The implementation of a design of experiments strategy to increase recombinant protein yields in yeast (review)

Nagamani Bora[1], Zharain Bawa[1], Roslyn M. Bill[1] and Martin D. B. Wilks[2†]

[1]School of Life and Health Sciences, Aston University, Aston Triangle, Birmingham, B4 7ET, UK

[2]Smallpeice Enterprises Ltd, 27 Newbold Terrace East, Leamington Spa, Warwickshire, CV32 4ES, UK

[†]To whom correspondence should be addressed:

Martin D. B. Wilks

Tel: +44 (0) 01926 336423

E-mail: martinw@smallpeice.co.uk

Abstract

Biological processes are subject to the influence of numerous factors and their interactions, which may be non-linear in nature. In a recombinant protein production experiment, understanding the relative importance of these factors, and their influence on the yield and quality of the recombinant protein being produced, is an essential part of its optimisation. In many cases, implementing a design of experiments (DoE) approach has delivered this understanding. This chapter aims to provide the reader with useful pointers in applying a DoE strategy to improve the yields of recombinant yeast cultures.

Key words: Design of experiments; process optimisation; process development; process characterisation.

1. Implementing a design of experiments approach

The design of experiments (DoE) approach involves the systematic application of statistics to an experimental set-up in order to determine how combinations of a series of input parameters or ‘factors’ set at different ‘levels’ (such as culture temperatures of 20 ºC, 25 ºC, 30 ºC, pH of 5, 6, 7 and dissolved oxygen concentrations of 30%, 40%, 50%) affect an output or ‘response’ (such as recombinant protein yield) (***1***). DoE is therefore an effective way of investigating the impact of multiple conditions whilst reducing the overall number of experiments, without compromising the quality of the data. Information on the relationship between the factors and the response is extracted in the form of an equation: the use of a statistically-robust design means that it is not necessary to perform experiments to examine all possible combinations of factors and levels in order to obtain the equation. In section 2.5, we discuss a recent study exploring three factors set at three levels. The statistical design used required only 13 experimental combinations out of a possible 27 to be examined in order to identify the optimal the relationship between the response (in this case, the yield of recombinant green fluorescent protein secreted from the yeast, *Pichia pastoris*) and the factors (temperature, pH and dissolved oxygen concentration) (***2***).

In a typical DoE set-up, the factors to be tested, the number of levels, the number of replicates to be performed (e.g. n = 3) and the layout of the experiment are specified in a design matrix (*see* section 2.5 and **Note 1**). Statistical analysis then fits the response, derived by running the specific experimental combinations defined by the matrix, to a model (which may be linear or non-linear) and quantitatively determines the effect of each factor on this response. The use of replicates means that the amount of error in the model can be determined as well as whether, or not, any lack of fit present is statistically significant. DoE therefore offers many benefits over more traditional experimental approaches of varying one factor at a time (OFAT), which are typically inefficient, expensive and time consuming (***3***).

DoE was first proposed as an alternative to OFAT by Sir Ronald A. Fisher in 1935 (***4***), who based his approach on the statistical method known as 'analysis of variance' (ANOVA). It was later used by Genichi Taguchi in the 1950s to improve the quality of manufactured goods and is now widely implemented in modern biotechnological applications (***5***). DoE as a general strategy is typically involved in both the early and late stages of industrial bioprocess development. More specifically, DoE is seen as being integral to the process of securing regulatory approval for products from organisations such as the US Food and Drug Administration (*see* http://www.fda.gov/regulatoryinformation/guidances/ucm128003.htm). In the following sections the application of DoE to screening, characterisation and optimisation of protein production experiments are introduced, followed by an overview of appropriate experimental set-ups.

1.1 Screening for key factors

Screening designs are used to reduce the factors under initial consideration (which could be 7–12 or more, based on previous experimentation and guidance from the literature (***6***)) to a shortlist of 3–5 that warrant further, more detailed study (***2***). Typically, fractional factorial designs are used at this screening stage where a 'fraction' of the experimental runs are selected from a full factorial design. This allows for a cheap and rapid investigation but may affect the data quality. The compromise between the size of the fraction and the quality of the data can be judged by checking the resolution of the design: a design of at least resolution V is typically chosen (see Chapter 7 of reference (***1***) for a further explanation of design resolutions). Implicit in this type of design is that information on how the interactions between factors affect the response is confounded (i.e. distorted). However, data on the main effect of each factor on the response are of sufficient quality to make a judgement about a factor's inclusion or exclusion from subsequent experimentation. Overall, the outcome of a screening exercise should be the identification of the factors that warrant further study, as well as an understanding of their appropriate experimental ranges.

1.2. Process characterisation

The primary goal of process characterisation is to identify and quantify the influence of the key factors, typically as part of a plan for process improvement. Characterisation confirms the identities of the factors influencing the response of a process (e.g. protein yield, functional activity or stability) and enables a prediction of the optimal response under a range of operating conditions. The investment of time and resources at this stage results in better process understanding, improved reproducibility and may reduce delays in costly regulatory procedures.

1.3. Process optimisation

Since a recombinant protein production experiment is a multi-phase, multi-component process, protein yield, as well as other responses such as stability and activity, can be influenced by a wide range of factors including the composition of the culture medium, its pH, the culture temperature, the availability of dissolved oxygen in the medium and the details of the induction regime (e.g. concentration of inducer as well as the point and duration of induction). In the process optimisation stage, the goal is to 'zoom in' on a particular portion of the design space or, by changing the design used, to model any non-linear behaviour observed in the previous stages (e.g. by using the 'response surface method'; section 2.4.3). By using DoE, process optimisation becomes more systematic and informative by enabling different levels of each factor and their interactions to be related to the response. In an iterative process, data from one round of DoE results in a model that provides the information for an improved design in subsequent rounds. **Table 1** gives some examples of how DoE has been used to improve a range of different bioprocesses, including recombinant protein production experiments.

(Insert Table 1 here)

2. Experimental set-up

Devising and analysing a DoE has been considerably simplified in recent years with the advent of a range of specialist software packages including MiniTab® (www.minitab.com), Modde® (www.umetrics.com), ECHIP® (www.experimentationbydesign.com/index.php) and Design-Expert® (http://www.statease.com/software.html). These packages are well supported by their providers (see the websites above for further information).

Before starting a DoE, the experimental goals and criteria for success should be clearly articulated. A relevant example would be the goal of determining the key factors influencing a response such as recombinant protein yield and then to use that information to maximise the yield, as measured in mg L^{-1} (**Table 1**). Only once these goals and criteria are defined can a valid DoE strategy be developed, including a plan of action in the event that the experiments do not turn out as expected. The effect of selected factors on a process response is then examined at a number of levels, depending on the experimental design chosen. It should be noted that the temptation to add a large number of factors or responses just to see how they change should be tempered by the fact that this may divert focus from those that are critical to meeting the goals of the DoE, and therefore should be avoided. In the following sections, the key components in setting up a DoE are considered.

2.1 Factor selection

Factors are usually variables, which can have defined set-points. They might include pH, temperature, dissolved oxygen concentration or the concentration of medium components. Input factors may also be an 'attribute', e.g. the presence or absence of a medium component at a level that does not vary. Other factors, which may or may not be controllable and which are referred to as 'noise factors', should be considered in the DoE (*see* **Note 2**). The presence of noise during the experiment can distort the results to the extent that incorrect conclusions are drawn. Their effect may therefore be minimised by using 'blocking' or 'randomisation' in the design (see Chapter 2 of reference (***1***) for further details). 'Blocking' mitigates 'categorical' noise, e.g. that introduced by using 'bioreactor 2' in some experimental runs instead of 'bioreactor 1'. 'Randomisation' mitigates 'variable' noise, e.g. day-to-day variations in laboratory temperature.

2.2 Level selection

For the simplest designs, known as 2^k designs, each of k factors is examined at two different levels, coded as -1 (for the low level) and +1 (for the high level) in a design matrix. This type of design can also be modified to accommodate many more levels. However, it is important to bear in mind that examining certain levels may not be biologically practical. For example a growth medium with a very low pH may inhibit the growth of the organism being studied, while maintaining very high dissolved oxygen concentrations may not be experimentally feasible. Since the difference in response observed experimentally is related to the difference in the levels of each factor, an equation can be derived that describes the relative importance of each factor on any change to that response.

2.3 Response selection

It is possible to carry out DoE where the response is an attribute (*7*) (e.g. the protein produced is functional or not), but most commonly the response can be measured on a continuous, variable scale. Protein yield, protein activity and culture density fall into this category.

2.4 Experimental design selection

The choice of experimental design, as discussed in the next section, is dependent on the purpose of the DoE (screening, characterisation or optimisation) and the number of factors under consideration, as summarised in **Table 2.**

(Insert Table 2 here)

2.4.1 Factorial designs

Full factorial designs (e.g. 2^k designs) can be used for screening a small number of factors (≤4) in order to identify the most significant ones, but can also be used sequentially to model and refine a process. Each factor can have two or more levels and the design generated will include all possible combinations of the factors and levels. In contrast, fractional factorials are more efficient designs used to screen a large number of factors (≥5) to find the few that are significant, but compromise on the quality of the information on the interactions between the factors. Consequently, full factorials should be used to estimate the effects of interactions, which may be missed in a fractional design.

In cases where a large number of factors is to be studied, whilst minimizing the experimental runs, Plackett-Burmann designs (***8***) may be considered. Alternatively, a D-optimal approach may be suitable, as it allows a subset of experimental runs to be selected (***9***). The D-optimal design also allows the inclusion of both quantitative and qualitative (attribute) factors with a mixed number of levels. Analysis of factorial designs are typically done using ANOVA (***1***), which lead to a first-degree polynomial equation describing the factors that influence the response of interest. However, full factorial designs may also be analysed using regression (see section 2.4.3).

2.4.2 Taguchi designs

Taguchi's orthogonal arrays (***1***), which were originally created before the widespread use of DoE software, are highly fractional designs that can be used to estimate main effects using only a few experimental runs. These designs are not only applicable to two level factorial experiments, but also can investigate the main effect of a factor with more than two levels. Designs are also available to investigate the effects when the factors do not have the same number of levels. As with Plackett-Burman designs, these designs require the experimenter to compromise on data describing any interaction effects. Taguchi designs are often focused on reducing the sensitivity of a response to noise. A recent example of the use of this type of approach is in the improvement of biological assays (***10***).

2.4.3 Response surface method designs

Factorial designs are sufficient to determine which factors have an impact on the response of interest. Once these have been identified, a more complex design can be implemented to generate a second-degree polynomial equation, which can be used to maximise, minimise, or achieve a specific response. Regression models are used for analysis of the response, as quantifying the relationship between the response and the factors is of the most interest, rather than the identification of the important factors: this is known as the response surface method (RSM) (***11***). Once the resultant equation has been validated, the behaviour of a process can be predicted, for example in maximising protein yield (**Table 1**).

In order to analyse response surfaces, special experimental designs are used that help the experimenter fit the second order equation to the response in the minimum number of runs. Examples of these designs include the Central Composite design (CCD) and the Box-Behnken design (BBD) (***11***). CCD is a two level, full or fractional factorial design augmented with a number of centre points and other chosen runs (***12***). BBD is similar in concept to Plackett-Burman designs, but with factors at 3 levels. Note that a full factorial design, with all factors at three levels, would also provide all required regression parameters. However, this type of design is expensive to use, requiring 27 runs compared, for example, with the 13 required in a BBD and 15 in a CCD (***13***). One further advantage of BBD in biological applications is that it does not contain factor combinations for which all factors are simultaneously at their highest or lowest levels, thus avoiding experiments that need to be performed under extreme conditions. However, if an experimenter is interested in the responses at the extremes, BBD may not be suitable.

2.5 Data analysis: a case study

A BBD was used to optimise the yield of recombinant green fluorescent protein (secreted from *P. pastoris*) as a function of the three most commonly-varied process parameters: culture temperature (T), pH and the percentage of dissolved oxygen in the culture medium (DO) (***2***). Based on the results of a first optimisation run (see http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2717918/#supplementary-material-sec), the three factors (T, pH, DO) were each varied at three levels, coded as –1 (lowest value), 0 (middle value) and +1 (highest value); MiniTab® statistical software (version 15.1.1.0) was used to construct the experimental matrix shown in **Table 3**.

2.5.1 Model building

The predictive model generated from the outputs of the matrix is described by **Equation 1** and **Figure 1**.

Equation 1: Yield (ng mL^{-1} OD_{595}^{-1}) = $(-21814.9 + (328.6\times T) + (5502.1\times pH) - (37.8\times DO) - (325.6\times pH^2) - (47.9\times T\times pH) + (6.4\times pH\times DO))\times\gamma$, where; T = temperature (°C), DO = dissolved oxygen (%) and $\gamma = 0.3$ and is the conversion factor from RFU to ng of protein

This model was derived in Minitab® (see www.minitab.com for a detailed description of its use) by removing insignificant terms from the full model based on their *p*-values (***14***). The adjusted R^2 value (R^2_{adj}) for the regression changed as each term was removed, R^2_{adj} being a modification of R^2 that adjusts for the number of terms in the model (***14***). R^2_{adj} values of

0.160 (full model), 0.115 (1 term removed), 0.274 (2 terms removed), 0.324 (3 terms removed) and 0.292 (4 terms removed) indicated that the model with 3 terms was statistically soundest. In **Equation 1**, the yield was converted to ng mL^{-1} OD_{595}^{-1} from RFU mL^{-1} OD_{595}^{-1} using an experimentally-derived factor.

Yields improved at lower T and higher pH, although at the temperatures tested, T did not have a large effect on yield (**Figure 1A**), which was highest around pH 7 (**Figure 1B**). Yields also increased with increasing DO (**Figure 1C**). **Figure 1D** shows the ε^2 results, which indicate the influence of each of the factors and their interactions within **Equation 1**. The data support the view that pH is a key factor as the ε^2 values for pH, pH^2 and the interactions of pH with both T and DO are substantial. DO alone is also important, while in contrast the effect of T alone makes a relatively small contribution, in agreement with the main effects plots (**Figure 1**).

(Insert Figure 1 here)

2.5.2 Model validation

The results of the statistical validation of this model by ANOVA are shown in **Table 4**. A recent report suggests that this type of analysis is often missing in published models and that good models from the literature have R^2 values > 0.75 with values below 0.25 being considered poor (***15***). This suggested that the model was of acceptable quality in line with recent DoE studies of protein production in *E. coli* (***15***).

(Insert Table 4 here)

The model was also validated experimentally by running the factor combinations shown in **Table 5**, which had not been used in the model building process, and comparing the fit of the experimental output to the predicted response from the model (**Figure 2**).

(Insert Table 5 and Figure 2 here)

Nine of the twelve data points were within 40 ng mL^{-1} OD_{595}^{-1} (i.e. within 5–15 %) of the predicted value. The three data points outside this range (with T, pH, DO values of 20, 7.5, 60; 28, 7.5, 90 and 27.5, 6.7, 80), were within 16–25 % of the predicted value, and were not correlated in any obvious manner. The experimental conditions leading to the maximum yield were predicted to be 21.5°C, pH 7.6, DO 90 % (**Figure 3**), which was confirmed experimentally (***2***).

(Insert Figure 3 here)

3. Notes

1. Before starting any experimentation, ensure the reliability of any gauges or measurement devices to be used and record any process drifts or changes (such as a change of operator) during the experiment. A minimum of three replicates should be done per experiment. Where possible, use the same starting materials for all experiments. Document all raw output data as well as the averaged data.

2. Noise factors may be categorical (such as noise associated with a change in operator or item of equipment) or random (the ambient temperature or humidity).

4. References

1. Anthony, J. (2003) *Design of Experiments for Engineers and Scientists*, Butterworth-Heinemann, Oxford.
2. Holmes, W. J., Darby, R. A. J., Wilks, M. D. B., Smith, R. and Bill, R. M. (2009) Developing a scalable model of recombinant protein yield from *Pichia pastoris*: the influence of culture conditions, biomass and induction regime, *Microb Cell Fact* ***8***, 35.
3. Czitrom, V. (1999) One-factor-at-a-time versus designed experiments, *American Statistician* ***53***, 2.
4. Fisher, R. A. (1971) *The Design of Experiments*, 9th ed., Macmillan, London.
5. Rao, R. S., Kumar, C. G., Prakasham, R. S. and Hobbs, P. J. (2008) The Taguchi methodology as a statistical tool for biotechnological applications: a critical appraisal, *Biotechnol J* ***3***, 510-523.
6. Knospel, F., Schindler, R. K., Lubberstedt, M., Petzolt, S., Gerlach, J. C. and Zeilinger, K. (2010) Optimization of a serum-free culture medium for mouse embryonic stem cells using design of experiments (DoE) methodology, *Cytotechnol* ***62***, 557-571.
7. Bisgaard, S. and Fuller, H.T. (1994-95) Analysis of factorial experiments with defects or defectives as a response, *Quality Eng* ***7***, 429-443
8. Yuan, L.-L., Li, Y.-Q., Wang, Y., Zhang, X.-H. and Xu, Y.-Q. (2008) Optimization of critical medium components using response surface methodology for phenazine-1-carboxylic acid production by *Pseudomonas* sp. M-18Q, *J Biosci Bioeng* ***105***, 232-237.
9. de Aguiar, P. F., Bourguignon, B., Khots, M. S., Massart, D. L. and Phan-Than-Luu, R. (1995) D-optimal designs, *Chemometrics Internat Lab Sys* ***30***, 199-210.

10. Luo, W., Pla-Roca, M. and Juncker, D. (2011) Taguchi design-based optimization of sandwich immunoassay microarrays for detecting breast cancer biomarkers, *Anal Chem* ***83***, 5767-5774.

11. Myers, R. H. and Montgomery, D. C. (1995) *Response Surface Methodology: Process and Product Optimization using Designed Experiments*, 1st ed., Wiley, New York.

12. Einsfeldt, K., Severo Junior, J. B., Correa Argondizzo, A. P., Medeiros, M. A., Alves, T. L., Almeida, R. V. and Larentis, A. L. (2011) Cloning and expression of protease ClpP from *Streptococcus pneumoniae* in *Escherichia coli*: Study of the influence of kanamycin and IPTG concentration on cell growth, recombinant protein production and plasmid stability, *Vaccine,* doi:10.1016/j.vaccine.2011.05.073.

13. Ferreira, S. L., Bruns, R. E., Ferreira, H. S., Matos, G. D., David, J. M., Brandao, G. C., da Silva, E. G., Portugal, L. A., dos Reis, P. S., Souza, A. S. and dos Santos, W. N. (2007) Box-Behnken design: an alternative for the optimization of analytical methods, *Anal Chim Acta* ***597***, 179-186.

14. Montgomery, D. C. and Peck, E. A. (1982) *Introduction to Linear Regression Analysis*, John Wiley & Sons.

15. Mandenius, C. F. and Brundin, A. (2008) Bioprocess optimization using design-of-experiments methodology, *Biotechnol Prog* ***24***, 1191-1203.

16. Shi, F., Xu, Z. and Cen, P. (2006) Efficient production of poly-gamma-glutamic acid by *Bacillus subtilis* ZJU-7, *Appl Biochem Biotechnol* ***133***, 271-282.

17. Garcia-Arrazola, R., Dawson, P., Buchanan, I., Doyle, B., Fearn, T., Titchener-Hooker, N. and Baganz, F. (2005) Evaluation of the effects and interactions of mixing and oxygen transfer on the production of Fab' antibody fragments in *Escherichia coli* fermentation with gas blending, *Bioprocess Biosyst Eng* ***27***, 365-374.

18. Wang, Y.-H., Yang, B., Ren, J., Dong, M.-L., D., L. and Xu, A. L. (2005) Optimization of medium composition for the production of clavulanic acid by *Streptomyces clavuligerus*, *Process Biochem* ***40***, 1161-1166.

19. Pritchett, J. and Baldwin, S. A. (2004) The effect of nitrogen source on yield and glycosylation of a human cystatin C mutant expressed in *Pichia pastoris*, *J Ind Microbiol Biotechnol* ***31***, 553-558.

20. Adinarayana, K., Ellaiah, P., Srinivasulu, B., Bhavani Devi, R. and Adinarayana, G. (2003) Response surface methodological approach to optimize the nutritional parameters for neomycin production by *Streptomyces marinensis* under solid-state fermentation, *Process Biochem* ***38***, 1565-1572.

Figure legends

Figure 1: A main effects plot showing the influence of each of the factors (A) T, (B) pH and (C) DO on the response (specific yield). Panel D shows the ε^2 analysis which indicates the influence of each of the factors and their interactions on the model. The value reported for ε^2 is the quotient of the sum of squares for the factor and the total sum of squares (from Table 4) expressed as a percentage.

Figure 2: Demonstration of the predictive capacity of the model. A scatter plot of the predicted versus experimental response is shown. Each check point condition was from within the model design space, but had not been used to build the model. The fit to the line of parity (y=x) is shown with $R^2 = 0.57$.

Figure 3: A response surface contour plot showing how yield per cell changes with each of the input factors. T = temperature (°C), pH = pH and DO = dissolved oxygen tension (%). All hold values are the “0” mid-point values in the DoE matrix.

Table 1: Examples of DoE in bioprocess improvement

Protein	Goal of DoE	Statistical method used	Reference
Recombinant erythropoietin (from *P. pastoris* culture)	Maximising protein yield as a function of the temperature, pH and dissolved oxygen concentration of the culture medium	Response surface method (Box-Behnken)	Bora, N and Bill, RM, unpublished
Recombinant green fluorescent protein (from *P. pastoris* culture)	Maximising protein yield as a function of the temperature, pH and dissolved oxygen concentration of the culture medium	Response surface method (Box-Behnken)	(***2***)
Polyglutamic acid isolated from *Bacillus subtilis*	Maximizing polyglutamic acid yield as a function of the composition of the growth medium	Fractional factorial design and response surface method	(***16***)
Recombinant Fab' fragment (from *Escherichia coli* culture)	Maximising yield as a function of agitation rate and dissolved oxygen concentration	Full factorial (2^2) design	(***17***)
Clavulanic acid from *Streptomyces clavuligerus*	Maximizing clavulanic acid yield by optimizing the composition of the growth medium	Screening by fractional factorial design and optimisation by response surface method	(***18***)
Recombinant cystatin C mutant (from *P. pastoris cultures*)	Maximizing yield and protein glycosylation as a function of three nitrogen sources	Full factorial (2^3) design	(***19***)
Neomycin isolated from *Streptomyces marinensis*	Maximizing neomycin yield by optimizing the composition of the growth medium	Full factorial design and response surface method	(***20***)

Table 2: An overview of statistical designs and when to use them

Number of Factors	Screening	Characterisation	Optimisation
1	Not applicable for a single factor	Linear regression or, in cases where there is no linear fit, non-linear regression	Linear or non-linear regression
2–4	Full factorial	Full factorial	Full factorial (for linear response) or response surface method (for non-linear response)
5 or more	Fractional factorial	Full factorial on selected factors (usually <4)	Full factorial (for linear response) or response surface method (for non-linear response)

Table 3: Factors and measurable responses for the model building experiments

INPUTS (controlled on-line)			MEASURABLE RESPONSES (measured offline)				
T (°C)	pH	DO (%)	OD_{595}	RFU (mL^{-1})	Specific RFU (mL^{-1} OD_{595}^{-1})	Specific yield (ng mL^{-1} OD_{595}^{-1})	SD; n=3 (ng mL^{-1} OD_{595}^{-1})
19	6	60	20.3	8651	426.2	127.9	3.2
19	8	60	0.8	1015	1268.8	380.6	3.9
19	7	30	13.1	10984	838.5	251.6	1.3
19	7	90	12.4	9259	746.7	224.0	2.1
24	6	30	24.4	8061	330.4	99.1	1.6
24	6	90	16.2	11951	737.7	221.3	5.6
24	8	30	4.7	1564	332.8	99.8	1.1
24	8	90	1.3	1954	1503.1	450.9	1.3
24	7	60	17.6	21382	1214.9	364.5	10.1
29	7	30	24.8	25392	1023.9	307.2	0.2
29	8	60	4.4	1413	321.1	96.3	1.5
29	6	60	21.7	10349	476.9	143.1	0.3
29	7	90	15.1	17495	1158.6	347.6	3.5

The input factors were temperature (T), pH and % dissolved oxygen (DO). Relative fluorescent units (RFU) and the optical density at 595 nm (OD_{595}) were measured in triplicate 48 h post induction. The mean values are reported for 1 mL of culture. The standard deviation (SD; n=3) is given for the specific yield of the culture, where the conversion factor from RFU to ng was determined by generating a standard curve (Adapted from reference (2)).

Table 4: Statistical significance of the predictive model by ANOVA

Source	Degrees of Freedom	Sum of Squares	Mean Square	*F* statistic	*p* value
Regression	6	1288405	214734	1.96	0.217
Linear	3	586988	223083	2.04	0.21
Square	1	326196	326196	2.98	0.135
Interaction	2	375221	187610	1.71	0.258
Residual	6	657203	109534		
Total	12	1945608			

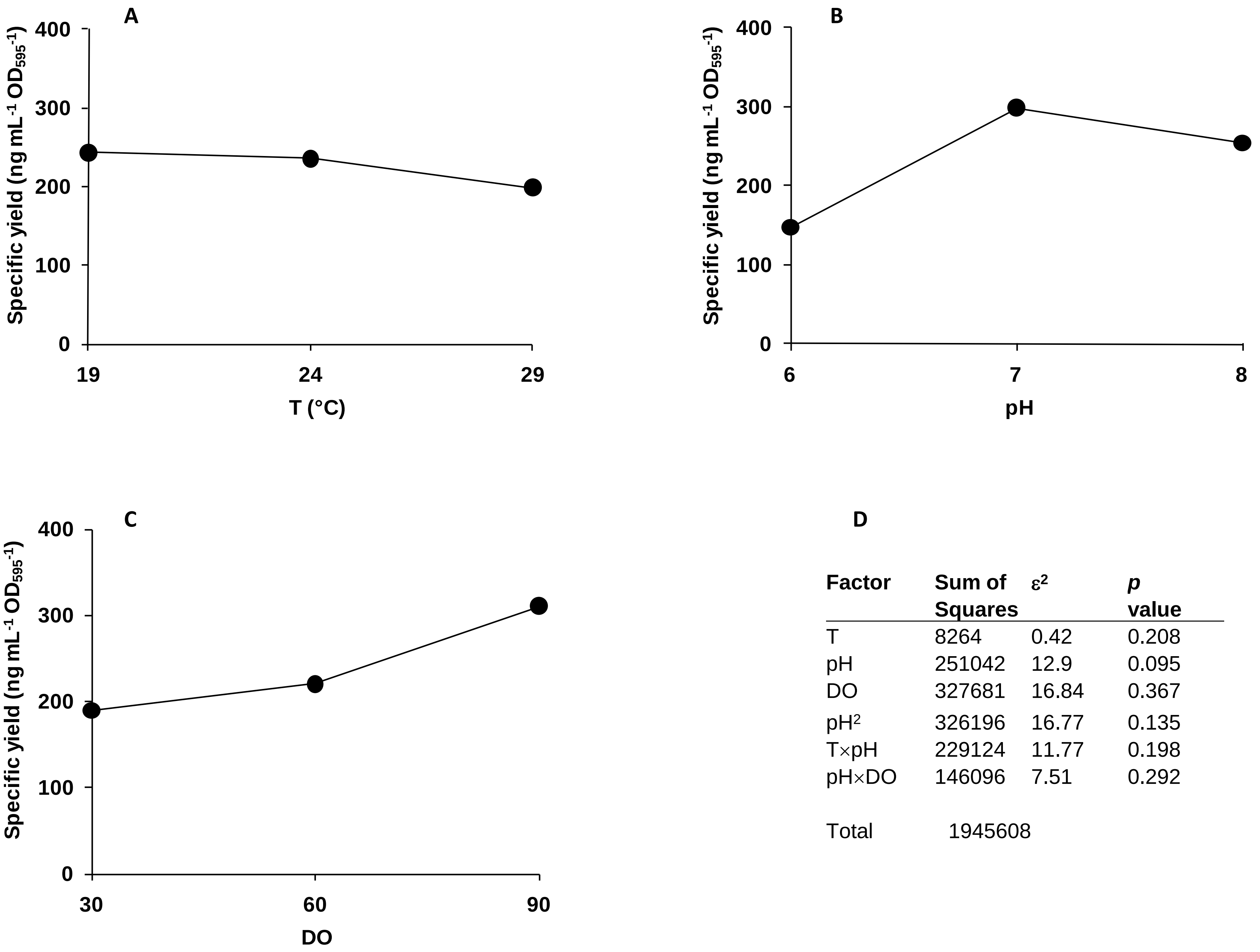
The statistical significance of the relationship between the predictors and the response of the model was assessed using ANOVA, which employs Fisher's *F*-test. The goodness of fit of the model is 66 %, as determined by the quotient of residual sum of squares/total sum of squares (R^2 = 0.66).

Table 5: Specification of the input factors for the model validation experiments

T(°C)	pH	DO(%)
20	7.5	60
20	7.7	80
27	8	50
28	7.5	90
28	6	80
23.6	7.25	60
27.5	6.7	80
27.5	6.5	60
27.5	6.3	60
21.5	7.6	20
21.5	7.6	40
21.5	7.6	60

The input factors were temperature (T), pH and dissolved oxygen (DO)

Figure 1



D

Factor	Sum of Squares	ε^2	*p* value
T	8264	0.42	0.208
pH	251042	12.9	0.095
DO	327681	16.84	0.367
pH^2	326196	16.77	0.135
T×pH	229124	11.77	0.198
pH×DO	146096	7.51	0.292
Total	1945608		

Figure 2

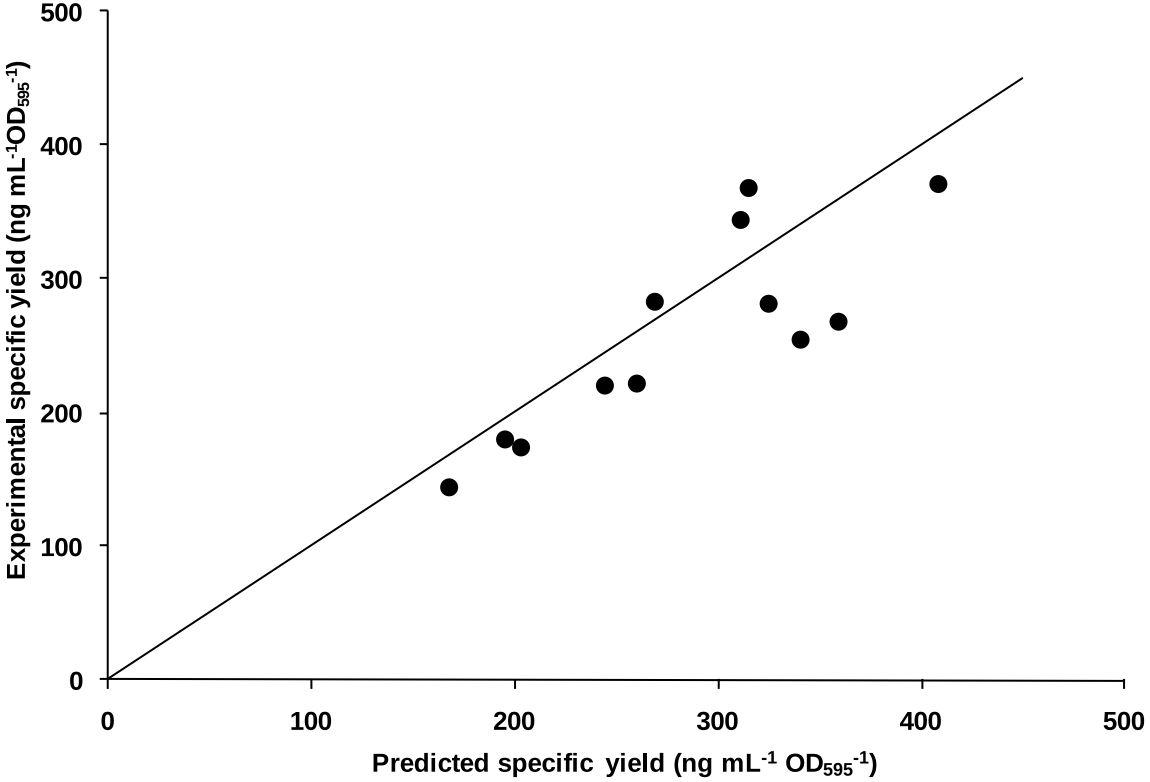
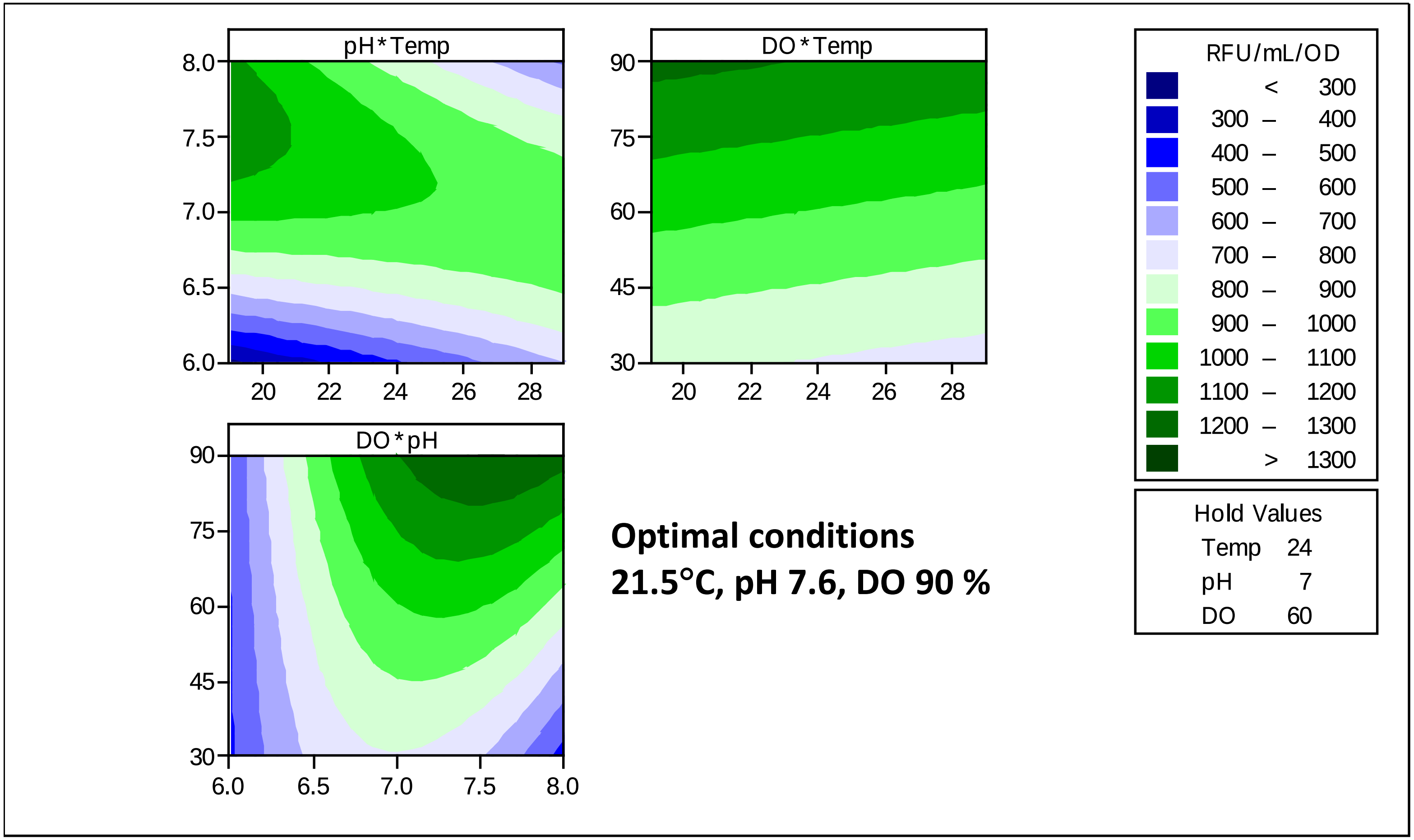

500
400
300
200
100
0
Experimental specific yield (ng mL-1OD595-1)
0
100
200
300
400
500
Predicted specific yield (ng mL-1 OD595-1)

Figure 3



Optimal conditions
21.5°C, pH 7.6, DO 90 %