Valero Cubas, S.; Argente, E.; Botti V.; Serra Alfaro, JM.; Serna Merino, PM.; Moliner Marin, M.; Corma Canós, A. (2009). DoE framework for catalyst development based on soft computing techniques. Computers & Chemical Engineering. 33(1):225-238. https://doi.org/10.1016/j.compchemeng.2008.08.012

New DoE Framework for Catalyst Development based on Soft Computing Techniques

S. Valero [a,1,*], E. Argente [a], V. Botti [a], J. M. Serra [b], P. Serna [b,2], M. Moliner [b] and A. Corma [b]

[a] *Departamento de Sistemas Informáticos y Computación, Universidad Politécnica de Valencia, 46020 Valencia, Spain*

[b] *Instituto de Tecnología Química, Universidad Politécnica de Valencia, 46020 Valencia, Spain*

Abstract

A new experimental design framework has been developed for the discovery and optimization of catalytic materials when exploring a high-dimensional space. This is based on a new Soft Computing architecture in which neural networks and a genetic algorithm are combined for optimizing the discovery of new materials and process conditions in catalytic reactors at industrial scale. Considering the high temporal and financial costs required for synthesizing and testing potential solid catalysts, the application of Soft Computing techniques in this field appears as an interesting alternative to reduce the number of experiments. The proposed Soft Computing framework has been employed to optimize a hypothetical function based on the modelled behaviour of multi-component catalysts explored in the field of combinatorial catalysis. Moreover, this experimental design framework has been

applied to a problem with industrial interest such as the optimization of a Ti-silicate catalyst for the epoxidation ofolefins.

Key words: Soft Computing, Genetic Algorithms, Neural Networks, High-throughput, Combinatorial Chemistry

1 Introduction

The chemical industry is continuously searching for new efficient catalysts that would enable to reduce operating costs, while decreasing resid. This involves in most cases a real task force since we are dealing with high-dimensional spaces with large number of samples to be prepared, characterized and tested. Recently, high throughput experimentation techniques are being used in heterogeneous catalysis thought a methodology, called Combinatorial Chemistry, in which large diversities of materials are prepared, processed and tested in parallel. Nevertheless, the experimentation of a very large number of materials involves high financial cost that justifies to take into consideration software techniques, such as artificial intelligence, in order to reduce the number of samples and to better understand the relationships among data.

We have developed a new Design of Experiments (DoE) framework for the discovery and optimization of catalytic materials when exploring a high-dimensional space. This framework allows to use the knowledge extracted from the previous experimentation in the design of the new subset of catalysts to be exper-

* Corresponding author: svalero@dsic.upv.es

[1] Financial support from Spanish government under TIC2003-07369-C02-01 is gratefully acknowledge

[2] Partially supported by FPU AP2003-4635

imentally screened in next optimization iteration. It is based on a new soft computing architecture in which neural networks and a genetic algorithm are combined for optimizing the discovery of new materials. In this architecture an artificial neural network (ANN) is employed as an approximated model for fitness evaluation; and a genetic algorithm (GA) tries to find the optimal solution by investigating several catalysts simultaneously. This constitutes a novel approach, which is based on the previous good results obtained in this area with the application of ANNs (Corma et al. (2002a); Serra et al. (2003b,a); Moliner et al. (2005)).

In section 2, a deeper description of the combinatorial chemistry problem is given. A revision of the state of the art in artificial neural networks and genetic algorithms is carried out in section 3, making emphasis in the latest works in the combinatorial chemistry field. In section 4, the soft computing framework that we propose is detailed, explaining its five main steps: (i) setting up process (in which soft computing parameters are set and a suitable ANN model is built); (ii) ANN re-training (in which ANN model is improved with new experimental data); (iii) GA operators (in which the candidates for the new generation are designed by the GA operators); (iv) pre-screening (where the number of samples to be finally tested is reduced) and (v) experimental testing. It should be pointed out that the genetic algorithm presented here employs a new developed codification which allows dealing with optimizations considering simultaneously complex catalyst formulations and different synthesis/testing conditions as variables. Moreover, in section 5 some application examples are given, including results with an hypothetical function and a real industrial problem.

2 Problem Description

One of the main objectives of the chemical industry is the design and synthesis of a more active and selective catalysts that will allow chemical processes to be more efficient and environmentally friendly from the economy, security and versatility process point of view. However, the discovery of new catalysts still is in many cases an arduous and rather unpredictable trial-and-error process. Moreover the tendency today is to consider the chemical process in a global manner, trying to optimize simultaneously the catalyst and reaction conditions. Therefore, it is worth studying the composition of the catalyst and the reaction conditions that will boost its performance.

Traditionally, the processing and analysis of the experimental results from characterization and catalytic testing was normally carried out by the researchers, who applied previous experiences, fundamental knowledge and intuition in order to (i) design new catalyst libraries; and (ii) analyze the data, and establish relationships between the different experimental results. Nowadays, the new development of high throughput experimentation (HTE) techniques in the frame of heterogeneous catalysis (Senkan (1998, 2001); Derouane (2002)) is enabling the screening of large number of new materials and, therefore, it is increasing exponentially the number of catalytic data, derived from the parallel synthesis, characterisation and catalytic testing. These HTE techniques are employed in a new methodology, called Combinatorial Chemistry, in which large diversities of materials are prepared, processed and tested in parallel.

Software techniques can help to the Combinatorial Chemistry methodology

with (i) the efficient administration and schedule of large amounts of experimental data, (ii) the comprehension and modelling of the organised data and (iii) the global search strategy to optimise the catalytic performance.

A crucial item in combinatorial catalysis is how to design the experiments in order to explore and optimize the high-dimensional solution space with the minimum costs (number of experiments). The techniques employed for experimental design are: (i) statistics procedures like factorial designs (Montgomery (2001)), (ii) deterministic optimization procedures like simplex, holographic search (Végvári et al. (2003)) or split & pool (Sun et al. (2002); Aramendía et al. (2002)) and (iii) stochastic procedures like simulated annealing or genetic algorithms (GA). Stochastic procedures are well-suited procedures for the optimization of multi-dimensional problems, being the application of genetic algorithms fruitful in the discovery of new heterogeneous catalysts (Kirsten and Maier (2004)).

On the other hand, the large number of variables involved and the application of complex optimization algorithms for the experimental design makes difficult the direct human interpretation of data derived from high throughput experimentation. Data mining techniques allow to analyze thoroughly raw multidimensional data (Corma et al. (2002a); Gedeck and Willett (2001); Weaver (2004); Wang et al. (2001); Yamada et al. (2001)) in such a way that knowledge can be systematically extracted, establishing multifactor relationships and patterns amongst input variables (catalyst composition, preparation and reaction conditions), output variables (catalyst characterization and catalytic performance) and also theoretical parameters concerning the catalyst components. Hereafter, this available *knowledge* could be applied to design the new subset of materials to be screened in a more intelligent and rational

way (Corma et al. (2002a); Klanner et al. (2004); Omata et al. (2004)). Several data mining techniques have been applied in combinatorial chemistry like clustering models and non-linear regression models, statistical models, association rules and decision trees, rule induction, Kohonen and artificial neural networks (ANNs).

A further-step optimization approach (Serra et al. (2003c); Klanner et al. (2003); Gilardoni et al. (2003)) is the combined use of high multidimensional optimization algorithms with prediction models obtained by data mining, in such a way that the knowledge extracted from all the previous experimentation can be applied in the design of the new subset of catalysts to be screened in the next optimization cycle.

In the following section, these specific artificial intelligent techniques, and more specifically artificial neural networks and genetic algorithms are discussed.

3 State of the art

Many efforts have been done in the development and optimization of several artificial intelligent techniques in combinatorial catalysis, in order to allow the extraction of information and knowledge from high-throughput experimentation raw data, establishing relationships and patterns between the input and output variables. Among those AI approaches, both data mining and soft computing techniques are mainly employed.

Data mining techniques have been applied (Gedeck and Willett (2001); Wang et al. (2001); Yamada et al. (2001); Rajan et al. (2001)) in order to find relationships and patterns between the input and output data derived from

accelerated experimentation.

Regarding soft computing, it is a collection of methodologies that deal with tolerance for imprecision, uncertainty and partial truth in order to achieve tractability, robustness and low solution cost (Zadeh (1994)). The principal constituents of soft computing are fuzzy logic, neural computing, evolutionary computation, machine learning and probabilistic reasoning. What is particularly important about soft computing is that it facilitates the use of those techniques in combination, leading to the concept of hybrid intelligent systems.

In this paper we present a novel soft computing framework for experimental design in which both artificial neural networks and genetic algorithms are combined in order to help in the research of new catalysts.

3.1 Artificial Neural Networks

Artificial neural networks (Bishop (1996); Ripley (1996)) are high performance non-linear analytical tools, which are capable of establishing the relationship between the input/output data without prior knowledge of the correlation between the variables involved in the system. They consist of a number of artificial neurons inter-connected together by synaptic weights to form a network, analogously to biological neurons. The basic unit of a neural network is the neuron, or node, composed of: a set of connections or inputs, $x_j(t)$, each of which is characterized by a synaptic weight w_{jk} , that represents the intensity of interactions between each neuron j of a previous layer and the actual neuron k; a propagation rule (1), which determines the effective input of neuron k from all individual inputs to this neuron; an activation function F_k, that

determines the output y_k (2) of neuron k by means of its level of excitation; and an additional external input, called polarization or bias (b_k), which increases or decreases the excitation threshold of the neuron. The most common activation functions are sigmoidal functions, such as logistic (3) and tangential (4), which present an equilibrium between lineal and non-lineal behaviour. Formulaes associated to a neuron are as follow:

$$S_k = \sum_{j=1}^{n} w_{jk} * x_j + b_k \tag{1}$$

$$y_k = F_k(S_k) \tag{2}$$

$$F_k(S_k) = 1/(1 + e^{-S_k}) \tag{3}$$

$$F_k(S_k) = (e^{S_k} - e^{-S_k})/(e^{S_k} + e^{-S_k}) \tag{4}$$

Two important features of neural networks are the ability to supply fast answers to a problem and the capability of generalizing their answers, providing acceptable results for unknown patterns. In this way, they need to learn about the problem under study and this learning is commonly named the training process. During this training process, neural networks are supplied with a set of samples belonging to the problem domain and they establish mathematical correlations between the samples (Ripley (1996)). A large quantity of information and time are required for analysis and processing. Supervised learning consists of supplying the neural network with training patterns having information about both input and output desired values. During the training process, the neural network will fit its neurone weights to minimise the error between the output calculated by the neurone and the desired value. On the other hand, non-supervised learning consists of supplying the neural network with only the input values of patterns, without the desired output values.

Then, the neural network has to classify the inputs and outputs on the basis of its similarity with other inputs.

One of the most well-known structures of neuronal networks for supervised learning is the multilayer perceptron, which is generally used for classification and prediction problems. In the multi-layer perceptron, neurons are grouped into layers or levels, so each input of a neuron is compose of the outputs of the neurons of the previous level, except for the neurons in the input layer, which have as input values those ones belonging to the problem to consider. The number of nodes at the input and output layers are determined by the problem features. However, the number of hidden layers, and even the number of nodes in each of these layers is unpredictable, so it is necessary to evaluate different structures to establish the neuronal network topology most suitable for the problem under study. An example of a multilayer perceptron with one input layer, two hidden layers and one output neuron can be described by the equation (5) where n is the number of nodes at the input layer, K and L are the number of hidden nodes and $f(\cdot)$ is the activation function.

$$y = \sum_{l=1}^{L} v_l f(\sum_{k=1}^{K} w_{kl}^{(2)} f(\sum_{i=1}^{n} w_{ik}^{(1)} x_i)) \tag{5}$$

Neural networks have shown to be effective tools for function approximation (Jin (2005)). Both feed-forward multi-layer perceptrons and radial-basis-function networks have widely been used. Moreover, an ANN model is preferred if the input space (design space) is high-dimensional and the number of samples is limited. It is recalled that to estimate the unknown parameters of second-order polynomial model, at least $(n+1) \times (n+2)/2$ data samples are required. Otherwise, the model will be undetermined. Furthermore, if a multilayer perceptron is used, it is necessary to consider regulating the model

complexity to avoid overfitting. Regarding the performance of the approximation model, diverse factors have to be taken into account. The most important factor is accuracy, in both on training data and testing data.

Concerning catalysis field, ANNs have successfully been applied to conventional catalytic modelling and design of solid catalysts. Two types of ANNs applications have been described up to now in the frame of combinatorial catalysis: (i) ANN catalyst compositional models, correlating composition and synthesis variables with catalytic performance and (ii) ANN kinetic models, correlating reaction conditions with catalytic performance. The first reported applications include the design of solid catalyst (Hattori and Kito (1995)) for different reactions of interest, such as design of ammoxidation of propylene catalyst (Hou et al. (1997)), design of methane oxidative decoupling catalyst (Huang et al. (2001)), analysis and prediction of results of the composition of NO over zeolites (Sasaki et al. (1995)); and the integration of ANNs techniques with evolutionary strategies for the design of propane ammoxidation catalyst (Cundari et al. (2001)) or for the material discovery in the oxidative dehydrogenation of ethane (ODHE) reaction (Corma et al. (2002a)), allowing the analysis and prediction of catalytic results within a population of catalysts produced by combinatorial techniques. Moreover, modelling of multi-phase crystalline systems in zeolite synthesis by means of ANNs has been described in (Moliner et al. (2005)). Regarding ANN kinetic models, applications (Bulsari (1995); Biniwale et al. (2002); Alaradi and Rohani (2002); Serra et al. (2003b)) refer to modelling experimental kinetic data in order to obtain rapidly black box models. These ANN kinetic models could be promptly obtained for a series of catalysts and rapidly determine which are the reaction conditions for optimal catalytic performance of each material. In addition, those models

can be applied for further catalyst scale up, process control and optimization. For example, in Serra et al. (2003b,a) several ANNs were trained to predict reaction results based on reactor conditions for the hydroisomerisation of linear alkanes.

3.2 Genetic Algorithms

Genetic algorithms (GAs) are adaptive methods used to solve search and optimization problems, based on genetic processes of biological organisms. GA was developed in the 1970's by John Holland and students at the University of Michigan (Holland (1992)). Their aim was to simulate adaptive process of natural systems and to develop artificial systems that reattain features of natural systems. The canonical form of the GA encodes each candidate solution to a given problem as a binary, integer or real-valued string, referred to as a chromosome. GAs simulate the genetic state (chromosomes) of a population of individuals using recombination operators (crossover and mutation). Crossover exchanges genetic material between two parents. Mutation flips a bit in a chromosome. It is conduced to prevent the premature convergence of the design variables, that is all the bit structures of strings in the mating pool become identical in an early stage of evolution. However, mutation may also slow down the searching process, affecting speed convergence. Each individual is evaluated and fitness assigned in proportion to the value of the objective function for the individual. New individuals created by these operators are selected on the basis of their fitness for the next generation.

GAs have had a great measure of success in search and optimization problems (Oduguwa et al. (2005)). The main reason for their success is their ability to

exploit the information accumulated about an initially unknown search space in order to bias subsequent searches into useful subspaces. This is their key feature, particularly in large, complex, and poorly understood search spaces, where classical search tools (enumerative, heuristic, etc.) are inappropriate, offering a valid approach to problems requiring efficient and effective search techniques.

Regarding fitness evaluations, they are not always straightforward in many real-word applications (Jin (2005)), like in the field of catalysis in where an explicit fitness function does not exist, or the evaluation of the fitness is computationally expensive. In both cases, it is necessary to estimate the fitness function by constructing an approximate model. Several models have been used for fitness approximation, for instance such as multi-layer perceptrons, radial-basis-function networks, etc (Jin (2005)). In our case, we have employed multi-layer perceptrons.

There are two major concerns in using approximate models for the fitness evaluation. First, it should be ensured that the evolutionary algorithm converges to the global optimum or a near-optimum of the original fitness function. Second, the computational cost should be reduced as much as possible. One essential point is that it is very difficult to construct an aproximate model that is globally correct due to the high dimensionality, inhomogeneous distribution and limited number of training samples. If an approximate model is used for fitness evaluation, it is very likely that the evolutionary computation will converge to a false optimum. Therefore, it is very essential in most cases that the approximate model should be used together with the original fitness function. So in the evolution control, the original fitness function is used to evaluate some of the individuals or all individuals in some generations. An

individual or a generation that is evaluated using the original fitness function is called a controlled individual or generation respectively. There are different approaches from the viewpoint of evolution control. For example, it is possible to follow a generation-based evolution control where the evolution control is carried out once in a fixed number of generations.

Moreover, due to the lack of data and the high dimensionality of input space, the quality of the approximate model should be improved as much as possible given a limited number of data. In our case, a previous study of the multi-layer perceptron topology, training algorithm and error measures selection is needed.

Genetic algorithms seem quite appropriate for heterogeneous catalysis since a) GA tolerates noisy data (experimental data) with considerable error, b) GA uses a population of points to conduct the search, which fits quite well with the application of HTE techniques, and c) the goal is to find an approximate global maximum in a high-dimensional space, minimizing the number of trials. GAs are very powerful tools but they could be dangerous if the problem codification is not appropriate. If the selected codification for the problem was wrong, it would be possible that the algorithm would solve a different optimization problem from the one under study.

Among their different applications, it should be mentioned the development of novel gasoline isomerization catalysts (Corma et al. (2002b)), carbon monoxide oxidation catalysts (Pereira et al. (2005)) and propane oxidative dehydrogenation catalysts (Wolf et al. (2000)). Moreover, in Valero et al. (2003, 2004a) a softcomputing technique allows to discover the best kinetic values for several n-paraffin reactions inside an specific range of input values. In Corma et al.

(2005) the synthesis variables of mesoporous Tisilicate materials are intensively and simultaneously explored with the aim of optimising the catalytic performance of the resulting catalysts for the epoxidation of olefins.

3.3 Discussion

The current iterative optimization algorithms, especially genetic algorithms, applied to the discovery and optimization of heterogeneous catalyst suffer mainly from two important points. On one hand, the optimization convergence is poor, and, therefore, the number of experimental rounds required to reach a maximum is far too large and sometimes the final catalytic performance of the found optimum does not satisfies the requirements for scaling-up. That is especially true for optimizations dealing with high-dimensional spaces, such as catalyst formulations comprising pools of 30 elements/variables. Moreover, through the optimization process many samples are experimentally tested, even when it could be said a priori that the expected catalytic performance would be poor - fair. On the other hand, the current optimization approaches do not apply a general and flexible codification method, being high the danger of divergence towards false objectives. An adequate codification should integrate (a) complex catalyst formulation comprising different components such as supports, active phase promoters, enhancers, etc. and (b) associated (chemical/thermodynamic or final-application-oriented) rules or constraints. In addition, other variables such as preparation procedure or catalytic testing conditions (real or categorical) must be also defined in the design tool.

An objective of a new DoE [3] tool would be to increase the convergence rate but

[3] Design of Experiment

maintaining the exploration function. Such a tool would integrate an optimizer (GA) and an approximate multivariate model (ANN), in such a way that in each experimental iteration (a) the optimizer will make use of the model for reducing the number of experiments and (b) the model will be improved by adding the new available data to the training/fitting dataset. Moreover, the codification tool should allow the proper definition of the "chemical problem" and the associated constraints.

4 SoftComputing Framework

A new optimization tool has been developed for the discovery and optimization of catalytic materials when exploring a high-dimensional space. Hence, the knowledge extracted from the previous experimentation can be applied in the design of the new subset of catalysts to be experimentally screened in next optimization iteration. This tool follows a new optimized architecture based on our previous works in which soft computing techniques were applied (Valero et al. (2004a); Corma et al. (2005); Serra (2004); Valero et al. (2004b)).

The proposed soft computing framework combines a genetic algorithm (GA) and an artificial neural network (ANN). Thus, ANN is employed as an approximated model for fitness evaluation, whereas the developed GA tries to find the optimal solution by investigating several catalysts simultaneously. Specifically, this soft computing technique (Figure 1) consists of the following steps: (i) setting up; (ii) ANN re-training; (iii) GA operators; (iv) pre-screening; and (v) experimental testing. The steps ii to iv are repeated till the convergence criteria is satisfied. However, the final actions performed in each step depend on the configuration of the tool implemented following the suggested frame-

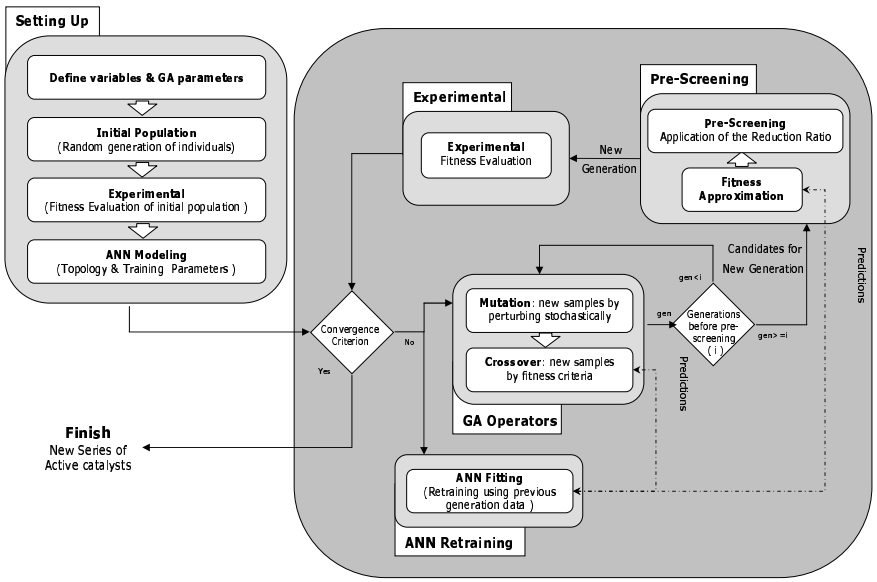

Fig. 1. Structure of the proposed soft computing technique

work. For example, it is possible to get different generations (employing fitness approximation) before carrying out the pre-screening step.

This software tool has been developed using the Borland DELPHI Enterprise 7 IDE. The computers used are Intel(R) Pentium(R) 4 CPU 2.80GHz processor, with a system memory of 1024MB RAM, and with the operating system Windows XP Professional Service Pack 2. Moreover, the batch system of the SNNS (Stuttgart Neural Network Simulator) is employed in all the tasks involving ANNs (Zell et al. (1995)).

In the following subsections, a more detailed explanation about the most relevant aspects of each algorithm step are shown.

4.1 Setting up

In the setting up process, the problem under study must be codified properly; next, the soft computing approach parameters are set; following, the starting

generation of the optimization process is calculated; and finally a suitable ANN model is obtained for predicting the catalytic performance.

4.1.1 Codification Problem.

The way in which the variables under study are codified is a crucial aspect, because GAs are very powerful tools but they could be dangerous if the problem codification is not appropriate. A bad codification selection should cause the tool to achieve a wrong solution, because it would be possible that the algorithm would solve a different optimization problem from the one under study. In the problem faced in this paper, each variable belongs to a continuous domain so it has been decided to adopt real codification (Goldberg (1991)).

In this kind of problems, the optimization variables can represent concentrations of the active compounds (catalyst formulation), preparation conditions or reaction conditions. The developed codification allows the simultaneous optimization of these variables. In addition, it is possible to define some rules that guide and restrict the optimization procedure, i.e., the maximum and minimum quantities of each optimization variable, compatibility between elements and/or conditions, the number of elements that can be selected simultaneously, etc.

Specifically, each sample or chromosome (Figure 2) is formed by zero or more compounds and conditions. The compounds describe the chemical elements (ingredients) included in the formulation of the material. Each compound can have one or more sections. Each section groups together those elements that obey specific characteristics. Moreover, sections are divided into subsections,

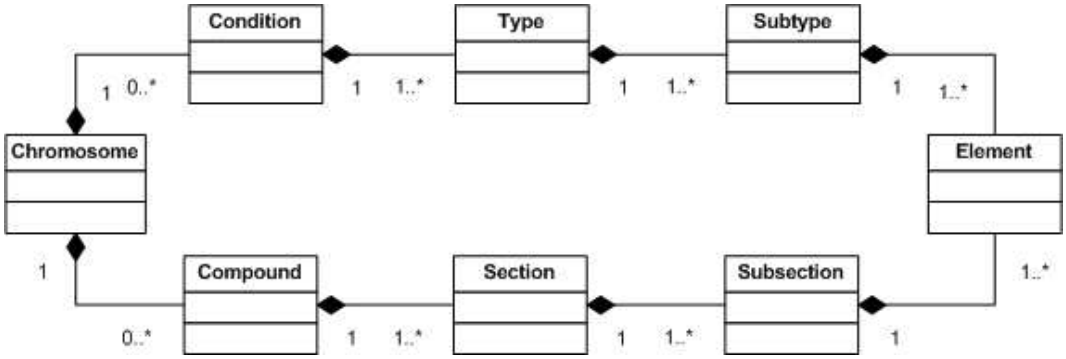

Fig. 2. Chromosome codification schema.

which contain the elements of the material. On the other hand, all the optimization variables concerning to preparation and reaction conditions will be represented as conditions in the codification. Conditions are also divided into types, subtypes and final variable values. For both compounds and conditions, it is possible to define different guiding rules in each division level, enabling to determine the number of elements to be selected from the lower level, and the maximum and minimum values of each element. In Figure 3 an example of a chromosome for a general formulation of a gold-based catalysis is displayed, showing its hierarchical structure.

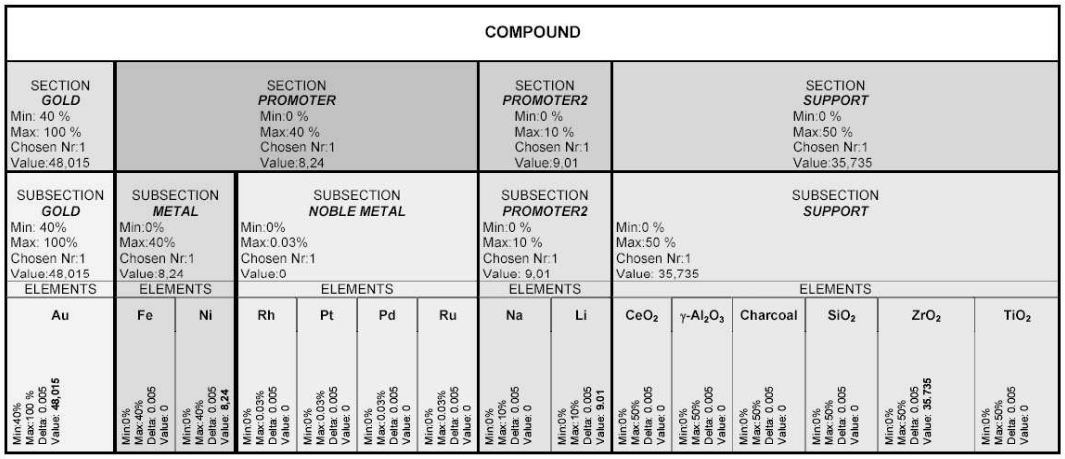

Fig. 3. Example of the codification of the general formulation of gold-based catalyis, including different rules.

4.1.2 Setting up the soft computing parameters.

This tool allows different ways to carry out the optimization process, being the users in charge of selecting the most suitable one to their interests. The different soft computing parameters that must be set are:

- **Generations before pre-screening step.** This parameter represents the fixed number of calculated generations before the evolution control is carried out (when samples are experimentally tested). The GA provides with the fixed generations, employing the ANN to approximate the fitness of each individual when needed. It should be pointed out that the frequency of the evolution control must depend on the fidelity of the approximate model. Thus, it is convenient to set low values in this parameter while the performance of the ANN model obtained is not good enough. Notice that users can modify this frequency along the optimization process, so the developed tool allows to follow an adaptive evolution control (Jin (2005)).
- **Virtual population.** This factor indicates the number of individuals proposed by the GA in each generation. The higher the population size, the faster the convergence. On the contrary, if the population size is not large enough, the search would not be able to converge to the global maximum.
- **Reduction ratio.** One important parameter in the pre-screening strategy is the reduction ratio (6) of the GA-proposed generation in order to obtain the final generation to be experimentally tested.

$$Reduction\ Ratio\% = \frac{Size\ of\ Controlled\ Population}{Size\ of\ Virtual\ Population} \tag{6}$$

Notice that if the reduction ratio is 0%, the pre-screening step will not be carried out. It should be also noticed that reduction rates higher than 40% are not convenient, since the error introduced in the optimization system by

the ANN predictions strongly interferes in the GA behaviour. In this case, a lot of the fitness approximations are not contrasted with the controlled values obtained in the experimental process.

- **Mutation probability.** The mutation is an explorer operator which looks for new solutions and prevents the premature convergence, ensuring the diversity of the population. However, mutation may also play a detrimental role to achieve fast convergence. For this reason, low values of this parameter are desired (5%..20%).
- **Number of genes to be mutate.** It is possible to set the number of genes that will be modified when a sample has been selected by the mutation operator. With higher values, the mutation operator has more impact.
- α **value**. Determines the size of the confidence interval used by the crossover operator and it also affects the performance of this operator. Thus, an α of 0.5 represents that exploring and exploiting capacities are balanced, whereas higher α values represent an increase in the exploration aptitude and vice versa.
- **Parents proportion.** By means of this parameter, the final number of selected individuals as progenitors is set. The progenitors (best individuals) are employed to calculate the confidence intervals used by the crossover operator. A desired value for this parameter will be lower than 30% of the virtual population.

4.1.3 Starting generation

An initial set of individuals (materials) is obtained following a process that guarantees the initial population diversity. So, this process consists of creating several random generations and carrying out a statistical population study

in order to select the most diverse population. The diversity of the initial generation ensures that the optimization process gains information for the whole search space. The starting generation is empirically tested in the reactor. Therefore, the experimental results of the behaviour of the samples are known.

4.1.4 Getting an ANN model

In order to establish a suitable ANN model for the problem under study, different factors involved in the ANN prediction performance are analyzed: ANN topology, training algorithms and activation functions. Several experiences following this methodological procedure and the results obtained are explained in Serra et al. (2003a); Valero et al. (2004a); Moliner et al. (2005); Corma et al. (2002a).

The experimental results of the starting generation are employed to carry out these studies. Usually, the initial number of samples is small. So these samples are grouped in different subsets of training (80%) and testing (20%)samples, in order to avoid that problem. Thus, each experiment is carried out with different combinations of training and testing subsets.

Using supervised learning, an incremental method is applied, testing different neural network topologies based on the multilayer perceptron. Starting with one single layer and few neurons, the topology is modified by increasing the number of neurons and the number of hidden layers. Different experiments are also carried out with those algorithms that turn out to be more suitable for the multilayer perceptron according to the literature (Bishop (1996)). Specifically, neural networks are trained with backpropagation and backpropagation with momentum, with different parameters (learning factor η and momentum μ).

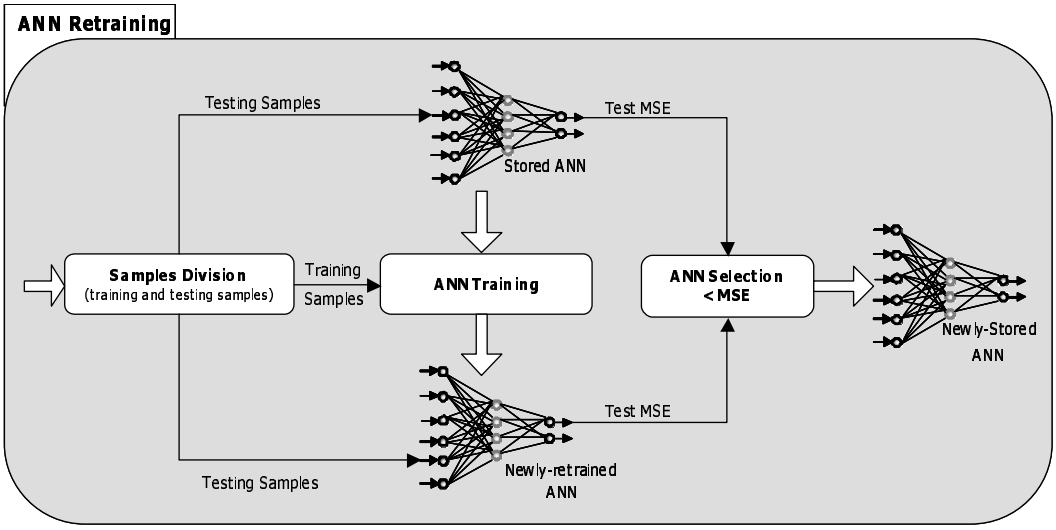

Fig. 4. Steps of the ANN retraining process

Finally, a study of the activation functions is also carried out, using logistic sigmoidal or tangential sigmoidal functions in the hidden units (explanation in 3.1).

4.2 ANN re-training

Regarding the ANN re-training step, new experimental data derived from the testing of each succeeding generation is divided into training and testing data. The training set is used to retrain the stored ANN, whereas the testing set is employed to compare the stored ANN and the newly-retrained ANN. So that one with the best predicting performance is selected and stored (figure 4).

This procedure prevents the generalizing capacity of the ANN model from being diminished. Notice the fact that the training data is derived from an experimental optimization process, so the fitness of the individuals of every new controlled generation is increased and consequently, the diversity of the population is progressively reduced. Other training procedures were studied

(Serra et al. (2005)), being this one selected as the best option.

4.3 GA operators

The candidates for the new generation are designed by the GA operators, taking into account the previous experimental results. Specifically, the GA employs mutation and crossover operators, requiring the last one the assistance of the ANN. Mutation operator modifies in a haphazard way genes with a new value, jumping randomly anywhere within the allowed gene domain. It do not only acts modifying genes values, but also modifying the elements selection. Thus, the mutation is an explorer operator which looks for new solutions and prevents the system to converge quickly on a local maximum, avoiding the loss of genetic diversity. The samples modified by the mutation operator are not further modified by the crossover operator.

The crossover operator proposed by Ortiz et al. (2001) has been adapted to our interests, considering the rules defines in the developed codification. This operator based on confidence intervals is associated with the capacity of interpolation (exploitation), related to the belonging of an individual to a confidence interval built from the best individuals of the population (parents). It is also associated with the capacity of extrapolation (exploration), derived from its not belonging to the same confidence interval. To obtain that confidence interval (7), three new individuals [4] formed by the lower ends ($CILL$), upper ends ($CIUL$) and means (CIM) of the parent samples genes are calculated. Therefore, the individuals CILL and CIUL divide each gene's domain, D_i, into

[4] CILL= Confidence Interval Lower Limit; CIUL= Confidence Interval Upper Limit; CIM= Confidence Interval Mean

three subintervals I_1, I_2 and I_3, being Min_i and Max_i the lower and upper limits of the domain D_i respectively.

$$D_i \equiv I_1 \bigcup I_2 \bigcup I_3;$$
$$I_1 \equiv [Min_i, CILL]; I_2 \equiv]CILL, CIUL[; I_3 \equiv [CULL, Max_i,] \quad (7)$$

Specifically, the confidence interval I_2 (the exploitation interval) is built from the best individuals of the population under the hypothesis that they are distributed following a Student's τ distribution, and there is a probability $1-\alpha$ of their genes' values belonging to that interval. Therefore, α is the probability that the genes' values belong to intervals I_1 or I_3 (exploration intervals). Thus, an α of 0.5 represents that exploring and exploiting functions are balanced, whereas higher values represent an increase in the exploration function and vice versa.

When an individual is crossed, the genes of the new individual are obtained from the original ones, following crossover rules (Ortiz et al. (2001)) and taking into account the confidence interval to which it belongs, the fitness values of the three confidence interval individuals and the quality of the original sample.

In order to calculate the fitness value of the three confidence interval individuals ($CILL$, $CIUL$, CIM), it is required to predict their catalytic results by means of the ANN, which simulates the experimental catalytic testing. Figure 5 shows two illustrating examples of mutation and crossover.

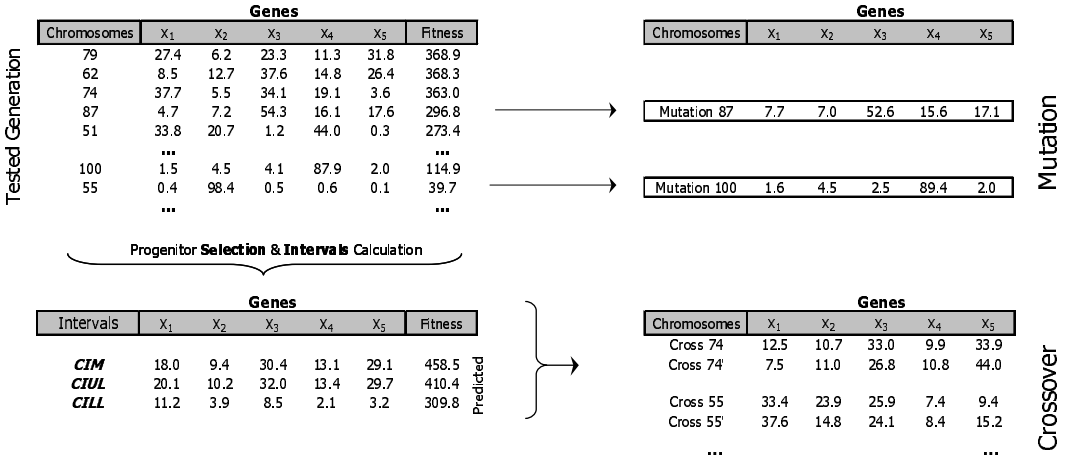

Fig. 5. Examples of mutation and crossover operations

4.4 Pre-screening

For very complex optimizations like those under study, the selection of the population size appears to be a crucial issue, and a compromise between experimental effort and convergence performance to global maximum should be reached. Therefore, this tool allows to increase the optimization performance, reducing the real number of catalysts to be experimentally evaluated (controlled samples to be synthesized and tested) by means of the pre-screening process.

The pre-screening process is done in two phases. Firstly an approximation fitness value for each sample is calculated by means of the ANN predictions. Secondly, a controlled generation is obtained from the virtual population, reducing the number of samples according to the value of the reduction ratio parameter. The individuals that will belong to the controlled generation are selected with the *roulette wheel* method (Kecman (2001)). This method consists of a random selection in which the samples with a higher fitness value have more possibilities to be selected.

Due the lack of data and the high dimensionality of the input space, it is very difficult to obtain a perfect global functional approximation of the original fitness function. For this reason, the approximated model should be used together with the original fitness function, i.e the reactor in our case. Thus, in this step, the pre-screened generation is experimentally evaluated, getting the controlled fitness evaluations of the last generation proposed by the GA. This controlled values are updated in the system and later, they are employed in the next optimization loop.

5 Results

5.1 Application examples

An hypothetical function (8), based on the modelled behaviour of multicomponent catalyst explored in the field of combinatorial catalysis, has been applied to illustrate the diverse optimization possibilities offered by the developed tool.

$$Y(x_1, x_2, x_3, x_4, x_5) = z_i(x_1, x_2) + z_j(x_2, x_3) z_k(x_3, x_4, x_5) \tag{8}$$

where:

$$\sum x_i = 100, x_i \geq 0$$
$$z_i(u, v) = 0.6g(100u - 35, 100v - 35) + 0.75g(100u - 10, 100v - 10) + 1g(100u - 35, 100v - 10)$$
$$z_j(u, v) = 0.4g(100u - 10, 100v - 30)$$

$$z_k(u,v,w) = 5 + 25(1 - (1 + (u - 0.3)^2 + (v - 0.15)^2 + (w - 0.1)^2)^{0.5})$$
$$g(u,v) = 100 - (u^2 + v^2)^{0.5} + 50(sin(1(u^2 + v^2)^{0.5}))/((u^2 + v^2 + 0.001)^{0.5})$$

In figure 6, a representation of this hypothetical function is shown. This function has a complex topology with several local maximums. As it can be seen, this function presents three high-activity areas while presenting some periodicity. This behaviour is common for heterogeneous catalysts, when varying their composition and synthesis conditions. The maximum values are in the lighter areas (values closer to 550).

A suitable ANN model of this function was obtained, training and testing several ANN topologies with different training algorithms. A multi-layer perceptron with 5 input nodes, 4 nodes in the 1st hidden layer, 3 nodes in the 2nd hidden layer and 1 output node, trained with Backpropagation algorithm with momentum (learning factor=0.8, momentum term=0.8) was selected.

The influence of the population size on the convergence performance of the DoE tool was analyzed. For this purpose, a battery of tests has been accomplished, using different population sizes within the range typically employed in the experimental screening of catalysts (i.e. from 15 to 95 samples), repeating the optimization process ten times for each size. In particular, the soft computing optimization tool parameters (4.1.2) were set to: generation before pre-screening step=1; reduction ratio=0% (without pre-screening); mutation probability=5%; genes to mutate=1; $\alpha = 0.9$; parents proportion=10%. Moreover, a unique starting random generation was used for each population size, but considering that the average fitness of each starting generation was very similar among the different population sizes. That condition was forced in or-

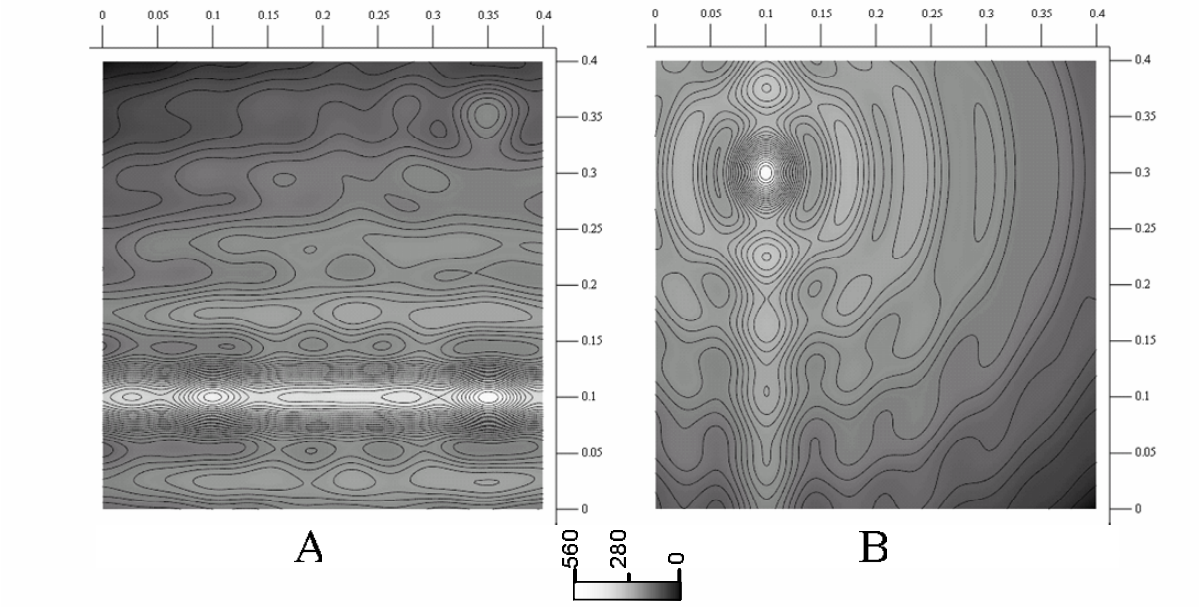

Fig. 6. Representation of the hypothetical function in the planes containing the absolute maximum of the function: A) Varying x_1 and x_2 with $(x_3, x_4, x_5) = (0.3, 0.15, 0.1)$; B) Varying x_2 and x_3 with $(x_1, x_4, x_5) = (0.11, 0.15, 0.1)$

der to minimize the effect of the quality of the initial generation, since this aspect was previously postulated to be very important (Valero et al. (2004a)).

Figure 7 shows the average and maximum quality achieved (normalized average of ten runs) in the optimization process for the different population sizes. It is clear that the higher the population size, the faster the convergence. Especially, values closer to the maximum are got with population sizes bigger than 35. However, when the size is not large enough (for instance 15 samples), the search does not converge to the global maximum and seems to be blocked. So, for complex optimization processes, the selection of the population size appears to be a crucial issue, and a compromise between experimental effort and convergence performance to global maximum should be reached. Therefore, it seems to be interesting to use the pre-screening process (4.4), which would

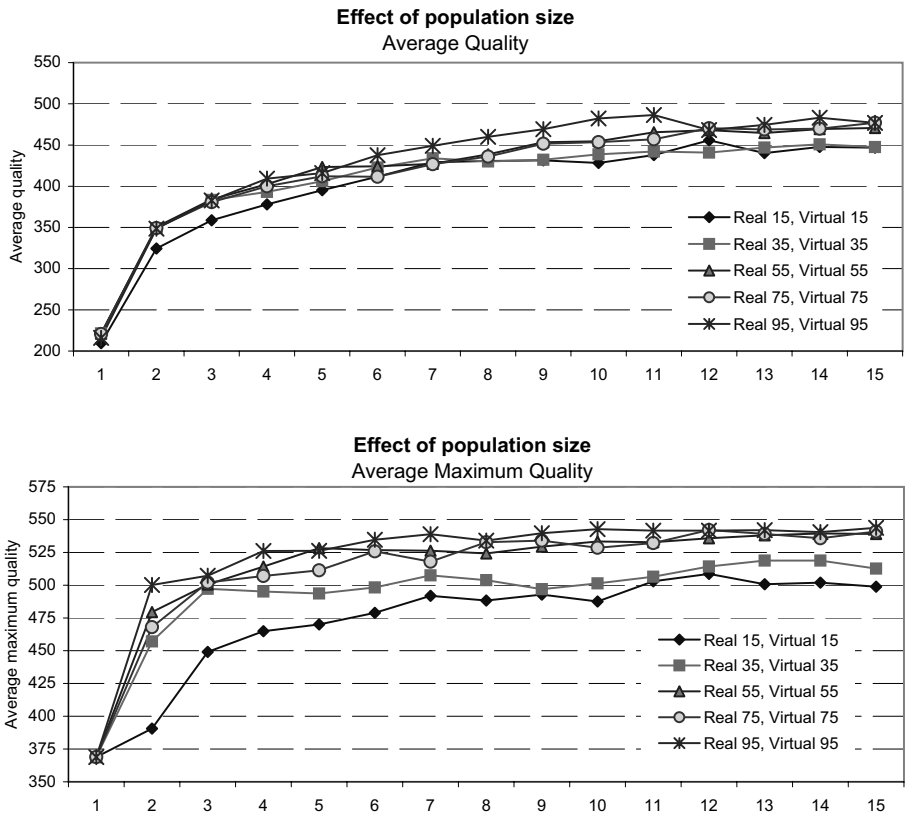

Fig. 7. Effect of population size on the soft computing tool performance.

allow to increase the optimization performance by reducing the real number of catalysts to be experimentally tested.

Thus, different reduction rates have been studied together with diverse population sizes with the aim of proving the pre-screening usefulness. As it was previously mentioned (4.1.2), this configuration of the DoE tool allows the genetic algorithm to work with virtual population sizes larger than those ones experimentally tested. Specifically, reduction rates of 20, 30 and 40% have been tested in order to get real populations of 35, 45 and 55 samples (i.e. normal reactor capacities). The other parameters of the soft computing tool were set with the same values used in the previous study about the influence of the population size. In the same way, the starting generation of each proof was established following the above-mentioned process. Ten runs were made

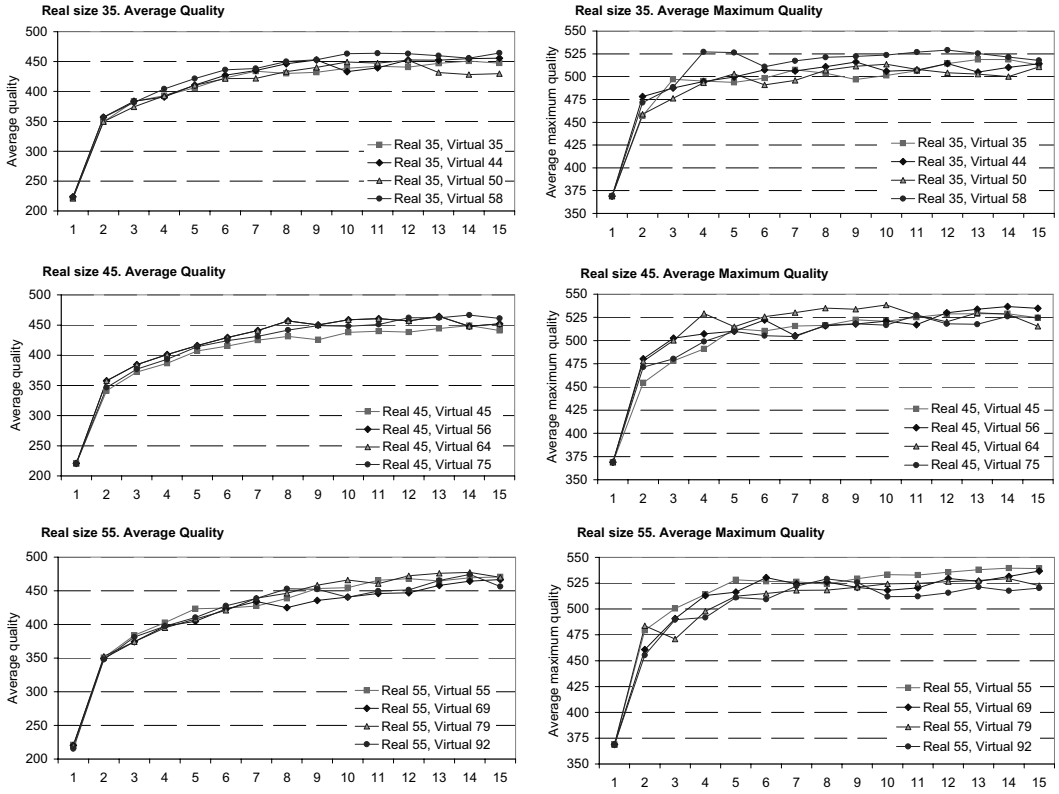

Fig. 8. Effect of the reduction ratio for different real or experimental population sizes on the soft computing tool performance.

for each combination. In figure 8 the average of the obtained results are shown, grouped for the same real population size (experimental samples). Specifically, the average quality obtained by each generation is displayed in left graphics, whereas the maximum quality average achieved by each generation is exposed in right graphics.

An improvement in the convergence with regard to using the optimization tool without pre-screening is observed in figure 8. This improvement is particularly visible in the initial steps of the optimization, specifically when smaller real size populations are used (i.e. 35 samples). That behaviour can be explained considering that the space explored by the GA is larger when pre-screening procedure is applied. In fact the GA works directly with the virtual population

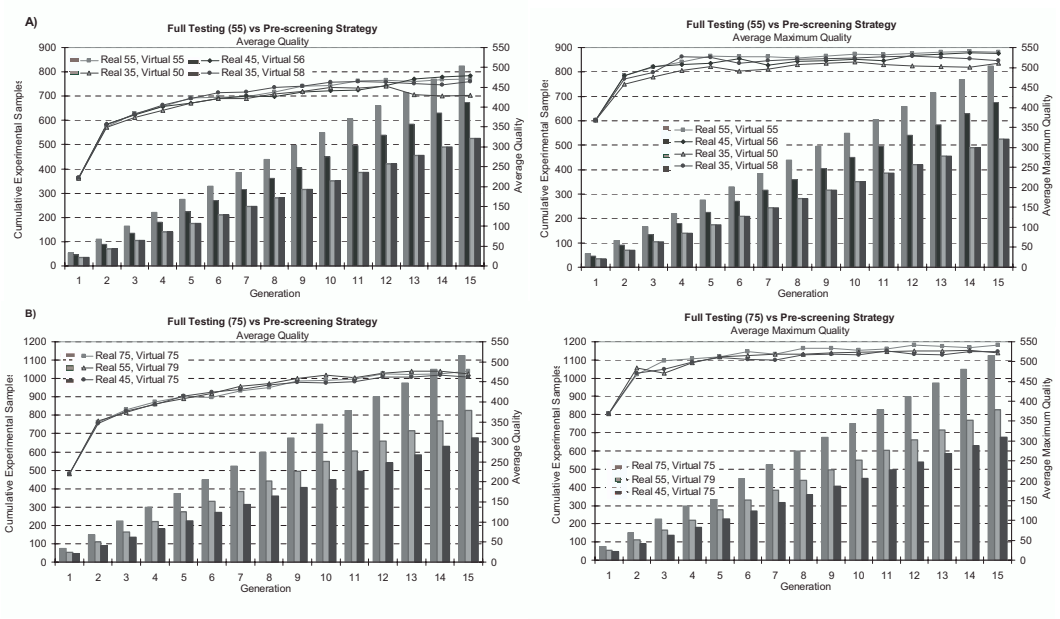

Fig. 9. Full testing versus pre-screening strategy. Correlation between performance and experimental costs

($virtual > real$); therefore the convergence to high-quality areas of the search space is faster. However, reduction rates higher than 40% are not convenient, since the error introduced in the optimization system by the predicted information not contrasted with experimental data strongly interferes in the GA behaviour.

Finally, figure 9 shows the benefits of using the pre-screening step regarding the experimental effort. Particularly, the correlations between the average and maximum of the quality achieved by the different optimization strategies (in lines) versus the experimental effort needed in each case (in bars) are shown. In the first case (figure 9- A), four different strategies are compared: virtual population of 55 samples without pre-screening; virtual population of 56 samples with a reduction ratio of 20%; virtual population of 50 samples with a reduction ratio of 30%; and finally, virtual population of 58 samples with a reduction ratio of 40%. In the second case (figure 9- B), three different strategies

are compared: virtual population of 75 samples without pre-screening; virtual population of 79 samples with a reduction ratio of 30%; and virtual population of 75 samples with a reduction ratio of 40%. In all cases the performance of the optimization process is similar for all the employed strategies, but the needed experimental effort is lower when a pre-screening strategy is followed. For instance, the strategy with a virtual population of 75 samples without pre-screening achieves a maximum of 511.31 in the fifth generation (average of ten runs), needing 375 experimental samples; whereas when employing a virtual population of 75 samples with a reduction ratio of 40%, a maximum of 509.93 is got in the fifth generation (maximum values of the function are closer to 550), but only 225 experimental samples are needed. So, a meaningful reduction of experimental samples is achieved, obtaining similar qualities.

5.2 Industrial application

The Soft Computing Tool has been applied to a problem with industrial implications, trying to optimize Ti-silicate catalysts for the epoxidation of olefins. In this case, the value of four variables have to be established.

The soft computing tool was parameterized as follows:Generations before pre-screening step = 1; Virtual population = 37; Reduction ratio = 0%; Mutation probability = 10%; Number of genes to mutate = 1; $\alpha = 0.7$; and Parents proportion = 20%. Moreover, following the procedure explained above(see 4.1.4), a multi-layer perceptron with 4 input nodes, 2 nodes in first hidden layer, 1 node in second hidden layer and 2 output nodes, trained with Back-propagation algorithm with Momentum (learning factor=0.8 and momentum term=0.5) was selected and successfully employed.

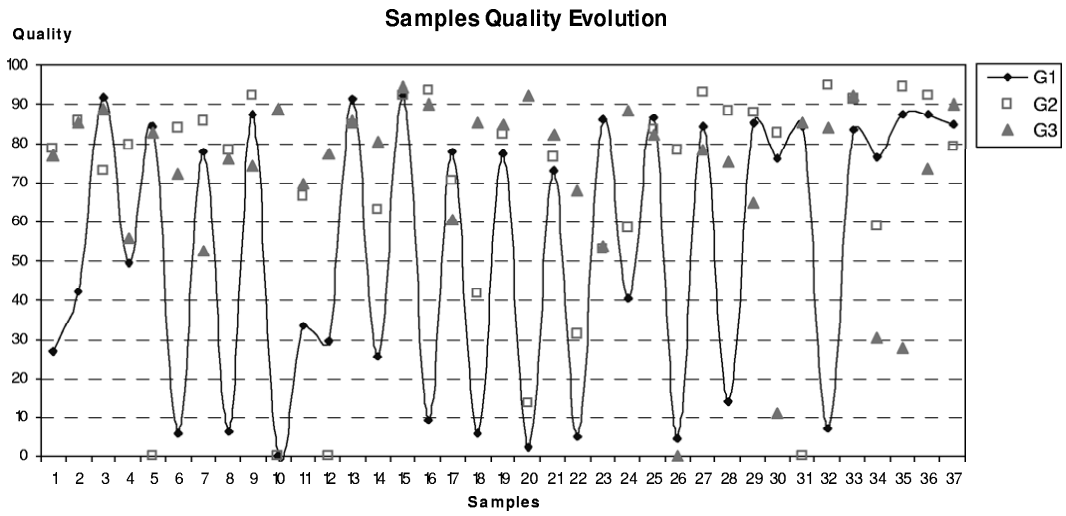

Fig. 10. Quality evolution of samples of the Ti-silicate catalyst for the epoxidation of oleofins

Through the optimization process three generations of 37 samples were synthesised and tested. Along this procedure, an important improvement in the activity and selectivity of the starting materials has been achieved as it can be observed in Figure 10. This figure shows the cyclohexene epoxide yields for the 3 evolved generations (3x37 samples). Moreover, the best catalyst found (2nd generation, sample 32) improves in 15% the catalytic performance (epoxide yield) with regard to the best previously reported catalyst (Corma et al. (1998)).

To summarize, a highly active and selective catalyst for the epoxidation of cyclohexene has been found, that can be applied to the epoxidation of other olefins, specially propylene. Epoxides are starting materials for commodity products like plastics or drugs (Taramasso et al. (1983)). The best materials have low titanium contents, and were extracted and silylated. These materials have a Ti-MCM-41 structure and a very hydrophobic surface.

6 Conclusions

A new Design of Experiments framework has been proposed for the intelligent discovery of new catalytic materials when exploring a high dimensional space. It has appropriate tools for high-dimensional optimization but maintains in memory the whole "history" of the search. So, in the future this method could be used to reduce the screening of statistically-poor active materials.

The optimization tool implemented following the proposed soft computing framework [5] could be employed practically in the design of experiments, since (i) the reduced population size required by the application is in the range employed by material characterization/testing techniques and (ii) the fast converge exhibited by the developed tool requires a low number of generations to identify high quality areas in the search space. Therefore, the application of this soft computing system to high throughput experimentation would reduce time and costs during HT experimentation.

Another important contribution is the new developed codification, which allows dealing with optimizations considering simultaneously complex catalyst formulations and different synthesis/testing conditions as variables, being possible to define different rules between variables. Although we have applied our technique to catalysis, our proposed codification is general enough to be applied to other fields such as materials science or drug discovery. Furthermore, the codification enables to carry out more deeper studies, not only optimizing the configuration of the samples, but also allowing the analysis of the elements that compose the samples (for example, identifying the elements that appear

[5] http://www.dsic.upv.es/users/ia/sma/tools/doE/index.html

in the best samples more frequently, in which percentage, etc.).

Finally, an industrial and relevant application of this novel optimization framework has allowed to find a highly active and selective catalyst for the epoxidation of cyclohexene, that can be applied to the epoxidation of other olefins, specially propylene. This catalyst improves by 15% the catalytic performance of the best previously reported catalyst (Corma et al. (1998)). Moreover, it should be pointed out that only three optimization generations (3x37 samples) have been necessary to obtain this new catalyst.

References

Alaradi, A., Rohani, S., 2002. Identification and control of a riser-type fcc unit using neural networks. Computers & Chemical Engineering 26, 401–421.

Aramendía, M., Borau, V., Jiménez, C., Marinas, J., Romero, F., Urbano, F., 2002. An approach to the construction of indexed libraries for the combinatorial selection of heterogeneous catalysts. Journal of Catalysis 209 (2), 413–416.

Biniwale, R., Labhsetwar, N., Kumar, R., Hasan, M., March 2002. Catalytic converter modelling: Artificial neural networks for perovskite-based catalyst. Society of Automotive Engineers, Technical Papers 1676, 49–54.

Bishop, C., 1996. Neural Networks for Pattern Recognition. Oxford Clarendon Press, Oxford.

Bulsari, A., 1995. Neural Networks for Chemical Engineers. Elsevier, Amsterdam.

Corma, A., Domine, M., Gaona, J., Jorda, J., Navarro, M., Rey, F., Perez-Pariente, J., Tsuji, J., McCulloch, B., Nemeth, L., 1998. Strategies to

improve the epoxidation activity and selectivity of Ti-MCM-41. Chem. Comm. 2211.

Corma, A., Serra, J., Argente, E., Botti, V., Valero, S., November 2002a. Application of artificial neural networks to combinatorial catalysis: Modelling and prediction of odhe catalysts. ChemPhysChem 3 (11), 939–945.

Corma, A., Serra, J., Chica, A., 2002b. Application of genetic algorithms to the development and optimisation of light paraffin isomerisation catalysts. In: Derouane, E., Parmon, V., Lemos, F., Ramôa Ribeiro, F. (Eds.), Principles and Methods for Accelerated Catalyst Design and Testing. Kluwer Academic Press, Dordrecht, pp. 153–172.

Corma, A., Serra, J., Serna, P., Argente, E., Botti, V., Valero, S., 2005. Optimization of olefin epoxidation catalysist applying high-throughput and genetic algorithms assisted by artificial neural networks (softcomputing techniques). Journal of Catalysis 225, 513–524.

Cundari, T., Deng, J., Zhao, Y., 2001. Design of a propane ammoxidation catalyst using artificial neural networks and genetic algorithms. Industrial & Engineering Chemistry Research 40 (23), 5475–5480.

Derouane, E., 2002. Principles and Methods for Accelerated Catalyst Design & Testing. Kluwer Academic Publisher.

Gedeck, P., Willett, P., 2001. Visual and computational analysis of structure activity relationships in high-throughput screening data. Current Opinion in Chemical Biology 5, 389–395.

Gilardoni, F., Graham, A., McKay, B., Brown, B., March 2003. Rational design, an alternative to the combinatorial explosion. In: Proceeding of 225th ACS National Meeting. New Orleans.

Goldberg, D., 1991. Real-coded genetic algorithms, virtual alphabets, and blocking. Complex Systems 5, 139–157.

Hattori, T., Kito, S., 1995. Neural network as a tool for catalyst development. Catalysis Today 23, 347–355.

Holland, J. H., 1992. Adaptation in natural and artificial systems. MIT Press, Cambridge, MA, USA.

Hou, Z., Dai, Q., Wu, X., Chen, G., 1997. Artificial neural network aided design of catalyst for propane ammoxidation. Applied Catalysis A: General 161, 183–190.

Huang, K., Chen, F., Lu, D. W., 2001. Artificial neural network-aided design of a multi-component catalyst for methane oxidative coupling. Applied Catalysis A: General 219, 61–68.

Jin, Y., January 2005. A comprehensive survey of fitness approximation in evolutionary computation. Soft Computing-A Fusion of Foundations, Methodologies and Applications 9 (1), 3–12.

Kecman, V., 2001. Learning and Soft Computing: Support Vector Machines, Neural Networks, and Fuzzy Logic Models. MIT Press, Cambridge, MA, USA.

Kirsten, G., Maier, W., 2004. Strategies for the discovery of new catalysts with combinatorial chemistry. Applied Surface Science 223, 87–101.

Klanner, C., Farruseng, D., Baumes, L., Mirodatos, C., Schüth, F., 2003. How to design diverse libraries of solid catalysts? QSAR and Combinatorial Science 22 (7), 729–736.

Klanner, C., Farruseng, D., Baumes, L., Mirodatos, C., Schüth, F., 2004. The development of descriptors for solids: Teaching catalytic intuition to a computer. Angewandte Chemie International Edition 43 (40), 5347–5349.

Moliner, M., Serra, J., Corma, A., Argente, E., Valero, S., Botti, V., 2005. Application of artificial neural networks to high-throughput synthesis of zeolites. Microporous and Mesoporous Materials. Zeolits, Clays 78, 73–81.

Montgomery, D., 2001. Design and analysis of experiments, 5th Edition. John Wiley.

Oduguwa, V., Tiwary, A., Roy, R., March 2005. Evolutionary computing in manufacturing industry: and overview of recent applications. Applied Soft Computing 5 (3), 281–299.

Omata, K., Watanabe, Y., Hashimoto, M., Umegaki, T., Yamada, M., 2004. Simultaneous optimization of preparation conditions and composition of methanol synthesis catalyst by all-encompassing calculation on an artificial neural network. Industrial and engineering chemistry research 43 (13), 3282–3288.

Ortiz, D., Hervas, C., Muñoz, J., 2001. Genetic algorithm with crossover based on confidence interval as an alternative to traditional nonlinear regression methods. In: 9th European Symposium On Artificial Neural Networks,ESANN'2001. Bruges,Belgium, pp. 193–198.

Pereira, S., Clerc, F., Farrusseng, D., Van der Waal, J., Marchmeyer, T., Mirodatos, C., 2005. Effect of the genetic algorithm parameters on the optimisation of heterogeneous catalysts. QSAR & Combinatorial Science 24, 45–57.

Rajan, K., Zaki, M., Bennet, K., 2001. Searching techniques for structure-property relationships in materials. Abstracts 221st National Meeting of the American Chemical Society.

Ripley, B., 1996. Pattern Recognition and Neural Networks. Cambridge University Press, Cambridge.

Sasaki, M., Hamada, H., Kintaichi, Y., Ito, T., 1995. Application of a neural network to the analysis of catalytic reactions analysis of no decomposition over cu/zsm-5 zeolite. Applied Catalysis A: General 132, 261–270.

Senkan, S., 1998. High-throughput screening of solid-state catalyst libraries.

Nature 394 (6691), 350–353.

Senkan, S., 2001. Combinatorial heterogeneous catalysis - a new path in an old field. Angewandte Chemie International Edition 40 (2), 312–329.

Serra, J., April 2004. Combinatorial catalysis: Development of new techniques and applications of interest. Ph.D. thesis, D.Q., I.T.Q, Polythecnic University of Valencia.

Serra, J., Corma, A., Argente, E., Valero, S., Botti, V., November 2003a. Neural networks for modelling of kinetic reaction data applicable to catalyst scale up and process control and optimization in the frame of combinatorial catalysis. Applied Catalysis A: General 254 (1), 133–145.

Serra, J., Corma, A., Chica, A., Argente, E., Botti, V., June 2003b. Can artificial neural networks help the experimentation in catalysis? Catalysis Today 81 (3), 393–403.

Serra, J., Corma, A., Farruseng, D., Baumes, L., Mirodatos, C., Flego, C., Perego, C., 2003c. Styrene from toluene by combinatorial catalysis. Catalysis Today 81 (3), 425–436.

Serra, J., Corma, A., Valero, S., Argente, E., Botti, V., 2005. Soft computing techniques applied to combinatorial catalysis: A new approach for the discovery and optimization of catalytic materials. QSAR & Combinatorial Science, In press.

Sun, Y., Chan, B., Ramnarayanan, R., Leventry, W., Mallouk, T. E., Bare, S. R., Willis, R., 2002. Split-pool method for synthesis of solid-state material combinatorial libraries. Journal of Combinatorial Chemistry 4, 569–575.

Taramasso, M., Perego, G., Notari, B., 1983. Us patent 4410501.

Valero, S., Argente, E., Botti, V., Serra, J., Corma, A., 2003. Softcomputing techniques applied to catalytic reactions. In: Actas X Conferencia de la Asociacin Espaola para la Inteligencia Artificial. Vol. I. pp. 213–222.

Valero, S., Argente, E., Botti, V., Serra, J., Corma, A., 2004a. Soft compting techniques applied to catalytic reactions. Lecture Notes in Artificial Intelligence 3040, 550–559.

Valero, S., Argente, E., Botti, V., Serra, J., Corma, A., August 2004b. A soft computing technique applied to industrial catalysis. In: López de Mántaras, R., Saitta, L. (Eds.), ECAI2004. ECCAI, IOS Press, pp. 765–769.

Végvári, L., Tomposb, A., Gobölös, S., Margitfalvi, J., 2003. Holographic research strategy for catalyst library design: Description of a new powerful optimisation method. Catalysis Today 81 (3), 517–527.

Wang, K., Wang, L., Yuan, Q., Luo, S., Yao, J., Yuan, S., Zheng, C., Brandt, J., 2001. Construction of a generic reaction knowledge base by reaction data mining. Journal of Molecular Graphics and Modelling 19 (5), 427–433.

Weaver, D., 2004. Applying data mining techniques to library design, lead generation and lead optimization. Current Opinion in Chemical Biology 8 (3), 264–270.

Wolf, D., Buyevskaya, O., Baerns, M., 2000. An evolutionary approach in the combinatorial selection and optimization of catalytic materials. Applied Catalysis A: General 200, 63–77.

Yamada, Y., Kobayashi, T., Mizuno, N., 2001. Recent advances on parallel synthesis and utilization of artificial intelligence for combinatorial chemistry of solid catalysts. Catalysts & Catalysis 43 (5), 310–315.

Zadeh, L., March 1994. Fuzzy logic, neural networks and soft computing. Communications of the ACM 37 (3), 77–84.

Zell, A., Mamier, G., Vogt, M., 1995. SNNS-Stuttgart Neural Network Simulator. User Manual, Version 4.1. University of Stuttgart, Stuttgart.