Design Before Prediction

A hybrid DOE-ML workflow for reducing complexity in biological models

Chee Ping Ng, Founder
AvatarsBio, The Netherlands

ABSTRACT

Biological experiments and simulations often begin with too many plausible variables. In endothelial sprouting, dendritic-cell migration, permeability assays, organ-on-chip systems, and omics studies, outcomes may depend on matrix properties, growth factors, enzyme activity, cell density, gradients, timing, batch effects, and many other inputs.

Machine learning is widely used to analyze complex biological data, but when the variable space is large and data are limited, a prediction-first approach can become expensive, data-hungry, and difficult to interpret.

This white paper presents a hybrid DOE–ML workflow in which DOE helps structure the experimental or simulation campaign, and machine learning is applied after data generation to refine patterns among the most promising variables. The aim is to identify likely key drivers first, then use machine learning on a smaller and more meaningful set of variables.

The goal is not to replace biological expertise or experimental validation, but to help scientists move from too many possible variables to a smaller set of testable drivers.

1. THE PLANNING PROBLEM: TOO MANY PLAUSIBLE VARIABLES

Many biological systems are shaped by multiple interacting factors, including biological variables, preparation steps, and protocol conditions [1]. In endothelial sprouting, for example, outcomes may depend on biophysical, biochemical, cellular, and process-related variables - from matrix stiffness and VEGF concentration to cell density, culture duration, and batch effects (Figure 1). Similar complexity appears in cell migration studies, permeability assays, toxicity screens, organ-on-chip models, and omics perturbation experiments.

The practical challenge is not simply that many variables exist, but that not all variables matter equally. Some factors may dominate the response, while others only matter under specific conditions, appear important because they are correlated with another variable, or add complexity without improving understanding.

Early experimental plans are often built from expert judgement: prior literature, biological intuition, available reagents, and practical experience. But when the number of plausible variables becomes large, judgement alone can become difficult to scale. DOE does not replace biological intuition. It helps turn that intuition into a structured experimental design.

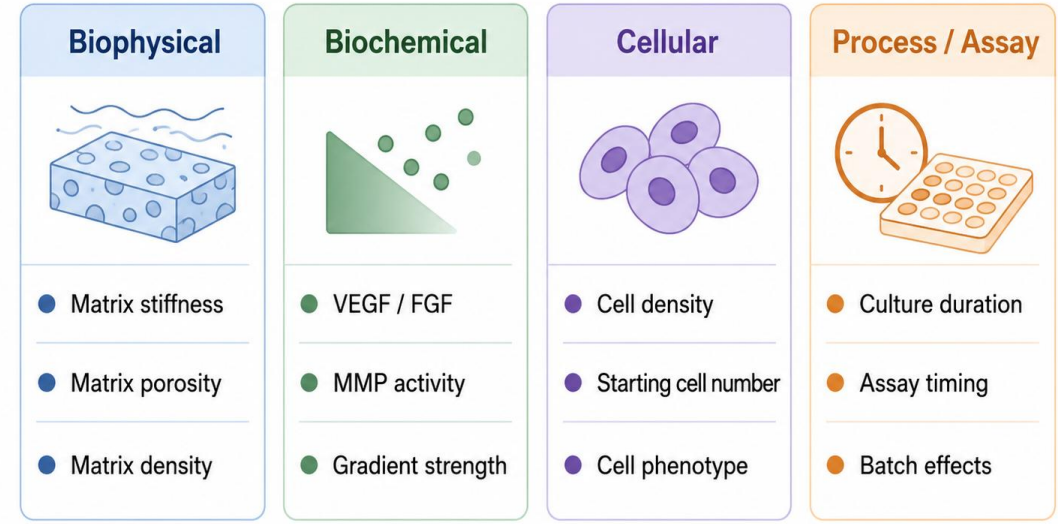

Figure 1. *Example variables in endothelial sprouting. Many plausible variables may influence the outcome. The figure illustrates why a structured planning method is useful before deeper modelling.*

2. DESIGN BEFORE PREDICTION

Machine learning is powerful when enough relevant data already exist. But in many biological projects, the challenge comes earlier: researchers have a biological question, candidate variables, practical constraints, and limited resources, but not yet a large, well-structured dataset.

If the first goal becomes "generate enough data for machine learning," the experimental plan may become large and unfocused. A DOE-first approach asks a different question:

> ***What is the smallest structured set of experiments or simulations that can tell us which variables deserve deeper attention?***

This changes the role of machine learning. Rather than designing a large experiment or simulation campaign first and asking ML to make sense of everything afterward, DOE helps structure the work before it begins. ML can then be applied later, where it adds more value: refining patterns among the most promising variables.

In this workflow, DOE is not mainly a post-hoc analysis layer. It is part of the planning step.

3. DOE AS A FOCUSING LENS

Design of Experiments is a structured way to plan experiments or simulations so that useful information can be gained with fewer runs [2]. Instead of changing one variable at a time, DOE tests carefully chosen combinations of variables, making it possible to estimate which factors have the strongest effect on the outcome. Classical screening designs, such as Plackett-Burman designs, were developed to explore many factors efficiently with relatively few experimental runs [3].

Related approaches, including Taguchi orthogonal arrays and robust-design thinking [4], also emphasize learning efficiently under practical variation.

For biological modelling, DOE can act as a focusing lens. It helps move from a broad list of possible variables to a smaller set of likely key drivers. This workflow (Figure 2) does not introduce DOE or machine learning as new methods. Both are well established. The value lies in the framing: using DOE before prediction, so machine learning is applied after the biological question has been structured.

In the strongest use case, DOE helps create the dataset by turning candidate variables and realistic ranges into a structured run plan. After those experiments or simulations are performed, the results can be analyzed to rank the most influential factors. ML can then be applied to the reduced set of variables, making the refinement step more focused, interpretable, and less data-hungry.

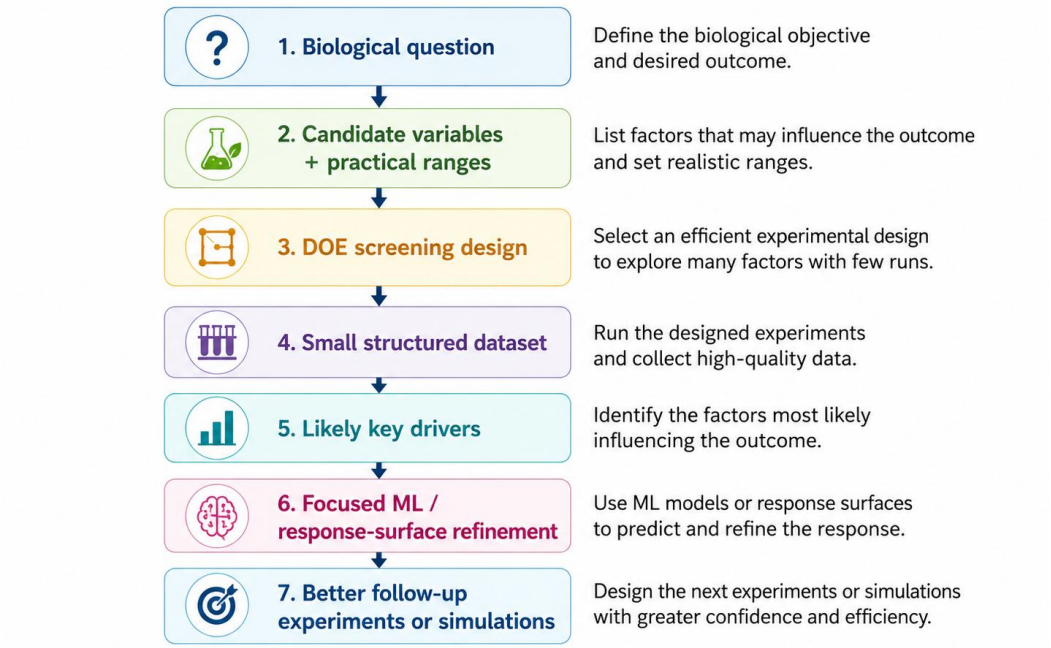

Figure 2. *DOE-ML planning workflow. DOE structures the first screen; machine learning or response-surface modelling is applied after the factor space has been narrowed.*

4. WHAT THE HYBRID WORKFLOW ADDS

DOE and machine learning answer different questions. DOE is useful when the first challenge is planning: deciding which variables to test, how to combine them, and how to learn as much as possible from a limited number of runs. Its strength is interpretability. DOE can estimate signed effects, showing whether increasing a factor is associated with a higher or lower response.

Machine learning is useful when the challenge is refinement: detecting nonlinear patterns, modelling interactions, or improving prediction once enough focused data exist. Its strength is flexibility. However, ML feature importance often gives relative importance rather than a direct biological direction. Random forests are one common machine-learning approach for estimating variable importance in predictive models [5].

The hybrid workflow combines these strengths rather than replacing one method with the other. DOE helps narrow the search space; ML then explores the reduced space in more detail. This distinction matters for biologists: knowing that a variable matters is useful, but knowing how it matters is often more useful.

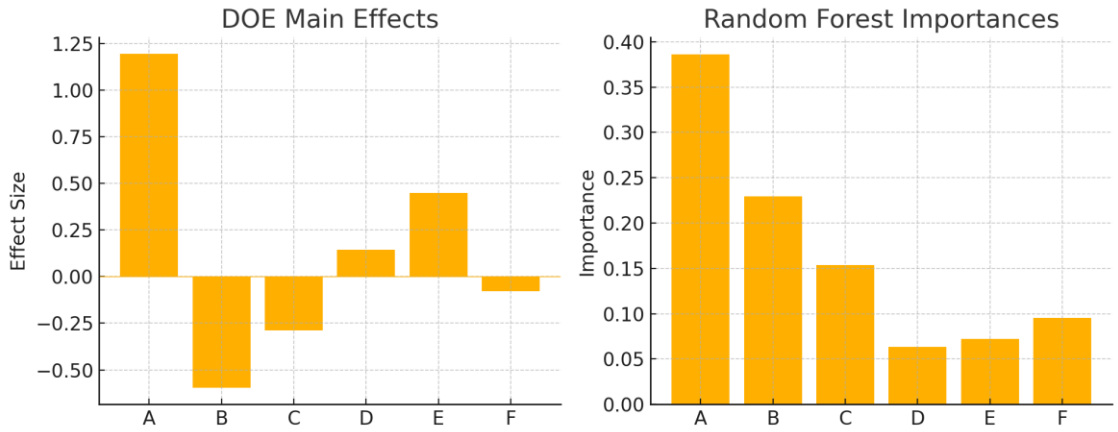

Figure 3. *DOE effects versus random forest importance. DOE-style screening provides signed effect estimates, while random forest feature importance provides relative predictive importance. The two approaches can be complementary rather than competing.*

A second advantage is efficiency. If the number of candidate variables is large, a prediction-first ML approach may require many observations before it becomes reliable. A DOE screening step can reduce the number of variables first, allowing the later ML step to focus on fewer, more meaningful factors.

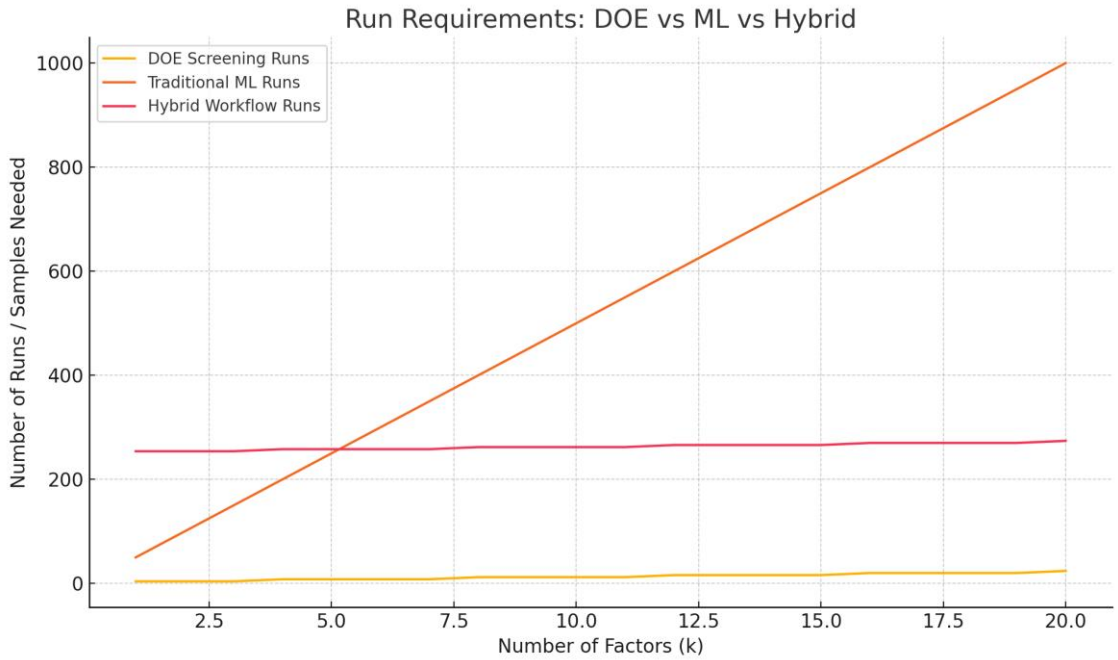

Figure 4. *Illustrative run requirements for DOE, ML, and hybrid workflows. This comparison is illustrative, not universal. It assumes DOE screening scales slowly with the number of candidate factors, while a prediction-first ML approach requires more samples as the number of variables increases. Exact numbers depend on noise, replication, model complexity, and the biological system.*

5. PROSPECTIVE FIRST, RETROSPECTIVE WHEN USEFUL

The strongest version of this workflow is prospective: DOE is used before data generation to plan a structured set of experiments or simulations, helping ensure that the resulting dataset is balanced, interpretable, and useful for follow-up modelling.

The same thinking can also be useful when data already exist. In post-experimental analysis, DOE-inspired screening can help assess variation, confounding, and which variables appear worth testing in a follow-up study. These results should be treated as exploratory rather than confirmatory, because existing datasets may not be randomized, balanced, or designed to separate factor effects cleanly.

In short, ***design first where possible, structure existing data where necessary, and predict only after the problem has been focused.***

6. LIMITATIONS

This pipeline should not be oversold. DOE screening does not prove final biological mechanisms by itself. It identifies likely drivers under the tested conditions.

Machine learning does not remove the need for biological interpretation. It can refine patterns, but those patterns still need to be checked experimentally or mechanistically.

Several practical issues remain important. Factors should be selected thoughtfully, ranges should be biologically realistic, and known batch effects should be blocked or controlled where possible. Experiments or simulations should be randomized when feasible. Hidden confounders can mislead interpretation, and variables that look unimportant in an initial screen may still matter through interactions.

The pipeline is therefore best described as a planning and prioritization tool, not a final proof engine.

CONCLUSION

Complex biological models often begin with too many possible variables. A DOE-first workflow offers a structured way to identify likely key drivers before applying deeper modelling.

Machine learning can then be used where it adds the most value: exploring nonlinear patterns, interactions, and prediction within a better-defined factor space.

Design first. Predict second. Validate always.

For biological modelling and experimental planning, this approach supports better decisions before larger investments are made. It also reflects the broader AvatarsBio approach: better biological modelling starts with better planning.

NOTE

This planning perspective also informs AvatarsBio's work on Antonii, an experimental-design assistant focused on structured DoE planning before deeper modelling.

SELECTED REFERENCES

1. Bohyn, A., Schoen, E. D., Ng, C. P., Bishard, K., Haarmans, M., Trietsch, S. J., & Goos, P. (2023). Design and analysis of a microplate assay in the presence of multiple restrictions on the randomization. https://doi.org/10.1007/s13253-023-00570-1

2. Montgomery, D. C. (2019). Design and Analysis of Experiments (10th ed.). Wiley.

3. Plackett, R. L., & Burman, J. P. (1946). The design of optimum multifactorial experiments. Biometrika, 33(4), 305-325. https://doi.org/10.1093/biomet/33.4.305

4. Taguchi, G. (1986). Introduction to Quality Engineering: Designing Quality into Products and Processes. Asian Productivity Organization.

5. Breiman, L. (2001). Random forests. Machine Learning, 45, 5-32. https://doi.org/10.1023/A:1010933404324