Article

MAP: An Iterative Experimental Design Methodology for the Optimization of Catalytic Search Space Structure Modeling

Laurent A. Baumes

J. Comb. Chem., **2006**, 8 (3), 304-314• DOI: 10.1021/cc050130+ • Publication Date (Web): 07 March 2006

MAP: An Iterative Experimental Design Methodology for the Optimization of Catalytic Search Space Structure Modeling

Laurent A. Baumes*

Max-Planck-Institut für Kohlenforschung, Mülheim, Germany, and CNRS-Institut de Recherche sur la Catalyse, Villeurbanne, France

Received September 27, 2005

One of the main problems in high-throughput research for materials is still the design of experiments. At early stages of discovery programs, purely exploratory methodologies coupled with fast screening tools should be employed. This should lead to opportunities to find unexpected catalytic results and identify the "groups" of catalyst outputs, providing well-defined boundaries for future optimizations. However, very few new papers deal with strategies that guide exploratory studies. Mostly, traditional designs, homogeneous covering, or simple random samplings are exploited. Typical catalytic output distributions exhibit unbalanced datasets for which an efficient learning is hardly carried out, and interesting but rare classes are usually unrecognized. Here is suggested a new iterative algorithm for the characterization of the search space structure, working independently of learning processes. It enhances recognition rates by transferring catalysts to be screened from "performance-stable" space zones to "unsteady" ones which necessitate more experiments to be well-modeled. The evaluation of new algorithm attempts through benchmarks is compulsory due to the lack of past proofs about their efficiency. The method is detailed and thoroughly tested with mathematical functions exhibiting different levels of complexity. The strategy is not only empirically evaluated, the effect or efficiency of sampling on future Machine Learning performances is also quantified. The minimum sample size required by the algorithm for being statistically discriminated from simple random sampling is investigated.

Introduction

High throughput experimentation (HTE) has become an accepted and important strategy in the search for novel catalysts and materials.[1] However, one of the major problems is still the design of experiments (DoE). At early stages of a discovery research program, only pure exploratory computer science methodologies coupled with very fast (i.e., qualitative response) screening tools should be employed. This aims at discovering the different "groups" of catalyst outputs to provide well-defined boundaries for future optimizations. Therefore, the prescreening strategy will extract information or knowledge from a restricted sampling of the search space to provide guidelines for further screenings. The chemist's knowledge should be used to define a "poorly explored" parameter space, leading to opportunities of surprising or unexpected catalytic results, especially when considering that HTE tools for synthesis and reactivity testing already restrict much the experimental space. However, very few new papers deal with the strategies that should be used to guide such an exploratory study. In most cases, either systematic methods for homogeneous covering[2−6] or simple random sampling (SRS)[7] are exploited, whereas other traditional DoE[8−10] are neglected due to their specificities and constraints, that is, restrictions. The typical distribution of catalytic outputs usually exhibits unbalanced datasets for which an efficient learning can hardly be carried out. Even if the overall recognition rate may be satisfactory, catalysts belonging to rare classes are usually misclassified. On the other hand, the identification of atypical classes is interesting from the point of view of the potential knowledge gain. SRS or homogeneous mapping strategies seem to be compulsory when no activity for the required reaction is measurable and the necessary knowledge for guiding the design of libraries is not available.[11]

In this study, classes of catalytic performances are unranked since the objective is not to optimize catalytic formulations but, rather, to provide an effective method for selecting generations of catalysts permitting (i) An increase in the quality of a given learning method performed at the end of this first exploratory stage. The aim is to obtain the best overall model of the whole search space investigated while working independently from the choice of the supervised learning system. (ii) A decrease in the misclassification rates of catalysts belonging to small frequency classes of performance (i.e., false negatives). (iii) Handling of all types of features at the same time, that is, both quantitative and qualitative. (iv) Integration of inherent constraints, such as a priori-fixed reactor capacity constraint and a maximum number of experiments, to be conducted (so-called deadline). (v) Proceeding iteratively and capturing the information contained in all previous experiments.

A new iterative algorithm called MAP (because it performs an improved MAPping) is suggested for the characterization

* Current address: Instituto de Tecnologia Quimica, (UPV-CSIC), Av. de Los Naranjos s/n, E-46022 Valencia, Spain. Phone: +34-963-877-806. Fax: +34-963-877-809. E-mail: baumesl@itq.upv.es.

10.1021/cc050130+ CCC: $33.50
Published on Web 03/07/2006

of the space structure. It works independently of the chosen learning process as a filter on the sampling to enhance the recognition rate by transferring selected catalysts to be synthesized and tested from "stable" search space zones to "unsteady" ones, which necessitates more experimental points be well-modeled within the search space. This new stochastic approach is a group sequential biased sampling.

The organization of the paper is as follows: First, the motivations of the present work are detailed, then some notations and representations used throughout the text are presented. In the third section, the MAP method and its central criterion are investigated. The fourth section details the creation of the benchmarks and discusses the great interest of using such testing methodology, then in the fifth section, the method is evaluated on different benchmarks identified by mathematical functions that exhibit different levels of difficulty. Finally, the sixth section emphasizes quantifying the strength of such a method through statistical analysis, and the results are thoroughly discussed.

Planning Methodologies

Since the entire research space is far too broad to be fully explored, three methodologies, namely, mapping, screening, and optimization, are used for selecting experiments to be conducted. These strategies are different from a point of view of their respective objectives relative to exploration and exploitation of the search space. Thus, each of them is more or less appropriated to a given HT step. At the beginning of a research program, the space must be explored, and approaching the end, exploitation of past results must be enhanced to obtain optimized formulations. In ref 6, mapping is described as to develop relationships among properties, such as composition, synthesis conditions, etc. while these interactions may be obtained without searching for hits or lead materials. Then the results of mapping studies can be used as input to guide subsequent screening or optimization experiments. The purpose of screening experiments is to identify iteratively, by accumulation of knowledge, hits or small space regions of materials with promising properties. The last way to guide the chemist, called optimization, is when experiments are designed to refine material properties.

Mapping has received relatively little attention, being too often subsumed under screening because of the rigidity of the different methods available. In MAP, the exploration is improved through its iterative behavior, and its flexibility is superior since all major constraints (iterative process, deadline, reactor capacity, and no a priori hypothesis) are handled. However, it remains entirely a mapping, since it does not perform any search for hits or leads, but instead makes use of performance levels by focusing on irregularity or variability, also called the "wavering" behavior of class distribution. Therefore, the given methodology should be tested versus an analogous algorithm under the same conditions.

In classical statistical DoE, the fundamental objective is hypothesis testing. An experiment is designed to generate statistically reliable conclusions to specific questions. Therefore, the hypothesis must be clearly formulated, and experiments are chosen according to the given supposition to verify it in a best statistical manner (see ref 12 for an example). This strategy is particularly suited to domains that are known sufficiently well that appropriate questions can be formed and models can be predefined.

In contrast, combinatorial methods are often employed for the express purpose of exploring new and unknown domains. Considering the different requirements defined earlier many drawbacks remain, such as (i) the mapping methods are rarely iterative; (ii) in case of complementary experiments (named plans), they have to be chosen by the user; (iii) the number of experiments to be performed is usually imposed by the method, it may not fit the reactor capacity, and selection of points to be added or removed is not a trivial task; (iv) usually, only continuous variables are considered, and high-order effects can be quantified to highlight synergism between factors, but the assessments have to be paid through a drastic increase in the number of experiments.

Even if carefully planned, statistically designed experiments which offer clear advantages over traditional one-factor-at-a-time alternatives do not guarantee the constraints imposed by the domain of application. Classical sampling procedures may appear as alternatives. The SRS gives a subset of possible catalysts among the whole predefined search space, for which each element of the population is equally likely to be included in the sample. A *stratified random sample* is obtained by partitioning the population then taking a SRS of specified size from each stratum. A *weighted random* sample is one in which the inclusion probabilities for each element of the population are not uniform. The use of weight or strata in samplings necessitate an a priori knowledge of the structure of the search space according to the given feature from which sampling is biased, and this is precisely what is lacking. Concerning SRS, an improved sampling strategy should control its random character to avoid obviously useless distributions. Finally, SRS remains the only strategy for comparison with our method; however, SRS should not be underestimated. See ref 13 for a detailed explanation of SRS robustness.

Simple Example: A Natural Behavior. Let us suppose that K catalysts have to be chosen iteratively by a given process. To obtain reliable statistics from the search space, a common strategy is to proceed with a first relatively broad random sample. Then this dataset is analyzed, and then smaller samples are used, fitting exactly the reactor capacity and taking into account the knowledge gained from the first evaluation. The number of iterations, noted g, is fixed a priori. $K = k_1 + gk_2$ is the total number of evaluated catalysts, with k_1 a given amount of points for the initialization step, whereas k_2 represents the size of the secondary samples. Let us draw a search space with five classes of catalyst performances, each represented by a different combination of color and shape. In this example, $k_1 = k_2 = 5$, $g = 10$, and $K = 55$. Figure 1a shows the k_1 first random points. Figure 1b and c show, respectively, the following generations ($g = 1$ and 2). One can observe that a natural way to proceed is to try to cover the space at the beginning of the exploratory study. At the third generation, 40 points still have to be placed, and on the other hand, the bottom-left corner is entirely filled with red triangles (Figure 2). Both "careful-

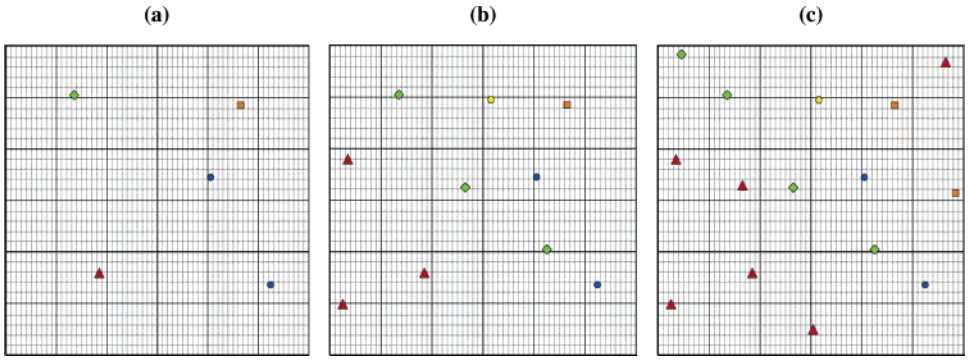


Figure 1. (a) k_1 first random points for initialization. (b) First generation of k_2 points with $k_2 = 5$. (c) Second generation.

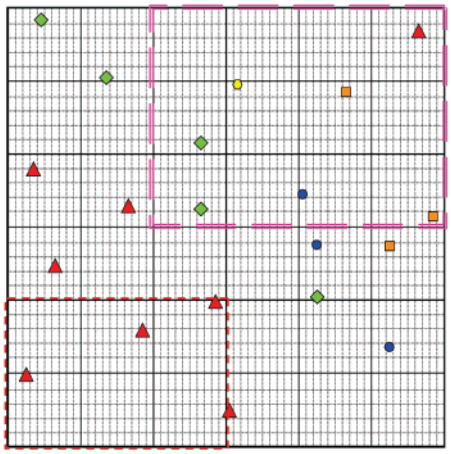


Figure 2. Third generation. The dotted line on the bottom left (- - -) defines a zone which is said "stable", as only red triangles appear in this area. On the other hand, the top right corner (== ==) is said to be "puzzling", as many (here, all) different classes of performances emerge in this region.

ness" and remaining time allow adding new points in this area, as shown in Figure 3, for the following generations. However, a decreasing number of points will be allocated in this area as the red triangle trend is confirmed. The top-right "zone" in Figure 2 (i.e., the dotted rectangle) appears as the most "puzzling" region, since the five different classes emerged for only seven points, making the space structure blurry and without a clear distribution at this step. As soon as the emergence of a confusing region is detected, a "natural behavior" is to select relatively more catalysts belonging to the given region to better capture the space structure. A better recognition of space zones in which a relatively high dynamism is detected should permit the understanding of the underlying or causal phenomenon and, therefore, could be extrapolated for localizing hit regions. In the last generation (Figure 4), 27 points are located the top third, 18 in the middle part and 10 in the bottom one. If a SRS was considered, the probability for obtaining such an average distribution would be very low. The present distribution and, consequently, such a "natural behavior" seem motivating for better modeling the structure of the search space.

This simple example emphasizes the intuitive iterative assignment of individuals onto the search space when the structure of the landscape has to be discovered. MAP can be performed with any types of features, and no distance measure is required; however, the learning system, also called machine learning (ML), should handle this flexibility, and this is the main reason (i) neural network (NN) approach has been chosen for future comparisons, and (ii) that the search space is supposed to be bidimensional in the simple previous example, since the *1-n*earest *n*eighbor (1-nn) method has been applied for modeling the entire search space (Figure 5). 1-nn, is a special case of k-nn[14] and necessitates a distance for assigning labels.

Notations and Iterative Space Structure Characterization. The MAP method is a stochastic group sequential biased sampling. Considering a given ML, it iteratively proposes a sample of the search space which fits user requirements for obtaining better ML recognition rates. Figure 6 depicts the whole process. The search space is noted Ω and $\omega_p \in \Omega$ ($p \in [1..P]$) corresponds to an experiment. The output set of variables is [Y]. A process $\mathscr{P}_{\text{partition}}$ is chosen by the user to provide a partition of [Y] in $H \geq 2$ classes, noted C_h, h $\in [1..H]$. $\mathscr{P}_{\text{partition}}$ can be a clustering, which, in some sense, "discovers" classes by itself by partitioning the examples into clusters, which is a form of unsupervised learning. Note that, once the clusters are found, each cluster can be considered as a "class" (see ref 11 for an example). [X] is the set of independent variables noted v_i, and x_{ij} is the

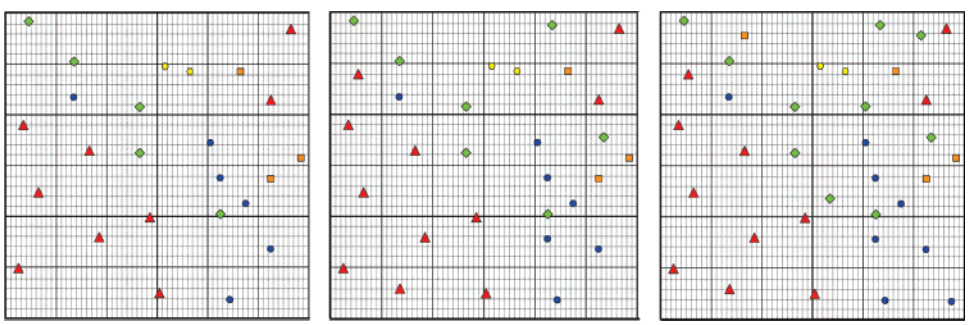


Figure 3. Simple example of intuitive iterative distribution of points: generations 4−7.

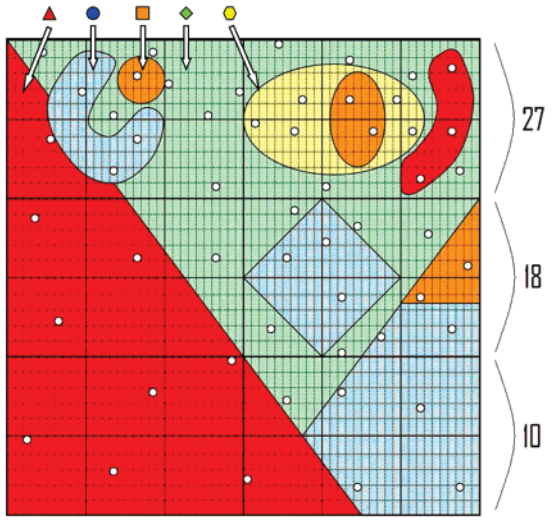


Figure 4. Last generation. $k_1 = 5$, $k_2 = 5$, $g = 10 \Rightarrow K = k_1 + g$. $k_2 = 55$. All the points are represented by white dots, whatever the corresponding class, and the structure of the entire search space is drawn as background. The number of points in each vertical third is noted on the right-hand side of the picture to underline the difference between a simple random sampling and such an intuitive sampling.

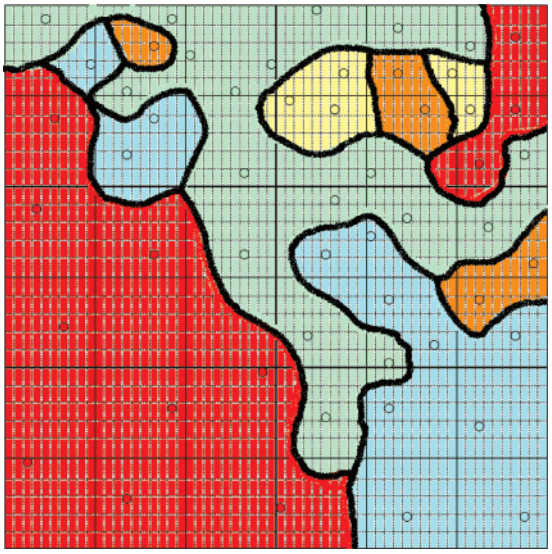


Figure 5. Here, the search space is supposed to be bidimensional and continuous. Considering this hypothesis, the modeled search space with (1 − nearest neighbor) algorithm is drawn. By overlapping perfectly Figures 6 and 7, the recognition rates of *1* − NN could be calculated for each class.

value of v_i for the individual j. Each v_i can be either qualitative or quantitative. A given quantitative feature, v_i discretized by a process ρ_{discr} provides a set of modalities m_i, with Card(m_i) = m_i, m_{i_j}. $j \in [1..m_i]$ is the modality j of v_i. For any variable, the number of modality m is of arbitrary size. MAP is totally independent from the choice of the ML. A classifier c, $C(.) = c(v_1(.), v_2(.), ..., v_n(.))$ is utilized (here, c is a NN for the reasons previously mentioned), which can recognize the class using a list of predictive attributes.

Criterion

The method transfers the points from potentially stable zones of the landscape to unsteady or indecisive ones. The following questions will be answered: How is the "difficulty" of a space zone assessed? How is the necessity to confirm trends and exploration balanced while bearing in mind that deadline is approaching?

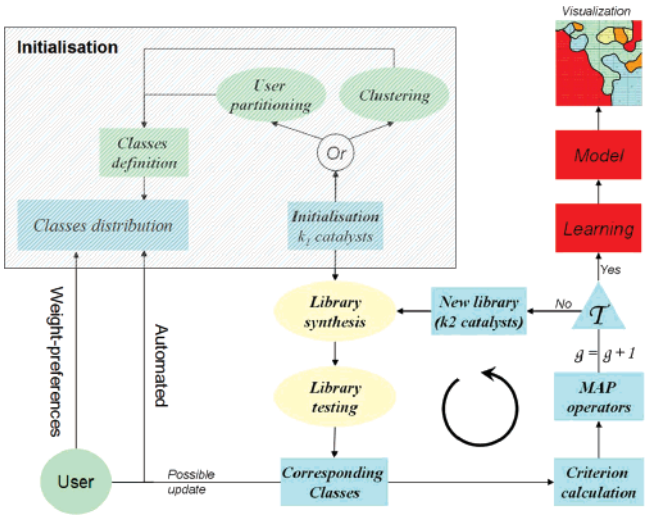


Figure 6. Scheme representing the MAP methodology.

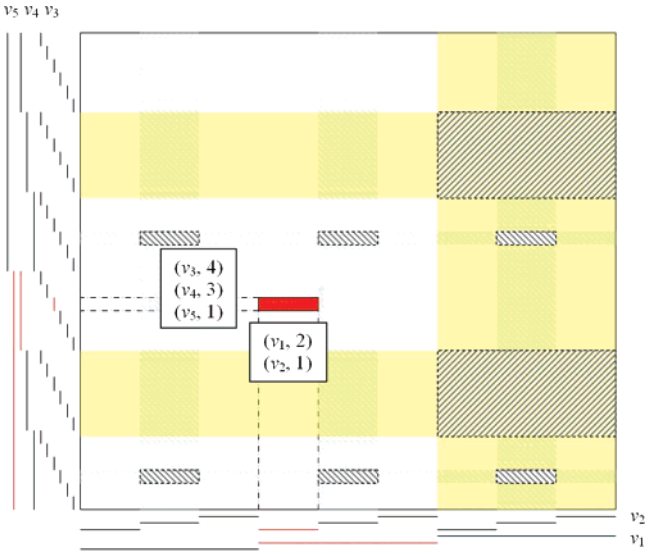


Figure 7. Space zones.

Contingency Analysis and Space Zones. The development of active materials mainly relies on the discovery of strong interactions between elements or, more generally speaking, on the finding of synergism between factors. Each cell of a bidimensional (2D) contingency table, say in row *i* and column *j*, represents the number of elements that have been observed to belong simultaneously to modality *i* of the first variable and to modality *j* of the second variable. The contingency analysis can be extended to higher dimensions and provides results in a format that is straightforward to be transformed in rules for the chemists.[15,16] A zone is defined as a set of *o* variables.[17] Examples are given in Figure 7. The dark area is the "smallest", that is, the most accurate, possible zone, since it is defined on the whole set of variables. "Larger", that is, more general, zones defined by 2 variables are drawn on the Figure 7: {(v_2, 1); (v_3, 3)} in (\\\\) and {(v_1, 3); (v_4, 2)} in (////), where (v_i, j) = m_{ij}. def(v_1, ..., v_n) = {{1..m_{i1}}, {1..m_{i2}}, ..., {1..m_{in}}}. A zone for which only some modalities are specified is noted *s* with

s: def → {m_i, −}, m_i def(v_i), where "−" is the unspecified modality. $o(s)$ is the function that returns the number of defined modalities in *s* (called "order"). Let us consider a search space partitioned into *H* classes and *N* catalysts already evaluated. v_i contains m_i modalities, and n^{i_j} corresponds to the amount of catalysts with m_{ij}. The number of catalysts belonging to the class *h* possessing the modality *j* of the variables v_i is $n_h^{i_j}$. The general notation is summarized in eq 1.

$$n^{i_j} = \sum_{h=1}^{H} n_h^{i_j}, \quad N_i = \sum_{j=1}^{m_i} n_h^{i_j},$$

$$N = \begin{cases} \sum_{h=1}^{H} N_h = \sum_{h=1}^{H} \left(\sum_{j=1}^{m_i} n_h^{i_j} \right) \\ \sum_{j=1}^{m_i} n^{i_j} = \sum_{j=1}^{m_i} \left(\sum_{h=1}^{H} n_h^{i_j} \right) \end{cases} \quad \begin{array}{|cccc|c} n_1^{i_1} & n_1^{i_2} & \cdots & n_1^{i_{m_i}} & N_1 \\ n_2^{i_1} & \ddots & & n_2^{i_{m_i}} & N_2 \\ \vdots & & \ddots & \vdots & \vdots \\ n_H^{i_1} & n_H^{i_2} & \cdots & n_H^{i_{m_i}} & N_H \\ \hline n^{i_1} & n^{i_2} & \ldots & n^{i_{m_i}} & N \end{array} \quad (1)$$

The Chi-Square. The calculation of the statistic called χ^2(chi-square, eq 2) is used as a measure of how far a sample distribution deviates from a theoretical distribution. This type of calculation is referred to as a measure of goodness of fit (GOF).

$$\chi_{i_j}^2 = \sum_{h=1}^{H} \frac{(\text{freq}_h - \text{freq}_h)^2}{h} = \sum_{h=1}^{H} \frac{\left(\frac{n_h^{i_j}}{n^{i_j}} - \frac{N_h}{N} \right)^2}{\frac{N_h}{N}} = N \times \sum_{h=1}^{H} \frac{\left(\frac{n_h^{i_j}}{n^{i_j}} - \frac{N_h}{N} \right)^2}{N_h} \geq 0$$

$$\left. \begin{array}{l} 0 \leq n_h^{i_j} \leq n^{i_j} \Rightarrow 0 \leq \dfrac{n_h^{i_j}}{n^{i_j}} \leq 1 \\ 0 \leq N_h \leq N \Rightarrow 0 \leq \dfrac{N_h}{N} \leq 1 \end{array} \right\} \Rightarrow$$

$$-1 \geq \frac{n_h^{i_j}}{n^{i_j}} - \frac{N_h}{N} \geq 1 \Rightarrow \left(\frac{n_h^{i_j}}{n^{i_j}} - \frac{N_h}{N} \right)^2 \in [0..1] \quad (2)$$

The chi-square can be used for measuring how the classes are disparate into zones as compared to the distribution one gets after the random initialization (k_1 points) or the updated one after successive generations. Therefore, a given number of points can be assigned into zones proportionally to the deviation between the overall distribution and observed distributions into zones. Figure 8 shows a given configuration with $H = 4$, $N = 1000$, and v_i (with $m_i = 5$) that splits the root (i.e., the overall distribution on the left-hand side). For equal distributions between the root and a leaf, chi-square is null (χ^{i5^2}, ■ in Figure 8). Chi-square values are equals for two leaves with the same distribution between each other (● in Figure 8). One would prefer to add a point with the third modality (bottom ●) to increase the number of individuals, which is relatively low. This is confirmed by the fact that χ^2 is relatively more "reactive" for leaves with

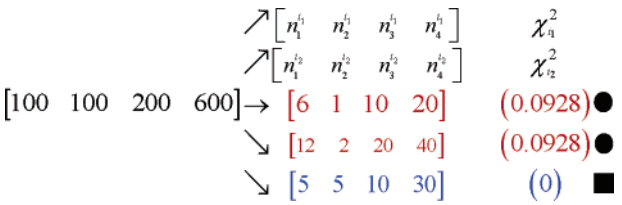


Figure 8. Criterion settings, first configuration. On the left-hand side is represented the entire search space. This given root has been split into five leaves, for which the distributions are given for the last three. Each leaf and the root are partitioned into five classes. The first class has received 100 elements, and among these, 12 belong to the fourth leaf. The chi-square statistic is given on the right-hand side of each leaf between brackets.

smaller populations (see the absolute variations (■ → ● and □ → ○) of two successive χ^2 in Figure 9). To obtain a significant impact, that is, information gain, by adding a new point, it is more interesting to test new catalysts possessing a modality which has been poorly explored (i.e., □). Chi-square does not make any difference between leaves that exhibit exactly the same distribution (i.e., □ and ■). Therefore, n^{i_j} must be minimized at the same time to support relatively empty leaves.

The MAP Criterion. On the basis of the chi-square behavior, the MAP criterion is defined as $(\chi^2 + 1) \times (n^{i_j} + 1)^{-1}$. Extremely unstable and small zones may have distributions that are very far from the overall distribution. With this criterion, they may continuously attract experiments; however, this may not be due to a natural complex underlying relationship but, rather, to lack of reproducibility, uncontrolled parameters, noise, etc. Therefore, the maximum number of experiments a zone can receive is bounded by the user. X^o_{rndk2} is the calculated average number of individuals that a zone of order *o* receives from a SRS of k_2 points. A maximum number of points noted $\rho X^o_{\text{rndk2}+k1}$ that MAP is authorized to allocate in a zone compared to X^o_{rndk2} can be decided. ρ is a parameter the user has to set.

After the distribution zone analysis done after each new selected generation, the algorithm ranks them on the basis of the MAP criterion. Among the whole set of zone, t_s, called (tournament size) zones, are selected randomly and compete together following the GA-like selection operator, called a "tournament".[18,19] A zone with rank *r* has a $2r \times [k_2(k_2 + 1)]^{-1}$ chance to be selected. As the criterion is computed on subsets of modalities (i.e., zone of order *o*), when a given zone is selected for receiving new points, the modalities that do not belong to *s* are randomly assigned.

The class concept is of great importance, since the criterion deeply depends on the root distribution. Enlarging or splitting classes permits an indirect control of the sampling. It is recommended that "bad" classes be merged and good ones be split to create relatively unbalanced root distributions. A reasonable balance must be respected; otherwise, small and interesting classes hidden in large ones will have less chance of being detected. In the experiments presented in the next section, *o* remains fixed and is a priori set. For each zone of order *o*, the corresponding observed distribution and the related MAP criterion value are associated.

Figure 9. Criterion settings, second configuration.

Benchmarking

In most cases, benchmarking is not performed with a sufficient number of different problems. Rarely can the results presented in articles be compared directly. Often, the benchmark setup is not documented well enough to be reproduced. It is impossible to say how many datasets would be sufficient (in whatever sense) to characterize the behavior of a new algorithm. With a small number of benchmarks, it is impossible to characterize the behavior of a new algorithm in comparison to known ones. The most useful setup is to use both artificial datasets,[20] whose characteristics are known exactly, and real datasets, which may have some surprising and very irregular properties. Ref 21 outlines a method for deriving additional artificial datasets from existing real datasets with known characteristics; the method can be used if insufficient amounts of real data are available or if the influence of certain dataset characteristics are to be explored systematically. Here, two criteria are emphasized to test this new algorithm: (i) Reproducibility. In a majority of cases, the information about the exact setup of the benchmarking tests is insufficient for other researchers to exactly reproduce it. This violates one of the most basic requirements for valid experimental science.[22] (ii) Comparability. A benchmark is useful if results can be compared directly with results obtained by others for other algorithms. Even if two articles use the same dataset, the results are most often not directly comparable, because either the input/output encoding or the partitioning of training versus test data is not the same or is even undefined.

The efficiency of the MAP method is thoroughly evaluated with mathematical functions. These benchmarks may be represented on multidimensional graphics transformed to bidimensional charts called "maps". Their construction is first carefully detailed, and then benchmarks are presented.

Creation of Benchmarks. A benchmark is built after 3 steps: (i) *n*-Dimension functions are traced onto a first bidimensional series plot. (ii) Classes of performances are constructed by setting thresholds on the *y* axis of the series plot. (iii) Between two thresholds, every point corresponding to a given class is labeled. On the basis of these classes, the map is created. Each variable of a given function f is continuous $f(x_i) \rightarrow y \in \mathbb{R}$. For simplicity, all of the variables for a given function are defined on the same range $\forall i$, $x_i \in [a \ldots b]$, $(a,b) \in \mathbb{R}$. The range is cut into pieces. $\mathscr{P}_{\text{discr}}$ splits $[a \ldots b]$ into m_i equal parts ($\forall i$, $m_i = m$). All the boundaries ($m + 1$) are selected as points to be plotted in the series plot. On the *x* axis, an overlapped loop is applied taking into account the selected values of each variable. As example, let us consider Baumes f_g function (eq 3). Figure 10 shows the associated series plot with $n = 6$ and $x_i \in [-1..1]$. An overlapped loop is used on each feature with nine points for each, that is, 531 441 points in total. This procedure permits one to simply determine the different levels that will be used for partitioning the performance and, thus, to establish the different classes. The size of each class (i.e., the number of points between two thresholds) is, thus, easily visualized by splitting the series plot with thresholds (horizontal lines in Figure 10). One color and form is assigned to each class: blue $\leq$ 2,2 $<$ aqua $\leq$ 6,6 $<$ green $\leq$ 10, 10 $<$ yellow $\leq$ 15, red $>$ 15. Figure 11 gives an example of the placement of points for creating the map. Figure 12 shows the map corresponding to Figure 10 (eq 3).

Selection of Benchmarks. Five different benchmarks (De Jong f_1 and De Jong f_3,[23] Schwefel f_7,[24] Baumes f_a, and Baumes f_g; see eq 3) have been selected to test the algorithm. Among them, some new ones (Baumes f_a and Baumes f_g) have been specially designed to trap the method and, thus, to reveal MAP limits. The maps are presented in the Supporting Information.

$$f_a(x_i) = \left|\tan\left\{\sum_{i=1}^{n}\left(\sin^2\left((x_i^2 - 1/2)/\left[1 + \frac{x_i}{1000}\right]^2\right)\right)\right\}\right| \quad 0 \leq x_i \leq 2$$

$$f_g(x_i) = \sum_{i=1}^{n}((n - i + 1) \times x_i^2) \quad -1 \leq x_i \leq 1$$

$$f_1(x_i) = \sum_{i=1}^{n} x_i^2 \quad 0 \leq x_i \leq 6$$

$$f_3(x_i) = A + \sum_{i=1}^{n} \text{int}(x_i) \quad A = 25 \text{ (option)} \quad 0 \leq x_i \leq 3$$

$$f_7(x_i) = nV + \sum_{i=1}^{n} - x_i \times \sin(\sqrt{|x_i|}) \quad -500 \leq x_i \leq 500 \quad (3)$$

Results

MAP samples and the corresponding effect on NN learning are compared to SRS. An introduction of NNs as a classifier for catalysts is thoroughly depicted in ref 25. The dataset is always separated into a training test and a selection test to prevent overfitting. The problem of overfitting is discussed in ref 26. The use of analytical benchmarks permits the utilization of test sets with an arbitrary number of cases. For each sample (both from MAP and SRS), 10 000 individuals are randomly chosen as test set. As an example, 1500 points have been sampled on De Jong f_1[23] search space ($9_{\text{var.}}/4_{\text{mod.}}$) (see Table 1 for both SRS and MAP). When using MAP, the number of good individuals (class A, the smallest) is increased from 4 with SRS (training + selection A) to 27 with MAP. The distribution on the search space with MAP permits one to increase both the overall rate of recognition and the recognition of small classes. For the other benchmarks, the distributions in the merged training and selection sets are given in Table 2, whereas the distribution in the test sets are shown in Table 3. It can be seen in the respective distributions of every tested benchmark that small classes

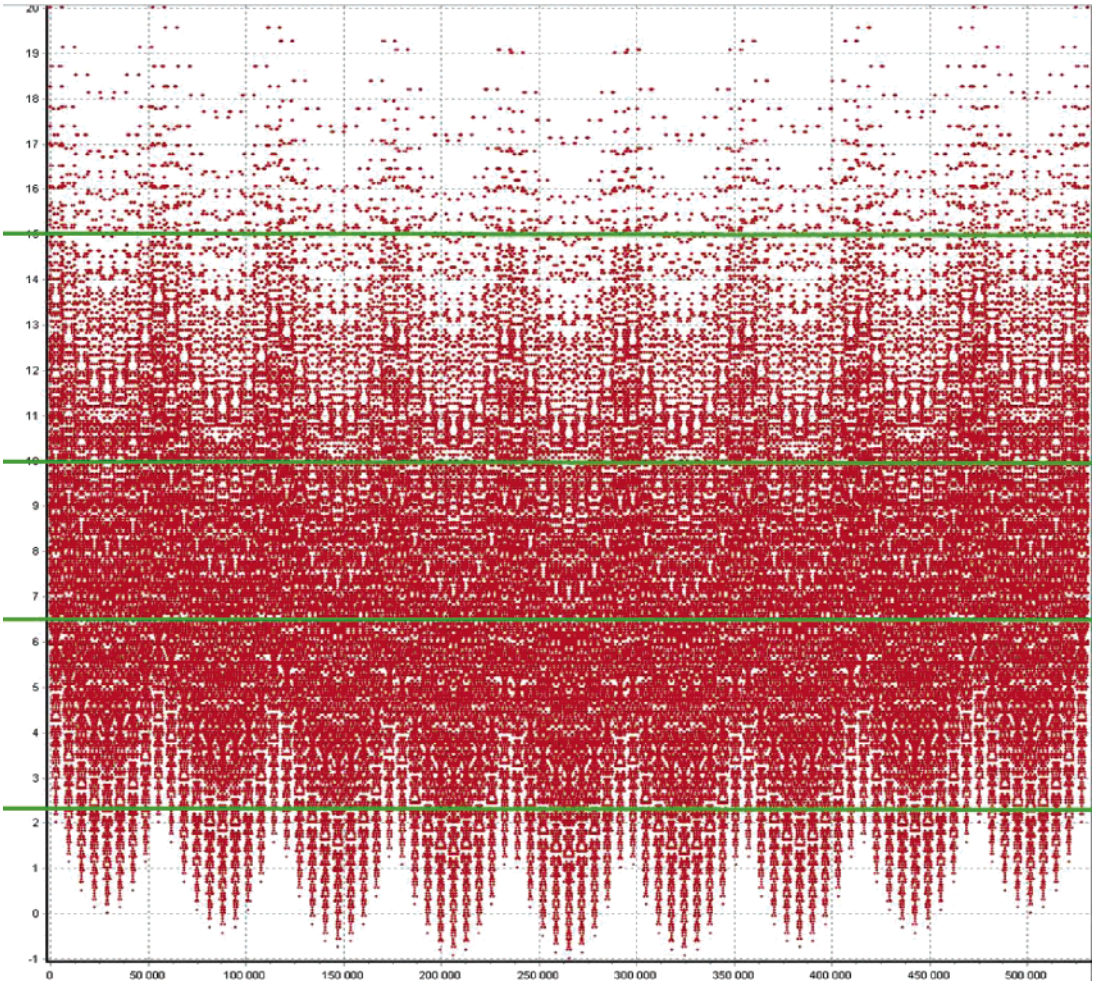
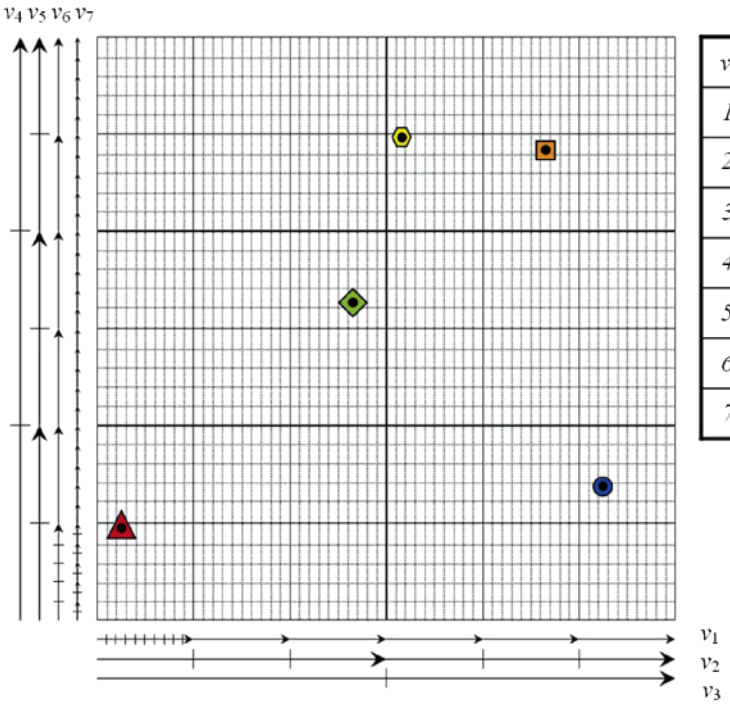


Figure 10. Series plot Baumes f_g. The number of variables noted, $n = 6$, and the number of points represented for each feature is 9.



v	Modalities or levels	⬡	▲	◆	●	■
1	10	2	3	7	3	7
2	3	1	1	3	3	2
3	2	2	1	1	2	2
4	3	3	1	2	1	3
5	2	1	1	2	2	1
6	5	5	5	2	2	5
7	2	2	2	1	2	1

Figure 11. Multidimensional function represented onto 2D space called "map".

received more experiments (the smallest class is in gray), and the larger ones lost a part of their effect (the largest is in black), as was expected if using MAP instead of SRS.

Results show clearly that MAP permits a better characterization of small zones than does SRS while exploration of the search space is perfectly maintained. The gain of

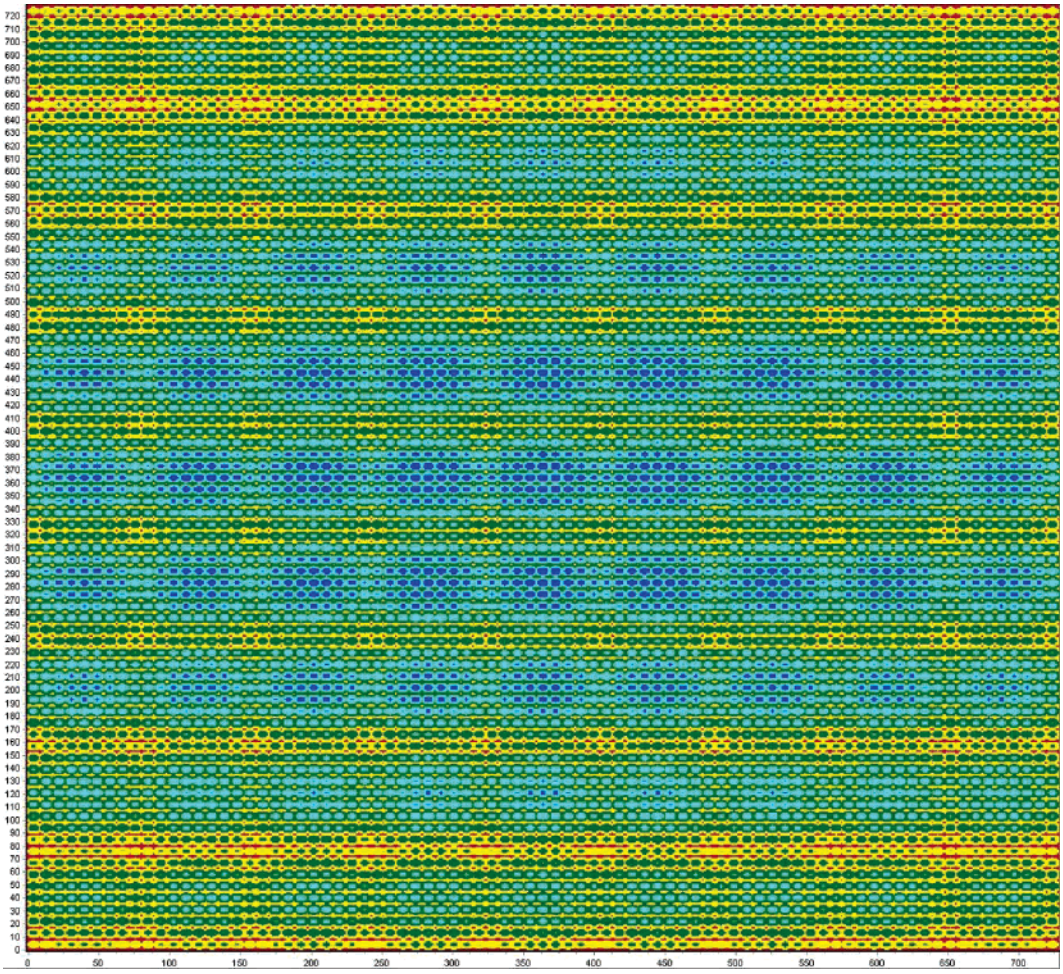


Figure 12. 2D map for function Baumes f_g ($n = 6$; 9 pts/var).

Table 1. Training, Selection, and Test Sets of De Jong f_1 from SRS (Upper Array) and MAP (Lower Array)

SRS

	Training					Selection					Test				
	A	B	C	D	E	A	B	C	D	E	A	B	C	D	E
A	1	0	0	0	0	1	0	0	0	0	9	4	0	0	0
B	0	47	0	0	0	2	30	10	0	0	51	476	175	0	0
C	0	0	102	0	0	0	8	94	6	0	0	62	1139	111	0
D	0	0	0	231	0	0	0	0	181	12	0	0	143	2509	164
E	0	0	0	0	392	0	0	0	14	361	0	0	0	186	4971

MAP

	Training					Selection					Test				
	A	B	C	D	E	A	B	C	D	E	A	B	C	D	E
A	13	0	0	0	0	11	0	0	0	0	20	12	0	0	0
B	0	110	1	0	0	4	75	14	0	0	40	486	93	0	0
C	0	0	178	0	0	0	7	148	11	0	0	44	1242	113	0
D	0	0	0	208	0	0	0	14	206	3	0	0	122	2592	93
E	0	0	0	0	263	0	0	0	3	231	0	0	0	101	5042

Table 2. Merged Training and Selection Sets after Sampling from MAP and SRS Considering All Other Benchmarks[a]

	De Jong F3		Baumes Fa		Baumes Fg		Schwefel F7	
	SRS	MAP	SRS	MAP	SRS	MAP	SRS	MAP
A	15	58	772	737	58	111	5	25
B	31	123	300	308	397	448	80	180
C	85	196	273	284	592	452	200	320
D	139	213	85	82	402	386	412	402
E	1230	910	70	89	51	103	803	573

[a] In each case, five classes are present (A−E).

recognition by NN on both the smallest and the largest classes for each benchmark using MAP instead of SRS is given in Figure 13. It can be seen that the gains on the smallest classes are tremendously increased, varying from 18% to an infinite gain. In Figure 13 and for the benchmark called "Schwefel f7",[24] a value of 600 is indicated (for infinite) since we have assigned one experiment into the smallest zone so as not to obtain a zero division. The loss of recognition rate for the largest classes (if there is loss) is very low compared to the high gain on small ones. Such loss is <22%, showing clearly that the exploration remains nearly perfect. The overall recognition rate being deeply influenced by the relative size of classes does not represent an adequate criterion; however, MAP outperforms SRS in most of the cases.

During all the experiments, the root distribution has been fixed from the beginning. One has to note that the user could intentionally not respect the real distribution in order to give weights on selected classes as presented earlier in the definition of the criterion. The methodology for reevaluating the root distribution is quickly presented in the available Supporting Information.

Discussion and Further Analysis

Is MAP Distribution Significantly Different from SRS Sampling? Because MAP is not influenced by the choice

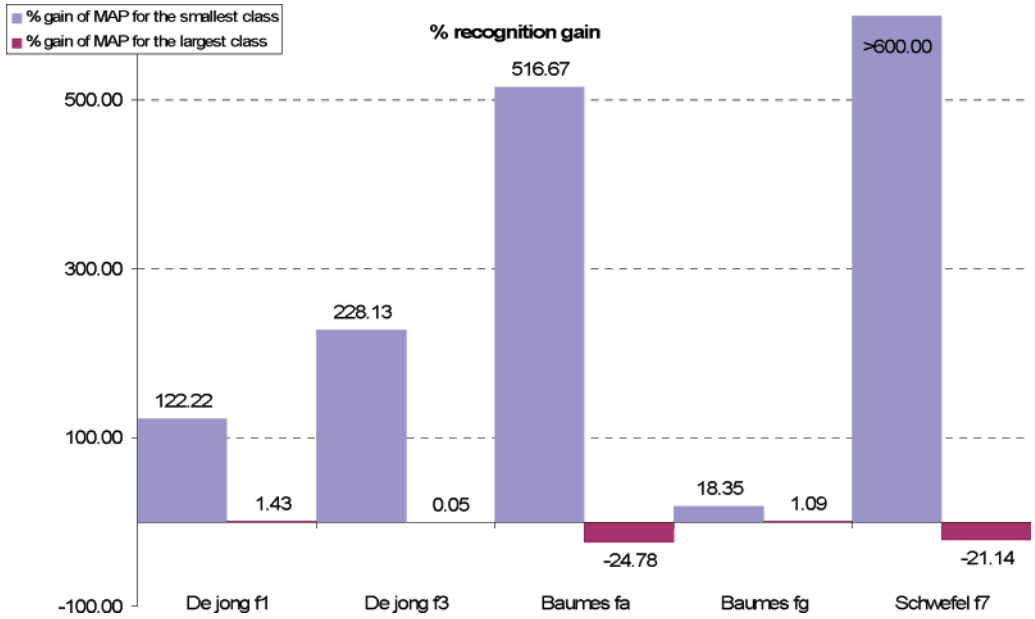


Figure 13. Percentage recognition gain for both the smallest and largest class considering every benchmark when using MAP methodology instead of SRS.

Table 3. Distribution of Classification by Neural Network in Test Depending on the Sample (SRS or MAP) for All Benchmarks

			Real																							
			De Jong F1					De Jong F3					Baumes Fa					Baumes Fg					Schwefel F7				
			A	B	C	D	E	A	B	C	D	E	A	B	C	D	E	A	B	C	D	E	A	B	C	D	E
Predicted	SRS	A	9	4	0	0	0	32	31	0	0	0	3471	917	642	188	201	248	109	0	0	0	52	36	0	0	0
		B	51	476	175	0	0	73	217	30	0	0	695	469	456	174	125	97	2360	231	0	0	50	3212	223	0	0
		C	0	62	1139	111	0	0	7	517	0	0	756	433	517	130	137	0	362	3488	243	0	0	279	5072	160	0
		D	0	0	143	2509	164	0	0	0	880	4	166	177	157	74	53	0	0	231	2225	103	0	0	155	741	9
		E	0	0	0	186	4971	0	0	0	5	8204	20	11	17	8	6	0	0	0	85	218	0	0	0	11	0
	MAP	A	20	12	0	0	0	105	0	6	0	0	2611	976	862	300	265	275	84	0	0	0	67	45	0	0	0
		B	40	486	93	0	0	0	255	0	0	0	1107	466	377	113	98	70	2481	181	0	0	35	3141	373	0	0
		C	0	44	1242	113	0	0	0	541	0	0	835	351	345	85	98	0	266	3526	175	0	0	341	4800	108	0
		D	0	0	122	2592	93	0	0	0	885	0	290	121	100	51	24	0	0	243	2273	63	0	0	275	789	5
		E	0	0	0	101	5042	0	0	0	0	8208	265	93	105	25	37	0	0	0	105	258	0	0	2	15	4

of the ML applied on selected points, another way to gauge the influence of MAP is to analyze the distribution of points. Therefore, if the overall distribution of classes on the whole search space is statistically similar to an SRS, the MAP method does *not* transfer a point from zone to zone.

The chi-square test[27] is used to test if a sample of data comes from a population with a specific distribution. The chi-square GoF test is applied to binned data (i.e., data put into classes) and is an alternative to the Anderson−Darling[28] and Kolmogorov−Smirnov[29] GOF tests, which are restricted to continuous distributions. In statistics, the researcher states as a "statistical null hypothesis", noted H_0, something that is the logical opposite of what it is believed. Then, using statistical theory, it is shown from the data that H_0 is false and should be rejected. This is called "reject−support testing" (RS testing) because rejecting the null hypothesis supports the experimenter's theory. Consequently, before undertaking the experiment, one can be certain that only 4 possible things can happen. These are summarized in the Table 4. Therefore, statistic tests with υ df ($\upsilon = (l-1)(c-1) = 4$) are computed from the data (Table 5). H_0: MAP = SRS, H_1 MAP ≠ SRS is tested. For such an upper one-sided test, one finds the column corresponding to α in the upper critical values table

Table 4. Statistical Hypothesis Acceptances and Rejections

		State of the world	
		H_0	H_a
Decision	H_0	Correct acceptance	Type 2 error β
	H_a	Type 1 error α	Correct rejection

and rejects H_0 if the statistic is greater than the tabulated value. The estimation and testing results from contingency tables hold regardless of the distribution sample model. Top values in Table 5 are frequencies calculated from Table 1. The chi-square $\chi_\upsilon^2 = \sum (f_{\text{observed}} - f_{\text{theoritical}})^2 \times f_{\text{theoritical}}^{-1}$ is noted in red and the critical values at a different level are in blue. Yes (Y) or no (N) correspond to answers to the question, "Is H_0 rejected?". Table 5 shows that MAP distribution differs from SRS for some cases only. One can note that negative answers are observed on two benchmarks, called Baumes f_a and Baumes f_g (the black cell is discussed later). These benchmarks have been created to check MAP efficiency on extremely difficult problems; however, the analysis of the results in the previous section clearly shows that MAP modifies the distributions and, thus, implies improvement of search space characterization through ML. Therefore, the sample size is thought not to be large enough to discriminate both approaches.

Table 5. Chi-Square GOF Test

		Baumes fg		De jong f3		De jong f1		Baumes fa		Schwefel f7	
		SRS	MAP	SRS	MAP	SRS	MAP	SRS	MAP	SRS	MAP
		3,86666667	7,4	1	3,86666667	0,53333333	1,86666667	51,4666667	49,1333333	0,33333333	1,66666667
		26,4666667	29,8666667	2,06666667	8,2	5,33333333	12,8	20	20,5333333	5,33333333	12
		39,4666667	30,1333333	5,66666667	13,0666667	14,5333333	23,6666667	18,2	18,9333333	13,3333333	21,3333333
		26,8	25,7333333	9,26666667	14,2	28,3333333	28,5333333	5,66666667	5,46666667	27,4666667	26,8
		3,4	6,86666667	82	60,6666667	51,2666667	33,1333333	4,66666667	5,93333333	53,5333333	38,2
Alpha	Critical value	6,75937209		20,11939707		18,75811101		0,430795216		13,94169525	
0,1	7,779	N		Y		Y		N		Y	
0,5	9,488	N		Y		Y		N		Y	
0,25	11,143	N		Y		Y		N		Y	
0,01	13,277	N		Y		Y		N		Y	
0,001	18,467	N		Y		Y		N		N	

Does MAP Really Move Points from Zones to Zones in the Search Space? Moving points into search space is a fact, but transferring individuals from stable zones to puzzling ones is different. Therefore, new tests have been performed. The overall distribution is split on the basis of a set of modalities or a given number of variables, and a new chi-square GoF is evaluated (eq 4).

$$\begin{array}{cccccc} & A & & B & C\ D & E \\ m_1 & n_1^A(\tilde{E}(n_1^A)) & & n_1^B(\tilde{E}(n_1^B)) & \cdots\ \cdots & n_1^E(\tilde{E}(n_1^E)) \\ \vdots & \vdots & & & & \vdots \\ m_i & n_i^A(\tilde{E}(n_i^A)) & & n_i^B(\tilde{E}(n_i^B)) & \cdots\ \cdots & n_i^E(\tilde{E}(n_i^E)) \end{array}$$

$$\Rightarrow \chi^2 = \sum_{j=1}^{j=i} \sum_{h=A}^{h=E} \frac{[n_j^h - \tilde{E}(n_j^h)]^2}{\tilde{E}(n_j^h)} \quad (4)$$

If $i = 3$, then $\upsilon = 6$, and the critical value is $\chi_{0.05(6)}^2 = 12.5916$. H_0 is accepted when no difference in zone size is observed for the considered variables on a given benchmark and also that H_0 is rejected when a clear difference appears. Tables from these tests are not presented. With "easy" benchmarks, it appears clearly that MAP acts as expected. However, for one case, H_0 is accepted, but this does not imply that the null hypothesis is true; it may mean that this dataset is not strong enough to convince that the null hypothesis is not true. To conclude that MAP action is not statistically significant when the null hypothesis is, in fact, false is called a "type II error". Thus, the power of the test is finally discussed.

Chi-Square Power. There are two kinds of errors represented in the Table 5. The power testing procedure is set up to give H_0 "the benefit of the doubt"; that is, to accept H_0 unless there is strong evidence to support the alternative. Statistical power $(1 - \beta)$ should be at least 0.80 to detect a reasonable departure from H_0. The conventions are, of course, much more rigid with respect to α than with respect to β. Factors influencing power in a statistical test include (i) What kind of statistical test is being performed; some statistical tests are inherently more powerful than others. (ii) Sample size. In general, the larger the sample size, the larger the power.

To ensure a statistical test will have adequate power, one usually must perform special analyses prior to running the experiment to calculate how large a sample size (noted n) is required. One could plot power against sample size, under the assumption that the real distribution is known exactly. The user might start with a graph that covers a very wide range of sample sizes to get a general idea of how the statistical test behaves; however, this work goes beyond the topic of this paper. The minimum required sample size that permits one to start discriminating (significantly, with a fixed error rate α) MAP from SRS is dependent on the search space landscape. This simulation will be investigated in future work. It needs to be noted that 1500 points have been selected for each benchmark; however, the search spaces are extremely broad, and thus, such a sample size represents only a very small percentage of the entire research space.

Conclusion

There are several motivations for wanting to alter the selection of samples. In a general sense, we want a learning system to acquire knowledge. In particular, we want the learned knowledge to be as generally useful as possible while retaining high performance. If the space of the configurations is very large with much irregularity, then it is difficult to adequately sample enough of the space. Adaptive sampling, such as MAP, tries to include the most productive samples. Such adaptive sampling allows selecting a criterion over which the samples are chosen. The learned knowledge about the structure is used for biasing the sampling.

MAP has been thoroughly presented and tested. As such, this methodology was developed to propose formulations that are relevant for testing at the very first stage of a HT program when numerous but inherent constraints are taken into account for the discovery of new performing catalysts. No comparative study has been found in the literature for when such a methodology is flexible enough to be applied on a broad variety of domains. The main advantages are the following: The number of false negatives is highly decreased while the number of true positives is tremendously increased. MAP is totally independent of the classifier and creates more balanced learning sets, permitting both preventing over-learning, and gaining higher recognition rates. All previous experiments can be integrated, giving more strength to the method, and any type of feature is taken into account. The method is tunable through the modification of the root distribution.

Acknowledgment. Ferdi Schueth and Katarina Klanner from Max-Planck-Institut für Kohlenforschung, Mülheim, Germany, and Claude Mirodatos and David Farrusseng from CNRS-Institut de Recherche sur la Catalyse, Villeurbanne, France, are gratefully acknowledged for the discussions which have permitted elaboration of such an approach. EU

Commission (TOPCOMBI Project) support is gratefully acknowledged for this research.

Supporting Information Available. Supporting Information as noted in text. This material is available free of charge via the Internet at http://pubs.acs.org.

References and Notes

(1) Senkan, S., Ed. *Angew. Chem., Int. Ed.* **2001**, *40* (2), 312−329.
(2) Bem, D. S.; Erlandson, E. J.; Gillespie, R. D.; Harmon, L. A.; Schlosser, S. G.; Vayda, A. J. In *Experimental Design for Combinatorial and High Throughput Materials Development*; Cawse, J. N., Ed.; Wiley and Sons, Inc.: Hoboken, NJ, 2003; pp 89−107.
(3) Cawse, J. N.; Wroczynski, R. In *Experimental Design for Combinatorial and High Throughput Materials Development*; Cawse, J. N., Ed.; Wiley and Sons, Inc.: Hoboken, NJ, 2003; pp 109−127.
(4) Serra, J. M.; Corma, A.; Farrusseng, D.; Baumes, L. A.; Mirodatos, C.; Flego, C.; Perego, C. *Catal. Today* **2003**, *81*, 425−436.
(5) Sjöblom, J.; Creaser, D.; Papadakis, K. *11th Nordic Symposium on Catalysis*, Oulu, Finland, 2004.
(6) Harmon, L. A. *J. Mater. Sci.* **2003**, *38*, 4479−4485.
(7) Farrusseng, D.; Klanner, C.; Baumes, L. A.; Lengliz, M.; Mirodatos, C.; Schüth, F. *QSAR Comb. Sci.* **2005**, *24*, 78−93.
(8) Deming, S. N.; Morgan, S. L. *Experimental Design: A Chemometric Approach*, 2nd ed.; Elsevier Science Publishers B.V., Amsterdam, The Netherlands, 1993.
(9) Montgomery, D. C. *Design and Analysis of Experiments*, 3rd ed.; Wiley: New York, 1991.
(10) Tribus, M.; Sconyi, G. *Qual. Prog.* **1989**, *22*, 46−48.
(11) Klanner, C.; Farrusseng, D.; Baumes, L. A.; Lengliz, M.; Mirodatos, C.; Schüth, F. *Angew. Chem., Int. Ed.* **2004**, *43*, 5347−5349.
(12) Fernández, J.; Kiwi, J.; Lizama, C.; Freer, J.; Baeza, J.; Mansilla, H. D.; *J. Photochem. Photobiol., A* **2002**, *151*, 213−219.
(13) Sammut, C.; Cribb, J. In *7th Int. Machine Learning Conf.*, Austin, Texas, , 1990.
(14) Cover, T. M.; Hart, P. E. *IEEE Trans. Inf. Theory* **1967**, *13*, 21−27.
(15) Farrusseng, D.; Baumes, L. A.; Mirodatos, C. In *High-Throughput Analysis: A Tool For Combinatorial Materials Science*; Potyrailo, R. A., Amis, E. J., Eds.; Kluwer Academic/Plenum Publishers: Norwell, MA, 2003; pp 551−579.
(16) Farrusseng, D.; Tibiletti, D.; Hoffman, C.; Quiney, A. S.; Teh, S. P.; Clerc, F.; Lengliz, M.; Baumes, L. A.; Mirodatos, C. *13th ICC*, Paris, France, July 11−16, 2004.
(17) Baumes, L. A.; Jouve, P.; Farrusseng, D.; Lengliz, M.; Nicoloyannis, N.; Mirodatos, C. *7th Int. Conf. on Knowledge-Based Intelligent Information & Engineering Systems* (KES '2003), University of Oxford, U.K.; Springer-Verlag: New York, 2003; Palade, V., Howlett, R. J., Jain, C., Eds.; In Lecture Notes in AI; LNCS/LNAI Series.
(18) Blickle, T.; Thiele, L. *6th Int. Conf. Genet. Algorithms*; Morgan Kaufmann: San Mateo, 1995.
(19) Thierens, D. *Proc. 7th Int. Conf. Genet. Algorithms*, ICGA-97, 1997; pp 152−159.
(20) Hickey, R. J. Machine Learning. In *Proc. 9th Int. Workshop*; Sleeman, D., Edwards, P., Eds.; Morgan Kaufman: San Mateo, CA, 1992, pp 196−205.
(21) Aha, D. W. Machine Learning. In *Proc. 9th Int. Workshop*; Sleeman, D., Edwards, P., Eds.; Morgan Kaufman: San Mateo, CA, 1992, pp 1−10.
(22) Christensen, L. B. *Experimental Methodology*, 6th ed.; Allyn and Bacon: Needham Heights, MA, 1994.
(23) De Jong, K. A. Doctoral dissertation, University of Michigan, 1975; Dissertation Abstract International, 36(10), 5140(B); University of Michigan Microfilms no. 76-9381.
(24) Whitley, D.; Mathias, K.; Rana, S.; Dzubera, J. *Artif. Intell.* **1996**, *85*, 245−276.
(25) Baumes, L. A.; Farruseng, D.; Lengliz, M.; Mirodatos, C. *QSAR Comb. Sci.* **2004**, *29*, 767−778.
(26) Baumes, L. A.; Serra, J. M.; Serna, P.; Corma, A. *J. Comb. Chem.*, submitted.
(27) Snedecor, G. W.; Cochran, W. G. *Statistical Methods*, 8th ed.; Iowa State University Press: Ames, IA, 1989.
(28) Stephens, M. A. *J. Am. Stat. Assoc.* **1974**, *69*, 730−737.
(29) Chakravarti, L.; Roy, H. L. *Handbook of Methods of Applied Statistics*; John Wiley and Sons: New York, 1967; Vol. 1, pp 392−394.

CC050130+