Catalysis Today 159 (2011) 55–63

Contents lists available at ScienceDirect

Catalysis Today

journal homepage: www.elsevier.com/locate/cattod

Efficient discovery and optimization of complex high-throughput experiments

James N. Cawse[a,b,*], Gianluca Gazzola[a], Norman Packard[a,c]

[a] ProtoLife Inc., 57 Post St. #513, San Francisco, CA 94014, USA
[b] Cawse and Effect LLC, 132 Kittredge Rd., Pittsfield, MA 01201, USA
[c] European Center for Living Technology, Calle del Clero 2940, 30124 Venice, Italy

ARTICLE INFO

Article history:
Available online 12 August 2010

Keywords:
Evolutionary design of experiments
High-throughput
Response surface
Experimental space
Optimization
Machine learning

ABSTRACT

As the pace of experimentation in materials science and catalysis has increased, experimental tactics and strategies have had to adapt to meet the demands of goals of experimentalists, and the spaces they explore. This pace has increased from runs/year to runs/day and sometimes to runs/minute in high-throughput experimentation. Although much of this capacity is used to simply speed up conventional experimental designs, the leading-edge application is discovery of low-probability, high-value occurrences (hits) by searching extensive, complex experimental spaces. Conventional design of experiments (DoE) is not capable of dealing with these issues. Instead, more advanced experimental tactics and strategies must be implemented. After introducing the elements that make an experimental campaign complex, here we present a novel statistical model-based evolutionary experimental strategy and apply it to the optimization of a family of artificial complex systems. With our experiments, we show that such a strategy may significantly reduce the experimental effort required for finding the optima compared to other state-of-the-art evolutionary strategies.

1. Introduction

Since the modern era of high-throughput chemistry and materials science began in the 1990s, experimenters have innovated novel experimental designs to attack the new, larger problems that became accessible with large numbers of experiments [1,2]. These included novelties such as fractal designs [3], "unpeeled" dense mixture designs [4], edge-sharing [5] and split-plot designs [6]. Although enthusiasm for the more elaborate of these designs has abated somewhat as the field has matured [7], problems still appear in which the experimental space is large and complex. These have the issues of:

- High-dimensional experimental spaces (many system constituents and experimental parameters);
- Complex constraints on the independent variables;
- Synergies, or beneficial nonlinear interactions between system constituents;
- Unpredictable behavior and the inability to derive experimental results from basic chemical and physical laws.

Addressing these complex problems requires a strategic approach to experimentation. Strategy (the overall approach) is distinct from tactics (the conduct of an individual experiment). The designs mentioned above can be tactical elements in the strategic plan, but one-design-at-a-time experimentation should be considered as obsolete as one-factor-at-a-time experimentation.

In this paper we first formulate the problem, defining experimental spaces with both qualitative and quantitative variables, and with a response surface representing experimental measurement. We then examine both tactical and strategic approaches to the problem, including experimental space sampling via a genetic algorithm (GA), and we introduce a novel form of evolutionary design of experiments (Evo-DoE), which combines statistical modeling and stochastic sampling. We finally illustrate Evo-DoE with a numerical case study, comparing it with a GA.

1.1. Experimental spaces and experimental response

The size of an experimental space is a fuzzy concept that requires exact definition for each experimental context. For the definition of the experimental space, consider all controllable variables (factors) that could affect an experimental result, both quantitative (including concentration levels) and qualitative. Each factor is one dimension of the experimental space. Typically the quantitative factors are explored for a specific set of levels, so that there are a finite number of levels for each factor. The number of possible

* Corresponding author at: Cawse and Effect LLC, 132 Kittredge Rd., Pittsfield, MA 01201, USA. Tel.: +1 413 822 5006.
E-mail address: cawse@cawseandeffect.com (J.N. Cawse).

0920-5861/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.cattod.2010.05.043

Table 1
Factors for one-step DPC synthesis ($1 \times 15 \times 15 \times 15 \times 9 \times 4 \times 3 \times 3 \times 3 = 3{,}280{,}500$ potential experiments).

Factors in typical screening experiment	No. of levels
Principal metal catalyst	1
Inorganic co-catalysts	15
Metal ligands	15
Organic co-catalysts	15
Anion	9
Associated cation	4
Reaction time	3
Reaction temperature	3
Reaction pressure	3

experiments, or size of the experimental space, is obtained by multiplying the number of levels for each factor, both qualitative and quantitative. This may be reduced by system constraints (e.g., in mixture experiments). For each experiment, we assume that an experimental response may be measured, thus defining a response surface over the experimental space.

Experimenters have some flexibility in determining the experimental space, in that they must choose which variables will be varied in a particular experiment. Out of all possible controllable variables, the experimenter may choose to fix some and vary others. In fact, some potentially controllable variables, such as intermediates in a complex reaction, may be unknown at the beginning of an experimental program. The experimenter's choice of experimental space may depend on his or her perception of the curvature or irregularity of the response surface. We will take as an operational definition of experimental space, for a given experimental program, the space of variables the experimenter chooses to explore, i.e., the factors chosen to vary.

The GE Global Research team determined just such an experimental space in their classic study of the carbonylation of phenol to diphenyl carbonate (DPC) (Table 1) [8]. The immensity of this space and the value of the potential catalyst justified a massive effort exceeding 22,000 experimental runs, subsampling the full experimental space in various non-systematic ways [9].

In approaching experimentation in these spaces, there are two critical questions that must be asked. First, is the space "too large" for the tools at hand? Is the number of factors, levels, and interactions too great, and is the response surface too irregular? If so, the experiment must be pruned using the best chemical knowledge available. Second, if the resources available appear to meet the challenge, what is the best strategy for attacking the project at hand? Some of these questions are addressable only with preliminary experimentation, e.g., to gain knowledge of the response surface.

1.2. Experimental response surface topology

One way to help our thinking in this area is to consider some idealized response surfaces and the consequences of the shapes of these surfaces. There are two extremes in this area. One is a surface over a space consisting of multilevel entirely qualitative factors. The other is a surface over a space with purely quantitative factors.

1.2.1. Qualitative spaces

The ultimate difficulty for a discovery program is the true "needle in the haystack," illustrated in Fig. 1 as a response surface over a 2-dimensional experimental space consisting of two qualitative factors with many levels each. In such a system, the optimum result appears only when all of the factors are at their precisely correct levels. No experimental design can find this peak more reliably than

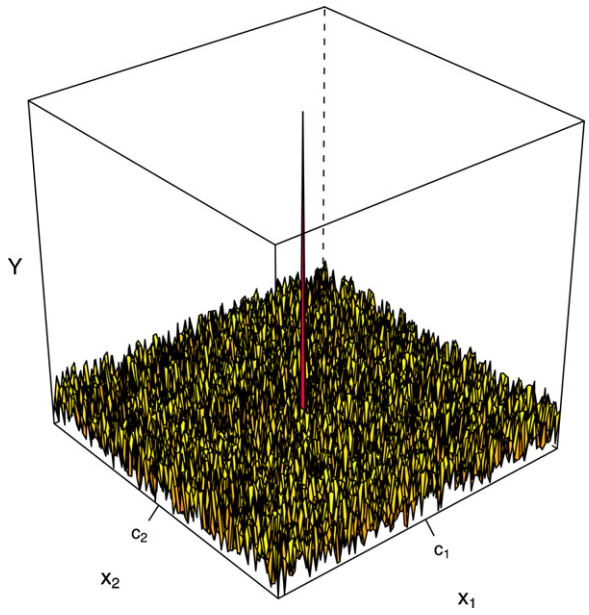

Fig. 1. Representation of a needle function where the only real response is at a single combination (interaction) of two factors, each with a large number of potential levels.

random search. Mathematically, this system can be represented by a function

$$Y = \begin{cases} 1+\varepsilon & \text{if } (\mathbf{x} = \mathbf{c}) \\ \varepsilon & \text{if } (\mathbf{x} \neq \mathbf{c}) \end{cases}$$

where $\mathbf{x}$ is a point in the space, $\mathbf{c}$ is the vector of optimal levels, and ε a normally distributed random variable with zero mean.

For a rational discovery program of any kind to succeed, there must be lower-dimensional "ridges" which can be followed to the optimum (Fig. 2). The equation then can become

$$Y = \frac{1}{n}\sum_{i=1}^{n}\delta(x_i) + \varepsilon, \quad \delta(x_i) = \begin{cases} 1 & \text{if } x_i = c_i \\ 0 & \text{if } x_i \neq c_i \end{cases}$$

where n is the dimensionality of the space, and x_i and c_i are the level of the ith component (factor) in a point $\mathbf{x}$ in the space and in the vector $\mathbf{c}$ of optimal levels, respectively. Here the lower-order ridges

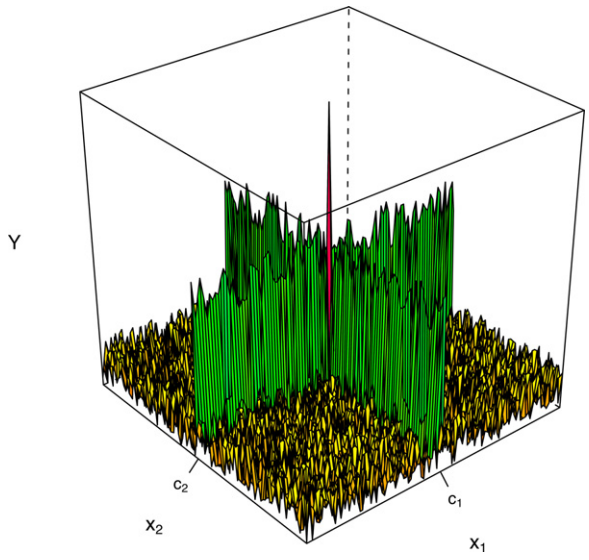

Fig. 2. Representation of a needle function that also has lower-dimensional "ridges".

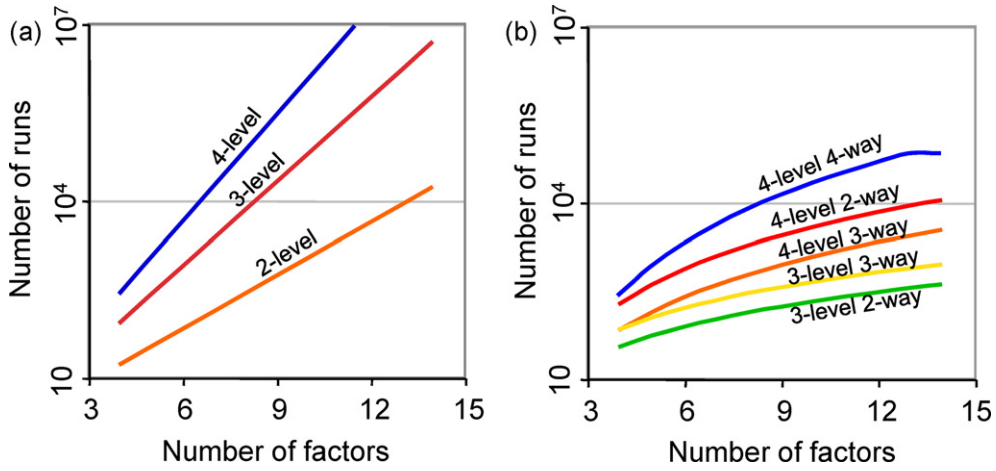

Fig. 3. (a) The number of runs required for a full factorial design of systems with large number of factors and levels per factor. Calculated from runs = (levels)$^{\text{factors}}$. (b) The number of runs required for D- or I-optimal designs for systems with large numbers of factors and levels, depending on the depth of interactions to be searched. Constructed from test systems using the "Custom Design" feature in JMP.

give the clues that identify the factors and levels that participate in the optimum.

We can imagine searching for such optima using full factorial and optimal designs. Full factorial designs suffer greatly from the "curse of dimensionality" (Fig. 3(a)). Optimal designs are far more efficient (Fig. 3(b)), but they require making a reasonably accurate guess as to the level of interactions, and they too will strain the laboratory resources when the number of factors, levels, and interactions gets too high.

Yet another way of considering this problem is to set a reasonable budget of runs for an experiment and examine the size of the spaces that can be examined with different approaches. This is given for a ~1000-run budget in Fig. 4. This is divided into three classes: systems in which every possible qualitative combination can be tested ("brute force"); those that can be handled by relatively conventional DoE; and those that can be handled with computer-generated optimal designs (D- or I-optimal) which search for second-order or higher interactions.

A real problem of the qualitative type is the discovery of synergetic drug combinations. In [10], e.g., Lehár et al. describe a screening experiment aimed at discovering 2-way synergetic effects in a 90-drug library, for inhibiting the proliferation of

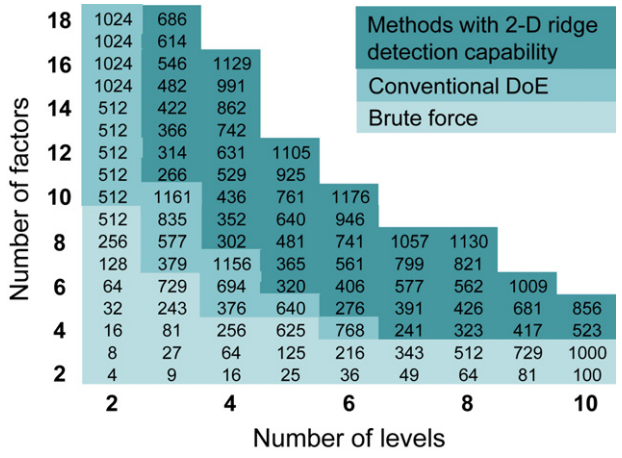

Fig. 4. Estimation of the size of experimental space that can be searched with a 1000-run budget under three scenarios: brute force (full factorial design); conventional DoE (fractional factorial design capable of detecting 2-way interactions); and methods with ridge detection capability (D- or I-optimal designs capable of detecting 2-way interactions).

HTC116 cancer cells. The response of each of the $C(90,2) = 4095$ combinations was an aggregate value extracted from a 6×6 inhibition matrix of combinations of different amounts of the two drugs, thus requiring a total of $4095 \times 36 = 147{,}420$ measurements. Now, if we consider the possibility of extending this search to drug triples, the number of possible qualitative combinations that need to be tested becomes $C(90,3) = 125{,}580$, corresponding to 4,520,880 response measurements. If the experimental effort scales linearly with the number of combinations examined, then screening for drug triples would take roughly 30 times as long as screening for drug pairs. If, e.g., screening pairs took one month of effort, then screening for triples would take two and a half years of effort. This would be a daunting task even for the resources of a major pharmaceutical company. However, there are only 4095 possible 2-way combinations, and if the "ridge" hypothesis is correct, three of those could be ridges to a 3-way peak.[1] The key to the problem will be use of a search algorithm that can detect those ridges.

1.2.2. *Quantitative spaces*

In quantitative systems, the optima will almost always be regions of some width rather than a single point. In these systems, the question will be whether the distance between any two points interrogating the space is on the same scale as the width of the optimum region, which can be modelled as a Gaussian, described by the following equation:

$$Y = \exp\left(\left(\frac{\sum(\mathbf{x}-\mathbf{c})}{\mathbf{s}}\right)^2\right)$$

where **c** and **s** are the mean and the standard deviation of the Gaussian, respectively. These types of peaks are shown in 2D and 3D in Figs. 5 and 6.

We propose a sense of the needed "resolution" of a search design by considering the volume of that part of the optimum region that has a response greater than the overall system noise. An estimate of the necessary resolution is that volume divided by the total space volume, and an estimate of the number of points needed is the inverse of that ratio. This could be done, e.g., by taking a "range-finding" set of points and considering all experimental points that are 3 standard deviations over the average for the set as candidates directing the search toward an optimum.

[1] If the peak is a 3-way A–B–C interaction, then there are three subsidiary 2-way interactions: A–B, A–C, and B–C.

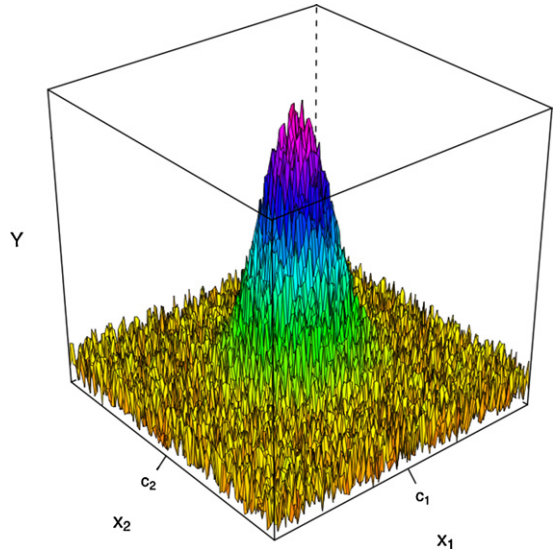

Fig. 5. Stochastically perturbed Gaussian peak in a 2-dimensional factor space.

Alternatively, before the experiments begin, we can consider the ratio of the volume of the anticipated optimum (e.g., calculated within the contour lines connecting the points on the Gaussian that are 2 standard deviations from its center) to the volume of the entire space. This must include consideration of the number of dimensions in which the optimum is defined.

The real problem is that none of these calculations can be made without making important assumptions about the optimum region. In a real discovery project, we have little or no information about these assumptions. It is still worthwhile to use preliminary experimentation to make some rough estimates of the needed resolution.

These ideas can be considered an extension of standard sequential response surface methodology [11], where detection of an "upward" slope leads to the optimum by steepest ascent methods. UOP Inc., a major catalyst manufacturer, has applied these methods extensively in the discovery of new zeolite catalysts, where a new optimum is at a narrowly defined apex of compositional and processing variables. A typical zeolite synthesis study includes factors such as oxide source, template, mineralizer, buffer, mixing order, temperature, temperature ramp, time, and seeding method. A modest estimate of 3–4 levels/factor quickly generates an experimental space exceeding 10^4 runs [12]. After a first space-filling experiment, a nonlinear multivariate regression was used to locate potential optima, followed by sampling in those regions with their optimal coverage algorithm [13].

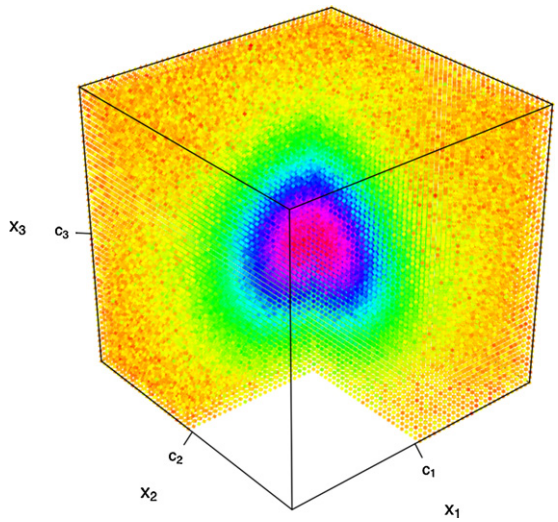

Fig. 6. Stochastically perturbed Gaussian peak in a 3-dimensional factor space. One quadrant has been cut out to make the peak structure visible.

1.3. Experimental space sampling tactics and strategies

When designing complex systems defined on very large, high-dimensional spaces, an evolutionary (or "adaptive"), iterative DoE strategy is usually preferred to a classic DoE tactic. Evolutionary DoE strategies abandon the idea of analyzing the entire space of combinations and all relationships between factors in only one experiment and with only one design. Instead, they iteratively build new designs as a function of the experimental results collected during the exploration. This allows the search to gradually converge on a limited optimal region of experimental space by drawing on clues gathered in the initial stages (or "generations"), sampling for observation only a tiny fraction of all possible points. Choosing the points of the first generation of an iterative program is a tactical issue that then feeds into the general strategy, which determines how points will be chosen in all following generations.

1.3.1. First generation sampling tactics

For sampling quantitative spaces, it is tempting to use a Cartesian design, in which all dimensions are divided equally and points are placed at every intersection. This is very simple to set up, but it has been shown to be far worse than the mathematically optimal method for sampling an unknown experimental region, which is a packing or covering lattice design [14]. This optimality is predicated on the assumption of an infinite space; for practical experimental spaces, lattices have fitting problems at the edges. Cartesian designs will also work in spaces that are composites of quantitative and multilevel qualitative factors. Lattices will not because they typically define points with non-integer coordinates.

For a real space, it is often preferable to use a stochastic method for generating a sampling set. As spaces grow more complex, fully deterministic designs like lattices become increasingly difficult to generate, because of system constraints and lack of good distance measures. In short, spaces become more irregular. This does not bother stochastic methods. There are several flavors of stochastic methods. A purely random design is by far the easiest to generate, and has the great advantage of being applicable to all types of spaces – quantitative, qualitative, and composite. True randomness, however, is actually somewhat "lumpy" (Abelson's first law) [15] – points are surprisingly likely to be clustered, leaving relatively large unsampled gaps. Weighted random sampling techniques, in which the probability of a point being sampled is biased toward its distance from previously tested points, should be preferred [16].

There are several other stochastic sampling methods for quantitative factors in the more sophisticated statistics packages such as JMP [17]. The best of these for chemical experimentation appear to be:

- Sphere packing, which emphasizes the spread of points;
- Latin Hypercube, which is a compromise between spread of points and uniform spacing;
- I-optimal, which minimizes the average variance of prediction over the region of the data.

A comparison of some of these methods is given in Fig. 7, showing the distribution of minimum distances of 250 random points in a 4-dimensional square space divided in 10^2 units in each dimension (10^8 points). The best methods are those with the least representation on the right hand (large gap) side, with Latin Hypercube the

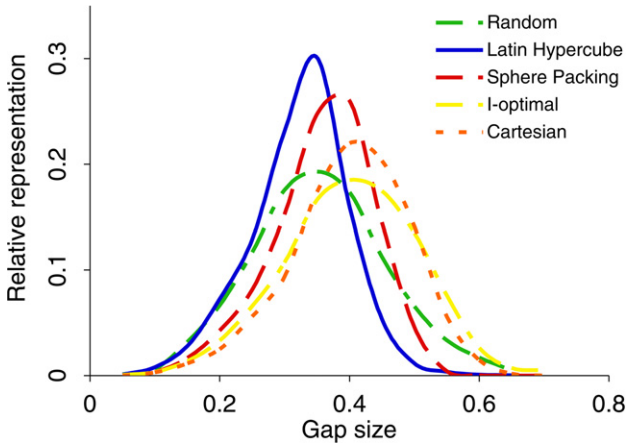

Fig. 7. Estimates of gap size between experimental points for several stochastic strategies.

best in this case, closely followed by Sphere Packing.[2] Johnson et al. [18] have shown that Sphere Packing and Latin Hypercubes are "the best choice of designs" for computer simulation experiments where high-order polynomial designs are likely to be appropriate. In these situations "responses from computer experiments can be complex" and "filling the design region is important because little is known about what portions of the region will provide the most informative and interesting effects". These are the same properties often observed in a complex combinatorial chemistry space, so it is reasonable to suggest that the same designs will be effective.

1.3.2. Sampling strategies for subsequent generations

Genetic algorithms are by far the most popular adaptive strategy employed for the optimization of physical and chemical systems [19–25]. GAs assign each point in a space a genetic code, and then progress from one generation of experiments to the next by applying genetic operators analogous to mutation and crossover to winning solutions. One of the main advantages of GAs in this context is that they do not make any specific assumption about the topology of the response surface, which makes them adaptable to a large variety of problems. As the effectiveness of a GA increases with increasing population size and number of generations, GAs are particularly powerful when coupled with high-throughput technologies, which allow experimenters to quickly create and assess a large number of configurations of the system to optimize.

Some authors, however, underline that traditional GAs tend to generate new candidate solutions without efficiently leveraging beneficial interactions between factors, due to the randomness of their operators [26]. With this premise, over the last years, a growing interest has been directed toward new adaptive optimization strategies that learn the structure of factor interactions from experimental observations, and exploit this information to speed up the optimization process. One of these strategies consists in building statistical models (often in the form of second-order polynomials) that explicitly estimate the relationship between location of points in the space and experimental response from previously collected observations, and use such models to guide or substitute genetic operators in the selection of new points to observe [27–29]. Another area that combines exploitation of structure with random search is the estimation of distribution algorithms [30], where Bayesian methods are the primary tools for detecting structure, with an overlay of various techniques, such as clustering and GAs, to cope with the complexity of multi-factor interactions [31].

[2] These results will probably vary with sample size and dimensionality.

Recent literature on design and optimization of experimental systems includes a number of reports in which GAs have been combined with artificial neural networks (ANNs) trained on experimental observations [32]. In some of these, the ANN is trained after several generations of a GA have identified a promising region. This is followed by virtual experimentation, e.g., for optimally adjusting some of GA parameters to the system at hand, using the ANN as artificial response function [33], or for studying how the efficiency of the evolutionary strategy could be improved using the ANN as a response predictor [34]. In other studies, a GA generates a large number of candidate points; then the ANN, trained on all previous experimental observations, selects the most promising points for the next generation of experiments [35]. Our contribution to this research thread is a new strategy of evolutionary design of experiments (Evo-DoE) that combines prediction of fruitful points from nonlinear regression models of previous experimental observations with stochastic exploration of the experimental space based on weighted random sampling. Thus, the ANN or other nonlinear model is, on the one hand, intimately embedded in the evolutionary process at each generation and, on the other hand, a tool for point selection which does not rely on a GA. Elsewhere, we report on the application of this approach to the problem of optimizing the cargo capacity of a complex liposomal drug formulation [36]. Here, we describe the Evo-DoE procedure in detail and assess its performance on a family of artificial problems. We then benchmark such performance with that of a standard GA.

2. Materials and methods

2.1. Artificial response surface and experimental space

To compare the performances of Evo-DoE and the GA, we ran three batches of simulations, in which the two experimental strategies were applied to the problem of optimizing three different complex artificial response surfaces. In previous work [24], we partially resolved the topology of the response landscape of a real high-dimensional mixture amphiphile system. The experimental data provided evidence of a multi-peaked response surface, perturbed by several components of experimental noise. This information was incorporated here in the artificial response surfaces, which were designed by superimposing several stochastically perturbed Gaussian peaks on a simplex-lattice space [37].

The response $F_{\mathbf{x}}$ of a point $\mathbf{x}$ in the q-dimensional simplex-lattice space was defined as:

$$F_{\mathbf{x}} = \max_k(G(\mathbf{x})), \quad \text{with} \quad G(\mathbf{x}) = a_k \exp\left(\left(\sum_{i=1}^{q} \frac{(x_i - c_{i_k})}{s_k}\right)^2\right)$$

where x_i is the level of the ith component (factor) of $\mathbf{x}$; k is the Gaussian index, varying from 1 to p; c_{i_k}, a_k and s_k are respectively: the coordinate of the mean $\mathbf{c}_k$ in the ith dimension, the height at $\mathbf{c}_k$, and the standard deviation of the kth Gaussian. The formula then calculates the heights of the p Gaussians at $\mathbf{x}$, and the highest of these is associated to $\mathbf{x}$ as response value. A given number $l < q$ of dimensions are supposed to be "neutral" with respect to one or more of the p Gaussians, as if in such dimensions the distance of $\mathbf{x}$ from $\mathbf{c}_k$ were null independently of where $\mathbf{x}$ is located. The response of a combination, with respect to each Gaussian, therefore depends on the interaction of l variables.

The measured response $F^m_{\mathbf{x}}$ of $\mathbf{x}$ is simulated as:

$$F^m_{\mathbf{x}} = F_{\mathbf{x}} + N(0, rF_{\mathbf{x}}) + N(0, t)$$

The "true" response $F_{\mathbf{x}}$ is then perturbed by two error components: one proportional to and one independent from $F_{\mathbf{x}}$. The two quantities are sampled from a normal distribution with zero

Table 2
Coordinates of the mean of each Gaussian contained in the artificial response surfaces used in the three experiments (neutral dimensions not shown). In every surface, one or more relevant factors are shared by more Gaussians (e.g., in experiment 2 factors 1 and 2 are relevant for all three Gaussians). Note that the optimal level of all relevant factors is always a positive value.

Height (a)	Optimal levels
Experiment 1	
1	$c_4 = 10$, $c_5 = 10$
3	$c_1 = 4$, $c_2 = 16$
8	$c_1 = 15$, $c_3 = 5$
Experiment 2	
1	$c_1 = 4$, $c_2 = 16$
3	$c_1 = 6$, $c_2 = 5$, $c_6 = 9$
8	$c_1 = 10$, $c_2 = 3$, $c_3 = 3$, $c_4 = 3$, $c_5 = 1$
Experiment 3	
1	$c_1 = 1$, $c_2 = 1$, $c_3 = 2$, $c_4 = 2$, $c_5 = 2$, $c_6 = 2$, $c_7 = 2$, $c_8 = 3$, $c_9 = 2$, $c_{10} = 3$
3	$c_8 = 5$, $c_9 = 3$, $c_{10} = 2$, $c_{11} = 3$, $c_{12} = 2$, $c_{13} = 1$, $c_{14} = 1$, $c_{15} = 1$, $c_{16} = 1$, $c_{17} = 1$
8	$c_2 = 4$, $c_9 = 1$, $c_{11} = 2$, $c_{18} = 3$, $c_{19} = 3$, $c_{20} = 2$, $c_{21} = 2$, $c_{22} = 1$, $c_{23} = 1$, $c_{24} = 1$

mean and standard deviation determined by parameters r and t, respectively.

The number q of factors in the space was set to 100. In every given point of the simplex, each factor can be present in one of $m = 20$ possible relative amounts (0, 0.05, 0.1, ..., 1), and the relative amounts of all factors must sum to 1. This means that up to 20 factors at a time can be present in positive quantities. Because this is a simplex-lattice space, its cardinality is equal to $C(q-1+m, m) \approx 2.46 \times 10^{22}$.

We considered three different response surfaces defined on this space, each composed by $p = 3$ Gaussians with the same standard deviation ($s_1 = s_2 = s_3 = 5$) but with different heights ($a_1 = 1$, $a_2 = 3$, $a_3 = 8$), and perturbed with parameters $r = t = 0.05$. The three functions differ from each other for the number of non-"neutral" variables, namely for the number of factors on whose interaction the response of a given point depends. In the first response surface, interactions are binaries with respect to each Gaussian; in the second, they are binary for the Gaussian with height 1, ternary for the Gaussian with height 3 and quinary for the Gaussian with height 8; in the third, interactions involve 10 factors for each Gaussian (Table 2).

2.2. Evolutionary sampling strategies

2.2.1. Evo-DoE

The Evo-DoE process used here started with a tactic consisting of a set of 384 randomly selected points.[3] The nth generation (iteration) of the Evo-DoE process consisted of the following steps:

(1) Measure the experimental response of the nth set of points;
(2) Optimize the metaparameters of the nth model, as described below;
(3) Build a model of the entire response surface from the experimentally measured responses for the first n sets of points;
(4) Starting from each of 400 randomly selected points, hill-climb the modeled response surface, keeping track of all the points visited that were not tried in the first n generations;
(5) Randomly choose 336 points from the untried ones found at the previous step with the top decile of responses predicted by the model,[4] and add those points to the $n+1$th set of points;
(6) Add 48 more randomly selected untried points to generate 384 points for the $n+1$th set of points.

The initial random sample, and the random samples chosen at steps 4–6 of every generation, were drawn from a probability distribution biased toward the unsampled regions of the space, in order to favor global exploration of the experimental space. Specifically, the probability of a point being sampled was proportional to the Euclidean distance between that point and the closest already sampled point. The probability distribution was recomputed after sampling every point. A previously sampled point could not be resampled.

The hill-climbing algorithm mentioned at step 4 started with the prediction, based on the model learned at step 3, of the response level of all nearest neighbors of the jth initial point, and the following selection of the neighbor with the highest predicted response. This procedure was then iteratively repeated, selecting the predicted best nearest neighbor of the point selected at the previous iteration, and so on, and it was stopped when either: (a) all nearest neighbors had lower predicted response than the point selected at the previous iteration; (b) the selected point had already been selected in previous hill-climbing runs starting from any of the first $j-1$ initial points.

The balance between the model-based and random points was determined in such a way that most of the experimental effort was aimed at intelligent sampling, and only a small fraction of it at pure exploration. The specific figures, however, should be considered arbitrary.

The models used here were feed-forward, single hidden-layer ANNs [38] (learned with back-propagation using nnet in the R language after standardizing all inputs and normalizing the output to the [0,1] interval), with 100 inputs and 1 output.

Each ANN was constructed with particular metaparameter values (weight decay constant and number of hidden-layer nodes). At step 2 of the Evo-DoE cycle, the model's metaparameters were selected using a bagging process [39], repeating the model learning on 20 different data sets, each being a different random sample of 80% of the observed points, and 10 times on each data set. Each configuration of metaparameters was then assigned a quality measure, calculated as the mean linear correlation between the remaining 20% observations and the corresponding predictions over all the repeats.

2.2.2. Benchmark genetic algorithm

The genetic algorithm used here is a variation of the one described in [24]. It started with a tactic consisting of a set of 384 randomly selected points. The nth generation of the genetic algorithm consisted of the following steps:

(1) Measure the experimental response of the nth set of points;
(2) Choose 168 parent points via tournament selection as described below;
(3) Generate 1 mutant child from every parent, as described below;
(4) Divide parent points in pairs and generate 2 child points from every pair by crossover, as explained below;
(5) Add 48 more randomly selected untried points to the $n+1$th set of points.

[3] The exact number of points is arbitrary and will vary with the experimental circumstances. High-throughput experimentation commonly makes use of 384-well plates. One such plate would then fit one entire generation of the evolutionary strategy.

[4] If less than 336 points fall within the top decile, the top 336 points should be selected.

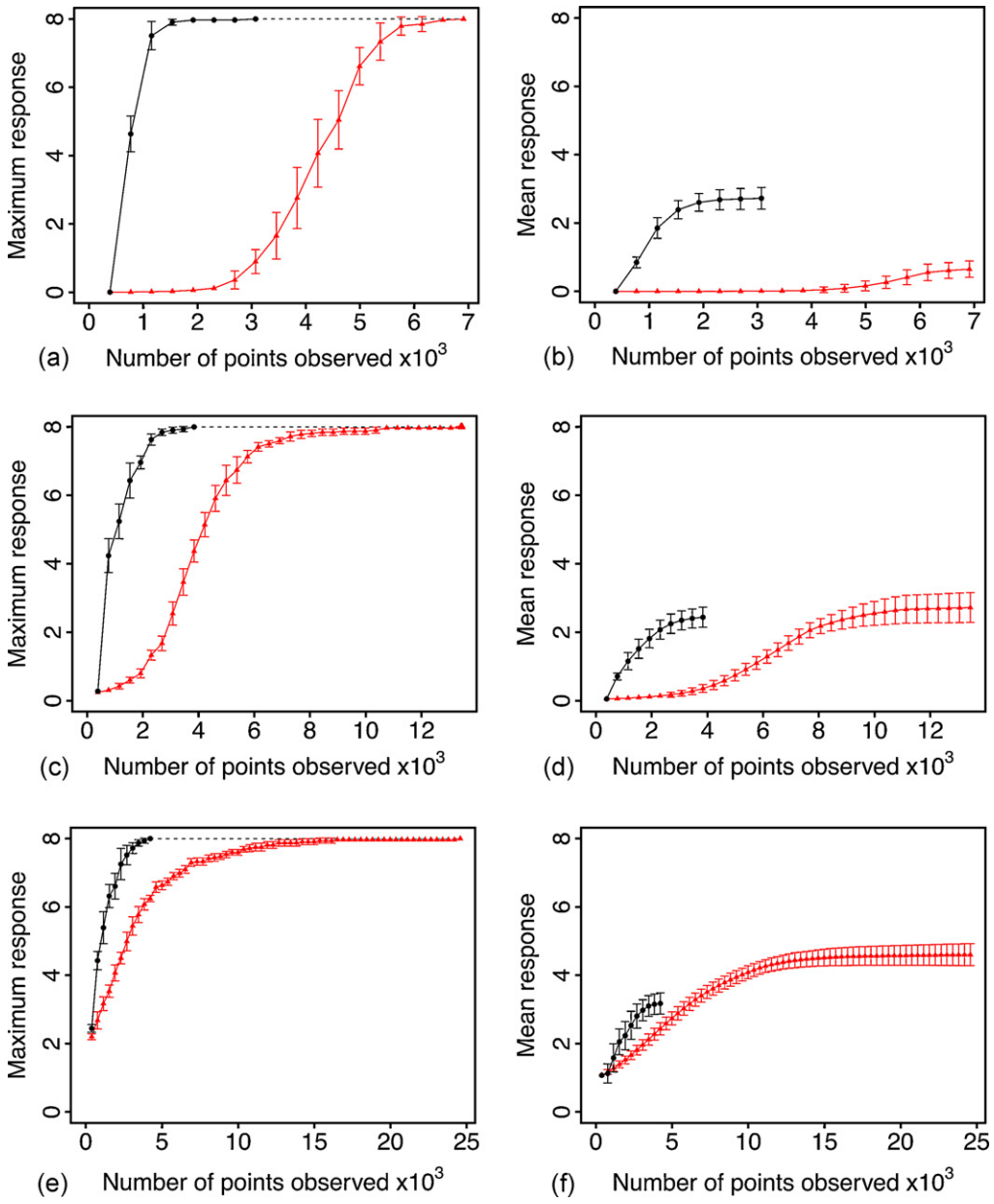

Fig. 8. Expected value of mean and maximum response for Evo-DoE (circles) and for the GA (triangles), in experiments 1 (a, b), 2 (c, d), 3 (e, f). Both statistics were calculated at every generation on all observed points (which also include those observed at the current generation). The time series stop at the generation at which the global optimum was found in all replicates. By definition, the maximum response would have remained the same if the simulations were run further (dotted line), while it is not possible to accurately predict what the dynamics of the mean response would have been. Error bars show 95% confidence intervals, estimated on 10 replicates. Confidence intervals smaller than the size of the symbols are not shown.

The initial random sample, and the random sample chosen at step 5 of every generation, were drawn from a uniform probability distribution.

Parent points were selected with the so-called "tournament" criterion [40], as follows. Let b the number of already observed points, d the number of parent points to select and e an integer such that $1 < e < b - d + 2$. The tournament criterion requires sampling without replacement for d times of e of the b points, and the selection as parents of those with the highest observed response in the d samples. In the experiments we set $d = e = 168$.

Mutant child points were generated following this scheme: one integer number u was randomly sampled from the values 1, 2, 3, 4, 5, with probability 5/15, 4/15, 3/15, 2/15, and 1/15 respectively; u genes (factors) were sampled with replacement among those present in positive quantity; from those factors, the u previously sampled units were removed and added to u other factors, sampled with replacement among the remaining ones. Any mutant point coincident with a previously sampled point was discarded and regenerated.

Parent points were coupled following the order in which they were sampled (the first with the second, the third with the fourth, etc.), and each couple generated 2 crossed child points following this scheme:

(a) One integer number v was sampled from a uniform distribution in the [1,19] interval: v corresponded to the number of volume units, indexed with the factor they referred to, provided by the first parent. The remaining w units (with $w = 20 - v$) were provided by the second parent. The combination of these $v + w$ units corresponded to the first child created with crossover.

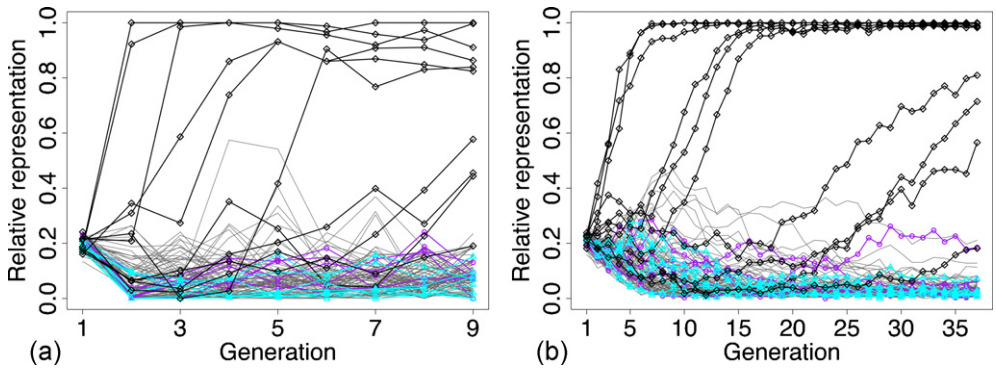

Fig. 9. Representation of every factor in the population of 336 non-random points for a representative replicate of experiment 3 with Evo-DoE (a) and GA (b). The representation of a factor can vary from 0 (in all points the level is equal to 0) to 336 (in all points the level is >0). The time series are shown: with diamonds for the factors relevant to the highest Gaussian; with circles for the factors relevant to the Gaussian with height 3; with triangles for the factors relevant for the Gaussian with height 1; with a grey line for all other (neutral) factors. If a factor is relevant for more than one Gaussian, its time series is shown with the type of line associated to the highest Gaussian. E.g., the time series of factor 2, relevant both for the Gaussian with height 8, and for that with height 1, is shown with a diamond-point line.

(b) The remaining units in the two parents corresponded to the second child.

The parent selection scheme assured that all children were different from their parents and different from each other.

3. Results and discussion

The simulations were stopped at the generation at which the best point, corresponding to the mean of the highest Gaussian, was found. Every experiment was repeated 10 times. To evaluate the behavior of the two evolutionary experimental strategies we used two different performance measures. The first consists in the computation of the maximum response reached on average in the different replicates, after testing a given number of points. We can interpret this measure as the expected value of the maximum response, for a given level of experimental effort. Analogously, the second measure calculates the expected value of the mean response as a function of the number of observed points. Both measures are calculated at every generation on all observed points and not only on those observed at the current generation.

The results of the first set of experiments show that both strategies optimize the response surface very quickly, considering the vastness of the experimental space (Fig. 8). The exploration operated by the Evo-DoE approach, however, is significantly more efficient, for any level of experimental effort. After only one generation the maximum response determined by Evo-DoE reaches a level of roughly 4.5, which implies that the strategy has already found the highest Gaussian and has selected points that are located in the neighborhood of its center (by construction, no combination that does not belong to the counter-image of this Gaussian can have a response higher than 3). To reach the same level of maximum response, the GA requires an experimental effort roughly 5 times larger (note that in the second generation the maximum response is almost unvaried and close to 0). Evo-DoE identifies the best combination within 8 generations (corresponding to 3072 observations), as opposed to 18 required by the GA (6012 observations). The dynamics of the mean response is, as expected, well correlated with that of the maximum response.

The relative performances of the two algorithms are qualitatively confirmed in the two other experiments. However, it is particularly interesting to note that the GA is affected much more significantly by the increase in the complexity level of the problem, compared to Evo-DoE. The optimization of the response surface composed by the 2-, 3- and 5-dimensional Gaussians (experiment 2) by the GA requires 35 generations (corresponding to 13,440 observations), as opposed to 10 required by Evo-DoE (3840 observations). As for the response surface composed by the 10-dimensional Gaussians (experiment 3), the necessary experimental effort increases to 64 generations (corresponding to 24,576 observations) for the GA and only 11 for Evo-DoE (4224).

Fig. 9 allows us to visualize approximately the path followed by the two algorithms in the exploration of the experimental space, up to the selection of the best combination. The figure shows, for every generation and for each of the 100 factors, the time series of the number of combinations that contain such factors at a level greater than zero, for a representative replicate of experiment 3. The 48 random points are excluded from this calculation in order to isolate the behavior of the only "intelligent" components of the strategies (ANN/parent selection, crossover and mutation). The path followed by the GA shows rather gradual changes of direction, with relatively small differences in the population of points between successive generations. The path followed by Evo-DoE, on the other hand, changes direction much more rapidly, suggesting that the topology of the response surface predicted by the ANNs tends to vary substantially from one generation to another (at least in terms of where the best decile of hill-climbed points is located).

Fig. 9(a) shows that the Evo-DoE has identified the six most important of the ten participating factors by generation 6, and the additional four appear by generation 9. The GA (Fig. 9(b)) had identified only three by generation 6, three more by generation 14, and the next four by generation 35. It must also be emphasized that the GA erroneously selects points with positive levels in the non-relevant factors[5] much more frequently than Evo-DoE.

4. Conclusion

Understanding the structure of the experimental space is critical to the planning process of a high-throughput experiment. Only when the experimental space is small enough to search exhaustively, or when interactions among system components are weak enough that the space can be simplified, are traditional DoE methods an appropriate optimization tool. In the presence of complex, synergistic systems defined on large, high-dimensional experimental spaces, more sophisticated optimization techniques are required. The project must be viewed with a holistic mindset. There should be a lively discussion of the interrelationships among

[5] Note that, since this is a mixture system, allocating positive levels to non-relevant factors constrains exploration of the relevant ones.

- The factor space (defined as all possible combinations of the controllable factors, with constraints included);
- The response surface (the anticipated degrees of interaction or irregularity);
- The understanding of the underlying chemistry and physics and;
- The high-throughput experimental system (its capacity and limitations).

Only when all these have been considered (and reconsidered as the experiment progresses) should the tactical elements be addressed. The tactics also should be considered as part of an iterative strategy. This is an update of classical DoE thinking, as in Box's famous dictum to limit the first experiment to no more than 25% of the planned effort [41].

Tactical designs such as sphere packing will serve as tools, but evolutionary strategies are required to find optima in truly complex systems. Stochastic methods such as GAs are effective, but we have demonstrated in our simulations that incorporation of nonlinear modeling and predictive power (Evo-DoE) can enhance the efficiency with which the global system optimum is successfully detected and climbed. In a live experimental context, one may never be sure of reaching a global maximum of the response without exhaustive sampling of the experimental space, but Evo-DoE is the most effective available strategy for finding a good result with a given experimental budget.

Acknowledgements

Thanks to the ProtoLife team, especially Mark Bedau, Martin Hanczyc, and Andrew Buchanan, for valuable discussion, and Emily Parke for thorough and sensitive editing.

References

[1] J.N. Cawse (Ed.), Experimental Design for Combinatorial and High Throughput Materials Development, John Wiley & Sons, Hoboken, NJ, 2003, pp. 1–38.
[2] J.N. Cawse, Experimental designs in high throughput systems, in: B. Narasimhan, S.K. Mallapragada, M.D. Porter (Eds.), Combinatorial Materials Science, John Wiley & Sons, Hoboken, NJ, 2007, pp. 21–50.
[3] T. Sun, Fractional masking methods in combinatorial synthesis of functional materials, in: J.N. Cawse (Ed.), Experimental Design for Combinatorial and High Throughput Materials Development, John Wiley & Sons, Hoboken, NJ, 2003, pp. 39–54.
[4] T.E. Mallouk, E.S. Smotkin, Combinatorial catalyst development methods, in: W. Vielstch, A. Lamm, H. Gasteiger (Eds.), Handbook of Fuel Cells – Fundamentals, Technology and Applications, Wiley, Chichester, UK, 2002, pp. 334–347.
[5] J.N. Cawse, R. Wroczynski, Combinatorial materials development using gradient arrays: designs for efficient use of experimental resources, in: J.N. Cawse (Ed.), Experimental Design for Combinatorial and High Throughput Materials Development, John Wiley & Sons, Hoboken, NJ, 2003, pp. 109–127.
[6] J.N. Cawse, M. Gardner, Split-plot designs, in experimental design for combinatorial and high throughput materials development, in: J.N. Cawse (Ed.), Experimental Design for Combinatorial and High Throughput Materials Development, John Wiley & Sons, Hoboken, NJ, 2003, pp. 129–146.
[7] M. Fasolka, E. Amis, Combinatorial materials science: measures of success, in: B. Narasimhan, S.K. Mallapragada, M.D. Porter (Eds.), Combinatorial Materials Science, John Wiley & Sons, Hoboken, NJ, 2007, pp. 1–20.
[8] D. Whisenhunt, G.L. Soloveichik, New catalysts for the carbonylation of phenol: discovery using high-throughput screening and leads scale-up, in: R.A. Potyrailo, W. Maier (Eds.), Combinatorial and High-Throughput Discovery and Optimization of Catalysts and Materials, CRC, Boca Raton, FL, 2007, p. 132.
[9] J.L. Spivack, J.N. Cawse, D.W. Whisenhunt Jr., B.F. Johnson, K.V. Shalyaev, J. Male, E.J. Pressman, J.Y. Ofori, G.L. Soloveichik, B.P. Patel, T.L. Chuck, D.J. Smith, T.M. Jordan, M.R. Brennan, R.J. Kilmer, E.D. Williams, Combinatorial discovery of metal co-catalysts for the carbonylation of phenol, Appl. Catal. A 254 (1) (2003) 5–25.
[10] J. Lehár, G.R. Zimmermann, A.S. Krueger, R.A. Molnar, J.T. Ledell, A.M. Heilbut, G.F. Short III, L.C. Giusti, G.P. Nolan, O.A. Magid, M.S. Lee, A.A. Borisy, B.R. Stockwell, C.T. Keith, Chemical combination effects predict connectivity in biological systems, Mol. Syst. Biol. 3 (80) (2007) 1–14.
[11] D.C. Montgomery, Design and Analysis of Experiments, sixth ed., John Wiley & Sons, Hoboken, NJ, 2005, p. 405 ff.
[12] W. Schmidt, Synthesis of Zeolites (accessed 01.05.10), http://www.mpi-muelheim.mpg.de/kofo/institut/arbeitsbereiche/schmidt/zeolites_ed.html, 2010.
[13] D.S. Bem, R.D. Gillespie, E.J. Erlandson, L.A. Harmon, S.G. Schlosser, A.J. Vayda, Combinatorial experimental design using the optimal-coverage algorithm, in: J.N. Cawse (Ed.), Experimental Design for Combinatorial and High Throughput Materials Development, John Wiley & Sons, Hoboken, NJ, 2003, pp. 103–105.
[14] F.A. Hamprecht, E. Agrell, Exploring a space of materials: spatial sampling design and subset selection, in: J.N. Cawse (Ed.), Experimental Design for Combinatorial and High Throughput Materials Development, John Wiley & Sons, Hoboken, NJ, 2003, pp. 277–307.
[15] R.P. Abelson, Statistics as Principled Argument, Psychology Press, London, UK, 1995, p. 18 ff.
[16] S. Kimura, K. Matsumura, Proceedings of the 2005 Conference on Genetic and Evolutionary Computation, Washington, DC, USA, 2005, pp. 1341–1346.
[17] JMP software, www.jmp.com (accessed 01.05.10).
[18] R. Johnson, D.C. Montgomery, B. Jones, P.A. Parker, Comparing computer experiments for fitting high-order polynomial metamodels, J. Qual. Technol. 42 (1) (2010) 86–102.
[19] J. Singh, A.A. Ator, E.P. Jaeger, M.P. Allen, D.A. Whipple, J.E. Soloweij, S. Chowdhary, A.M. Treasurywala, Application of genetic algorithms to combinatorial synthesis: a computational approach to lead identification and lead optimization, J. Am. Chem. Soc. 118 (7) (1996) 1669–1676.
[20] V. Venkatasubramanian, K. Chan, J.M. Caruthers, Computer-aided molecular design using genetic algorithms, Comput. Chem. Eng. 18 (9) (1994) 833–844.
[21] R.P. Sheridan, S.K. Kearsley, Using a genetic algorithm to suggest combinatorial libraries, J. Chem. Inf. Comput. Sci. 35 (1995) 310–320.
[22] D. Wolf, M. Baerns, Evolutionary strategy for the design and evaluation of high throughput experiments, in: J.N. Cawse (Ed.), Experimental Design for Combinatorial and High Throughput Materials Development, John Wiley & Sons, Hoboken, NJ, 2003, pp. 147–162.
[23] L. Weber, S. Wallbaum, C. Broger, K. Gubernator, Optimisation of the biological activity of combinatorial compound libraries by a genetic algorithm, Angew. Chem. Int. Ed. 34 (20) (1995) 2280–2282 (in English).
[24] M. Theis, G. Gazzola, M. Forlin, I. Poli, M.M. Hanczyc, N.H. Packard, M.A. Bedau, Optimal formulation of complex chemical systems with a genetic algorithm, in: Online Proceedings of the European Conference on Complex Systems ECCS '06, Oxford, UK, 2010, Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.125.1921&rep=rep1&type=pdf (accessed 07.25.10).
[25] M. Forlin, I. Poli, D. De March, N. Packard, G. Gazzola, R. Serra, Evolutionary experiments for self-asembling amphiphilic systems, Chemom. Intell. Lab. Syst. 90 (2) (2008) 153–160.
[26] M. Pelikan, D.E. Goldberg, F.G. Lobo, A survey of optimization by building and using probabilistic models, Comput. Optim. Appl. 21 (2002) 5–20.
[27] M. Pelikan, D.E. Goldberg, E. Cantu-Paz, BOA: the Bayesian optimization algorithm, in: W. Banzhaf, J. Daida, A.E. Eiben, M.H. Garzon, V. Honavar, M. Jakiela, R.E. Smith (Eds.), Proceedings of the Genetic and Evolutionary Computation Conference GECCO '99, Morgan Kaufmann, San Francisco, CA, 1999, pp. 525–532.
[28] G. Harik, Linkage learning via probabilistic modeling in the ECGA Technical Report 99010, University of Illinois at Urbana-Champaign, Illinois Genetic Algorithms Laboratory, Urbana, IL, 1999.
[29] L. Bull, On model-based evolutionary computing, Soft Comput. 3 (1999) 183–190.
[30] P. Larrañaga, J.A. Lozano, Estimation of Distribution Algorithms, Kluwer Academic Publishers, Dordrecht, Germany, 2001.
[31] M. Pelikan, D. Goldberg, Genetic algorithms, clustering and the breaking of symmetry, in: M. Schowenauer, G. Rudolph, X. Yao, E. Lutton, J.J. Merelo, H.-P. Schwefel (Eds.), Parallel Problem Solving from Nature – PPSN VI. Lecture Notes in Computer Science, vol. 1917, 2000, pp. 385–394.
[32] Farrusseng, High-throughput heterogeneous catalysis, Surf. Sci. Rep. 63 (2008) 487–513.
[33] U. Rodemerck, M. Baerns, M. Holena, D. Wolf, Application of a genetic algorithm and a neural network for the discovery and optimization of new solid catalytic materials, Appl. Surf. Sci. 223 (2004) 168–174.
[34] L. Baumes, D. Farrusseng, M. Lengliz, C. Mirodatos, Using artificial neural networks to boost high-throughput discover in heterogeneous catalysis, QSAR Comb. Sci. 23 (2004) 767–778.
[35] A. Corma, J.M. Serra, P. Serna, S. Valero, E. Argente, V. Botti, Optimisation of olefin epoxidation catalysts with the application of high-throughput and genetic algorithms assisted by artificial neural networks (softcomputing techniques), J. Catal. 229 (2) (2005) 513–524.
[36] F. Caschera, G. Gazzola, M.A. Bedau, C. Bosch Moreno, A. Buchanan, J. Cawse, N. Packard, M.M. Hanczyc, Automated discovery of novel drug formulations using predictive iterated high throughput experimentation, PLoS ONE 5 (1) (2010) e8546, doi:10.1371/journal.pone.0008546.
[37] J.A Cornell, Experiments with Mixtures: Designs, Models, and the Analysis of Mixture Data, third ed., John Wiley & Sons, New York, NY, 2002.
[38] S. Haykin, Neural Networks: A Comprehensive Foundation, second ed., Prentice Hall, NJ, 1998.
[39] L. Breiman, Bagging predictors, Mach. Learn. 24 (2) (1996) 123–140.
[40] B.L. Miller, D.E. Goldberg, Genetic algorithms, tournament selection, and the effects of noise, Compl. Sys. 9 (3) (1995) 193–212.
[41] G.E.P. Box, J.S. Hunter, W.G. Hunter, Statistics for Experimenters, second ed., Wiley–Interscience, Hoboken, NJ, 2005.