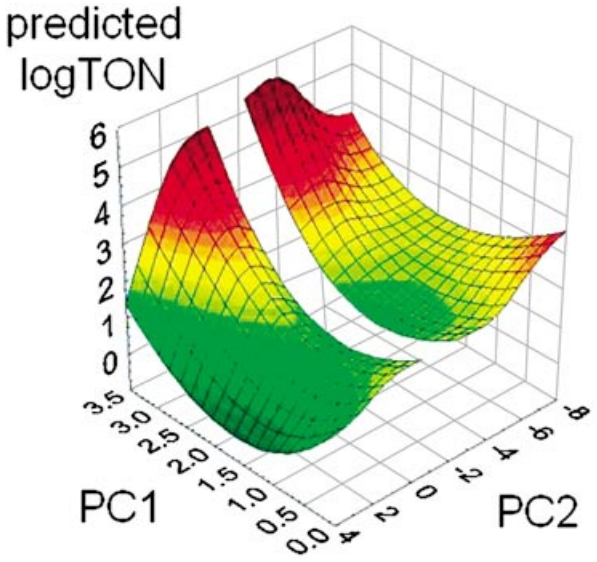


Figure 9. Predicted TON values for 60,000 virtual cross-coupling reactions are plotted *versus* the first two PCs calculated for all the reaction descriptors. The first PC is correlated mainly with the Pd loading and the electronic descriptors of the organic residue on the alkene, R^2, while the second represents mainly the ligand's electronic descriptors.

From Statistical Results to Mechanistic Conclusions

Our results show that Pd loading is the most significant descriptor in determining catalytic TON and TOF. Considering the diversity of the data set, this is a statistically significant and general result, and the negative correlation suggests that something is fishy. There are two possible explanations, and both are to do with the formation of palladium clusters. As shown by de Vries,[34] Reetz,[35] and us,[36] nanometric Pd clusters are good catalysts for many C–C coupling reactions, including the Heck reaction. We maintain that every Heck reaction that is catalyzed by 'homogeneous Pd complexes' also involves Pd clusters. In most cases, the reaction stops after the catalyst deactivates as micrometric Pd black, which means that Pd clusters must form somewhere during the deactivation process. These clusters can enter the catalytic cycle, either as catalysts or as 'reservoirs' for monoatomic palladium.[37] If clusters are involved, one would indeed expect an inverse relationship between Pd concentration and TON, for example, because a high concentration of clusters would lead to faster deactivation *via* Pd black.

Applying Virtual Screening to High-Throughput Experimentation

The above methods can be used to complement high-throughput experimentation (HTE) workflows in catalysis research. In the 1990s, the limiting factor was the

number of experiments that could be performed in the lab. Today, automated reaction set-ups are common laboratory tools. Chemists can now perform more experiments, but they cannot test all of the possible combinations, and even examining yesterday's output becomes a problem if a robot performs 1000 reactions a day. The bottleneck is now designing the next generation of experiments. Combining virtual libraries, selection algorithms, and HTE systems can solve this problem.

All the possible catalysts and conditions for a specific reaction (solvent, temperature, substrate type and concentration, etc.) form together a multidimensional virtual space. Most of the points in this space pertain to "bad" combinations, so a random choice process is unlikely to hit a "good" spot. Instead, by using the virtual library concept outlined above, combined with a selection algorithm, it is possible to concentrate on those regions of the space where "good" catalysts and reaction conditions are likely to be found. For example, classification can function as a coarse filter, to discriminate between "good" and "bad" experimental candidates.

Conclusions

In this study we have shown how to model and predict reaction data with a combination of regression and classification statistical methods. Due to the complex nature of the cross-coupling reaction under study, we found out that non-linear methods, namely neural networks and classification trees, perform better than simple linear regression techniques in modeling and predicting the catalytic activity. These results can be easily generalized to other homogeneously catalyzed reactions. Once a regression or a classification model is built, the performance of virtual catalysts can be predicted. In this way, one can decide which regions of the catalyst space are worth to be sampled experimentally. The power of this approach lies in the possibility of driving the future experiments by screening *in silico* thousands of virtual catalysts in search for the optimal combinations, discarding those which are predicted to perform poorly, and thus saving time and materials. Combining this approach with state-of-the-art HTE and combinatorial chemistry tools will lead chemists to new and powerful homogeneous catalysts.

Experimental Section – Computational Methods

Geometries of molecules were first subjected to a fast minimization using the MMFF94 molecular mechanics force field and then further optimized using the PM3 semi-empirical method within the Spartan program package.[38] The minimized structures were then used for the calculation of all the descriptors. The Steric program[39] was employed to calculate the solid angle of all compounds (ligands, reagents and solvent molecules) attached to the metal center as well as its related properties. Hammett parameters were used to characterize the electronic effect of *para* and *ortho* substituents on the phenyl rings. Additional empirical and theoretical scales were used to characterize solvent activities. A complete list of the descriptors is included in the Supporting Information. The PCA and PLS treatment of this dataset is described in detail in our previous paper.[9]

Regression Analysis

Two different regression methods were used in this study: artificial neural networks (ANN) and multiple linear regression analyses.[7] ANNs attempt to mimic the fault-tolerance and capacity to learn of biological neural systems by modeling the low-level structure of the brain. To capture the essence of biological neural systems, an artificial neuron receives a number of inputs, either from original data or from the output of other neurons in the network.

A single artificial neuron can be implemented in many different ways. The general mathematic definition is showed in Eq (2):

$$y(x) = G\left(\sum_{i=0}^{n} w_i x_i\right) \quad (2)$$

where x is a neuron with n input dendrites $(x_0 \ldots x_n)$ and one output axon $y(x)$ and where $w_0 \ldots w_n$ are weights that determine how much the inputs should be weighted. G is an activation function (i.e., a sigmoid or a hyperbolic tangent), based on the sum of the n inputs, which weights how powerful the output should be from the neuron.

ANNs have two different phases: a training phase and an execution phase. In the training phase the ANN is trained to return a specific output when given a specific input (variables describing the system under study). In the execution phase the ANN is fed with new cases (in this study, the virtual catalysts) and returns an output (prediction) on the basis of the network generated in the previous phase. The training process can be seen as an optimization problem, where we wish to minimize the mean square error of the entire set of training data.

In this study, we used the multilayer perception (MLP) topology; neurons are grouped into one or two hidden layers so that each input of a neuron is composed of the outputs of the neurons of the previous layer. The network was trained using the back propagation principles with the conjugated gradient descent, as implemented in the Statistica™ software package.

When an ANN is used for regression analysis, confidence levels determine how the neural network is able to predict a quantitative value of performance (output value) for each input case. The efficiency of the network is evaluated by means of the standard deviation ratio, which is the ratio of the prediction error standard deviation to the original output data standard deviation. The lower the ratio, the better is the prediction. The error is evaluated by the root mean square of the network errors on individual cases. The individual errors are generated by the network error function, which is a function of the observed and expected output neuron activation levels (sum-squared). To prevent overfitting, the dataset used for the learning stage was split into two groups, namely the training and ver-

ification datasets, that correspond to 50% and 25% of the whole dataset, respectively. The model assessment was performed on the remaining 25% test data.

Multiple linear regression analysis fits a response variable y as a linear combination of multiple x variables by the method of least squares. The goal of this procedure is to fit a straight line through the points in a multivariable space. Specifically, the program will compute a line so that the squared deviations of the observed points from that line are minimized. In general, multiple regression procedures will estimate a linear equation of the form

$$y = b_0 + b_1x_1 + b_2x_2 + \ldots + b_nx_n \quad (3)$$

where b_i are the regression coefficients and represent the independent contributions of each independent variable to the prediction of the dependent variable. Another way to express this fact is to say that, for example, variable x_i is correlated with the y variable, after controlling for all other independent variables. This type of correlation is also referred to as a partial correlation.

Classification Methods

Three different classification methods were used: artificial neural networks (ANN), linear discriminant analyses and induction tree according to the C&RT algorithm.[10]

When an ANN is used for classification, confidence levels determine how the neural network assigns the input cases to the classes. In the case of two-class classification (positive/negative), the output class is indicated by default by a single output neuron, with high output (e.g., 1) corresponding to one class and low output (e.g., 0) to the other. In the present work, the "accept" and "reject" thresholds were set as equal, forcing all cases to be classified either as "positives" or as "negatives".

A linear discriminant analysis (LDA) produces a set of coefficients defining the single linear combination of descriptors that best differentiates negative (low TON or TOF) from positive (high TON or TOF) cases. A score for each experiment on the linear discriminant function is computed as a composite of each experiment on the discriminant descriptor, weighted by the respective discriminant function coefficients. The predicted classification of each experiment as "negative" or "positive" is made by simultaneously considering the experiment scores on the discriminant variables. If the decision equation (e.g., of the form $aA + bB + cC - d$) is less than or equal to the negative/positive threshold, the case is considered "negative", else it is "positive".[40] Multivariate F-tests were employed in order to determine whether there are any significant differences (with regard to all variables) between groups. A cross-validation procedure was used to warn against overlearning.

Classification trees are used to predict responses on a categorical dependent variable by identifying discriminant sets of variables. In contrast to the "black-box" ANN, the output of classification tree analysis provides a clear scheme on the model and cases assignation *via* sets of if-then logical conditions (splits). These splits are selected one at time, starting with the split at the root node, and continuing with splits of resulting child nodes until the splitting stops. The child nodes that have not been split become terminal nodes. We used the C&RT algorithm as implemented in Statistica software for an exhaustive search of univariate splits which allows operating for categorical or ordered predictor variables. All possible splits for each predictor variable at each node are examined to find the split producing the largest improvement in goodness of fit (using the Gini measure). When univariate splits are performed, the predictor variables can be ranked on a 0–100 scale in terms of their potential importance in accounting for responses on the dependent variable.

The results for all classification analysis can be expressed in a 2×2 contingency table (also called a confusion matrix) where counts of positive classification and misclassification are reported. The loss coefficient gives the relative "cost" of the two possible misclassifications (false-positive *versus* false-negative). A loss coefficient of 1.0 indicates that the two classes are equally important whereas a loss coefficient above 1.0 indicates that it is relatively more important to correctly recognize true positive cases, even at the expense of misclassifying more false negative cases. In all studies we used a loss coefficient of 1.0 in order to have a fair model comparison.

Supporting Information Available

Complete list of the descriptors used in the data analysis and structures of the building blocks used to construct the 61 monophosphine ligands for the virtual library.

References and Notes

[1] P. P. Pescarmona, J. C. van der Waal, I. E. Maxwell, T. Maschmeyer, *Catal. Lett.* **1999**, *63*, 1.
[2] M. T. Reetz, *Angew. Chem. Int. Ed.* **2001**, *40*, 284.
[3] G. Rothenberg, H. F. M. Boelens, D. Iron, J. A. Westerhuis, *Catalysis Today* **2003**, *81*, 359.
[4] G. Rothenberg, H. F. M. Boelens, D. Iron, J. A. Westerhuis, *Chim. Oggi* **2003**, *21*, 80.
[5] J. M. Caruthers, J. A. Lauterbach, K. T. Thomson, V. Venkatasubramanian, C. M. Snively, A. Bhan, S. Katare, G. Oskarsdottir, *J. Catal.* **2003**, *216*, 98.
[6] L. Eriksson, E. Johansson, *Chemom. Intell. Lab. Sys.* **1996**, *34*, 1.
[7] M. Karelson, V. S. Lobanov, A. R. Katritzky, *Chem. Rev.* **1996**, *96*, 1027.
[8] G. R. Famini, L. Y. Wilson, *J. Phys. Org. Chem.* **1999**, *12*, 645.
[9] E. Burello, G. Rothenberg, *Adv. Synth. Catal.* **2003**, *345*, 1334.
[10] D. L. Massart, B. G. M. Vandeginste, L. M. C. Buydens, S. De Jong, P. J. Lewi, J. Smeyers-Verbeke, *Handbook of Chemometrics and Qualimetrics*, Elsevier, Amsterdam, **1997**.
[11] I. P. Beletskaya, A. N. Kashin, N. B. Karlstedt, A. V. Mitin, A. V. Cheprakov, G. M. Kazankov, *J. Organomet. Chem.* **2001**, *622*, 89.
[12] C. B. Ziegler, R. F. Heck, *J. Org. Chem.* **1978**, *43*, 2941.
[13] W. A. Herrmann, C. Brossmer, K. Oefele, M. Beller, H. Fischer, *J. Mol. Catal. A: Chem.* **1995**, *103*, 133.
[14] A. F. Littke, G. C. Fu, *J. Org. Chem.* **1999**, *64*, 10.
[15] A. F. Littke, G. C. Fu, *J. Am. Chem. Soc.* **2001**, *123*, 6989.

[16] A. M. Magill, D. S. McGuinness, K. J. Cavell, G. J. P. Britovsek, V. C. Gibson, A. J. P. White, D. J. Williams, A. H. White, B. W. Skelton, *J. Organomet. Chem.* **2001**, *617*, 546.
[17] F. Miyazaki, K. Yamaguchi, M. Shibasaki, *Tetrahedron Lett.* **1999**, *40*, 7379.
[18] A. Spencer, *J. Organomet. Chem.* **1983**, *258*, 101.
[19] A. Spencer, *J. Organomet. Chem.* **1984**, *270*, 115.
[20] J. P. Stambuli, S. R. Stauffer, K. H. Shaughnessy, J. F. Hartwig, *J. Am. Chem. Soc.* **2001**, *123*, 2677.
[21] G. P. F. van Strijdonck, M. D. K. Boele, P. C. J. Kamer, J. G. de Vries, P. van Leeuwen, *Eur. J. Inorg. Chem.* **1999**, 1073.
[22] C. L. Yang, H. M. Lee, S. P. Nolan, *Org. Lett.* **2001**, *3*, 1511.
[23] K. H. Shaughnessy, P. Kim, J. F. Hartwig, *J. Am. Chem. Soc.* **1999**, *121*, 2123.
[24] C. A. Tolman, *Chem. Rev.* **1976**, #76#*77*, 313.
[25] D. White, B. C. Tavener, P. G. L. Leach, N. J. Coville, *J. Organomet. Chem.* **1994**, *478*, 205.
[26] D. White, B. C. Taverner, P. G. L. Leach, N. J. Coville, *J. Comput. Chem.* **1993**, *14*, 1042.
[27] C. Hansch, A. Leo, R. W. Taft, *Chem. Rev.* **1991**, *91*, 165.
[28] K. Kira, L. A. Rendell, *Traditional methods and a new algorithm*. In: *Proceedings of the Ninth National Conference on Artificial Intelligence*. AAAI Press (American Association for Artificial Intelligence), **1992**, Amherst, MA, pp. 129–134.
[29] K. Kira, L. A. Rendell, *A practical approach to feature selection*. In: *Proceedings of the Ninth International Workshop on Machine Learning*. Morgan Kaufmann, Aberdeen, **1992**.
[30] Because we could not invert the dataset matrix, we were not able to perform a similar linear regression analysis for the TOF.
[31] Cross-validation was used to verify that the model is not overlearning.
[32] To obtain reliable regression predictions of virtual experiments, descriptor values should fall within the same range of the learning dataset.
[33] Working in collaboration with industry, on large catalytic datasets obtained using high-throughput robotic systems, we can confirm that the model's performance improves under controlled conditions.
[34] A. H. M. de Vries, J. M. C. A. Mulders, J. H. M. Mommers, H. J. W. Henderickx, J. G. de Vries, *Org. Lett.* **2003**, *5*, 3285.
[35] M. T. Reetz, E. Westermann, *Angew. Chem. Int. Ed.* **2000**, *39*, 165.
[36] M. B. Thathagar, J. Beckers, G. Rothenberg, *J. Am. Chem. Soc.* **2002**, *124*, 11858.
[37] G. Rothenberg, S. C. Cruz, G. P. F. van Strijdonck, H. C. J. Hoefsloot, *Adv. Synth. Catal.* **2004**, *346*, 467.
[38] Spartan is distributed by Wavefunction Inc., 18401 Von Karman Ave., #370, Irvine, CA 92612 USA, © Wavefunction Inc.
[39] Copyright 1994,1995 C. T. Steric, All rights reserved.
[40] Here *a*, *b*, and *c* are the corresponding linear discriminant function coefficients of descriptors *A*, *B*, and *C*, respectively, and *d* is the "cut point" on the discriminant function for separating the two classes.

