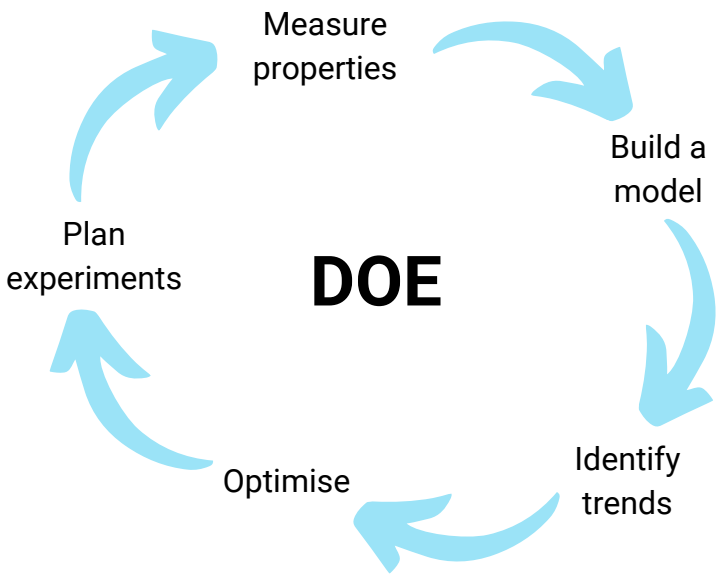
intellegens

WHITE PAPER

Machine Learning for Guided Design of Experiments

Executive Summary

Design of experiments can be used to guide experimentation and find the best combination of parameters in the fewest number of steps. The advent of machine learning approaches has enabled innovative companies to augment their design of experiments with a more guided approach to not only find the 'answer' the quickest but also identify experiments to best improve the underlying model, leading to a continual cycle of improved operational performance. Here we highlight the traditional approaches to experimental design and how Intellegens is using machine learning to disrupt this methodology, resulting in significant savings in time and money in the product development lifecycle.



Traditional approaches to experimental design

Traditional R&D is limited by the human inability to interpret high-dimensional data and make unbiased decisions. Experiments and computational modelling can consume vast quantities of time and resources. The development of new methodologies that accelerate the discovery and design of new formulations is therefore crucial for achieving time efficiency and cost reductions.

The design of new formulations is being largely carried out through traditional experimentation and intuition. The limitations attached to such processes include cost, manual effort and long times.

There are many different approaches to experimental design. Here we will highlight the most simple COST (Change One Separate variable at a Time) based approach and combinatorial DOE (Design Of Experiments) approach, a statistical method that allows the modification of different variables simultaneously. We will demonstrate how traditional approaches to experimental design can be significantly improved with machine learning.

Exploring target property with two variables

For example, in a simple chemical reaction when the goal is to find optimal parameters to maximise yield, we can alter variables such as:

1. The volume of the reaction vessel (between 500 and 700 ml) and,
2. The pH level of the solution (between 2.5 and 5)

In the COST framework we **alter a variable and monitor the response outcome**, which in this case is the yield. So, for instance, we might initially fix the pH level at 3 and vary the volume of the reaction medium from 500 ml to 700 ml. By varying one of the variables at a time, we will see what effect it has on the yield. Figure 1 shows a plot of the measured yield against several reaction volumes in the range of interest. It is clear that the volume that results in maximum yield is 550 ml.

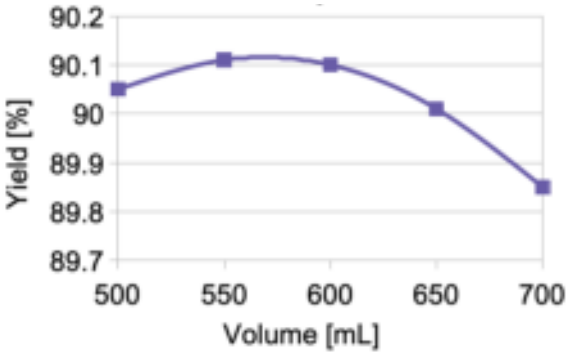


Figure 1. *Table and plot of the measured reaction yield for different reaction volumes*

The next stage is to evaluate the case when the reaction volume is fixed at 550 ml (the optimal level) and change the second variable (pH). In the second set of experiments the pH is raised in 0.5 increments from 2.5 to 5.0 units and the yield is measured. Figure 2 shows that the optimal pH level is around 4.5.

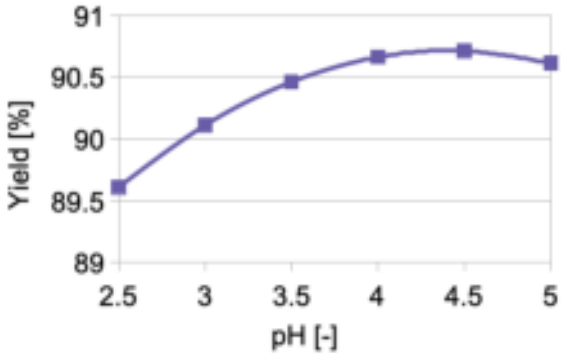


Figure 2. *Table and plot of measured reaction yield for different pH levels*

From this example we might conclude that the best combination of parameters to achieve maximum yield would be a volume of 550 ml and a pH of 4.5.

However, the actual relationship between pH and volume is presented by the Contour Plot in Figure 3. The optimal combination would correspond to a set of pH and volume parameters located towards the top-right, in the large red area, which was not identified. This clearly highlights the problem with designing experiments using the COST-based approach: we can reach a different result depending on the parameter ranges of choice, and this may not be the optimal combination.

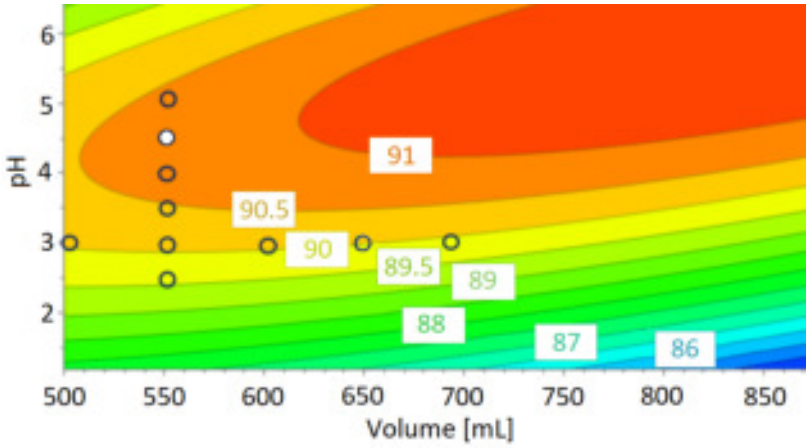
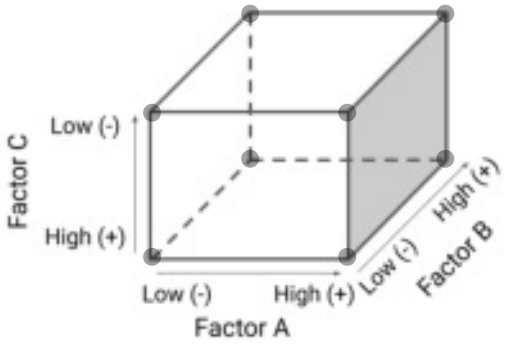


Figure 3. *Contour plot for the reaction yield (indicated in the white squares at the contour lines) at different pH-volume coordinates. The open circles are the pH and volume values investigated in Figures 1 and 2, and the solid white circle is the pH-volume combination identified as resulting in maximum yield.*

This method can be improved with a combinatorial Design of Experiments (DOE) approach, which allows to analyse more than two variables at once and makes it simpler to assess a large number of variables, albeit in a large number of experiments.

Predicting a target property in a high-dimensional space

A typical combinatorial design of experiment (DOE) approach is to **sample a fixed number of points for each input variable**, and then to evaluate all possible combinations of these variables. The advantage of adopting the combinatorial DOE approach is that multiple input variables (e.g. volume and pH) can be varied simultaneously to show the conditions at which the outputs (e.g. yield) reach an optimum value. When the experimental evaluation process begins, we will obtain very valuable information about the direction in which to continue to vary the parameters to improve the yield. When looking at multiple dimensions, we want to understand the values of the response for all possible combinations of variable levels. As the number of variables increases, the total number of experiments increases exponentially (see Figure 4 for a 3-variable example).



Run	Factor A	B	C
1	-	-	-
2	+	-	-
3	-	+	-
4	+	+	-
5	-	-	+
6	+	-	+
7	-	+	+
8	+	+	+

Figure 4. *Visual (left) and tabular (right) representation of a combinatorial DOE approach. The grey circles on the cube denote the combination of parameters represented in the table.*

Even when the number of variables is small, many runs are needed if a full factorial design is to be used. For instance, for five variables, $2_5 = 32$ experiments are needed in the base run only. If more samples are required per variable, for instance to capture nonlinear responses, the number of experiments required rapidly increases, to the point where the combinatorial DOE approach becomes prohibitively expensive. For example, if five samples are required per variable to obtain accurate sampling over 3000 experiments will be required, even if there are only five variables to consider.

The main drawback of the combinatorial DOE approach is the number of runs of the experiment you need to do to obtain a reliable covering of the search space.

Because multiple data points are needed for each combination of variables, the **testing requirement can be quite large**. A combinatorial DOE evaluation of a process can take considerable investment. Other drawbacks of the combinatorial DOE include an inability to control certain variables, problems identifying all the variables that affect the process, and difficulty in identifying the optimal settings with non-linear or correlated variables.

These problems tend to limit the frequency with which combinatorial DOE is used to solve problems. To effectively search for new materials and their properties concurrently, machine learning must be adopted to achieve accuracy and scale to encompass the large chemical space of materials discovery.

Machine learning: A better alternative for experimental design

Machine learning is a powerful tool for finding patterns in high-dimensional data. It uses algorithms to learn from experimental data by modelling the relationships between material properties and related variables. Machine learning automates analytical model building using algorithms that iteratively learn from data, thereby allowing computers to discover hidden patterns and insights without being programmed or biased in where to search.

Several machine learning methods such as deep learning have attracted significant interest across multiple industries as they show good applicability to tasks involving high-dimensional data. This approach is particularly useful when **data is unstructured, incomplete, and highly correlated**. The major advantage of machine learning compared to classical experimental design is that it reduces much of the randomness concerning the choice of which variables to stratify on.

> By adopting machine learning approaches, the researcher can simply take all available data into account without repeatedly thinking about which experiments are most valuable.

Machine learning makes **minimal assumptions** about the systems that are generating the data. They can be effective even when the data are gathered without a controlled experimental design and in the presence of complicated nonlinear interactions.

Feature search optimization

Grid search and **manual search** are the most widely used strategies for feature search optimization. However, when dealing with a high-dimensional space (over 5 dimensions), a **random search** algorithm will return insights more efficiently than a grid search by covering the search space a lot more efficiently (Figure 5).

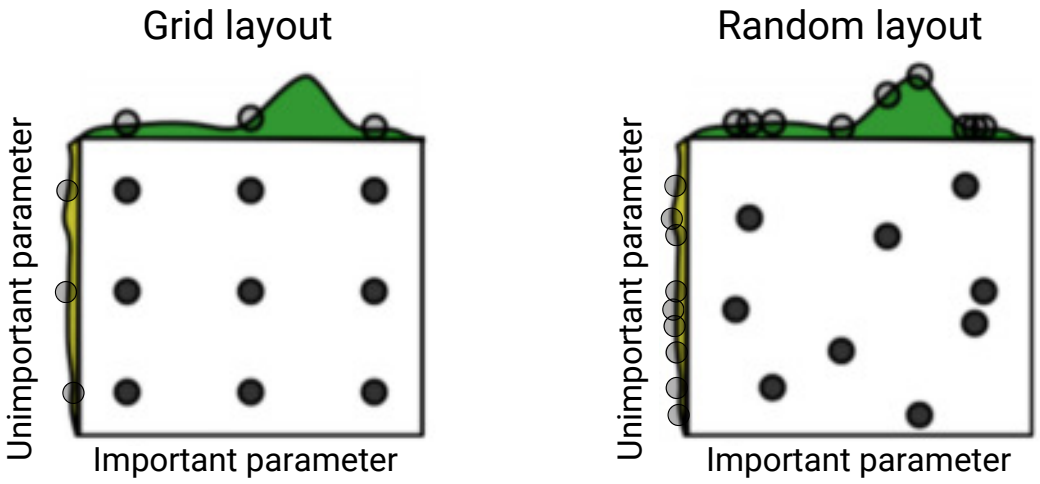


Figure 5. *Schematic of the grid and random feature search layouts. The "important parameter" (x-axis) is the one whose value affects the desired outcome the most (green curve). The green and yellow curves denote the importance of the value with regards to the target property. The black circles show the parameter combinations that have been characterised and the thick and thin grey circles denote the values of the important and unimportant parameter that were investigated, respectively.*

One of the advantages of random search is that if two variables are a little correlated, this approach enables the optima of each variable to be found more precisely. Random search is also valuable where some variables are more important than others: **random search does not waste effort on examining unimportant variables**.

> By adopting machine learning approaches, the researcher can simply take all available data into account without repeatedly thinking about which experiments are most valuable.

Alchemite™ outperforms random search

Predicting components is a way to map the landscape and discover new formulations. Machine learning can be used to guide which compositions are most likely to form effective compounds. The current bottlenecks of component prediction are that the search space for components is limited and such searches require several verification calculations and experiments, which can delay the discovery of new materials.

In the example shown in Figure 6, our machine learning approach, Alchemite™, gives better predictions when compared to random search as more information becomes available.

> The machine learning model understands the formulation landscape better as the number of experiments increases.

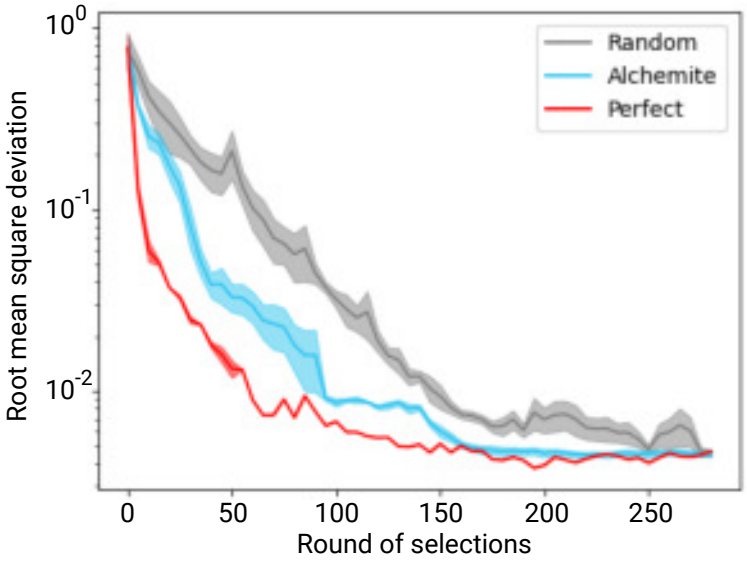


Figure 6. *Accuracy of the model (root mean square error) versus number of experiments performed. The grey curve shows randomly chosen experiments, the blue curve when experiments were chosen by the Alchemite™ algorithm, and the red curve shows the outcome from a perfect selection chosen with hindsight. The shaded areas correspond to the uncertainty in the predictions.*

The Alchemite™ method, plotted in blue in Figure 6, more accurately models the data for every training set chosen via this method. To analyze the curve further we ask the question of how many experiments are required to attain a certain level of accuracy, and plot that in Figure 7.

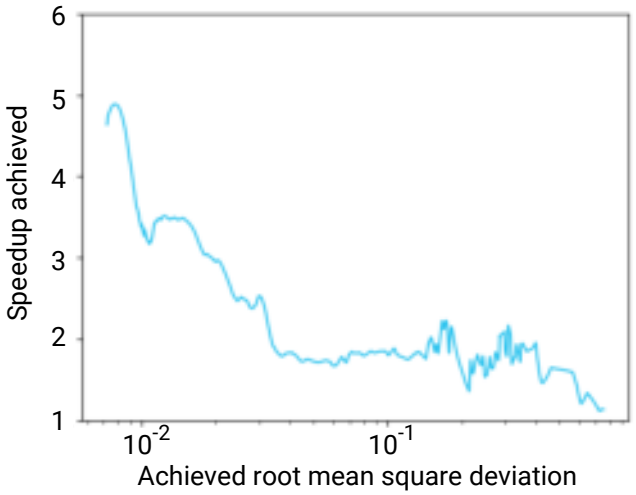


Figure 7. *Speed-up ratio (number of experiments performed when selected by machine learning versus randomly) to achieve a target root mean square deviation.*

Starting from 10 randomly chosen points, both methods achieve the same accuracy and so there is no speed-up in Figure 7, where accuracy is measured by the root mean square deviation. However, with more available data points, or in other words, as the machine learning model understands the material landscape better, the speed-up available becomes more pronounced. The difficulty in reaching the highest level of accuracy (root mean square deviation of <0.01) means that we can achieve it with between 4 and 5 times fewer experiments if we select them carefully.

The Alchemite™ approach not only offers higher accuracy at greater speed, but is also capable of dealing with sparse and noisy data. By predicting and mapping the formulations landscape with attached confidence levels, the approach allows scientists to effectively choose the next best experiment to run in the optimal direction.

Guide your experimental design using the Alchemite™ Analytics platform

With the Alchemite™ Analytics platform, you transform R&D with machine learning by easily experimenting, modelling and visualising sparse and noisy real-world data. Choose the next best experiment to run next by quickly assessing the accuracy and confidence levels of your results.

Why is **Alchemite™** a better approach?

1. **Suggests the most important experiments needed (significantly fewer than the DOE)**
2. **Improves understanding of specific properties**
3. **Accurately maps the landscape of formulation space**
4. **Provides a model for the direction to follow**
5. **Many variables with many levels can be used**
6. **Data can be sparse, noisy and unstructured**

About Intellegens

Intellegens has developed a unique artificial intelligence engine, Alchemite™ for training neural networks from sparse, and noisy data, typical of real-world data. The technique was first developed at the University of Cambridge where it has been used to develop several superalloys, guide the design of new drugs and help optimise battery pack design. The tool is now being used to solve a wide range of industrial customer problems leading to accelerated development, reduced environmental impact, and lower cost.

Want to learn more about how our AI technology can be applied to your specific needs? Contact us to learn more at info@intellegens.ai

Case study data from:

Box, G. E. P., Hunter, W. G., & Hunter, J. S. (1978). Statistics for experimenters: an introduction to design, data analysis, and model building. New York: Wiley.

Click on the link below to subscribe to our latest news and upcoming events

https://intellegens.ai/subscribe

intellegens

intellegens.ai
info@intellegens.ai
@intellegensai