Molecular Biotechnology (2019) 61:873–891
https://doi.org/10.1007/s12033-019-00218-x

REVIEW

Design of Experiments As a Tool for Optimization in Recombinant Protein Biotechnology: From Constructs to Crystals

Christos Papaneophytou[1]

Published online: 29 October 2019

Abstract

In this review, the basic concepts and applications of design of experiments (DoE) in recombinant protein biotechnology will be discussed. The production of recombinant proteins usually begins with the construction of an expression vector that is then introduced into a microbial host. The target protein is overexpressed in the host's cells and subsequently, it is isolated using a suitable purification method, its activity is assessed using a biological assay, while its crystallization is often required. Because each protein is unique and due to the complex interactions among the reagents in experiments, it is impossible that one set of reaction conditions would be optimal for all cases. Optimization of experimental conditions is usually carried out by the inefficient one-factor-at-a-time approach that does not take into account the combined effects of factors on a process. On the other hand, DoE approaches with a carefully selected small set of experiments, and therefore with a reduced cost and in a limited amount of time predict the effect of each factor and the effects of their interactions on a process. Importantly, several software packages are available that facilitate the choice of the DoE approach, design of the experiments, and analysis of the results.

Keywords Design of experiments · Recombinant proteins · Optimization · Response surface methodology

Introduction

Recombinant proteins are widely used in diverse fields in laboratory and industry, while many applications require sufficient amounts of high-quality proteins in terms of purity and activity [1]. In addition, the use of therapeutic recombinant proteins (biopharmaceuticals) has significantly increased since the introduction of recombinant human insulin in 1982 [2]. Nowadays, biopharmaceuticals are also being used for the treatment of a variety of diseases including cancer and metabolic disorders [3]. Interestingly, approximately 50% of all new medicines are classified as biopharmaceuticals [4, 5] and according to the report *Biopharmaceuticals Market by Type and Application: Global Opportunity Analysis and Industry Forecast, 2018–2025* (https://www.researchandmarkets.com/research/qh3vxr/global?w=4), in 2017 the total marker of biopharmaceuticals has reached US$186,470 million, while it is estimated that it will exceed $500,000 million by 2025. In addition, there are over 400 marketed recombinant pharmaceutical products while more than 1300 are undergoing clinical trials [4, 6]. Recombinant proteins are also a prerequisite and vital component of several drug design projects while crystallographic studies in these projects require hundreds of milligrams of purified protein samples [6, 7]. Therefore, the main goal of industrial and academic research laboratories is to produce high amounts of pure and functional proteins at a reasonable cost [8, 9].

Even though recombinant proteins can be expressed in both prokaryotic and eukaryotic systems [10, 11], *Escherichia coli* is the first choice for the production of non-glycosylated recombinant proteins at industrial scale due to its ability to easily replicate, low cost, simplicity, and Food and Drug Administration (FDA)-approved status for human applications [12, 13]. In addition, PCR-based cloning represents one of the most essential tools in recombinant protein production technology. Theoretically, using *E. coli* as an expression host and PCR as a cloning method, the production of a recombinant protein is a straightforward process, i.e., the gene of interest (GOI) is cloned into an expression

✉ Christos Papaneophytou
papaneophytou.c@unic.ac.cy

[1] Department of Life and Health Sciences, School of Sciences and Engineering, University of Nicosia, 46 Makedonitissas Ave., 2417 Nicosia, Cyprus

plasmid vector and, subsequently, it is inserted into an expression host. Following the induction of protein expression in the host's cells with the appropriate inducer (e.g., IPTG), the recombinant protein is purified and, subsequently, its biophysical properties and activity are determined [14]. Moreover, recombinant DNA synthesis techniques and technologies facilitate the synthesis of recombinant DNA. The construction of a recombinant plasmid requires only a thermocycler, primers, and DNA polymerases and, therefore, any basic molecular biology laboratory can synthesize several kilobase pairs of DNA in less than a week [15]. However, in practice several things may go wrong: low amounts of PCR product, insufficient ligation, formation of inclusion bodies, background proteins during purification, low activity or difficulties to obtain protein crystals [16]. In addition, protein stability and low purification yields are challenges that must be resolved when *E. coli* and other microbes are used as hosts to produce recombinant proteins [17]. Therefore, the optimization of reaction conditions may be needed in one or more of the following processes that are used in recombinant protein biotechnology: (i) amplification of the GOI using PCR; (ii) ligation of the GOI with an expression plasmid vector; (iii) expression of the target protein in high amounts and in a soluble form; (iv) isolation of the target protein in a pure and active form; (v) assessment of protein activity; (vi) identification of conditions to obtain a single protein crystal.

The majority of methods that are used in recombinant protein biotechnology are developed and subsequently optimized using the time-consuming and inefficient one-factor-at-the-time (OFAT) approach [18], which examines the effect of only one factor at a time. However, several biochemical processes are affected by the interactions of the experimental variables (factors). The best approach to examine the effect of multiple factors, as well as the effect of their interactions on a process, is the statistical design of experiments (or simply Design of Experiments—DoE) approach. DoE approaches have been successfully used in the development, optimization, and assessment of the robustness of many biochemical processes [19], including those that are employed in recombinant protein biotechnology [20]. However, when browsing through the literature using the keywords "optimization" and "recombinant proteins" from 2017 to 2019, I found that more 2000 papers have in their title the word "optimization" or "effect of," while less than 10% of them used statistical-based approaches to optimize a process. Interestingly, the process optimization in most cases was carried out using the OFAT approach.

The aim of this review is to discuss the potential applications of DoE approaches at every step of recombinant protein biotechnology, from the construction of an expression plasmid vector to crystallographic studies, and the recent progress in this growing field is discussed. Initially, the basic principles of DoE are presented. Then, the main factors affecting each step of recombinant protein production, purification, and characterization are discussed. This review focuses on the production of recombinant proteins specifically in *E. coli* that is one of the organisms of choice for the production of recombinant proteins including biopharmaceuticals. This will probably be the first review that extensively examines the use of DoE in all steps of recombinant protein biotechnology from the construction of a plasmid vector to crystallographic studies.

Optimization of a Process

An essential question, to begin with, is *What is the optimization of a process?* According to the English Oxford dictionary [21], optimization is *the action or process of making the best of something* or *the action or process of rendering optimal; the state or condition of being optimal*. Moreover, the online business dictionary (http://www.businessdictionary.com) defines optimization as *Finding an alternative with the most cost effective or highest achievable performance under the given constraints, by maximizing desired factors and minimizing undesired one*. Thus, according to these definitions, in order to optimize a process in recombinant protein biotechnology the experimenter should (i) examine the effect of multiple factors (variables) on the response (e.g., protein expression in a soluble form, purity, activity), in order to exclude the unimportant ones and subsequently (ii) find the optimum combination of the important factors that maximize the response.

Optimization Approaches: One-Factor-at-a-Time Approach Versus Statistically Designed Experiments

The majority of methods that are used in recombinant protein technology are developed and optimized using the traditional OFAT approach [18]. However, using the OFAT design, the experimenter gets information about one factor in each experimental trial [22] and, therefore, this approach is time-consuming especially when a large number of factors must be evaluated. The main disadvantage of the OFAT approach is that it does not examine the effects of the interactions among the experimental factors on a process, i.e., whether one factor influences the effect of another factor on a process (response). On the other hand, optimization studies can be conducted by varying several factors at the same time and examining both their effects and the effects of their interactions on a process (response) using statistical-based experimental approaches [23]. Overall, DoE is an organized approach that provides more reliable and useful information per experiment compared to the OFAT approach.

Theory and Steps of Design of Experiments

DoE approaches are employed in both the early and late stages of bioprocess development [24]. DoE uses statistical experimental methods and varies the factors that affect a process simultaneously over a specific set of experiments. The results are subsequently analyzed using a mathematical model, which gives significant information about the effect of each factor and the effect of the interactions between factors on the process (response) facilitating optimization of the process [25]. The main advantage of DoE approaches is that they use only a minimum number of experiments to examine simultaneously the effect of many parameters on a process, while biases are avoided [26]. Most importantly DoE is not only cheaper but it is also faster than OFAT because it provides optimized information content by using a small number of experiments [26, 27]. The theory and potential applications of DoE approaches are extensively described in many textbooks [28–30].

Several statistical software packages, such as Design-Expert (Stat-Ease Inc, MN, USA), JMP (SAS Institute Inc., Cary, NC, USA), Minitab (Minitab Inc, PA, USA), and ECHIP (ECHIP Inc, DE, USA), are available that lead the user during the design of experiments and analysis of the results. It should be noted that these software packages require only basic knowledge of statistics, design of experiments principles, elementary optimization methods, and regression modeling techniques. In addition, these software packages are able to develop mathematical models that demonstrate the relationships between factors and the response(s). In my laboratory, we routinely use Design-Expert software (https://www.statease.com/software/design-expert/) to design the experiments, analyze the data, and visualize the results. Design-Expert has been specifically developed for performing DoE and includes a variety of experimental designs including full factorial, fractional factorial, Plackett–Burman design, Taguchi Orthogonal Array, several types of response surface methodology, mixture designs, combined designs, etc., while it contains test matrices for testing up to 50 factors. The statistical significance of the test factors on the response is assessed using analysis of variance (ANOVA). The data are fitted on a mathematical model and graphical tools are employed to identify the impact of each factor on the response and reveal abnormalities in the data. It should be pointed out that other software packages, such as Minitab (http://www.minitab.com), JMP (https://www.jmp.com), and ECHIP (http://www.experimentationbydesign.com), contain the same/similar tools and features as Design-Expert and the selection of a software package for DoE purposes is a matter of personal choice.

Usually, a bioprocess is affected by a large number of factors and, therefore, DoE is carried out in two stages. During the first stage (screening experiments), the factors that have a statistically significant effect on the process are identified using a factorial design (discussed below), in order to reduce the number of factors to a manageable one [25]. Once the important parameters are identified, an optimization step is performed using the response surface methodology (RSM) in order to identify the optimum combination of factors that maximize the response (discussed below). The exclusion of insignificant factors during the first stage reduces the number of experiments and helps in the reduction of experimental effort required in the second step [31].

Before beginning the discussion on the applications of DoE in recombinant protein biotechnology, it is essential to give a short description of the terms that are widely used in DoE and they are summarized in Table 1.

The specific steps taken to optimize a process are described in the following paragraphs.

Stage 1: Screening Experiments—Identification of Significant Factors

A fundamental question that should be answered is *which is the most suitable experimental design for optimization studies?* The answer is that the choice of experimental design

Table 1 Vocabulary of DoE

Term	Description
Experimental design	The actual experimental plan composed of the different combinations of the variables (factors) to be tested
Variable (or factor)	An independent factor that may affect a process (response) and can take different values in different experiments
Categorical variable	A qualitative variable that is non-numerical. Categorical variables do not have a logical order, e.g., the growth medium is either Luria broth or Terrific broth
Continuous variable	A numerical variable that its values are numbers, e.g., the pH ofa buffer is either 7.0 or 7.5, etc.
Level	The numerical value of a continuous factor or the type of a categorical factor
Response	The response or depended variable is the quantity to be measured in an experiment (e.g., enzyme activity) and it depends on the independent variables
Run	An experiment composed of a specific combination of variables and levels to be tested
Full factorial design	A full factorial design examines all possible combinations of factors and levels
Fractional factorial design	A fractional factorial examines only a fraction, e.g., 1/2, 1/4 of the full factorial design

is dependent on the number and the type (categorical or continuous) of factors that have to be evaluated as well as on the previous knowledge about the protein of interest [18]. Another important question that should also be answered is *what factors should be tested*? In general, the choice of initial factors and the range of their values should be based on either literature examples with the same or similar proteins or previous experience expressing recombinant proteins [18].

Table 2 Experimental matrix of a 2_{III}^{7-4} fractional factorial for 7 variables studied at two levels

Run	Factor						
	A	*B*	*C*	*D*	*E*	*F*	*G*
1	−	+	−	−	+	−	+
2	+	+	+	+	+	+	+
3	+	−	−	−	−	+	+
4	−	−	+	+	−	−	+
5	+	−	+	−	+	−	−
6	−	−	−	+	+	+	−
7	+	+	−	+	−	−	−
8	−	+	+	−	−	+	−

Factors' levels "−" and "+" indicate the minimum and maximum possible value for each factor

Factorial Designs

Usually to identify the variables (factors) that significantly affect a process (response), a 2-level factorial approach is employed. In general, 2-level approaches are those in which all factors have only two values (a high- and a low-value) and these approaches are often referred to as "screening or preliminary experiments." A 2-level approach could be either full or fractional depending on whether all or a fraction of all possible combinations of factors are tested. A 2-level full factorial design (2^k, k is the number of factors and $k > 2$) examines all the possible combinations of factors. In a 2-level fractional factorial design (2_R^{k-p} fractional factorial, k is the number of factors, p indicates the size of the fraction of the 2^k full factorial, and R is the resolution of the method) only the $(1/2)^p$ fraction of the total number (2^k) of the combinations is examined (e.g., the 2^{k-1} and 2^{k-2} designs require only the half and the one quarter, respectively, of the experiments). The resolution (R) of the method illustrates how clearly the effects can be separated in a design (the higher the better) and resolution IV designs are usually employed.

Two-level approaches are useful for highlighting the critical factors for further detailed study using RSM (discussed further below). An example of a 2_{III}^{7-4} fractional approach that is widely being used in screening experiments and that requires only 8 experiments (a 2^7 full factorial design requires 128 experiments) is illustrated in Table 2. Using a fractional factorial approach is also beneficial when a large number (>4) of variables must be examined, and thus a 2^k full factorial demands a high number of experiments and, therefore, a high cost [32]. The fraction of experiments to be carried out is defined by the aforementioned software packages based on the number and type of variables that need to be examined. More details about fractional factorial designs could be found in Ref. [33].

A variety of algorithms, such as Plackett–Burman [34] and Taguchi orthogonal array [35], are also available which guide the selection of the fraction to be tested. Plackett–Burman design (PBD) is a small-sized two-level factorial experimental design that is widely used to identify large main effects. PBD identifies the important effectors from N number of variables in $N+1$ experiments (where N is a multiple of 4) without recourse to the interaction effects between and among the variables. Thus, PBD just screens the design space to detect large main effects [36]. As the number of factors increases, full and fractional factorial approaches become impractical and expensive since a large number of experiments must be carried out. To overcome this problem, Taguchi introduced the orthogonal array, a specially designed method, to study the entire space of variables using a smaller number of experiments. Taguchi proposed to use signal-to-noise (S/N) ratio as a measurable value instead of standard deviation because, as the mean decreases, the standard deviation also decreases and vice versa [37]. However, Taguchi designs are more complicated and should only be used by experimenters who are familiar with the complex aliasing issues behind the designs.

Identification of Significant Factors

As aforementioned, an essential step in the optimization of a process is the identification of the factors that have a statistically significant effect on the response. For example, the statistical significance of the seven variables of Table 2, i.e., *A*, *B*, *C*, *D*, *E*, *F*, and *G* on a response can be initially evaluated using a half-normal probability plot (Fig. 1a). A half-normal probability plot is a plot of the absolute value of the effect estimated with respect to their cumulative normal probabilities [38] and unimportant factors are those that have near-zero effects (i.e., they have a normal distribution centered near zero; factors *D*, *E*, *C* in the example of Fig. 1a), while important factors are those whose effects are significantly removed from zero, i.e., factors *A*, *G*, *B*, and *F* in the example of Fig. 1a.

The magnitude of the effect of each factor on the response is more clearly illustrated in a Pareto chart (Fig. 1b) which is a bar chart that rank-orders the effect of each factor by its magnitude. Pareto charts establish the *t*-value of the effect that is examined by two limit lines, i.e., the Bonferroni limit

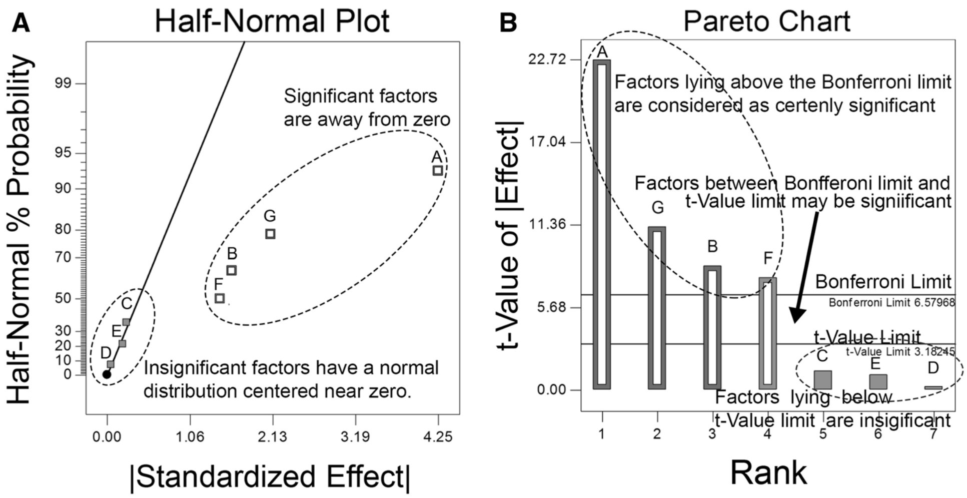

Fig. 1 An example of identification of the factors that have a statistically significant effect on a process (response). **a** Adjusted half-normal probability plot with significant factors selected: *A*, *G*, *B*, and *F*. The farther that a factor is from the diagonal line, the greater its influence on the response. **b** The Pareto chart identified factors *A*, *G*, *B*, and *F* lying above the Bonferroni limit and are designated as certainly significant coefficients

line (top line) and *t*-limit line (bottom line). Coefficients with a *t*-value of effect above the Bonferroni line are the significant factors; coefficients with a *t*-value of effect between Bonferroni line and *t*-limit line are termed as "coefficients likely to be significant," while coefficients with a *t*-value of effect below the *t*-limit line are statistically insignificant and should be excluded from the analysis [39]. The Pareto Chart of the example in Fig. 1b shows that factors *A*, *G*, *B*, and *F* lie above the Bonferroni limit and are considered as highly probable to be statistically significant.

Table 3 Central composite design of 4 independent variables that are examined at 5 levels (− 2, − 1, 0, + 1, + 2) for process optimization

Run	Coded values of factors			
	A	*B*	*C*	*D*
1	− 1	− 1	− 1	− 1
2	1	− 1	− 1	− 1
3	− 1	1	− 1	− 1
4	1	1	− 1	− 1
5	− 1	− 1	1	− 1
6	1	− 1	1	− 1
7	− 1	1	1	− 1
8	1	1	1	− 1
9	− 1	− 1	− 1	1
10	1	− 1	− 1	1
11	− 1	1	− 1	1
12	1	1	− 1	1
13	− 1	− 1	1	1
14	1	− 1	1	1
15	− 1	1	1	1
16	1	1	1	1
17	− 2	0	0	0
18	2	0	0	0
19	0	− 2	0	0
20	0	2	0	0
21	0	0	− 2	0
22	0	0	2	0
23	0	0	0	− 2
24	0	0	0	2
25	0	0	0	0
26	0	0	0	0
27	0	0	0	0
28	0	0	0	0
29	0	0	0	0
30	0	0	0	0

Stage 2: Optimization Experiments Using Response Surface Methodology

Following identification of factors that have a statistically significant effect on a process, RSM is usually applied to identify the best combination of these factors that maximize the response. The goals of RSM are to (1) develop a mathematical model that describes how the variables (factors) and the interactions between variables affect the response and (ii) determine the values of all variables that optimize the response [40].

Response surfaces are typical second-order polynomial models and the central composite design (CCD) is usually used. A CCD is composed of [1] a fractional factorial (or full factorial) design; [2] an additional design (often a star design in which experimental points are at a distance from its center); and [3] a central point. CCD is an efficient design that is ideal for sequential experimentation and allows a reasonable amount of information to test the "lack of fit" using a small number of design points. Besides CCD, there are many experimental designs for RSM such as Box–Behnken design (BBD), and small composite design (SCD). In a CCD, all factors are studied at five levels ($-\alpha$, − 1, 0, 1, $+\alpha$) and for two, three, and four variables the value of alpha is 1.41 ($\sqrt{2}$), 1.68 ($\sqrt{3}$), and 2 ($\sqrt{4}$), respectively. The theory and the mathematical part of DoE approaches including RSM are extensively discussed in many books (see [41] and references cited therein).

Table 3 illustrates the experimental setup of a four-factor-five level CCD (four factors are examined at 5 levels; −2, −1, 0, 1, 2). It should be noted that a limitation of RSM is that only continuous factors can be examined, while if categorical factors are added the design will be duplicated for every combination of the categorical factor levels.

The experimental data obtained from the design (e.g., Table 3) are subsequently fitted on a second-order polynomial model (Eq. 1).

$$Y = \beta_o + \sum \beta_i x_i + \sum \beta_{ij} x_i x_j + \sum \beta_{ii} x_{ii}^2, \tag{1}$$

where Y is the measured response variable, β_o is a constant, β_i, β_{ij}, and β_{ii} are the regression coefficients of the model, and x_i and x_j represent the independent variables in coded values.

The second-order polynomial coefficients are estimated using a software package, e.g., Design-Expert. An example of a second-order equation (mathematical model) obtained during the optimization of a process is illustrated below (Eq. 2). In this example, the effects of four variables, namely A, B, C, and D, as well as the effects of their interactions, on the response (Y) were examined.

$$\begin{aligned} Y[\text{response}] = {} & +12.41 - 0.86A - 1.39B - 1.61C \\ & + 1.04D + 0.076AB - 0.19AC - 0.14AD \\ & + 0.13BC - 0.71BD - 0.69CD - 1.09A^2 \\ & - 0.18B^2 - 1.90C^2 - 0.79D^2. \end{aligned} \tag{2}$$

In Eq. (2), plus (+) and minus (−) symbols show whether a model term has a positive or negative effect on the response.

Graphical Representation of the Interactive Effects of Variables on the Response

After the generation of the mathematical model, it is possible to predict the response for any possible combination of the factors that are tested within the experimental region (domain) even for those experiments that have not been actually carried out. Thus, the response at any point in the experimental domain can be predicted and, therefore, a graphical representation can be easily obtained. Usually, the fitted second-order equations (e.g., Eq. 2) are presented as a two-dimensional representation (contour plots, Fig. 2a) or as a three-dimensional plot (response surface plots, Fig. 2b).

These plots are the graphical representation of the relationship among three variables, i.e., two independent variables (while the others are kept at their zero points) and the response. A 3D response surface plot (Fig. 2b) is obtained by plotting two independent variables on the x- and y-axes (the other are kept at their center points), while the response is shown by a smooth 3D surface plot. The plots indicate the direction in which the original design must be displayed in order to achieve the optimal conditions. By looking at the plots of Fig. 2, it is easy to understand that the best reaction time changes according to the reaction temperature (and vice versa), while a reaction time above 8 h and a temperature above 26 °C have a negative effect on the response.

Validation of the Mathematical Model

The evaluation of the quality of the fitted mathematical model, as well as the effect of the factors that are examined and the effect of their interactions on the response(s), is usually carried out with the analysis of variance (ANOVA). Briefly, ANOVA uses F-tests (F = variation between sample means/variation within the samples) to compare the mean values between the factors that are examined and determines whether any of those means are statistically significantly different from each other (for more information above ANOVA

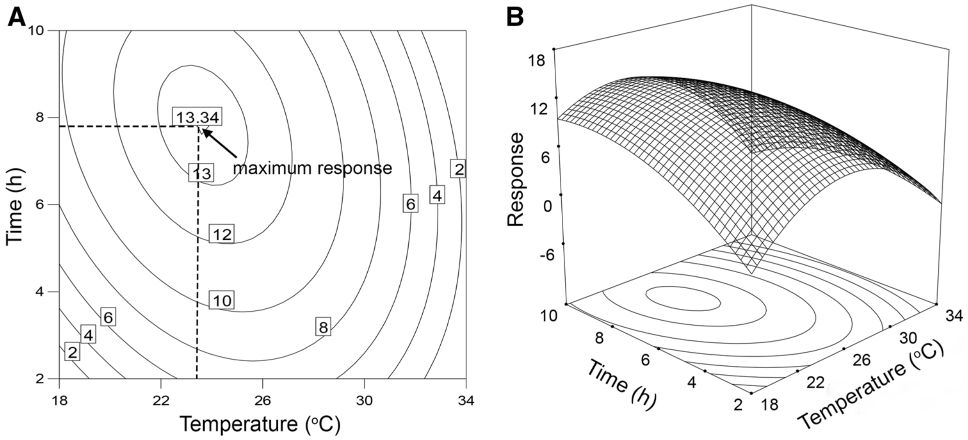

Fig. 2 An example of a contour plot (**a**) and a response surface plot (**b**) showing the effect of reaction temperature and time on the response, adapted from [129] (modified). The figure illustrates that the one factor influences the effect of the other factor on the response. In this example, the maximum response (~ 13.3 relative units) was obtained at a reaction temperature between 23 and 26 °C when the reaction time was set at ~ 8 h

see Ref. [42]). To determine whether each main effect is statistically significant, the p value for each term is compared to the significance level to assess the null hypothesis, while a significance level of 0.05 is usually used. Overall, the p value of each factor (term) should be < 0.05 to be significant while in several cases the insignificant factors ($p > 0.05$) are excluded from the model. Initially, the quality of the model is evaluated by the F value and p value of the model. In general, a high model F value indicates that more of the variance can be explained by the model (the higher the better), whereas the p value of the model should be strongly significant (< 0.05). The lack of fit is also used to determine whether the model fits the data well. If the model does not fit the data well, then the lack of fit will be significant ($p > 0.05$). The insight of mathematical model significance is also assessed from two determination coefficients (R-squared or R^2), namely "adjusted" R^2 and "predicted" R^2. The adjusted R^2 indicates the amount of variation around the mean explained by the model, while the predicted R^2 indicates how well a response value is predicted by the model. In general, the higher the R^2, the better the model fits the data while an $R^2 > 0.6$ is required. Finally, the quality of the model is evaluated by the Adequate Precision that is the signal-to-noise ratio and a ratio greater than 4 is required (for more information about ANOVA see https://www.statease.com/docs/v11/navigation/anova-rsm.html).

Overall, the adequacy of the mathematical model is evaluated using the "lack-of-fit" test and the "Adj R-squared," as well as using:

(i) The normal (%) probability plot of the "Studentized" residuals that shows whether the data are normally distributed or not. Figure 3a shows an example of the evaluation of a mathematical model using the normal (%) probability plot and, as can be seen, the errors are normally distributed.

(ii) The "predicted *vs* actual" plot which shows whether the actual values (experimental data) are in agreement with the predicted values. In other words, predicted vs. actual plots detect how well the model fits the data. For a perfect fit, all the points would be on a straight line [30, 43]. Figure 3b shows an example where experimental values and predicted values are in good agreement.

Incomplete Fractional Factorial Designs: An Alternative Approach

In several cases, a high number of variables have to be tested, and therefore a high number of experiments is required. Thus, performing a full or fractional factorial as well as RSM is impractical especially when a large number of categorical factors must be examined. To this end, incomplete factorial (IF) designs were developed to test only a part of a large full factorial design when a large number of combinations of factors must be examined [44, 45]. Thus, any design that is developed by removing experimental conditions from a full factorial design is an IF. Even though according to this definition, a fractional factorial design can be designated as an IF design, a main dereference between the two designs is that the term "fractional factorial" refers to incomplete factorials that share the balance property of the corresponding full factorial approach, while an IF involves fewer experimental combinations which are not balanced. Thus, all fractional factorials are IF, but not all IF designs are fractional factorials [46]. IF that are not fractional factorials involve fewer experimental conditions, and they provide an economical and effective way to assess the effect of different possible factors and identify those most likely to be essential, which is beneficial especially when experimental costs are high [46]. Another major advantage of IF designs is that both categorical and continuous factors can be simultaneously examined while each factor can be examined at more two levels [18, 47]. Most importantly, a freeware online software called SAmBA (http://www.igs.cnrs-mrs.fr/samba) has been developed for the design of IF

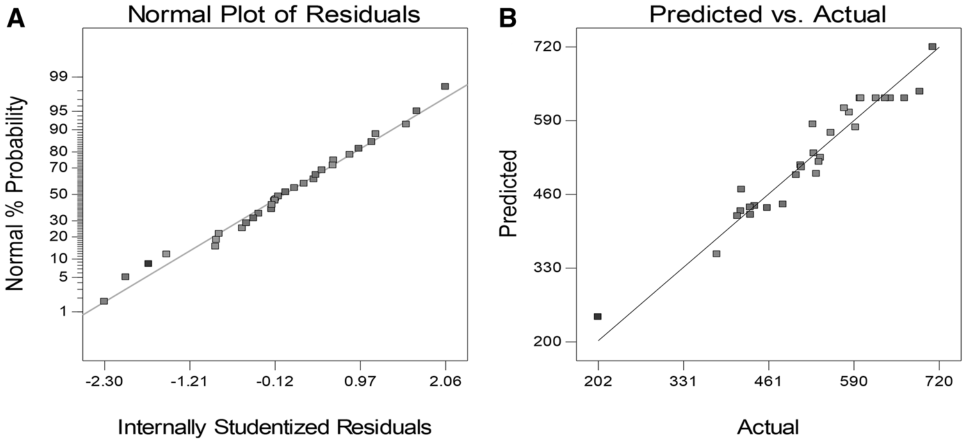

Fig. 3 An example of diagnostic plots that are used for the evaluation of the accuracy of mathematical models in RSM adapted from [129]. **a** Normal (%) probability plot of the "Studentized" residuals for the model. **b** Predicted (by the model) values of the response versus actual values (experimental)

approaches. An example of an IF design is provided and discussed in section "Construction of Recombinant Plasmids": *Application of design of experiments in recombinant protein biotechnology*; paragraph *Ligation of the insert with the vector.*

Application of Design of Experiments in Recombinant Protein Biotechnology

Table 4 summarizes the DoE approaches that are commonly employed in the optimization of processes that are used in recombinant protein biotechnology. It should be noted that for screening experiments, two-level factorial designs are very common and more economical compared to the 3- or higher level factorial designs, and due to do their simpler structure are more interpretable in practice [30]. IF designs are probably a better choice when the effect of 2 or 3 categorical factors must be examined; however, previous experience with the protein of interest is required. Overall, a statistical design should be carefully selected based on (1) the availability in resources and equipment and (2) the existing information about the protein of interest.

The potential applications of DoE designs in recombinant protein biotechnology are extensively discussed in the following paragraphs.

Construction of Recombinant Plasmids

The first step in producing recombinant DNA is to identify and isolate the target DNA and vector DNA. Despite several techniques have been developed for generating recombinant DNA sequences including TA cloning [48], ligation-independent cloning [49, 50], recombinase-dependent cloning [51–53], and PCR-mediated cloning [54–56], PCR-based cloning is routinely being used in molecular cloning [57–59]. In addition, PCR primers that introduce restriction enzyme sites on the insert's sequence are usually employed. This review will be focused on the traditional PCR-based cloning and the basic steps that are followed during this technique are described below:

Table 4 DoE designs that are commonly used in recombinant protein biotechnology

Method	Choose when	Advantages	Disadvantages
Full factorial	– Limited information about the process is available – 4 or fewer factors must be examined	– Both categorical and continuous variables can be simultaneously tested – Results from the whole set of experiments are utilized	– Can be only used for screening experiments – Can only exclude the unimportant factors
Fractional factorial	– 5 or more factors must be tested – The number of factors that are examined must be reduced to the significant ones	– Both categorical and continuous variables can be simultaneously tested – Information can be obtained by testing only a fraction of all possible combinations of factors (full factorial)	
Plackett–Burman	The gaps found in fractional factorial must be reduced	Plackett–Burman method screens the design space to detect a large main effect	Should only be used when is known that there are no interactions present in the design
Taguchi orthogonal	Multiple factors at multiple levels must be examined	Taguchi orthogonal is a highly fractional orthogonal design allowing to examine a selected subset of combinations of multiple factors at multiple levels with the fewest number of experiments	– Very complicated – Specific knowledge about the complex aliasing issues behind the designs is required
RSM	The optimum values of variables that significantly affect a process should be identified to maximize the response (process)	– The only "real" optimization process – Can be used to fine-tune the optimum conditions	– Only continuous values can be tested – A high number of runs are needed – Dedicated statistical packages are required
Incomplete fractional	– Multiple factors must be tested – A quick outcome is required	– Can easily be set up using a freeware package (SAmBA) – Advance knowledge in statistics is not necessary – Factors can be examined in more than two levels	Previous information about the process is required

PCR Amplification of the Gene of Interest

Because of the potential interactions among the components of PCR, usually optimization of PCR conditions is carried out by changing one or more factors that are known to affect the primer–DNA interaction and primer extension. The isolation of pure, intact, and high-quality DNA is essential for molecular biology studies [60]. Although PCR cloning is routinely being used in molecular cloning [57], there are not general guidelines on setting up a PCR reaction. It is well known that the concentration of Mg^{2+} ions, the pH of the reaction buffer, and the annealing temperature influence the amplification of a DNA fragment in PCR. In addition, the interactions of some reagents (factors) influence the amplification of the GOI (response). For example, dNTPs chelate Mg^{2+} ions, and therefore an increase in the concentration of dNTPs will reduce the concentration of Mg^{2+} ions in the reaction mixture [61]. To this end, DoE approaches have been successfully employed in many cases to optimize the reaction conditions of PCR reactions. Boleda et al. [62] used a two-step approach to optimize PCR of DNA blood spots. In the first step of optimization and by using a 2^5 fractional factorial approach, the DNA concentration and Mg^{2+} were identified as the factors that significantly affect the response (DNA amplification). Subsequently, an RSM was employed to identify the optimum concentration of the two factors. In another study, the optimum concentrations of dNTPs, Mg^{2+}, and primers were identified by using a full factorial approach and by a three-dimensional Simplex [63]. DoE approaches have also been employed for the optimization of reaction conditions of quantitative PCR [64], real-time PCR [65, 66], and digital PCR [67] assays. Therefore, it is suggested that the conditions of any PCR assay, including cloning PCR, could be optimized by statistically designed experiments. Following PCR, the amplified GOI is analyzed on an agarose gel and recovered using a commercially available gel extraction kit.

Digestion with Restriction Enzymes

Subsequently both the PCR-amplified GOI (insert) and vector are digested with the same restriction enzymes in order to create complementary cohesive sticky ends. Restriction digestions are carried out according to the manufacturer's instructions and the optimization of at this point is limited to the duration of reaction and/or the amount of enzyme [68].

Ligation of the Insert with the Vector

This step is usually catalyzed by a DNA ligase and optimizing ligation efficiency is essential to cloning experiments [69]. Even though some ligation mathematical models have previously been reported [70–73], they are either too specific or too general to be used as a universal tool to improve ligation efficiency. The factors that significantly affect ligation are the molar insert-to-vector ratio, the temperature, and duration of ligation, and total DNA concentration. However, optimization of ligation reactions is usually carried out using the OFAT approach. It has been suggested that a more generic but easily altered strategy is needed to improve DNA ligation [69]. In my laboratory, we have developed an IF design composed of 16 combinations of 3 insert-to-vector ratios, 3 ligation temperatures, 3 durations of ligation, and 3 total DNA amounts as illustrated in Table 5, to identify the best combination of these factors for the ligation of a plasmid vector with a PCR-amplified gene. The ligation efficiency is monitored using the Lig-PCR method (i.e., the ligation reactions are monitored using PCR and primers that are present in the majority of vectors) as previously described in [74]. This straightforward approach examines all factors affecting ligation efficiency and provides in less than 2 days a positive answer to the ligation query. In the case of a negative result (no ligation), a significant amount of time can be saved.

Transformation of Competent Cells

The ligation product is subsequently inserted into a cloning strain (e.g., DH5α), to ensure the stable amplification of recombinant DNA [75] using a standard transformation protocol (see [76] and references cited therein). In general, the introduction of foreign DNA into bacteria using either electroporation or chemical transformation is affected by

Table 5 Incomplete factorial approach for the ligation of a plasmid vector with the gene of interest (insert)

Run	Insert: vector	Temperature (°C)	Duration (h)	DNA (ng)
1	7:1	4	2	50
2	3:1	16	2	100
3	3:1	4	16	50
4	5:1	20	16	100
5	3:1	16	16	75
6	5:1	16	0.5	50
7	7:1	16	16	50
8	3:1	20	2	75
9	5:1	4	16	75
10	3:1	4	0.5	100
11	3:1	20	0.5	50
12	7:1	20	16	100
13	5:1	20	2	50
14	7:1	4	0.5	75
15	7:1	20	0.5	75
16	5:1	16	2	100

many factors including electrical parameters [77] (only for electroporation), washing buffer, cell wall weakening agents [78], the cell density (optical density at 600 nm – OD_{600nm}) [79], duration of heat-shock, medium composition, and by the presence of some co-factors (e.g., DMSO) [76]. However, optimization of transformation conditions is carried out using OFAT approaches. Even though the optimization of transformation conditions for cloning experiments using DoE has not been reported, fractional factorial approaches have been successfully employed to identify the factors that significantly affect the transformation efficiency of bacteria for other purposes (e.g., drug development). For example, Yildirim et al. [80] evaluated the effect of five factors (cell density, voltage, resistance, plasmid DNA concentration, and Mg^{2+} concentration) on the transformation efficiency of *Acinetobacter baumannii* using a three-level fractional factorial approach and the transformation efficiency was increased by four times. Thus, DoE approaches could be probably used as a tool to maximize the transformation efficiency of bacteria during cloning experiments.

Expression of Recombinant Proteins in *E. coli*

E. coli expression systems provide an inexpensive, robust, and flexible platform appropriate for the production of recombinant proteins at both industrial and laboratory scales. Optimum expression conditions for each construct must be identified for the maximal production of soluble protein. A recombinant protein is expressed in the microenvironment of *E. coli* which may differ from that of its native source in terms of folding mechanism, pH, co-factors, ionic strength, and redox potential. These factors affect both protein stability and solubility, while in several cases recombinant proteins are expressed in the form of inclusion bodies [81–83].

In general, soluble expression of recombinant proteins is affected not only by the expression host strain and expression vector, but also by expression conditions including induction temperature and time, the concentration of inducer, and the composition of the culture medium [18, 84]. The yield and solubility of recombinant proteins can be therefore increased by optimizing these factors and several expression conditions are usually tested [85]. One of the standard procedures, when setting out to express a recombinant protein, is to test different culture conditions and media because this is easy, cheap, and has been proven to have an impact on protein solubility levels [20]. Among the factors, affecting protein expression and/or solubility in *E. coli*, the induction temperature and time are probably the most important ones, because these variables, in most cases, interact. In bacterial, a slower and longer induction promotes the expression of several proteins in a soluble form, and this approach requires a low temperature [86]. The magnitude of induction is also an important factor that affects both the expression and solubility of recombinant proteins. Insufficient concentration of inducer (e.g., IPTG) may result in low protein expression, whereas the addition of a high concentration of inducer can result in reduced cell growth and/or recombinant protein yield [87]. The soluble expression of recombinant protein is also affected by the expression host [88] and thus multiple *E. coli* strains that facilitate the expression of membrane proteins, proteins with rare codons, proteins with disulfide bonds, proteins that are otherwise toxic to the cell, etc., are commercially available (see [18] and references cited therein). Cell density before induction has also an impact on the soluble expression levels of recombinant proteins. Despite induction is usually performed at early mid-log phase, some proteins require induction at late-log phase [89] or stationary phase [90]. To facilitate recombinant protein expression in a soluble form and to accelerate the characterization of protein structure and function, a variety of affinity and solubility tags have also been employed [17], including small peptides [e.g., hexahistidine (6 × His-tag), FLAG] and large peptides/proteins [e.g., Glutathione-*S*-transferase (GST), maltose-binding protein (MBP)] [91]. The main factors affecting recombinant protein soluble expression have been extensively reviewed elsewhere [14, 18].

Even though several criteria must be considered when expressing recombinant proteins in *E. coli*, optimization experiments are usually carried out using the traditional OFAT approach [92]. On the other hand, DoE approaches have successfully been employed for the identification of factors that have a statistically significant effect on the soluble expression of recombinant proteins. Usually, optimization of soluble expression goes from screening experiments to ascertain what variables have an effect (i.e., a full or fractional factorial design is initially employed looking at cell line, and possibly media and additives, along with OD_{600nm} at induction, temperature, time and IPTG concentration, as variables to determine which factors are significant) and then to RSM to determine the optimum values (levels) of the most significant variables [16, 18].

DoE designs and especially RSM have been successfully employed to optimize the culture conditions and/or culture medium composition for a variety of recombinant proteins such as recombinant scFv antibody [93], human Interferon-beta [94], DT386-BR2 [95], receptor activator of nuclear factor (NF)-κB ligand (RANKL) [96], superoxide dismutase [97], pneumolysin from *Streptococcus pneumoniae* [32], pneumococcal surface adhesin A [98, 99], *Taq*I endonuclease [100], pyruvate oxidase [101], sea anemone neurotoxin [102], heparinase I [103], lipase KV1 [104], and glutaryl-7-aminocephalosporanic acid acylase [105]. The potential applications of DoE in the soluble expression of recombinant proteins have been extensively reviewed elsewhere [2, 116]. In my laboratory, we have successfully employed

RSM for the optimization of soluble expression of several recombinant proteins including tumor necrosis factor-alpha (TNF-α) [106], RANKL [107], heme oxygenase-1 (HO-1) [108], and human rhinovirus-3C protease (HRV3CP) [109]. In each case, preliminary experiments were performed in order to identify the best expression host for each target protein as well as to identify the factors that have a statistically significant effect on both the yield and soluble expression of the protein of interest. Subsequently, the culture conditions that maximize the soluble expression of each recombinant protein (i.e., TNF-α, RANKL, HO-1, and HRV3CP) were identified using RSM. The use of DoE approaches in recombinant protein expression with examples of media and culture conditions optimization has been recently reviewed in Ref. [40].

Even though RSM has been successfully employed to maximize the soluble expression of several recombinant proteins, this design has several limitations, especially when the number of the test variables is high [98]. Moreover, RSM is a fine-tuning technique, i.e., it is used to identify the optimum combination of the independent variables that maximize the response. Therefore, as mentioned above, preliminary experiments are essential to identify the factors that significantly affect the response, while only continuous factors can be examined [11]. To this end, in my laboratory we have recently developed an IF approach that we called IF-STTI (Incomplete Factorial-Strain/Time/Temperature/Inducer) [109] to identify the best combination of the four most important factors (i.e., expression host, temperature and duration of induction, and IPTG concentration) affecting the soluble expression of recombinant proteins in *E. coli* [18] in a single experiment. In detail, IF-STTI is composed of 24 different combinations of three expression strains, three post-induction temperatures, four induction times, and three IPTG concentrations. The design was validated with three GST-tagged recombinant proteins, i.e., TNF-α, RANKL, and HRV3CP. The results obtained from this design were subsequently compared with those obtained using RSM and interestingly the soluble expression levels of the three tested proteins were close to those obtained by RSM. Most importantly, we demonstrated that the IF-STTI design is an accurate and straightforward method as it provides, in only 24 experiments, the same information regarding the interactions of variables on the soluble expression of recombinant proteins, as would do a full factorial design (108 experiments) or RSM (30 experiments). Another advantage of the IF-SSTI compare to the 2^k factorial designs is that all variables may be examined in more than two levels.

Two incomplete factorial designs called "InFFact" [110] that is made of 12 combinations of 4 *E. coli* strains, 3 media, and 3 expression temperatures (full factorial 36 combinations) and "Fusion-InFFact" [111] that is composed of 24 combinations of 4 expression strains, 3 media, 3 expression temperatures, and 5 N-terminal tags (full factorial 180 combinations) have also been reported. Both methods have been successfully employed to determine the conditions that maximize the soluble expression of several recombinant proteins in *E. coli.*

Purification of Recombinant Proteins

The ultimate objective of protein purification for therapeutic or analytical applications is to achieve both high yield and purity [112]. Thus, it has been suggested that the protein of interest should be produced as a fusion to an affinity tag because tags facilitate the purification of any protein, in one step, without any prior knowledge about the protein of interest and do not affect its biochemical or biological activity [113]. To this end, a variety of affinity purification methods have been developed (for a review on affinity tags see Ref. [17]). In general, affinity purification of recombinant proteins depends on two factors: (i) the ability of a protein to bind to an affinity matrix which is composed of a substrate attached to a solid support, e.g., Sepharose, agarose, or resin and (ii) the ability to recover the protein from the affinity matrix. Elution is usually carried out either using a soluble substrate that competes for binding sites (competitive elution) or, in some cases, by cleavage between the protein and affinity matrix with a specific protease [114, 115].

A typical protein purification includes several operating parameters that can affect the yield and purity of the protein of interest. To achieve both high purity and yield of the target protein, it is essential to examine the relationship between these two goals and the purification factors and to optimize purification conditions accordingly. The final yield and purity of a protein are affected by multiple factors including the composition of the sample to be loaded, chromatography medium, purification method, binding, wash, and elution conditions [116]. Moreover, the final purity and recovery of the protein can be optimized by controlling the operating conditions such as flow rate, ionic strength gradient, sample load, physical properties of the adsorbent matrix, column dimensions, and the ratio of the protein to the column size [112]. In addition, it is important to take into account the effect of buffer composition in protein stability and purification yield while it is beneficial to decide what the ideal final buffer would be. Therefore, optimization of purification processes can be time-consuming [116] and despite the obvious advantages of DoE approaches over OFAT approach, optimization of purification conditions of recombinant proteins is usually carried out using the latter method. For example, during optimization of immobilized metal affinity chromatography (IMAC) protocols using the OFAT approach, a different volume of metal-chelated resin and concentration of imidazole in the washing and elution buffer are tested in each experiment. However, OFAT approach does not take into account the effect of interactions among purification factors, on purity and yield of the target protein. Because every protein is different, the optimum purification conditions must be identified for each protein [9].

Nevertheless, the significant advantages of DoE over OFAT approach in recombinant protein purification are highlighted in recent publications. A two-step DoE approach has been used for the affinity purification of recombinant 6 × His human erythropoietin (hEPO). During the first step of this approach, it has been demonstrated that the ratio of loaded protein to resin significantly affects both protein purity and yield. Subsequently, in the second step, the optimal purification conditions (i.e., the amount of resin and wash/elution conditions) of 6 × His- hEPO were identified using RSM. This two-step DoE-optimized purification approach resulted in a 45% yield and a 90% purity of recombinant 6 × His-tagged hEPO [117].

A two-step DoE approach has also been employed for the optimization of purification conditions of recombinant single-chain variable fragment against type 1 insulin-like growth factor receptor (IGF-1R) using the capto-L affinity chromatography medium [118]. In an initial step, the effect of seven variables including the pH value of the buffer, and the concentration of the following additives: NaCl, urea, arginine, trehalose, polyethylene glycol (PEG), and dextran on both IGF-IR aggregation and recovery were evaluated using a 2-level fractional factorial approach. Trehalose concentration and pH were identified as the main factors and, subsequently, the purification conditions were optimized using a central composite circumscribed design. Overall, a total yield of 77% and a 98.5% purity of the final product were achieved.

In another study, Amadeo et al. [119] employed RSM in order to identify the best combination of the critical factors, i.e., sample pH, the ratio of loaded protein to resin, and residence time, that affect the purity and yield of recombinant human erythropoietin using Blue Sepharose as an affinity matrix. An 88% recovery and a 71.5% purity of the protein of interested were achieved after optimization of purification conditions. RSM has been successfully employed for the optimization of purification conditions (PEG and salt concentration, pH value and/or concentration of the purification buffer), with aqueous two-phase system of several enzymes including glucose dehydrogenase (GDH) from *Bacillus subtilis* [120] and d-galactose dehydrogenase (GalDH) from *Pseudomonas fluorescens* AK92 glucose [120, 121].

In my laboratory, we have recently reported an IF approach composed of 16 different combinations of three resin volumes, three glycerol and four DTT concentrations in purification buffers, and three incubation times of cell lysate with resin, in order to determine the optimal purification conditions for GST-tagged HRV3CP [109]. The 16 combinations of these factors were selected out of the 108 combinations (3 × 3 × 4 × 3) of the full factorial design using the SAmbA freeware. The results revealed that the recovery of the protease was increased by 15% (compared to the protease recovery before optimization), while the proteins that were previously co-purified (before optimization of purification conditions) with the target protein (GST-HRV3CP) were eliminated. Our method was validated further using another two GST-tagged recombinant proteins, i.e., GST-TNF-α and GST-RANKL and the yields of two proteins were increased by 11% and 10%, respectively [109].

Based on the examples described above, DoE approaches could overcome the limitations of the traditional OFAT approach for the optimization of purification conditions of any recombinant protein. As purification of any protein is affected by multiple factors (variables), the OFAT approach often fails to identify the optimal purification conditions because this approach examines only a limited part of the experiment space and most importantly it does not examine the combined effects of all the factors involved. The specific steps that are followed during the optimization of purification conditions of recombinant proteins using DoE, as well as specific guidelines for execution and analysis of experiments, are described in Ref. [116].

Assessment of Protein Activity

Following purification of the target protein, in several cases, its activity should be assessed. Moreover, enzymes are important drug targets and in areas such as drug development, clinical diagnosis, and biotechnology research the determination of kinetic parameters of enzymes is essential. To identify potential therapeutics that inhibit the function of proteins/enzymes that have been implicated in the pathogenesis and development of diseases is essential to design, develop, and validate biological assays for high-throughput screening (HTS). Developing sensitive biological assays suitable for HTS requires identification of factors affecting assay performance and robustness and the correct design of a biological assay is essential to derive the correct information and to collect data suitable for analysis and modeling [122]. Thus, the development of reliable biological assays for the identification and validation of potential therapeutics is essential in the various stages of drug development [123–125].

Depending on the assay format and the nature of the protein that is studied in each case, different variables, i.e., assay conditions, should be examined. For example, typical factors that affect the enzyme activity include the composition, concentration, and pH of the reaction buffer, type and concentration of enzyme, ionic strength, as well as the type and concentration of substrate, reaction conditions (assay incubation time and temperature), and appropriate assay technology. Likewise, in non-enzymatic protein assays, e.g., ligand-binding assays, several factors should also be examined including the dilution of the protein, assay incubation time and temperature, viscosity, and ionic strength, while in several cases buffer additives should be included in order to facilitate protein stability or to improve ligand solubility [126, 127]. Several factors can be optimized during the development, optimization, and assessment of the robustness of ELISA-based ligand-binding assays including conditions associated with samples and calibrators and conditions associated with the detection of the analyte, such as substrate development time [125, 128].

A major concern during the development and optimization of a biological assay is the selection of factors to be tested as well as their ranges to be used. In general, assay optimization determines how a range of experimental conditions can affect assay performance and is an essential step to find the value that each variable should have to produce the best possible response. Usually, if there is literature available, the experimenter begins with the reaction conditions and factors published previously to be needed for the activity of the same or a similar protein/enzyme. The reaction conditions and the concentration range of the selected factors to be examined should then be selected carefully and should be large enough to cause a clear alteration in the measured response, but not so large that the process will 'fall off a cliff' and produce unusable data [129].

Despite methodologies for assays have been extensively reviewed, however, because each protein/enzyme is different, a further modification of procedure is often required, (i) to adjust the assay conditions to the special features of the protein/enzyme of interest or (ii) in order to develop an assay for a newly discovered protein/enzyme [130]. It has been suggested that a DoE study must be carried out before the validation of an assay for early identification of factors that significantly affect assay performance and robustness [128]. However, a survey carried out by HTStec in 2009 [131] revealed that optimization of assays is carried out using the traditional OFAT approach because most researchers believe that DoE designs are very difficult to be employed. However, using the traditional OFAT approach usually takes at least 4 months to develop an assay, and therefore assay development can become a bottleneck in drug discovery projects [122]. To this end, we have recently reported the steps any researcher could follow to develop, optimize, and define the design space for determination of enzyme activity using a two-step DoE methodology, including guidelines for (i) identification of factors that significantly affect the activity of the enzyme to be studied, and (ii) execution of experiments and data analysis, using HRV3CP in a 96-well plate format assay, as an example [129]. Briefly, a 2^{8-4} fractional factorial design was initially employed to assess the effect of seven factors: one categorical factor, i.e., buffer composition (Tris–HCl and HEPES) and seven continuous factors including reaction pH, temperature, and time as wells as the concentration of NaCl, DTT, EDTA, and glycerol on protease activity. The results of the screening DoE were used to eliminate non-significant factors using the half-normal probability and Pareto charts as described in section *Theory and steps for design of experiments* and particularly in paragraph *Identification of significant factors*. Our analysis revealed that only the pH of the buffer, the incubation time, and the concentrations of both DTT and glycerol produced significant effects on the activity of HRV3CP. Subsequently, we employed RSM to determine the optimal combination of the four statistically significant variables that produce the maximum HRV3CP activity and a 1.5-fold increase in the activity of the protease was achieved [129].

It should be noted that the quality of an HTS assay is usually assessed using the Z-prime (Z') statistical test that takes into account both the signal window and assay viability [132]. The Z'-factor is calculated based on the following equation (Eq. 3):

$$Z' = 1 - 3 \times \frac{\sigma_p + \sigma_n}{\mu_p - \mu_n}, \tag{3}$$

where σ_p and σ_c are the stand deviations of the positive and negative controls, respectively, and μ_p and μ_n are their respective average values.

In general, the Z' test is the most important statistical test to assess the quality of an assay. A Z' equal to 1 is ideal, though an assay can never have a Z' of 1.00000, while Z' can never be greater than 1.0. When an assay has a Z' between 0.5 and 1.0, it means that it is excellent, while when Z' is between 0 and 0.5 it means that the assay is marginal. A Z' factor less than 0 means that the assay is not suitable for HTS screening [132]. In the aforementioned example, an increase of Z' factor from 0.78 (before optimization) to 0.92 (after optimization) was achieved and thus the assay is suitable for HTS of HRV3CP inhibitors [129].

Even though DoE designs have significant advantages over the OFAT approach in assay optimization and in assessing the robustness of a method, there are only a limited number of publications in the literature that utilize these designs in assay development, optimization, and validation. Nevertheless, DoE approaches have been successfully used for the optimization of several assay conditions and for the determination of kinetic constant of various enzymes including glucose oxidase [133], the enzymes involved in the synthesis of precorrin-2 [134], and hydrolases [135], for the development and validation of a cell-based bioassay for the detection of anti-drug neutralizing antibodies in human serum [136], for the optimization of various immunoassays [137–140], as well as to evaluate the robustness of a ligand-binding assay [128] and other assays [141]. A detailed tutorial that describes the use of DoE approaches in non-enzymatic assay optimization has been previously reported [122].

Protein Crystallography

In drug discovery projects, crystallization of the target protein(s) that is (are) implicated in the pathogenesis of a disease with a potential therapeutic is an essential step in order to identify the interactions between the two molecules. These interactions are translated into a picture where a drug molecule binds to the target protein(s) and acts as an inhibitor, an agonist or a modulator [142].

A major issue in protein crystallization is that a high number of parameters must be tested to identify the conditions that yield a single large crystal for the collection of X-ray data [143, 144]. Biochemical, chemical, and physical factors such as genetic modifications of the protein, the type of precipitants, type of salts, concentrations, pH value of the buffer, and the temperature of the environment may have an impact on the crystallization process. Because each protein has a unique primary structure it is quite challenging to determine the crystallization conditions that can yield a crystal for a protein a priori [145], and therefore adapted methods are employed to enable the growth of the appropriate crystals [146].

To this end, the conditions for protein crystallization have been traditionally identified using two DoE designs, namely incomplete factorial experiments (IFE) [145, 147] and sparse matrix sampling (SMS) [144, 148]. The incomplete factorial approach was introduced in protein crystallography in 1979 [44] as a powerful tool for identifying the factors and conditions that need to be varied to obtain crystals. The goals of this approach are to (i) identify the important factors that influence the crystallization of the target protein and (ii)

Table 6 Applications of DoE in the main processes that are used in recombinant protein biotechnology

Step/method	Examples of factors that should be optimized	Proposed optimization strategy
PRC amplification	1. Mg^{2+} concentration 2. Annealing temperature 3. Template DNA (ng) 4. Concentration of primers	– Incomplete factorial approach that examines all factors at more than 2 levels or – Small composite design (RSM; 15 runs)
Ligation	1. Insert-to-vector ratio 2. Temperature 3. Duration 4. Total DNA amount	– Incomplete factorial approach that examines all factors at more than 2 levels in 16 experimental runs (see also Table 5) or – Small composite design (RSM; 15 runs)
Expression in *E. coli*	1. Vector type 2. Expression host 3. Affinity Tag 4. Temperature of induction 5. Duration of induction 6. Culture medium 7. IPTG concentration 8. Cell density before induction	– *Information about the protein is not available:* Identification of the factors that significantly affect soluble expression (or purity and recovery) of the target protein using a fractional factorial design and then optimization of the most important factors using RSM – *Information about the protein is available*: Incomplete factorial approach that examines the main factors affecting the soluble expression (or purity and recovery) of the protein in more than 2 levels
Purification	1. Column size 2. Buffer composition 3. Buffer pH 4. Protein-to-resin ratio 5. Buffer additives 6. Flow rate 7. Ionic strength	
Functional assay	1. Buffer composition 2. Buffer pH 3. Buffer additives 4. Ionic strength 5. Co-factors 6. Incubation time 7. Reaction temperature 8. Substrate concentration	Identification of factors that significantly affect the activity of the protein/enzyme of interest using a fractional factorial design and subsequently optimization of most important factors using RSM
Crystallography	1. Type of precipitant 2. Type of salt 3. Concentration of salt 4. Buffer type 5. Buffer pH 6. Temperature	– Incomplete factorial experiments (IFE) or – Sparse matrix sampling (SMS) *Note* several crystallization screening conditions are commercially available; however, crystallization conditions may be further optimized

reduce the total number of crystallization conditions compared to full factorial designs [44]. The IFE is an essential tool especially in the case there is not enough protein to test a high number of crystallization conditions while at the same time it provides sufficient information regarding the important factors in a small number of experiments [149]. The SMS method was initially reported in 1991 by Jancarik and Kim [144], and interestingly their original screen, plus a wide range of variations, has been commercialized [150]. The sparse matrix approach uses three categories of major variables: pH and buffer materials, additives, and precipitating agents. These ranges of the buffer, pH, additives, and precipitant conditions are empirically derived based on past experience to have resulted in protein crystallization.

Following screening crystallization experiments, a set of optimization methods are usually applied to improve the quality of the crystals. Further details regarding the IFE and SMS methods and optimization techniques can be found in the literature and are beyond the scope of this review. Importantly, crystallization techniques that are based on DoE are continually optimized. For example, Dinć et al. [151] reported the "Associative Experimental Design (AED)" approach for the optimization of crystallization conditions for proteins. The main advantage of this approach is that following analysis of preliminary experiments, the AED generates candidate cocktails, i.e., novel conditions, leading to crystals (see also [151] and references cited therein).

Conclusion

Recombinant proteins are essential tools in biomedical, pharmaceutical, and biological industries and the production of soluble and functional recombinant proteins is the ultimate goal in protein biotechnology. Several recombinant proteins are being used as drugs (biopharmaceutical) and their demand in the pharmaceutical industry will be increased in the next years because biopharmaceuticals have been successfully used for the prevention, detection, and treatment of diseases. To meet the growing demand for recombinant proteins, it is essential to produce them in high amounts and in a pure and active form. Due to the unique properties of each protein and the complex interactions among the reagents in the experiments, it is almost impossible that one set of reaction conditions would be optimal for all cases. Optimization of several processes that are used in recombinant protein biotechnology is usually carried out using the traditional OFAT approach that is not only time-consuming, but also incapable of identifying the true optimal conditions as it does not examine the interactions between the factors affecting the desired response(s). On the contrary, DoE designs are gaining success for optimization of all processes of recombinant protein biotechnology including construction of recombinant plasmid vector, protein production, expression, purification, assessment of activity, and crystallography as summarized in Table 6, because they require fewer experiments and therefore less time, for the amount of information obtained, while in the case of negative results a significant amount of time can be saved. Most importantly, DoE designs can provide models that may assist to (i) identify the factors that have a statistically significant effect on a process and (ii) study interactions between different variables and predict the maximized response in all processes of recombinant protein technology.

References

1. Palomares, L. A., Estrada-Mondaca, S., & Ramirez, O. T. (2004). Production of recombinant proteins: Challenges and solutions. *Methods in Molecular Biology, 267,* 15–52.
2. Leader, B., Baca, Q. J., & Golan, D. E. (2008). Protein therapeutics: a summary and pharmacological classification. *Nature Reviews Drug Discovery, 7,* 21–39.
3. Kesik-Brodacka, M. (2018). Progress in biopharmaceutical development. *Biotechnology and Applied Biochemistry, 65,* 306–322.
4. Jozala, A. F., Geraldes, D. C., Tundisi, L. L., Feitosa, V. A., Breyer, C. A., Cardoso, S. L., et al. (2016). Biopharmaceuticals from microorganisms: From production to purification. *Brazilian Journal of Microbiology, 47*(Suppl 1), 51–63.
5. Basu, A., Li, X., & Leong, S. S. (2011). Refolding of proteins from inclusion bodies: Rational design and recipes. *Applied Microbiology and Biotechnology, 92,* 241–251.
6. Sanchez-Garcia, L., Martín, L., Mangues, R., Ferrer-Miralles, N., Vázquez, E., & Villaverde, A. (2016). Recombinant pharmaceuticals from microbial cells: A 2015 update. *Microbial Cell Factories, 15,* 33.
7. Kim, Y., Bigelow, L., Borovilos, M., Dementieva, I., Duggan, E., Eschenfeldt, W., et al. (2008). High-throughput protein purification for x-ray crystallography and NMR. *Advances in Protein Chemistry and Structural Biology, 75,* 85–105.
8. Tralau-Stewart, C. J., Wyatt, C. A., Kleyn, D. E., & Ayad, A. (2009). Drug discovery: New models for industry–academic partnerships. *Drug Discovery Today, 14,* 95–101.
9. Structural Genomics, C., Architecture et Fonction des Macromolécules, B., Berkeley Structural Genomics, C., China Structural Genomics, C., Integrated Center for, S., Function, I., Israel Structural Proteomics, C., Joint Center for Structural, G., Midwest Center for Structural, G., New York Structural Genomi, X. R. C. f. S. G., Northeast Structural Genomics, C., Oxford Protein Production, F., Protein Sample Production Facility, M. D. C. f. M. M., Initiative, R. S. G. P. and Complexes, S. (2008). Protein production and purification. *Nature Methods, 5,* 135–146.
10. Khan, K. H. (2013). Gene expression in mammalian cells and its applications. *Advanced Pharmaceutical Bulletin, 3,* 257–263.
11. Walsh, G. (2014). Biopharmaceutical benchmarks 2014. *Nature Biotechnology, 32,* 992–1000.
12. Marisch, K., Bayer, K., Cserjan-Puschmann, M., Luchner, M., & Striedner, G. (2013). Evaluation of three industrial *Escherichia coli* strains in fed-batch cultivations during high-level SOD protein production. *Microbial Cell Factories, 12,* 58.

13. Long, X., Gou, Y., Luo, M., Zhang, S., Zhang, H., Bai, L., et al. (2015). Soluble expression, purification, and characterization of active recombinant human tissue plasminogen activator by auto-induction in *E. coli*. *BMC Biotechnology, 15,* 13.
14. Rosano, G. L., & Ceccarelli, E. A. (2014). Recombinant protein expression in *Escherichia coli*: Advances and challenges. *Frontiers in Microbiology, 5,* 172.
15. Hughes, R. A., Miklos, A. E., & Ellington, A. D. (2011). Gene synthesis: Methods and applications. *Methods in Enzymology, 498,* 277–309.
16. Jia, B., & Jeon, C. O. (2016). High-throughput recombinant protein expression in *Escherichia coli*: current status and future perspectives. *Open Biology, 6,* 160196.
17. Young, C. L., Britton, Z. T., & Robinson, A. S. (2012). Recombinant protein expression and purification: A comprehensive review of affinity tags and microbial applications. *Biotechnology Journal, 7,* 620–634.
18. Papaneophytou, C. P., & Kontopidis, G. (2014). Statistical approaches to maximize recombinant protein expression in *Escherichia coli*: A general review. *Protein Expression and Purification, 94,* 22–32.
19. Lee, K. M., & Gilmore, D. F. (2006). Statistical experimental design for bioprocess modeling and optimization analysis: Repeated-measures method for dynamic biotechnology process. *Applied Biochemistry and Biotechnology, 135,* 101–116.
20. Chambers, S. P., & Swalley, S. E. (2009). Designing experiments for high-throughput protein expression. In S. A. Doyle (Ed.), *High throughput protein expression and purification: Methods and protocols* (pp. 19–29). Totowa, NJ: Humana Press.
21. Oxford English Dictionary. (2008). Oxford: Oxford University Press.
22. Jeff Wu, C.-F., & Hamada, M. (2000). *Experiments: Planning, analysis, and parameter design optimization*. Hoboken: Wiley.
23. Rodrigues, M., & Francisco Iemma, A. (2014). *Experimental design and process optimization*. Boca Raton, FL: CRC Press.
24. Bora, N., Bawa, Z., Bill, R. M., & Wilks, M. D. (2012). The implementation of a design of experiments strategy to increase recombinant protein yields in yeast (review). *Methods in Molecular Biology, 866,* 115–127.
25. Mandenius, C. F., & Brundin, A. (2008). Bioprocess optimization using design-of-experiments methodology. *Biotechnology Progress, 24,* 1191–1203.
26. Weissman, S. A., & Anderson, N. G. (2015). Design of experiments (DoE) and process optimization. A review of recent publications. *Organic Process Research & Development, 19,* 1605–1633.
27. Swalley, S. E., Fulghum, J. R., & Chambers, S. P. (2006). Screening factors effecting a response in soluble protein expression: Formalized approach using design of experiments. *Analytical Biochemistry, 351,* 122–127.
28. Hicks, C. R., & Turner, K. V. (1999). *Fundamental concepts in the design of experiments*. New York: Oxford University Press.
29. Box, G. E., Hunter, J. S., & Hunter, W. G. (2005). *Statistics for experimenters*. Hoboken: Wiley.
30. Montgomery, D. C. (2008). *Design and analysis of experiments*. New York: Wiley.
31. Vijesh, K., Akriti, B., & Rathore, A. S. (2014). Design of experiments applications in bioprocessing: Concepts and approach. *Biotechnology Progress, 30,* 86–99.
32. Marini, G., Luchese, M. D., Argondizzo, A. P. C., de Góes, A. C. M. A., Galler, R., Alves, T. L. M., et al. (2014). Experimental design approach in recombinant protein expression: determining medium composition and induction conditions for expression of pneumolysin from *Streptococcus pneumoniae* in *Escherichia coli* and preliminary purification process. *BMC Biotechnology, 14,* 1.
33. Box, G. E. P., & Hunter, J. S. (1961). The 2^{k-p} fractional factorial designs. *Technometrics, 3,* 311–351.
34. Plackett, R. L., & Burman, J. P. (1946). The design of optimum multifactorial experiments. *Biometrika, 33,* 305–325.
35. Phadke, M. S. (1989). *Quality engineering using robust design*. Upper Saddle River, NJ: Prentice Hall.
36. Cavazzuti, M. (2013). *Optimization methods: From theory to design*. Berlin: Springer.
37. Karna, S., & Sahai, R. (2012). An overview on Taguchi method. *IJEMS, 1,* 1–7.
38. Daniel, C. (1959). Use of half-normal plots in interpreting factorial two-level experiments. *Technometrics, 1,* 311–341.
39. Shah, M., & Pathak, K. (2010). Development and statistical optimization of solid lipid nanoparticles of simvastatin by using 2^3 full-factorial design. *An Official Journal of the American Association of Pharmaceutical Scientists, 11,* 489–496.
40. Uhoraningoga, A., Kinsella, G. K., Henehan, G. T., & Ryan, B. J. (2018). The goldilocks approach: A Review of employing design of experiments in prokaryotic recombinant protein production. *Bioengineering (Basel), 5,* E89.
41. Leardi, R. (2009). Experimental design in chemistry: A tutorial. *Analytica Chimica Acta, 652,* 161–172.
42. Bezerra, M. A., Santelli, R. E., Oliveira, E. P., Villar, L. S., & Escaleira, L. A. (2008). Response surface methodology (RSM) as a tool for optimization in analytical chemistry. *Talanta, 76,* 965–977.
43. Luzier, W. D. (1992). Materials derived from biomass biodegradable materials. *Proceedings of the National academy of Sciences of the United States of America, 89,* 839–842.
44. Carter, C. W., Jr., & Carter, C. W. (1979). Protein crystallization using incomplete factorial experiments. *Journal of Biological Chemistry, 254,* 12219–12223.
45. Carter, C. W. (1990). Efficient factorial designs and the analysis of macromolecular crystal growth conditions. *Methods, 1,* 12–24.
46. Collins, L. M., Dziak, J. J., & Li, R. (2009). Design of experiments with multiple independent variables: A resource management perspective on complete and reduced factorial designs. *Psychological Methods, 14,* 202–224.
47. Byar, D. P., Herzberg, A. M., & Tan, W. Y. (1993). Incomplete factorial designs for randomized clinical trials. *Statistics in Medicine, 12*(17), 1629–1641.
48. Marchuk, D., Drumm, M., Saulino, A., & Collins, F. S. (1991). Construction of T-vectors, a rapid and general system for direct cloning of unmodified PCR products. *Nucleic Acids Research, 19,* 1154.
49. Weeks, S. D., Drinker, M., & Loll, P. J. (2007). Ligation independent cloning vectors for expression of SUMO fusions. *Protein Expression and Purification, 53,* 40–50.
50. Rashtchian, A., Thornton, C. G., & Heidecker, G. (1992). A novel method for site-directed mutagenesis using PCR and uracil DNA glycosylase. *PCR Methods and Applications, 2,* 124–130.
51. Walhout, A. J. M., Temple, G. F., Brasch, M. A., Hartley, J. L., Lorson, M. A., van den Heuvel, S., et al. (2000). GATEWAY recombinational cloning: Application to the cloning of large numbers of open reading frames or ORFeomes. *Methods in Enzymology, 328,* 575–592.
52. Cheo, D. L., Titus, S. A., Byrd, D. R., Hartley, J. L., Temple, G. F., & Brasch, M. A. (2004). Concerted assembly and cloning of multiple DNA segments using in vitro site-specific recombination: Functional analysis of multi-segment expression clones. *Genome Research, 14,* 2111–2120.
53. Court, D. L., Sawitzke, J. A., & Thomason, L. C. (2002). Genetic engineering using homologous recombination. *Annual Review of Genetics, 36,* 361–388.

54. Zuo, P., & Rabie, B. M. (2010). One-step DNA fragment assembly and circularization for gene cloning. *Current Issues in Molecular Biology, 12,* 11–16.
55. Shuldiner, A. R., Scott, L. A., & Roth, J. (1990). PCR-induced (ligase-free) subcloning: A rapid reliable method to subclone polymerase chain reaction (PCR) products. *Nucleic Acids Research, 18,* 1920–1920.
56. Shuldiner, A. R., Tanner, K., Scott, L. A., Moore, C. A., & Roth, J. (1991). Ligase-free subcloning: A versatile method to subclone polymerase chain reaction (PCR) products in a single day. *Analytical Biochemistry, 194,* 9–15.
57. Mitchell, D. B., Ruggli, N., & Tratschin, J. D. (1992). An improved method for cloning PCR fragments. *PCR Methods and Applications, 2,* 81–82.
58. Mead, D. A., Pey, N. K., Herrnstadt, C., Marcil, R. A., & Smith, L. M. (1991). A universal method for the direct cloning of PCR amplified nucleic acid. *Biotechnology, 9,* 657–663.
59. Guo, B., & Bi, Y. (2002). Cloning PCR products. In B.-Y. Chen & H. W. Janes (Eds.), *PCR cloning protocols* (pp. 111–119). Totowa, NJ: Humana Press.
60. Tan, S. C., & Yiap, B. C. (2009). DNA, RNA, and protein extraction: The past and the present. *Journal of Biomedicine and Biotechnology, 2009,* 10.
61. Roux, K. H. (2009). *Optimization and troubleshooting in PCR*. New York: Cold Spring Harbor Protocols.
62. Boleda, M. D., Briones, P., Farrés, J., Tyfield, L., & Pi, R. (1996). Experimental design: A useful tool for PCR optimization. *BioTechniques, 21,* 134–140.
63. Benčina, M. (2002). Optimisation of multiple PCR using a combination of full factorial design and three-dimensional simplex optimisation method. *Biotechnology Letters, 24,* 489–495.
64. Besseris, G. J. (2014). A fast-and-robust profiler for improving polymerase chain reaction diagnostics. *PLoS ONE, 9,* e108973.
65. Wadle, S., Lehnert, M., Rubenwolf, S., Zengerle, R., & von Stetten, F. (2016). Real-time PCR probe optimization using design of experiments approach. *Biomolecular Detection and Quantification, 7,* 1–8.
66. Hui, K., & Feng, Z. P. (2013). Efficient experimental design and analysis of real-time PCR assays. *Channels, 7,* 160–170.
67. Dorazio, R. M., & Hunter, M. E. (2015). Statistical models for the analysis and design of digital polymerase chain reaction (dPCR) experiments. *Analytical Chemistry, 87,* 10886–10893.
68. Smith, D. R. (1993). Restriction endonuclease digestion of DNA. *Methods in Molecular Biology, 18,* 427–431.
69. Ng, D. T. W., & Sarkar, C. A. (2012). Model-guided ligation strategy for optimal assembly of DNA libraries. *Protein Engineering, Design & Selection, 25,* 669–678.
70. Dugaiczyk, A., Boyer, H. W., & Goodman, H. M. (1975). Ligation of EcoRI endonuclease-generated DNA fragments into linear and circular structures. *Journal of Molecular Biology, 96,* 171–184.
71. Legerski, R. J., & Robberson, D. L. (1985). Analysis and optimization of recombinant DNA joining reactions. *Journal of Molecular Biology, 181,* 297–312.
72. Revie, D., Smith, D. W., & Yee, T. W. (1988). Kinetic analysis for optimization of DNA ligation reactions. *Nucleic Acids Research, 16,* 10301–10321.
73. Dardel, F. (1988). Computer simulation of DNA ligation: Determination of initial DNA concentrations favouring the formation of recombinant molecules. *Nucleic Acids Research, 16,* 1767–1778.
74. Chandra, P. K., & Wikel, S. K. (2005). Analyzing ligation mixtures using a PCR based method. *Biological Procedures Online, 7,* 93–100.
75. Thomason, L. C., Sawitzke, J. A., Li, X., Costantino, N., & Court, D. L. (2014). Recombineering: Genetic engineering in bacteria using homologous recombination. *Current Protocols in Molecular Biology, 106,* 39.
76. Chan, W.-T., Verma, Chandra S., Lane, David P., & Gan, Samuel K.-E. (2013). A comparison and optimization of methods and factors affecting the transformation of *Escherichia coli*. *Bioscience Reports, 33,* e00086.
77. Dower, W. J., Miller, J. F., & Ragsdale, C. W. (1988). High efficiency transformation of *E. coli* by high voltage electroporation. *Nucleic Acids Research, 16,* 6127–6145.
78. Aune, T. E. V., & Aachmann, F. L. (2010). Methodologies to increase the transformation efficiencies and the range of bacteria that can be transformed. *Applied Microbiology and Biotechnology, 85,* 1301–1313.
79. Tang, X., Nakata, Y., Li, H. O., Zhang, M., Gao, H., Fujita, A., et al. (1994). The optimization of preparations of competent cells for transformation of *E. coli*. *Nucleic Acids Research, 22,* 2857–2858.
80. Yildirim, S., Thompson, M. G., Jacobs, A. C., Zurawski, D. V., & Kirkup, B. C. (2016). Evaluation of parameters for high efficiency transformation of *Acinetobacter baumannii*. *Scientific Reports, 6,* 22110.
81. Hartley, D. L., & Kane, J. F. (1988). Properties of inclusion bodies from recombinant *Escherichia coli*. *Biochemical Society Transactions, 16,* 101–102.
82. Carrió, M. M., & Villaverde, A. (2002). Construction and deconstruction of bacterial inclusion bodies. *Journal of Biotechnology, 96,* 3–12.
83. Ferrer-Miralles, N., Domingo-Espín, J., Corchero, J. L., Vázquez, E., & Villaverde, A. (2009). Microbial factories for recombinant pharmaceuticals. *Microbial Cell Factories, 8,* 17.
84. Makino, T., Skretas, G., & Georgiou, G. (2011). Strain engineering for improved expression of recombinant proteins in bacteria. *Microbial Cell Factories, 10,* 32.
85. Sørensen, H. P., & Mortensen, K. K. (2005). Soluble expression of recombinant proteins in the cytoplasm of *Escherichia coli*. *Microbial Cell Factories, 4,* 1.
86. Schein, C. H., & Noteborn, M. H. M. (1988). Formation of soluble recombinant proteins in *Escherichia coli i*s favored by lower growth temperature. *Biotechnology, 6,* 291–294.
87. Ramirez, O. T., Zamora, R., Espinosa, G., Merino, E., Bolivar, F., & Quintero, R. (1994). Kinetic-study of penicillin acylase production by recombinant *Escherichia Coli* in batch cultures. *Process Biochemistry, 29,* 197–206.
88. Shaw, M. K., & Ingraham, J. L. (1967). Synthesis of macromolecules by *Escherichia coli* near the minimal temperature for growth. *Journal of Bacteriology, 94,* 157–164.
89. Galloway, C. A., Sowden, M. P., & Smith, H. C. (2003). Increasing the yield of soluble recombinant protein expressed in *E. col*i by induction during late log phase. *BioTechniques, 34,* 524–530.
90. Ou, J. X., Wang, L., Ding, X. L., Du, J. Y., Zhang, Y., Chen, H. P., et al. (2004). Stationary phase protein overproduction is a fundamental capability of *Escherichia coli*. *Biochemical and Biophysical Research, 314,* 174–180.
91. Waugh, D. S. (2005). Making the most of affinity tags. *Trends in Biotechnology, 23,* 316–320.
92. Czitrom, V. (1999). One-factor-at-a-time versus designed experiments. *American Statistician, 53,* 126–131.
93. Kasli, I. M., Thomas, O. R. T., & Overton, T. W. (2019). Use of a design of experiments approach to optimise production of a recombinant antibody fragment in the periplasm of *Escherichia coli*: selection of signal peptide and optimal growth conditions. *AMB Express, 9,* 5.
94. Morowvat, M. H., Babaeipour, V., Rajabi Memari, H., & Vahidi, H. (2015). Optimization of fermentation conditions for recombinant human interferon beta production by *Escherichia coli*

using the response surface methodology. *Jundishapur Journal of Microbiology, 8,* e16236.

95. Shafiee, F., Rabbani, M., & Jahanian-Najafabadi, A. (2017). Optimization of the expression of DT386-BR2 fusion protein in *Escherichia coli* using response surface methodology. *Advanced Biomedical Research, 6,* 22.
96. Maharjan, S., Singh, B., Bok, J.-D., Kim, J.-I., Jiang, T., Cho, C.-S., et al. (2014). Exploring codon optimization and response surface methodology to express biologically active transmembrane RANKL in *E. coli*. *PLoS ONE, 9,* 96259.
97. Wang, Y., Wang, Q., Wang, Y., Han, H., Hou, Y., & Shi, Y. (2017). Statistical optimization for the production of recombinant cold-adapted superoxide dismutase in *E. coli* using response surface methodology. *Bioengineered, 8,* 693–699.
98. Larentis, A. L., Argondizzo, A. P. C., Esteves, Jessouron, Galler, R., & Medeiros, M. A. (2011). Cloning and optimization of induction conditions for mature PsaA (pneumococcal surface adhesin A) expression in *Escherichia coli* and recombinant protein stability during long-term storage. *Protein Expression and Purification, 78,* 38–47.
99. Larentis, A. L., Nicolau, J. F. M. Q., Argondizzo, A. P. C., Galler, R., Rodrigues, M. I., & Medeiros, M. A. (2012). Optimization of medium formulation and seed conditions for expression of mature PsaA (pneumococcal surface adhesin A) in *Escherichia coli* using a sequential experimental design strategy and response surface methodology. *Journal of Industrial Microbiology and Biotechnology, 39,* 897–908.
100. Nikerel, İ. E., Toksoy, E., Kırdar, B., & Yıldırım, R. (2005). Optimizing medium composition for TaqI endonuclease production by recombinant *Escherichia coli* cells using response surface methodology. *Process Biochemistry, 40,* 1633–1639.
101. Zhao, J., Wang, Y., Chu, J., Zhang, S., Zhuang, Y., & Yuan, Z. (2008). Statistical optimization of medium for the production of pyruvate oxidase by the recombinant *Escherichia coli*. *Journal of Industrial Microbiology and Biotechnology, 35,* 257–262.
102. Wang, Y.-H., Jing, C.-F., Yang, B., Mainda, G., Dong, M.-L., & Xu, A.-L. (2005). Production of a new sea anemone neurotoxin by recombinant *Escherichia coli*: Optimization of culture conditions using response surface methodology. *Process Biochemistry, 40,* 2721–2728.
103. Chen, Y., Xing, X.-H., Ye, F., Kuang, Y., & Luo, M. (2007). Production of MBP–HepA fusion protein in recombinant *Escherichia coli* by optimization of culture medium. *Biochemical Engineering Journal, 34,* 114–121.
104. Batumalaie, K., Khalili, E., Mahat, N. A., Huyop, F. Z., & Wahab, R. A. (2018). A statistical approach for optimizing the protocol for overexpressing lipase KV1 in *Escherichia coli*: Purification and characterization. *Biotechnology and Biotechnological Equipment, 32,* 69–87.
105. Volontè, F., Marinelli, F., Gastaldo, L., Sacchi, S., Pilone, M. S., Pollegioni, L., et al. (2008). Optimization of glutaryl-7-aminocephalosporanic acid acylase expression in *E. coli*. *Protein Expression and Purification, 61,* 131–137.
106. Papaneophytou, C. P., & Kontopidis, G. A. (2012). Optimization of TNF-alpha overexpression in *Escherichia coli* using response surface methodology: Purification of the protein and oligomerization studies. *Protein Expression and Purification, 86,* 35–44.
107. Papaneophytou, C. P., Rinotas, V., Douni, E., & Kontopidis, G. (2013). A statistical approach for optimization of RANKL overexpression in *Escherichia coli*: Purification and characterization of the protein. *Protein Expression and Purification, 90,* 9–19.
108. Papaneophytou, C., & Kontopidis, G. (2016). A comparison of statistical approaches used for the optimization of soluble protein expression in *Escherichia coli*. *Protein Expression and Purification, 120,* 126–137.
109. Antoniou, G., Papakyriacou, I., & Papaneophytou, C. (2017). Optimization of soluble expression and purification of recombinant human rhinovirus type-14 3C protease using statistically designed experiments: Isolation and characterization of the enzyme. *Molecular Biotechnology, 59,* 407–424.
110. Abergel, C., Coutard, B., Byrne, D., Chenivesse, S., Claude, J. B., Deregnaucourt, C., et al. (2003). Structural genomics of highly conserved microbial genes of unknown function in search of new antibacterial targets. *Journal of Structural and Functional Genomics, 4,* 141–157.
111. Noguere, C., Larsson, A. M., Guyot, J. C., & Bignon, C. (2012). Fractional factorial approach combining 4 *Escherichia coli* strains, 3 culture media, 3 expression temperatures and 5 N-terminal fusion tags for screening the soluble expression of recombinant proteins. *Protein Expression and Purification, 84,* 204–213.
112. Asenjo, J. A., & Andrews, B. A. (2009). Protein purification using chromatography: selection of type, modelling and optimization of operating conditions. *Journal of Molecular Recognition, 22,* 65–76.
113. Uhlen, M., Forsberg, G., Moks, T., Hartmanis, M., & Nilsson, B. (1992). Fusion proteins in biotechnology. *Current Opinion in Biotechnology, 3,* 363–369.
114. Jenny, R. J., Mann, K. G., & Lundblad, R. L. (2003). A critical review of the methods for cleavage of fusion proteins with thrombin and factor Xa. *Protein Expression and Purification, 31,* 1–11.
115. Zheng, N., Perez Jde, J., Zhang, Z., Dominguez, E., Garcia, J. A., & Xie, Q. (2008). Specific and efficient cleavage of fusion proteins by recombinant plum pox virus NIa protease. *Protein Expression and Purification, 57,* 153–162.
116. Amadeo, I., Mauro, L., Ortí, E., & Forno, G. (2014). Establishment of a design space for biopharmaceutical purification processes using DoE. In N. E. Labrou (Ed.), *Protein downstream processing: Design, development and application of high and low-resolution methods* (pp. 11–27). Totowa, NJ: Humana Press.
117. Shin, H. S., & Cha, H. J. (2003). Statistical optimization for immobilized metal affinity purification of secreted human erythropoietin from *Drosophila* S2 cells. *Protein Expression and Purification, 28,* 331–339.
118. Song, Y. H., Sun, X. W., Jiang, B., Liu, J. E., & Su, X. H. (2015). Purification optimization for a recombinant single-chain variable fragment against type 1 insulin-like growth factor receptor (IGF-1R) by using design of experiment (DoE). *Protein Expression and Purification, 116,* 98–104.
119. Amadeo, I., Mauro, L. V., Orti, E., & Forno, G. (2011). Determination of robustness and optimal work conditions for a purification process of a therapeutic recombinant protein using response surface methodology. *Biotechnology Progress, 27,* 724–732.
120. Shahbaz Mohammadi, H., Mostafavi, S. S., Soleimani, S., Bozorgian, S., Pooraskari, M., & Kianmehr, A. (2015). Response surface methodology to optimize partition and purification of two recombinant oxidoreductase enzymes, glucose dehydrogenase and D-galactose dehydrogenase in aqueous two-phase systems. *Protein Expression and Purification, 108,* 41–47.
121. Azar, S. R., Naiebi, R., Homami, A., Akbari, Z., Kianmehr, A., Mahdizadehdehosta, R., et al. (2015). Expression and response surface optimization of the recovery and purification of recombinant D-galactose dehydrogenase from *Pseudomonas fluorescens*. *Indian Journal of Biochemistry & Biophysics, 52,* 68–74.
122. Altekar, M., Homon, C. A., Kashem, M. A., Mason, S. W., Nelson, R. M., Patnaude, L. A., et al. (2007). Assay optimization: A statistical design of experiments approach. *Clinics in Laboratory Medicine, 27,* 139–154.

123. Andricopulo, A. D., Salum, L. B., & Abraham, D. J. (2009). Structure-based drug design strategies in medicinal chemistry. *Current Topics in Medicinal Chemistry, 9,* 771–790.

124. Guido, R. V., Oliva, G., & Andricopulo, A. D. (2008). Virtual screening and its integration with modern drug design technologies. *Current Medicinal Chemistry, 15,* 37–46.

125. DeSilva, B., Smith, W., Weiner, R., Kelley, M., Smolec, J., Lee, B., et al. (2003). Recommendations for the bioanalytical method validation of ligand-binding assays to support pharmacokinetic assessments of macromolecules. *Pharmaceutical Research, 20,* 1885–1900.

126. Papaneophytou, C. P., Mettou, A. K., Rinotas, V., Douni, E., & Kontopidis, G. A. (2013). Solvent selection for insoluble ligands, a challenge for biological assay development: A TNF-alpha/SPD304 study. *ACS Medicinal Chemistry Letters, 4,* 137–141.

127. Papaneophytou, C. P., Grigoroudis, A. I., McInnes, C., & Kontopidis, G. (2014). Quantification of the effects of ionic strength, viscosity, and hydrophobicity on protein-ligand binding affinity. *ACS Medicinal Chemistry Letters, 5,* 931–936.

128. Cowan, K. J., Erickson, R., Sue, B., Delarosa, R., Gunter, B., Coleman, D. A., et al. (2012). Utilizing design of experiments to characterize assay robustness. *Bioanalysis, 4,* 2127–2139.

129. Onyeogaziri, F. C., & Papaneophytou, C. (2019). A general guide for the optimization of enzyme assay conditions using the design of experiments approach. *SLAS Discovery, 24,* 587–596.

130. Bisswanger, H. (2014). Enzyme assays. *Perspectives on Science, 1,* 41–55.

131. DOE in Assay Development Trends 2009 Report, published by HTStec Limited, Cambridge, UK, 18 July 2009.

132. Zhang, J. H., Chung, T. D., & Oldenburg, K. R. (1999). A simple statistical parameter for use in evaluation and validation of high throughput screening assays. *Journal of Biomolecular Screening, 4,* 67–73.

133. Boyacı, İ. H. (2005). A new approach for determination of enzyme kinetic constants using response surface methodology. *Biochemical Engineering Journal, 25,* 55–62.

134. Fang, H., Dong, H., Cai, T., Zheng, P., Li, H., Zhang, D., et al. (2016). In vitro optimization of enzymes involved in precorrin-2 synthesis using response surface methodology. *PLoS ONE, 11,* e0151149.

135. Buss, O., Jager, S., Dold, S. M., Zimmermann, S., Hamacher, K., Schmitz, K., et al. (2016). Statistical evaluation of HTS assays for enzymatic hydrolysis of beta-keto esters. *PLoS ONE, 11,* e0146104.

136. Chen, X. C., Zhou, L., Gupta, S., & Civoli, F. (2012). Implementation of design of experiments (DOE) in the development and validation of a cell-based bioassay for the detection of anti-drug neutralizing antibodies in human serum. *Journal of Immunological Methods, 376,* 32–45.

137. Sittampalam, G. S., Smith, W. C., Miyakawa, T. W., Smith, D. R., & McMorris, C. (1996). Application of experimental design techniques to optimize a competitive ELISA. *Journal of Immunological Methods, 190,* 151–161.

138. Hammond, O., Reynolds, J., Rubinstein, L. J., Sikkema, D., & Marchese, R. D. (2008). Complexities of clinical assay development and optimization prior to first-in-man immunization trials—A description of immunogenicity assay development for the testing of samples from a phase 1 Alzheimer's vaccine trial. *Journal of Immunoassay & Immunochemistry, 29,* 332–347.

139. Ray, C. A., Patel, V., Shih, J., Macaraeg, C., Wu, Y., Thway, T., et al. (2009). Application of multi-factorial design of experiments to successfully optimize immunoassays for robust measurements of therapeutic proteins. *Journal of Pharmaceutical and Biomedical Analysis, 49,* 311–318.

140. Mikulskis, A., Yeung, D., Subramanyam, M., & Amaravadi, L. (2011). Solution ELISA as a platform of choice for development of robust, drug tolerant immunogenicity assays in support of drug development. *Journal of Immunological Methods, 365,* 38–49.

141. Joelsson, D., Moravec, P., Troutman, M., Pigeon, J., & DePhillips, P. (2008). Optimizing ELISAs for precision and robustness using laboratory automation and statistical design of experiments. *Journal of Immunological Methods, 337,* 35–41.

142. Schmidt, T., Bergner, A., & Schwede, T. (2014). Modelling three-dimensional protein structures for applications in drug design. *Drug Discovery Today, 19,* 890–897.

143. Kwon, J. S., II, Nayhouse, M., Christofides, P. D., & Orkoulas, G. (2014). Protein crystal shape and size control in batch crystallization: Comparing model predictive control with conventional operating policies. *Industrial and Engineering Chemistry Research, 53,* 5002–5014.

144. Jancarik, J., & Kim, S.-H. (1991). Sparse matrix sampling: A screening method for crystallization of proteins. *Journal of Applied Crystallography, 24,* 409–411.

145. Stevens, R. C. (2000). High-throughput protein crystallization. *Current Opinion in Structural Biology, 10,* 558–563.

146. Giege, R. (2013). A historical perspective on protein crystallization from 1840 to the present day. *FEBS Journal, 280,* 6456–6497.

147. Abergel, C., Moulard, M., Moreau, H., Loret, E., Cambillau, C., & Fontecilla-Camps, J. C. (1991). Systematic use of the incomplete factorial approach in the design of protein crystallization experiments. *Journal of Biological Chemistry, 266,* 20131–20138.

148. Doudna, J. A., Grosshans, C., Gooding, A., & Kundrot, C. E. (1993). Crystallization of ribozymes and small RNA motifs by a sparse matrix approach. *Proceedings of the National Academy of Sciences of the United States of America, 90,* 7829–7833.

149. Luft, J. R., Newman, J., & Snell, E. H. (2014). Crystallization screening: the influence of history on current practice. *Archive of Acta Crystallographica Section F, Structural Biology Communications, 70,* 835–853.

150. Snell, E. H., Nagel, R. M., Wojtaszczyk, A., O'Neill, H., Wolfley, J. L., & Luft, J. R. (2008). The application and use of chemical space mapping to interpret crystallization screening results. *Acta Crystallographica. Section D, Biological Crystallography, 64,* 1240–1249.

151. Dinc, I., Pusey, M. L., & Aygun, R. S. (2016). Optimizing associative experimental design for protein crystallization screening. *IEEE Transactions on NanoBioscience, 15,* 101–112.