WHITE PAPER

Machine learning for adaptive experimental design

Reduce experimental time and costs by 50-80%

intellegens.com | info@intellegens.com
Intellegens, The Studio, Chesterton Mill, Cambridge, CB4 3NP, UK

intellegens

Executive Summary

What are the optimal composition, chemistry, or processing parameters to achieve commercial performance goals? Answering such questions as quickly as possible is key to the design of **formulations, chemicals, materials**, and **biopharmaceuticals**. Traditional approaches suffer from key disadvantages. Expert-driven design is labor-intensive and time-consuming. Single-factor analysis misses the effects of correlation between factors. Conventional **Design of Experiments** is exhaustive but focused on covering the design space rather than rapidly achieving performance goals.

Machine learning identifies improved products and processes much faster than traditional methods by enabling an **adaptive** approach that focuses experimental effort on those routes most likely to be successful. Since experimental costs associated with a typical industrial R&D project run to hundreds of thousands of dollars, the resulting **50-80% reduction** in experimental workload delivers significant return on investment.

Approaches to experimental design

The goal of R&D is to design new products and processes to meet commercial needs as quickly and efficiently as possible. Experimental campaigns to optimize formulations, chemistry, materials, or manufacturing processes can consume vast amounts of time and resources. New methodologies to accelerate this innovation are crucial for time-to-market and cost reductions.

Experimental campaigns to optimize formulations, chemistry, or materials consume vast amounts of time and resources

Historically, such R&D has been driven by the **knowledge of domain experts**, who leverage years of experience to decide on the next experiments. Relying on experts to direct experimental campaigns has limitations for commercial development: the bottleneck of single experts' availability, variability across an organization as different experts make inconsistent decisions, and the risk of expertise being lost when valuable members of staff move on.

The road to more reproducible, methodical experimental design begins with systematically optimizing single factors in the so-called **COST (Change One Separate variable at a Time) framework**. This approach requires the identification of the most important factor for a given system. This factor is then optimized and all other factors held constant. The procedure is repeated for the next most important factor, with all other factors held constant, and so on.

Although COST is more systematic than expert-driven design, it is unsuitable for complex systems where there are interactions between the factors or non-linear responses.

Conventional **Design of Experiments (DOE)** overcomes these shortcomings by efficiently covering design space to expose how each factor influences the responses. The conceptually simplest method is 'full factorial' DOE, where *N* levels for each of *F* factors are considered, and all possible combinations are measured. For example, one factor may be temperature, measured at 4 different levels (0°C, 25°C, 50°C, 75°C). The approach is illustrated by the grid of blue points in Figure 1a for F=2 and *N*=4. This provides exhaustive coverage of the experimental possibilities and often enables powerful insights into the relationships between factors, their combinations, and properties of interest. However, a full factorial design requires N^F experiments, which becomes prohibitively expensive as *N* and *F* increase.

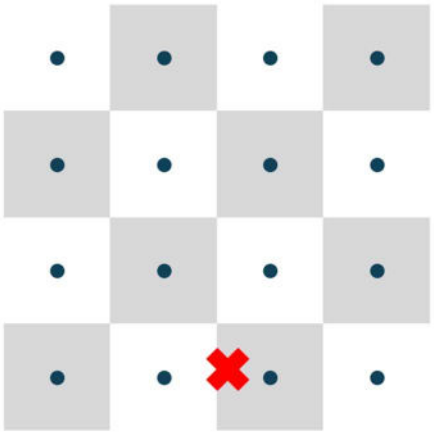


Figure 1a. Full factorial DOE in this case requires 16 experiments (blue points) to cover the space in search of the optimum (red cross).

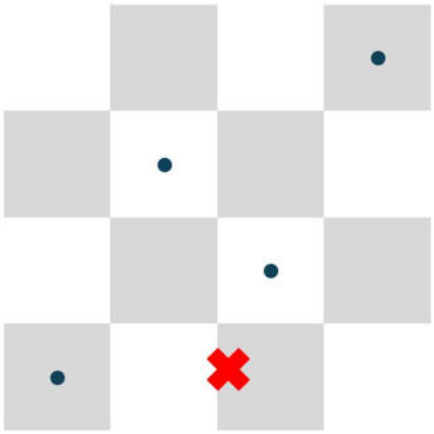


Figure 1b. Latin hypercube sampling of two factors requires fewer experiments but may sacrifice accuracy in locating the optimum.

More advanced traditional DOE techniques cover the design space with fewer experiments. One popular approach is Latin Hypercube sampling, where instead of using every combination of factors, one measurement is proposed per level for each factor, ensuring that this is achieved simultaneously for all factors. This may, however, sacrifice accuracy (Figure 1b). More advanced **statistical DOE methods** include Box-Behnken and Plackett-Burman designs, central composite designs, Taguchi arrays, and definitive screening designs. Implementing these, however, requires statistical knowledge and insight into the likely relationships between the input variables and the response.

Traditional DOE works well for understanding linear effects in a system and some of the more advanced statistical DOE methods are good for quadratic (second order) effects. However, these conventional approaches to DOE are not, in general, practical for **higher-order, non-linear effects**. Yet most real experimental problems are characterized by such interactions.

An alternative is to apply **machine learning** to experimental design, proceeding iteratively from a starting point to identify the experiments that provide the quickest route to the optimum, achieving project goals with substantially less testing (Figure 2). Exploring more complicated

design spaces with higher order effects (Figure 3) is a strength of machine learning, which can understand and dynamically leverage relationships between inputs, in contrast to more static DOE approaches.

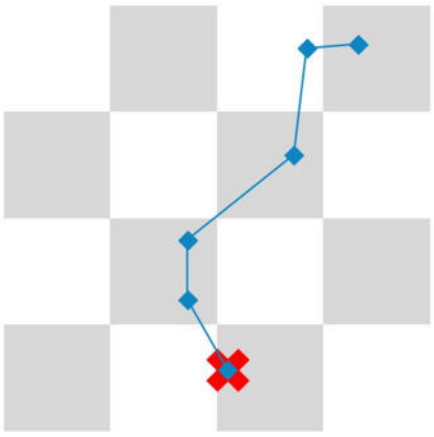


Figure 2. Machine learning enables an iterative approach to find the fastest route to the optimum.

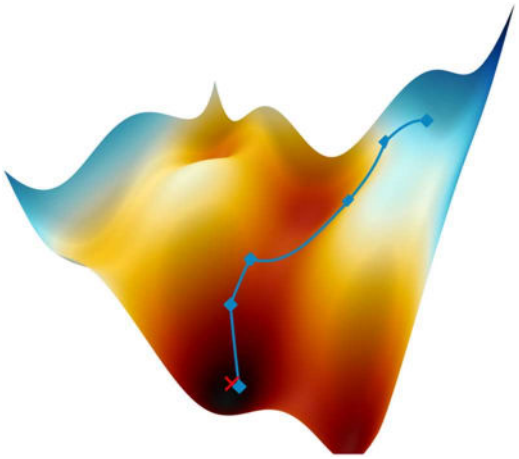


Figure 3. Machine learning can also navigate more complex design spaces.

The key point is that all conventional DOE approaches fundamentally attempt to answer the same question: how to sample the design space most efficiently to understand the way that each factor influences each response. But this is not the most commercially relevant question for experimental design: instead of covering all available options, the true aim of a design project is to find the most effective product in as few experiments as possible.

Adaptive experimental design

Machine learning (ML) enables more sophisticated experimental design strategies that move from simply covering the design space to directly targeting optimal formulations, chemicals, materials, or processes to achieve a project's goals.

> Machine learning can determine which experiments are most likely to succeed

An ML approach can learn from some initial data, with no need to apply advanced statistical knowledge to select the right methods. A 'training' process constructs an ML model – a mathematical representation of the system being studied. The model is then used to select which experiments to carry out next by determining which measurement will be most likely to succeed against the project's goals, while also helping to improve the ML model itself. This results in an **adaptive** approach that creates a virtuous cycle (Figure 4) of a rapidly improving model suggesting increasingly performant new products.

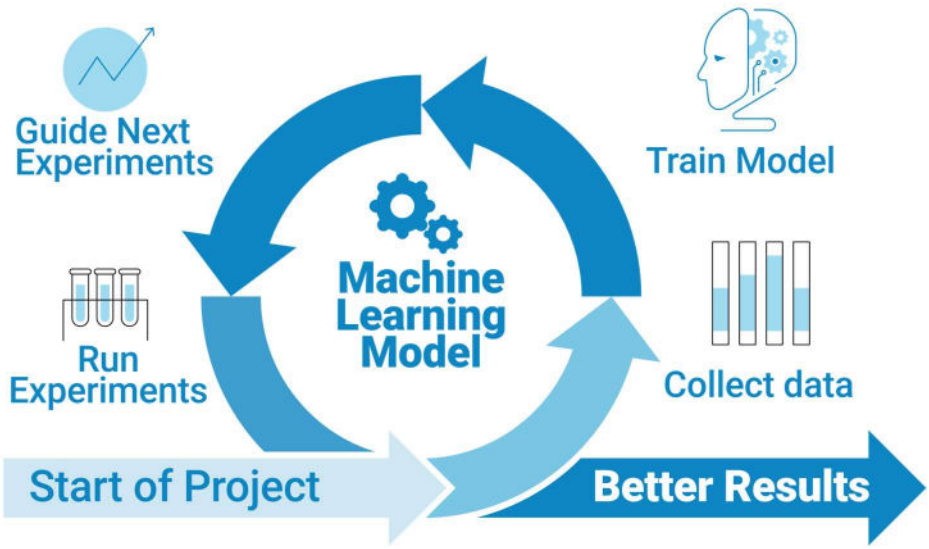


Figure 4. A schematic of the adaptive experimental design approach.

Figure 5 shows the performance of a variety of standard DOE approaches at finding the optimal point of a simple 4-factor analytic function. Almost all the standard approaches achieve comparable results, finding similarly good values in similar numbers of experiments.

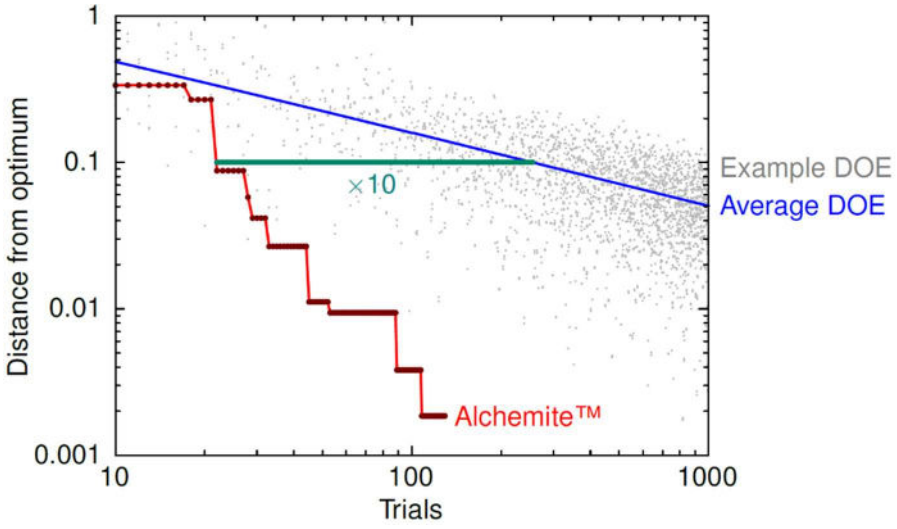


Figure 5. Performance of a variety of standard DOE approaches (gray points) at finding the optimal point of a 4-factor analytic function. The vertical axis shows how far the best result is from the optimum. The horizontal axis shows how many experiments are required to achieve this result. Alchemite™ (red) gives superior accuracy in 10 x fewer experiments.

Machine learning-driven adaptive experimental design, such as that enabled by the **Alchemite™** ML method, finds better results quicker. It requires, in theory, ten times fewer experimental measurements to find formulations much closer to the optimal result than those identified by standard DOE. In discovery or development projects, this translates directly to many-fold savings in the time and resources required to achieve project goals, improving efficiency and productivity of the R&D process.

Case study examples

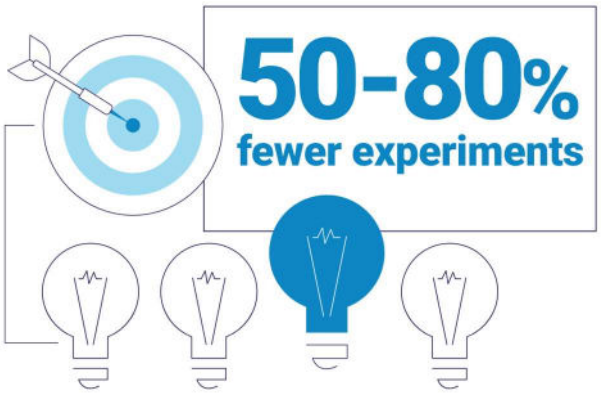
Figure 5 showed a theoretical comparison of adaptive and traditional DOE, proposing a 90% advantage through the machine learning approach. What is achievable in practice?

Johnson Matthey, in their studies of catalysis formulations using Alchemite™, have reported halving of experimental workload in one project and identification of a new experimental route in another project that would lead to a five-fold reduction in the number of experiments.

The **Advanced Manufacturing Research Center (AMRC)** used Alchemite™ to plan a test program for manufacturing research in composite tooling that achieved project objectives with 80% fewer tests.

Domino Printing Sciences saw a dramatic reduction in the amount of experimentation needed in an ink reformulation project. They found new formulations after two batches of experiment on 12 formulations proposed by Alchemite™. Traditional DOE would have required 1,800 formulations.

The exact savings in time and cost in an experimental program taking an adaptive DOE approach will vary based on the details of the system being studied and project objectives. But, based on their experience of many such projects, including those above, the Intellegens team expects savings in the range 50-80%.



More on these and other case studies can be found at **intellegens.com/casestudies**.

Guide your experimental design using Alchemite™

With the **Alchemite™** software, you transform R&D with machine learning by easily experimenting, modeling, and visualizing real-world data. Choose the best experiment to run next by quickly assessing the accuracy and confidence levels of results.

Alchemite™ has several advantages for DOE when compared to other ML methods. Unlike most ML methods, it can train models using sparse data (i.e., data with missing measurements) and data that is noisy. In other words, it works with typical experimental and process datasets. This means that researchers can get started much sooner on the iterative cycles of adaptive experimental design, without needing to first generate 'clean' datasets.

Alchemite™ also incorporates accurate uncertainty quantification methods. This enables it to predict and map the design landscape together with associated confidence levels, thus enabling scientists and engineers to more effectively identify the next best experiments to run to most rapidly succeed in discovery projects.

All of this makes Alchemite™ a powerful DOE tool. Its benefits when compared with traditional DOE methods are summarized in Table 1.

Table 1. Comparison of traditional DOE with Alchemite™

Limitation of traditional DOE	Alchemite™ approach
Still result in a high experimental burden	Suggests the most important experiments, giving a 50-80% reduction in the number needed
It's hard to address cross-correlations; these methods often model one parameter at a time	Captures complex, high dimensional, non-linear relationships: ideal to map design space for materials, formulations, chemicals, & processes
Can require statistical expertise	Method learns from the data provided to build a model – the user does not need statistical knowledge to set it up
Maps out a set of experiments, but the analysis and understanding of results is a separate task	Creates a model that can be used to understand what drives specific properties, and as a predictive tool

Alchemite™ can be viewed as an evolution from traditional DOE. But familiar DOE methods remain widely used and valuable in many scenarios. The approaches can be used in a complementary fashion.

Table 2. How traditional DOE and Alchemite™ offer complementary capabilities

Traditional DOE	Alchemite™
Use it to explore all options	Use it to efficiently achieve a goal
Use it to gather first data on a new problem	Use it to gain value from existing data, including merging data from previous projects
All experiments can be run simultaneously	Typically iterate through experiments
Useful if you require a guarantee on number (not quality) of experiments	Aim for the highest quality experiments
Uses / requires advanced statistical knowledge	Pose question, get an answer in the domain language (no statistical / ML knowledge needed)

Alchemite™ Software

The **Alchemite™ Suite** is a range of easy-to-use R&D tools, each focused on a key challenge for R&D managers, scientists, experimentalists, or data scientists. Give the right app to the right team member, speeding and informing their work. Then share results and collaborate across your team, creating an integrated machine learning solution for your R&D organization.

For Design of Experiments, **Alchemite™ Designer** provides a simple web browser-based user interface enabling you to set up and run DOE projects in just a few button clicks – no lengthy training courses, no need for advanced statistics, and no need for coding.

More at **intellegens.com/doe/**.

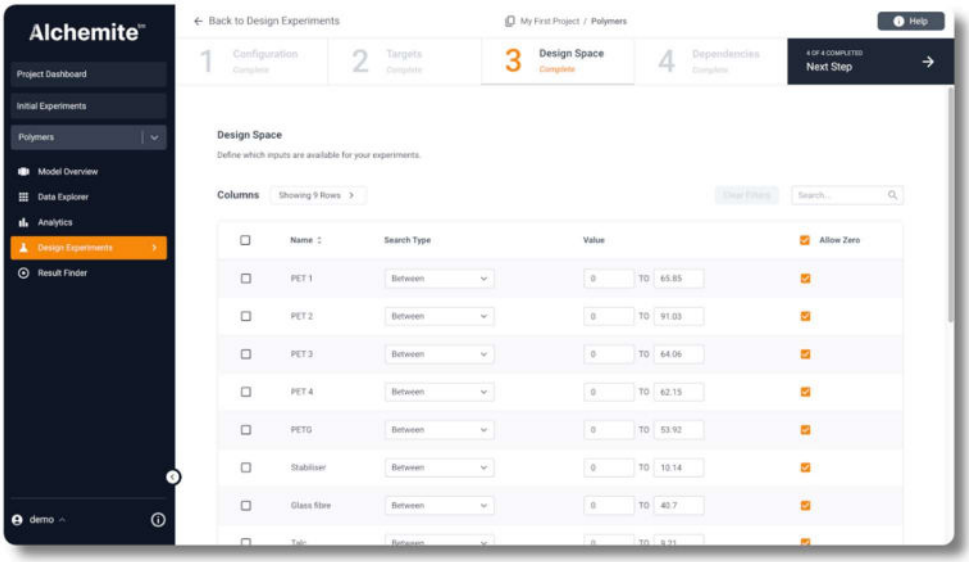


Figure 6. Alchemite™ Designer – DOE made easy.

About Intellegens

Our vision is that machine learning will drive innovation and deliver value wherever data is used in R&D. We aim for best-in-class easy-to-use machine learning software for data analysis in chemicals, materials, life science, and manufacturing. Our Alchemite™ technology originated at the University of Cambridge and development is on-going at Intellegens, in close collaboration with our growing community of customer organizations. These represent sectors including additive manufacturing, aerospace, alloys, batteries, biopharmaceuticals, ceramics, chemical processes, composites, consumer products, cosmetics, drug discovery, energy, food and beverage, formulated products, paints, plastics, and printing technology.

www.intellegens.com | info@intellegens.com