Kriging Based Data Analysis and Experimental Design in Biotechnology

Inaugural-Dissertation

zur Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultät
der Heinrich-Heine-Universität Düsseldorf

vorgelegt von

Lars Freier
aus Aschersleben

Korschenbroich, Februar 2018

aus dem Institut für
der Heinrich-Heine-Universität Düsseldorf

Gedruckt mit der Genehmigung der
Mathematisch-Naturwissenschaftlichen Fakultät der
Heinrich-Heine-Universität Düsseldorf

Referent: Prof. Dr. Oliver Ebenhöh

Korreferent: Prof. Dr. Wolfgang Wiechert

Tag der mündlichen Prüfung: 08.12.2017

Selbstständigkeitserklärung

Ich versichere an Eides Statt, dass die Dissertation von mir selbständig und ohne unzulässige fremde Hilfe unter Beachtung der „Grundsätze zur Sicherung guter wissenschaftlicher Praxis an der Heinrich-Heine-Universität Düsseldorf

Bisher habe ich keine erfolgslosen Promotionsversuche unternommen und diese Dissertation nicht an einer anderen Fakultät vorgelegt.

Ort, Datum	Lars Freier

“I have no knowledge of gold mines or of mining problems; I have therefore read Mr. Krige’s paper only as a statistician frequently concerned with a variety of sampling problems.”

(Dr. David John Finney)

Danksagung

Insbesondere möchte ich mich bei Dr. Eric von Lieres bedanken, der mir als Leiter der Modellierungs- und Simulationsgruppe am IBG-1 Jülich stets für Rat und Tat zur Seite stand. Des Weiteren möchte ich Jun.- Prof. Dr. Oliver Ebenhöh und Prof. Dr. Wolfgang Wiechert danken, die sich bereit erklärt haben das Gutachten für diese Arbeit zu verfassen.
Weiterer Danke gehören Johannes Hemmerich und Dr. Holger Morschett mit denen ich zwei großartige Publikationen schreiben durfte. Auch möchte ich meine Kollegen aus der Modellierungs- und Simulationsgruppe erwähnen, die mir stets zur Seite standen und mir gern Feedback bzgl. meiner wissenschaftlichen Arbeit gegeben haben.
Ich möchte außerdem beim CLIB-Graduate Cluster Industrial Biotechnology für die Finanzierung bedanken.
Natürlich möchte ich mich ebenfalls bei meiner Familie und meiner Freundin Yvonne Mertens bedanken, die mich auch abseits der akademischen Karriere unterstützt haben.

Danke, Lars Freier

List of publications, conference talks, conference proceedings, and open source software

Publications in Journals

L. Freier, J. Hemmerich, K. Schöler, W. Wiechert, M. Oldiges, E. von Lieres
"Framework for Kriging-based iterative experimental analysis and design: Optimization of secretory protein production in Corynebacterium glutamicum," *Eng. Life Sci.*, vol. 16, no. 6, pp. 538–549, Sep. 2016.

H. Morschett, L. Freier, J. Rohde, W. Wiechert, E. von Lieres, M. Oldiges
"A framework for accelerated phototrophic bioprocess development: integration of parallelized microscale cultivation, laboratory automation and Kriging-assisted experimental design," *Biotechnol. Biofuels*, vol. 10, no. 1, p. 26, Dec. 2017.

L. Freier and E. von Lieres
"Multi-objective global optimization (MOGO): Algorithm and case study in gradient elution chromatography," *Biotechnol. J.*, p. 1600613, Mar. 2017.

L. Freier and E. von Lieres
"Kriging with nonlinear trend functions: Theory and application in enzyme kinetics,", submitted in *Eng. Life Sci.* Feb. 2017

Conference Talks

L. Freier and E. von Lieres (2015)
Kriging based iterative parameter estimation procedure for biotechnology applications with nonlinear trend functions
8th Vienna International Conference on Mathematical Modelling, Vienna, Austria

L. Freier and E. von Lieres (2016)
Robust multi-objective process design
European Congress on Biotechnology, Krakow, Poland

L. Freier and E. von Lieres (2016)
Robust multi-objective process design
European Symposium on Biochemical Engineering Sciences, Dublin, Irland

L. Freier and E. von Lieres (2017)
Robust multi-objective process design
13th International PhD Seminar on Chromatographic Separation Science, Annweiler am Trifels, Germany

Conference Proceedings

L. Freier and E. von Lieres
“Kriging based iterative parameter estimation procedure for biotechnology applications with nonlinear trend functions,” *IFAC-PapersOnLine*, vol. 48, no. 1, pp. 574–579, 2015

Open Source Software

L. Freier and E. von Lieres
Kriging ToolKit (KriKit)
Available at: https://github.com/modsim/KriKit

Abstract

Process optimization problems emerge frequently in industry as well as in academia. Here, parameter sets must be found that maximize or minimize defined quality criteria. In particular in biotechnology, the high complexity of the investigated system and the high costs of the experiments make the usage of mathematical models attractive. These enable estimating the effect of process parameters on targeted objectives as well as the localization of potential optima.

In practice, empirical models, such as artificial neural networks and Kriging, have proven useful for solving process optimization problems. These kinds of models have in common that they are primarily based on the provided experimental data set but not on mechanistic knowledge. Using Kriging, the functional relationship between input and output variables is modeled as Gaussian process. Beside the interpolation capability, the stochastic ansatz allows further the direct estimation of the model uncertainty, which can be used for statistical approaches such as hypothesis tests.

Different to other empirical modelling approaches, Kriging also cares the possibility to integrate mechanistic models into the prediction. This feature makes Kriging to a hybrid modeling approach and is in particular helpful if the data density is low. However, at the present state of research, only models can be integrated that are linear in their parameters. As many models in biotechnology do not match this criterion, the aim of this thesis is *inter alia* to extent the Kriging methodology by the capability to integrate also nonlinear models.

The here presented works focusses further on the development and application of Kriging based optimization strategies. State-of-the-art Kriging based optimization algorithms are restricted on the single-objective case. However, in particular in biotechnology, there exists a need for multi-objective optimization regarding objectives such as purity, yield, and productivity. A further contribution of this thesis is therefore the development of the “Multi-Objective Gaussian Optimization” (MOGO) that integrates latest methods for Kriging based Design of Experiment and for the estimation of the model uncertainty.

Further, state-of-the-art Kriging based optimization algorithms are not flexible with respect to the changes in the range of input variables nor do they support parallel

experimentations. However, in biotechnology, there is a clear trend forward parallelization and the high complexity of biological systems makes it hard to define appropriated the range of input variables *a priori*. An additional contribution of this thesis is therefore the development of an optimization strategy that cares in its core an iterative Kriging based optimization but is extended by elements to tackle the mentioned deficiencies.

The effectiveness, efficiency as well as the practical feasibility of the here introduced approaches are examined and discussed in three case studies. Moreover, the reproducibility as well as the convergence behavior of the MOGO-algorithm are investigated in a comprehensive *in silico* study.

Zusammenfassung

Sowohl in der Industrie als auch in der Forschung sind Prozessoptimierungen allgegenwärtig. Optimale Prozessparameterwerte, die definierte Zielkriterien maximieren oder minimieren, müssen gefunden werden. Die, insbesondere in der Biotechnologie auftretende, hohe Komplexität des zu optimierenden Systems und der hohe Kostenaufwand der Versuchsdurchführung machen die Nutzung von mathematischen Modellen attraktiv. Diese erlauben die Abschätzung des Einflusses der Prozessparameter auf die Zielgrößen als auch die Lokalisierung potentieller Optima.

In der Praxis haben sich bzgl. initialer Prozessoptimierung insbesondere empirische Modelle, wie künstliche neuronale Netze und Kriging, bewährt. Diese Art von Modellen teilen die Eigenschaft, dass sie hauptsächlich auf experimentellen Information aufbauen, nicht aber auf mechanistischen Wissen.

Die hier vorliegende Doktorarbeit beschäftig sich mit der Entwicklung und Anwendung von Kriging-basierten Optimierungsstrategien. Kriging ist ein empirisches Schätzverfahren bei dem der Zusammenhang zwischen Eingangs- und Ausgangsgröße als Gaußprozesses beschrieben wird. Der stochastische Ansatz erlaubt neben der Interpolation auch die direkte Abschätzung der Modellunsicherheit und somit die Anwendung von statistische Verfahren wie Hypothesentests. Anders als bei anderen empirischen Modellen besteht bei Kriging die Möglichkeit, zusätzlich mechanistische Modelle in die Vorhersage zu integrieren. Dies ist insbesondere bei geringer Datendichte hilfreich und macht Kriging somit zu einem hybriden Modellierungsansatz. Zum derzeitigen Stand der Technik können jedoch nur Modelle integriert werden, die linear in ihren Parametern sind. Da dies aber für viele in der Biotechnologie verwendeten Modelle nicht der Fall ist, beschäftigt sich die vorliegende Doktorarbeit u.a. mit der Fragestellung wie die Kriging-Methodologie um die Integration nichtlinearer Modelle erweitert werden kann.

Des Weiteren beschränken sich gängige Kriging-Optimierungsmethoden auf einzelne Zielkriterien. Insbesondere in der Biotechnologie besteht jedoch der Bedarf nach mehrkriterieller Optimierung bzgl. Zielgrößen, wie Reinheit, Ertrag und Produktivität. Ein wesentlicher Beitrag dieser Arbeit ist deshalb die Entwicklung des „Multi-Objective

Gaussian Optimization“ (MOGO) Algorithmus, welcher neueste Methoden zur Kriging-basierten Versuchsplanung und zur Abschätzung der Modellunsicherheit in sich vereint. Derzeitige Kriging-basierte Optimierungsalgorithmen bieten weder die Flexibilität, gewählte Parameterbereiche zu erweitern, noch unterstützen sie paralleles Experimentieren. Insbesondere in der Biotechnologie zeichnet sich jedoch derzeit ein Trend zur Parallelisierung ab, und die hohe Komplexität biologischer Systeme erschwert zusätzlich eine geeignete *a priori* Abschätzung von Parameterbereichen. Ein weiterer Beitrag dieser Arbeit ist deshalb die Entwicklung einer Optimierungsstrategie die im Kern eine iterative Kriging-basierte Optimierung ist, aber um Elemente erweitert wurde, um die oben genannten Schwächen zu überwinden.

Die Effektivität, Effizienz sowie die Praxistauglichkeit der in dieser Arbeit entwickelten Verfahren werden anhand von drei Fallstudien geprüft und diskutiert. Des Weiteren wird die Reproduzierbarkeit sowie das Konvergenzverhalten des MOGO-Algorithmus in einer umfangreichen *in silico* Studie untersucht.

Table of Contents

List of Abbreviations

DoE	Design of Experiments
UK	Universal Kriging
GPR	Gaussian Process Regression
$Z(\boldsymbol{x})$	Output Variable
$\boldsymbol{x}$	Input vector
$m(\boldsymbol{x})$	Kriging Trend Function
$C(\boldsymbol{x}_i, \boldsymbol{x}_j)$	Covariogram
$\boldsymbol{\lambda}$	Kriging/GPR coefficient for data point
MLE	Maximum Likelihood Estimation
$\boldsymbol{\theta}$	Covariogram parameter set
n	Number of measured data points
m	Number of input variables
k	Number of trend functions
$\boldsymbol{\mu}$	Vector of Lagrange multipliers
KBDoE	Kriging Based Design of Experiment
PI	Probability of Improvement
EI	Expected Improvement
$\mathrm{I}\big(Z(\boldsymbol{x})\big)$	Improvement at location $\boldsymbol{x}$
$\mathrm{P_I}(\boldsymbol{x})$	Probability of an improvement at location $\boldsymbol{x}$
$\sigma^2_{\mathrm{UK}}(\hat{\boldsymbol{x}})$	Kriging prediction variance
KriKit	Kriging toolKit
CV	Cross-Validation
MCMC	Markov-Chain Monte-Carlo
DRAM	Delayed Rejection Adaptive Metropolis
n_{Obj}	Number of objectives of interest
n_{Pareto}	Number of Pareto optimal points
P	Set of Pareto optimal points
MOGO	Multi-Objective Gaussian Optimization
MOO	Multi-Objective Optimization
MOEA	Multi-Objective Evolutionary Algorithm

EHVI	Expected HyperVolume Improvement
EGO	Efficient Global Optimization
KBO	Kriging Based Optimization
GFP	Green Fluorescent Protein
HV	HyperVolume
GA	Genetic Algorithm

1. Introduction

1.1 Aim of the Thesis

This thesis has three major aims. First, it focusses on the extension of the Kriging methodology by the ability to integrate trend functions that are non-linear in their parameters. As discussed in section 1.3, using appropriate trend functions leads potentially to better model predictions, in particular in areas with low data density. In particular, this extension is essential for biotechnology, as here many mechanistic models are non-linear in their parameters, such as the Michaelis-Menten enzyme kinetic. The approach is derived in section 2.1 and is based on Taylor linearization, leading to an iterative parameter estimation procedure. This iterative procedure is subsequently interpreted as a root-finding problem making it accessible to numeric solvers specialized on this type of problems.

Furthermore, in section 1.5, the practice feasibility of state-of-the-art Kriging based optimization algorithms is discussed with focus on biotechnological problems. It is explained that current procedures would be more attractive to biotechnology if they would support the three following aspects: 1) an initial screening for input variables with (non-)significant effect 2) support of parallel experimentation and 3) the possibility to change the range of input variables. As a result, a Kriging based optimization strategy is introduced that considers these aspects with particular focus on parallel experimentation in section 1.5.1. The application of this optimization strategy is demonstrated in two experimental case studies in section 2.2&2.3.

A further goal of this thesis is the development of a Kriging based algorithm for multi-objective optimization. The resulting “Multi-Objective Gaussian Optimization” (MOGO) algorithm is described in section 2.4 and transfers the state of the art algorithm “Efficient Global Optimization” (EGO) [1] to the multi-objective case. In a comprehensive *in silico* case study from the field of preparative chromatography, the convergence behaviors as well as the reproducibility of the results and the parallelization capability are investigated. Moreover, the MOGO is discussed as alternative to currently used multi-objective optimization algorithms that are based on genetic algorithms.

1.2 General Introduction

Mathematical models have been proven useful tools for generating knowledge and for efficiently optimizing complex and nonlinear systems. In particular, biological processes show a high degree of complexity making an intuitive overall understanding impossible. Additionally, using mathematical simulations instead of “real” experiment often represents an economical advantage. Moreover, effective mathematical optimization algorithms, such as the simplex algorithm or genetic algorithms, require in many cases a high number of samples. Therefore, only the mathematical abstraction of the physical system enables the application of these optimization algorithms.

Mechanistic models are constructed based on a priori knowledge, such as physical, chemical or biological laws. These laws define the overall structure of the model including several adjustable parameters. They find applications in a variety of biotechnological fields, for example in chromatography [2], microfluidic single-cell cultivation [3], metabolic engineering [4], or for modeling enzyme kinetics [5]. Mechanistic models can be used for estimating quantifiable variables that are not directly measureable, such as concentration values or diffusion coefficients. Furthermore, by using appropriate mechanistic models and optimal experimental design, the number of needed experiments for the modeling procedure can be kept low.

However, the development of an adequate mechanistic model and the associated parameter identification are non-trivial problems. Franceschini et al. [6] provide a good overview of parameter identification strategies with focus on applications regarding biochemical networks and biological processes. In addition, complex mechanistic models, as used in Computational Fluid Dynamic (CFD), often suffer from long simulation times.

For process optimization, it is often sufficient to use empirical models, which are mainly data driven. In comparison with mechanistic models, they carry the advantage to be universally applicable and computationally cheap. Typical empirical modeling approaches are fitting polynomials [7], deep learning via artificial neural network [8], or Kriging [9].

In particular, the polynomial modeling based Design of Experiment (DoE) is very popular and finds a variety of applications in academia and industry. The associated full factorial, fractional factorial or central composite designs help to efficiently estimate the effects of

input variables on the objective of interest. Mandenius et al. [10] reviewed common applications of this “classic” DoE in the development of biotechnological processes. However, for highly nonlinear systems and wide ranges of input values, the prediction accuracy can be insufficient.

In these cases, more sophisticated modeling approaches should be preferred, such as Kriging. Simpson [11] and Cock [12] have demonstrated that Kriging typically outperforms the polynomial fit with respect to prediction accuracy when nonlinear systems are investigated. Kriging based optimization is already applied in many fields, in particular in fluid dynamics [13], [14].

In biotechnology, however, Kriging has only rarely been used with main focus on data visualization, e.g. [15], [16]. Therefore, this thesis concentrates on developing Kriging based optimization approaches adapted to the needs in biotechnology. This includes both single-objective optimization as well as multi-objective optimization. The approaches are applied to experimental and *in silico* studies. Moreover, the convergence behavior, the reproducibility of the optimization results as well as the use of parallel experimentation is investigated. Further, the integration of mechanistic models into the Kriging methodology is discussed and a method for integrating mechanistic models that are nonlinear in their parameter is introduced in this thesis.

1.3 Brief Introduction to Kriging

When Daniel Krige first published his prediction method (later called Kriging) in 1951 [17], he could not foresee how popular and important his method would become in the future. At the time point of writing this thesis, the original publication was cited 2118 times! However, it should also be noted that Georges Matheron later developed the mathematical derivation of Kriging that it is used nowadays [18].

In the following, if not stated otherwise, Kriging refers to the most commonly used approach: Universal Kriging (UK). However, it is worth mentioning that there are many different approaches associated with the term Kriging (e.g. Indicator Kriging [19], Bayesian Kriging [20], etc.). UK originates from the statistical frequentist perspective and its prediction results are identical to the outcome of the Gaussian Process Regression (GPR) methodology originating from the perspective of the Bayesian statisticians.

Although the model predictions are identical for UK and GPR, the mathematical derivations differ from each other. In all my publications (section 2) I provide the derivation of UK from the frequentist viewpoint. Here, I want to take the opportunity to provide the derivation of GPR in order to emphasize differences and similarities between these two different ansatzes.

In both cases, it is assumed that the measured sample data result from a Gaussian process, as visualized in Figure 1. That is, each measured output originates from a stochastic process $Z(\boldsymbol{x})$ comprising a basic trend $m(\boldsymbol{x})$ and the stochastic variable $Y(\boldsymbol{x})$, with $\boldsymbol{x} = (x_1, \dots, x_m)^T \in \mathbb{R}^m$. Here, $Y(\boldsymbol{x})$ has zero mean, and a defined covariance $C(\boldsymbol{x}_i, \boldsymbol{x}_j)$.

$$Z(\boldsymbol{x}) = m(\boldsymbol{x}) + Y(\boldsymbol{x}) \tag{1}$$

$$\mathrm{E}[Y(\boldsymbol{x})] = 0 \tag{2}$$

$$C(\boldsymbol{x}_i, \boldsymbol{x}_j) = \mathrm{Cov}\left(Y(\boldsymbol{x}_i), Y(\boldsymbol{x}_j)\right) = \mathrm{Cov}\left(Z(\boldsymbol{x}_i), Z(\boldsymbol{x}_j)\right) \tag{3}$$

In the frameworks of UK and GPR, the basic trends can be furthermore represented by a linear combination of k functions $f_l(\boldsymbol{x})$ with the coefficients a_l:

$$m(\boldsymbol{x}) = \sum_{l=1}^{k} a_l f_l(\boldsymbol{x}) \tag{4}$$

The coefficients a_l are implicitly estimated by Kriging, which is in detail discussed in publication I, see section 2.1, and not repeated here.

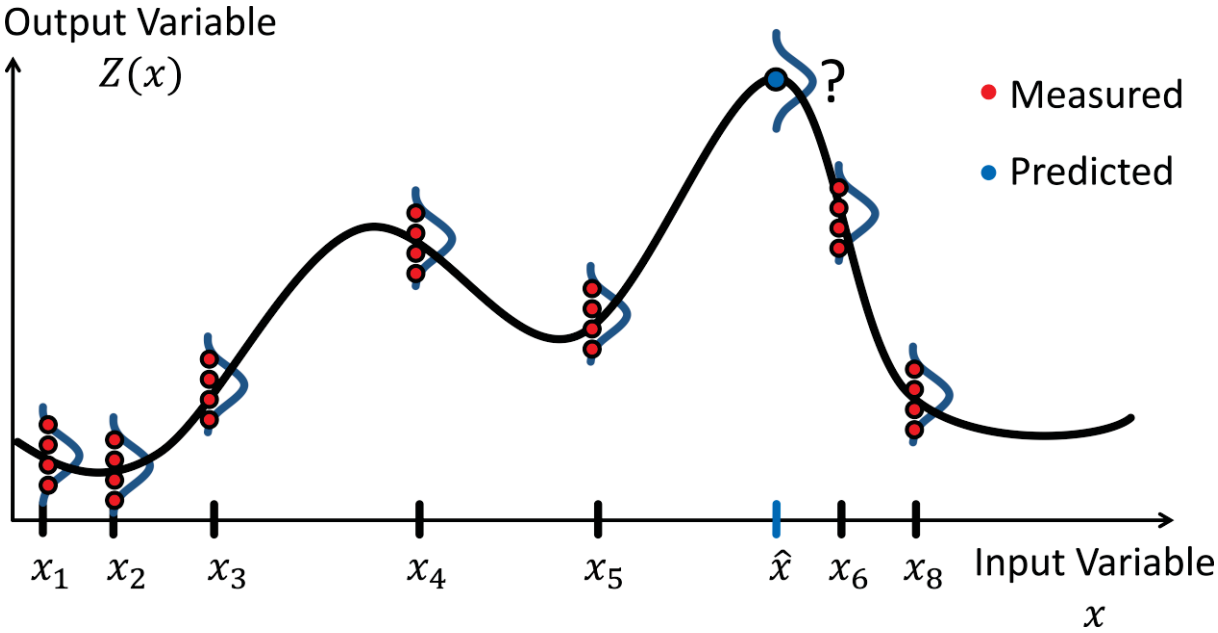


Figure 1: Schematic illustration of a Gaussian process. Red dots depict sample data and the blue dot is the model prediction $Z^*(\hat{x})$ at the point $\hat{x}$. The black curve represents the "true" functional relationship. The blue curves indicate the Gaussian probability distribution $Y(\boldsymbol{x})$.

1.3.1 Gaussian Process Regression

GPR originates from the Bayesian perspective and aims to estimate the conditioned output distribution $Z^*(\hat{X})|Z(X), X = Z(\hat{\boldsymbol{x}}_1, \dots, \hat{\boldsymbol{x}}_n)^T \in \mathbb{R}^n$ at arbitrary input locations $\hat{X} = (\hat{\boldsymbol{x}}_1, \dots, \hat{\boldsymbol{x}}_{\hat{n}})^T \in \mathbb{R}^{\hat{n} \times m}$ for a given set of measured output values $Z(X) = \big(Z(\boldsymbol{x}_1), \dots, Z(\boldsymbol{x}_n)\big)^T \in \mathbb{R}^n$ at the input locations $X = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_n)^T \in \mathbb{R}^{n \times m}$, with n as the number of measured data points, $\hat{n}$ the number of points of interest and m as the number of input variables. Rasmussen [21] provides a detailed overview about the mathematical derivation, different covariance models, and applications of GPR.

The mathematical ansatz of GPR is to interpret the Gaussian process in eqs. (1)-(3) as prior distribution for $Z^*(\hat{X})$, eq. (5).

$$Z^*(\hat{X}) \sim N\left(\boldsymbol{m}(\hat{X}), C(\hat{X}, \hat{X})\right) \tag{5}$$

The term prior refers to the Bayesian approach and implies that eq. (5) represents the probability distribution of $Z(\hat{X})$ before the measured observations are taken into account. The prior distribution is determined by choosing an appropriate trend function and covariance model $C(\hat{X}, \hat{X})$ and by estimating the associated parameters. The influence

of the trend function $m(\hat{X})$ is later discussed in section 1.3.3. In the following, the estimation of the covariance model is described.

The covariance model $C(\boldsymbol{x}_i, \boldsymbol{x}_j)$, also known as covariogram model, describes covariance of the output values, $Z(\boldsymbol{x}_i), Z(\boldsymbol{x}_j)$, as function of the associated locations $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$. Three covariogram model types are commonly applied: the spherical, the Matérn, and the exponential model [22]. All three models have in common that they are secondary stationary, i.e. they depend on the distance between the inputs $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ and not on the actual input values.

$$C(\boldsymbol{x}_i, \boldsymbol{x}_j) = C(|\boldsymbol{x}_i - \boldsymbol{x}_j|) = C(\Delta \boldsymbol{x}_{i,j}) \tag{6}$$

The final choice for one of these covariograms depends on the provided data set. Details about the chosen covariogram functions for my studies can be found in the respective publication (section 2), and is not discussed here.

Each covariogram contains model parameters $\boldsymbol{\theta}$ that have to be estimated from the given data set. In context of GPR, the covariogram parameters are determined using the Maximum Likelihood Estimation (MLE) [23]. Here, in agreement with the Gaussian process assumption, it is assumed that the set of measurements $Z(X) = \big(Z(\boldsymbol{x}_1), \dots, Z(\boldsymbol{x}_n)\big)^T \in \mathbb{R}^n$ follows a multivariate Gaussian distribution with the expected values $\boldsymbol{m}(X) = \big(m(\boldsymbol{x}_1), \dots, m(\boldsymbol{x}_n)\big)^T \in \mathbb{R}^n$ and the covariance of $C(X, X, \boldsymbol{\theta}^*) \in \mathbb{R}^{n \times n}$. Eq. (7) describes the associated probability density function.

$$p(Z(X)|\boldsymbol{\theta}^*) = \frac{\exp\left(-\frac{1}{2}\big(Z(X) - \boldsymbol{m}(X)\big)^T C(X, X, \boldsymbol{\theta}^*)^{-1}\big(Z(X) - \boldsymbol{m}(X)\big)\right)}{(2\pi)^{\mathrm{n}/2}\det\big(C(X, X, \boldsymbol{\theta}^*)\big)^{1/2}} \tag{7}$$

The entries in the covariance matrix $C \in \mathbb{R}^{n \times n}$ are calculated by the covariance function using the parameters $\boldsymbol{\theta}^*$. For a fixed data set $\{X, Z(X)\}$, the probability in eq. (7) only depends on the covariogram parameters and can be used for determining the best parameter set. In practice, the best parameter set maximizes the logarithm of eq. (7).

$$\log\big(p(Z|\boldsymbol{\theta}^*)\big) = -\frac{1}{2}(Z - \boldsymbol{m})^{\mathrm{T}} C(\boldsymbol{\theta}^*)^{-1}(Z - \boldsymbol{m}) - \frac{1}{2}\log(|C(\boldsymbol{\theta}^*)|) - \frac{n}{2}\log(2\pi) \tag{8}$$

Figure 2A depicts the prior distribution $N\left(\boldsymbol{m}(\hat{X}), C(\hat{X}, \hat{X})\right)$ for a schematic example using a constant trend function and a Matérn class covariogram with a smoothing parameter of 3/2, given in eq. (9)

$$C(\Delta \boldsymbol{x}_{i,j}) = \theta_{\epsilon}^2 + \theta_{\sigma}^2 \left(1 + \sqrt{3}r\right) \exp\left(-\sqrt{3}r\right), \text{ with } r = \sqrt{\sum_{l=1}^{m} \frac{\left(\Delta \boldsymbol{x}_{i,j}^{l}\right)^2}{\theta_l^2}} \quad (9)$$

With $\Delta \boldsymbol{x}_{i,j} = \left(\left|\boldsymbol{x}_1^{(i)} - \boldsymbol{x}_1^{(j)}\right|, \ldots, \left|\boldsymbol{x}_m^{(i)} - \boldsymbol{x}_m^{(j)}\right|\right)^T$ representing the absolute dististance between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$.

Gray curves in Figure 2A represent stochastic simulations following the prior distribution, eq. (5). The covariogram parameters were estimated using the maximum likelihood approach. The smoothness of the curves is controlled by the covariance function while the trend function defines the expectation values over the input space.

A B

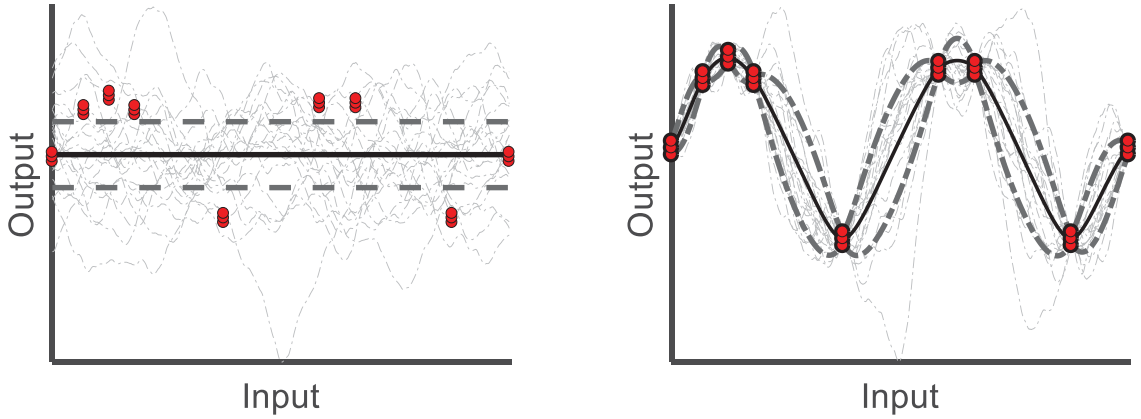


Figure 2: Stochastic simulation of A) the prior distribution B) the posteriori distribution. Gray curves represent 20 stochastic simulations over a grid with 1000 sample points. The black bold lines indicate the expectation curves and the dashed lines represent the 67% confidence tube. Red dots are the given sample points used for the covariagram estimation and for the conditioning.

In the second step, the posterior distribution is calculated by conditioning the prior distribution with the information about the measured samples. The joint distribution of the measured and estimated output is given by eq. (10).

$$\begin{bmatrix} Z(X) \\ Z^*(\hat{X}) \end{bmatrix} \sim N\left(\begin{bmatrix} \boldsymbol{m}(X) \\ \boldsymbol{m}(\hat{X}) \end{bmatrix}, \begin{bmatrix} C(X,X) & C(X,\hat{X}) \\ C(X,\hat{X})^T & C(\hat{X},\hat{X}) \end{bmatrix}\right) \quad (10)$$

It can be shown [24] that the resulting conditioned posteriori distribution is given by

$$Z^*(\hat{X})|Z(X), X \sim N\left(\mu_{\mathrm{GPR}}^*(\hat{X}), \Sigma_{\mathrm{GPR}}^*(\hat{X},\hat{X})\right) \quad (11)$$

$$\text{with } \boldsymbol{\mu}_{\mathrm{GPR}}^*(\hat{X}) = \boldsymbol{m}(\hat{X}) + C(X,\hat{X})^T C(X,X)^{-1}(Z(X) - \boldsymbol{m}(X)) \quad (12)$$

$$\text{and } \Sigma_{\mathrm{GPR}}^*(\hat{X},\hat{X}) = C(\hat{X},\hat{X}) - C(X,\hat{X})^T C(X,X)^{-1} C(X,\hat{X}) \quad (13)$$

Eq. (12)-(13) can be rewritten into a more readable form, by defining the GPR coefficients $\boldsymbol{\lambda}_{\mathrm{GPR}}$.

$$\boldsymbol{\mu}^*_{\mathrm{GPR}}(\widehat{X}) = \left(\boldsymbol{m}(\widehat{X}) - \boldsymbol{\lambda}^T_{\mathrm{GPR}}\boldsymbol{m}(X)\right) + \boldsymbol{\lambda}^T_{\mathrm{GPR}}Z(X) \tag{14}$$

$$\Sigma^*_{\mathrm{GPR}}(\widehat{X},\widehat{X}) = C(\widehat{X},\widehat{X}) - \boldsymbol{\lambda}^T_{\mathrm{GPR}}C(X,\widehat{X}) \tag{15}$$

$$\text{With } \boldsymbol{\lambda}_{\mathrm{GPR}} = C(X,X)^{-1}C(X,\widehat{X}) \tag{16}$$

As discussed in further detail in section 1.3.3, with increasing distance between the test point and the sample points, GPR coefficients $\boldsymbol{\lambda}_{\mathrm{GPR}}$ converge to zero. Consequently, the posteriori variance Σ^* and the posteriori expectation value $\boldsymbol{\mu}^*$ converge to their prior correspondents. That is, the trend function $\boldsymbol{m}(\widehat{\boldsymbol{x}})$ and the modeled variance $C(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{x}})$.

Figure 2B depicts the posterior distribution $N\left(\boldsymbol{\mu}^*_{\mathrm{GPR}}(\widehat{X}), \Sigma^*_{\mathrm{GPR}}(\widehat{X},\widehat{X})\right)$ for the schematic example. Stochastic simulations were calculated in the same way as already done for the prior distribution. The conditioning step causes the expectation curve to follow the sample points.

1.3.2 Universal Kriging

The most commonly cited literature regarding Universal Kriging (UK) is the book *Statistics for Spatial Data* from Noel Cressie [9]. From the frequentist perspective, UK aims at constructing a linear predictor $z^*_{uk}(\widehat{\boldsymbol{x}})$, that is unbiased, and has minimal error variance.

$$z^*_{\mathrm{UK}}(\widehat{\boldsymbol{x}}) = \mathrm{E}[Z^*(\widehat{\boldsymbol{x}})] = \boldsymbol{\lambda}^T_{\mathrm{UK}}Z(X) \tag{17}$$

$$\mathrm{E}[Z(\widehat{\boldsymbol{x}}) - z^*_{\mathrm{UK}}(\widehat{\boldsymbol{x}})] = 0 \tag{18}$$

$$\mathrm{Var}[Z(\widehat{\boldsymbol{x}}) - z^*_{\mathrm{UK}}(\widehat{\boldsymbol{x}})] \rightarrow \min \tag{19}$$

With the estimation point $\widehat{\boldsymbol{x}} = (\hat{x}_1, \dots, \hat{x}_m)^T \in \mathbb{R}^m$, the sample locations $X = (\boldsymbol{x}_1^T, \dots, \boldsymbol{x}_n^T) \in \mathbb{R}^{n\times m}$, and the Kriging coefficients $\boldsymbol{\lambda}_{\mathrm{UK}} = \left(\boldsymbol{\lambda}^{(1)}_{\mathrm{UK}}, \dots, \boldsymbol{\lambda}^{(n)}_{\mathrm{UK}}\right)^T \in \mathbb{R}^n$.

The unbiasedness condition leads to k additional constraints regarding the trend function.

$$\begin{aligned} &\mathrm{E}[Z(\widehat{\boldsymbol{x}}) - z^*_{\mathrm{UK}}(\widehat{\boldsymbol{x}})] = 0 \\ &\Leftrightarrow \mathrm{E}[Z(\widehat{\boldsymbol{x}})] = \mathrm{E}[z^*_{\mathrm{UK}}(\widehat{\boldsymbol{x}})] \end{aligned} \tag{20}$$

$$\Leftrightarrow a_l f_l(\hat{\boldsymbol{x}}) = \boldsymbol{\lambda}_{\mathrm{UK}}^T a_l f_l(X)$$

With $l = 1, \dots, k$ and $f_l(X) = \left(f_l(\boldsymbol{x}_1), \dots, f_l(\boldsymbol{x}_n)\right)^T \in \mathbb{R}^n$. Eqs. (4) and (20) then lead to:

$$m(\hat{x}) = \boldsymbol{\lambda}_{\mathrm{UK}}^T m(\hat{X}). \tag{21}$$

By comparing eq. (14) and eq. (17), it becomes clear that the unbiasedness condition is necessary in order to make the result from GPR and UK equivalent.

Calculating the Kriging coefficients $\boldsymbol{\lambda}_{\mathrm{UK}}$ represents a minimization problem with k equality constraints. The method of Lagrange multipliers can be applied for finding the optimal $\boldsymbol{\lambda}_{\mathrm{UK}}$ vector that minimizes the prediction variance and fulfils the conditions given by eq. (21). In this context, the constraint minimization problem is translated to an unconstrained extreme point problem of the Lagrange function:

$$\begin{aligned} L(\boldsymbol{\lambda}_{\mathrm{UK}}, \boldsymbol{\mu}) &= \mathrm{Var}[Z(\hat{\boldsymbol{x}}) - z_{\mathrm{UK}}^*(\hat{\boldsymbol{x}})] - \sum_{l=1}^{k} \mu_l \left(f_l(\hat{\boldsymbol{x}}) - \boldsymbol{\lambda}_{\mathrm{UK}}^T f_l(X)\right) \\ &= \mathrm{Var}[Z(\hat{\boldsymbol{x}}) - \boldsymbol{\lambda}_{\mathrm{UK}}^T Z(X)] - (\boldsymbol{f}^T - \boldsymbol{\lambda}_{\mathrm{UK}}^T F)\boldsymbol{\mu} \end{aligned} \tag{22}$$

With $\mu_l \in \mathbb{R}$ as Lagrange multiplier, $F \in \mathbb{R}^{n \times k}$ is a matrix with entries $F_{i,l} = f_l(\boldsymbol{x}_i)$, and $\boldsymbol{f} \in \mathbb{R}^k$ is a vector with entries $f_l(\hat{\boldsymbol{x}})$. The second term in eq. (22) represents the equality constraints and has to be zero. The first term can be reformulated to eq. (23).

$$\mathrm{Var}[Z(\hat{\boldsymbol{x}}) - \boldsymbol{\lambda}_{\mathrm{UK}}^T Z(X)] = C(\hat{\boldsymbol{x}}, \hat{\boldsymbol{x}}) - 2\boldsymbol{\lambda}_{\mathrm{UK}}^T C(\hat{\boldsymbol{x}}, X) + \boldsymbol{\lambda}_{\mathrm{UK}}^T C(X, X)\boldsymbol{\lambda}_{\mathrm{UK}} \tag{23}$$

For an extreme point of $L(\boldsymbol{\lambda}_{\mathrm{UK}}, \boldsymbol{\mu})$ it holds that all partial derivatives, eq. (22), w.r.t. λ_i and μ_l are equal to zero.

$$\frac{\partial L(\boldsymbol{\lambda}_{\mathrm{UK}}, \boldsymbol{\mu})}{\partial \boldsymbol{\lambda}_{\mathrm{UK}}} = -2C(\hat{\boldsymbol{x}}, X) + 2C(X, X)\boldsymbol{\lambda}_{\mathrm{UK}} + F\boldsymbol{\mu} = \mathbf{0} \tag{24}$$

$$\frac{\partial L(\boldsymbol{\lambda}_{\mathrm{UK}}, \boldsymbol{\mu})}{\partial \boldsymbol{\mu}} = \boldsymbol{f}^T - \boldsymbol{\lambda}_{\mathrm{UK}}^T F = \mathbf{0} \tag{25}$$

With $\mathbf{0}$ representing a zero vector. Eqs. (24)-(25) can be reformulated in a matrix presentation eq. (26).

$$\begin{bmatrix} C & F \\ F^T & \mathbf{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{\lambda}_{\mathrm{UK}} \\ \boldsymbol{\mu}^* \end{bmatrix} = \begin{bmatrix} \boldsymbol{c} \\ \boldsymbol{f} \end{bmatrix} \tag{26}$$

Here, $C = C(X, X) \in \mathbb{R}^{n \times n}$ is a matrix with entries $C_{i,j} = C(\boldsymbol{x}_i, \boldsymbol{x}_j)$, $\boldsymbol{c} = C(\hat{\boldsymbol{x}}, X) \in \mathbb{R}^n$ is a vector with entries $c_i = C(\boldsymbol{x}_i, \hat{\boldsymbol{x}})$, and $\boldsymbol{\mu}^* = \frac{1}{2}\boldsymbol{\mu}$.

An explicit formula for an inverse of a 2x2-block matrix can be found in the literature [25] and leads to eq. (27).

$$\begin{bmatrix} C & F \\ F^T & \mathbf{0} \end{bmatrix}^{-1} = \begin{bmatrix} C^{-1} - C^{-1}F(F^TC^{-1}F)^{-1}F^TC^{-1} & C^{-1}F(F^TC^{-1}F)^{-1} \\ (F^TC^{-1}F)^{-1}C^{-1}F & -(F^TC^{-1}F)^{-1} \end{bmatrix}^{-1} \tag{27}$$

The explicit calculation formula for the Kriging coefficients $\boldsymbol{\lambda}_{\mathrm{UK}}$ can consequently be formulated.

$$\begin{aligned} \boldsymbol{\lambda}_{\mathrm{UK}} &= C^{-1}\boldsymbol{c} - C^{-1}F\left(F^TC^{-1}F\right)^{-1}F^TC^{-1}\boldsymbol{c} + C^{-1}F\left(F^TC^{-1}F\right)^{-1}\boldsymbol{f} \\ &= C^{-1}\boldsymbol{c} + C^{-1}F\left(F^TC^{-1}F\right)^{-1}(-F^TC^{-1}\boldsymbol{c} + \boldsymbol{f}) \end{aligned} \tag{28}$$

Using eq. (28) a direct conversion between the GPR coefficients $\boldsymbol{\lambda}_{\mathrm{GPR}} = C^{-1}\boldsymbol{c}$, eq. (16), and the UK coefficients $\boldsymbol{\lambda}_{\mathrm{UK}}$, eq. (29), can be formulated.

$$\begin{aligned} \boldsymbol{\lambda}_{\mathrm{UK}} &= \boldsymbol{\lambda}_{\mathrm{GPR}} + R(-F^T\boldsymbol{\lambda}_{\mathrm{GPR}} + \boldsymbol{f}) \\ &= (1 - RF^T)\boldsymbol{\lambda}_{\mathrm{GPR}} + R\boldsymbol{f} \end{aligned} \tag{29}$$

$$\text{With } R = C^{-1}F(F^TC^{-1}F)^{-1} \tag{30}$$

The reader might be confused at this point why different coefficients are needed for GPR and UK. The reason lies in the different formulation of the prediction formulas, i.e. eq. (14) for GPR and eq. (17) for UK. While the GPR prediction formula explicitly includes the trend functions, the UK prediction formula does not. That is, the effect of the trend functions $f_{1,\dots,k}(\boldsymbol{x})$ has to be implicitly included through the calculation of the UK coefficients $\boldsymbol{\lambda}_{\mathrm{UK}}$.

In contrast to the Bayesian statisticians, the frequentists estimate the covariogram model parameters by least square fitting based on the related stationary variogram [26]:

$$\mathrm{Var}\left(\Delta\boldsymbol{x}_{i,j}\right) = 2\left(C(0) - C\left(\Delta\boldsymbol{x}_{i,j}\right)\right) \tag{31}$$

Sample data for the least square fit are approximated using Matheron's estimator [18]:

$$\mathrm{Var}\left(\Delta\boldsymbol{x}_{i,j}\right) \cong \left(Z(\boldsymbol{x}_i) - Z\left(\boldsymbol{x}_j\right)\right)^2 \tag{32}$$

In case of replicates, $\mathrm{Var}\left(\Delta\boldsymbol{x}_{i,j}\right)$ is calculated by averaging eq. (32).

In order to demonstrate the UK prediction approach, the same schematic example is applied as in section 1.3.1. The results are visualized in Figure 3. The red dots in Figure 3A depict the data set that is used for constructing the Kriging model. At each of the nine sample locations, three replicates are measured. The associated least-square fitted

variogram is depicted in Figure 3B. It is typical that the estimated variogram shows a better agreement with the data in the near vicinity to the coordinate origin. The Kriging prediction curve is depicted in Figure 3A in black and the associated 67% confidence tube has been drawn in blue. The Kriging prediction curve as well as the confidence tube are approximately the same as for GPR, in section 1.3.1. However, it should be noted that in some cases, the estimated covariogram parameter values depend on the estimation approach, i.e. MLE or variogram fitting. In these cases, the different parameter values will automatically also influence the model prediction. In practice, the variogram fitting approach has proven less robust than MLE. A possible reason for this phenomenon represents the Matheron's estimator for the variance, eq. (32), that introduces an additional source of inaccuracy.

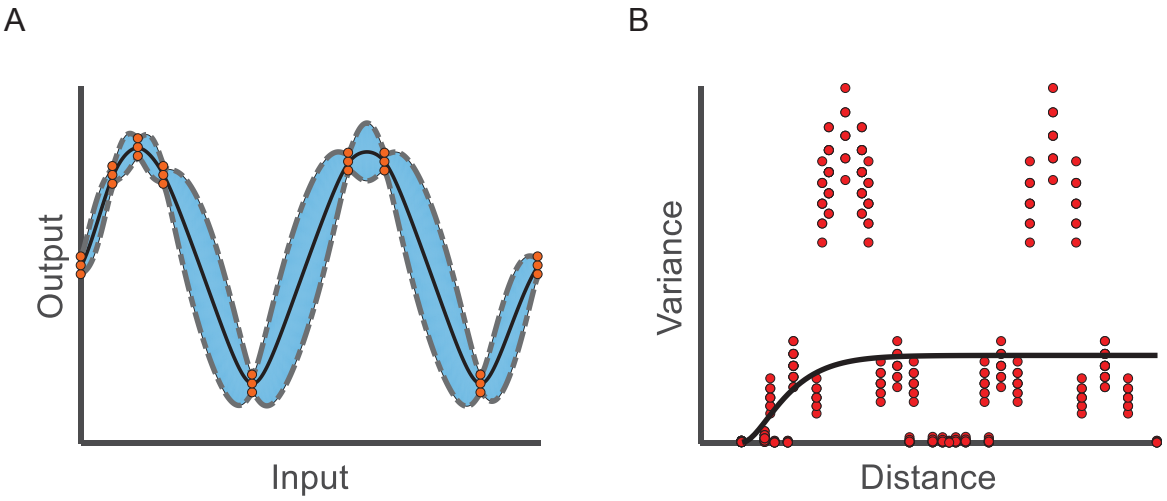


Figure 3: Schematic example. A) Kriging approximation. Sample data are indicated by red dots and the black line represents the Kriging prediction $Z^*(\hat{x})$. The 67% confidence is visualized in blue. B) Variogram. Red dots indicate the Matheron estimation of the variance. The black line represents the estimated variogram function.

1.3.3 Kriging with Nonlinear Basis Functions

The results from the GPR derivation in section 1.3.1 demonstrate that the posteriori distribution converges towards the prior distribution when the distance between the test locations $\hat{X}$ and the sample points X increases. This statement can be easily concluded from eq. (14)-(15) as with increasing distance, the covariance $C(X, \hat{X})$ converges towards a zero matrix. As a result, the posteriori expectation curve $\mu^*(\hat{X})$ and covariance $\Sigma^*(\hat{X}, \hat{X})$ converge towards eqs. (33)-(34).

$$\mu^*(\hat{X}) \to \boldsymbol{m}(\hat{X}) \tag{33}$$

$$\Sigma^*(\hat{X}, \hat{X}) \to C(\hat{X}, \hat{X}) \tag{34}$$

This explains why in sparsely sampled areas an appropriate trend function is necessary for sufficiently modeling of the underlying stochastic process.

However, universal Kriging and other state of the art Kriging approaches are limited to the use of trend functions that are linear in their parameters, see eq. (4). A workaround is to estimate the nonlinear parameter values via least-square-fitting. That is, using a local or global optimizer for minimizing the sum of squares of the residuals between trend function output and measurement data. Kriging can then use the originally nonlinear trend function by fixing the model parameters to their previously estimated values. However, in publication I (section 2.1), a more elegant way is provided by integrating the nonlinear parameter estimation into the framework of Kriging. The approach is based on Taylor linearization, leading to an iterative parameter estimation procedure. This iterative procedure is subsequently interpreted as a root-finding problem making it accessible to numeric solvers specialized on this type of problems.

1.4 Kriging Based Optimization

In addition to using Kriging for visualization purposes, the Kriging approach also represents a useful tool for optimization tasks, i.e. maximizing or minimizing an objective of interest. For the sake of clarity, in this thesis, the term optimization always refers to a maximization problem. Note that any minimization problem can easily be converted to a maximization problem by multiplying the output values by minus one. A Kriging based optimization procedure comprises in general an iterative cycle. In each iteration, the functional relationship between the input variables and the output variable is modeled

by applying the Kriging methodology, based on the current data set. Utilizing the Kriging prediction as well as the prediction error variance, new experiments can be designed for the optimization process. The designed experiments are performed afterwards and the Kriging model is updated. As a Kriging model also takes measurement noise into account, Kriging based optimization can conceptually be applied to both wet-lab and *in silico* experimentation.

Using the Kriging model for designing new experiments at promising regions during the iterative optimization process is in the following called “Kriging Based Design of Experiment” (KBDoE). The aim of KBDoE is to determine a promising sample location that leads to an improvement compared to the current data set. That is, an improvement $\mathrm{I}\big(Z(\boldsymbol{x})\big)$ is achieved if the measured output $Z(\boldsymbol{x})$ at the new designed sample location $\boldsymbol{x}$ is higher than the best output value Z_{max} in the current data set, eq. (35).

$$\mathrm{I}\big(Z(\boldsymbol{x})\big) = \begin{cases} 0 & \text{if } Z(\boldsymbol{x}) < Z_{max} \\ Z_{max} - Z(\boldsymbol{x}) & \text{otherwise} \end{cases} \tag{35}$$

In this section, existing KBDoE approaches are introduced that are the basis for the novel Kriging based optimization strategies, developed in this thesis, see section 1.5 and 1.6. Jones [27] provides a good overview about existing KBDoE approaches. Four main types of KBDoE can be found:

1) maximizing the Kriging prediction
2) maximizing the upper confidence bound
3) maximizing the probability of improvement
4) maximizing the expected improvement.

These maximization tasks can be solved using numerical optimization algorithms, such as gradient based, population based, or Markov-Chain Monte-Carlo based algorithms.

For the first approach, the numerical optimization algorithm is applied directly to the Kriging model prediction and a new experiment is performed at its maximum. However, as illustrated in a schematic example in Figure 4A, this approach can be misleading. The black bold line visualizes the Kriging prediction curve. The maximal prediction value can be found at location x_1. However, three measurements have already been performed at this location. Adding an additional one would be pointless as it would lead most likely to no improvement. In fact, the first approach fails in many cases, i.e. new data points are

designed around the same spot over many iterations and consequently, no improvement is achieved.

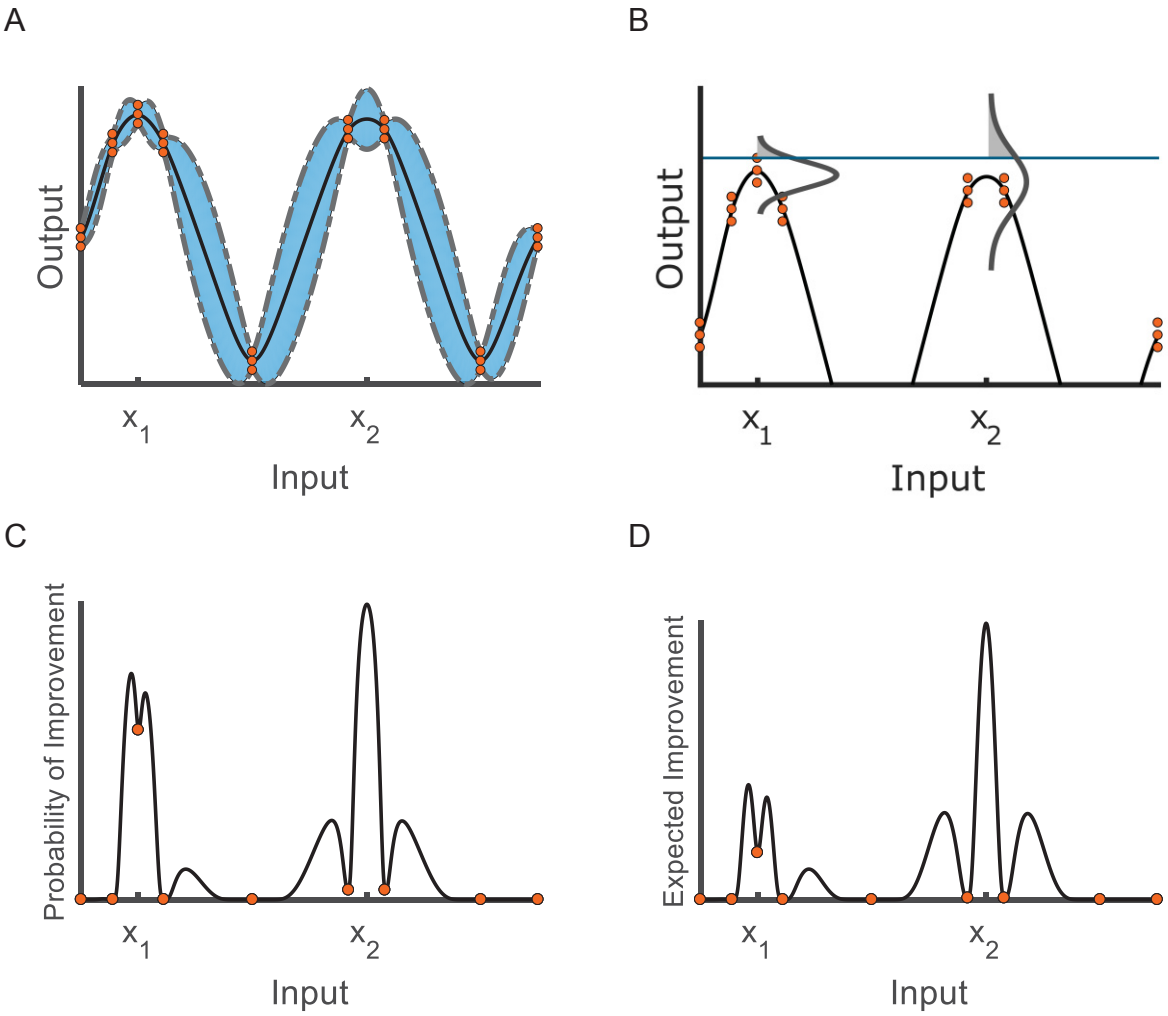


Figure 4: Schematic illustration of A) Kriging approximation. Black line represents Kriging prediction $Z^*(\hat{\boldsymbol{x}})$. The 67% confidence is visualized in blue. B) Probability of improvement calculation. The blue horizontal line is the best so far found output value Z_{max}. Gray curves depict the respective probability density function of the Kriging prediction. C) Probability of improvement at different input locations D) Expected improvement at different input locations. Red dots indicate sample points

In addition to the prediction value, the prediction variance is considered in the second approach by maximizing the upper bound of the Kriging confidence tube. Figure 4A illustrates the 67% confidence tube in blue. The maximum of the upper bound is located at x_2. As the prediction value at x_2 is not maximal but relatively high and the confidence interval is broad, the chance is high that a new measurement at this location will lead to an improvement. However, the outcome of the upper bound maximization depends on

the choice of the confidence level. For instance, in the extreme case of choosing a confidence level of 0%, the second approach is identical to the first approach. In fact, it is generally not a trivial task to define an appropriate confidence level.

A more sophisticated approach is to maximize the probability of improvement (PI). As illustrated in Figure 4B, the probability of an improvement is the integral under the Gaussian curve for output values Z that are bigger than the best output value in the current data set Z_{max}, eq. (36).

$$\mathrm{PI}(\boldsymbol{x}) = \int_{Z>Z_{max}} \mathrm{PDF}_{\boldsymbol{x}}(Z)dZ \tag{36}$$

With $\mathrm{PDF}_{\boldsymbol{x}}(Z)$ as the probability density function of a Gaussian distribution with the expected value $z^*_{\mathrm{UK}}(\hat{\boldsymbol{x}})$, eq. (17), and the variance $\sigma^2_{\mathrm{UK}}(\hat{\boldsymbol{x}})$, eq. (23). As shown in eq. (37), the calculation of PI eq. (36) is related to the calculation of the cumulative probability $\int_{Z<Z_{max}} \mathrm{PDF}_{\boldsymbol{x}}(Z)dZ$. After normalizing the output values, i.e. subtracting $z^*_{\mathrm{UK}}(\hat{\boldsymbol{x}})$ and dividing by $\sigma_{\mathrm{UK}}(\boldsymbol{x})$, the cumulative probability of the standard normal distribution $\Phi()$ can be used instead.

$$\begin{aligned}
\mathrm{PI}(\boldsymbol{x}) &= 1 - \int_{Z<Z_{max}} \mathrm{PDF}_{\boldsymbol{x}}(Z)dZ \\
&= 1 - \Phi\left(\frac{Z_{max} - z^*_{\mathrm{UK}}(\boldsymbol{x})}{\sigma_{\mathrm{UK}}(\boldsymbol{x})}\right) \\
&= \Phi\left(-\frac{Z_{max} - z^*_{\mathrm{UK}}(\boldsymbol{x})}{\sigma_{\mathrm{UK}}(\boldsymbol{x})}\right) \\
&= \Phi\left(\frac{z^*_{\mathrm{UK}}(\boldsymbol{x}) - Z_{max}}{\sigma_{\mathrm{UK}}(\boldsymbol{x})}\right)
\end{aligned} \tag{37}$$

Figure 4B illustrates the calculation of $\mathrm{PI}(\boldsymbol{x})$ at the locations x_1 and x_2, for the schematic example. While at x_1 the expected value is bigger than at x_2, the prediction variance is smaller, leading to a smaller area under $\mathrm{PDF}_{\boldsymbol{x}}(Z)$ for $Z > Z_{max}$. Figure 4C depicts the PI values over the full input range with its maximum value at x_2. These results look promising but Jones [27] points out that a PI based optimization is sometimes very inefficient as, in some cases, an excessively high number of experiments is designed around “good” data points.

Figure 5 demonstrates this effect assuming a scenario in which the first data set on the right hand-side of x_2 is missing. In this case, the prediction curve is already decreasing at x_2 and the PI value at this location will be much smaller than before. Consequently, a location in the near vicinity of x_1 is preferred.

The fourth KBDoE approach is more robust regarding this described scenario. Here, the new experiments are designed by maximizing the chance of achieving a “large” improvement value. This chance is quantified by the expected improvement EI, meaning the integration over the improvement $\mathrm{I}(Z)$ weighted with the probability $\mathrm{PDF}_{\boldsymbol{x}}(Z)$ that the output value Z is achieved, eq. (38).

$$\mathrm{EI}(\boldsymbol{x}) = \int_{Z\in\mathbb{R}} \mathrm{I}(Z)\cdot \mathrm{PDF}_{\boldsymbol{x}}(Z)dZ \tag{38}$$

Similar to eq. (37), after a normalization step, eq. (39), the cumulative distribution function and the probability density function of the standard normal distribution can be used for the calculation of $\mathrm{EI}(\boldsymbol{x})$, eq. (40).

$$u = \frac{z^*_{\mathrm{UK}}(\boldsymbol{x}) - Z_{max}}{\sigma_{\mathrm{UK}}(\boldsymbol{x})} \tag{39}$$

$$\mathrm{EI}(\boldsymbol{x}) = (z^*_{\mathrm{UK}}(\boldsymbol{x}) - Z_{max})\Phi(u) + \phi(u) \tag{40}$$

The derivation of eq. (40) is out of the scope of this thesis and can be found in [28].
Figure 4D and
Figure 5C visualize the EI function over the input space for the schematic example for both scenarios, with the full and reduced data set, respectively. In contrast to PI, the maximum of the EI function is in both scenarios around x_2. This demonstrates that the EI is more robust with respect to variations in the data set.

An EI based optimization, also known as Efficient Global Optimization (EGO) [1], represents in fact the most common used Kriging based optimization strategy. EGO is an iterative optimization procedure, i.e. a Kriging model is constructed based on the current data set. A new experiment is designed at the location that is associated with the maximal EI value. After performing the designed experiment, the new data point is added to the current data set and the Kriging model is updated. This way, the data set as well as the Kriging model prediction accuracy is successively increasing. The optimization stops as soon as the maximal EI value drops under a defined threshold.

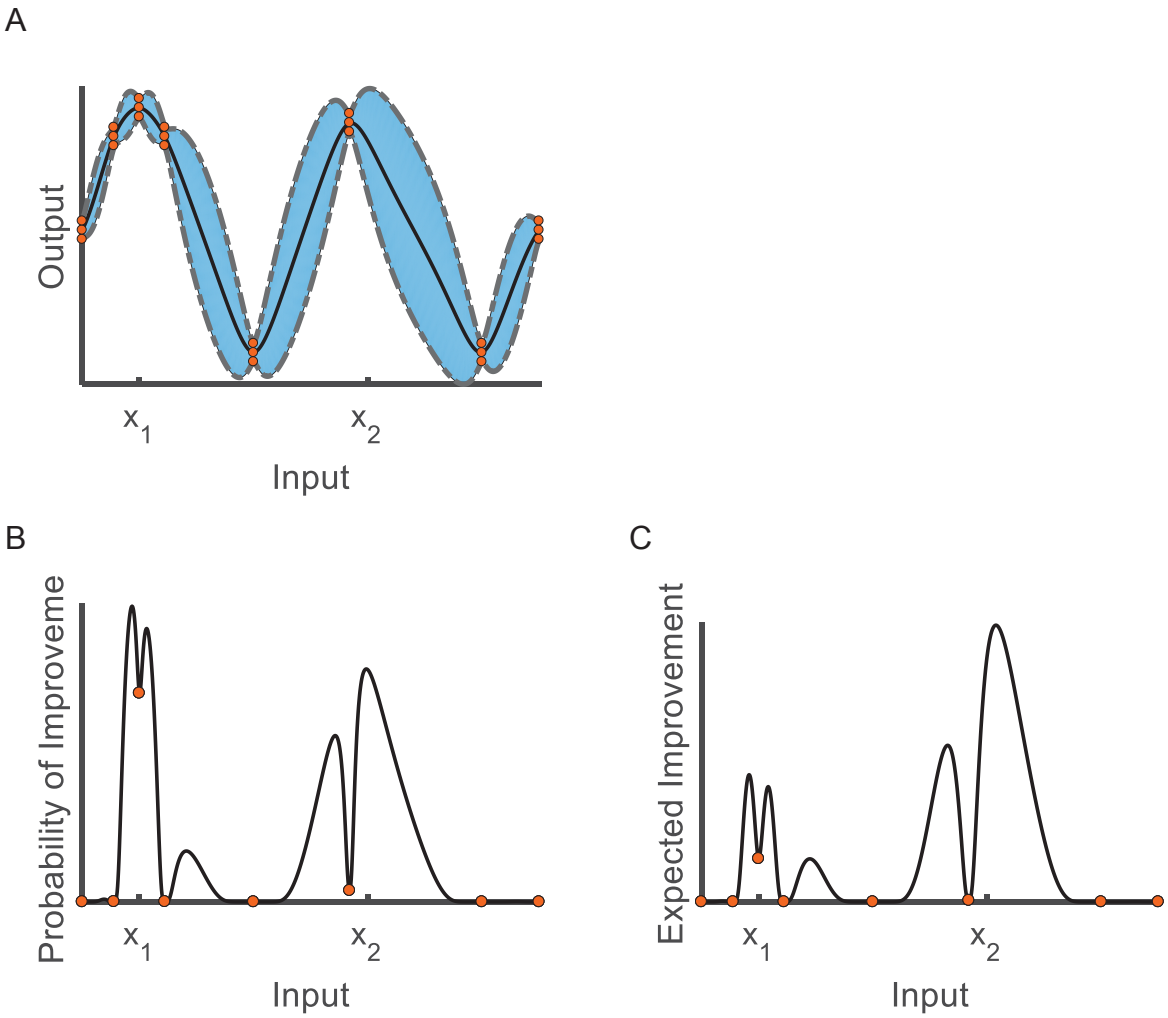


Figure 5: Modified schematic example with one missing data set: A) Kriging approximation B) Probability of improvement C) Expected Improvement.

1.5 Kriging Based Optimization in Biotechnology

The EGO algorithm has proven to be useful for optimizing computer simulations [29]. Here, a mechanistic model is used for simulating the input-output relationship and the optimization goal is to maximize or minimize the model output. Since mechanistic models can be computationally expensive and the EGO algorithm is used for keeping the number of required function evaluations at a minimum. However, three characteristics of the EGO algorithm make it disadvantageous for applications to biotechnological optimization tasks.

First, when designing a biotechnological study, relevant input variables are often not clearly identifiable. Instead, based on experiences and literature research, many input variables are chosen that have a potential effect on the output variable. On the other

hand, the number of possible experiments is limited caused by long experiment duration and by high costs. Under these conditions, an iterative EI based optimization is often not possible, as it requires too many experiments for determining the optimum with sufficient accuracy. Consequently, a prescreening for input variables with significant effects is necessary.

Second, using the EGO algorithm, fixed ranges for the input variables must be given. However, caused by the high complexity of biological systems, it is in general not trivial to identify an appropriate range for the chosen input variables *a priori*. If the input variable ranges are chosen too narrow, the optimum might lie outside of this defined range. If the input variable range is chosen to broad, a high number of experiments are needed to locate the optimum. An appropriate optimization procedure therefore should be flexible in defining these ranges.

A third disadvantage of the EGO algorithm is the fact that only one experiment is designed at each iteration. However, motivated by the long experiment duration, there exists a clear trend towards parallel experimentation in biotechnology [30]–[32].

In conclusion, an appropriated Kriging Based Optimization (KBO) strategy should consider the following three points

1. A sensitivity analysis is performed before entering an iterative optimization cycle
2. During the optimization cycle, the range of the input variables can be adjusted
3. In each iteration, multiple experiments can be designed

A major aim of this thesis is to introduce a KBO strategy that considers all the above-mentioned concerns. Section 1.5.1 focuses on a novel KBDoE approach for designing multiple experiments on the basis of EI, while sections 1.5.2-1.5.3 introduce the developed KBO strategy. The introduced KBO strategy represents rather a framework than a defined algorithm. That is, some steps in the framework require to be manually adapted to the experimental circumstances. For example, in the “Statistic Analysis” step, several tools are mentioned. The experimentalist must decide which of the tools are appropriate. There is no strict rule but rather some advices supporting this decision, see section 1.5.3. The framework consequently provides main instructions for the experimentalist and the data analyst on how to conduct the optimization.

1.5.1 Designing Multiple Experiments Using Expected Improvement and Markov Chain Monte Carlo

As pointed out earlier, in order to be attractive for biotechnology, a KBO strategy must allow parallel optimization, i.e. the design of multiple experiments in each iteration. The original EGO algorithm [1] does not include this feature. The here introduced approach comprises a combination of the concepts of Markov-Chain Monte-Carlo (MCMC) sampling and the expected improvement (EI) based experimental design. Both, MCMC and EI, are already established approaches but, the combination of both concepts for designing multiple experiments is, to the best of my knowledge, novel in its application w.r.t. designing new experiment in the field of biotechnology.

Figure 6A illustrates the EI curve for the same schematic example as used in section 1.4. The curve has a clear maximum that represents the best sample location for a sequential optimization strategy. However, other locations are also associated with high EI values and are consequently promising. Therefore, in case of designing multiple experiments, sample points with high EI values should have a higher chance to be chosen than other points.

In other words, the expected improvement value at location $\boldsymbol{x}$, $\mathrm{EI}(\boldsymbol{x})$, is proportional to the probability density of sampling at this location. This interpretation leads to the MCMC sampling approach that aims at approximating the probability density distribution.

A popular MCMC algorithm is the Metropolis-Hasting algorithm [33]. It is an iterative procedure where in each iteration one sample point is generated. As visualized in Figure 6A, for a sufficient number of iterations, the histogram of sample locations approximates well the actual probability density distribution.

A pseudo code of the Metropolis-Hasting algorithm is given in Figure 6B. The procedure is initialized by generating a sample point $\boldsymbol{x}_0$ from an a priori defined proposal distribution $p(\boldsymbol{x})$. The proposal distribution can for example be a multivariate Gaussian distribution with an expectation point and a covariance matrix that are defined from the measurement data. For each of the following iterations, a new sample is generated following a modified proposal distribution $p(\boldsymbol{x}|\boldsymbol{x}_{i-1})$ by taking into account the last drawn sample location $\boldsymbol{x}_{i-1}$. For example, the expectation point of the multivariate Gaussian distribution might shift to $\boldsymbol{x}_{i-1}$. The new drawn sample point $\boldsymbol{x}'$ is accepted as part of the Markov chain with the probability $p_{accept}(\boldsymbol{x}', \boldsymbol{x}_{i-1})$, eq. (41). If the candidate was

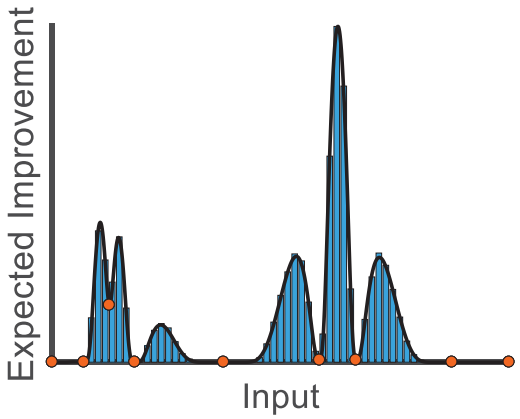
rejected, the proposal distribution stays untouched and a new candidate is drawn. The iterative sampling continues until the MCMC comprises n_{MCMC} locations.

$$p_{accept}(\boldsymbol{x}', \boldsymbol{x}_{i-1}) = \begin{cases} 1 & if\ \mathrm{EI}(\boldsymbol{x}') > \mathrm{EI}(\boldsymbol{x_{i-1}}) \\ \dfrac{\mathrm{EI}(\boldsymbol{x}')/p(\boldsymbol{x}'|\boldsymbol{x}_{i-1})}{\mathrm{EI}(\boldsymbol{x}_{i-1})/p(\boldsymbol{x}_{i-1}|\boldsymbol{x}')} & else \end{cases} \quad (41)$$

A more advanced version of MCMC represents the Delayed Rejection Adaptive Metropolis algorithm (DRAM) [34]. The term “delayed rejection” indicates that rejected candidates have an additional chance of being accepted. The second term “Adaptive Metropolis” refers to the adaptation of the proposal distribution $p(\boldsymbol{x}|\boldsymbol{x}_{i-1})$. In context of DRAM, $p(\boldsymbol{x}|\boldsymbol{x}_{i-1})$ represents a Gaussian distribution with a covariance matrix that is calibrated using the sample path of the MCMC chain.

Multiple experiments can therefore be designed by first approximating the sample distribution using DRAM. Afterwards, an arbitrary number of sample points are drawn uniformly from the Markov chain. Sample locations with high EI values are more often represented in the Markov chain and have therefore a higher chance to be drawn. This way, designed multiple experiments are distributed over the entire input space but concentrated at promising areas. In the publication III and IV (section 2.3-2.4), DRAM in combination with EI could be usefully applied to two case studies.

A



B

```
Initialization: x_0~p(x), i = 1
while i < n_MCMC
    x'~p(x|x_{i-1})
    if r~u(0,1) < p_accept(x', x_{i-1})
       % with r_Random ∈ [0,1]
       x_i = x'
       i = i + 1
    end
end
```

Figure 6: A) Histogram of MCMC sampling for EI. The heights of bars are normalized to the EI values for better visualization. The black line is the EI curve and red dots are the locations of the measurement data used for constructing the Kriging model. B) Pseudo code of the Metropolis-Hasting algorithm.

1.5.2 Designing the Study and Performing the Sensitivity Analysis

The framework, visualized in Figure 7, starts by designing the fundamentals of the optimization study. That is, the experimentalist must choose the objective of interest, input variables with a potential effect on the chosen objective, and an initial range for the input variables. The objective should be easily measurable with sufficient accuracy. If the actual objective is not easily detectable, the experimentalist can think about indirect measurement procedures such as fluorescence detection for the determination of protein concentration. The accuracy and reproducibility of the measurements can often be improved by applying robot automation [35]. The choice of input variables and their initially investigated ranges are usually based on expert knowledge. It is common practice to define a reference experiment to which the optimization results are compared. The defined ranges of the input variables should comprise the reference values.

In general, the experimentalist can find a high number of input variables that potentially affect the measured output. However, limitations in the number of performable experiments allow only the detailed optimization of a limited number of input variables. Therefore, a sensitivity analysis is needed that helps to identify input variables with significant effect. Afterwards, a detailed optimization is performed based on the reduced number of input variables.

Classical design of experiments (DoE) [7] is applicable for this task as it aims at efficiently estimating the main and combinatorial effects of the input variables on the output. This approach was already invented 1935 by Ronald Fisher and is nowadays well established in biotechnology [10]. Classical DoE is based on the estimation of the input-output relationship by a polynomial, eq. (42).

$$Z(x_1, x_2, \dots) = a_0 + a_1 x_1 + a_2 x_2 + \cdots + a_{12} x_1 x_2 + \cdots \qquad (42)$$

The interpretation of the estimated polynomial coefficients $\boldsymbol{a}$ allow the identification of significant main factor effects (e.g. x_1, x_2) and combinatorial effects (e.g. $x_1 \cdot x_2$). In classic DoE, an optimal experimental design refers to the accurate estimation of $\boldsymbol{a}$, i.e. minimization of the estimation error. These optimal experimental designs can be found in textbooks [36] or can be looked up in the internet [37].

As a result, the sensitivity analysis reveals for which input variables a variation inside the defined range has a significant effect on the output variable. In many cases, only a minority of the initially chosen input variables affects the objective of interest in a statistically relevant manner. For the further optimization, the values of these not relevant input variables are set to their reference value.

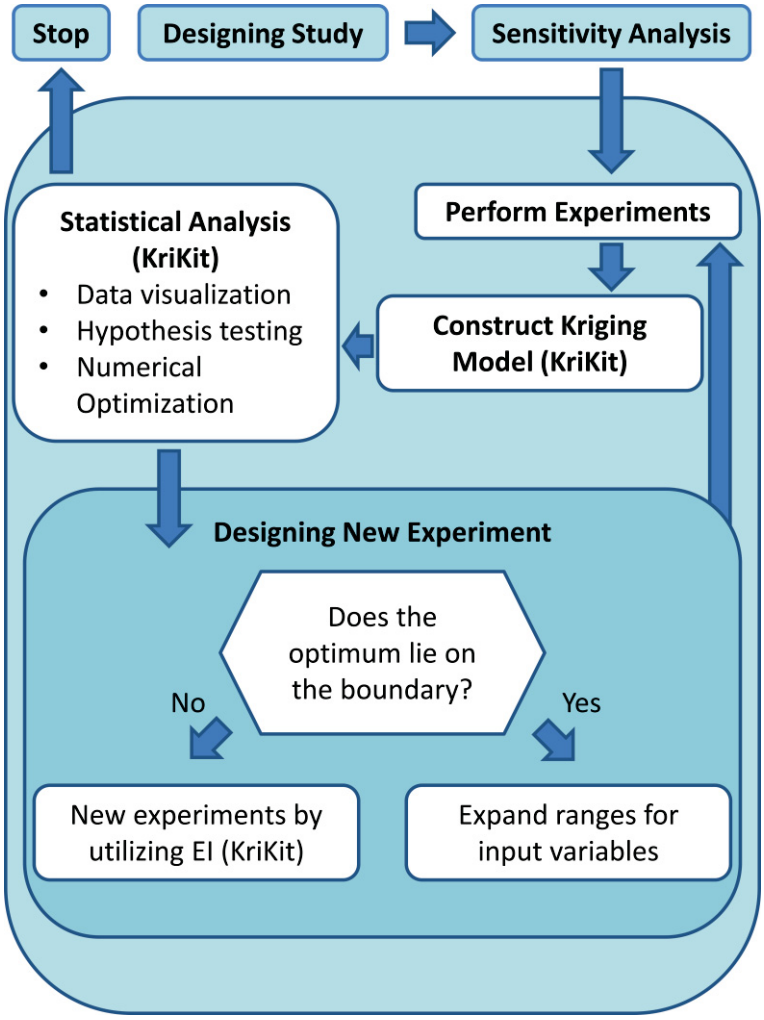


Figure 7: Framework for KBDoE in biotechnology

1.5.3 Iterative Optimization Cycle

After conducting the sensitivity analysis and reducing the number of relevant input variables the Kriging based optimization starts. In the first iteration, initial experiments are conducted. The experimental results from the sensitivity analysis cannot be used for the further optimization as all experiments were performed using the extreme values of the defined range for the input variables. However, during the further study, the not

relevant input variables are fixed to their reference values and consequently, the results are not comparable anymore to the previously performed experiments.

In this context, an optimal design for the initial experiments refers to optimal model prediction using Kriging. In general, it is recommend use a space-filling design. As explained in section 1.3, the Kriging prediction is calculated by a weighted sum of the sample points, eq. (17), whereby sample points in the near environment are stronger weighted than others. A high overall prediction accuracy is achieved by distributing the sample points equally over the hypercube. A broader discussion about good initial experimental design for constructing Kriging models can be found elsewhere [38].

From my experience, it is recommended to plan the initial designs according to the classical DoE approach. These designs are space-filling as the experiments are located in the corners of a hypercube with the axes representing the individual input variables. Consequently, the extreme cases are examined during the first iteration. If more experiments can be performed than the classical DoE suggests, the remaining points should be distributed in a space filling manner using for example Latin hypercube sampling.

After conducting the experiments, the data set is used for constructing a Kriging model, as explained in section 1.3. In order to keep the effort for the experimentalist and the data analyst to a minimum, the Kriging toolKit (KriKit) was developed, which is implemented in MATLAB and is freely available under https://github.com/modsim/KriKit. KriKit not only enables the user to construct a Kriging model, but also provides tools for the effective use of the Kriging model for next step; the statistical analysis. In this context, the most important features comprise data visualization, numerical optimization, and hypothesis testing. The various visualization tools of KriKit support the gain of an intuitive understanding about the functional relationship between the input variables and the measured output. Moreover, potential optima can be found applying local and global optimization algorithms on the Kriging prediction. These potential optima can furthermore be tested for significance using statistical hypothesis testing.

After the statistical analysis was performed, new experiments are designed for the next iteration. Usually, the optimum lies inside the defined range of the input variables and new experiments are designed based on the EI. That is, a numerical optimizer is applied to find the set of input variables that maximizes the EI function. The numerical optimizer

can be a gradient, population based, or MCMC based algorithm. The new experiment is then performed at the identified point. If a parallel experimental strategy is applied, multiple experiments are designed using the KBDoE approach explained in section 1.5.1. The use of MCMC and EI for the experimental designs is an integrated part of KriKit. This multiple experimental design was already successfully applied to a medium optimization of a lipid production process using *Chlorella vulgaris* and is documented in publication III (section 2.3).

It is possible that the initial range of the input variables has been chosen too narrow. In this case, the optimum, identified during the statistical analysis, lies on the boundary of the defined range and the range has to be expanded. New experiments are placed in the not yet considered area using a space filling design, i.e. similar to the initial experimental design in the first iteration. In publication II (section 2.2), the range had to be extended several times. In fact, it was necessary to increase the maximum value of one input variable to the 32-fold of the initial maximum value.

After completing the experimental design, the iteration loop is closed by performing the designed experiments. The optimization proceeds until the maximum number experiments has been reached. Alternatively, the optimization can be stopped as soon as the data density is sufficiently high or the Kriging model prediction is overall accurate enough. For the last criterion, the cross-validation CV is helpful. CV is calculated by the squared difference between Kriging prediction $z^*_{\mathrm{UK}}(\boldsymbol{x}_i)$, after removing the point $\boldsymbol{x}_i$ from the data base, and the measured value $Z(\boldsymbol{x}_i)$. The squared differences are summed over all points, eq. (43).

$$CV = \frac{1}{n}\sum_{i=1}^{n}\left(Z(\boldsymbol{x}_i) - z^*_{\mathrm{UK}}(\boldsymbol{x}_i)\right)^2 \tag{43}$$

1.6 Multi-Objective Gaussian Optimization

In section 1.5, a framework for biotechnological optimization tasks was introduced that aims at exploiting the full potential of Kriging based design of experiment and data analysis. However, the framework is only designed for single-objective optimization problems.

In many biotechnological applications, multiple objectives need to be optimized simultaneously. Typical examples can be found in the field of biocatalysis [39], where the reaction quality is assessed by purity, yield, and productivity. Similar performance indicators are also known from preparative chromatography [40] as well as from fed-batch bioprocess reactors [41]. This proves that there is a demand for effective and efficient multi-objective optimization algorithms.

A scientific goal of this thesis was the development of a multi-objective optimization algorithm, called MOGO, that utilizes the characteristics of Kriging [42]. In contrast to the framework for single-objective optimization in section 1.5, MOGO can be used for automated processes. This comes at the cost that the input ranges have to be fixed over the entire optimization procedure.

In the following, section 1.6.1 provides an overview about the concept of multi-objective optimization. Section 1.6.2-1.6.3 concentrates on already published work in the field of multi-objective KBDoE that represents the basis of MOGO. Details about the MOGO algorithm can be found in publication IV (section 2.4).

1.6.1 Concept of Multi-Objective Optimization

Multi-Objective Optimization (MOO) aims at finding the best compromises between several competing objectives $Z(\boldsymbol{x}) = \left(Z_1(\boldsymbol{x}), \dots, Z_{n_{Obj}}(\boldsymbol{x})\right)^T \in \mathbb{R}^{n_{Obj}}$, $n_{Obj} > 1$. For the sake of simplicity, the optimization problem is in the following considered as maximization problem.

$$\max_{\boldsymbol{x}} Z(\boldsymbol{x}) = \max_{\boldsymbol{x}} \left(Z_1(\boldsymbol{x}), \dots, Z_{n_{Obj}}(\boldsymbol{x})\right) \tag{44}$$

As illustrated in Figure 8, these objectives are competing with each other, meaning that not all objectives can be optimized simultaneously without compromising each other.

The best compromises of a data set $\{X, Z(X)\}$, with the output values $Z(X) = \left(Z(\boldsymbol{x}_1), \dots, Z(\boldsymbol{x}_n)\right)^T \in \mathbb{R}^n$ at the input locations $X = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_n)^T \in \mathbb{R}^{n \times m}$, are known as a set of Pareto optimal points $\mathrm{P} = \left(\boldsymbol{p}_1, \dots, \boldsymbol{p}_{n_{Pareto}}\right)^{\mathrm{T}} \in \mathbb{R}^{n_{Pareto}}$ with $\forall \boldsymbol{p} \in Z(X)$. In context of Pareto optimization a point $\boldsymbol{u} = (u_1, \dots, u_k)$ dominates a point $\boldsymbol{v} = (v_1, \dots, v_k)$, denoted by $\boldsymbol{u} < \boldsymbol{v}$, iff $\forall i \in \{1, \dots, k\}$, $u_i < v_i$ and $\boldsymbol{u} \neq \boldsymbol{v}$. A measured output that is not a member of the Pareto set, $Z(\boldsymbol{x}) \notin \mathrm{P}$, is dominated by at least one point in P.

$$Z(\boldsymbol{x}) < \mathrm{P}, \text{ if } Z(\boldsymbol{x}) \notin \mathrm{P}, \text{ with } Z(\boldsymbol{x}) \in Z(X) \tag{45}$$

Vice versa, for each Pareto optimal point $Z(\boldsymbol{x}) \in \mathrm{P}$, it holds true that there is no point in the overall data set $Z(X)$ that is better in all objectives.

$$Z(\boldsymbol{x}) \prec Z(X), \text{ with } Z(\boldsymbol{x}) \in \mathrm{P} \tag{46}$$

Conceptually, there exists a curve or surface where all Pareto optimal points are located, called the Pareto front $\mathrm{P}_{\mathrm{front}}$. A multi-objective optimization algorithm aims at approximating $\mathrm{P}_{\mathrm{front}}$ and comprises in general an iterative optimization procedure. Starting with an initial Pareto front approximation $\mathrm{P}_{i=1}$, the approach will converge towards $\mathrm{P}_{\mathrm{front}}$, eq. (47).

$$\lim_{i \to \infty} \mathrm{P}_{\mathrm{i}} \to \mathrm{P}_{\mathrm{front}} \tag{47}$$

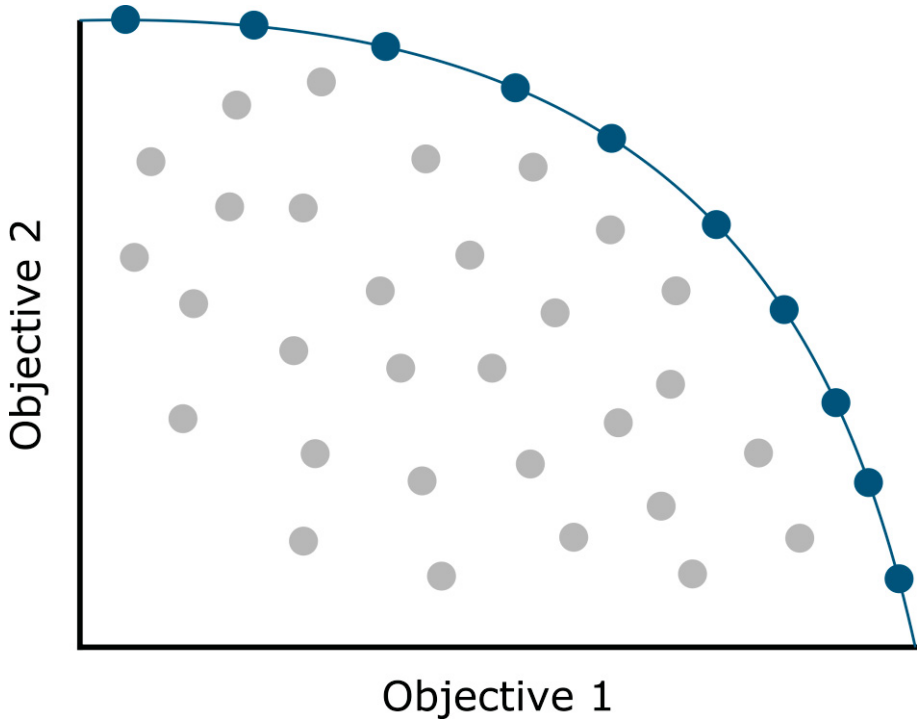


Figure 8: Schematic illustration of a Pareto optimization problem. Gray and blue dots indicate, respectively, the non-optimal and Pareto optimal solutions. The blue curve represents the “true” Pareto front P_{true}.

1.6.2 Multi-Objective KBDoE

The vast majority of MOO algorithms belongs to the class of Multi-Objective Evolutionary Algorithms (MOEA). Zhou et al. [43] and Lücken et al. [44] provide comprehensive surveys about MOEA. The most popular algorithm of this class is the Non-dominated Sorting Genetic Algorithm II (NSGA-II) [45]. Although MOEAs are very effective, the high number of experiments they need for convergence makes them impracticable for many biotechnological applications. The experimental effort can be reduced by using surrogate models, such as Kriging [46]. For example, Li et al. [47] could reduce the experimental effort for the NSGA-II by partially replacing the costly experiment by the Kriging model prediction. Here, the Kriging model prediction is accepted if the associated prediction error is lower than a defined threshold.

A major aim of this thesis is the development of the Multi-Objective Global Optimization (MOGO) algorithm, publication IV (section 2.4). MOGO is based on only recently published state-of-the-art methods that allow the transfer of the established EGO algorithm [1] to the multi-objective case.

As explained in section 1.4, EGO is the most common KBDoE approach for single-objection optimization. Using EGO, new experiments are planned based on the Expected Improvement (EI) and the optimization stops as soon as the Kriging model is reliable enough. The transfer of these two key parts to MOO was only accomplished in the recent years. While in 2006, Emmerich et al. already have developed the concept of Expected HyperVolume Improvement (EHVI) as equivalent to EI [48], only 2015, Hupkens et al. [49] provided an algorithm that allows the fast calculation of EHVI. Also in 2015, Binois et al. [50] published an approach for quantifying the prediction uncertainty of the Kriging based Pareto front estimation. This quality indicator is successfully used in publication IV for detecting convergence (section 2.4). While in publication IV the procedure for quantifying the prediction uncertainty is derived and discussed in detail, the derivation of the EHVI is held compact there and is discussed in more detail in section 1.6.3.

MOGO [42] comprises these recently developed concepts as well as the Markov chain Monte Carlo based approach for designing multiple experiments, introduced in section 1.5.1. This makes MOGO an effective and efficient algorithm for automated parallel multi-objective optimization.

1.6.3 Expected Hypervolume Improvement

The Expected HyperVolume Improvement (EHVI) is the multi-objective equivalent to the Expected Improvement (EI). The EI quantifies the potential of achieving an improvement by adding an additional data point to the current data set. As the term indicates, in context of EHVI, the HyperVolume (HV) is used for quantifying the improvement during the MOO. For the calculation of the HV, it is first necessary to determine the Pareto set $\mathrm{P} = \left(\boldsymbol{p}_1, \dots, \boldsymbol{p}_{n_{Pareto}}\right)^{\mathrm{T}} \in \mathbb{R}^{n_{Pareto}}$ and to define a reference point $\boldsymbol{r} = \left(r_1, \dots, r_{n_{Obj}}\right)^{T} \in \mathbb{R}^{n_{Obj}}$. P is the set of Pareto optimal points of the current data set and $\boldsymbol{r}$ is dominated by all points in P. The HV measures the volume V of the subspace between P and r.

$$\mathrm{HV}(\mathrm{P}, \boldsymbol{r}) \colon V\left(\bigcup_{\boldsymbol{p}_i \in \mathrm{P}} space(\boldsymbol{p}_i, \boldsymbol{r})\right), \text{with } \boldsymbol{r} < \boldsymbol{p}_i, \forall i \in \{1, \dots, n_{Pareto}\} \tag{48}$$

With $space(\boldsymbol{p}_i, \boldsymbol{r})$ representing a 2-D rectangles or a hyper-rectangles with $\boldsymbol{p}_i$ and $\boldsymbol{r}$ located in the corners and V is the Lebesgue measure.

As visualized in Figure 9A, an improvement during the MOO is achieved if the added data point either completes the current Pareto set P, such as $Z^{(1)}$ does, or replaces members of P, as in case of $Z^{(2)}$. In both cases, after adding the new data point Z, the HV is extended and the extension represents the associated improvement $\mathrm{I}(Z)$, eq. (49).

$$\mathrm{I}(Z) = \mathrm{HV}(\mathrm{P} \cup Z, \boldsymbol{r}) - \mathrm{HV}(\mathrm{P}, \boldsymbol{r}) \tag{49}$$

In general, MOO represents an iterative optimization procedure where in each iteration new data points are added and the Pareto front is updated. During the iterative procedure, the approximated Pareto front P_i will converge to the “true” Pareto front $\mathrm{P}_{\mathrm{true}}$, eq. (50), and the HV is increasing monotonically, eq. (51).

$$\lim_{i \to \infty} \mathrm{P}_i = \mathrm{P}_{\mathrm{true}} \tag{50}$$

$$\mathrm{HV}(\mathrm{P}_i, \boldsymbol{r}) \geq \mathrm{HV}\left(\mathrm{P}_j, \boldsymbol{r}\right) \quad \forall i \geq j \tag{51}$$

In fact, the hypervolume reaches its maximum value if P_i converges to the “true” Pareto front P_{true} [51], eq. (52).

$$\lim_{i \to \infty} \mathrm{HV}(\mathrm{P}_i, \boldsymbol{r}) = \mathrm{HV}(\mathrm{P}_{\mathrm{true}}, \boldsymbol{r}) \tag{52}$$

In analogy to the single-objective EI, EHVI is defined as the integration of the improvement $\mathrm{I}(Z)$ weighted by the probability that the improvement value will be achieved at point $\boldsymbol{x}$, eq. (53).

$$\mathrm{EHVI}(\boldsymbol{x}) = \int_{Z \in \mathbb{R}^{n_{Obj}}} I(Z) \cdot \mathrm{PDF}_{\boldsymbol{x}}(Z) dZ \tag{53}$$

$\mathrm{PDF}_{\boldsymbol{x}}(Z)$ represents a multivariate Gaussian distribution originating from the Kriging models for the individual objectives with the expected point $\boldsymbol{\mu} = \left(\mu_1, \dots, \mu_{n_{Obj}}\right)^T \in \mathbb{R}^{n_{Obj}}$, eq. (54), and the covariance matrix $\Sigma \in \mathbb{R}^{n_{Obj} \times n_{Obj}}$, eq. (55).

$$\mu = \begin{bmatrix} z_{\mathrm{UK}}^{(1)}(\boldsymbol{x}) \\ \vdots \\ z_{\mathrm{UK}}^{(n_{Obj})}(\boldsymbol{x}) \end{bmatrix} \tag{54}$$

$$\Sigma = \begin{bmatrix} {\sigma_{\mathrm{UK}}^{(1)}}^2(\boldsymbol{x}) & & \\ & \ddots & \\ & & {\sigma_{\mathrm{UK}}^{(n_{Obj})}}^2(\boldsymbol{x}) \end{bmatrix} \tag{55}$$

$z_{\mathrm{UK}}^{(i)}(\boldsymbol{x})$ is the Kriging prediction value, eq. (17), and ${\sigma_{\mathrm{UK}}^{(i)}}^2(\boldsymbol{x})$ is the Kriging variance, eq. (23), of ith objective at location $\boldsymbol{x}$. As eq. (55) indicates, it is assumed that all Kriging models are independent although it might be useful to overcome this assumption if the objectives do influence each other. However, to the best of my knowledge, there is currently no analytical calculation of EHVI published for this scenario.

As depicted in Figure 9B, algorithms for the calculation of EHVI partition the output space into rectangles, or hyper-rectangles in case of more than two objectives. For each hyper-rectangle, the associated contribution to the EHVI can be calculated individually and summed up afterwards. Based on this ansatz, several algorithms exist that mainly differ in how the rectangles are defined. For example, the efficiency of the algorithm can be increased by choosing only rectangles that are not dominated and consequently have an EHVI contribution larger than zero. In publication IV (section 2.4), an algorithm was used that is in particularly efficient for three objectives [49], but it is noteworthy that very recently an algorithm was published that handles the bi-objective case more efficiently [52].

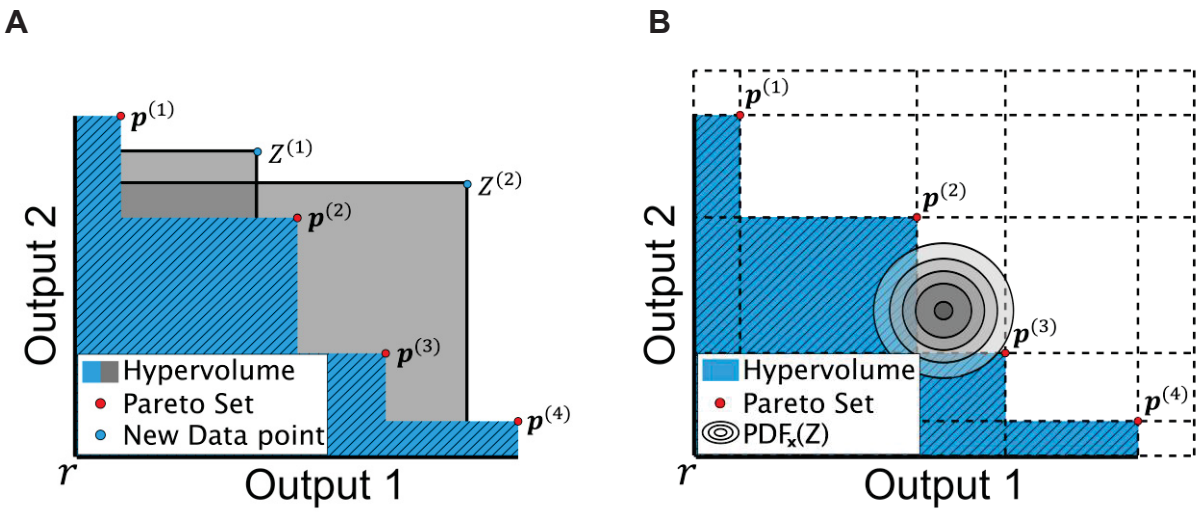


Figure 9: Visualization of A) HV associated with an extended data set and B) Calculation of the EHVI. Red dots indicate Pareto optimal points of the current data set and the blue area the associated HV. Gray area indicates the HV extension after adding the respective blue dots to the data sets. Dashed lines indicate the separation of the output space into rectangles for efficient calculation of the EHVI.

1.7 References

[1] Jones, D. R., Schonlau, M., and William, J., Efficient Global Optimization of Expensive Black-Box Functions, 1998, 455–492.

[2] von Lieres, E. and Andersson, J., A Fast and Accurate Solver for the General Rate Model of Column Liquid Chromatography, *Comput. Chem. Eng.*, 2010, 34, 1180–1191.

[3] Westerwalbesloh, C., Grünberger, A., Stute, B., Weber, S., *et al.*, Modeling and CFD Simulation of Nutrient Distribution in Picoliter Bioreactors for Bacterial Growth Studies on Single-Cell Level, *Lab Chip*, 2015, 15, 4177–4186.

[4] Mendes, P. and Kell, D., Non-Linear Optimization of Biochemical Pathways: Applications to Metabolic Engineering and Parameter Estimation, *Bioinformatics*, 1998, 14, 869–883.

[5] Straathof, A. J. J., Rakels, J. L. L., and Heijnen, J. J., Kinetics of the Enzymatic Resolution of Racemic Compounds in Bi-Bi Reactions, *Biocatal. Biotransformation*, 1992, 7, 13–27.

[6] Franceschini, G. and Macchietto, S., Model-Based Design of Experiments for Parameter Precision: State of the Art, *Chem. Eng. Sci.*, 2008, 63, 4846–4872.

[7] Fisher, R. A., The Design of Experiments. Oliver and Boyd, Edinborough 1935.

[8] Bhadeshia, H. K. D. H., Neural Networks in Materials Science., *ISIJ Int.*, 1999, 39, 966–979.

[9] Cressie, N. A. C., Statistics for Spatial Data, 3rd ed. Wiley New York, 1993.

[10] Mandenius, C.-F. and Brundin, A., Bioprocess Optimization Using Design-of-Experiments Methodology, *Biotechnol. Prog.*, 2008, 24, 1191–1203.

[11] Simpson, T. W., Mauery, T. M., Korte, J. J., and Mistree, F., Kriging Models for Global Approximation in Simulation-Based Multidisciplinary Design Optimization, *AIAA J.*, 2001, 39, 2233–2241.

[12] Cock, D. R. De, Kriging as an Alternative to Polynomial Regression in Response Surface Analysis(doctoral thesis), Iowa State University, 2003.

[13] Jeong, S., Murayama, M., and Yamamoto, K., Efficient Optimization Design Method Using Kriging Model, *J. Aircr.*, 2005, 42, 413–420.

[14] Bellary, S. A. I., Samad, A., Couckuyt, I., and Dhaene, T., A Comparative Study of Kriging Variants for the Optimization of a Turbomachinery System, *Eng. Comput.*,

2016, 32, 49–59.

[15] Gatti, M. N., Milocco, R. H., and Giaveno, A., Modeling the Bacterial Oxidation of Ferrous Iron with Acidithiobacillus Ferrooxidans Using Kriging Interpolation, *Hydrometallurgy*, 2003, 71, 89–96.

[16] Bonowski, F., Kitanovic, A., Ruoff, P., Holzwarth, J., *et al.*, Computer Controlled Automated Assay for Comprehensive Studies of Enzyme Kinetic Parameters, *PLoS One*, 2010, 5, e10727.

[17] Krige, D. G., A Statistical Approach to Some Basic Mine Valuation Problems on the Witwatersrand, *J. South. African Inst. Min. Metall.*, 1951, 52, 119 – 139.

[18] Matheron, G., The Theory of Regionalized Variables and Its Applications. Ecole Nationale Superieure des Mines, Paris 1971.

[19] Solow, A. R., Mapping by Simple Indicator Kriging, *Math. Geol.*, 1986, 18, 335–352.

[20] Omre, H., Bayesian Kriging - Merging Observations and Qualified Guesses in Kriging, *Math. Geol.*, 1987, 19, 25–39.

[21] Rasmussen, C. E., Gaussian Processes for Machine Learning. MIT Press, 2006.

[22] Marchant, B. P. and Lark, R. M., The Matérn Variogram Model: Implications for Uncertainty Propagation and Sampling in Geostatistical Surveys, *Geoderma*, 2007, 140, 337–345.

[23] Mardia, K. V and Marshall, R. J., Maximum Likelihood Estimation of Models for Residual Covariance in Spatial Regression, *Biometrika*, 1984, 71, 135–146.

[24] Morris, E. L., Chapter 3: The Normal Distribution on a Vector Space, in: *Multivariate Statistics*, Institute of Mathematical Statistics, Beachwood, Ohio, USA 2007, pp. 103–131.

[25] Lu, T. and Shiou, S., Inverses of 2x2 Block Matrices, *Comput. Math. with Appl.*, 2000, 43, 119–129.

[26] Cressie, N. A. C., Covariogram and Correlogram, in: *Statistics for spatial data*, 3rd ed., Wiley New York, 1993, pp. 67–68.

[27] Jones, D. R., A Taxonomy of Global Optimization Methods Based on Response Surfaces, *J. Glob. Optim.*, 2001, 21, 345–383.

[28] Schonlau, M. and Welch, W. J., Global Optimization with Nonparametric Function Fitting, in *Proceedings of the American Statistical Association, Section on Physical*

and Engineering Sciences, 1996, 183–186.

[29] Sacks, J., Welch, W. J., Mitchell, T. J., and Wynn, H. P., Design and Analysis of Computer Experiments, *Stat. Sci.*, 1989, 4, 409–423.

[30] Gernaey, K. V., Baganz, F., Franco-Lara, E., Kensy, F., *et al.*, Monitoring and Control of Microbioreactors: An Expert Opinion on Development Needs, *Biotechnol. J.*, 2012, 7, 1308–1314.

[31] Baumann, P., Hahn, T., and Hubbuch, J., High-Throughput Micro-Scale Cultivations and Chromatography Modeling: Powerful Tools for Integrated Process Development, *Biotechnol. Bioeng.*, 2015, 112, 2123–2133.

[32] Jacques, P., Béchet, M., Bigan, M., Caly, D., *et al.*, High-Throughput Strategies for the Discovery and Engineering of Enzymes for Biocatalysis, *Bioprocess Biosyst. Eng.*, 2017, 40, 161–180.

[33] HASTINGS, W. K., Monte Carlo Sampling Methods Using Markov Chains and Their Applications, *Biometrika*, 1970, 57, 97–109.

[34] Haario, H., Laine, M., Mira, A., and Saksman, E., DRAM: Efficient Adaptive MCMC, *Stat. Comput.*, 2006, 16, 339–354.

[35] Rohe, P., Venkanna, D., Kleine, B., Freudl, R., and Oldiges, M., An Automated Workflow for Enhancing Microbial Bioprocess Optimization on a Novel Microbioreactor Platform, *Microb. Cell Fact.*, 2012, 11, 144.

[36] Myers, R. H., Montgomery, D. C., and Anderson-Cook, C. M., Response Surface Methodology: Process and Product Optimization Using Designed Experiments, 3rd ed. Wiley, New York 2009.

[37] National Institute of Standards and Technology, Summary Tables of Useful Fractional Factorial Designs.

[38] Huang, D., Experimental Planning and Sequential Kriging Optimization Using Variable Fidelity Data(doctoral thesis), The Ohio State University, 2005.

[39] Baraibar, Á. G., von Lieres, E., Wiechert, W., Pohl, M., and Rother, D., Effective Production of (S)-α-Hydroxy Ketones: An Reaction Engineering Approach, *Top. Catal.*, 2014, 57, 401–411.

[40] Degerman, M., Westerberg, K., and Nilsson, B., A Model-Based Approach to Determine the Design Space of Preparative Chromatography, *Chem. Eng. Technol.*, 2009, 32, 1195–1202.

[41] Sarkar, D. and Modak, J. M., Pareto-Optimal Solutions for Multi-Objective Optimization of Fed-Batch Bioreactors Using Nondominated Sorting Genetic Algorithm, *Chem. Eng. Sci.*, 2005, 60, 481–492.

[42] Freier, L. and von Lieres, E., Multi-Objective Global Optimization (MOGO): Algorithm and Case Study in Gradient Elution Chromatography, *Biotechnol. J.*, 2017, 1600613.

[43] Zhou, A., Qu, B.-Y., Li, H., Zhao, S.-Z., *et al.*, Multiobjective Evolutionary Algorithms: A Survey of the State of the Art, *Swarm Evol. Comput.*, 2011, 1, 32–49.

[44] von Lücken, C., Barán, B., and Brizuela, C., A Survey on Multi-Objective Evolutionary Algorithms for Many-Objective Problems, *Comput. Optim. Appl.*, 2014, 58, 707–756.

[45] Deb, K., Pratap, A., Agarwal, S., and Meyarivan, T., A Fast and Elitist Multiobjective Genetic Algorithm: NSGA-II, *IEEE Trans. Evol. Comput.*, 2002, 6, 182–197.

[46] Ong, Y. S., Nair, P. B., and Keane, A. J., Evolutionary Optimization of Computationally Expensive Problems via Surrogate Modeling, *AIAA J.*, 2003, 41, 687–696.

[47] Li, M., Li, G., and Azarm, S., A Kriging Metamodel Assisted Multi-Objective Genetic Algorithm for Design Optimization, *J. Mech. Des.*, 2008, 130, 031401.

[48] Emmerich, M. T. M., Giannakoglou, K. C., and Naujoks, B., Single-Objective and Multiobjective Evolutionary Optimization Assisted by Gaussian Random Field Metamodels, *IEEE Trans. Evol. Comput.*, 2006, 10, 421–439.

[49] Hupkens, I., Deutz, A., Yang, K., and Emmerich, M., Faster Exact Algorithms for Computing Expected Hypervolume Improvement, in: A. Gaspar-Cunha, C. Henggeler Antunes, and C. C. Coello, (Eds.), *Evolutionary Multi-Criterion Optimization*, 9019, Springer International Publishing, Cham 2015, pp. 65–79.

[50] Binois, M., Ginsbourger, D., and Roustant, O., Quantifying Uncertainty on Pareto Fronts with Gaussian Process Conditional Simulations, *Eur. J. Oper. Res.*, 2015, 243, 386–394.

[51] Fleischer, M., The Measure of Pareto Optima. Applications to Multi-Objective Metaheuristics, in *Evolutionary Multi-Criterion Optimization. Second International*

Conference, EMO 2003, 2003, 519–533.

[52] Emmerich, M., Yang, K., Deutz, A., Wang, H., and Fonseca, C. M., A Multicriteria Generalization of Bayesian Global Optimization, in: *Springer Optimization and Its Applications*, 107, 2016, pp. 229–242.

[53] Chromatography Analysis and Design Toolkit(CADET).

[54] Tan, K. C., Lee, T. H., and Khor, E. F., Evolutionary Algorithms with Dynamic Population Size and Local Exploration for Multiobjective Optimization, *IEEE Trans. Evol. Comput.*, 2001, 5, 565–588.

2 Publications

2.1 Kriging with nonlinear trend functions: Theory and application in enzyme kinetics

Engineering in Life Sciences

Eng. Life Sci. **2017**, *17*, 916–922 www.els-journal.com

Lars Freier
Wolfgang Wiechert
Eric von Lieres

IBG-1: Biotechnology, Forschungszentrum Jülich, Jülich, Germany

Research Article

Kriging with trend functions nonlinear in their parameters: Theory and application in enzyme kinetics

Kriging is an interpolation method commonly applied in empirical modeling for approximating functional relationships between impact factors and system response. The interpolation is based on a statistical analysis of given data and can optionally include a priori defined trend functions. However, Kriging can so far only be used with trend functions that are linear with respect to the parameters. In this contribution, we present an extension of the Kriging approach for handling trend functions that are nonlinear in their parameters. Our approach is based on Taylor linearization combined with an iterative parameter estimation procedure whose solution is practically computed via a root finding problem. We demonstrate our novel approach with measurement data from the application field of biocatalysis.

Keywords: Enzyme kinetics / Kriging / Nonlinear trend function

Received: February 3, 2017; *revised:* June 13, 2017; *accepted:* July 3, 2017

DOI: 10.1002/elsc.201700022

1 Introduction

In chemistry and biochemistry, technical production processes are often to be optimized with respect to the maximization of product yield, productivity and/or selectivity. Such processes typically depend on operating conditions such as solvent concentrations, pH-values and temperature. These tunable factors often have unclear and complex impacts on system performance.

Empirical models, for instance response surface methodology (RSM), are commonly applied to approximate relationships between impact factors and system output. In RSM, these relations are mathematically described by polynomials whose coefficients are typically estimated by least squares fitting of the model to experimental data. The articles [1] and [2] provide good overviews of the use of RSM in chemistry and biochemistry.

Alternatively, mechanistic models can be used for describing the dependency of the system output on input factors. The mathematical structure of mechanistic models is based on a priori knowledge. For example, the Michaelis–Menten equation describes the impact of substrate concentration on the reaction rate of enzymes. Compared to empirical models, appropriately chosen mechanistic models generally have the advantage of providing more reliable predictions. On the other hand, choosing appropriate model structures can be challenging. Moreover, parameter estimation for nonlinear models typically requires computationally expensive global optimization.

Correspondence: Dr. Eric von Lieres (e.von.lieres@fz-juelich.de), IBG-1: Biotechnology, Forschungszentrum Jülich, 52428 Jülich, Germany

Kriging is a statistical interpolation method which allows integrating both empirical and mechanistic modeling approaches. While Kriging is commonly applied in geostatistics and fluid dynamics [3, 4], the method is only recently recognized in biotechnology [5,6]. The empirical part of Kriging analyzes first how the covariance of given measurement data depends on the distance of the respective measurement points. Based on this statistical information, a smooth prediction curve/surface and the associated confidence tube is calculated. In addition, basic trends can be incorporated by mechanistic models which allows, to some extent, to extrapolate over the range of data used for model calibration.

However, the standard Kriging framework only allows linear trend functions in their parameters but the majority of mechanistic models used in biochemistry are nonlinear. For instance, the well-known Michaelis–Menten equation (1) is linear with respect to its maximal conversion rate $V_{\max}$ but, nonlinear with respect to the second parameter K_m which describes the substrate concentration at half saturation ($v = V_{\max}/2$)

$$v = \frac{V_{\max}\,[S]}{K_m + [S]}. \tag{1}$$

The next sections are structured as following: First, Sections 2.2 and 2.3 introduce the theoretical background of the standard Kriging framework that is illustrated by example data sampled from the Michaelis–Menten equation (1). In Section 2.4, the parameter estimation procedure for linear trend functions in the Kriging framework is described. In Section 3, our Kriging based interpolation approach with nonlinear trend functions is introduced in two steps. At first, the nonlinear parameters are

determined by fixed point iteration using Taylor based linearization [7]. In the second step, this approach is extended by reformulating the parameter estimation problem into a root-finding problem. Finally, in Section 3.1, the Kriging approach with nonlinear trend functions is demonstrated using the two-substrate Michaelis–Menten equation.

Kriging with nonlinear trend functions should not be confused with disjunctive Kriging, which is also referred to as nonlinear Kriging [8,9]. Disjunctive Kriging describes the measurement data using a nonlinear estimator, while our approach preserves the linear characteristics of the predictor but integrates trend functions that are nonlinear with respect to their parameters.

2 Materials and methods

2.1 Universal Kriging

Kriging, also referred to as Gaussian process regression [10], is a statistical approach for interpolating functional relationships. Its origins goes back to the work of Krige (1951) [11]. This section provides a brief introduction to Kriging, further details can be found in [12]. In general, a Kriging prediction is based on n observations $Z(\boldsymbol{x}_1), \ldots, Z(\boldsymbol{x}_n)$ each of which depends on m input variables, $\boldsymbol{x}_i \in \mathbb{R}^m$. Each observation $Z(\boldsymbol{x}) \in \mathbb{R}$ is assumed to follow a Gaussian process that can be decomposed into a deterministic trend function $m(\boldsymbol{x}) \in \mathbb{R}$ and a random function $Y(\boldsymbol{x}) \in \mathbb{R}$ with zero mean and standard deviation $\sigma(\boldsymbol{x}) \in \mathbb{R}$

$$Z(\boldsymbol{x}) = m(\boldsymbol{x}) + Y(\boldsymbol{x}) \tag{2}$$

$$\mathrm{E}[Y(\boldsymbol{x})] = 0 \tag{3}$$

$$\sigma(\boldsymbol{x}) = \sqrt{\mathrm{Var}(Y(\boldsymbol{x}))}. \tag{4}$$

Basic trends can be described by a linear combination of k functions $f_l(\boldsymbol{x}) \in \mathbb{R}$ with coefficients a_l

$$m(\boldsymbol{x}) = \sum_{l=1}^{k} a_l f_l(\boldsymbol{x}). \tag{5}$$

Kriging is a linear smoother [13] computing predictions Z^* of unknown observations at points $\hat{\boldsymbol{x}}$ by a linear combination of the given data $Z(\boldsymbol{x}_i)$ with coefficients $\lambda_i \in \mathbb{R}$

$$Z^*(\hat{\boldsymbol{x}}) = \sum_{i=1}^{n} \lambda_i(\hat{\boldsymbol{x}}) Z(\boldsymbol{x}_i). \tag{6}$$

In the Kriging context, the coefficients λ_i, which depend on $\hat{\boldsymbol{x}}$, are determined such as to provide unbiased predictions (7) and to minimize the prediction variance (8)

$$\mathrm{E}\left[Z(\hat{\boldsymbol{x}}) - Z^*(\hat{\boldsymbol{x}})\right] = 0 \tag{7}$$

$$\mathrm{Var}\left[Z(\hat{\boldsymbol{x}}) - Z^*(\hat{\boldsymbol{x}})\right] \to \min. \tag{8}$$

Using (2) and (6), the Kriging unbiasedness condition (7) can be expressed as

$$\begin{aligned}\mathrm{E}\left[Z(\hat{\boldsymbol{x}}) - Z^*(\hat{\boldsymbol{x}})\right] &= \mathrm{E}\left[Z(\hat{\boldsymbol{x}}) - \sum_{i=1}^{n} \lambda_i Z(\boldsymbol{x}_i)\right] \\ &= \mathrm{E}\left[m(\hat{\boldsymbol{x}}) + Y(\hat{\boldsymbol{x}}) - \sum_{i=1}^{n} \lambda_i (m(\boldsymbol{x}_i) + Y(\boldsymbol{x}_i))\right] \\ &= \mathrm{E}\left[m(\hat{\boldsymbol{x}}) - \sum_{i=1}^{n} \lambda_i m(\boldsymbol{x}_i)\right] = 0. \end{aligned} \tag{9}$$

Inserting the trend functions (5), this can be further reformulated to

$$\mathrm{E}[Z(\hat{\boldsymbol{x}}) - Z^*(\hat{\boldsymbol{x}})] = \mathrm{E}\left[\sum_{l=1}^{k} a_l f_l(\hat{\boldsymbol{x}}) - \sum_{i=1}^{n} \lambda_i \sum_{l=1}^{k} a_l f_l(\boldsymbol{x}_i)\right] = 0. \tag{10}$$

A sufficient condition for (10) to hold is

$$a_l f_l(\hat{\boldsymbol{x}}) = \sum_{i=1}^{n} \lambda_i a_l f_l(\boldsymbol{x}_i) \text{ with } l = 1, \ldots k. \tag{11}$$

The Kriging prediction variance furthermore is given by (12) [14]:

$$\begin{aligned}\mathrm{Var}[Z(\boldsymbol{x}) - Z^*(\boldsymbol{x})] = {} & \mathrm{Cov}(Z(\hat{\boldsymbol{x}}), Z(\hat{\boldsymbol{x}})) \\ & -2\sum_{i=1}^{n} \lambda_i \mathrm{Cov}(Z(\boldsymbol{x}_i), Z(\hat{\boldsymbol{x}})) \\ & +\sum_{i=i}^{n}\sum_{j=1}^{n} \lambda_i \lambda_j \mathrm{Cov}\left(Z(\boldsymbol{x}_i), Z\left(\boldsymbol{x}_j\right)\right). \end{aligned} \tag{12}$$

Coefficients λ_i that satisfy the fundamental Kriging conditions (7) and (8) are calculated by minimizing the prediction variance given by (12) under the linear constraints posed by (11) using the method of Lagrange multipliers. This leads to the following system of linear equations:

$$\begin{aligned}\sum_{i=1}^{n} \lambda_i \mathrm{Cov}\left(Z(\boldsymbol{x}_i), Z\left(\boldsymbol{x}_j\right)\right) + \sum_{l=1}^{k} \mu_l f_l(\boldsymbol{x}_i) \\ = \sum_{i=1}^{n} \lambda_i \mathrm{Cov}(Z(\boldsymbol{x}_i), Z(\hat{\boldsymbol{x}})) \end{aligned} \tag{13}$$

$$\sum_{i=1}^{n} \lambda_i f_l(\boldsymbol{x}_i) = f_l(\hat{\boldsymbol{x}}). \tag{14}$$

Equations (13) and (14) can be expressed more compactly in matrix notation:

$$\begin{bmatrix} \boldsymbol{C} & \boldsymbol{F} \\ \boldsymbol{F}^T & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{\lambda} \\ \boldsymbol{\mu} \end{bmatrix} = \begin{bmatrix} \boldsymbol{c} \\ \boldsymbol{f} \end{bmatrix}. \tag{15}$$

Here, $C \in \mathbb{R}^{n \times n}$ is a matrix with entries $C_{i,j} = \mathrm{Cov}(Z(\boldsymbol{x}_i), Z(\boldsymbol{x}_j))$, $F \in \mathbb{R}^{n \times k}$ is a matrix with entries $F_{i,l} = f_l(\boldsymbol{x}_i)$, $\boldsymbol{c} \in \mathbb{R}^n$ is a vector with entries $c_i = \mathrm{Cov}(Z(\boldsymbol{x}_i), Z(\hat{\boldsymbol{x}}))$, and $\boldsymbol{f} \in \mathbb{R}^k$ is a vector with entries $f_l(\hat{\boldsymbol{x}})$. The vectors $\boldsymbol{\lambda} \in \mathbb{R}^n$ and $\boldsymbol{\mu}$ contain the n coefficients λ_i and the k Lagrange multipliers μ_l, respectively.

2.2 Covariogram model

The coefficient $\boldsymbol{\lambda}$, as determined by (15), depends on the mutual covariances between data points $C_{i,j} = \mathrm{Cov}(Z(\boldsymbol{x}_i), Z(\boldsymbol{x}_j))$ and on the covariances between data points and the point of

interest $\hat{\boldsymbol{x}}$, $c_i = \mathrm{Cov}(Z(\boldsymbol{x}_i), Z(\hat{\boldsymbol{x}}))$. These covariances are generally unknown and, consequently, are approximated using a covariogram model. The covariogram model, $C(\boldsymbol{h})$, depends only on the distance $\boldsymbol{h} = \boldsymbol{x}_i - \boldsymbol{x}_j$ of any two points $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ and not on their absolute positions.

Mainly three types of covariogram models can be found in the literature [15]: the spherical, the Matérn, and the exponential model. The Matérn function is known to be particularly suitable for representing various covariance–distance relationships [16] and is used in this study, with a smoothing parameter of 5/2, (16).

$$C(\boldsymbol{h}) = \theta^2_{\mathrm{Nugget}} + \theta^2_\sigma \left(1 + \sqrt{5}r + \frac{5}{3}r^2\right) \exp\left(-\sqrt{5}r\right), \quad \text{with } r = \sqrt{\sum_{l=1}^{m} \frac{h_l^2}{\theta_l^2}}\,. \tag{16}$$

The parameter $\theta^2_{\mathrm{Nugget}}$, which is historically referred to as nugget factor, introduces an extra offset at $\boldsymbol{h} = 0$, which provides more flexibility in modeling the measurement error. The covariogram converges toward θ_σ for $\boldsymbol{h} \to 0$ and toward zero for $\boldsymbol{h} \to \infty$, and θ_l determines the characteristic length-scale for input variable x_l.

In this study, the covariogram parameters are determined by Maximum Likelihood Estimation (MLE) [17]. In this context, it is assumed that the set of measured outputs $Z(X) = (Z(\boldsymbol{x}_1), \ldots, Z(\boldsymbol{x}_n))^T \in \mathbb{R}^n$ follows a multivariate Gaussian distribution with the expected values $\boldsymbol{m}(X) = (m(x_1), \ldots, m(x_n))^T \in \mathbb{R}^n$ and the covariance $\boldsymbol{C} \in \mathbb{R}^{n\times n}$, (17)

$$p\left(Z(X)\,|\boldsymbol{\theta}\right) = \frac{1}{(2\pi)^{\mathrm{n}/2}\det(\boldsymbol{C})^{1/2}} \exp\left(-\frac{1}{2}(Z(X) - \boldsymbol{m}(X))^T \boldsymbol{C}^{-1}(Z(X) - \boldsymbol{m}(X))\right). \tag{17}$$

The entries in the covariance matrix $\boldsymbol{C} \in \mathbb{R}^{n\times n}$ are calculated by the covariance function using parameters $\boldsymbol{\theta}$. Using (17) the resulting log-likelihood function for a given set of covariogram parameters $\boldsymbol{\theta}$ is defined by (18)

$$\log p(\boldsymbol{\theta}|Z(\boldsymbol{x})) = -\frac{1}{2}(Z(X) - \boldsymbol{m}(X))^{\mathrm{T}} \boldsymbol{C}^{-1}(Z(X) - \boldsymbol{m}(X)) - \frac{1}{2}\log(|\boldsymbol{C}|) - \frac{n}{2}\log(2\pi)\,. \tag{18}$$

2.3 Estimating parameters of linear trend functions

The Kriging principles of unbiasedness and minimal prediction variance are also used for determining the coefficients a_l in which the trend functions are linear (5). Matheron (1971) [18] introduced this ansatz to estimate each individual parameter $a_l^* \in \mathbb{R}$ as a linear combination of the observations $Z(\boldsymbol{x}_i)$:

$$a_l^* = \sum_{i=1}^{n} \lambda_i^l Z(\boldsymbol{x}_i)\,. \tag{19}$$

In analogy to calculating the coefficients $\lambda_i(\boldsymbol{x})$ of the Kriging prediction in Section 2.2, the values of λ_i^l are determined such as to achieve unbiasedness and to minimize variance. The estimate a_l^* and the "true" parameter a_l have the following expected difference:

$$\begin{aligned} E\left[a_l - a_l^*\right] &= E\left[a_l - \sum_{i=1}^{n} \lambda_i^l Z(\boldsymbol{x}_i)\right] \\ &= E\left[a_l - \sum_{i=1}^{n} \lambda_i^l \left(m(\boldsymbol{x}_i) + Y(\boldsymbol{x}_i)\right)\right] \\ &= E\left[a_l - \left(\sum_{j=1}^{k} a_j \sum_{i=1}^{n} \lambda_i^l f_j(\boldsymbol{x}_i) + \sum_{i=1}^{n} \lambda_i^l Y(\boldsymbol{x}_i)\right)\right] \\ &= a_l - \sum_{j=1}^{k} a_j \sum_{i=1}^{n} \lambda_i^l f_j(\boldsymbol{x}_i)\,. \end{aligned} \tag{20}$$

Sufficient conditions for the unbiasedness condition (20) to hold are given by:

$$\begin{aligned} \sum_{i=1}^{n} \lambda_i^l f_j(x_i) &= 1 \text{ for } j = l \\ \sum_{i=1}^{n} \lambda_i^l f_j(x_i) &= 0 \text{ for } j \neq l. \end{aligned} \tag{21}$$

Using these conditions the variance can be estimated by (22)

$$\mathrm{Var}\left(a_l - a_l^*\right) = \sum_{i=1}^{n}\sum_{q=1}^{n} \lambda_i^l \lambda_q^1 \mathrm{Cov}\left(Z(\boldsymbol{x}_i), Z\left(\boldsymbol{x}_q\right)\right). \tag{22}$$

The optimal coefficients λ_i^l are also calculated using the method of Lagrange multipliers, resulting in the following system of linear equations:

$$\begin{bmatrix} \boldsymbol{C} & \boldsymbol{F} \\ \boldsymbol{F}^T & \boldsymbol{O} \end{bmatrix} \begin{bmatrix} \boldsymbol{\lambda}_l \\ \boldsymbol{\mu}_l \end{bmatrix} = \begin{bmatrix} 0 \\ \boldsymbol{\delta}_l \end{bmatrix}. \tag{23}$$

The left-hand side of (23) is identical to that of (15), and hence, the inverse of the block matrix can be reused here. The Dirac symbol $\boldsymbol{\delta_l} \in \mathbb{R}^k$ denotes a vector with entry 1 at index l and 0 elsewhere. The vectors $\boldsymbol{\lambda} \in \mathbb{R}^n$ and $\boldsymbol{\mu}$ contain the n sought coefficients λ_i^l and the k Lagrange multipliers μ_j^l, respectively.

This entire procedure needs to be repeated for each of the k parameters a_l which are calculated from the resulting coefficients $\boldsymbol{\lambda}_l$ (19). Alternatively, all k parameters can be determined simultaneously by solving the following matrix equation:

$$\begin{bmatrix} \boldsymbol{C} & \boldsymbol{F} \\ \boldsymbol{F}^T & \boldsymbol{O} \end{bmatrix} \begin{bmatrix} \boldsymbol{\Lambda} \\ \mathbf{M} \end{bmatrix} = \begin{bmatrix} \boldsymbol{O} \\ \boldsymbol{I} \end{bmatrix}. \tag{24}$$

Here, $\boldsymbol{\Lambda} \in \mathbb{R}^{n\times k}$ is a matrix with entries $\boldsymbol{\Lambda}_{i,l} = \lambda_i^l$, $\boldsymbol{M} \in \mathbb{R}^{k\times k}$ is a matrix with entries $\boldsymbol{M}_{j,l} = \mu_j^l$, and $\boldsymbol{I} \in \mathbb{R}^{k\times k}$ is the identity matrix.

Computing Λ from (24) requires inversion of a 2×2 block matrix, for which Lu and Shiou [19] provide an explicit formula. Exploiting the specific structure of the right-hand side of (24) and applying the dataset $\boldsymbol{z} = (Z(\boldsymbol{x}_1), \ldots, Z(\boldsymbol{x}_n))^T \in \mathbb{R}^n$, the coefficients $\mathbf{a}$ can be estimated by (25)

$$\boldsymbol{a} = \Lambda \boldsymbol{z} = \left(\boldsymbol{F}^T \boldsymbol{C}^{-1} \boldsymbol{F}\right)^{-1} \boldsymbol{F}^T \boldsymbol{C}^{-1} \boldsymbol{z}. \tag{25}$$

2.4 Kriging with nonlinear trend functions (KNT)

In this section, we extend the Kriging method to nonlinear trend functions. Consider a trend function $m(\boldsymbol{x})$ that depends nonlinearly on some parameters $p \in \mathbb{R}^s$. For technical reasons, the trend functions must linearly depend on at least one parameter a_0

$$m(\boldsymbol{x}) = a_0 f(\boldsymbol{x}, \boldsymbol{p}) . \tag{26}$$

In order to make the nonlinear parameters accessible to the Kriging method, we apply Taylor expansion of $m(\boldsymbol{x}, \boldsymbol{p})$ around an initial guess $\boldsymbol{p}^{(0)}$, resulting in a linear approximation $\tilde{m}(\boldsymbol{x}, \boldsymbol{p})$:

$$\tilde{m}(\boldsymbol{x}, \boldsymbol{p}) = a_0 m(\boldsymbol{x}, \boldsymbol{p}) + \sum_{l=1}^{s} a_0 \left(p_l - p_l^{(0)}\right) \frac{\partial m(\boldsymbol{x}, \boldsymbol{p})}{\partial p_l} \Bigg|_{p_l = p_l^{(0)}} \tag{27}$$

Now, the procedure from Section 2.4 can be applied for estimating the linear coefficients a_l of the linearized trend function $\tilde{m}(\boldsymbol{x}, \boldsymbol{p})$, as defined by (28) for $l \geq 1$

$$a_l = a_0 \left(p_l - p_l^{(0)}\right) . \tag{28}$$

The resulting estimates a_l are then used to compute new estimates of the nonlinear parameters $p_l^{(1)}$ from the initial guesses $p_l^{(0)}$:

$$p_l^{(1)} = \frac{a_l}{a_0} + p_l^{(0)} . \tag{29}$$

For linear trend functions, the estimated values in $\boldsymbol{p}^{(1)}$ would be the final result. However, for nonlinear trend functions the above steps (27) to (29) have to be repeated with $\boldsymbol{p}^{(1)}$ as new linearization point [7]. This leads to an iterative procedure which continues until the estimation is sufficiently accurate, i.e. the difference of $\boldsymbol{p}^{(i+1)}$ and $\boldsymbol{p}^{(i)}$ drops below a predefined threshold ϵ:

$$\|\boldsymbol{p}^{(i+1)} - \boldsymbol{p}^{(i)}\| \leq \epsilon \tag{30}$$

It should be noted here that the procedure described in this section depends on the choice of the initial guess and may result in local optima for the estimated parameters. For this reason, a multi-start strategy is suggested. In the following, the iterative procedure is reformulated into a root-finding problem. This allows the application of efficient standard algorithms [20].

Convergence of the iterative estimation procedure for the nonlinear parameters requires that for an infinite number of iterations the linearization point $p^{(i)}$ and the previous parameter estimation $p^{(i+1)}$ become identical (31)

$$\lim_{i \to \infty} \|\boldsymbol{p}^{(i+1)} - \boldsymbol{p}^{(i)}\| = 0. \tag{31}$$

From (29) it follows, that (31) is satisfied if and only if all coefficients a_l of the Taylor polynomial with $l \geq 1$ are zero (28) and (29). This can be used to reformulate the nonlinear parameter estimation problem as a root finding problem:

$$\sum_{l=1}^{k} a_l^2 = \boldsymbol{a}_{2:k+1}^T \boldsymbol{a}_{2:k+1} = 0. \tag{32}$$

Applying (25) we can rewrite (32) as:

$$\left(\tilde{\boldsymbol{F}}^T \boldsymbol{C}^{-1} \tilde{\boldsymbol{F}}\right)^{-1} \tilde{\boldsymbol{F}}^T \boldsymbol{C}^{-1} \boldsymbol{z} = \begin{bmatrix} a_0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \tag{33}$$

where $\tilde{\boldsymbol{F}} \in \mathbb{R}^{n \times (s+1)}$ contains the evaluation of the trend function $m(\boldsymbol{x}, \boldsymbol{p})$ and its s partial derivatives at each of the n given observation points.

As the values of a_0 do not enter in 2.31, the first row of (33) can be ignored:

$$\tilde{I}\left(\tilde{\boldsymbol{F}}^T \boldsymbol{C}^{-1} \tilde{\boldsymbol{F}}\right)^{-1} \tilde{\boldsymbol{F}}^T \boldsymbol{C}^{-1} \boldsymbol{z} = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix} \tag{34}$$

where $\tilde{\boldsymbol{I}}$ represents the identity matrix but with zero at the first entry.

3 Results and discussion

3.1 Numerical case study

In this section, we apply Kriging with nonlinear trend functions (KNT) to the two-substrate Michaelis–Menten equation (35) and study the robustness of the parameter estimation under the influence of noise

$$m(x, a, b, p) = \frac{a x_1 x_2}{b + p_1 x_1 + p_2 x_2 + x_1 x_2} \tag{35}$$

The model is linear with respect to the parameter a and nonlinear with respect to the parameters b, p_1 and p_2. The linear parameter a is not considered in the following analysis, since its value does not depend on the linearization point, and the parameter b is set to a fixed value for clarity and better visualization.

The nonlinear parameters p_1 and p_2 are estimated by KNT from a series of data sets with increasing noise levels. Artificial data are generated using the two-substrate Michaelis-Menten equation (35) with parameters $a = 10$, $b = 15$, $p_1 = 10$, $p_2 = 5$, following a Gaussian process with varied noise level $r \in \mathbb{R}^+$:

$$N(m(\boldsymbol{x}, \boldsymbol{p}), r \cdot m(\boldsymbol{x}, \boldsymbol{p})) . \tag{36}$$

Several datasets are generated on a 5×5 grid, with both variables x_1 and x_2 (35) taking any of the values $\{1, 5, 10, 15, 20\}$.

We first consider data without artificial noise ($r = 0$). Figure 1A shows the generated data set. The nonlinear trend function parameters p_1 and p_2 are determined as described in Section 3: The two-substrate Michaelis-Menten equation (35) is linearized and the corresponding parameters p_1 and p_2 are determined by Kriging according to (25). The optimal parameter values are then determined such that the corresponding Taylor coefficients a_1 and a_2 become zero. Figure 1B illustrates that both zero contour lines intersect at the "true" parameter set $p_1 = 10$ and $p_2 = 5$.

Eng. Life Sci. **2017**, *17*, 916–922

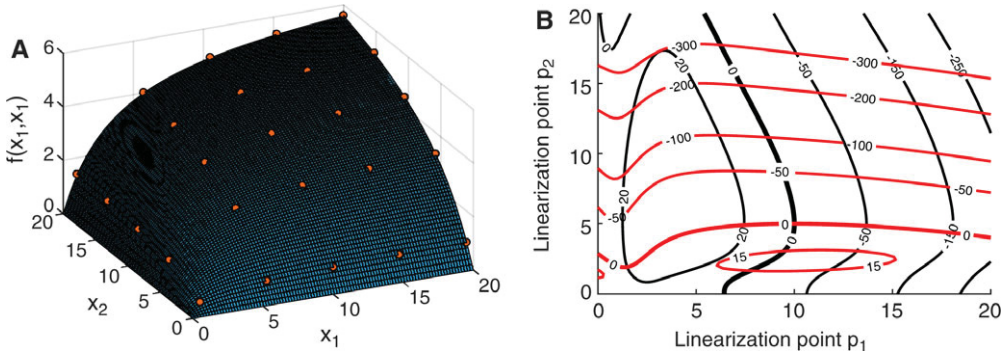


Figure 1. (A) Artificial dataset without noise. Generated data points are indicated as red dots. The mesh represents the two-substrate Michaelis–Menten kinetic (35) with parameters $a = 10$, $b = 15$, $p_1 = 10$ and $p_2 = 5$. (B) Contour plot of the Taylor coefficients a_1 (black) and a_2 (red) as functions of the linearization point for p_1 and p_2.

Next, the influence of noise on KNT is investigated using the following workflow:

(i) Value of r is chosen and 50 data sets are generated as previously described following (36).
(ii) Parameters of the covariogram model are estimated for each data set by maximum likelihood estimation applying the fitrgp implementation from MATLAB [21].
(iii) The nonlinear parameters p_1 and p_2 are determined for each data set as previously described, using the trust-region Dogleg method in the MATLAB function *fsolve* and a multi-start strategy for solving the root finding problem.
(iv) Points 1-3 are repeated for all noise levels of interest.

In total, $10 \times 50 = 1000$ data sets are generated with 10 different noise levels of r and 50 independent data sets at each noise level. The chosen noise levels are equally distributed in the range $0 \leq r \leq 0.09$.

Figure 2 A and B visualize the residuum distribution in dependency of the noise level. Variation in the parameter estimation increases with the noise level while the median stays near 0 and indicates an unbiased estimation. For comparison, the parameter estimation was repeated using the least squares (LSQ) method implemented by the MATLAB function lsqnonlin. In the numerical case study, LSQ leads to a smaller variance of the estimated parameters, as shown in Fig. 2 C and D. However, the linear parameter estimated by the LSQ method is not used in the Kriging prediction, as the Kriging method includes another estimate of this parameter. Hence, the LSQ estimates of the non-linear parameters might not be consistent with the Kriging estimate of the linear parameter. In the KNT framework, the non-linear parameters are estimated independently of the linear parameter. Moreover, Kriging does generally not aim at determining the most accurate parameter estimates but at minimizing the prediction variance.

3.2 Case study with real measurement data

Finally, we demonstrate the application of our nonlinear Kriging procedure to real measurement data from the field of biocatalysis. Kulig et al. [22] characterized an alcohol dehydrogenase from *Ralstonia sp.* (RADH) which can reduce sterically hindered ketones. The authors have investigated *inter alia* the influence of pH and temperature on the acceptance of different substrates (aldehydes, ketones and alcohols) and of the resulting kinetic parameters of the enzyme. In this section, we use the measured data of specific RADH activities for varying concentrations of Cyclohexanol (CycloHex) and $NADP^+$. The authors of [22] have varied only one factor at a time, assuming that no cooperative effects between the two impact factors are present.

The parameters of the Matern covariogram model are again determined by maximal likelihood estimation. Figure 3A shows the ordinary Kriging interpolation of the specific RADH activity using a constant as trend function. This leads to unphysical Kriging predictions in the poorly sampled regions. In particular, negative values such as predicted at the coordinate origin, are biologically not feasible. These deficiencies can be overcome by using a suitable trend function, for example the two-substrate Michaelis-Menten equation (35), introduced in Section 3.1 with CycloHex as x_1 and $NADP^+$ as x_2. This study does not aim at identifying the correct mechanistic model for this enzyme. However, the trend function reflects some fundamental properties of enzyme-catalyzed reactions, i.e. no conversion at zero substrate concentrations, initial reaction velocities at low substrate concentrations, and saturation at high substrate concentrations. A more detailed mechanistic model is not required, as the Kriging model captures deviations from the fundamental trend.

In this case study, using KNT the parameter set $a = 7.36$ U/mg, $b = 0.14$ mM^2, $p_1 = 0.05$ mM, $p_2 = 3.13$ mM was estimated. In the two-substrate Michaelis-Menten equation, a is the specific activity, p_1 and p_2 are the Michaelis constants of both substrates, and b is the mathematical product of the Michaelis constant of one substrate and the inhibition constant of the other substrate. Figure 3b shows the Kriging prediction surface using the estimated two-substrate Michaelis-Menten kinetic as trend function. The results show that the application of KNT clearly leads to much more realistic predictions. Furthermore, the prediction accuracy was evaluated by cross-validation. That is, the Kriging prediction $Z^*(\boldsymbol{x}_i)$ is calculated after removing the point $\boldsymbol{x}_i$ from the data base and compared to the Kriging prediction $Z(\boldsymbol{x}_i)$ with point $\boldsymbol{x}_i$. The squared differences are summed over all points, (37)

$$CV = \frac{1}{n} \sum_{i=1}^{n} \left(Z\left(\boldsymbol{x}_i\right) - Z^*\left(\boldsymbol{x}_i\right)\right)^2. \tag{37}$$

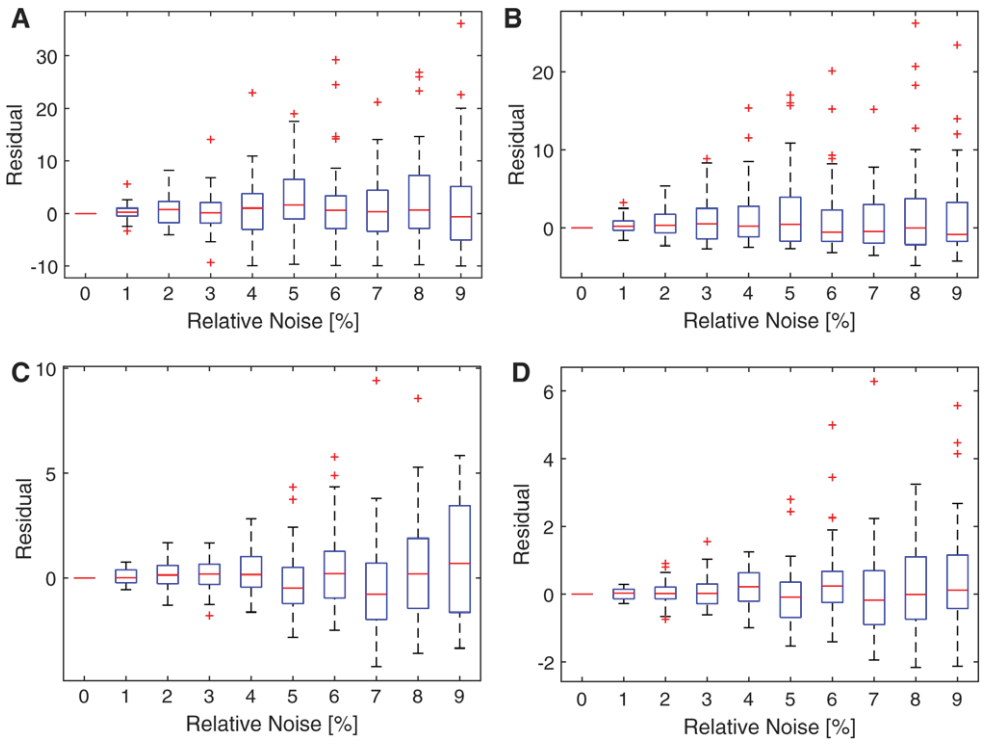


Figure 2. Dependency of the estimation error on the relative noise using KNT (A: p_1 and B: p_2) and LSQ (C: p_1 and D: p_2). Red line indicates the median, the bottom and top edge blue box represent the 25 and the 75% quantile. The bottom and top edge of the whiskers represent the 1.5-fold of the 25% and the 75% quantile. In case of a normal distribution, approximately 99.3% of all samples are contained in this range. Red stars represent outliers.

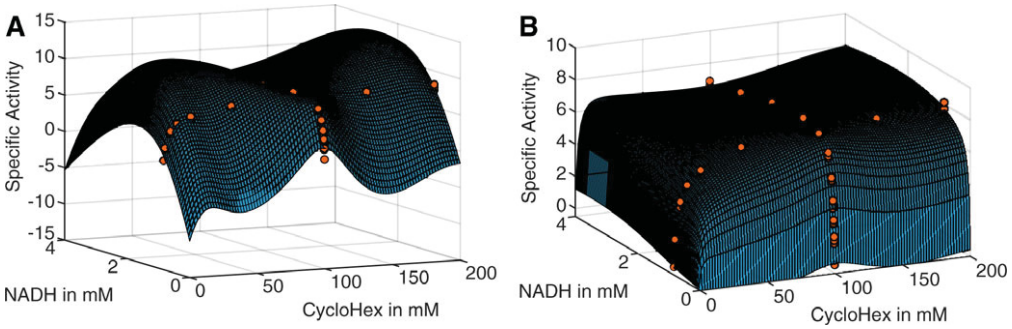


Figure 3. (A) Kriging interpolation surface with a constant as trend function. (B) Kriging interpolation surface with two-substrate Michaelis–Menten kinetic as trend function.

The prediction error of KNT, $CV_{KNT} = 2.45$, is significantly lower than of ordinary Kriging, $CV_{\text{Ordinary}} = 36.7$. This demonstrates that, in case of small data sets, KNT can outperform ordinary Kriging by combining the best of both worlds, statistical data analysis and mechanistic modeling.

4 Conclusions

Kriging is a powerful tool for approximating functional relationships between input and output factors in complex system analysis, combining features of statistical data analysis and mechanistic modeling. In this context, deterministic trend functions can be particularly useful for compensating lack of experimental information in poorly sampled regions. However, Kriging so far only allows trend functions that are linear with respect to their parameters. In this contribution, we have presented a practical procedure for applying Kriging with nonlinear trend functions, based on Taylor linearization and leading to a root finding problem, shortly referred to as Kriging with nonlinear trend functions (KNT). In two case studies, KNT is applied on the two substrate Michaelis–Menten equation. First the method is studied using synthetic data with increasing noise levels. Then, KNT is applied for analyzing real measurement data with rather low and irregular sample density. The results demonstrate that KNT outperforms ordinary Kriging, as the mechanistic trend function can compensate for missing experimental information.

Practical application

Kriging is a statistical interpolation method which allows integrating both empirical and mechanistic modeling approaches. While Kriging is commonly applied in geostatistics and fluid dynamics, the method is only recently recognized in biotechnology. The empirical part of Kriging analyzes first how the covariance of given measurement data depends on the distance of the respective measurement points. Based on this statistical information a smooth prediction curve/surface and the associated confidence tube can be calculated. In addition, basic trends can be incorporated by mechanistic models which allows, to some extent, to extrapolate over the range of data used for model calibration.

The authors gratefully acknowledge preliminary work by Robert Dalitz and thank Samuel Leweke for fruitful discussions. Moreover, the authors gratefully thank Dörte Rother and Martina Pohl for providing data files. Lars Freier gratefully acknowledges a PhD scholarship by the Ministry of Innovation, Science and Research of North Rhine-Westphalia and the Heinrich Heine University Düsseldorf in the CLIB Graduate Cluster Industrial Biotechnology.

The authors have declared no conflict of interest.

Nomenclature

Z	Output/process variable
Z^*	Kriging estimation
$\boldsymbol{x}$	Input/design variable
$\hat{\boldsymbol{x}}$	point of interest
E	Expected Value
Cov	Covariance
Var	Variance
$\boldsymbol{C}$	Covariance matrix
H_0	Null hypothesis
$\sigma(\boldsymbol{x})$	Standard deviation of the random process
$m(\boldsymbol{x},\boldsymbol{p})$	Complete trend function of random process
$f_l(\boldsymbol{x})$	Sub-trend function
a_l	Coefficient associated with $f_l(x)$
λ	Kriging coefficient
μ	Lagrange multiplier
Km	Michaelis–Menten constant
p	Parameter of trend function

5 References

[1] Bezerra, M. A., Santelli, R. E., Oliveira, E. P., Villar, L. S. et al., Response surface methodology (RSM) as a tool for optimization in analytical chemistry. *Talanta* 2008, *76*, 965–77.

[2] Baş, D., Boyacı, İ. H., Modeling and optimization I: Usability of response surface methodology. *J. Food Eng.* 2007, *78*, 836–845.

[3] Diggle, P. J., Tawn, J. A., Moyeed, R. A., Model-based geostatistics. *J. R. Stat. Soc. Ser. C (Applied Stat.)* 2002, *47*, 299–350.

[4] Jeong, S., Murayama, M., Yamamoto, K., Efficient optimization design method using Kriging model. *J. Aircr.* 2005, *42*, 413–420.

[5] Bonowski, F., Kitanovic, A., Ruoff, P., Holzwarth, J. et al., Computer controlled automated assay for comprehensive studies of enzyme kinetic parameters. *PLoS One* 2010, *5*, e10727.

[6] Freier, L., Hemmerich, J., Schöler, K., Wiechert, W. et al., Framework for Kriging-based iterative experimental analysis and design: Optimization of secretory protein production in Corynebacterium glutamicum. *Eng. Life Sci.* 2016, *16*, 538–549.

[7] Freier, L., von Lieres, E., Kriging based iterative parameter estimation procedure for biotechnology applications with nonlinear trend functions. *IFAC-PapersOnLine* 2015, *48*, 574–579.

[8] Yates, S. R., Warrick, A. W., Myers, D. E., Disjunctive Kriging: 1. Overview of estimation and conditional probability. *Water Resour. Res.* 1986, *22*, 615–621.

[9] Asa, E., Saafi, M., Membah, J., Billa, A., Comparison of linear and nonlinear Kriging methods for characterization and interpolation of soil data. *J. Comput. Civ. Eng.* 2012, *26*, 11–18.

[10] Rasmussen, C. E., *Gaussian Processes for Machine Learning*, MIT Press, Cambridge 2006.

[11] Krige, D. G., A statistical approach to some basic mine valuation problems on the Witwatersrand. *J. South. African Inst. Min. Metall.* 1951, *52*, 119–139.

[12] Cressie, N. A. C., *Statistics for Spatial Data*, 3rd ed., Wiley, New York 1993.

[13] Buja, A., Hastie, T., Tibshirani, R., Linear smoothers and additive models. *Ann. Stat.* 1989, *17*, 453–510.

[14] Cressie, N. A. C., Kriging variance and prediction intervals, in: *Statistics for Spatial Data*, 1993, pp. 154–155.

[15] Marchant, B. P., Lark, R. M., The Matérn variogram model: Implications for uncertainty propagation and sampling in geostatistical surveys. *Geoderma* 2007, *140*, 337–345.

[16] Minasny, B., McBratney, A. B., The Matérn function as a general model for soil variograms. *Geoderma* 2005, *128*, 192–207.

[17] Mardia, K. V, Marshall, R. J., Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika* 1984, *71*, 135–146.

[18] Matheron, G., *The Theory of Regionalized Variables and its applications*, Ecole Nationale Superieure des Mines de, Paris, 1971.

[19] Lu, T., Shiou, S., Inverses of 2x2 block matrices. *Comput. Math. with Appl.* 2000, *43*, 119–129.

[20] Hansen, E., Patrick, M., A family of root finding methods. *Numer. Math.* 1976, *27*, 257–269.

[21] MathWorks, fitrgp. 2016.

[22] Kulig, J., Frese, A., Kroutil, W., Pohl, M. et al., Biochemical characterization of an alcohol dehydrogenase from Ralstonia sp. *Biotechnol. Bioeng.* 2013, *110*, 1838–1848.

2.2 Framework for Kriging-based iterative experimental analysis and design: Optimization of secretory protein production in Corynebacterium glutamicum

Engineering in Life Sciences

Eng. Life Sci. **2016**, *0*, 1–12 www.els-journal.com

Lars Freier[1]*
Johannes Hemmerich[1,2]*
Katja Schöler[1,2]
Wolfgang Wiechert[1,2]
Marco Oldiges[1,2,3]†
Eric von Lieres[1,3]

[1]Institute for Bio- and Geosciences, Biotechnology (IBG-1), Forschungszentrum Jülich, Jülich, Germany

[2]BioEconomy Science Center (BioSC)

[3]Institute of Biotechnology, RWTH Aachen University, Aachen, Germany

Research Article

Framework for Kriging-based iterative experimental analysis and design: Optimization of secretory protein production in *Corynebacterium glutamicum*

The production of bulk enzymes used in food industry or organic chemistry constitutes an important part of industrial biotechnology. The development of production processes for novel proteins comprises a variety of biological engineering and bioprocess reaction engineering factors. The combinatorial explosion of these factors can be effectively countered by combining high-throughput experimentation with advanced algorithms for data analysis and experimental design. We present an experimental optimization strategy that merges three different techniques: (1) advanced microbioreactor systems, (2) lab automation, and (3) Kriging-based experimental analysis and design. This strategy is demonstrated by maximizing product titer of secreted green fluorescent protein (GFP), synthesized by *Corynebacterium glutamicum*, through systematic variation of CgXII minimal medium composition. First, relevant design parameters are identified in an initial fractional factorial screening experiment. Then, the functional relationship between selected media components and protein titer is investigated more detailed in an iterative procedure. In each iteration, Kriging interpolations are used for formulating hypotheses and planning the next round of experiments. For the optimized medium composition, GFP product titer was more than doubled. Hence, Kriging-based experimental analysis and design has been proven to be a powerful tool for efficient process optimization.

Keywords: Bioprocess engineering / Design of experiments / High-throughput experimentation / Iterative experimental optimization / Kriging

Additional supporting information may be found in the online version of this article at the publisher's web-site

Received: December 1, 2015; *revised:* February 16, 2016; *accepted:* March 9, 2016

DOI: 10.1002/elsc.201500171

1 Introduction

1.1 Protein production in industrial biotechnology

The production of proteins that are needed in high quantities is a major part of industrial biotechnology. Such bulk enzymes (e.g. lipases, proteases, amylases) are found widely in daily applications [1].

Developing a protein production process includes both biological engineering (i.e. the expression host) and the bioprocess reaction engineering (i.e. cultivation control). The first aspect covers methods that have evolved with the advent of modern recombinant gene technology, e.g. promotor libraries for fine-tuning of gene expression [2], homologous, and heterologous signal peptide libraries for enhanced protein secretion [3], adjustment of codon usage to control translation speed [4], or customizing the glycosylation profile of recombinant proteins [5]. The second above-mentioned aspect, the bioprocess engineering, includes also a wealth of parameters that are known

Correspondence: Dr. Eric von Lieres (e.von.lieres@fz-juelich.de) IBG-1: Biotechnology, Forschungszentrum Jülich, Wilhelm-Johnen-Straße 1, 52425 Jülich, Germany

Abbreviations: DoE, design of experiments; **GFP**, green fluorescent protein; **MBR**, microbioreactor

*These authors have contributed equally to this work.

†Additional correspondence: Prof. Marco Oldiges (m.oldiges@fz-juelich.de), IBG-1: Biotechnology, Forschungszentrum Jülich, Wilhelm-Johnen-Straße 1, 52425 Jülich, Germany

affect overall productivity, e.g. induction strategy [6], the feeding rate [7], controlling the specific growth rate [8], adjustment of cultivation temperature [9], or the nutrient supply in basal and feed media [10–12].

Clearly, the high number of biological and process engineering aspects leads to a combinatorial explosion of possible experiments. This problem can be addressed applying high-throughput experimentation in combination with sophisticated algorithms for interpretation and planning of cultivation experiments.

1.2 Microbioreactor systems for high-throughput experimentation

Nowadays, microbioreactor (MBR) systems are emerging tools in bioprocess development. Such systems provide the possibility to conduct several cultivation experiments in parallel on a small footprint and thus, increase experimental throughput. Typically, reaction volumes are in the range of several hundred μL to several mL. Scalability to lab-scale stirred tank reactors has been demonstrated several times for different MBR [13–16]. Research and development on MBR is still ongoing, with the ultimate goal to minimize laborious and low-throughput conventional laboratory scale experimentation (see also literature reviews in [17–19]). Within this context, several demands on high-throughput process development have been formulated [20], including the incorporation of mathematical methods for experimental planning and data evaluation, both of which is addressed in this study. Due to the high number of generated samples using MBR, the current bottleneck in bioprocess development is shifted toward analytics. Furthermore, a simultaneously upcoming limitation is the interpretation of results and subsequent planning of new experiments during iterative optimization of bioprocesses.

1.3 Cultivation medium optimization

Optimization of medium is a typical task in bioprocess design. In general, several strategies are found in studies concerning medium optimization focusing on various optimization targets. A comprehensive overview and discussion is given by Kennedy and Krouse [21], Weuster-Botz [22], and Zhang and Greasham [23]. The here presented study aims to maximize product titer (i.e. amount of secreted recombinant GFP (green fluorescent protein)) as this is a typical objective of industrial process optimization rather than, e.g. biomass-specific product concentrations, which reflect a biological property of the cell. GFP was chosen as model protein due to its easy detection by fluorescence of the active protein and by doing so, analytics was prohibited from becoming a bottleneck and further impact of noise on the optimization response caused by wet lab assays for product estimation is minimized. Besides, use of GFP as model protein is widespread in characterization studies on MBR [13, 24, 25].

1.4 Design of experiments and Kriging

"Classic" design of experiments (DoE) is a powerful tool for efficiently estimating single and combinatorial effects of input variables on noisy system outputs. It has a long tradition and was introduced by Fisher in 1935 [26]. DoE is established in several biotechnological fields and has been applied for optimization of, e.g. nutrition media or process operation (cultivation, product recovery, purification) as summarized by Mandenius and Brundin [27].

Although DoE is popular for identifying significant input variables, the methodology usually lacks in approximating highly nonlinear functional relationships. For this purpose, the Kriging approach is often more appropriate [28]. Kriging provides a data driven, unbiased linear estimator with minimal mean square prediction error. Technical details of the Kriging methodology can be found in Sections 2.3–2.5.

In this study, a combination of "classic" DoE and Kriging is demonstrated to enable efficient iterative experimental optimization. First, a fractional factorial DoE is applied for screening for media components that have a major impact on the optimization objective, the GFP signal. On this basis, the functional relationship between selected key components and the GFP signal is iteratively analyzed using the Kriging methodology. In each iteration, Kriging interpolations are used for planning experiments for the next iteration. Afterwards, the Kriging interpolation and its corresponding error estimate are used in a statistic test for identifying promising media compositions. A final DoE is applied for checking if the results of the initial screening with all media components are also valid around this optimal medium composition.

1.5 Aim of the study

The study aims to demonstrate how current technology of MBR, lab automation, and algorithm-based planning and evaluation of cultivation experiments can be hyphenated to boost bioprocess development. A Kriging-based DoE approach is applied to investigate the influence of media composition on secretory protein production in *Corynebacterium glutamicum*. The here described process optimization framework yielded a considerably raised overall process performance in a statistically secured fashion. Moreover, the findings of the study indicate that the optimized medium provides a general trigger to enhance secretory protein production with *C. glutamicum*.

2 Materials and methods

2.1 Strain and cultivation conditions

C. glutamicum ATCC13032 with pEKEx2 expression plasmid containing the fusion of phoD signal peptide from *B. subtilis* and GFP (resulting plasmid: pCGPhoD^{Bs}-GFP) was used as model protein secretion strain [29]. The phoD signal peptide mediates secretion of folded GFP via Tat-pathway, whereas pEKEx2 plasmid gives Kanamycin resistance. In all cultivations selection

pressure for maintaining the expression plasmid was established by adding 25 mg/L of Kanamycin.

For the study, a working cell bank was made by overnight culturing of the model strain in complex BHI-medium containing 37 g/L of brain heart infusion broth powder. Afterwards, one volume of overnight culture was combined with one volume of sterile glycerol solution (500 g/L) as cyro-preservative and deep-frozen at -80°C as cryo-stock 1-mL-aliqouts. Conditions for this overnight cultivation were as follows: 50 mL culture volume, four-baffled shaking flask with a nominal volume of 500 mL, 30°C, 250 rpm at 25 mm shaking diameter.

For each medium optimization experiment, a preculture was inoculated with a complete cryo-stock aliqout and grown in BHI-medium until an optical density (OD_{600}) of 3–4 was reached after an incubation time of approx. 5 h. Then, then preculture was used to inoculate the main culture as described below. Conditions for precultures were as described above for generation of the working cell bank.

All main cultivations were carried out using a microbioreactor system ("BioLector," Art.-No. G-BL-100, m2p-labs GmbH, Baesweiler, Germany), capable of sensing biomass concentration via backscatter light at 620 nm, GFP concentration via fluorescence (Ex.: 488 nm, Em.: 520 nm) and pH and dissolved oxygen via optical sensing spots located at the bottom of 48-well flower-shaped microtiterplates ("Flowerplate," Art.-No. MTP-48-BOH, m2p-labs GmbH, Baesweiler, Germany). Cultivation conditions were as described in [15].

For main cultivations CgXII minimal medium and variants according to the experimental planning were used. The composition of CgXII minimal medium used in several other studies (e.g. [30] and references therein) is referred to as "reference" throughout the study. Variations of the CgXII medium included different concentrations of medium components under investigation, including a zero concentration (i.e. omitting the specific components). Modifications of the medium composition are shown in the results section. Glucose as main carbon source was fixed to 10 g/L. For induction of recombinant gene expression, all main cultivation media were prepared with 100 μM Isopropyl-β-D-1-thiogalactopyranoside (IPTG) final. Recipes for CgXII standard medium ("reference") and finally optimized composition with respect to highest achieved GFP titer are given in the supplement (Supporting Information Table S3). In each MBR cultivation five replicates using CgXII reference medium composition were conducted as internal standard to correct for possible systemic effects during automated medium preparation or resulting from different production batches of stock solutions. After automated media preparation and inoculation, the MBR cultivation was started and terminated after a minimum incubation time of 18 h. At this time point, all cultivations throughout the study had reached stationary growth phase, indicated by online measured biomass and dissolved oxygen signal.

The normalized GFP fluorescence signal after 17 h of incubation was used for data analysis. The data were normalized by division with the mean GFP fluorescence signal of all five reference cultivations to enhance plate to plate comparability.

All reagents used for media preparation were purchased from Carl Roth, Merck, or Sigma and were of analytical grade.

2.2 Analytics

Cultivation supernatants were obtained by centrifugation for 10 min at 13 000 rpm ("Biofuge pico," Heraeus). Protein content of supernatant was determined with a Bradford assay kit according to the supplier's instructions using BSA as protein standard ("Bradford Reagent," Sigma). Amount of secreted GFP was visualized by denaturing SDS-page of supernatants in 12% Bis-Tris Gel according to the suppliers instructions ("TruPage," Sigma), with a microwave-based staining protocol [31].

2.3 Automated media preparation

Variations of CgXII medium were made using appropriate stock solutions of medium components. Depending on target concentration of selected medium components, different amounts of stock solutions were pipetted by an automated liquid handling station, which has been described and characterized earlier [15]. For the liquid handling station used in this study, the pipetting accuracy was determined to be better than 2.5% for 10 μL, better than 2% for 50 μL and better than 0.5% for 100 μL and above. The pipetting precision was determined to be better than 5% for 10 μL, better than 2% for 50 μL and better than 1% for 100 μL and above (personal communication with A. Radek, [32]).

Additionally, the minimal volume to be pipetted was set to 10 μL with increments of 5 μL up to 1000 μL. A volume of 950 μL of the completely prepared medium was filled into each well of a Flowerplate. Then, 50 μL of freshly grown preculture was added to each well with the liquid handling robot, yielding the final cultivation volume of 1000 μL. To assess positional effects, reference cultivations were located in the corners and in the center of each Flowerplate; no positional effects were observed throughout the whole study.

2.4 Ordinary Kriging

Data analysis and experimental design were performed using the Kriging method for data interpolation/extrapolation and visualization. In this section, a brief introduction to Kriging is given, while further details and mathematical deviations are described elsewhere [33].

In Kriging, the system output $Z(\boldsymbol{x})$ is assumed to follow a basic trend $m(\boldsymbol{x})$ and the noise in the data is assumed to be caused by a random process $Y(\boldsymbol{x})$ with zero mean and standard deviation $\sigma(\boldsymbol{x})$.

$$Z(\boldsymbol{x}) = m(\boldsymbol{x}) + Y(\boldsymbol{x}) \tag{1}$$

$$\mathrm{E}[Y(\boldsymbol{x})] = 0 \tag{2}$$

$$\sigma(x) = \sqrt{\mathrm{Var}(Y(\boldsymbol{x}))} \tag{3}$$

Even though more complex trend functions can be used, it is often sufficient to assume a constant trend, $m(\boldsymbol{x}) = c$, which is referred to as ordinary Kriging.

In Kriging, predictions Z^* of unknown observations at points $\hat{\boldsymbol{x}}$ are calculated by linear combination of the given data $Z(\boldsymbol{x}_i)$ with coefficients $\lambda_i \in \mathbb{R}$.

$$Z^*(\hat{\boldsymbol{x}}) = \sum_{i=1}^{n} \lambda_i(\hat{\boldsymbol{x}})\, Z(\boldsymbol{x}_i) \tag{4}$$

The coefficients λ_i, which depend on $\hat{\boldsymbol{x}}$, are determined such as to provide unbiased predictions, eq. (5), and to minimize the prediction variance, eq. (6).

$$\mathrm{E}\left[Z(\hat{\boldsymbol{x}}) - Z^*(\hat{\boldsymbol{x}})\right] = 0 \tag{5}$$

$$\mathrm{Var}\left[Z(\hat{\boldsymbol{x}}) - Z^*(\hat{\boldsymbol{x}})\right] \rightarrow \min \tag{6}$$

For finding appropriate Kriging coefficients that satisfy eq. (5) and eq. (6), the covariance $\mathrm{Cov}(Z(\boldsymbol{x}_i), Z(\boldsymbol{x}_j))$ between two observations at positions x_i and x_j needs to be approximated by a covariogram model, as described in Section 2.4. Using the covariogram model, Kriging coefficients can be calculated by eq. (7) [33].

$$\begin{bmatrix} \boldsymbol{\lambda} \\ \mu \end{bmatrix} = \begin{bmatrix} C & \mathbf{1} \\ \mathbf{1}^T & 0 \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{c} \\ 1 \end{bmatrix} \tag{7}$$

In eq. (7), $C \in \mathbb{R}^{n\times n}$ is a matrix with entries $C_{i,j} = \mathrm{Cov}(Z(\boldsymbol{x}_i), Z(\boldsymbol{x}_j))$, $\mathbf{1} \in \mathbb{R}^n$ is a vector with entries 1, and $\boldsymbol{c} \in \mathbb{R}^n$ is a vector with entries $c_i = \mathrm{Cov}(Z(\boldsymbol{x}_i), Z(\widehat{\boldsymbol{x}}))$. The vector $\boldsymbol{\lambda} \in \mathbb{R}^n$ contains the n coefficients λ_i, and μ is a Lagrange multiplier. The Lagrange multiplier is required for solving eq. (6) constrained by eq. (5), but not required for calculating the Kriging prediction in eq. (4) [33].

By minimizing the prediction variance, Kriging inherently provides an estimation of the confidence interval. The variance of the Kriging prediction in eq. (4), indicating the prediction accuracy, can be estimated by eq. (8).

$$\begin{aligned} \mathrm{Var}\left[Z(\boldsymbol{x}) - Z^*(\boldsymbol{x})\right] &= \mathrm{Cov}\left(Z(\hat{\boldsymbol{x}}), Z(\hat{\boldsymbol{x}})\right) \\ &\quad -2\sum_{i=1}^{n} \lambda_i \mathrm{Cov}\left(Z(\boldsymbol{x}_i), Z(\hat{\boldsymbol{x}})\right) \\ &\quad +\sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i\lambda_j \mathrm{Cov}\left(Z(\boldsymbol{x}_i), Z(\boldsymbol{x}_j)\right) \end{aligned} \tag{8}$$

The first term in eq. (8) represents measurement noise. The second term contains information about the significance of individual measured observations $Z(\boldsymbol{x}_i)$ for predicting the observation at $\hat{\boldsymbol{x}}$. Usually, the absolute value of this term is larger for measurement points $\boldsymbol{x}_i$ that are closer to $\hat{\boldsymbol{x}}$. The third term quantifies the information content of the entire dataset as its value increases for high correlation between the data points. Consequently, the Kriging prediction variance increases with increasing noise and decreasing information content of the data, and it adapts to the location of the points of interest.

2.5 Covariogram model

The Kriging coefficients, as computed by eq. (7), depend on the mutual covariances between the data points $C_{i,j} = \mathrm{Cov}(Z(\boldsymbol{x}_i), Z(\boldsymbol{x}_j))$ and on the covariances between the data points and the point of interest $\hat{\boldsymbol{x}}$, $c_i = \mathrm{Cov}(Z(\boldsymbol{x}_i), Z(\hat{\boldsymbol{x}}))$. These covariances are generally unknown and, consequently, need to be approximated using a covariogram model. The covariogram model $C(\boldsymbol{h})$ depends only on the distance $\boldsymbol{h} = \boldsymbol{x}_i - \boldsymbol{x}_j$ of any two points $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ and not on their absolute positions.

Mainly three types of covariogram models can be found in the literature [34]: the spherical, the Matérn and the exponential model. In the following sections, the exponential model is used, eq. (9), as suggested in [35].

$$C(\boldsymbol{h}) = \begin{cases} \sigma + \sigma_{Nugget} & for \;\; \boldsymbol{h} = \mathbf{0} \\ \sigma \exp\left(-\sum_{q=1}^{d} \theta_q |h_q|^{p_q}\right) & for \;\; \boldsymbol{h} \neq \mathbf{0} \end{cases} \tag{9}$$

Here, p_q characterizes the steepness and θ_q the width of the transition for varying h_q. The parameter σ represents the maximal covariance value for $\boldsymbol{h} \rightarrow \mathbf{0}$, and σ_{Nugget}, which is historically referred to as nugget factor, introduces an extra offset at $\boldsymbol{h} = \mathbf{0}$. This extra offset provides more flexibility in modeling the measurement error. The values of p_q, θ_q, σ, and σ_{Nugget} must be larger than 0, and p_q is furthermore restricted to the interval $0 < p_q < 2$, since otherwise the resulting covariance matrix is not strictly positive definite [36]. In the present study, the parameters of the Covariogram model are estimated using the cross-validation method.

2.6 Cross-validation

In cross-validation, the quality of the estimated covariogram parameters is directly assessed by the quality of the associated Kriging interpolation. A good Kriging interpolation should approximate the measurement data and the normalized residual should follow a standard normal distribution. Consequently, two optimization criteria are formulated, as described by [37]. First, minimizing the sum of the mean squared error between the observation $Z(\boldsymbol{x}_i)$ and the Kriging estimation $Z^{*i}(\boldsymbol{x}_i)$, computed with the point $\boldsymbol{x}_i$ excluded from the data base:

$$c_1 = \frac{1}{n}\sum_{i=1}^{n} \left(Z(\boldsymbol{x}_i) - Z^{*i}(\boldsymbol{x}_i)\right)^2 \tag{10}$$

Second, minimizing the difference of the corresponding square error normalized by the estimated standard deviation, eq. (8), to its ideal value of one:

$$c_2 = \left| 1 - \frac{1}{n}\sum_{i=1}^{n} \left(\frac{Z(\boldsymbol{x}_i) - Z^{*i}(\boldsymbol{x}_i)}{\sigma^{*i}(\boldsymbol{x}_i)}\right)^2 \right| \tag{11}$$

Finally, the optimization problem for estimating the covariogram model parameters σ, σ_{Nugget}, $\boldsymbol{\theta}$, and $\boldsymbol{p}$ is formulated as:

$$\min_{\sigma,\sigma_{\mathrm{Nugget}},\boldsymbol{\theta},\boldsymbol{p}} \left(\frac{c_1}{\max(\boldsymbol{z})^2} + c_2\right) \tag{12}$$

Here, $\boldsymbol{z} \in \mathbb{R}^n$ is a vector containing the provided observations. Minimizing the criterion c_1 leads to an accurate interpolation and minimizing c_2 leads to a reliable estimation of the confidence interval. However, the actual values of c_1 and c_2 may vary by

orders of magnitude that may lead to a bias. Hence, c_1 is scaled by the squared maximal observation value.

Equation (12) is minimized by subsequent application of, first, the genetic algorithm implemented in the MATLAB global optimization toolbox [38] as global optimizer, and then the sequential quadratic programming algorithm implemented in the MATLAB optimization toolbox [39] as local optimizer.

2.7 Statistical identification of optimal regions

As described in Section 1.2, goal of the optimization is to maximize the measured GFP signal by varying the medium composition. However, noisy data make it difficult to identify optimal composition(s). The acquired data are always associated with some noise, and thus the identification of an optimal medium composition is in consequence associated with some uncertainty. Statistical hypothesis tests can help to overcome this problem, as they are based on statistical properties such as expected value and variance.

Such statistical hypothesis tests, e.g. Student *t*-test and *z*-test, allow to compare two random variables and to decide if they are significantly different from each other. In general, two variables are considered to be significantly different if their expected values differ with a probability of at least $p = 1 - \alpha$. Typical values for α, which is also referred to as significance level, are 0.10, 0.05, and 0.01. An overview about statistical hypothesis tests can be found in [40].

Although the Student *t*-test is often used for experiments with small samples sizes, the *z*-test is better suited to be used in combination with Kriging. It is suggested to apply the *z*-test on random variables that follow a normal distribution and when the associated variance σ^2 is known [40]. In Kriging, it is generally assumed that the output value $Z(\boldsymbol{x})$ is normally distributed, and Kriging provides an estimation of the prediction error variance $\sigma_{\boldsymbol{x}}^2 = \mathrm{Var}[Z(\boldsymbol{x}) - Z^*(\boldsymbol{x})]$, see eq. (8).

Since the optimization goal is to maximize the GFP signal, it seems reasonable to avoid parameter values (i.e. medium compositions) that correspond to significantly smaller GFP signals than the maximum that was found so far. Statistically speaking, it is tested if the output prediction $Z^*(\boldsymbol{x}_i)$ at point $\boldsymbol{x}_i$ is significantly smaller than the maximal predicted GFP signal $Z^*(\hat{\boldsymbol{x}}_{opt})$. The hypothesis H_0 is consequently formulated as follows:

$$H_0 : Z^*(\boldsymbol{x}_i) \leq Z^*\left(\hat{\boldsymbol{x}}_{opt}\right) \tag{13}$$

That is, the point of interest is significantly worse if H_0 is accepted.

Following the framework of the *z*-test, the standard score z_{x_i} is calculated as:

$$z_{\boldsymbol{x}_i} = \frac{Z^*(\boldsymbol{x}_i) - Z^*\left(\hat{\boldsymbol{x}}_{opt}\right)}{\sigma_{\boldsymbol{x}_i}} \tag{14}$$

With a given significance level α, the parameter set $\boldsymbol{x}_i$ is considered to be "suboptimal" if z_{x_i} is smaller than the (1-α)-quantile of the standard normal distribution. In this study, α was set to a value of 0.01, i.e. a confidence of 99%.

3 Results and discussion

3.1 Optimization of medium composition

Despite having the possibility to conduct 48 experiments in parallel, it is still needed to focus on a certain number of design variables in order to cope with the combinatorial explosion mentioned earlier. The here used CgXII reference recipe comprises 16 components [41], which would lead to a total number of $2^{16} \approx 65000$ experiments using a full factorial experimental layout.

To narrow the window of design variables, it was decided not to alter the following medium components: glucose as the main carbon source fixed to 10 g/L, nonmetabolizable 3-(N-morpholino)propanesulfonic acid (MOPS) to provide buffering capacitiy, KH_2PO_4 and K_2HPO_4 providing also buffering capacity and serving as phosphate source, urea as basal nitrogen source and pH-stabilizing agent, biotin for complementation of biotin auxotrophy, protocatechuic acid (PCA) as iron chelating agent. Thus, there are still nine medium components left for investigation, namely $(NH_4)_2SO_4$, $FeSO_4 \cdot 7\ H_2O$, $MnSO_4 \cdot H_2O$, $ZnSO_4 \cdot 7\ H_2O$, $CuSO_4 \cdot 5\ H_2O$, $NiCl_2 \cdot 6\ H_2O$, $CaCl_2 \cdot 2\ H_2O$, $MgSO_4 \cdot 7\ H_2O$, and $CoCl_2 \cdot 6\ H_2O$ (cf. Supporting Information Table S3). Additionally, $Na_2MoO_4 \cdot 2\ H_2O$ and H_3BO_3 were included as design variables, as those were described as additives in another CgXII formulation [42]. The medium components of interest (i.e. design variables) can be divided into three groups:

First, $(NH_4)_2SO_4$ as standard nitrogen source is a major nutrient. As mentioned before, the basal nitrogen source urea was not altered, thus the nitrogen supply was expected not to become growth limiting. In another study, modulating the nitrogen source was determined as promising target, although the overall optimization goal was to maximize biomass-specific GFP signal [15].

Second, $FeSO_4 \cdot 7\ H_2O$, $MnSO_4 \cdot H_2O$, $ZnSO_4 \cdot 7\ H_2O$, $CuSO_4 \cdot 5\ H_2O$, $NiCl_2 \cdot 6\ H_2O$, and $CoCl_2 \cdot 6\ H_2O$ represent the group of trace elements, which seem to be inherited from the first publication of the CgXII medium recipe by Keilhauer et al. [41]. To add more variations of this theme, $Na_2MoO_4 \cdot 2\ H_2O$ and H_3BO_3 were also investigated as mentioned before.

$MgSO_4 \cdot 7\ H_2O$ and $CaCl_2 \cdot 2\ H_2O$ constitute the third group, as those exceed clearly the concentration range of trace elements and thus, their necessary presence to promote growth is most likely from other nature as for trace elements. Teramoto et al. reported that increased Ca^{2+} concentration correlates with increased GFP and Amylase secretion using the Tat pathway [43]. It was speculated that varying concentrations of Mg^{2+} and Ca^{2+} show significant effects on secretion of GFP, despite different conditions were used than by Teramoto et al. who applied *C. glutamicum* R as background strain, medium containing yeast extract and casamino acids, and CgR0949 as signal peptide. In this study, *C. glutamicum* ATCC13032, minimal medium, and PhoD as signal peptide was employed. It should be pointed out that even the choice of a signal peptide determines secretion efficiency heavily and in a nonpredictable manner [3, 44].

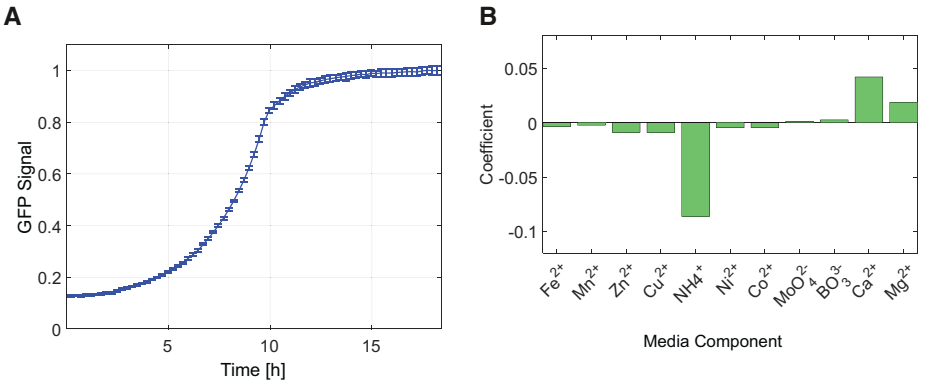


Figure 1. (A) Time course of normalized GFP signal during growth of reference cultivation experiments in one microbioreactor run ($n = 16$ biological replicates, depicted as mean value ± SD). (B) Effect of investigated medium components on GFP signal.

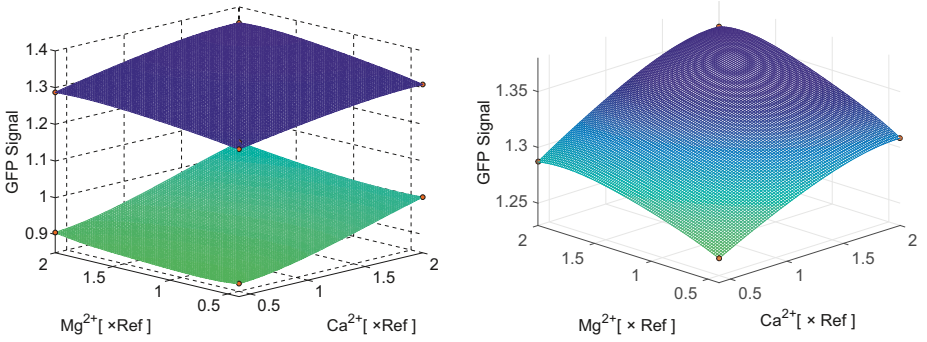


Figure 2. (A) Kriging interpolations of GFP signal when maximal amount of NH_4^+ (lower plane) or no NH_4^+ (upper plane) is added to medium. Sample points are indicated by red dots. *x* and *y* axes denote relative concentration of ions used for medium preparation compared to CgXII-Medium. (B) Detailed view of Kriging interpolation using medium without NH_4^+.

3.2 Screening analysis

The initial experiments were designed such as to identify components, here also referred to as design variables, with potentially significant impact on the GFP signal. As described in Section 1.4, classic design of experiment (DoE) is well suited for efficiently estimating single and combinatorial effects. Therefore, DoE is used for an initial screening in order to reduce the number of medium components for further, more detailed, analysis. A fractional factorial with 32 experiments was chosen, which allows an analysis of main factor effects without confounding with pairwise interaction. In order to determine biological reproducibility, 16 experiments were additionally run using the reference medium CgXII. The complete experimental design can be found in the Supporting Information.

The statistics of all 16 normalized GFP signal curves are depicted Fig. 1A. The relative standard deviation of the curves increases with time but never exceeds 1.7%. Screening analysis results are listed and visualized in Fig. 1B. The results indicate that NH_4^+ has a strong negative effect on GFP signal due to lower biomass production, whereas Ca^{2+} and Mg^{2+} show a positive effect on GFP signal. Thus, these three medium components are investigated more in detail in an iterative optimization procedure.

3.3 Iterative medium optimization

The experiments of the first iteration were planned such as to extend the knowledge obtained by the initial screening by investigating also potential interactions between the previously identified medium components of influence. For sake of clarity, all following concentrations refer to the concentrations used in the original reference medium (CgXII), indicated by ×Ref. Furthermore, since the screening revealed that NH_4^+ has a significant negative effect on the GFP-Signal, it was also of interest if NH_4^+ could be omitted. A full factorial experimental design with eight experiments was applied. Concentration ranges for Ca^{2+} and Mg^{2+} are chosen to be the same as in the screening analysis, i.e. minimal concentration was 0.4 ×Ref and maximal concentration was 2 ×Ref. For NH_4^+, the concentrations were set in the range of 0 ×Ref to 2 ×Ref. The GFP signal between the sample points was estimated using Universal Kriging as described in Section 2.3 (see Fig. 2). As predicted by the screening analysis, Ca^{2+} and Mg^{2+} have a positive influence on the GFP signal, with the effect of Ca^{2+} being more pronounced. Furthermore, the negative effect of NH_4^+ was confirmed. Consequently, by omitting NH_4^+, the maximal GFP signal could be increased by ca. 30% (from 1.03 to 1.36, see Fig. 2A and B). The Kriging interpolations in Fig. 2A indicate that high concentrations of Ca^{2+} and Mg^{2+} lead to high GFP signals. In order to validate this hypothesis, the maximal concentrations of Ca^{2+} and Mg^{2+} were doubled in the next experimental iteration.

Zimmermann [45] formulated two criteria that satisfy a good experimental design for Ordinary Kriging: First, the design should be space filling for exploring the system and second, the design should comprise some points in close proximity for studying the dependency of the covariance between these neighboring sample points. Following these two criteria, the new experiments for the second iteration were designed using a full

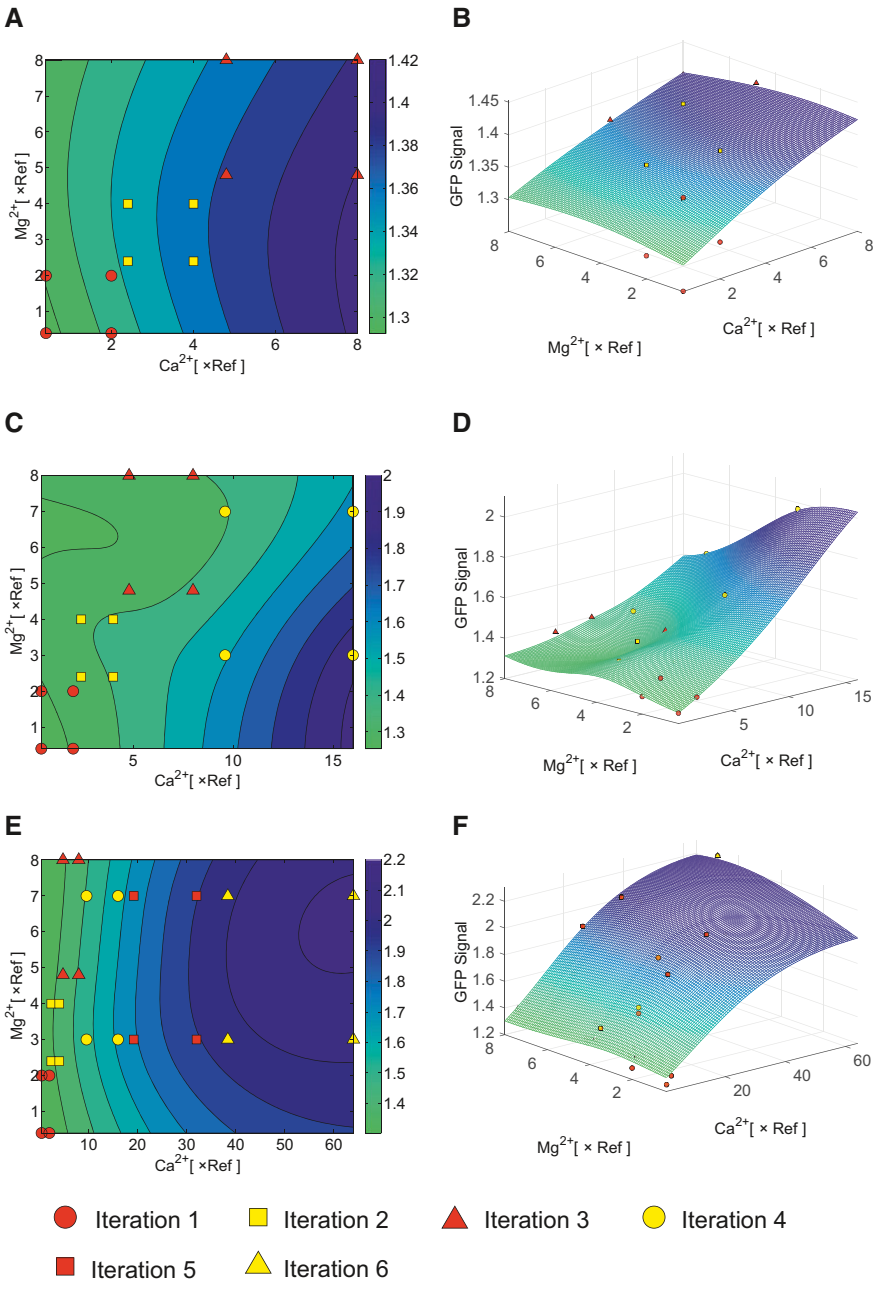


Figure 3. Contour plots (left column) and 3D plots (right column) of the Kriging-interpolated GFP signal. Results of iterations 3, 4, and 6 are shown in rows 1, 2, and 3.

factorial in the upper right quarter of the extended parameter space with one point located near to the maximal concentration level of Ca^{2+} and Mg^{2+} in iteration 1, see Fig. 3A. Experimental results and the Kriging interpolation of iteration 2 confirmed the positive correlation of the GFP signal with Ca^{2+} and Mg^{2+}. Therefore, further enhancement of GFP signal was expected for increased concentrations of both Ca^{2+} and Mg^{2+}.

The experiments of the third iteration were planned in a similar manner as in the second iteration: The maximal concentration of Ca^{2+} and Mg^{2+} was again doubled and a full factorial was placed in the upper right part of the new parameter space. Experimental data of the iterations 1–3 and the associated Kriging interpolation are presented in Fig. 3A and B. Apparently, samples with a Mg^{2+} concentration of 8 × Ref do not lead to higher GFP signals compared to samples with a Mg^{2+} concentration of 4 × Ref, indicating that an optimal concentration range of Mg^{2+} has potentially been found. However, still no indication for a saturation of the response of GFP for increasing Ca^{2+} concentrations could be found in the third iteration. Consequently, the experimental design for iteration 4 was placed around the potential optimal Mg^{2+} concentration (4 × Ref) and an increase in the Ca^{2+} concentration range from 9.6 × Ref to 16 × Ref.

The results of iteration 4 are visualized in Fig. 3C and D and show that the GFP signal could be doubled compared to the reference cultivations when applying a Ca^{2+} concentration of 16 × Ref and a Mg^{2+} concentration of 3 × Ref. Kriging

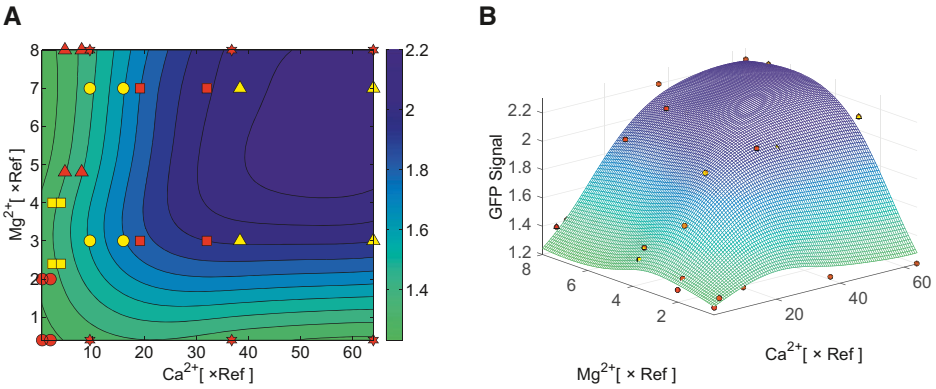


Figure 4. Contour plot (left column) and 3D plot (right column) of the Kriging-interpolated GFP signal after iteration 7.

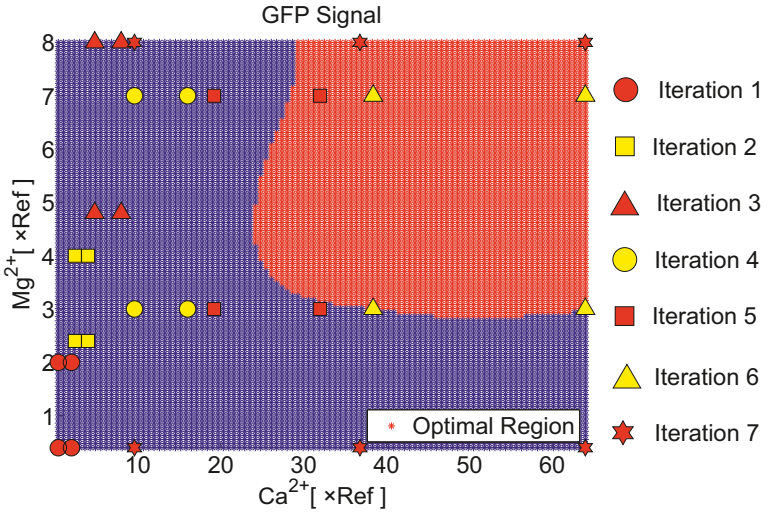


Figure 5. Identification of optimal parameter region based on *z*-test and Kriging interpolation.

interpolation predicts a steep increase in the GFP signal in the Ca^{2+} concentration range from 8 × Ref to 16 × Ref. Although a clear explanation for this phenomenon cannot be given, increased Ca^{2+} concentrations have been described to enhance recombinant protein secretion before [43]. Furthermore, the optimum with respect to Mg^{2+} found in the previous iteration could be confirmed.

Since an optimal Ca^{2+} concentration was not detected yet, the experimental design from iteration 4 was again shifted toward a higher Ca^{2+} concentration range in iteration 5. That is, Ca^{2+} concentration was varied in a range from 19.2 × Ref to 32 × Ref, and range of Mg^{2+} was not changed due to the confirmed optimum. Experiments of iteration 5 and associated Kriging interpolation gave first indications for a limitation of the positive effect of Ca^{2+} on GFP signal (Fig. 3E and F).

In order to verify this finding, the experimental plan for iteration 6 was constructed likewise as for iterations 4 and 5 by doubling the concentration range for Ca^{2+}, see Fig. 3E. Experimental data from iterations 1–6 and the associated Kriging interpolation are depicted in Fig. 3F. The Kriging interpolation clearly shows a flattening of the GFP signal at high Ca^{2+} concentrations. As will be discussed in more detail in Section 3.6, the plateau is most likely caused by precipitation of solid Ca-complexes in combination with other medium components, which leads to limited accessibility of soluble Ca^{2+} ions to the cells. Hence, no further increase in GFP signal is expected for higher Ca^{2+} concentrations due to the formation of solid non-bioavailable Ca-complexes and not due to saturated uptake of Ca^{2+} by the biological system. Moreover, it appears that the negative effect of high Mg^{2+} concentrations can be neutralized by high Ca^{2+} concentrations.

The final iteration 7 was planned with the intention to explore the boundaries of the parameter space, i.e. for the applied minimal and maximal Mg^{2+} concentrations. Results are shown in Fig. 4. The Kriging interpolation reveals that very low Mg^{2+} concentrations lead to a significant decrease in the GFP signal, especially in the case of high Ca^{2+} concentrations. Consequently, the data of Fig. 4 justifies the assumption that both cations Ca^{2+} and Mg^{2+} cannot replace each other.

3.4 Identifying optimal parameter regions

The results of Section 3.3 show that an optimal region with respect to Ca^{2+} and Mg^{2+} concentration exists. Identifying the boundaries of this region is a nontrivial task since noisy measurements and low number of data points may lead to inaccuracies in the Kriging prediction model, but Kriging provides an estimation of the prediction error. This allows to make a statement about the significance of differences in predicted output signals.

In Section 2.6, a method is described that is based on a *z*-test and uses the Kriging interpolation as well as its estimated prediction error. Figure 5 visualizes the results of the *z*-test: The blue region indicates medium compositions that lead to significantly lower GFP signals than the best prediction. Parameter values in the red region will most likely lead to high GFP signals. The identified optimal region is located in the upper right corner of the investigated design space that is, high Mg^{2+} and Ca^{2+} concentrations. It is not expected that higher concentrations of Mg^{2+} and Ca^{2+} ions will lead to further improvements of GFP secretion. Considering Mg^{2+}, the results of iteration 3 showed that for a low Ca^{2+} concentration, high amounts of Mg^{2+} lead eventually to decreasing values in the GFP signal. It seems likely that this also holds for high Ca^{2+} concentration, although the positive effect is believed to be limited by precipitation of Ca^{2+} salts.

3.5 Validation screening

The identified optimal concentrations for Ca^{2+} and Mg^{2+} cause a doubled GFP signal compared to the reference composition of CgXII medium. To confirm the optimized medium composition finally, a validation screening was conducted to show that the results of the screening analysis in Section 3.2 are also valid with optimized concentrations of Ca^{2+} and Mg^{2+}. Consequently, all initial medium components of interest were varied except for Ca^{2+} and Mg^{2+}. These were fixed at values of $Ca^{2+}_{opt} = 32 \times Ref$ and $Mg^{2+}_{opt} = 6.8 \times Ref$ that yield optimal GFP signals and result in feasible pipetting volumes for the liquid handling system (compare Section 2.2). This sample point is part of the predicted optimal region in Section 3.4, near the left border (see Fig. 5), as higher Ca^{2+} concentrations are not expected to cause a significant increase in GFP signal. The applied fractional factorial is similar to the experimental design of the initial screening with a resolution of IV and allows an analysis of main factor effects without confounding with pairwise interaction. Additionally, eleven samples were placed at the defined optimal point for assessing biological reproducibility.

Figure 6A depicts the statistics of ten normalized GFP signal curves at the optimal point. The remaining eleventh curve was excluded as an experimental outlier as it differed significantly from the other curves. It can be seen that the mean normalized GFP value from the cultivations with optimized Ca^{2+} and Mg^{2+} concentration is doubled compared to the reference cultivations. The mean value of the curves converges to $2 \times Ref$. The relative standard deviation of the GFP signal increases with time but never exceeds 3.7%, which is below the maximum standard deviation (5%) of the BioLector device measurements, according to manufacturer data.

The screening analysis results are visualized in Fig. 6B and indicate that NH_4^+ still has a strong negative effect on the GFP signal. Remaining components seem to have a very low or no effect. It can consequently be concluded that changes in Ca^{2+} and Mg^{2+} concentrations do not affect the impact of the other medium components on the GFP signal.

For the used expression strain, it was reported that the majority of the mature GFP is located extracellularly [29]. Consequently, the GFP fluorescence of the fermentation suspension can be considered to be appropriate for capturing the amount of

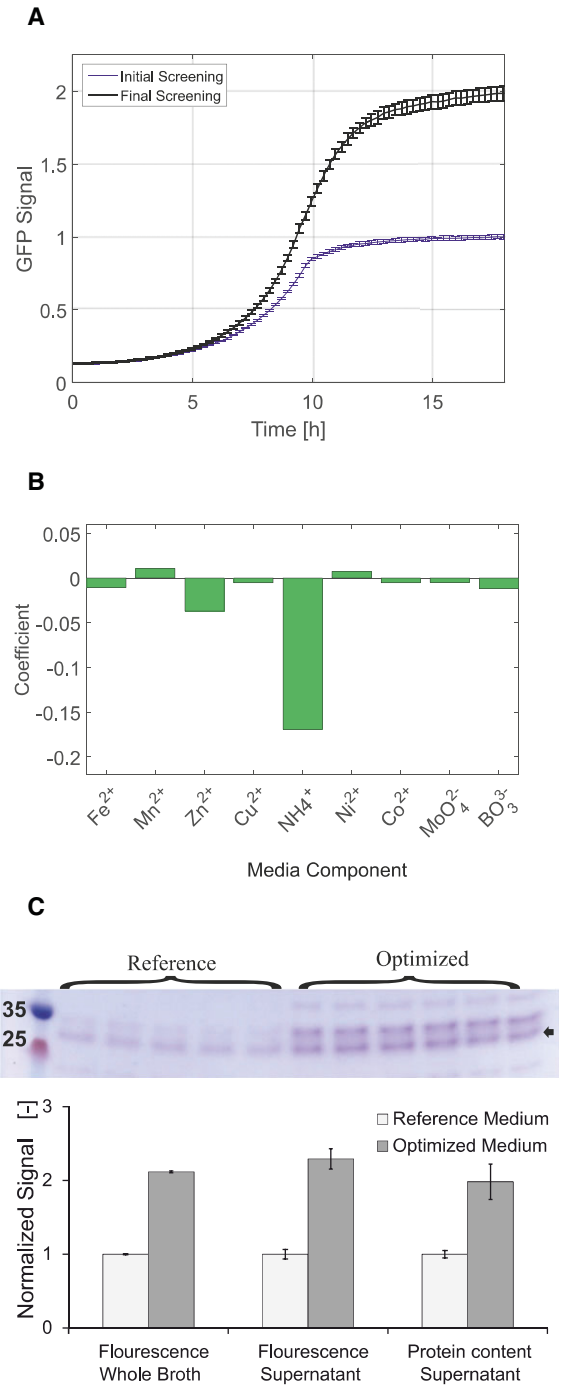


Figure 6. (A) Comparison normalized GFP signal over time for reference cultivations ($n = 16$ replicates) and cultivations with optimized medium composition as determined after iterative optimization supported by Kriging interpolation ($n = 10$ replicates). Mean values ± standard deviation are shown. (B) Impact of medium components on GFP signal at optimized medium composition. (C, top) SDS-page analysis of cultivation supernatants in reference medium ($n = 5$) and in optimized medium ($n = 6$), GFP band (26 kDa) is indicated by an arrow. (C, bottom) Comparison of GFP fluorescence from whole cultivation broth and supernatant after cultivation and protein content in supernatant, made from reference ($n = 6$) and optimized medium ($n = 6$).

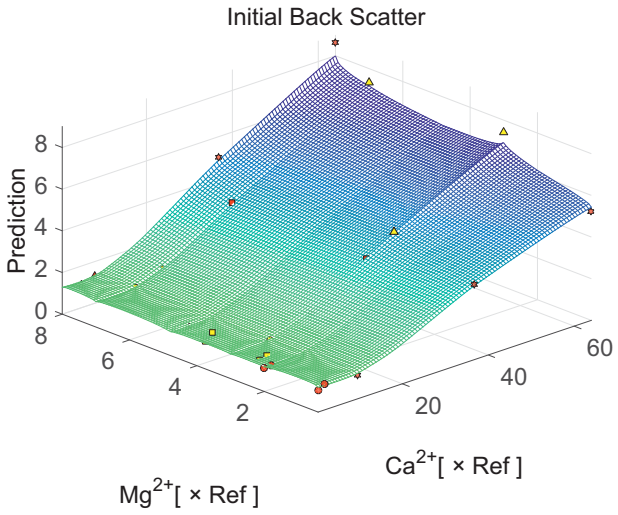


Figure 7. Kriging interpolation of functional relationship between medium component Ca^{2+} and Mg^{2+} and backscatter as indicator for precipitation.

secreted GFP. However, additional analytics have been conducted from a separate cultivation run with six biological replicates for both reference and optimized medium composition. As depicted in Fig. 6C, both fluorescence and protein content of cultivation supernatants from optimized cultivations have doubled compared to those from reference medium. This is in agreement with GFP signals from the whole fermentation suspension. Indeed, analysis by SDS-page verifies that the amount of secreted GFP is higher for the optimized medium.

3.6 Interpretation of results

Response data of the optimized medium composition was reevaluated to give indication of the previously speculated underlying reason of the limited effect of increased Ca^{2+} and Mg^{2+} addition. In cultivation wells with high $CaCl_2$ concentrations (approximately 0.1–5.5 mM, i.e. ~ 10 to 50 × Ref), white turbidity of the medium was observed by optical inspection. Therefore, it can be assumed that Ca^{2+} forms complexes with other medium components like PO_4^{3-}, which results in precipitation as soon as Ca^{2+} concentration reaches a certain threshold in CgXII medium. Consequently, associated calcium ions are temporarily not accessible to the microorganism, but may redissolve during growth when soluble Ca^{2+} is incorporated into increasing biomass. However, associated reduction in turbidity will be covered by increasing turbidity due to biomass formation.

For the optimized medium, only concentrations of Ca^{2+} and Mg^{2+} have been increased, thus cultivation data for all iterations regarding initial backscatter have been related to the applied concentration of those cations. We used Kriging for interpolating the functional relationship between the concentration of Ca^{2+} and Mg^{2+} and backscatter. As illustrated in Fig. 7, the backscatter shows a positive correlation with increasing Ca^{2+} concentration, which is most likely caused by precipitation. Mg^{2+} shows only a positive influence for high Ca^{2+} concentration. Thus, the conclusion is hardened by combinatorial output from qualitative optical properties showing turbidity of medium, experimental output in terms of limited effect of increased Ca^{2+} addition on GFP secretion and statistical evaluation of presumed correlation between initial backscatter and Ca^{2+} concentration.

Despite the verified finding that increased concentrations of Ca^{2+} promote enhanced secretion of recombinant GFP in *C. glutamicum*, the underlying reason is not clear. In the following, three hypotheses are presented:

In some cases, an increase of Ca^{2+} ions in the medium has been reported to increase the amount of active (i.e. correctly folded) secreted protein. This phenomenon was explained by the need for divalent cations to support correct folding of the target protein that incorporates those cations as cofactor. In case of GFP, such an effect can be ruled out here, as GFP does not need divalent cations as folding cofactors [46].

Another hypothesis considers the cell wall as several layers of a molecular sieve acting as a depth filter, whose pores are permeable for endogenous extracellular proteins due to evolutionary adaption. This may not be the case for secretory heterologous proteins during their passage to the outside of the cell, but the incorporation of Ca^{2+} may enlarge the pore size of this assumed depth filter around the cell which the facilitates the release of recombinant GFP.

Furthermore, Ca^{2+} ions might neutralize negative charges of the "continuum of anionic charge" formed by lipoteichonic acids and wall teichonic acids found in the cell wall of gram positives [47]. In combination with indications that GFP presents positive charges on the outer side [48], reduction of electrostatic interaction could facilitate the passage of GFP to the extracellular medium.

For *Bacillus brevis* it was shown that with increasing concentrations of $MgSO_4$, $MgCl_2$, or $CaCl_2$ (up to 5 mM) the expression of certain cell wall proteins is remarkably decreased [49]. In our study, we identified optimal concentrations of ≈ 7 mM (6.8 × Ref) for $MgSO_4 \cdot 7\ H_2O$ and ≈ 2.9 mM (32 × Ref) of $CaCl_2 \cdot H_2O$, which are in the same order of magnitude. Maybe a similar phenomenon as described for *B. brevis* occurs also in *C. glutamicum* as it is also a gram-positive organism. Thus, it is speculated that a high load of Ca^{2+} causes downregulation of cell wall protein synthesis and induces a "leaky" cell wall.

However, the last two assumptions imply that fluorescence of cell wall retained GFP is shadowed compared to GFP located in the medium and clearly, the presented hypotheses remain speculative.

4 Concluding remarks

In this study, we have presented an experimental optimization strategy that merges three different aspects: (1) Current technology of microbioreactor systems, (2) lab automation, and (3) Kriging-based experimental analysis and design. The overall optimization goal was defined by maximization of the product titer of secreted GFP, synthesized by *C. glutamicum*, through variation of media component concentrations.

Starting with 11 media components, a fractional factorial-based screening analysis was performed in order to reduce the number of design parameters. While too high concentrations of

NH_4^+ had a strong negative impact, only the divalent metal ions Ca^{2+} and Mg^{2+} showed a clear positive effect on the performance criterion.

In an iterative Kriging-based procedure, the functional relationship between the divalent metal ions and the protein titer was investigated in more detail. In each iteration, Kriging was used for approximating and visualizing this relationship. Based on Kriging estimation, hypotheses were formulated and used for planning new experiments.

Instead of defining one optimal point, an optimal parameter region was identified based on a statistical hypothesis test. In this region, high protein titer values are expected. For one of the optimal parameter sets, it was verified that the results of the initial screening are still valid and consequently, the optimization based on only two components is justified.

As major result, by increasing the Ca^{2+} and Mg^{2+} concentrations, the performance criterion could be more than doubled compared to its initial value. However, the positive effect of both components is limited. In case of Ca^{2+}, the limitation was caused by precipitation.

Compared to other statistical optimization procedures, such as response surface by polynomial approximation, the suggested Kriging-based optimization approach has the advantage that all successively collected data are used. This allows a global approximation of the dependency of the performance criterion on the input variables and consequently a robust optimization. Moreover, Kriging is well suited to approximate even highly nonlinear functional relationships and the provided prediction-uncertainty can be used for statistical analysis. The presented workflow can easily be adapted to other experimental settings, e.g. bioprocess operation or genetic engineering of recombinant expression hosts.

Practical application

Advanced microbioreactor systems, lab automation, Kriging-based data analysis, and experimental design are combined for process optimization. An iterative strategy allows to effectively cope with the combinatorial explosion of impact factors typically encountered in optimizing industrial production processes. The approach is demonstrated by maximizing the product titer of an example cultivation through variation of growth medium composition.

The authors thank the Ministry of Innovation, Science, and Research of North Rhine-Westphalia and the Heinrich Heine University Düsseldorf for a scholarship to Lars Freier within the CLIB-Graduate Cluster Industrial Biotechnology. The scientific activities of the Bioeconomy Science Center were financially supported by the Ministry of Innovation, Science, and Research within the framework of the NRW Strategieprojekt BioSC (No. 313/323-400-002 13*).*

The authors have declared no conflicts of interest.

Nomenclature

Z	[-]	Output/process variable
Z^*	[-]	Kriging estimation
$\boldsymbol{x}$	[-]	Input/design variable
$\hat{\boldsymbol{x}}$	[-]	Point of interest
Cov	[-]	Covariance
Var	[-]	Variance
C	[-]	Covariance matrix
H_0	[-]	Null hypothesis
i×Ref	[-]	i-times reference concentration of the CgXII medium
$\sigma(\boldsymbol{x})$	[-]	Standard deviation of the random process
$m(\boldsymbol{x})$	[-]	Mean/trend function of random process
λ	[-]	Kriging coefficient
α	[-]	Significance level

5 References

[1] Li, S., Yang, X., Yang, S., Zhu, M. et al., Technology prospecting on enzymes: Application, marketing and engineering. *Comput. Struct. Biotechnol. J.* 2012, *2*, 1–11.

[2] Hartner, F. S., Ruth, C., Langenegger, D., Johnson, S. N. et al., Promoter library designed for fine-tuned gene expression in *Pichia pastoris*. *Nucleic Acids Res.* 2008, *36*, e76–e76.

[3] Degering, C., Eggert, T., Puls, M., Bongaerts, J. et al., Optimization of protease secretion in *Bacillus subtilis* and *Bacillus licheniformis* by screening of homologous and heterologous signal peptides. *Appl. Environ. Microbiol.* 2010, *76*, 6370–6376.

[4] Zhang, G., Hubalewska, M., Ignatova, Z., Transient ribosomal attenuation coordinates protein synthesis and co-translational folding. *Nat. Struct. Mol. Biol.* 2009, *16*, 274–280.

[5] Hamilton, S. R., Gerngross, T. U., Glycosylation engineering in yeast: The advent of fully humanized yeast. *Curr. Opin. Biotechnol.* 2007, *18*, 387–392.

[6] Pinsach, J., de Mas, C., López-Santín, J., Induction strategies in fed-batch cultures for recombinant protein production in *Escherichia coli*: Application to rhamnulose 1-phosphate aldolase. *Biochem. Eng. J.* 2008, *41*, 181–187.

[7] Jian Li, Z., Zhao, Q., Liang, H., Jiang, S. et al., Control of recombinant human endostatin production in fed-batch cultures of *Pichia pastoris* using the methanol feeding rate. *Biotechnol. Lett.* 2002, *24*, 1631–1635.

[8] Puertas, J., Ruiz, J., de la Vega, M., Lorenzo, J. et al., Influence of specific growth rate over the secretory expression of recombinant potato carboxypeptidase inhibitor in fed-batch cultures of *Escherichia coli*. *Proc. Biochem.* 2010, *45*, 1334–1341.

[9] Pinsach, J., de Mas, C., López-Santín, J., Striedner, G. et al., Influence of process temperature on recombinant enzyme activity in *Escherichia coli* fed-batch cultures. *Enzyme Microb. Technol.* 2008, *43*, 507–512.

[10] Ramchuran, S. O., Holst, O., Karlsson, E. N., Effect of postinduction nutrient feed composition and use of lactose as inducer during production of thermostable xylanase in *Escherichia coli* glucose-limited fed-batch cultivations. *J. Biosci. Bioeng.* 2005, *99*, 477–484.

[11] Fong, B. A., Wood, D. W., Expression and purification of ELP-intein-tagged target proteins in high cell density *E. coli* fermentation. *Microb. Cell Fact.* 2010, *9*, 77.

[12] Huber, R., Roth, S., Rahmen, N., Büchs, J., Utilizing high-throughput experimentation to enhance specific productivity of an *E. coli* T7 expression system by phosphate limitation. *BMC Biotechnol.* 2011, *11*, 22.
[13] Kensy, F., Engelbrecht, C., Büchs, J., Scale-up from microtiter plate to laboratory fermenter: Evaluation by online monitoring techniques of growth and protein expression in *Escherichia coli* and *Hansenula polymorpha* fermentations. *Microb. Cell Fact.* 2009, *8*, 68.
[14] Hortsch, R., Stratmann, A., Weuster-Botz, D., New milliliter-scale stirred tank bioreactors for the cultivation of mycelium forming microorganisms. *Biotechnol. Bioeng.* 2010, *106*, 443–451.
[15] Rohe, P., Venkanna, D., Kleine, B., Freudl, R. et al., An automated workflow for enhancing microbial bioprocess optimization on a novel microbioreactor platform. *Microb. Cell Fact.* 2012, *11*, 144.
[16] Islam, R. S., Tisi, D., Levy, M. S., Lye, G. J., Scale-up of *Escherichia coli* growth and recombinant protein expression conditions from microwell to laboratory and pilot scale based on matched k_La. *Biotechnol. Bioeng.* 2008, *99*, 1128–1139.
[17] Bareither, R., Pollard, D., A review of advanced small-scale parallel bioreactor technology for accelerated process development: Current state and future need. *Biotechnol. Prog.* 2011, *27*, 2–14.
[18] Long, Q., Liu, X., Yang, Y., Li, L. et al., The development and application of high throughput cultivation technology in bioprocess development. *J. Biotechnol.* 2014, *192*, 323–338.
[19] Lattermann, C., Büchs, J., Microscale and miniscale fermentation and screening. *Curr. Opin. Biotechnol.* 2015, *35*, 1–6.
[20] Gernaey, K. V., Baganz, F., Franco-Lara, E., Kensy, F. et al., Monitoring and control of microbioreactors: An expert opinion on development needs. *Biotechnol. J.* 2012, *7*, 1308–1314.
[21] Kennedy, M., Krouse, D., Strategies for improving fermentation medium performance: A review. *J. Ind. Microbiol. Biotechnol.* 1999, *23*, 456–475.
[22] Weuster-Botz, D., Experimental design for fermentation media development: Statistical design or global random search? *J. Biosci. Bioeng.* 2000, *90*, 473–483.
[23] Zhang, J., Greasham, R., Chemically defned media for commercial fermentations. *Appl. Microbiol. Biotechnol.* 1999, *51*, 407–421.
[24] Lu, C., Bentley, W. E., Rao, G., A high-throughput approach to promoter study using green fluorescent protein. *Biotechnol. Prog.* 2004, *20*, 1634–1640.
[25] Zanzotto, A., Boccazzi, P., Gorret, N., Van Dyk, T. K. et al., In situ measurement of bioluminescence and fluorescence in an integrated microbioreactor. *Biotechnol. Bioeng.* 2006, *93*, 40–47.
[26] Fisher, R. A., *Design of Experiments*, Oliver and Boyd, Edinburgh 1936.
[27] Mandenius, C.-F., Brundin, A., Bioprocess optimization using design-of-experiments methodology. *Biotechnol. Prog.* 2008, *24*, 1191–1203.
[28] Cock, D. R. De., Kriging as an alternative to polynomial regression in response surface analysis, 2003.
[29] Meissner, D., Vollstedt, A., van Dijl, J. M., Freudl, R., Comparative analysis of twin-arginine (Tat)-dependent protein secretion of a heterologous model protein (GFP) in three different Gram-positive bacteria. *Appl. Microbiol. Biotechnol.* 2007, *76*, 633–642.
[30] Unthan, S., Grünberger, A., van Ooyen, J., Gätgens, J. et al., Beyond growth rate 0.6: What drives *Corynebacterium glutamicum* to higher growth rates in defined medium. *Biotechnol. Bioeng.* 2014, *111*, 359–371.
[31] Wong, C., Sridhara, S., Bardwell, J. C., Jakob, U., Heating greatly speeds coomassie blue staining and destaining. *Biotechniques* 2000, *28*, 426–428, 430, 432.
[32] Morschett, H., Wiechert, W., Oldiges, M., Automation of a Nile red staining assay enables high throughput quantification of microalgal lipid production. *Microb. Cell Fact.* 2016, *15*, 34.
[33] Cressie, N. A. C., Second-order stationarity, in: *Statistics for Spatial Data*, 3rd ed., Wiley, New York 1993, p. 53.
[34] Marchant, B. P., Lark, R. M., The Matérn variogram model: Implications for uncertainty propagation and sampling in geostatistical surveys. *Geoderma* 2007, *140*, 337–345.
[35] Gorsich, D., Genton, M., Variogram model selection via nonparametric derivative estimation. *Math. Geol.* 2000, *32*, 249–270.
[36] Diggle, P., Tawn, J., Moyeed, R., Model-based geostatistics. *J. R. Stat. Soc. Ser. C* 1998, *47*, 300.
[37] Dubrule, O., Cross validation of Kriging in a unique neighborhood. *J. Int. Assoc. Math. Geol.* 1983, *15*, 687–699.
[38] MathWorks, Genetic algorithm. 2015.
[39] MathWorks, Constrained nonlinear optimization algorithms. 2015.
[40] Mariappan, P., *Biostatistics: An Introduction*, Pearson Education, India 2013.
[41] Keilhauer, C., Eggeling, L., Sahm, H., Isoleucine Synthesis in *Corynebacterium glutamicum*: Molecular analysis of the ilvB-ilvN-ilvC operon. *J. Bacteriol.* 1993, *175*, 5595–5603.
[42] Weuster-Botz, D., Kelle, R., Frantzen, M., Wandrey, C., Substrate controlled fed-batch production of L-lysine with *Corynebacterium glutamicum*. *Biotechnol. Prog.* 1997, *13*, 387–393.
[43] Teramoto, H., Watanabe, K., Suzuki, N., Inui, M. et al., High yield secretion of heterologous proteins in *Corynebacterium glutamicum* using its own Tat-type signal sequence. *Appl. Microbiol. Biotechnol.* 2011, *91*, 677–687.
[44] Brockmeier, U., Caspers, M., Freudl, R., Jockwer, A. et al., Systematic screening of all signal peptides from *Bacillus subtilis*: A powerful strategy in optimizing heterologous protein secretion in gram-positive bacteria. *J. Mol. Biol.* 2006, *362*, 393–402.
[45] Zimmerman, D. L., Optimal network design for spatial prediction, covariance parameter estimation, and empirical prediction. *Environmetrics* 2006, *17*, 635–652.
[46] Zimmer, M., Green fluorescent protein (GFP): Applications, structure, and related photophysical behavior. *Chem. Rev.* 2002, *102*, 759–782.
[47] Silhavy, T. J., Kahne, D., Walker, S., Envelope, T. B. C. *Cold Spring Harb. Perspect. Biol.* 2010, *2*, a000414–a000414.
[48] Umakoshi, H., Nishida, M., Suga, K., Bui, H. T. et al., Characterization of green flourescent protein using aqueous two-phase systems. *Solvent Extr. Res. Dev.* 2009, *16*, 145–150.
[49] Adachi, T., Yamagate, H., Tsukagoshi, N., Udaka, S., Repression of the cell wall protein gene operon in *Bacillus brevis* 47 by magnesium and calcium ions. *J. Bacteriol.* 1991, *173*, 4243–4245.

2.3 A framework for accelerated phototrophic bioprocess development: integration of parallelized microscale cultivation, laboratory automation and Kriging-assisted experimental design

Morschett *et al. Biotechnol Biofuels (2017) 10:26*
DOI 10.1186/s13068-017-0711-6

Biotechnology for Biofuels

RESEARCH Open Access

A framework for accelerated phototrophic bioprocess development: integration of parallelized microscale cultivation, laboratory automation and Kriging-assisted experimental design

Holger Morschett[1†], Lars Freier[1†], Jannis Rohde[1], Wolfgang Wiechert[1], Eric von Lieres[1*] and Marco Oldiges[1,2*]

Abstract

Background: Even though microalgae-derived biodiesel has regained interest within the last decade, industrial production is still challenging for economic reasons. Besides reactor design, as well as value chain and strain engineering, laborious and slow early-stage parameter optimization represents a major drawback.

Results: The present study introduces a framework for the accelerated development of phototrophic bioprocesses. A state-of-the-art micro-photobioreactor supported by a liquid-handling robot for automated medium preparation and product quantification was used. To take full advantage of the technology's experimental capacity, Kriging-assisted experimental design was integrated to enable highly efficient execution of screening applications. The resulting platform was used for medium optimization of a lipid production process using *Chlorella vulgaris* toward maximum volumetric productivity. Within only four experimental rounds, lipid production was increased approximately threefold to 212 ± 11 mg L^{-1} d^{-1}. Besides nitrogen availability as a key parameter, magnesium, calcium and various trace elements were shown to be of crucial importance. Here, synergistic multi-parameter interactions as revealed by the experimental design introduced significant further optimization potential.

Conclusions: The integration of parallelized microscale cultivation, laboratory automation and Kriging-assisted experimental design proved to be a fruitful tool for the accelerated development of phototrophic bioprocesses. By means of the proposed technology, the targeted optimization task was conducted in a very timely and material-efficient manner.

Keywords: Biodiesel, *Chlorella vulgaris*, Design of experiments, Kriging, Lipid production

Background

By virtue of significant advantages offered over agricultural crops [1–6], microalgae are generally accepted as promising feedstock for bio-economy applications [7–9]. However, until now, their industrial exploitation remains mostly uneconomic, especially when lower-priced products like biofuels are targeted [10]. Currently, the integrated utilization of biomass is intensively investigated as a promising concept to improve the overall efficiency in terms of cost and energy [9, 11, 12]. In this context, intracellular lipids represent a compound class of special interest as they can be either transesterificated to biodiesel [7] or boost the nutritional quality of algae for functional food applications [13].

Regarding phototrophic bioprocess development, early-stage strain and parameter screening are of crucial importance to the successful set-up of economic

*Correspondence: e.von.lieres@fz-juelich.de; m.oldiges@fz-juelich.de
†Holger Morschett and Lars Freier contributed equally to this work
[1] Forschungszentrum Jülich GmbH, Institute of Bio- and Geosciences, IBG-1: Biotechnology, Wilhelm-Johnen-Straße, 52428 Jülich, Germany
Full list of author information is available at the end of the article

processes [14]. Today, these aspects are typically studied by means of only marginally parallelized reactor systems like shake flasks, test tubes or even single-vessel reactors [15–18]. Consequently, experimental throughput is fairly limited rendering screening tasks rather laborious and highly time consuming. Only recently a strong demand for high throughput micro-photobioreactors has been identified, based on which some prototype systems have been developed [19–27]. To take full advantage of phototrophic microscale cultivation, supporting methodologies and technologies, such as simplified strain maintenance [28], high throughput analytics [29, 30], and automated processing [20, 31], are needed.

In the medium term, these initial developments and especially further progress in high throughput technology and laboratory automation will clearly boost the efficiency of phototrophic process development. Phototrophic processes are characterized by their intrinsic complexity induced by a high number of potentially interacting input variables. Hence, experimental capacities, i.e., mainly cultivation, will be always a crucial factor due to the trade-off between throughput and the necessary laboratory resources. Current micro-photobioreactors mainly rely on standardized microtiter plates. Thus, a further rise of cultivation capacity by intensifying parallelization would need to be based on scale out and inevitably be accompanied by increasing cost. Hence, an alternative strategy to focus cultivation activities on only the most informative experiments is the ultimate solution to tackle the omnipresent challenge of restricted experimental throughput.

One approach to achieve this efficiently, which is already well-established for microbial bioprocess development [32], is the use of Design of Experiments (DoE) to focus on experiments providing the highest information content in a targeted parameter space. Despite having been established during the early twentieth century [33], there is still ongoing research into this methodology [34]. This approach is regarded to be particularly suitable to deal with the combinatorial explosion typically occurring when investigating multi-parameter relations [35]. Moreover, DoE overcomes a critical limitation of "conventional" one-factor-at-a-time experiments, as such approaches often fail in locating global optima by not taking potentially synergistic or antagonistic interactions of input variables into account [36]. Regarding bioprocess development, the most prominent application of DoE is the culture media optimization [32]. During such tasks, the omnipresent interactions between single compounds render locating a global optimum by "conventional" experimental planning to chance.

In the above context, the current study aims at the combination of emerging technologies for parallelized microscale cultivation and analytics to phototrophic microorganisms with elaborate experimental design as has previously been fruitfully applied for heterotrophic systems by [37]. Thereby, an integrated framework for the accelerated development of phototrophic bioprocesses is to be set up. Optimizing medium composition toward maximized lipid productivity of the unicellular microalga *Chlorella vulgaris* was chosen as a model process for the above purpose.

Methods

Chemicals, strain

All chemicals were purchased either from Sigma-Aldrich (Steinheim/Germany) or Roth (Karlsruhe/Germany) and were of analytical grade. The unicellular microalga *C. vulgaris* 211-11b [38], purchased from the Culture Collection of Algae at the University of Göttingen (Germany), was used throughout all cultivation experiments.

Medium

Cultivations were carried out in variations of an enriched Bold's Basal Medium [39] prepared from stock solutions. The previously established reference medium [27, 28] was composed of chemicals as follows: 9.76 g L^{-1} 2-(*N*-morpholino)ethanesulfonic acid (MES), 0.6 g L^{-1} K_2HPO_4, 1.4 g L^{-1} KH_2PO_4, 1.5 g L^{-1} $NaNO_3$, 187.5 mg L^{-1} $MgSO_4{\cdot}7\ H_2O$, 6.25 mg L^{-1} NaCl, 125 mg L^{-1} $CaCl_2{\cdot}2\ H_2O$, 17.64 mg L^{-1} $ZnSO_4{\cdot}7\ H_2O$, 2.88 mg L^{-1} $MnCl_2{\cdot}4\ H_2O$, 2.4 mg L^{-1} $Na_2MoO_4{\cdot}2\ H_2O$, 3.14 mg L^{-1} $CuSO_4{\cdot}5\ H_2O$, 0.94 mg L^{-1} $CoSO_4{\cdot}7\ H_2O$, 22.8 mg L^{-1} H_3BO_3, 9.96 mg L^{-1} $FeSO_4{\cdot}7\ H_2O$, 3.68 mg L^{-1} H_2SO_4, 100 mg L^{-1} $Na_2EDTA{\cdot}2\ H_2O$, 62 mg L^{-1} KOH and 100 mg L^{-1} penicillin-G sodium salt. The pH value was set to 6.5 with 5 M NaOH. Ultrapure water (type 1) was used for the preparation of all cultivation media.

During optimization experiments, the medium composition was varied according to the respective experimental plan by adjusting the applied volumes of the individual stock solutions. These media variants were prepared by a liquid-handling platform as previously described in literature. Medium preparation was carried out in a fully automated manner, while a surrounding laminar flow hood ensured sterile conditions [34, 40, 41]. Media were prepared at 2.5 mL scale in an MTP-R-48-B "Round Well Plate" (m2p-labs, Baesweiler/Germany) under continuous shaking at 500 rpm on an integrated Teleshake 95 (Inheco, Martinsried/Germany), while the minimum volume to be pipetted was set to 10 µL. Thereby, achieving sufficiently high accuracy (±0.3%) and precision (±0.3%) could be ensured [34]. Subsequently, 950 µL of each medium was transferred to a well of an MTP-48-B "FlowerPlate®" (m2p-labs, Baesweiler/Germany) in which the cultivation took place (see "Main cultivation" section).

Strain maintenance and pre-cultivation

Chlorella vulgaris was maintained as glucose-adapted cryocultures. Preserved cells were re-adapted to light during phototrophic pre-cultivation in illuminated shake flasks. A detailed description of strain maintenance and pre-cultivation strategy can be obtained from [28]. After 60 h of incubation, the cells were harvested by 5 min centrifugation at 3939×*g* and 4 °C in a Labofuge 400R (Heraeus Instruments, Hanau/Germany). The supernatant was discarded, and the pellet re-suspended in 0.9% (w v^{-1}) NaCl to a biovolume of 2 µL mL^{-1} to generate the stock solution required for the inoculation of subsequent main cultivations.

Main cultivation

Main cultivations were conducted in pre-sterilized, disposable 48-well MTP-48-B "FlowerPlates®" (m2p-labs, Baesweiler/Germany). Each well was filled with 950 µL of medium and inoculated to a biovolume of 0.1 µL mL^{-1} with 50 µL of the inoculation stock solution generated as described in "Strain maintenance and pre-cultivation" section. The plates were sealed using an F-R48-10 "perforated sealing foil for evaporation reduction" (m2p-labs, Baesweiler/Germany), pasted over with an F-GP-AB10 "gas-permeable seal" (m2p-labs, Baesweiler/Germany).

The microtiter plates were incubated using a microphotobioreactor prototype. The system relies on bottom-side illumination with a set of blue and white LEDs and indirect temperature control via placement of the plates in a tempered incubation chamber. A detailed description and the schematic representation of the system are given in [27]. The following cultivation conditions were applied: 25 °C, continuous shaking at 1200 rpm, 3 mm shaking diameter, 2.5% (v v^{-1}) CO_2, 200 µmol m^{-2} s^{-1} photon flux density (constant), and ≥85% relative humidity.

Biomass detection

Optical density (OD) was acquired using 10-mm polystyrene semi-micro cuvettes (ratiolab, Dreieich/Germany) and an UV-1800 photometer (Shimadzu, Duisburg/Germany) at 750 nm, while desalted water served as a blank. If needed, samples were diluted to $OD_{750} \leq 0.3$ using 0.9% (w v^{-1}) NaCl solution to fit the linear range of the photometer.

The biovolume was measured taking advantage of a particle counter (MultiSizer 3, Beckman Coulter, Krefeld/Germany) using the "Coulter principle" [42]. The device was equipped with a 30 µm capillary which had been calibrated using a suspension of 3 µm latex beads (Beckman Coulter, Krefeld/Germany) according to the manufacturer's specification and was operated in volumetric control mode. Prior to measurement, cell suspensions were diluted to $OD_{750} \leq 0.025$ in CASYton buffer (Omni Life Science, Bremen/Germany), and only particles in the range of 1.8–14 µm were analyzed.

The cell dry weight was determined by means of gravimetry. Culture liquid from two replicate wells of a microtiter plate was pooled to obtain sufficient sample amounts for the analysis. Cells were spun down in pre-dried and weighed 2-mL reaction tubes for 5 min at 16,060×*g* (Biofuge Pico, Heraeus Instruments, Hanau/Germany). The supernatants were discarded and the pellets freeze-dried in an LT-105 freeze dryer (Christ Gefriertrocknungsanlagen, Osterode am Harz/Germany) until attaining a constant weight. After acclimatization to room temperature in a desiccator, weighing was repeated, and the cell dry weight was derived from the resulting mass difference.

Lipid quantification

The intracellular accumulation of neutral lipids was quantitatively monitored by means of an automated high throughput Nile red staining assay as previously described in [29].

Nitrate quantification

Cells were removed by filtration using 0.2 µm cellulose acetate syringe filters (DIA-Nielsen, Düren/Germany), and the cell-free supernatant was stored at −20 °C prior to analysis, if needed. Nitrate was quantified using the Spectroquant 1.09713.0002 nitrate test (Merck, Darmstadt/Germany) according to the manufacturer's specifications, scaled down to one quarter of the recommended volume. Supernatants were pre-diluted with desalted water to fit the linear range of the assay, if needed. The measurements were conducted in UV semi-micro cuvettes (Brand, Wertheim/Germany) using an UV-1800 photometer (Shimadzu, Duisburg/Germany).

Acquisition of fatty acid fingerprints

Lyophilized biomass from cell dry weight determination (see Sect. 2.5) was in-situ transesterificated using acidic methanol [10% (w w^{-1}) H_2SO_4], and the resulting fatty acid methyl esters were subsequently extracted with heptane. Semi-quantitative fingerprints were accessed by gas chromatography time-of-flight mass spectrometry of the extracts. A detailed description of the methodology can be obtained from [43].

Experimental design

Media composition was optimized with respect to lipid productivity using a Design of Experiments methodology. The applied optimization strategy was adopted from [34]. Initially, fractional and full factorial experimental designs were applied for estimating single component effects and combinatorial interactions. Myers et al. [44] provide a good overview of these classical DoE methods.

Based on the initially collected data, the statistically more advanced concept of Kriging was applied for data analysis, visualization, and for designing further experiments with potentially improved lipid productivity. Kriging is an interpolation method that provides unbiased approximations of the underlying nonlinear functional relationships between media composition and lipid productivity with minimal prediction error. This method originates in geostatistics and has recently been adapted for optimizing biotechnology processes [34]. Further mathematical details of the Kriging method can be found in the monograph of Cressie [45]. The statistical analysis tools applied in this study are part of the open source Kriging toolkit "KriKit", which can be freely downloaded at https://github.com/modsim/KriKit.

Expected Improvement

Given a Kriging model of the current dataset, further experiments were designed to maximize the Expected Improvement (EI). This experimental design strategy seeks a compromise between maximizing lipid productivity and reducing prediction uncertainty of the Kriging approximation in relevant regions of the parameter space [46]. In a comparative study, EI has been found to outperform other sampling strategies in Kriging-based optimization [47].

In sequential optimization, new experiments are typically planned at maximal EI. Parallel experiments, as in the present study, are most efficiently planned by sampling from the EI distribution. In a non-deterministic sampling process, using the Markov Chain Monte Carlo (MCMC) method, new experiments are selected with probability proportional to their EI. Naturally, experiments with high EI are preferred over experiments with lower EI, which nonetheless have a reduced chance of being selected, while experiments with zero EI are strictly excluded. Freier et al. have demonstrated that MCMC sampling can significantly reduce the number of required experiments in process optimization [48]. In the present study, the Delay Rejection Adaptive Metropolis algorithm [49] was applied with a chain length of 10,000 elements, of which the first 1000 are discarded (burn in phase of the MCMC method).

Results and discussion

Choice of relevant media components

The medium targeted for optimization incorporates 17 different components (see "Medium") with phosphate salts counted as one compound due to their pH-dependent equilibrium. This number is too high to efficiently perform the experimental study with a manageable number of experiments, since a full factorial design with two concentration levels would result in $2^{17} \approx 130{,}000$ experiments. In order to keep the number of components of interest, preselection was completed based on the literature information. Table 1 summarizes the known biological effects of the individual components. Penicillin-G concentration was kept constant under all conditions, and all trace elements were clustered to one single input variable as a similar effect on cultivation was expected. Sulfuric acid and potassium hydroxide had to be varied together with $FeSO_4$ and Na_2EDTA, respectively, as they were needed to keep the latter two components dissolved in their stock solutions. Thereby, the number of input variables was reduced by almost 50% from 17 to 9.

Kriging-assisted optimization

Fractional factorial

Starting with the nine remaining media components of interest, a full factorial design would require $2^9 \approx 500$ experiments. Making full use of 48-fold parallelized microtiter plate cultivation (see "Main cultivation" section), this leads to a total of 11 experimental runs, equivalent to 4 months of cultivation time. In 12-fold parallelized shake flasks, the experiments would even take 14 months. Yet, such time scales are clearly far from feasible, underlining the necessity to effectively reduce the experimental effort.

Fractional factorial designs allow the reduction of the number of experiments by estimating only single component effects and a subset of combinatorial effects [44]. The chosen design (see Additional file 1 for both, design and corresponding measurement data) comprises 37 experiments, five of which represent the reference point using the $enBBM_{ref}$ medium (see Additional file 2 for medium composition). Taking reference points into account allows for the investigation of measurement noise and normalization. The other experiments allowed for a statistical analysis of the effect of single components, as well as the interaction with magnesium ions. The interaction with this divalent metal ion was analyzed, as it is reported to be an effector of the acetyl-CoA carboxylase, an enzyme essential for lipid biosynthesis responsible for the initial step of carbon dioxide fixation to malonyl-CoA (see Table 1). An overview of the functionality of this enzyme complex and its regulation is given by Ohlrogge and Browse [67]. Thus, any interactions with this input variable are of special interest with respect to product accumulation in the cells.

Figure 1a shows the resulting statistical analysis of the fractional factorial experiments. The green bars indicate the expected effect of varying the medium concentrations between their minimal and maximal values (see Additional file 1). The error bars indicate the uncertainty of the estimations. In the following, the main and combinatorial effects of the components are checked for significance using a t test with a significance level of $p = 0.1$. Using a lower significance level would increase the risk of

Table 1 Initial evaluation of the medium components' potentials for the optimization of lipid productivity

Component	Evaluation	Reference	Variation
$CaCl_2$	Versatile effector in plant cells; reported to be essential for induction of lipid synthesis	[50–52]	Yes
$FeSO_4$	Influence on growth and lipid metabolism reported	[51, 53–55]	Yes
H_2SO_4	Sulfur supply ensured by sulfate anions from diverse other medium components; nevertheless varied as provided together with $FeSO_4$ in one stock solution		Yes[b]
K_2HPO_4/KH_2PO_4	Essential phosphorus source (nucleic acid synthesis)	[54]	Yes
KOH	Potassium excess by phosphate salts; nevertheless varied as provided together with Na_2EDTA in one stock solution		Yes[c]
MES	Trade-off between osmotic inhibition and buffer capacity; alkaline pH may inhibit cell cycle	[56]	Yes
$MgSO_4$	Influence on growth and lipid production reported; effector of acetyl-CoA carboxylase, an essential enzyme during lipid biosynthesis; central atom of chlorophyll	[51, 57, 58]	Yes
NaCl	Reported to increase lipid production; excess may cause metabolic burden (ATP dependent sodium exporters) and thus inhibit growth	[59, 60]	Yes
Na_2EDTA	Commonly used metal chelator; excess may cause growth repression due to ion depletion	[51, 55]	Yes
$NaNO_3$	Essential nitrogen source (protein synthesis)	[54]	Yes
Penicillin-G	Support of long-time sterile conditions; not metabolized (data not shown)		No
Trace elements ($CoSO_4$, $CuSO_4$, H_3BO_3, $MnCl_2$, Na_2MoO_4, $ZnSO_4$)	Numerous studies about wastewater detoxification available, but only limited information concerning metabolism and lipid production; general pattern: little amounts essential, but high level cytotoxic (e.g., inhibition of photosynthesis); thus clustered to one input variable	[61–66]	Yes[a]

[a] All trace elements were clustered to one single input variable

[b] Varied together with $FeSO_4$ as provided in one single stock solution

[c] Varied together with Na_2EDTA as provided in one single stock solution

false negatives, i.e., excluding relevant components from the remaining study. The diagram shows that an increase in the concentration of $NaNO_3$ has a significant negative effect on lipid productivity. On the other hand, an increase in the trace element's concentration results in a significant ($p < 0.1$, t test) productivity improvement. Furthermore, the analysis indicates a positive tendency with the increasing $CaCl_2$ concentration and the lowering EDTA concentration. However, because of measurement noise and the comparably low number of experiments, the uncertainty of the estimation is relatively high and leads to no reliable statement about the effects of $CaCl_2$ and EDTA, respectively. Similarly, this holds true for $MgSO_4$, but here, the pairwise interaction with another component was additionally investigated. As shown in Additional file 3, a significant negative combinatorial effect was identified with the sodium salts, $NaNO_3$ and NaCl.

For visual inspection of the negative combinatorial effect, a Kriging model was constructed based on the given data. The predicted functional relationship between $MgSO_4$, $NaNO_3$, and the lipid productivity is displayed in Fig. 1b. In case of low $NaNO_3$ concentration, the interpolation reveals a positive correlation between an increase in $MgSO_4$ and that of the performance indicator. With the increasing $NaNO_3$ concentration, this positive effect is weakened.

In conclusion, significant effects of $NaNO_3$ and the trace elements were identified, as well as positive tendencies of $MgSO_4$ and $CaCl_2$. Furthermore, the effect of $MgSO_4$ appears to depend on the sodium salts, $NaNO_3$ and NaCl. The remaining components have only low potential to affect the lipid productivity and were thus excluded from further analysis.

Full factorial

In order to verify the observed tendencies and to investigate potential pairwise or higher combinatorial effects, a full factorial design was constructed for the remaining five input variables: $NaNO_3$, $MgSO_4$, $CaCl_2$, NaCl, and the clustered trace elements. This design again comprises five reference points and 32 experiments with minimal/maximal concentration (see Additional file 4 for the individual designs and the corresponding measurement data).

Figure 2a shows the updated statistical results after performing the full factorial design. The previously observed effects of $NaNO_3$ and the trace elements were confirmed. The positive tendency of $CaCl_2$ turned out to be significant, while the effect of NaCl remained insignificant. However, the interaction of $MgSO_4$ with the sodium salts could be investigated in more detail. Figure 2b shows the opposing effect of $MgSO_4$ dependent on $NaNO_3$. This

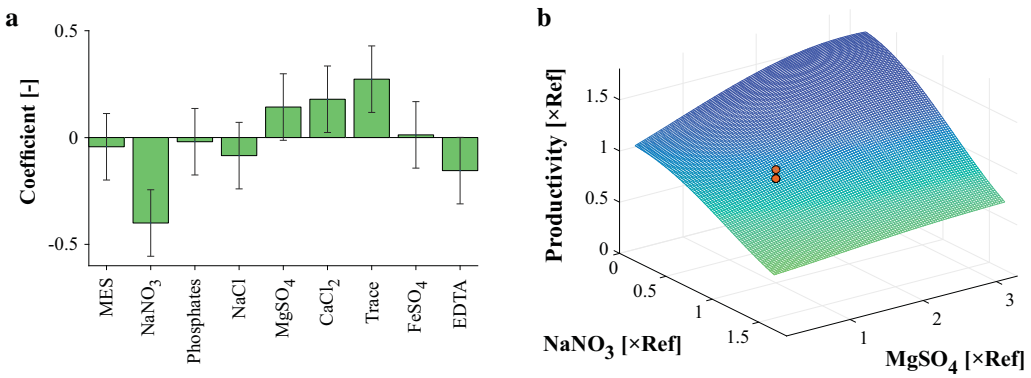


Fig. 1 **a** Estimated relative effects of single components on lipid productivity. **b** Kriging visualization of combinatorial effect of $MgSO_4$ and $NaNO_3$ on lipid productivity. Remaining medium components are set to their reference values. *Red dots* indicate measurements at the reference point

interaction leads to a non-distinguishable single component effect of $MgSO_4$, as indicated in Fig. 2a. The analysis also revealed a negative interaction between $CaCl_2$ and the trace elements, as indicated in the screening plot shown in Additional file 5.

Locating optimal medium composition

In "Full factorial" section, single and combinatorial effects of the media components were investigated on the basis of a full factorial design providing a rough estimate about optimal medium. The goal of the next step was to examine limitations of the particular effects and to identify potential optimal media compositions. To achieve this, the minimum and maximum concentrations were adjusted, and a more complex experimental design scheme was applied, comprising several nested factorial designs (see Additional file 6 for the full experimental design including the corresponding measurement data).

The maximum concentration of $NaNO_3$ was lowered from 1.7 × Ref to 1 × Ref. The upper bound of the concentration of the clustered trace elements was increased by 50% to 3.75 × Ref. The concentration of $CaCl_2$ could not be increased, as various types of precipitation effects were observed that distorted lipid analysis (see Additional file 7).

However, $MgSO_4$ was varied over three levels, as illustrated in Fig. 3a. For each level, the concentrations of $NaNO_3$, $CaCl_2$, and trace elements were distributed using a full factorial design. For the intermediate concentration of $MgSO_4$, the remaining components were varied only over half of their total ranges. A center point was located in each of these full factorial cubes. An additional nine points were space filling distributed over the edges of the cubes. In total, 39 experiments were performed and analyzed, including four reference replicates.

Figure 3b shows the Kriging interpolation based on all the data available after the third round of experiments. The figure shows three contour plots where the third component was fixed to the front, bottom–left corner of the inner cube as shown in Fig. 3a. The contour plots clearly show an interaction of the trace elements with $NaNO_3$ and $CaCl_2$, whereas $MgSO_4$ influences the lipid productivity only slightly positively. Moreover, an optimal region for the medium composition can be identified around $MgSO_4$ = 3.25 × Ref, $CaCl_2$ = 1.5 × Ref, Trace = 2 × Ref, and $NaNO_3$ = 0.3 × Ref (A three-dimensional plot of the Kriging model together with the measured data can be obtained from Additional file 8).

Refining the optimum

In the fourth and last round of the experiments, twelve experiments were placed around the optimum predicted by the Kriging interpolation. These experiments were planned by sampling the EI distribution, as described in "Expected Improvement" section, for maximizing the lipid productivity and minimizing the prediction uncertainty of the Kriging model. In addition, 23 experiments were uniformly distributed over the parameter space in a random manner, in order to improve prediction accuracy also in non-optimal regions. In total, 39 experiments were performed, including the four reference experiments (The full experimental design including the respective measurement data can be obtained from Additional file 9).

Figure 4 shows predictions of the updated Kriging model in the same fashion as described in "Locating optimal medium composition" section. The location of the optimum shifted toward $MgSO_4$ = 3.25 × Ref, $CaCl_2$ = 1.25 × Ref, trace elements = 2.5 × Ref, and $NaNO_3$ = 0.45 × Ref (A three-dimensional plot of the Kriging model together with the measured data can

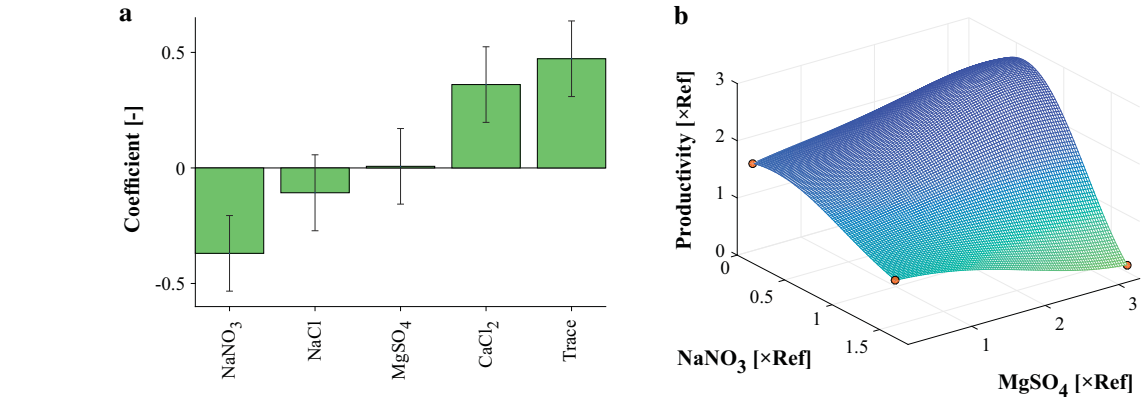


Fig. 2 **a** Estimated relative effects of single components on lipid productivity. **b** Kriging visualization of the combinatorial effect of $MgSO_4$ and $NaNO_3$ on lipid productivity. NaCl was set to lower value, and $CaCl_2$ and trace elements are set to upper values in the factorial design. *Red dots* indicate measurement data points

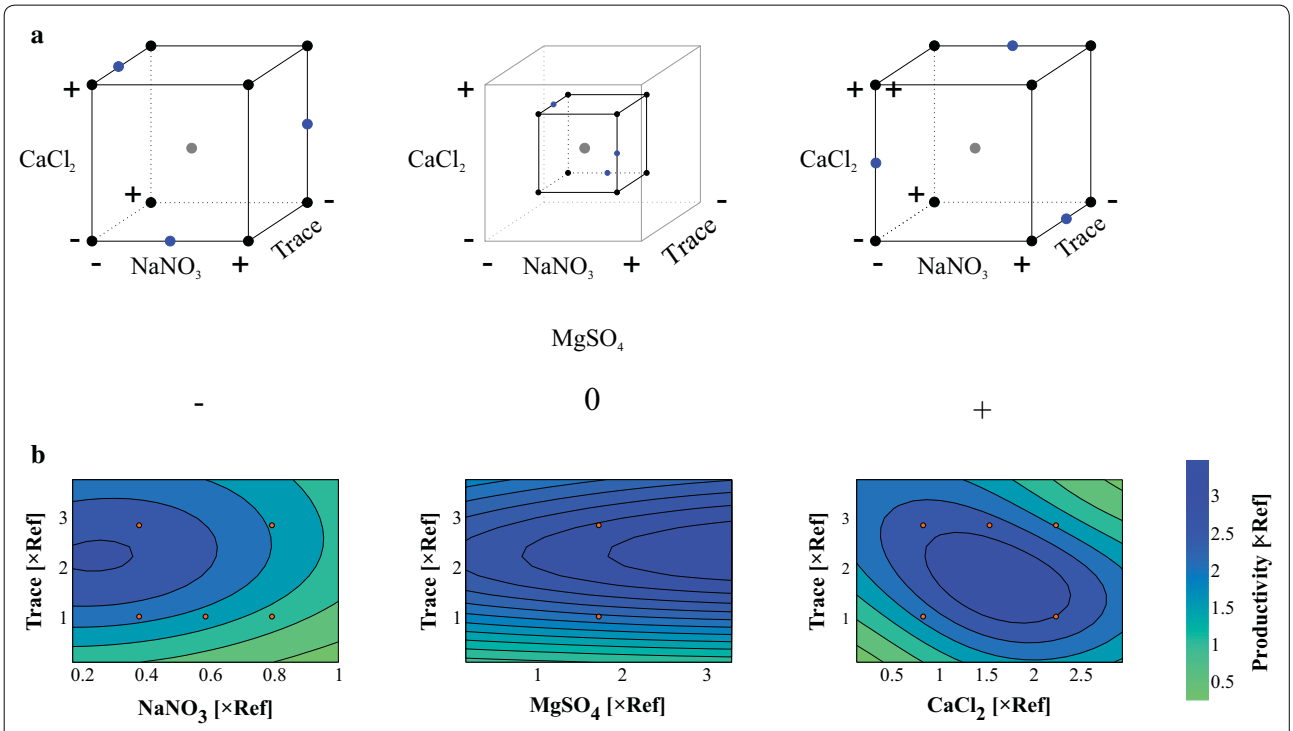


Fig. 3 **a** Illustration of experimental design in the third round of experiments. *Black dots* represent full factorial designs with respect to $NaNO_3$, $CaCl_2$, and trace elements. *Blue dots* indicate additionally added points. *The gray dots* represent the respective center points. *Minus* and *plus signs* symbolize minimum and maximum concentrations of the respective components. **b** Contour plots of the Kriging interpolation. The non-varied components were fixed at $MgSO_4 = 1.7 \times$ Ref, $CaCl_2 = 0.82 \times$ Ref, Trace $= 1 \times$ Ref, or $NaNO_3 = 0.37 \times$ Ref. *Red dots* indicate measurement data

be obtained from Additional file 10). For the optimal medium composition, the Kriging model predicts an increase by a factor of 3.03 ± 0.81 in lipid productivity compared with the reference medium.

Validating the optimal medium composition

In order to validate the determined optimal medium composition (see "Kriging-assisted optimization" section) and to highlight potential changes of process kinetics,

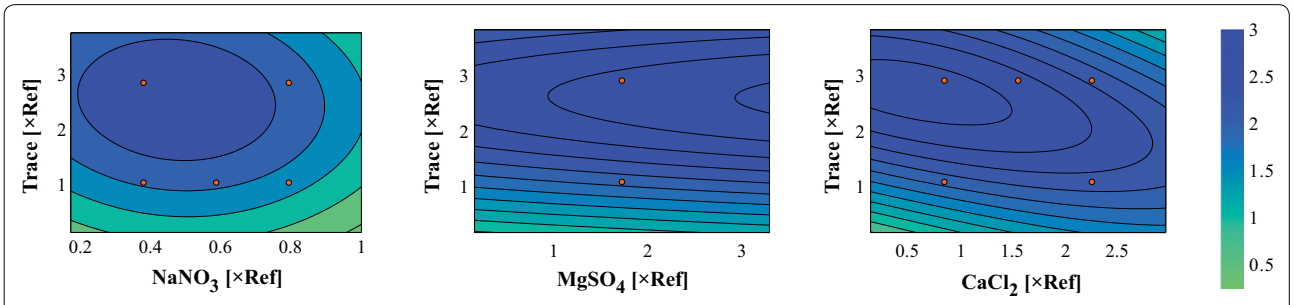


Fig. 4 Contour plots of the updated Kriging interpolation after the fourth round of experiments. The non-varied components were fixed at $MgSO_4 = 1.7 \times$ Ref, $CaCl_2 = 0.82 \times$ Ref, Trace $= 1 \times$ Ref, or $NaNO_3 = 0.37 \times$ Ref. *Red dots* indicate measurement data

cultivations using $enBBM_{ref}$ and $enBBM_{opt}$ were carried out (see Additional file 2 for medium composition). Both processes were monitored in-depth by sequential harvest of replicate wells from microtiter plate cultivations (see Fig. 5). To maximize comparability with the literature reports, biomass concentration at harvest was acquired as cell dry weight rather than biovolume in this context.

Medium optimization resulted in a series of significant changes in process performance as summarized in Table 2. While the exponential growth rates in both media did not differ significantly ($p < 0.05$, t test), times to nitrogen depletion were 84 h and 52 h for $enBBM_{ref}$ and $enBBM_{opt}$, respectively. This was due to the reduction of nitrate concentration during medium optimization down to 0.45 × Ref. In the reference process, exponential growth shifted to linear kinetics reaching an optical density of 4.94 ± 0.06 typically indicating the onset of light limitation and in clear accordance with prior experiments [27]. This effect was not observed for the optimized medium before nitrogen depletion. Neutral lipid accumulations started within 36 h ($enBBM_{ref}$) and 20 h ($enBBM_{opt}$) after nitrogen limitation which corresponds to a reduction of approx. 45%. Moreover, the biomass-specific lipid accumulation rate (estimated by linear fit) increased by approx. 32% from 4.87 ± 0.53% (w w^{-1}) d^{-1} to 6.43 ± 0.17% (w w^{-1}) d^{-1} due to medium optimization. Most probably, both effects are attributable to the increased availability of magnesium and calcium ions, as well as trace elements in the medium. This might result in a boost of the enzymatic turnover of lipid synthesis, especially regarding acetyl-CoA carboxylase (see Table 1).

Alternatively, a kinetic limitation of ion import into the cells at the low concentrations in the reference medium could be an explanation. Regarding downstream processing, the increased magnesium concentration offers another positive aspect, as it was previously reported to assist flocculation of the cells at high pH [68]. This mechanism is currently being investigated as an alternative to the comparably costlier biomass separation by centrifugation.

Most strikingly, cell dry weight at harvest did not differ significantly ($p < 0.05$, t test) for both media, despite the nitrate concentration being reduced to 45% in $enBBM_{opt}$. This indicates the nitrate-specific biomass yield as being a function of the initial nitrate availability, a phenomenon that has recently been recognized and discussed for a fairly comparable *Chlorella* process [27]. Together with an increase in the neutral lipid content from 10.55 ± 0.35% (w w^{-1}) to 23.9 ± 1.2% (w w^{-1}), this translated into a 2.3-fold increase of volumetric productivity up to 169 ± 7 mg L^{-1} d^{-1}.

Besides the evaluation of productivity-related issues, the relative composition of the fatty acids from the neutral lipid product fraction was compared by gas chromatography time-of-flight mass spectrometry (see Fig. 6).

The obtained fingerprints were in clear agreement with the previous literature reports [69] as palmitic, oleic, linoleic, and α-linolenic acids made up the major product fractions of 85% ($enBBM_{ref}$) and 89% ($enBBM_{opt}$). There are indications that the lipid fingerprint largely depends on cultivation conditions such as temperature [70], illumination [71], etc. However, our results demonstrate that changes in the medium composition can also lead to differences in the fatty acid fingerprint. The fractions of palmitoleic (16:1 Δ^9), hexadecadienoic (16:2 $\Delta^{7,10}$), hexadecatrienoic (16:3 $\Delta^{7,10,13}$), stearic (18:0), and linoleic (18:2 $\Delta^{9,12}$) remained nearly unchanged. On the contrary, the proportions of palmitic (16:0) and α-linolenic (18:3 $\Delta^{9,12,15}$) acids shrank by 22 and 42%, respectively, while linoleic (18:1 Δ^9) acid increased by 92% to a total share of 48 ± 1.8% using $enBBM_{opt}$. With respect to biodiesel synthesis, this reduction in the polyunsaturated fatty acids' fraction is clearly advantageous, increasing the fuel's oxidative stability [72].

Final medium simplification

In "Kriging-assisted optimization" section, several input variables were identified to be 'non-relevant' and thus kept at the respective reference values throughout the whole study. However, for MES and especially for

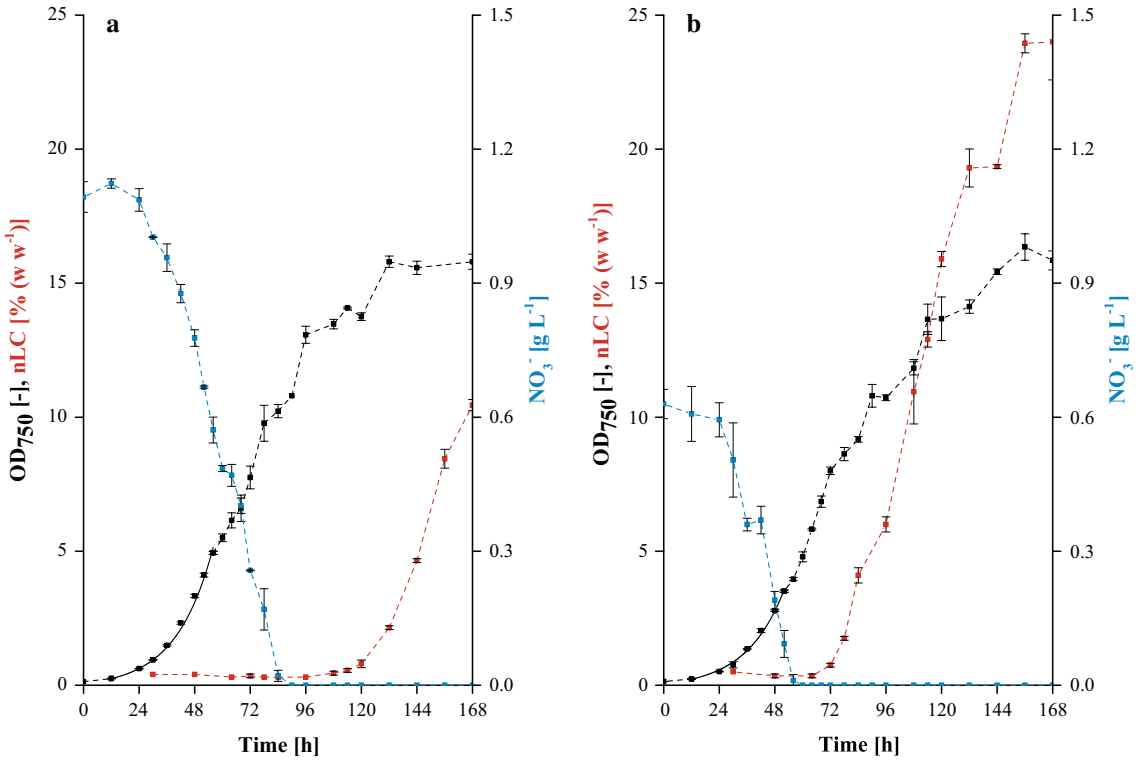


Fig. 5 Comparison of the processes using reference and optimized medium. **a** enBBM$_{ref}$, **b** enBBM$_{opt}$; 25 °C, 2.5% (v v^{-1}) CO_2, 200 µmol m^{-2} s^{-1} PAR, ≥85% relative humidity. *Error bars* represent min/max from biological replicates ($n = 2$)

Table 2 Comparison of process performance indicators using reference and optimized media. Error bars represent min/max from biological replicates ($n = 2$)

Parameter	enBBM$_{ref}$	enBBM$_{opt}$
Exponential growth rate (d^{-1})	1.49 ± 0.06	1.45 ± 0.1
Time to nitrate depletion (h)	84	52
Delay from nitrate depletion to onset of lipid synthesis [h]	36	20
Biomass-specific lipid accumulation rate [% (w w^{-1}) d^{-1}]	4.87 ± 0.53	6.43 ± 0.17
Cell dry weight at harvest (g L^{-1})	4.95 ± 0.06	4.93 ± 0.01
Neutral lipid content at harvest [% (w w^{-1})]	10.55 ± 0.35	23.9 ± 1.2
Volumetric productivity (mg L^{-1} d^{-1})	74 ± 1	169 ± 7

EDTA, a negative, but still non-significant ($p < 0.05$, t test) trend was observed. Besides economic aspects, culture media should only contain the necessary ingredients in appropriate concentrations to ensure high nutrient usage efficiency. Thus, an additional variant, in the following denoted as enBBM$_{opt,min}$, was investigated. Here, the concentrations of all 'non-relevant' components were set to the respective minimum values during the screening analysis. In particular, this included the complete omissions of MES buffer and the chelator EDTA, as well as NaCl (see Additional file 2 for medium composition), while phosphate availability was reduced to 0.125 × Ref.

In comparison with the results using enBBM$_{ref}$, as well as enBBM$_{opt}$, these adaptations did not change the overall obtained cell dry weight significantly ($p < 0.05$, t test) but led to an increase of the neutral lipid to 30.1 ± 1.6% (w w^{-1}), while the respective lipid fingerprint remained unchanged in comparison with enBBM$_{opt}$ (see Additional file 11). The resulting volumetric productivity of 212 ± 11 mg L^{-1} d^{-1} represents a total 2.9-fold improvement compared with the reference. Leaving out EDTA and especially the MES buffer drastically reduces the medium costs, so that the price per liter is lowered by 96%. Most probably, MES is not required as the phosphate salts offer sufficient pH stabilization capacity. Although EDTA is commonly used as a metal chelator to improve long-term stability of algae cultivation media, the results clearly indicate that its usage is not beneficial for this specific application. Moreover, the reduction of phosphate concentration to 12.5% is advantageous for large-scale application where the recovery of excess

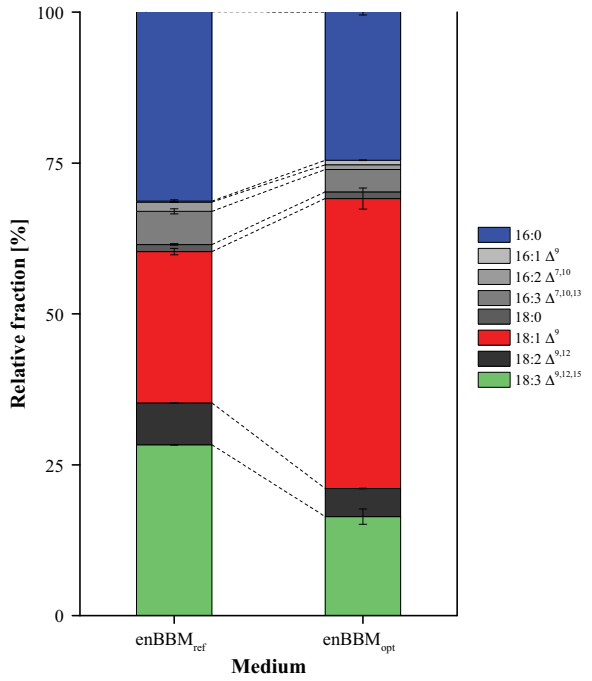


Fig. 6 Relative composition of the fatty acids from the neutral lipid product fractions for enBBM$_{ref}$ and enBBM$_{opt}$. Error bars represent min/max from biological replicates ($n = 2$)

nutrients to prevent overfertilization by wastewater is an important economic aspect. Yet, these results clearly confirm the validity of the initial screening analysis.

Assessment of achieved volumetric productivity

In the last decade, numerous studies addressed the lipid production of diverse *C. vulgaris* strains in different laboratory-scale batch processes [6, 17, 58, 70, 71, 73–81]. Among these, the average volumetric productivity was approximately 51 ± 36 mg L^{-1} d^{-1} and thus was comparable to the achieved value of 74 ± 1 mg L^{-1} d^{-1} using the enBBM reference medium. However, the reported values exhibit a wide spread, and it has to be assumed that these differences do not only originate from the different strains used, but from process conditions and reactor design as well. Some studies report productivities in the range of 130 mg L^{-1} d^{-1} when cultivating *C. vulgaris* in laboratory-scale batch processes with optimized nitrogen availability [17, 80]. Unfortunately, it is not generally clarified if productivities refer to the neutral lipid or the total lipid content. In this study, the volumetric productivity of neutral lipids of up to 212 ± 11 mg L^{-1} d^{-1} clearly exceeds previous reports and thereby underlines the importance of medium optimization not only for nitrate as commonly done, but especially for the concentrations of further salts and trace ions.

Conclusions

In this study, a blueprint strategy for the accelerated development of phototrophic bioprocesses is presented. This strategy is very efficient in terms of time and material, by incorporating state-of-the-art phototrophic cultivation and analytics with higher throughput that is closely linked to sophisticated experimental design strategies.

Taking neutral lipid production by the unicellular microalga *C. vulgaris* as a model process, the cultivation medium was optimized toward volumetric productivity. Fractional and full factorial designs in combination with Kriging-based approaches for data analysis, visualization, and experimental design allowed for an efficient and effective optimization in terms of time and cost. The optimized process has an approximately threefold increased lipid productivity of 212 ± 11 mg L^{-1} d^{-1}, which was achieved with only four experimental rounds with one microtiter plate each.

Besides the commonly addressed concentration of the nitrogen source (here nitrate), especially magnesium, calcium, and various trace elements were shown to be of crucial importance. Analysis tools furthermore revealed multi-parameter interactions that could have been overlooked otherwise. Over and above this, the concentration of non-relevant medium components was successfully minimized, contributing to reducing medium cost. Taking all the above results together, a smart combination of microscale phototrophic cultivation with sophisticated design of experiments led to a tremendous improvement of neutral lipid production with *C. vulgaris*, at the same time reducing cost for media components by 96%, while all other process performance indicators were kept constant.

Additional files

Additional file 1. Fractional factorial design for the initial screening analysis and corresponding measurement data of the individual cultivations.

Additional file 2. Medium composition of enBBM$_{ref}$, enBBM$_{opt}$ and enBBM$_{opt,min}$.

Additional file 3. Estimated effect of two factor interactions with $MgSO_4$. Estimations are based on the experiments in section "Fractional factorial" using the fractional factorial design and the corresponding measurement data given in Additional material 1.

Additional file 4. Experimental design and corresponding measurement data evaluated in section "Full factorial".

Additional file 5. Screening plot around reference point (enBBM$_{ref}$ medium). Estimation of the functional relationship between media components and lipid productivity was done by Kriging. The Kriging model is based on the experiments in section "Kriging-assisted optimization" using the open source software KriKit.

Additional file 6. Experimental design and corresponding measurement data evaluated in section "Locating optimal medium composition".

Additional file 7. Analysis of calcium precipitation by means of optical density measurements. Shaded area represents the parameter space covered during cultivation experiments. Error bars deviated from technical replicates (n = 3).

Additional file 8. Measured data and Kriging interpolation after the third round of experiments in three-dimensional representation. The non-varied components were fixed at $MgSO_4$=1.7xRef, $CaCl_2$=0.82xRef, Trace=1xRef, or $NaNO_3$=0.37xRef. Red dots indicate measurement data.

Additional file 9. Experimental design and corresponding measurement data evaluated in section "Refining the optimum".

Additional file 10. Measurement data and updated Kriging interpolation after the fourth round of experiments in three-dimensional representation. The non-varied components were fixed at $MgSO_4$=1.7xRef, $CaCl_2$=0.82xRef, Trace=1xRef, or $NaNO_3$=0.37xRef. Red dots indicate measurement data.

Additional file 11. Relative composition of the fatty acids from the neutral lipid product fraction for $enBBM_{opt}$ and $enBBM_{opt,min}$. Error bars represent min/max from biological replicates (n = 2).

Abbreviations

DoE: Design of Experiments; EI: Expected Improvement; enBBM: enriched Bold's Basal Medium; KriKit: Kriging toolKit; MCMC: Markov Chain Monte Carlo; nLC [% (w w^{-1})]: neutral lipid content; OD_{750} [—]: optical density at 750 nm; P_{vol} [mg L^{-1} d^{-1}]: volumetric productivity; Ref: reference value; μ [d^{-1}]: exponential growth rate.

Authors' contributions

HM and LF designed the study, evaluated the results, and prepared the manuscript. HM carried out the experimental work with assistance from JR, while LF performed the computational analysis of the acquired data. WW helped to finalize the manuscript and EvL and MO helped to finalize the manuscript and supervised the computational and experimental work, respectively. All authors read and approved the final manuscript.

Author details

[1] Forschungszentrum Jülich GmbH, Institute of Bio- and Geosciences, IBG-1: Biotechnology, Wilhelm-Johnen-Straße, 52428 Jülich, Germany. [2] Institute of Biotechnology, RWTH Aachen University, Aachen, Germany.

Acknowledgements

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Availability of supporting data

The datasets generated and analyzed during the current study are available from the corresponding author upon reasonable request.

Funding

The authors thank the Federal Ministry for Economic Affairs and Energy on the basis of a grant by the German Bundestag for the support and funding (grant no. KF2519304CS3) for Holger Morschett, as well as the Ministry of Innovation, Science and Research of North Rhine-Westphalia and the Heinrich Heine University Düsseldorf for the scholarship awarded to Lars Freier within the purview of CLIB-Graduate Cluster Industrial Biotechnology.

Received: 13 October 2016 Accepted: 13 January 2017
Published online: 31 January 2017

References

1. Mata TM, Martins AA, Caetano NS. Microalgae for biodiesel production and other applications: a review. Renew Sust Energ Rev. 2010;14:217–32.
2. Weyer KM, Bush DR, Darzins A, Willson BD. Theoretical maximum algal oil production. Bioenerg Res. 2010;3:204–13.
3. Gouveia L, Oliveira A. Microalgae as a raw material for biofuels production. J Ind Microbiol Biotechnol. 2009;36:269–74.
4. Gui MM, Lee KT, Bhatia S. Feasibility of edible oil vs. non-edible oil vs. waste edible oil as biodiesel feedstock. Energy. 2008;33:1646–53.
5. Yang J, Xu M, Zhang X, Hu Q, Sommerfeld M, Chen Y. Life-cycle analysis on biodiesel production from microalgae: water footprint and nutrients balance. Bioresour Technol. 2011;102:159–65.
6. Rodolfi L, Chini Zittelli G, Bassi N, Padovani G, Biondi N, Bonini G, et al. Microalgae for oil: strain selection, induction of lipid synthesis and outdoor mass cultivation in a low-cost photobioreactor. Biotechnol Bioeng. 2009;102:100–12.
7. Chisti Y. Biodiesel from microalgae. Biotechnol Adv. 2007;25:294–306.
8. Wijffels RH, Barbosa MJ. An outlook on microalgal biofuels. Science. 2010;329:796–9.
9. Wijffels RH, Barbosa MJ, Eppink MHM. Microalgae for the production of bulk chemicals and biofuels. Biofuel Bioprod Biorefin. 2010;4:287–95.
10. Norsker N-H, Barbosa MJ, Vermuë MH, Wijffels RH. Microalgal production—a close look at the economics. Biotechnol Adv. 2011;29:24–7.
11. Hariskos I, Posten C. Biorefinery of microalgae—opportunities and constraints for different production scenarios. Biotechnol J. 2014;9:739–52.
12. Brennan L, Owende P. Biofuels from microalgae—a review of technologies for production, processing, and extractions of biofuels and co-products. Renew Sust Energ Rev. 2010;14:557–77.
13. Adarme-Vega TC, Lim DKY, Timmins M, Vernen F, Li Y, Schenk PM. Microalgal biofactories: a promising approach towards sustainable omega-3 fatty acid production. Microb Cell Fact. 2012;11:96.
14. Van Wagenen J, Holdt SL, De Francisci D, Valverde-Pérez B, Plósz BG, Angelidaki I. Microplate-based method for high-throughput screening of microalgae growth potential. Bioresour Technol. 2014;169:566–72.
15. Radmann EM, Camerini FV, Santos TD, Costa JAV. Isolation and application of SO_X and NO_X resistant microalgae in biofixation of CO_2 from thermoelectricity plants. Energy Convers Manag. 2011;52:3132–6.
16. de Morais MG, Costa JAV. Isolation and selection of microalgae from coal fired thermoelectric power plant for biofixation of carbon dioxide. Energy Convers Manag. 2007;48:2169–73.
17. Breuer G, Lamers PP, Martens DE, Draaisma RB, Wijffels RH. The impact of nitrogen starvation on the dynamics of triacylglycerol accumulation in nine microalgae strains. Bioresour Technol. 2012;124:217–26.
18. Debska D, Potvin G, Lan C, Zhang Z. Effects of medium composition on the growth of *Chlorella vulgaris* during photobioreactor batch cultivations. J Biobased Mater Bio. 2010;4:68–72.
19. Radzun KA, Wolf J, Jakob G, Zhang E, Stephens E, Ross I, et al. Automated nutrient screening system enables high-throughput optimisation of microalgae production conditions. Biotechnol Biofuels. 2015;8:1–17.
20. Tillich UM, Wiolter N, Schulze K, Kramer D, Brödel O, Frohme M. High-throughput cultivation and screening platform for unicellular phototrophs. BMC Microbiol. 2014;14:239.
21. Chen M, Mertiri T, Holland T, Basu AS. Optical microplates for high-throughput screening of photosynthesis in lipid-producing algae. Lab Chip. 2012;12:3870–4.
22. Han W, Li C, Miao X, Yu G. A novel miniature culture system to screen CO_2-sequestering microalgae. Energies. 2012;5:4372–89.
23. Heo J, Cho D-H, Ramanan R, Oh H-M, Kim H-S. PhotoBiobox: a tablet sized, low-cost, high throughput photobioreactor for microalgal screening and culture optimization for growth, lipid content and CO_2 sequestration. Biochem Eng J. 2015;103:193–7.
24. Ojo EO, Auta H, Baganz F, Lye GJ. Design and parallelisation of a miniature photobioreactor platform for microalgal culture evaluation and optimisation. Biochem Eng J. 2015;103:93–102.
25. Kim HS, Weiss TL, Thapa HR, Devarenne TP, Han A. A microfluidic photobioreactor array demonstrating high-throughput screening for microalgal oil production. Lab Chip. 2014;14:1415–25.
26. Graham PJ, Riordon J, Sinton D. Microalgae on display: a microfluidic pixel-based irradiance assay for photosynthetic growth. Lab Chip. 2015;15:3116–24.
27. Morschett H, Schiprowski D, Müller C, Mertens K, Felden P, Meyer J, et al. Design and validation of a parallelized micro-photobioreactor enabling phototrophic bioprocess development at elevated throughput. Biotechnol Bioeng. 2017;114:122–31.

28. Morschett H, Reich S, Wiechert W, Oldiges M. Simplified cryopreservation of the microalga *Chlorella vulgaris* integrating a novel concept for cell viability estimation. Eng Life Sci. 2016;16:36–44.
29. Morschett H, Wiechert W, Oldiges M. Automation of a Nile red staining assay enables high throughput quantification of microalgal lipid production. Microb Cell Fact. 2016;15:34.
30. Pereira H, Barreira L, Mozes A, Florindo C, Polo C, Duarte C, et al. Microplate-based high throughput screening procedure for the isolation of lipid-rich marine microalgae. Biotechnol Biofuels. 2011;4:1–12.
31. Tillich UM, Wolter N, Franke P, Dühring U, Frohme M. Screening and genetic characterization of thermo-tolerant *Synechocystis* sp. PCC6803 strains created by adaptive evolution. BMC Biotechnol. 2014;14:1–15.
32. Mandenius C-F, Brundin A. Bioprocess optimization using design-of-experiments methodology. Biotechnol Progr. 2008;24:1191–203.
33. Fisher RA. The design of experiments. Edinborough: Oliver and Boyd; 1935.
34. Freier L, Hemmerich J, Schöler K, Wiechert W, Oldiges M, von Lieres E. Framework for Kriging-based iterative experimental analysis and design: optimization of secretory protein production in *Corynebacterium glutamicum*. Eng Life Sci. 2016. doi:10.1002/elsc.201500171.
35. Montgomery DC. Design and analysis of experiments. New York: Wiley; 2012.
36. Lee K-M, Gilmore DF. Statistical experimental design for bioprocess modeling and optimization analysis. Appl Biochem Biotechnol. 2006;135:101–15.
37. Islam RS, Tisi D, Levy MS, Lye GJ. Framework for the rapid optimization of soluble protein expression in *Escherichia coli* combining microscale experiments and statistical experimental design. Biotechnol Progr. 2007;23:785–93.
38. Beijerinck MW. Culturversuche mit Zoochlorellen und anderen niederen Algen. Btg Ztg. 1890;45:725–40.
39. Bold HC. The morphology of *Chlamydomonas chlamydogama*, sp. nov. Bull Torrey Bot Club. 1949;76:101–8.
40. Rohe P, Venkanna D, Kleine B, Freudl R, Oldiges M. An automated workflow for enhancing microbial bioprocess optimization on a novel microbioreactor platform. Microb Cell Fact. 2012;11:144.
41. Unthan S, Radek A, Wiechert W, Oldiges M, Noack S. Bioprocess automation on a Mini Pilot Plant enables fast quantitative microbial phenotyping. Microb Cell Fact. 2015;14:32.
42. Graham MD. The Coulter principle: foundation of an industry. JALA. 2003;8:72–81.
43. Paczia N, Nilgen A, Lehmann T, Gätgens J, Wiechert W, Noack S. Extensive exometabolome analysis reveals extended overflow metabolism in various microorganisms. Microb Cell Fact. 2012;11:122.
44. Myers RH, Montgomery DC, Anderson-Cook CM. Response surface methodology: process and product optimization using designed experiments. Hoboken: Wiley; 2016.
45. Cressie N. Statistics for spatial data. Hoboken: Wiley-Intersciences; 2015.
46. Jones DR, Schonlau M, Welch WJ. Efficient global optimization of expensive black-box functions. J Glob Optim. 1998;13:455–92.
47. Jones DR. A taxonomy of global optimization methods based on response surfaces. J Glob Optim. 2001;21:345–83.
48. Freier L, von Lieres E. Multi-objective global optimization (MOGO) Algorithm and case study in gradient elution chromatography. Biotechnol J. doi:10.1002/biot.201600613.
49. Haario H, Laine M, Mira A, Saksman EDRAM. Efficient adaptive MCMC. Stat Comput. 2006;16:339–54.
50. Chen H, Zhang Y, He C, Wang Q. Ca^{2+} signal transduction related to neutral lipid synthesis in an oil-producing green alga *Chlorella* sp. C2. Plant Cell Physiol. 2014;55:634–44.
51. Ren H-Y, Liu B-F, Kong F, Zhao L, Xie G-J, Ren N-Q. Enhanced lipid accumulation of green microalga *Scenedesmus* sp. by metal ions and EDTA addition. Bioresour Technol. 2014;169:763–7.
52. Jones RGW, Lunt OR. The function of calcium in plants. Bot Rev. 1967;33:407–26.
53. Liu ZY, Wang GC, Zhou BC. Effect of iron on growth and lipid accumulation in *Chlorella vulgaris*. Bioresour Technol. 2008;99:4717–22.
54. Ruangsomboon S, Ganmanee M, Choochote S. Effects of different nitrogen, phosphorus, and iron concentrations and salinity on lipid production in newly isolated strain of the tropical green microalga *Scenedesmus dimorphus* KMITL. J Appl Phycol. 2013;25:867–74.
55. Concas A, Steriti A, Pisu M, Cao G. Comprehensive modeling and investigation of the effect of iron on the growth rate and lipid accumulation of *Chlorella vulgaris* cultured in batch photobioreactors. Bioresour Technol. 2014;153:340–50.
56. Guckert JB, Cooksey KE. Triglyceride accumulation and fatty acid profile changes in *Chlorella* (*Chlorophyta*) during high pH-induced cell cycle inhibition. J Phycol. 1990;26:72–9.
57. Deng X, Fei X, Li Y. The effects of nutritional restriction on neutral lipid accumulation in *Chlamydomonas* and *Chlorella*. Afr J Microbiol Res. 2011;5:260–70.
58. Lv JM, Cheng LH, Xu XH, Zhang L, Chen HL. Enhanced lipid production of *Chlorella vulgaris* by adjustment of cultivation conditions. Bioresour Technol. 2010;101:6797–804.
59. Alyabyev AJ, Loseva NL, Gordon LK, Andreyeva IN, Rachimova GG, Tribunskih VI, et al. The effect of changes in salinity on the energy yielding processes of *Chlorella vulgaris* and *Dunaliella maritima* cells. Thermochim Acta. 2007;458:65–70.
60. Duan X, Ren GY, Liu LL, Zhu WX. Salt-induced osmotic stress for lipid overproduction in batch culture of *Chlorella vulgaris*. Afr J Biotechnol. 2012;11:7072–8.
61. Mallick N. Biotechnological potential of immobilized algae for wastewater N, P and metal removal: a review. Biometals. 2002;15:377–90.
62. Pittman JK, Dean AP, Osundeko O. The potential of sustainable algal biofuel production using wastewater resources. Bioresour Technol. 2011;102:17–25.
63. Muñoz R, Guieysse B. Algal–bacterial processes for the treatment of hazardous contaminants: a review. Wat Res. 2006;40:2799–815.
64. Chen M, Tang H, Ma H, Holland TC, Ng KYS, Salley SO. Effect of nutrients on growth and lipid accumulation in the green algae *Dunaliella tertiolecta*. Bioresour Technol. 2011;102:1649–55.
65. Yang J, Cao J, Xing G, Yuan H. Lipid production combined with biosorption and bioaccumulation of cadmium, copper, manganese and zinc by oleaginous microalgae *Chlorella minutissima* UTEX2341. Bioresour Technol. 2015;175:537–44.
66. Clijsters H, Van Assche F. Inhibition of photosynthesis by heavy metals. Photosynth Res. 1985;7:31–40.
67. Ohlrogge J, Browse J. Lipid biosynthesis. Plant Cell. 1995;7:957–70.
68. Vandamme D, Foubert I, Fraeye I, Meesschaert B, Muylaert K. Flocculation of *Chlorella vulgaris* induced by high pH: role of magnesium and calcium and practical implications. Bioresour Technol. 2012;105:114–9.
69. Petkov G, Garcia G. Which are fatty acids of the green alga *Chlorella*? Biochem Syst Ecol. 2007;35:281–5.
70. Converti A, Casazza AA, Ortiz EY, Perego P, Del Borghi M. Effect of temperature and nitrogen concentration on the growth and lipid content of *Nannochloropsis oculata* and *Chlorella vulgaris* for biodiesel production. Chem Eng Process. 2009;48:1146–51.
71. Atta M, Idris A, Bukhari A, Wahidin S. Intensity of blue LED light: a potential stimulus for biomass and lipid content in fresh water microalgae *Chlorella vulgaris*. Bioresour Technol. 2013;148:373–8.
72. Wadumesthrige K, Salley SO, Ng KYS. Effects of partial hydrogenation, epoxidation, and hydroxylation on the fuel properties of fatty acid methyl esters. Fuel Process Technol. 2009;90:1292–9.
73. Widjaja A, Chien C-C, Ju Y-H. Study of increasing lipid production from fresh water microalgae *Chlorella vulgaris*. J Taiwan Inst Chem E. 2009;40:13–20.
74. Yeh K-L, Chang J-S. Effects of cultivation conditions and media composition on cell growth and lipid productivity of indigenous microalga *Chlorella vulgaris* ESP-31. Bioresour Technol. 2012;105:120–7.
75. Mallick N, Mandal S, Singh AK, Bishai M, Dash A. Green microalga *Chlorella vulgaris* as a potential feedstock for biodiesel. J Chem Technol Biotechnol. 2012;87:137–45.
76. Liang Y, Sarkany N, Cui Y. Biomass and lipid productivities of *Chlorella vulgaris* under autotrophic, heterotrophic and mixotrophic growth conditions. Biotechnol Lett. 2009;31:1043–9.
77. Hsieh C-H, Wu W-T. Cultivation of microalgae for oil production with a cultivation strategy of urea limitation. Bioresour Technol. 2009;100:3921–6.
78. Griffiths M, Hille R, Harrison SL. The effect of nitrogen limitation on lipid productivity and cell composition in *Chlorella vulgaris*. Appl Microbiol Biotechnol. 2014;98:2345–56.

79. Gorain PC, Bagchi SK, Mallick N. Effects of calcium, magnesium and sodium chloride in enhancing lipid accumulation in two green microalgae. Environ Technol. 2013;34:1887–94.
80. Yeh K-L, Chang J-S. Nitrogen starvation strategies and photobioreactor design for enhancing lipid content and lipid production of a newly isolated microalga *Chlorella vulgaris* ESP-31: implications for biofuels. Biotechnol J. 2011;6:1358–66.
81. Yoo C, Jun SY, Lee JY, Ahn CY, Oh HM. Selection of microalgae for lipid production under high levels carbon dioxide. Bioresour Technol. 2010;101:71–4.

2.4 Multi-objective global optimization (MOGO): Algorithm and case study in gradient elution chromatography

Biotechnology Journal
www.biotechnology-journal.com

Biotechnol. J. 2017, 12, 1600613

DOI 10.1002/biot.201600613

Research Article

Multi-objective global optimization (MOGO): Algorithm and case study in gradient elution chromatography

Lars Freier and Eric von Lieres

IBG-1: Biotechnology, Forschungszentrum Jülich, Jülich, Germany

Biotechnological separation processes are routinely designed and optimized using parallel high-throughput experiments and/or serial experiments. Well-characterized processes can further be optimized using mechanistic models. In all these cases – serial/parallel experiments and modeling – iterative strategies are customarily applied for planning novel experiments/simulations based on the previously acquired knowledge. Process optimization is typically complicated by conflicting design targets, such as productivity and yield. We address these issues by introducing a novel algorithm that combines recently developed approaches for utilizing statistical regression models in multi-objective optimization. The proposed algorithm is demonstrated by simultaneous optimization of elution gradient and pooling strategy for chromatographic separation of a three-component system with respect to purity, yield, and processing time. Gaussian Process Regression Models (GPM) are used for estimating functional relationships between design variables (gradient, pooling) and performance indicators (purity, yield, time). The Pareto front is iteratively approximated by planning new experiments such as to maximize the Expected Hypervolume Improvement (EHVI) as determined from the GPM by Markov Chain Monte Carlo (MCMC) sampling. A comprehensive Monte-Carlo study with in-silico data illustrates efficiency, effectiveness and robustness of the presented Multi-Objective Global Optimization (MOGO) algorithm in determining best compromises between conflicting objectives with comparably very low experimental effort.

Received	06 OCT 2016
Revised	28 NOV 2016
Accepted	21 DEC 2016
Accepted article online	23 DEC 2016

Supporting information available online

Keywords: Gaussian process regression · Gradient elution chromatography · Multi-objective optimization

1 Introduction

Multiple performance indicators, also referred to as objectives, are often required in the assessment of industrial processes. In many cases, these are conflicting each other, i.e. one objective can only be improved at the cost of compromising with another. For such situations, the concept of Pareto optimization has been developed [1], i.e. instead of finding one unique optimum for all objectives, one aims at finding the full set of best compromises between them. Pareto optimization is commonly applied in (bio-)chemical engineering [2].

For instance, Zhang et al. [3] have used a genetic algorithm (GA) for the multi-objective optimization of continuous countercurrent chromatography units. In fact, different GA variants are very popular for solving Pareto optimization problems [4]. However, these algorithms generally require many experiments. The experimental effort can potentially be reduced by using surrogate models. Such models utilize the set of currently known data for estimating functional relationships between process parameters and objectives. Santana-Quintero et al. [5] provide a good review of techniques for applying surrogate modeling in Pareto optimization.

Emmerich et al. [6] propose the application of Gaussian Process Regression Modeling (GPM) in multi-objective optimization. The authors compare four different options for using GPM in this context: mean value, prob-

Correspondence: Dr. Eric von Lieres, IBG-1: Biotechnology, Forschungszentrum Jülich, Wilhelm-Johnen-Straße 1, 52425 Jülich, Germany
E-mail: e.von.lieres@fz-juelich.de

Abbreviations: EHVI, expected hypervolume improvement; **GPM,** Gaussian process regression model; **MOGO,** multi-objective global optimization

ability of improvement, lower confidence bound, and expected improvement prescreening. In single-objective optimization, it is well-known that the last option results in the most robust and effective optimization strategy [7]. However, only very recently an efficient algorithm has been published that allows fast calculation of the multi-objective equivalent to expected improvement [8].

Finally, Binois et al. [9] have developed a single-valued measure of model uncertainty using random set theory, in order to assess the quality of a given GPM. This measure allows observing convergence of the optimization process and consequently helps to avoid unnecessary experiments.

This publication presents an algorithm and software that combine key scientific advances made in GPM assisted multi-objective optimization in the recent last years. Starting with an initial set of experiments, a GPM is constructed and iteratively updated during the optimization process. In each iteration, new experiments are designed by maximizing expected improvement. The optimization algorithm is continued until model uncertainty drops under a defined threshold.

The presented Multi-Objective Global Optimization (MOGO) algorithm is a direct but conceptually and computationally non-trivial extension of the single-objective equivalent, Efficient Global Optimization (EGO) developed by Jones et al. [10]. Important properties of the MOGO algorithm, such as efficiency, effectiveness, and robustness are investigated by optimizing operating conditions (gradient shape, pooling strategy) in elution chromatography of a three-component system (lysozyme, cytochrome, ribonuclease) with respect to three conflicting objectives (purity, yield, process time).

2 Theory

2.1 Gaussian process regression

The proposed optimization algorithm is based on the prediction of a Gaussian Process Regression Model (GPM). Rasmussen [11] gives a good overview of the derivation and applications of GPM. In geostatistics, GPM is also referred to as Kriging [12]. We have implemented the GPM concept in MATLAB as part of the Kriging ToolKit (KriKit), which is open source and freely available at https://github.com/modsim/KriKit.

In GPM, it is assumed that the output $Z(\boldsymbol{x})$ follows a basic trend $m(\boldsymbol{x})$ and that random fluctuation in the data is caused by a stochastic process $Y(\boldsymbol{x})$ with zero mean, Eq. (2), and standard deviation $\sigma(\boldsymbol{x})$, Eq. (3).

$$Z(\boldsymbol{x}) = m(\boldsymbol{x}) + Y(\boldsymbol{x}) \tag{1}$$

$$\mathrm{E}[Y(\boldsymbol{x})] = 0 \tag{2}$$

$$\sigma(\boldsymbol{x}) = \sqrt{\mathrm{Var}\big(Y(\boldsymbol{x})\big)} \tag{3}$$

Even though more complex trend functions can be applied, it is often sufficient to assume a constant trend, Eq. (4).

$$m(\boldsymbol{x}) = c \tag{4}$$

Predictions Z^* of unknown observations at points $\hat{\boldsymbol{x}}$ are described by a linear combination of the given data $Z(\boldsymbol{x}_i)$ with coefficients $\lambda_i \in \mathbb{R}$, Eq. (5).

$$Z^*(\hat{\boldsymbol{x}}) = \sum_{i=1}^{n} \lambda_i(\hat{\boldsymbol{x}}) Z(\boldsymbol{x}_i) \tag{5}$$

The coefficients λ_i are determined such that the prediction is unbiased, Eq. (6), and has minimal variance, Eq. (7).

$$\mathrm{E}[Z(\hat{\boldsymbol{x}}) - Z^*(\hat{\boldsymbol{x}})] = 0 \tag{6}$$

$$\mathrm{Var}\big[Z(\hat{\boldsymbol{x}}) - Z^*(\hat{\boldsymbol{x}})\big] \to \min \tag{7}$$

The corresponding values of the coefficients $\boldsymbol{\lambda}$ are determined using statistical information on the covariance Cov ($Z(\boldsymbol{x}_i)$, $Z(\boldsymbol{x}_j)$) between observations at points $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$. This covariance is approximated using a covariogram model, as described in more detail in the following section. Using the covariogram model, the coefficients $\boldsymbol{\lambda}$ can be calculated using Eq. (8).

$$\begin{bmatrix} \boldsymbol{\lambda} \\ \mu \end{bmatrix} = \begin{bmatrix} C & \mathbf{1} \\ \mathbf{1}^T & 0 \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{c} \\ 1 \end{bmatrix} \tag{8}$$

In Eq. (8), $C \in \mathbb{R}^{n \times n}$ is a matrix with entries $C_{i,j} = \mathrm{Cov}(Z(\boldsymbol{x}_i), Z(\boldsymbol{x}_j))$, $\mathbf{1} \in \mathbb{R}^n$ a vector with entries 1, and $\boldsymbol{c} \in \mathbb{R}^n$ a vector with entries $c_i = \mathrm{Cov}(Z(\boldsymbol{x}_i), Z(\hat{\boldsymbol{x}}))$. The vector $\boldsymbol{\lambda} \in \mathbb{R}^n$ contains the n coefficients λ_i, and μ is a Lagrange multiplier. The Lagrange multiplier is required for solving Eq. (7) constrained by Eq. (6) but not used for calculating the prediction in Eq. (5).

GPM inherently provides an estimation of the confidence interval. The prediction error can be estimated by Eq. (9) [12].

$$\begin{aligned} \mathrm{Var}\big[Z(\hat{\boldsymbol{x}}) - Z^*(\hat{\boldsymbol{x}})\big] &= \mathrm{Cov}\big(Z(\hat{\boldsymbol{x}}), Z(\hat{\boldsymbol{x}})\big) \\ &\quad - 2\sum_{i=1}^{n} \lambda_i \mathrm{Cov}\big(Z(\boldsymbol{x}_i), Z(\hat{\boldsymbol{x}})\big) \\ &\quad + \sum_{i=i}^{n}\sum_{j=1}^{n} \lambda_i \lambda_j \mathrm{Cov}\big(Z(\boldsymbol{x}_i), Z(\boldsymbol{x}_j)\big) \end{aligned} \tag{9}$$

The variance usually increases with the distance between the point of interest $\hat{\boldsymbol{x}}$ and points $\boldsymbol{x}_i$ of the given measure-

ment data. It is further increased by measurement noise and correlations between measurements.

2.2 Covariogram model

The coefficients $\boldsymbol{\lambda}$, as computed by Eq. (8), depend on the mutual covariances between data points $C_{i,j} = \mathrm{Cov}\,(Z(\boldsymbol{x}_i), Z(\boldsymbol{x}_j))$ and on the covariances between data points and the point of interest $\hat{\boldsymbol{x}}$, $c_i = \mathrm{Cov}(Z(\boldsymbol{x}_i), Z(\hat{\boldsymbol{x}}))$. These covariances are generally unknown and need to be approximated using a covariogram model. The covariogram model, $C(\boldsymbol{h})$, depends only on the distance $\boldsymbol{h} = \boldsymbol{x}_i - \boldsymbol{x}_j$ of any two points $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ and not on their absolute positions.

Mainly three types of covariogram models can be found in the literature [13]: the spherical, the Matérn, and the exponential model. The Matérn function is known to be particularly suitable for representing various covariance-distance relationships [14] and is hence used in this study, with a smoothing parameter of 3/2, Eq. (10).

$$\mathrm{C}(\boldsymbol{h}) = \theta^2_{\mathrm{Nugget}} + \theta^2_{\sigma}\left(1+\sqrt{3}r\right)\exp\left(-\sqrt{3}r\right),\quad \text{with } r = \sqrt{\sum_{l=1}^{k}\frac{(h_l)^2}{\theta_l^2}} \tag{10}$$

The covariogram converges toward θ_σ for $\boldsymbol{h} \to \boldsymbol{0}$ and toward zero for $\boldsymbol{h} \to \infty$, and θ_l determines the characteristic length-scale for input variable $\boldsymbol{x}_l$. The parameter θ_{Nugget}, which is historically referred to as nugget factor, introduces an extra offset at $\boldsymbol{h} = \boldsymbol{0}$, which provides more flexibility in modeling the measurement error.

In this study, the covariogram parameters are estimated using the Maximum Likelihood Estimation (MLE) approach [15]. In this context, the output variables are considered to follow a multivariate Gaussian distribution with probability distribution function defined by Eq. (11).

$$p(\boldsymbol{z}\,|\,\boldsymbol{\theta}) = \frac{1}{(2\pi)^{n/2}\det(\mathrm{C})^{1/2}}\exp\left(-\frac{1}{2}(Z-\boldsymbol{m})^T\mathrm{C}^{-1}(Z-\boldsymbol{m})\right) \tag{11}$$

Here, the vector $\boldsymbol{m} \in \mathbb{R}^n$ contains the trend function evaluations and Z the measured values at the sample points. The entries in the covariance matrix $\mathrm{C} \in \mathbb{R}^{n\times n}$ are calculated by the covariance function using parameters $\boldsymbol{\theta}$. Using Eq. (11), the resulting log-likelihood function for a given set of covariogram parameters $\boldsymbol{\theta}$ is defined by Eq. (12).

$$\log p(\boldsymbol{\theta}\,|\,Z(\boldsymbol{x})) = -\frac{1}{2}(Z-\boldsymbol{m})^T\mathrm{C}^{-1}(Z-\boldsymbol{m}) - \frac{1}{2}\log\left(|\mathrm{C}|\right) - \frac{n}{2}\log\left(2\pi\right) \tag{12}$$

Optimal covariogram parameter values are determined by maximizing Eq. (12). In this study, the trend function is a constant $\boldsymbol{m}$, whose value is estimated by Eq. (13) [12].

$$m = \left(\mathbf{1}^T\mathrm{C}^{-1}\mathbf{1}\right)^{-1}\mathbf{1}^T\mathrm{C}^{-1}Z \tag{13}$$

2.3 Pareto optimality

In multi-objective optimization, each input vector $\boldsymbol{x} \in \mathbb{R}^m$ is associated with more than one objective, summarized in the vector $Z \in \mathbb{R}^d$ with $d > 1$ [1]. The aim of the optimization is to maximize all objectives, Eq. (14).

$$\max_{\boldsymbol{x}}\left(Z_1(\boldsymbol{x}),\ldots,Z_d(\boldsymbol{x})\right) \tag{14}$$

In case of conflicting objectives, no input vector maximizes all objectives at the same time. However, Pareto optimal solutions $Z\,(\boldsymbol{x}_{\mathrm{opt}})$ can be identified, i.e. points at which no objective can be improved without deteriorating at least one other objective. A set of Pareto optimal solutions is referred to as Pareto front, Eq. (15).

$$P = \left\{Z\left(\boldsymbol{x}_{\mathrm{opt}}\right)\right\} \tag{15}$$

Conceptually, there is a "true" Pareto front P_{true} with all optimal solutions. Multi-objective optimization aims at approximating this front, mostly using iterative procedures. Such algorithms typically start with an initial data set, of which the subset of Pareto optimal outputs is determined, here denoted by P_1. In contrast to the "true" Pareto front P_{true} which contains infinitely many points, the front P_1 contains finitely many points which are Pareto optimal only with respect to the other points in the initial database. In following iterations, this database is systematically extended by planning and performing further experiments following specific experimental design strategies that will be discussed later in more detail. The results of new measurements are added to the existing database, and the Pareto front is correspondingly updated. In each iteration i, the current Pareto front P_i weakly dominates all previous Pareto fronts P_j with $j < i$, Eq. (16).

$$P_i \geq P_j \;\forall i \geq j \tag{16}$$

New points that exceed the previous Pareto front, P_{i-1}, are added to the current set P_i. They complement the previous optimal points and potentially (but not necessarily) replace some of them. If a suitable experimental design strategy is used, the current Pareto front will eventually converge towards the "true" Pareto front, Eq. (17).

$$\lim_{i\to\infty} P_i = P_{true} \tag{17}$$

In order to monitor convergence of the optimization procedure, the progress between Pareto fronts P_j and P_i needs to be quantified. In a comprehensive study, Zitzler et al. [16] have assessed several quality indicators. Zitzler and Thiele [17] have developed the hypervolume indicator H, which is commonly applied. The indicator H is defined as the area, volume or hypervolume (in two, three and

higher dimensions, respectively) between the positive coordinate axes and the Pareto front. Obviously, H increases monotonically during the iterative optimization procedure, Eq. (18), as the successively updated Pareto curves dominate each other.

$$H(P_i) \geq H\left(P_j\right) \; \forall i \geq j \tag{18}$$

The indicator reaches its maximum as the Pareto front P_i converges towards the "true" Pareto front [18].

$$\lim_{i \to \infty} H\left(P_i\right) = H\left(P_{true}\right) \tag{19}$$

There are several algorithms for efficiently calculating the hypervolume. In this study, the dimension-sweep algorithm is used [19].

2.4 Expected Hypervolume Improvement (EHVI)

As visualized in Fig. 1A, an important part of the optimization algorithm is to new experiments. This requires to be able to quantify of the potential improvement of the current database when additional measurements are performed. In single-objective optimization, the improvement achieved through an experiment at point $\boldsymbol{x}$ can be quantified by comparing the associated output value $Z(\boldsymbol{x}) \in \mathbb{R}$ to the best value in the current database, $Z_{\max}$, Eq. (20).

$$\mathrm{I}\left(Z(\boldsymbol{x})\right) = \begin{cases} 0 & \text{if } Z(\boldsymbol{x}) < Z_{\max} \\ Z_{\max} - Z(\boldsymbol{x}) & \text{otherwise} \end{cases} \tag{20}$$

The output values of future experiments are generally unknown. However, the expected improvement $EI(\boldsymbol{x})$ can be estimated by taking the average of all possible output values, weighted by their respective probabilities, Eq. (21).

$$EI(\boldsymbol{x}) = \int_{Z \in \mathbb{R}} I(Z) \cdot PDF_{\boldsymbol{x}}(Z) dZ \tag{21}$$

In Eq. (21), $\mathrm{PDF}_{\boldsymbol{x}}(Z)$ is the probability density function of the GPM. The PDF is by definition normalized to $\int_{Z \in \mathbb{R}} PDF_{\boldsymbol{x}}(Z) dZ = 1$ and, hence, the integral in Eq. (21) can be interpreted as a weighted average. In the single objective case, this integral can be solved analytically [10]. Sasena et al. [7] have shown that the GPM based expected improvement is particularly effective for single-objective optimization, as not only model prediction but also the associated model uncertainty are considered.

Emmerich et al. [6] have extended the concept of expected improvement to multi-objective optimization. They propose using the hypervolume indicator H for defining a scalar measure of improvement when the output vector $Z(\boldsymbol{x}) \in \mathbb{R}^d$ of an experiment at point $\boldsymbol{x}$ is added to the current database, Eq. (22).

$$I\left(Z(\boldsymbol{x})\right) = H\left(P \cup Z(\boldsymbol{x})\right) - H(P) \tag{22}$$

By definition, $I\left(Z(\boldsymbol{x})\right)$ is zero if $Z(\boldsymbol{x})$ is dominated by the current approximation of the Pareto front P and a positive scalar otherwise. In analogy to the single objective case, the Expected Hypervolume Improvement (EHVI) of a point $\boldsymbol{x}$ is calculated by integrating the improvement, Eq. (22), weighted by the probability $PDF_{\boldsymbol{x}}(Z)$ that the respective output value Z is actually measured at point $\boldsymbol{x}$, over the solution space, Eq. (23).

$$EHVI(\boldsymbol{x}) = \int_{Z \in \mathbb{R}^d} I(Z) \cdot PDF_{\boldsymbol{x}}(Z) dZ \tag{23}$$

The probability density function $PDF_{\boldsymbol{x}}(Z)$ is a multidimensional Gaussian distribution from the GPM (Fig. 1B). The calculation of EHVI is difficult, even if statistical independency is assumed. Emmerich et al. [6] propose to approximate the integral by a Monte Carlo approach. However, many samples may be required to reach sufficient accuracy, which can be computationally demanding. In later work, Emmerich et al. [20] provide a more efficient algorithm for the exact calculation of EHVI in the two-dimensional case. This algorithm was most recently extended to higher dimensions by Hupkens et al. [8], who also provide a fast algorithm for the three dimensional case. An implementation of this algorithm can be downloaded at http://moda.liacs.nl/index.php. (We have tuned the original implementation for numerical performance without changing the mathematical algorithm.)

2.5 Conditional simulation

Conditional simulation (CS) is often used in spatial statistics for generating realizations of a given GPM. CS can be applied to assess extremes, to estimate uncertainties, and for multi-point statistics. The books of Goovaerts [21] and Chiles and Delfiner [22] provide several algorithms for CS, such as the sequential simulation algorithm, the p-field approach, the simulated annealing approach, and the residual algorithm. In this study, the residual algorithm is used. Detailed mathematical derivations can be found in [22] or [23]. The main idea is to calculate the CS in two steps, first generating samples from the prior distribution and then conditioning these samples using measurement data from the current database.

For a given Gaussian process, as introduced in Section 2.1, the set of non-conditioned samples $Z^{NC}(X_q)$ at points $X_q = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_q\}$ is defined by Eq. (24).

$$Z^{NC}(X_q) \sim N\left(Z^*\left(X_q\right), C\left(X_q\right)\right) \tag{24}$$

In Eq. (24), $Z^*(X_q)$ denotes the GPM prediction, and $C(X_q)$ the covariance matrix, as introduced in Section 2.2. Hansen (2016) [24] has shown that $Z^{NC}(X_q)$ can be calcu-

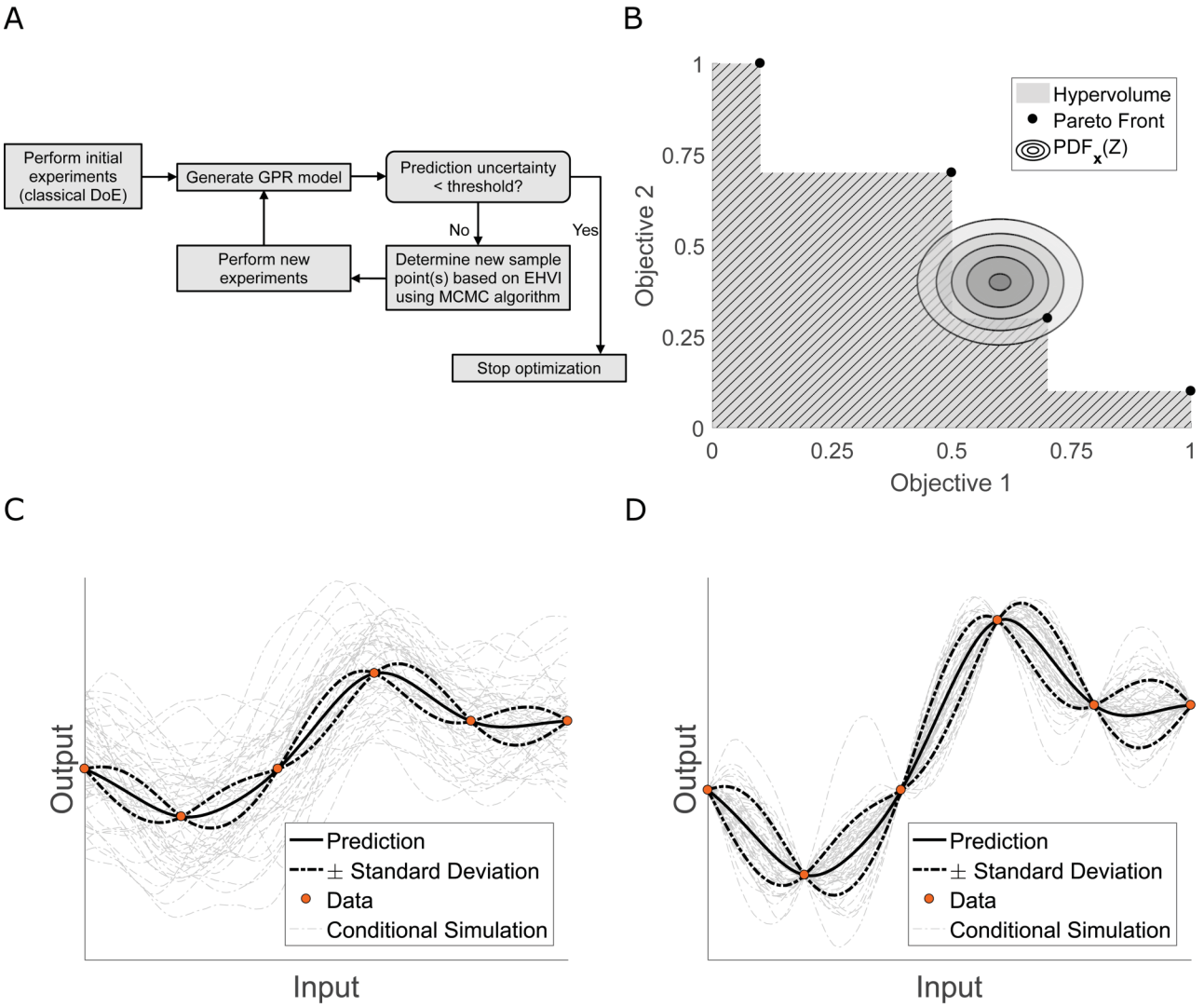


Figure 1. (**A**) MOGO algorithm flowchart, (**B**) Schematic of EHVI calculation. Contour lines indicate probability density function of the GPM predicted output Z at point $\boldsymbol{x}$, (**C**) Schematic of an example GPM with six samples and fifty CS (non-conditioned simulation), (**D**) Schematic of an example GPM with six samples and fifty CS (conditioned simulation)

lated using a standard normal random number generator, Eq. (25) with $Z^* = Z^*(X_q)$ and $C = C(X_q)$.

$$\begin{aligned} N\left(Z^*,C\right) &\sim Z^* + N(0,C) \\ &\sim Z^* + C^{1/2}N(0,I) \\ &\sim Z^* + BD^{1/2}B^TN(0,I) \\ &\sim Z^* + BD^{1/2}N(0,I) \end{aligned} \tag{25}$$

A schematic example for non-conditioned simulations is presented in Fig. 1C. In a second step, the conditioned samples $Z^C(X)$ are calculated using the data $Z(X_n)$, Eq. (26).

$$\left(Z^C\left(X_q\right)|Z^{NC}\left(X_q\right)\right) = Z^*(X_q) + Z^{NC}\left(X_q\right) - \boldsymbol{\lambda}^T Z^{NC}\left(X_n\right) \tag{26}$$

In Eq. (26), $\boldsymbol{\lambda}$ are the Kriging coefficients, Eq. (8), and $Z^{NC}(X_n)$ the non-conditioned values at points X_n. Hence, the measurement points, X_n, must be included in the CS sample points, X_q. Chevalier et al. [23] have proven that $Z^C(X_q)$ and the underlying GPM have equal distributions, and that different CSs are stochastically independent from each other. Figure 1D depicts a schematic example for conditioned simulations.

2.6 Uncertainty estimation

Asymptotic convergence is important, but practical applications also require an appropriate stopping criterion. In the case of single-objective optimization, criteria are applied that are based on the GPM prediction and error estimate [25] [10]. Chevalier et al. [26] have shown

that the GPM prediction and the associated model uncertainty can also be estimated using random set theory. Binois et al. [9] have extended this approach to the estimation of Pareto fronts and their variability using several GPMs for the individual objectives.

The distribution of the current Pareto front is estimated based on s different CSs of the underlying Gaussian processes, as introduced in Section 2.5. In contrast to the GPM prediction, each CS preserves not only the expected value but also the variance of the underlying Gaussian process. Pareto fronts $P_1^C,\ldots,P_S^C$ are determined for each of the CS. Then, a so-called coverage function $p(Z)$ is defined by counting the frequency of how often a point in the solution space is dominated by a CSs. Finally, the quantile surface Z_β is defined as the set of points with a coverage function value equal to β, Eq. (27).

$$Z^\beta : \left\{Z_1^\beta,\ldots,Z_j^\beta,\ldots,Z_k^\beta\right\} \text{ with } p\left(Z_j^\beta\right)=\beta \tag{27}$$

The expected Pareto front P_{exp} is estimated by adjusting the quantile β such that the hypervolume H of Z^β is equal to the mean hypervolume of the Pareto fronts $P_1^C,\ldots,P_n^C$ of all CSs, Eq. (28).

$$P_{\text{exp}} : Z^\beta, \text{ with } H\left(Z^\beta\right)=\frac{1}{\text{n}}\sum_{i=1}^{s}H(P_\text{i}) \tag{28}$$

A scalar stopping criterion G can be formulated that quantifies the global uncertainty of the Pareto front estimation. The global uncertainty G is estimated by the normalized symmetric difference between the CSs P_i and the expected Pareto front P_{exp}, Eq. (29).

$$G=\frac{1}{n}\sum_{i=1}^{n}\frac{\left(H\left(P_{\text{exp}}\cup P_i\right)-\frac{1}{2}H\left(P_{\text{exp}}\right)-\frac{1}{2}H\left(P_\text{i}\right)\right)}{H\left(P_{\text{exp}}\cup P_\text{i}\right)} \tag{29}$$

The optimization is stopped once G drops under a defined threshold. In Section 4.4, suitable values for G are discussed based on the calculation of 500 conditional simulation.

3 Case study

3.1 Example system

The proposed MOGO algorithm is demonstrated by optimizing ion-exchange chromatography of lysozyme (*lys*), cytochrome c (*cyt*), and ribonuclease a (*rna*) on the cation-exchanger SP Sepharose FF. The optimization aims to separate *cyt* from *lys* and *rna* by simultaneous variation of the operating conditions (gradient shape, pooling times). The quality of the resulting separation process is characterized by three conflicting objectives, namely purity, yield, and process time. Experiments are performed in silico using the Chromatography Analysis and Design Toolkit (CADET) [27, 28]. CADET is based on the general rate model of column liquid chromatography with steric mass action binding model [29]. The applied model parameters can be found in the Supporting information S1.

Table 1. Process parameter ranges

Parameter	Lower Bound	Upper Bound
p_1	10 mol/m³	300 mol/m³
p_2	10 mol/m³	1500 mol/m³
p_3	10 s	6990 s
Δt_1, Δt_2	5 s	1000 s

Figure 2A visualizes the investigated chromatography process. The system comprises four components (salt, lysozyme, cytochrome, and ribonuclease) with load, wash, and elution phases. For the combined load and wash phases (step I, 0–990 s), a salt concentration of 10 mol/m³ is applied. In the load phase, the three protein components are supplied at a concentration of 1 mol/m³. The elution phase starts with an immediate increase in salt concentration and has two parts with linear gradients but different slopes (step II and step III, 990–8000 s). The initial increase in salt concentration is specified by the process parameter. Both slopes are specified by parameters p_2, and the time point of transition between step II and step III by parameter. Step IV (8000–9000 s) has a fixed salt concentration that forces complete elution of all components at the end of the process. The pooling times t_1 and t_2, between which the effluent with the desired target component is collected, are parameterized by $t_{\text{peak}}-\Delta t_1$ and $\Delta t_2+t_{\text{peak}}$ where t_{peak} is the time point where the concentration of *cyt* reaches its maximum.

The process parameters are bound by ranges listed in Table 1. For numerical reasons, only nonzero values are used for p_3, Δt_1, and Δt_2. The salt concentration of the load and wash phases is used as lower bound for p_1 and p_2. p_3 is restricted by the start and end times of the elution phase. The salt concentration during step IV is used as upper bound for p_2.

3.2 Three conflicting performance indicators

The present case study aims at separating *cyt* from *lys* and *rna*. Purity and yield of *cyt* and the overall process time are used as performance indicators. All three objectives are conflicting each other, which leads to a Pareto optimization problem as defined in Section 2.3. Purity and yield are functions of the sampled amounts N_j of components $j \in \{lys, cyt, rna\}$ in the central effluent pool that is collected between times t_1 and t_2. The amounts N_j are determined by integrating the respective concentration peak c_j between the pooling times, Eq. (30).

$$N_j=\int_{t_1}^{t_2}c_j(t)\,dt \tag{30}$$

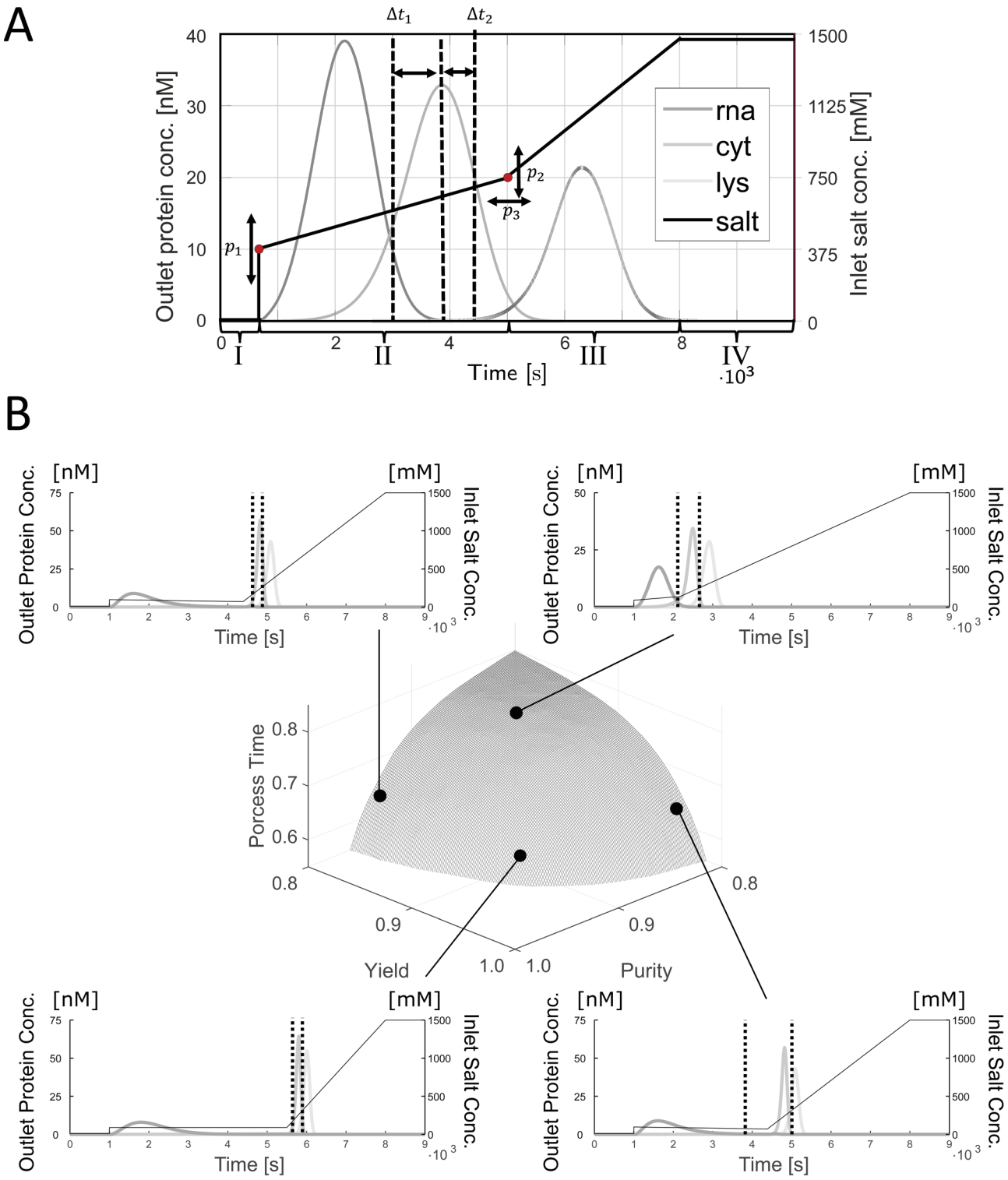


Figure 2. (**A**) Schematic of the gradient elution chromatography process with five design variables (gradient shape, pooling times) that are to be optimized. (**B**) Visualization of the Pareto Front as determined by brute force. The process time is scaled to the interval [0,1] and reversed. Values 0 and 1 indicate worst and best performance, respectively.

Purity is defined as the amount of the target component *cyt* relative to the total amount of all components in the collected pool, Eq. (31).

$$\text{Purity} = N_{\text{cyt}} \Big/ \sum_{j\in\{\text{lys,cyt,ma}\}} N_j \tag{31}$$

Yield is defined as the amount of the target component *cyt* in the collected pool relative to the amount of *cyt* loaded to the column, Eq. (32).

$$\text{Yield} = N_{\text{cyt}}/N_{\text{cyt,load}} \tag{32}$$

In this study, the loaded amount is $N_{\text{j,load}}$ is 10 mol for all components $j \in \{lys, cyt, ma\}$.

The separation is defined to end at time point t_{final} when 99.9% of $N_{\text{j,load}}$ of all components are eluted from the column.

Lu and Anderson-Cook [30] suggest to scale the performance indicators to the interval [0,1] for graphical representation of the Pareto front, where zero indicates the worst and one the best possible value. Purity and yield, as defined by Eq. (31) and Eq. (32), already fulfill this criterion. The process time t_{final} is correspondingly scaled with respect to the maximal simulation time $t_{\text{max}} = 9000$ s, Eq. (32).

$$\text{Process Time} = (1 - t_{\text{final}}/t_{\text{max}}) \tag{33}$$

Consequently, the Pareto front has a theoretical maximum hypervolume of $H_{\text{max}} = 1$.

3.3 Brute force analysis

For illustration, the "true" three-dimensional Pareto front P_{true} is first approximated in a brute force study using random sampling and manually chosen data points. In fact, approximately 11 million samples were calculated to make sure that the Pareto front approximation is reliable. The shape of the chromatogram and the process time only depend on the gradient parameters, p_1, p_2, p_3, while purity and yield additionally depend on the pooling times, t_1, t_2. Hence, the chromatography experiment is performed (in silico) once for each gradient shape, and shape, purity and yield are subsequently calculated for 100 different pairs of pooling times using Latin hypercube sampling for Δt_1 and Δt_1. The extremes are included by manually placing sample points at the upper and lower boundaries of Δt_1 and Δt_2.

Figure 2B shows the Pareto front determined by brute force in the region of purity and yield above 80%. The plot reveals that both purity and yield benefit from shallow elution gradients, which in turn requires long process times. For fixed process times, yield increases and purity decreases with increasing width of the collected product pool. The hypervolume of this brute force approximation to the "true" Pareto front is $H^*_{\text{max}} = 0.8123$, which is used as reference value in the following study.

4 Results and discussion

In this section, Multi-Objective Global Optimization (MOGO) is demonstrated using the previously introduced case study. MOGO is a direct extension of the well-known single-objective Efficient Global Optimization (EGO) algorithm [10] to the multi-objective case. The general MOGO procedure is illustrated in Fig. 1A, while the individual steps are detailed in the following sections. In particular, accelerating convergence is discussed in Section 4.3, and stopping criteria are investigated in Section 4.4. The impact of parallel experimentation on the required numbers of iterations and total experiments is studied in Section 4.5. MOGO represents a statistical optimization procedure and can be applied to systems with continues input and outputs.

4.1 Initial experimental design

GPM cannot be applied without database. Hence, functional relationships between process parameters and performance indicators are initially studied by classical design of experiments. After defining appropriate parameter ranges, a 2^3 full factorial design with center point is used for the gradient parameters p_1, p_2, and p_3. For each simulation, the effect of Δt_1 and Δt_2 is investigated by calculating purity and yield for ten pooling times that are placed by Latin hypercube sampling. The extremes are included by manually placing sample points at the upper and lower boundaries of Δt_1 and Δt_2. The resulting twelve samples per gradient shape were found sufficient for investigating the influence of Δt_1 and Δt_2 on purity and yield of the studied system.

Based on the resulting data set with 108 sample points in total, a GPM is created using the five process parameters ($p_{1,2,3}$ and $\Delta t_{1,2}$) as input variables and the three performance indicators (purity, yield, process time) as output variables.

4.2 Iterative optimization based on expected improvement

In further iterations of the MOGO algorithm, additional experiments are designed such as to maximize the EHVI, as defined in Section 2.4. In analogy to the single-objective equivalent, EHVI quantifies the potential improvement with respect to a scalar optimality criterion (the hypervolume under the Pareto front), which is based on the GPM prediction and its estimated error. The expected improvement represents a trade-off between refining promising areas with high predicted objectives and exploring areas with high uncertainty [7]. Jones et al. [10]

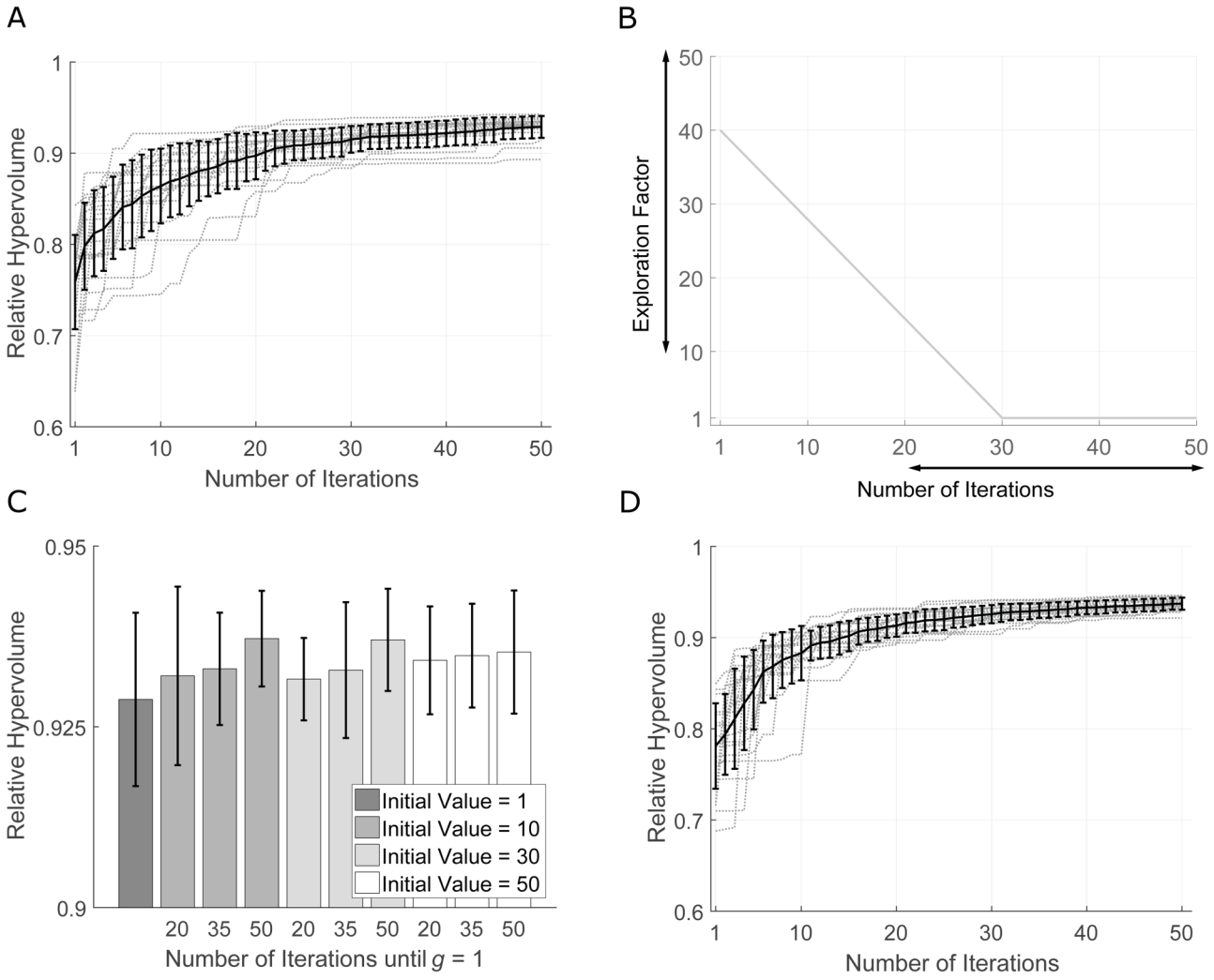


Figure 3. (**A**) Results of 20 complete optimization runs with 50 iterations each (grey lines). Black line and bars indicate mean value ± standard deviation. (**B**) Scheme for decreasing the exploration factor from iteration to iteration in order to speed up convergence (Section 4.3). (**C**) Comparison of different cooling strategies with varied initial values and slopes (iterations until g = 1 is reached). (**D**) Results of 20 complete optimization runs using the best cooling strategy (grey lines). Black line and bars indicate mean value ± standard deviation.

have demonstrated that optimization strategies which are based on expected improvement can handle highly nonlinear functional relationships.

As a key step in the MOGO algorithm, the additional experiments are designed according to their EHVI distribution, as determined by Markov Chain Monte Carlo (MCMC) sampling. I.e. the probability of new experiments to be chosen is directly proportional to their EHVI values. This approach allows to draw one or several samples in each iteration of the optimization procedure, elegantly enabling the design of both serial and parallel experiments with maximal information content. In this study, the Delayed Rejection Adaptive Metropolis (DRAM) algorithm which was developed and implemented by Haario et al. [31] is applied with a chain length of 10 000 and all other algorithm parameters at their default values. The pooling times are varied as in Section 4.1, using twelve samples for each gradient shape, as these function evaluations are much cheaper than calculating the chromatogram.

Purity, yield, and process time are calculated for each of the new samples, and the amended data set is used for updating the GPM. The experiments are performed in series, i.e. the GPM is updated for each gradient shape (serial experiments are discussed in Section 4.5). The iterative optimization procedure is continued until a defined stopping criterion is fulfilled (see Section 4.4 for details). For illustration, Fig. 3A compares the results of twenty MOGO optimizations with 50 iterations each, including the initial experimental design. The *y*-axis indicates the currently estimated hypervolume, H_{rel} relative to the reference value, H^{*}_{max}, from Section 3.3.

Variation of H_{rel} in the first iteration is caused by randomness in the Latin hypercube sampling, and in the other iterations additionally by stochasticity of the DRAM method. Technically, the uncertainty caused by Latin hypercube sampling can be reduced by refining the grid; however, the time for computing the EHVI increases exponentially with the number of samples. In the present study, twelve samples have been found to be a good compromise between computational effort and variability of the optimization process. Figure 3A demonstrates that the hypervolume increases with the number of iterations, indicating improvement of the corresponding estimated Pareto front. Options for speeding up convergence of the MOGO method are discussed in the next section. Optimizations resulted in 163.0 ± 14.5 Pareto optimal points. Pareto front is more densely sampled near to the extremes: high purity/low yield and vice versa. If a higher density is needed, the optimization can be continued for further iterations. Using a PC with four cores (Intel Core i5-4570), an optimization with 50 iterations took in average 8 h. The computation time is increasing with number of samples, mainly influenced by the calculation of the EHVI which scales with a complexity of $O(n^3)$.

4.3 Improving convergence speed and robustness

According to Ponweiser et al. [33], the GPM potentially underestimates the prediction error. Consequently, the optimizer can be caught by local optima until the prediction in their neighborhood is excessively accurate. This can dramatically increase the number of iterations required for finding the global optimum. Schonlau et al. [10, 34] have addressed this problem by introducing an exploration factor, g. This factor can be used to increase the weight of the prediction error in the GPM, and thus to improve exploration of the parameter space. However, appropriate values of the exploration factor are not trivial to determine. Sasena et al. [35] have suggested a cooling strategy, in analogy to the simulated annealing algorithm. This strategy starts with a high value of g which is then successively reduced in the course of the optimization procedure.

Interpreting the exploration factor g as an amplifier of the prediction error, Eq. (34), helps to define an analog in the multi-objective case. This concept can be directly transferred to the multi-objective case by adjusting $PDF_{\hat{\boldsymbol{x}}}(Z)$ that used for the EHVI calculation, see section 2.4.

$$Z(\hat{\boldsymbol{x}}) = m(\hat{\boldsymbol{x}}) + N\left(0, g\sigma(\hat{\boldsymbol{x}})\right) \tag{34}$$

The accelerating effect of g is studied for different cooling strategies. Generally, the exploration factor is linearly decreased with varying initial value and slope and a minimum value of $g = 1$ (Fig. 3B). The studied initial values are 10, 30, and 50, and the slope is specified by taking 20, 35, or 50 iterations until the exploration factor is decreased to one. Results of different cooling strategies are compared to the standard case, i.e. a constant exploration factor of $g = 1$ (Fig. 3C). Each cooling strategy is evaluated for 20 complete optimization runs with 50 iterations each.

Figure 3C reveals that the global optimizer performance is better when the exploration factor is decreased more slowly, in particular for smaller initial values. The cooling strategy with initial value $g = 10$ and a decrease over 50 iterations shows the best performance, i.e. lowest variation and highest hypervolume on average. Figure 3D shows the results of twenty optimization runs using this cooling strategy. The comparison to Fig. 3A clearly shows the benefit of a well-chosen cooling strategy for the exploration factor. The adjusted MOGO converges not only faster but is also more robust, i.e. has a reduced run-to-run variability.

4.4 Stopping criterion

Experimental effort can potentially be reduced by stopping the MOGO algorithm as soon as the result is sufficiently close to the optimum. In Section 2.6, the measure G was introduced for estimating global uncertainty of the multi-objective GPM prediction. In this section, a stop criterion based on estimated global uncertainty is studied using all 420 optimization runs from Section 4.3. The calculation of G is based on 500 conditional simulations. Here, the optimization progress is defined as the hypervolume reached after n iterations, relative to the maximal hypervolume reached after 50 iterations in the same optimization run. This approach also accounts for convergence towards local optima with reduced maximal hypervolume. The optimization progress stands in contrast to the relative hypervolume used in Fig. 4B. The relative hypervolume refers to the approximation of maximal achievable hypervolume H^*_{max} for an infinitive long optimization.

Figure 4A shows the correlation between the global model uncertainty G and the associated hypervolume value, based on 420 optimization runs and independent of the actual iteration number. Clearly, if lower threshold values for G are used, the optimization process continues longer and higher hypervolume values are achieved. However, the figure also illustrates that 98% of the reference value is already reached on average using an threshold level of $G = 0.043$. This result is independent of the cooling strategy (data not shown). Hence, the value $G = 0.04$ can be applied as a threshold which indicates that the optimization process can be terminated. This stopping criterion is tested in a study with 260 optimization runs, applying the best cooling strategy. On average, less than 26 iterations are required before the optimization is terminated, which is a significant reduction of experimental effort. According to the histogram in Fig. 4B, the stopping criterion is effective in determining when the

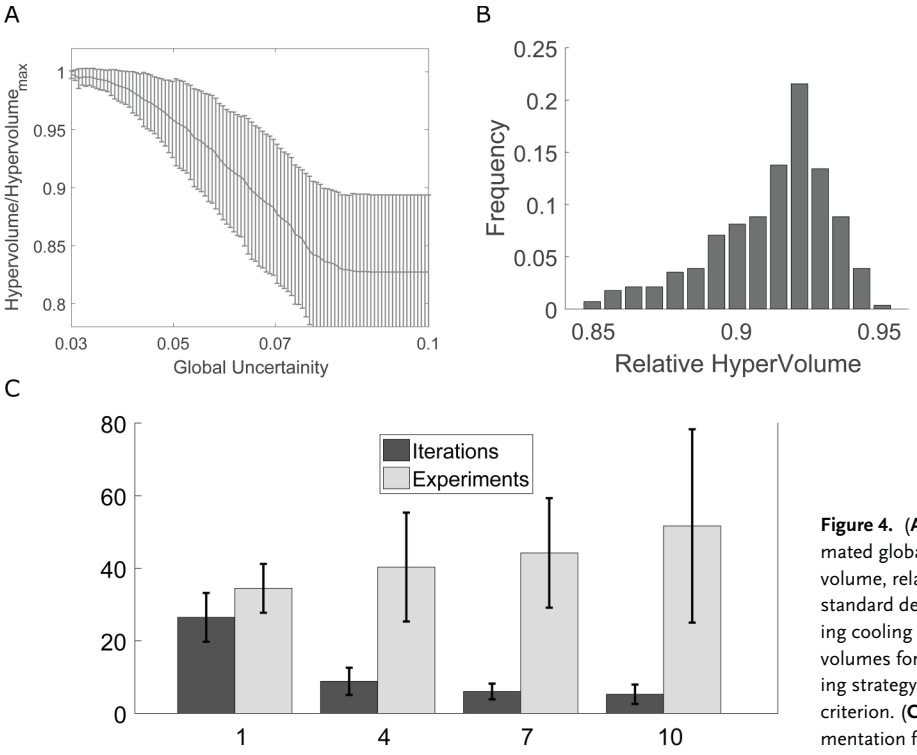


Figure 4. (**A**) Functional relationship between the estimated global uncertainty G and the reached hypervolume, relative to the reference value (mean value and standard deviation over 420 optimization runs with varying cooling strategies). (**B**) Histogram of reached hypervolumes for 260 optimization runs using the best cooling strategy and a threshold $G = 0.043$ for the stopping criterion. (**C**) Comparison of serial and parallel experimentation for 20 optimization runs each.

current hypervolume is sufficiently close to the optimum. In fact, 90% or more of the reference hypervolume is reached in ¾ of the optimization runs. Each optimization runs resulted in 120.5 ± 27.0 Pareto optimal points. The distribution of the number of Pareto optimal points is illustrated in the Supporting information S2.

4.5 Parallel experiments

In Section 4.2, the MOGO algorithm was explained to be equally suitable for designing serial or parallel experiments, using the DRAM method for MCMC sampling. In this section, the efficiency of parallelization is evaluated for 4, 7, and 10 parallel experiments, i.e. gradient shapes, in each iteration. As in the previous sections, each scenario is evaluated using 20 optimization runs with the best cooling strategy. The term experiment refers to a simulation with a unique gradient parameter set. Sample time variation after performing the simulations is not taken into account, as this procedure is computationally cheap.

Figure 4C shows that parallelization can effectively reduce the number of required iterations for reaching the stopping criterion. The first iteration contains the initial experimental design of nine experiments. The figure illustrates that the number of iterations on average decreases by factor of three when four experiments are performed in each iteration, and by a factor of five for ten parallel experiments. However, the experimental effort increases with the degree of parallelization. For ten parallel experiments, the total number of experiments increases on average by a factor of 1.5. Hence, parallel experiments are only advisable when the saved time is worth the increased experimental effort. In all cases, the histogram of reached hypervolumes for parallel optimization does not differ significantly from the sequential case (data not shown).

5 Conclusions

We have introduced a novel algorithm for multi-objective optimization, Multi-Objective Global Optimization (MOGO). The MOGO algorithm is a direct but non-trivial extension of the single-objective equivalent EGO, developed by Jones et al. [10], for efficiently determining multi-dimensional Pareto fronts. Experimental data are analyzed by Gaussian process regression modeling (GPM), starting with an initial data set. In an iterative procedure, additional experiments are designed such as to maximize the expected hypervolume under the Pareto front, as calculated by Markov Chain Monte Carlo (MCMC) sampling. This way, process performance is maximized while prediction uncertainty is minimized in promising regions of the parameter space. New performed experiments are added to the data set and used for updating the GPM. The algorithm allows efficient design of both serial and parallel experiments with maximal information content. A global uncertainty indicator can be used for stopping the algorithm when the optimizer has converged with sufficient accuracy.

Approach and performance of the MOGO algorithm are demonstrated with a relevant case study from chemical engineering, chromatographic separation of three chemical components with five process parameters and three conflicting objectives. Elution gradient shape and pooling strategy are varied such as to optimize purity, yield, and process time. The MOGO algorithm effectively approximates the Pareto front in only 26 iterations. Parallel experiments allow reducing the number of iterations even further.

Lars Freier gratefully acknowledges a PhD scholarship by the Ministry of Innovation, Science and Research of North Rhine-Westphalia and the Heinrich Heine University Düsseldorf in the CLIB Graduate Cluster Industrial Biotechnology.

The authors declare no conflicts of interest.

Nomenclature

Symbol (unit)		description
Δt_1, Δt_2	…	(s) distance leftwards/rightward to the maximum of the peak of Cyt
p_1	…	(mol/m³) offset variable
p_2	…	(mol/m³) concentration at point of transition
p_3	…	(s) time point of transition
t_1, t_2	…	(s) cut times for pooling
CS	…	Conditional Simulation
cyt	…	cytochrome c
EHVI	…	Expected HyperVolume Improvement
GA	…	Genetic Algorithm
GPM	…	Gaussian Process Regression Model
lys	…	lysozyme
MCMC	…	Markov Chain Monte Carlo
MLE	…	Maximum Likelihood Estimation
MOGO	…	Multi-Objective Global Optimization
rna	…	ribonuclease A

Math symbols

P	…	Pareto front in the current iteration
P_{true}	…	True (ideal) Pareto Front
$\boldsymbol{x}$	…	Input point
X_{opt}	…	Pareto optimal set
$Z(\boldsymbol{x})$	…	Output value
$H(X)$	…	Hypervolume of the set X
H_{max}	…	Maximal theoretical achievable hypervolume
H^*_{max}	…	Approximation of the "true" hypervolume resulting from random sampling
H_{rel}	…	Relative hypervolume of the data set w.r.t. h_{rand}
$I(\boldsymbol{x})$	…	Improvement of after adding $\boldsymbol{x}$ to the data set
G	…	Global uncertainty of the Pareto front estimation
$C(\boldsymbol{x}_i, \boldsymbol{x}_j)$	…	Covariogram model
C	…	Covariance matrix
g	…	Exploration factor
$PDF_{\boldsymbol{x}}(Z)$	…	Probability density function of the output $Z(\boldsymbol{x})$

6 References

[1] Zeleny, M., *Multiple criteria decision making*. McGraw Hill Higher Education, 1982.
[2] Bhaskar, V., Gupta, S. K., Ray, A. K., Applications of multiobjective optimization in chemical engineering. *Rev. Chem. Eng.* 2000, *16*, 1–54.
[3] Zhang, Z., Hidajat, K., Ray, A. K., Morbidelli, M., Multiobjective optimization of SMB and varicol processes for chiral separation. *AIChE J.* 2002, *48*, 2800–2816.
[4] Van Veldhuizen, D. A. Lamont, G. B., Multiobjective evolutionary algorithms: Analyzing the state-of-the-art. *Evol. Comput.* 2000, *8*, 125–147.
[5] Santana-Quintero, L. V, Montano, A. A., Coello, C. A., A review of techniques for handling expensive functions in evolutionary multiobjective optimization, in: Tenne, Y., Goh, C.-K. (Eds.), *Computational Intelligence In Expensive Optimization Problems* Vol. 2, Springer Berlin Heidelberg 2010, pp. 29–59.
[6] Emmerich, M. T. M., Giannakoglou, K. C., Naujoks, B., Single-objective and multiobjective evolutionary optimization assisted by Gaussian random field metamodels. *IEEE Trans. Evol. Comput.* 2006, *10*, 421–439.
[7] Sasena, M. J., Papalambros, P. Y., Goovaerts, P., Metamodeling sampling criteria in a global optimization framework. *Am. Inst. Aeronaut. Astronaut.* 2000, doi: 10.2514/6.2000-4921.
[8] Hupkens, I., Deutz, A., Yang, K., Emmerich, M., Faster exact algorithms for computing expected hypervolume improvement, in: António Gaspar-Cunha, A., Henggeler Antunes, C., Coello, C. (Eds.), *Evolutionary Multi-Criterion Optimization*, Vol. 9019, Springer Berlin Heidelberg 2015, pp. 65–79.
[9] Binois, M., Ginsbourger, D., Roustant, O., Quantifying uncertainty on Pareto fronts with Gaussian process conditional simulations. *Eur. J. Oper. Res.* 2015, *243*, 386–394.
[10] Jones, D. R., Schonlau, M., William, J., Efficient Global Optimization of Expensive Black-Box Functions. *J. Global. Optim.* 1998, *13*, 455–492.
[11] Rasmussen, C. E., *Gaussian Processes for Machine Learning*. MIT Press, 2006.
[12] Cressie, N. A. C., Statistics for Spatial Data, 3rd Edn. Wiley New York, 1993.
[13] Marchant, B. P., Lark, R. M., The Matérn variogram model: Implications for uncertainty propagation and sampling in geostatistical surveys. *Geoderma* 2007, *140*, 337–345.
[14] Minasny, B., McBratney, A. B., The Matérn function as a general model for soil variograms. *Geoderma* 2006, *128*, 192–207.
[15] Mardia, K. V., Marshall, R. J., Maximum likelihood estimation of models for residual covariance in spatial regression, *Biometrika* 1984, *71*, 135–146.
[16] Zitzler, E., Thiele, L., Laumanns, M., Fonseca, C. M., Grunert, V., Performance assessment of multiobjective optimizers: An analysis and review. *IEEE Trans. Evol. Comput.* 2002, *7*, 117–132.
[17] Zitzler, E., Thiele, L., Multiobjective optimization using evolutionary algorithm. *Lect. Notes Comput. Sci.* 1998, *1498*, 292–301.

[18] Fleischer, M., The measure of Pareto optima. Applications to multi-objective metaheuristics, in: Fonseca, C. M., Fleming, P. J., Zitzler, E., Thile, L., Deb, K. (Eds.) *Evolutionary Multi-Criterion Optimization*, Springer Berlin 2003, pp. 519–533.
[19] Fonseca, C. M., Paquete, L., López-Ibáñez, M., An improved dimension-sweep algorithm for the hypervolume indicator. *IEEE Congress on Evolutionary Computation* 2006, 3973–3979.
[20] Emmerich, M. T. M., Deutz, A. H., Klinkenberg, J. W., Hypervolume-based expected improvement: Monotonicity properties and exact computation Proc. IEEE CEC 2011, *2011*, 2147–2154.
[21] Goovaerts, P., *Geostatistics for Natural Resources Evaluation*. Oxford: University Press on Demand, 1997.
[22] Chiles, J.-P., Delfiner, P., *Geostatistics: Modeling Spatial Uncertainty*, John Wiley & Sons, New York 2009.
[23] Chevalier, C., Emery, X., Ginsbourger, D., Fast update of conditional simulation ensembles. *Math. Geosci.* 2015, *47*, 771–789.
[24] Hansen, N., *The CMA Evolution Strategy: A Tutorial*, Research centre Saclay–Ile-de-France, Universite Paris-Saclay 2016.
[25] Cox, D. D., John, S., SDO: A statistical method for global optimization, in: Alexandrov, N., Hussaini, M. Y. (Eds.), *Multidisciplinary Design Optimization*, SIAM, Philadelphia 1997, pp. 315–329.
[26] Chevalier, C., Ginsbourger, D., Bect, J., Molchanov, I., *Estimating and Quantifying Uncertainties on Level Sets Using the Vorob'ev Expectation and Deviation with Gaussian Process Models*. Springer 2013.
[27] Lieres, E. von, Andersson, J., A fast and accurate solver for the general rate model of column liquid chromatography. *Comput. Chem. Eng.* 2010, *34*, 1180–1191.
[28] Püttmann, A., Schnittert, S., Naumann, U., Lieres, E. von, Fast and accurate parameter sensitivities for the general rate model of column liquid chromatography. *Comput. Chem. Eng.* 2013, *56*, 46–57.
[29] Guiochon, G., Felinger, A., Shirazi, D. G., Katti, A. M., *Fundamentals of Preparative and Nonlinear Chromatography*, 2nd Edn., Elsevier Academic Press, Amsterdam 2006.
[30] Lu, L., Anderson-Cook, C. M., Adapting the hypervolume quality indicator to quantify trade-offs and search efficiency Pareto fronts, *Qual. Reliab. Eng. Int.* 2013, *29*, 1117–1133.
[31] Haario, H., Laine, M., Mira, A., DRAM: Efficient adaptive MCMC. *Stat. Comput.*, 2006, *16*, 339–354.
[32] Ponweiser, W., Wagner, T., Vincze, M., Clustered multiple generalized expected improvement: A novel infill sampling criterion for surrogate models, *IEEE Congress on Evolutionary Computation* 2008.
[33] Ponweiser, W., Wagner, T., Biermann, D., Vincze, M., Multiobjective optimization on a limited budget of evaluations using model-assisted S-metric selection. 2008, 784–794.
[34] Schonlau, M., Welch, W. J., Jones, D. R., Global versus local search in constrained optimization of computer models. in: *New Developments and Applications in Experimental Design, Lecture Notes – Monograph Series* 1998, pp. 11–25.
[35] Sasena, M. J., Papalambros, P., Goovaerts, P., Exploration of metamodeling sampling criteria for constrained global optimization, *Eng. Optim.*, 2002, *34*, 263–278.

3 Results and Discussion

In biotechnology, there exists a strong demand for algorithms that find optimal solutions with minimal experimental effort. In this context, optimality criteria refer to maximization or minimization of defined quantities such as protein concentration, biomass concentration or enzyme activity. The optimization is usually done iteratively. New experiments are designed based on the data set acquired in the previous iterations. In order to keep the number of performed experiments at a minimum, experiments can partially be replaced by mathematical models, which estimate the input-output relationship. There exist two different types of mathematical model, mechanistic and empirical models. Mechanistic mathematic models are based on mechanistic understanding and an appropriate chosen model results in accurate prediction outside of the data set. However, finding an appropriated mechanistic model and to estimate the associated model parameters is a not trivial task and, for process optimization, empirical models are therefore often preferred. Empirical models are constructed based on the experimental data but not on mechanistic knowledge.

An exception represents the Kriging approach that is mainly an empirical modeling approach but also provides the possibility to integrate mechanistic models, see section 1.3. These mechanistic models, called trend functions in this context, are in particular helpful in parameter regions where the sample density is rather low such that a purely data driven model would be very inaccurate. The current variants of Kriging however do only support models that are linear in their parameters, but many mechanistic models in biotechnology are typically nonlinear in their parameter, e.g. the Michaelis-Menten model for enzyme kinetics. In **publication I** (section 2.1), a methodology is introduced that allows the integration of nonlinear trend functions. Here, the parameter estimation problem is relaxed by Taylor linearization, which is further used for formulating an iterative parameter estimation approach. This iterative approach is then converted into a root-finding problem making it accessible for numerical solvers that are specifically developed for this kind of problems.

Kriging can not only be applied as surrogate model for “real” experiments but can also be used for designing the input values of new experiments with potentially promising

output values. For this goal, both the Kriging prediction as well as the estimated model uncertainty is used, leading to the concept of Expected Improvement (EI), see section 1.4. In particular, the state-of-the-art Efficient Global Optimization (EGO) algorithm, pusblished by Jones et al. [1], makes use of EI. EGO is an iterative optimization algorithm that constructs first a Kriging model based on an initial data set. The Kriging model is then used for designing new experiments by maximizing the EI, i.e. a new experiment is performed using the input variable values where EI has its maximum. The loop is afterwards closed by updating the Kriging model.

EGO is an effective optimization algorithm but, only one experiment is designed in each iteration and the algorithm is furthermore not flexible with respect to changes in the ranges of the input variables. However, for handling biotechnological optimization tasks, parallel optimization and flexibility are desired. Moreover, caused by the high complexity of a biological system, many input variables with potential influence on the output variables can be identified although only a minority has actually a significant effect. Therefore, in order to reduce the experimental effort, a sensitivity analysis is often required at the beginning for focusing on input variables with significant influence during the remaining optimization procedure. In section 1.5, I have introduced a framework for a Kriging based optimization strategy that tackles all of these mentioned demands. The design of multiple experiments is enabled by applying the Markov-Chain Monte-Carlo (MCMC) procedure. MCMC is an iterative procedure where in each iteration one sample point is generated. For a sufficient number of iterations, the histogram of sample locations approximates well the actual probability density distribution. In this specific case, the histogram is approximating the EI landscape. That is, the sample distribution is spread over the entire input space but concentrated on areas associated with high EI values. New experiments are designed by uniformly random drawing from the resulting MCMC. If only a sequential optimization is desired, the sample location with the highest EI value is chosen. The MCMC based experimental design represents therefore a powerful tool for both sequential and parallel experimentation.

Furthermore, a screening is conducted upstream in the iterative optimization cycle based on a “classical” design of experiment study. During the iterative optimization, new experiments are either designed based on the EI or by expanding the input space and placing new experiments in the added areas of the input parameter space.

The developed framework was successfully applied in two experimental studies. The first study is presented in **publication II** (section 2.2), and aimed to maximize the product titer of secreted Green Fluorescent Protein (GFP), which is secreted by a recombinant *Corynebacterium glutamicum* strain. The GFP titer was optimized by varying the composition of the growth medium. Cultivations were carried out in a microbioreactor system integrated with a liquid handling robot allowing parallel experimentation with online fluorescence detection as well as automated media preparation. For comparison, the established CgXII minimal medium composition, which is described in literature for cultivation of *C. glutamicum*, was used as reference medium. The sensitivity analysis was conducted using a fractional factorial design and helped to reduce the number of relevant medium components from eleven to three. The concentration of these three medium components was subsequently optimized in the described iterative cycle while the concentrations of the remaining components were kept constant at their values of the reference medium. During the optimization, it was necessary to extend the concentration ranges of the optimized components several times. In fact, the maximal concentration of one component was successively increased to the 32-fold of the original upper limit. This highlights that, caused by the high complexity of a biological system, it is often hard to define appropriate input variable ranges *a priori*. Consequently, there is a need for adapting these ranges during the optimization. After seven iterations, comprising a total number of 32 experiments, a plateau of the GFP titer was identified and the optimization was stopped. Medium compositions that are located at this plateau have a GFP titer double as high as for the reference medium.

The second experimental study is presented in **publication III** (section 2.3), and addresses also a medium optimization but aimed at maximizing lipid production of the photoautotrophic microalga *Chlorella vulgaris*. The reference medium was again adapted from literature. The application of a liquid handling robot-assisted micro-photo-bioreactor allowed to automate the medium preparation as well as the incubation and to conduct up to 48 experiments in parallel. Similar to the first experimental study, a fractional factorial design was applied for reducing the number of relevant medium component from nine to four. Following the sensitivity analysis, new experiments were designed by applying a full factorial design and adding manually sample points leading to a space filling experimental design. The resulting Kriging model showed a clear

optimum. In a second iteration, multiple experiments were designed applying the EI and MCMC based methodology, introduced in section 1.5.1. As the prediction of the Kriging model was similar to the previous iteration, the optimization was stopped at this point. The optimized medium composition leads to a lipid productivity that is three-fold as high as for the reference medium. This study demonstrates the importance of simultaneous design of multiple experiments combined with parallel experimentation. Parallel experimentation is in particular important in case of time-consuming experiments, such as the cultivation of *Chlorella vulgaris*, which needed one week per experimental run. Only through parallel experimentation, the number of iterations could be kept to the small number of two.

In both case studies, optimization was performed by maximizing only a single objective. In many biotechnological relevant optimization studies, however, there exist multiple competing objectives of interests, e.g. yield, purity, or productivity. Due to this conflict, no unique optimal solution exists, and multi-objective optimization (MOO) can be used for finding the best compromises between the solutions. Conceptually, these compromises lie on a curve referred to as Pareto front. A scientific achievement of this thesis is the development of a multi-objective equivalent to the EGO algorithm. The resulting Multi-Objective Gaussian Optimization (MOGO) algorithm is described in **publication IV** (section 2.4). MOGO integrates the only recently developed mathematical concept of Expected HyperVolume Improvement (EHVI) and latest algorithms for quantifying uncertainty of the Pareto front estimation. EHVI is applied analogously to EI for designing new experiments. While the concept of EHVI was first introduced in 2006 by Emmerich et al. [48], only 2015 an efficient algorithm for its analytic calculation was published and EHVI became available for practical applications. On the other hand, quantifiers of the Pareto front estimation uncertainty can be used as stopping criteria for preventing unnecessary iterations. As described in detail in **publication IV**, the prediction uncertainty is calculated by analyzing the results of many generated stochastic simulations that follow the Gaussian process modeled by Kriging. This calculation procedure was first published in 2015 by Binois [50].

In **publication IV**, the convergence behavior is furthermore studied as well as the reproducibility and the parallelization capability in an *in silico* case study of practical relevance. Here, the protein cytochrome C should be isolated from a three-component

mixture using ion-exchange chromatography. The simulation was conducted using the open-source Chromatography Analysis and Design Toolkit (CADET) [53]. The separation process has five input parameters and three competing objectives. The convergence behavior was investigated using the hypervolume as quality indicator for the approximated Pareto front. It was demonstrated that with high reproducibility more than 90% of the maximal possible hypervolume was achieved after 26 iterations, which corresponds to 34 experiments. It is worth mentioning at this point that other state-of-the-art MOO algorithms, such as the Non-dominated Sorting Genetic Algorithm (NSGA-II) [45], require already 34 experiments after much less iterations. This is caused by the fact that the majority of MOO algorithms are based on Genetic Algorithms (GA) that require a relatively high amount of data points in each iteration for converging. For example, Tan et al. [54] tested different GA based MOOs for an optimization problem with three objective, as also be done in **publication IV**. As result, the most efficient algorithm needed around 1000 experiments. The higher efficiency of MOGO is *inter alia* caused by the fact that MOGO keeps record of the data set in previous iterations leading to a successive increase in the Kriging model prediction power. MOGO does consequently “remember” suboptimal locations and no further experiment is designed here.

The stopping criterion, based on the prediction uncertainty, was used to detect convergence and the optimization was stopped after the uncertainty fell below a defined threshold. The results demonstrate again that by using parallel experimentation in the context of MOGO, the optimization converges faster. While for the sequential optimization, 26 iterations were on average needed, for instance nine iterations were on average sufficient performing four experiments in parallel. However, with increasing number of parallel experiments, the efficiency of the optimization algorithm decreases and a higher number of experiments are required in total. The experimentalist needs consequently to decide for a trade-off between a fast and an efficient optimization.

In order to make all introduced methods and algorithms available for the scientific community and to facilitate reproduction of the published results, I have developed and published the Kriging toolKit (KriKit), Figure 10. KriKit is implemented in MATLAB and freely available on github: https://github.com/modsim/KriKit. The software allows *inter alia* the construction of Kriging models based on given data sets using either linear or

nonlinear mechanistic models. Moreover, the toolbox allows the visualization of Kriging predictions, confidence tubes, and measurement data using 2D, 3D, and contour plots as well as animated videos. KriKit can also be used for hypothesis testing, taking into account the expected Kriging prediction and the model uncertainty. Further key features of KriKit are experimental design based on EI and EHVI for solving optimization tasks.

Figure 10; Logo of the Kriging toolkit

4 Conclusions and Outlook

The here presented PhD thesis provides two Kriging Based Optimization (KBO) approaches, for single-objective and multi-objective cases, respectively. Both algorithms are based on interactive procedures where new experiments are designed in each iteration utilizing the concepts of expected improvement and Markov chain Monte-Carlo. The combination of both concepts is novel and allows sequential as well as parallel experimentation.

The single-objective algorithm was successfully applied to two media optimization problems with respect for maximizing of protein titer and lipid productivity, respectively. Using KBO, the protein titer could be increased by a factor of two and the lipid productivity by a factor of three compared to using the initial medium composition. These experimental studies also demonstrated the effective use of parallel experimentation in the context KBO, which is discussed in detail in this thesis.

The multi-objective algorithms was applied to a practical relevant *in silico* optimization problem from the field of chromatography, aimed at the simultaneous optimization of three competing objectives, namely purity, yield and process time. The utilization of *in silico* experimentation allowed a comprehensive investigation of the convergence behavior, the reproducibility, and of the effective use of parallel experimentation.

However, it is noteworthy that the developed approaches are also applicable to optimization tasks from other fields. KBO is suitable whenever the input and output variables are continuous. The spectrum of optimization problems fulfilling these conditions is quite broad and comprises for example the estimation of model parameter values for modeling fitting problems, process design in process engineering, etc. It would be therefore interesting to demonstrate this universality by applying the developed approach to a variety of different optimization problems.

However, in some situations, it can be an advantage to adapt parts of the developed KBO strategies to the system specification. For example, considering a system with different accessibility of the input variables, where some input variables can be varied between individual experiments, while others can be varied only between parallel runs. For instance, using microtiter plates where several cultivations can be run in parallel on one plate and the concentrations of the medium components can be varied between the

different wells. However, cultivation conditions such as the temperature can only be varied for each plate. A suitable adaptation has consequently to classify the input variables to two categories: variable between individual experiments and variable between parallel runs.

As discussed in this thesis, integrating appropriate trend functions into the Kriging procedure can potentially increase accuracy of the model prediction and consequently also the accuracy of the expected improvement calculation. This can lead to faster convergence toward optima. As many trend functions are non-linear in their parameters, an approach was developed extending the Kriging methodology by this feature. Although the positive effect of utilizing appropriate trend functions was demonstrated for the prediction accuracy, the influence on the KBO has not yet been investigated and might fall within the scope of future research.

A further not yet investigated research question is how to handle stochastic variations in the input variables, which occur in some experimental scenarios. That is, some input variables might be measurable but not controllable and underlie random fluctuations. The question arises how to model the error propagation such that the Kriging prediction error also comprises the uncertainty regarding the input values and how to deal with it during the optimization. In other words, how to design a process that is robust to these stochastic variations and fulfils defined quality criteria with a pre-specified probability.

As overall conclusion, this thesis provides a solution of integrating trend functions that are nonlinear in their parameters into the Kriging methodology. Also, the MOGO algorithm is introduced that transfers the state of the art KBO algorithm “Efficient Global Optimization” to the multi-objective case. Further, in order to make KBO algorithms more attractive to biotechnological applications, a framework was introduce providing instruction for conducting an initial screening, handling variations in input variables ranges, and performing parallel experimentation. Methods and algorithms made available to the scientific community by embedding them in the Kriging toolKit (KriKit). KriKit is implemented in MATLAB and freely available on github: https://github.com/modsim/KriKit.

Appendix

Supplement to “Framework for Kriging-based iterative experimental analysis and design: Optimization of secretory protein production in Corynebacterium glutamicum”

1 Experimental Design for Initial Screening

Supplementary Table 1: Experimental design using coded values. -1 and 1 indicate respectively the minimal/maximal concentration.

Number of Experiment	Coded volume / concentration of component										
	Fe	Mn	Zn	Cu	NH4	Ni	Co	Mo	BO	Ca	Mg
1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
2	1	-1	-1	-1	-1	-1	-1	1	1	-1	1
3	-1	1	-1	-1	-1	1	-1	-1	-1	1	1
4	1	1	-1	-1	-1	1	-1	1	1	1	-1
5	-1	-1	1	-1	-1	1	1	1	-1	-1	1
6	1	-1	1	-1	-1	1	1	-1	1	-1	-1
7	-1	1	1	-1	-1	-1	1	1	-1	1	-1
8	1	1	1	-1	-1	-1	1	-1	1	1	1
9	-1	-1	-1	1	-1	1	1	1	1	1	1
10	1	-1	-1	1	-1	1	1	-1	-1	1	-1
11	-1	1	-1	1	-1	-1	1	1	1	-1	-1
12	1	1	-1	1	-1	-1	1	-1	-1	-1	1
13	-1	-1	1	1	-1	-1	-1	-1	1	1	-1
14	1	-1	1	1	-1	-1	-1	1	-1	1	1
15	-1	1	1	1	-1	1	-1	-1	1	-1	1
16	1	1	1	1	-1	1	-1	1	-1	-1	-1
17	-1	-1	-1	-1	1	-1	1	-1	1	1	1
18	1	-1	-1	-1	1	-1	1	1	-1	1	-1
19	-1	1	-1	-1	1	1	1	-1	1	-1	-1
20	1	1	-1	-1	1	1	1	1	-1	-1	1
21	-1	-1	1	-1	1	1	-1	1	1	1	-1
22	1	-1	1	-1	1	1	-1	-1	-1	1	1
23	-1	1	1	-1	1	-1	-1	1	1	-1	1
24	1	1	1	-1	1	-1	-1	-1	-1	-1	-1
25	-1	-1	-1	1	1	1	-1	1	-1	-1	-1
26	1	-1	-1	1	1	1	-1	-1	1	-1	1
27	-1	1	-1	1	1	-1	-1	1	-1	1	1
28	1	1	-1	1	1	-1	-1	-1	1	1	-1
29	-1	-1	1	1	1	-1	1	-1	-1	-1	1
30	1	-1	1	1	1	-1	1	1	1	-1	-1
31	-1	1	1	1	1	1	1	-1	-1	1	-1
32	1	1	1	1	1	1	1	1	1	1	1

Supplementary Table 2: Code for experimental Design in Supplementary Table 1

Code	Volume of stock solution in µL										
	Fe	Mn	Zn	Cu	NH4	Ni	Co	Mo	BO	Ca	Mg
-1	10	10	10	10	20	10	0	0	0	10	10
1	40	40	40	40	80	40	40	40	40	40	40

	Fe	Mn	Zn	Cu	NH4	Ni	Co	Mo	BO	Ca	Mg
Conc. in stock solution in g/L:	0.4	0.4	0.04	0.01252	400	0.0008	0.0052	0.0026	0.002	0.53	10

Code	Concentration of component in mmol/L										
	Fe	Mn	Zn	Cu	NH4	Ni	Co	Mo	BO	Ca	Mg
-1	1.4E-02	2.4E-02	1.4E-03	5.0E-04	1.2E+02	3.4E-05	0.0E+00	0.0E+00	0.0E+00	3.6E-02	4.1E-01
1	5.8E-02	9.5E-02	5.6E-03	2.0E-03	4.8E+02	1.4E-04	8.7E-04	4.3E-04	1.3E-03	1.4E-01	1.6E+00

2 Optimized Media

Supplementary Table 3: Comparison of concentration of medium components between standard CgXII medium ("reference") and optimized composition. $Na_2MoO_4 * 2 H_2O$ and H_3BO_3 are not included in the standard CgXII composition, but are listed in the table as these components were included in screening analyses before and after iterative medium optimization.

Component	**Concentration in reference medium composition**	**Concentration in optimized medium composition**	**Fold-change from reference to optimal**
Glucose	10 g/L	10 g/L	1 X (no change)
$(NH_4)_2SO_4$	**20 g/L**	**0 g/L**	**0 X**
Urea	5 g/L	5 g/L	1 X (no change)
KH_2PO_4	1 g/L	1 g/L	1 X (no change)
K_2HPO_4	1 g/L	1 g/L	1 X (no change)
MOPS	42 g/L	42 g/L	1 X (no change)
$FeSO_4 * 7 H_2O$	10 mg/L	10 mg/L	1 X (no change)
$MnSO_4 * H_2O$	10 mg/L	10 mg/L	1 X (no change)
$ZnSO_4 * 7 H_2O$	1 mg/L	1 mg/L	1 X (no change)
$CuSO_4 * 5 H_2O$	0.31 mg/L	0.31 mg/L	1 X (no change)
$NiCl_2 * 6 H_2O$	0.02 mg/L	0.02 mg/L	1 X (no change)
$CaCl_2 * 2 H_2O$	**0.01325 g/L**	**0.424 g/L**	**32 X**
$MgSO_4 * 7 H_2O$	**0.25 g/L**	**1.7 g/L**	**6.8 X**
$CoCl_2 * 6 H_2O$	0.31 mg/L	0.31 mg/L	1 X (no change)
$Na_2MoO_4 * 2 H_2O$	0 g/L	0 g/L	1 X (no change)
H_3BO_3	0 g/L	0 g/L	1 X (no change)
Protocatechuic acid (PCA)	30 mg/L	30 mg/L	1 X (no change)
Biotin	0.2 mg/L	0.2 mg/L	1 X (no change)

Supplement to "A framework for accelerated phototrophic bioprocess development: integration of parallelized microscale cultivation, laboratory automation and Kriging-assisted experimental design"

Additional File 1

Table 1: Experimental design and results, fractional factorial – Part 1

#experiment	MES [mmol L-1]	NaNO3 [g L-1]	K2HPO4 / KH2PO4 [g L-1]	NaCl [mmol L-1]	MgSO4 [mmol L-1]
1	5.00E+01	1.50E+00	2.00E+00	1.07E-01	7.61E-01
2	5.00E+01	1.50E+00	2.00E+00	1.07E-01	7.61E-01
3	5.00E+01	1.50E+00	2.00E+00	1.07E-01	7.61E-01
4	5.00E+01	1.50E+00	2.00E+00	1.07E-01	7.61E-01
5	5.00E+01	1.50E+00	2.00E+00	1.07E-01	7.61E-01
6	0.00E+00	2.50E-01	2.50E-01	5.00E+00	2.50E+00
7	0.00E+00	2.50E-01	2.50E+00	0.00E+00	2.50E+00
8	7.50E+01	2.50E+00	2.50E-01	0.00E+00	2.50E+00
9	0.00E+00	2.50E+00	2.50E+00	5.00E+00	2.50E+00
10	0.00E+00	2.50E+00	2.50E-01	5.00E+00	1.00E-01
11	7.50E+01	2.50E-01	2.50E+00	5.00E+00	1.00E-01
12	7.50E+01	2.50E-01	2.50E-01	5.00E+00	2.50E+00
13	7.50E+01	2.50E+00	2.50E+00	0.00E+00	1.00E-01
14	7.50E+01	2.50E-01	2.50E+00	0.00E+00	2.50E+00
15	0.00E+00	2.50E-01	2.50E-01	5.00E+00	2.50E+00
16	0.00E+00	2.50E+00	2.50E-01	5.00E+00	1.00E-01
17	7.50E+01	2.50E-01	2.50E-01	0.00E+00	1.00E-01
18	7.50E+01	2.50E+00	2.50E-01	5.00E+00	1.00E-01
19	7.50E+01	2.50E+00	2.50E+00	5.00E+00	2.50E+00
20	0.00E+00	2.50E+00	2.50E+00	5.00E+00	2.50E+00
21	0.00E+00	2.50E-01	2.50E-01	0.00E+00	1.00E-01
22	7.50E+01	2.50E-01	2.50E+00	0.00E+00	2.50E+00
23	7.50E+01	2.50E-01	2.50E+00	5.00E+00	1.00E-01
24	0.00E+00	2.50E+00	2.50E+00	0.00E+00	1.00E-01
25	0.00E+00	2.50E+00	2.50E-01	0.00E+00	2.50E+00
26	0.00E+00	2.50E+00	2.50E+00	0.00E+00	1.00E-01
27	7.50E+01	2.50E+00	2.50E-01	5.00E+00	1.00E-01
28	0.00E+00	2.50E-01	2.50E+00	0.00E+00	2.50E+00
29	0.00E+00	2.50E+00	2.50E-01	0.00E+00	2.50E+00
30	7.50E+01	2.50E+00	2.50E+00	0.00E+00	1.00E-01
31	7.50E+01	2.50E-01	2.50E-01	5.00E+00	2.50E+00
32	0.00E+00	2.50E-01	2.50E+00	5.00E+00	1.00E-01
33	0.00E+00	2.50E-01	2.50E-01	0.00E+00	1.00E-01
34	7.50E+01	2.50E-01	2.50E-01	0.00E+00	1.00E-01
35	0.00E+00	2.50E-01	2.50E+00	5.00E+00	1.00E-01
36	7.50E+01	2.50E+00	2.50E-01	0.00E+00	2.50E+00
37	7.50E+01	2.50E+00	2.50E+00	5.00E+00	2.50E+00

Table 2: Experimental design and results, fractional factorial – Part 2

#experiment	CaCl2 [mmol L-1]	trace [x fold]	FeSO4 [mmol L-1]	EDTA [mmol L-1]	Lipid Productivity [a.u.]
1	8.50E-01	1.00E+00	4.00E-03	2.97E-01	2.92E+02
2	8.50E-01	1.00E+00	4.00E-03	2.97E-01	3.17E+02
3	8.50E-01	1.00E+00	4.00E-03	2.97E-01	2.92E+02
4	8.50E-01	1.00E+00	4.00E-03	2.97E-01	2.75E+02
5	8.50E-01	1.00E+00	4.00E-03	2.97E-01	2.56E+02
6	2.50E+00	2.50E+00	4.00E-03	0.00E+00	2.38E+02
7	1.00E-01	2.50E+00	4.00E-02	0.00E+00	9.22E+02
8	2.50E+00	2.50E+00	4.00E-02	0.00E+00	5.65E+02
9	2.50E+00	1.00E-01	4.00E-02	0.00E+00	1.61E+01
10	2.50E+00	2.50E+00	4.00E-03	1.00E+00	5.21E+02
11	2.50E+00	2.50E+00	4.00E-02	0.00E+00	2.95E+02
12	1.00E-01	1.00E-01	4.00E-02	0.00E+00	3.57E+02
13	2.50E+00	1.00E-01	4.00E-03	1.00E+00	7.45E+01
14	2.50E+00	1.00E-01	4.00E-03	0.00E+00	6.07E+02
15	1.00E-01	2.50E+00	4.00E-02	1.00E+00	5.15E+02
16	1.00E-01	2.50E+00	4.00E-02	0.00E+00	2.86E-01
17	1.00E-01	2.50E+00	4.00E-03	0.00E+00	1.10E+02
18	1.00E-01	1.00E-01	4.00E-02	1.00E+00	1.91E-01
19	2.50E+00	2.50E+00	4.00E-02	1.00E+00	2.33E+01
20	1.00E-01	1.00E-01	4.00E-03	1.00E+00	3.66E-01
21	1.00E-01	1.00E-01	4.00E-03	1.00E+00	3.32E-01
22	1.00E-01	1.00E-01	4.00E-02	1.00E+00	3.32E+00
23	1.00E-01	2.50E+00	4.00E-03	1.00E+00	1.81E-01
24	1.00E-01	2.50E+00	4.00E-02	1.00E+00	5.65E+00
25	1.00E-01	1.00E-01	4.00E-03	0.00E+00	1.75E+01
26	2.50E+00	2.50E+00	4.00E-03	0.00E+00	3.71E-01
27	2.50E+00	1.00E-01	4.00E-03	0.00E+00	3.39E+02
28	2.50E+00	2.50E+00	4.00E-03	1.00E+00	7.21E+02
29	2.50E+00	1.00E-01	4.00E-02	1.00E+00	2.69E+01
30	1.00E-01	1.00E-01	4.00E-02	0.00E+00	1.28E+01
31	2.50E+00	1.00E-01	4.00E-03	1.00E+00	7.86E+01
32	2.50E+00	1.00E-01	4.00E-02	1.00E+00	1.64E-01
33	2.50E+00	1.00E-01	4.00E-02	0.00E+00	3.23E-02
34	2.50E+00	2.50E+00	4.00E-02	1.00E+00	7.71E+02
35	1.00E-01	1.00E-01	4.00E-03	0.00E+00	6.71E+02
36	1.00E-01	2.50E+00	4.00E-03	1.00E+00	6.41E+00
37	1.00E-01	2.50E+00	4.00E-03	0.00E+00	1.38E+01

Additional File 2

Table 3: Media Compositions

component	concentration enBBMref [mmol L-1]	concentration enBBMopt [mmol L-1]	concentration enBBMopt,min [mmol L-1]
CaCl2	8.50E-01	1.06E+00	1.06E+00
CoSO4	3.30E-04	8.25E-04	8.25E-04
CuSO4	1.00E-03	2.50E-03	2.50E-03
FeSO4	4.00E-03	4.00E-03	4.00E-03
H3BO3	3.70E-02	9.25E-02	9.25E-02
K2HPO4 / KH2PO4	1.37E+01	1.37E+01	1.72E+00
KOH	1.11E+00	1.11E+00	0.00E+00
MES	5.00E+01	5.00E+01	0.00E+00
MgSO4	7.61E-01	2.50E+00	2.50E+00
MnCl2	2.00E-03	5.00E-03	5.00E-03
NaCl	1.07E-01	1.07E-01	0.00E+00
Na2EDTA	2.97E-01	2.97E-01	0.00E+00
Na2MoO4	1.00E-03	2.50E-03	2.50E-03
NaNO3	1.76E+01	7.94E+00	7.94E+00
penicillin-G	2.81E-01	2.81E-01	2.81E-01
ZnSO4	6.00E-03	1.50E-02	1.50E-02

Additional File 3

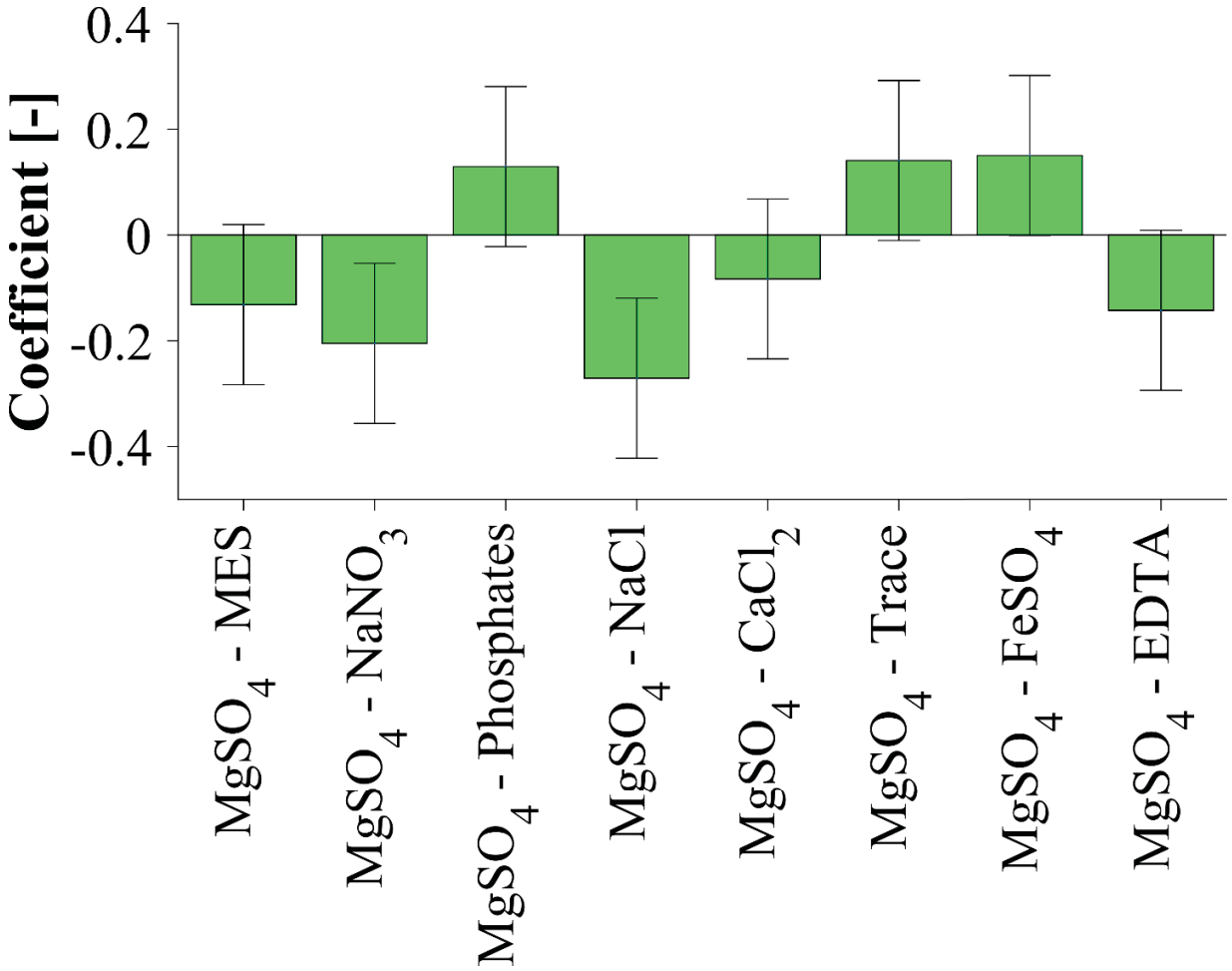


Figure 1: Coefficients for pair-wise interaction with $MgSO_4$

Additional File 4

Table 4: Experimental design and results, full factorial – Part 1

#experiment	MES [mmol L-1]	NaNO3 [g L-1]	K2HPO4 / KH2PO4 [g L-1]	NaCl [mmol L-1]	MgSO4 [mmol L-1]
1	50	1.5	2	0.107	0.761
2	50	1.5	2	0.107	0.761
3	50	1.5	2	0.107	0.761
4	50	1.5	2	0.107	0.761
5	50	2.5	2	0	0.1
6	50	0.25	2	5	2.5
7	50	2.5	2	0	2.5
8	50	2.5	2	0	2.5
9	50	0.25	2	2.5	1.9
10	50	2.5	2	5	0.1
11	50	0.25	2	5	2.5
12	50	0.25	2	0	2.5
13	50	0.25	2	5	2.5
14	50	0.25	2	5	2.5
15	50	2.5	2	5	2.5
16	50	0.25	2	5	0.1
17	50	2.5	2	5	0.1
18	50	0.25	2	0	2.5
19	50	2.5	2	0	2.5
20	50	2.5	2	5	2.5
21	50	2.5	2	0	0.1
22	50	0.25	2	0	0.1
23	50	2.5	2	0	0.1
24	50	2.5	2	0	0.1
25	50	0.25	2	0	0.1
26	50	2.5	2	5	2.5
27	50	0.8125	2	2.5	1.9
28	50	0.25	2	0	2.5
29	50	0.25	2	0	0.1
30	50	0.8125	2	0	2.5
31	50	0.25	2	5	0.1
32	50	0.25	2	0	0.1
33	50	0.25	2	5	0.1
34	50	0.25	2	0	2.5
35	50	2.5	2	5	2.5
36	50	2.5	2	0	2.5
37	50	2.5	2	5	0.1
38	50	0.25	2	5	0.1
39	50	2.5	2	5	0.1

Table 5: Experimental design and results, full factorial – Part 2

#experiment	CaCl2 [mmol L-1]	trace [x fold]	FeSO4 [mmol L-1]	EDTA [mmol L-1]	Lipid Productivity [a.u.]
1	0.85	1	0.004	0.297	356.5913129
2	0.85	1	0.004	0.297	373.2704665
3	0.85	1	0.004	0.297	338.6284309
4	0.85	1	0.004	0.297	365.9636463
5	2.5	2.5	0.004	0.297	361.1236215
6	0.1	0.1	0.004	0.297	8.854160233
7	2.5	2.5	0.004	0.297	44.65471473
8	2.5	0.1	0.004	0.297	29.53868928
9	2.5	2.5	0.004	0.297	831.9867314
10	2.5	0.1	0.004	0.297	468.9142965
11	2.5	0.1	0.004	0.297	329.6298815
12	2.5	0.1	0.004	0.297	316.2274568
13	2.5	2.5	0.004	0.297	767.1525227
14	0.1	2.5	0.004	0.297	817.1550468
15	2.5	2.5	0.004	0.297	25.07835108
16	0.1	0.1	0.004	0.297	0.329783553
17	0.1	0.1	0.004	0.297	5.421150708
18	0.1	2.5	0.004	0.297	816.2714476
19	0.1	2.5	0.004	0.297	23.62594111
20	2.5	0.1	0.004	0.297	23.74003025
21	2.5	0.1	0.004	0.297	711.3313574
22	0.1	2.5	0.004	0.297	148.3129184
23	0.1	0.1	0.004	0.297	0.331542145
24	0.1	2.5	0.004	0.297	636.4821649
25	0.1	0.1	0.004	0.297	3.663354735
26	0.1	0.1	0.004	0.297	7.870234186
27	1.9	1.9	0.004	0.297	1542.929642
28	2.5	2.5	0.004	0.297	796.7885052
29	2.5	2.5	0.004	0.297	628.545046
30	1.9	1.9	0.004	0.297	1621.528419
31	2.5	0.1	0.004	0.297	295.6224386
32	2.5	0.1	0.004	0.297	358.8520543
33	2.5	2.5	0.004	0.297	764.7553045
34	0.1	0.1	0.004	0.297	11.77646422
35	0.1	2.5	0.004	0.297	42.12366806
36	0.1	0.1	0.004	0.297	8.285368399
37	0.1	2.5	0.004	0.297	364.4682896
38	0.1	2.5	0.004	0.297	326.6295673
39	2.5	2.5	0.004	0.297	418.0846118

Additional File 5

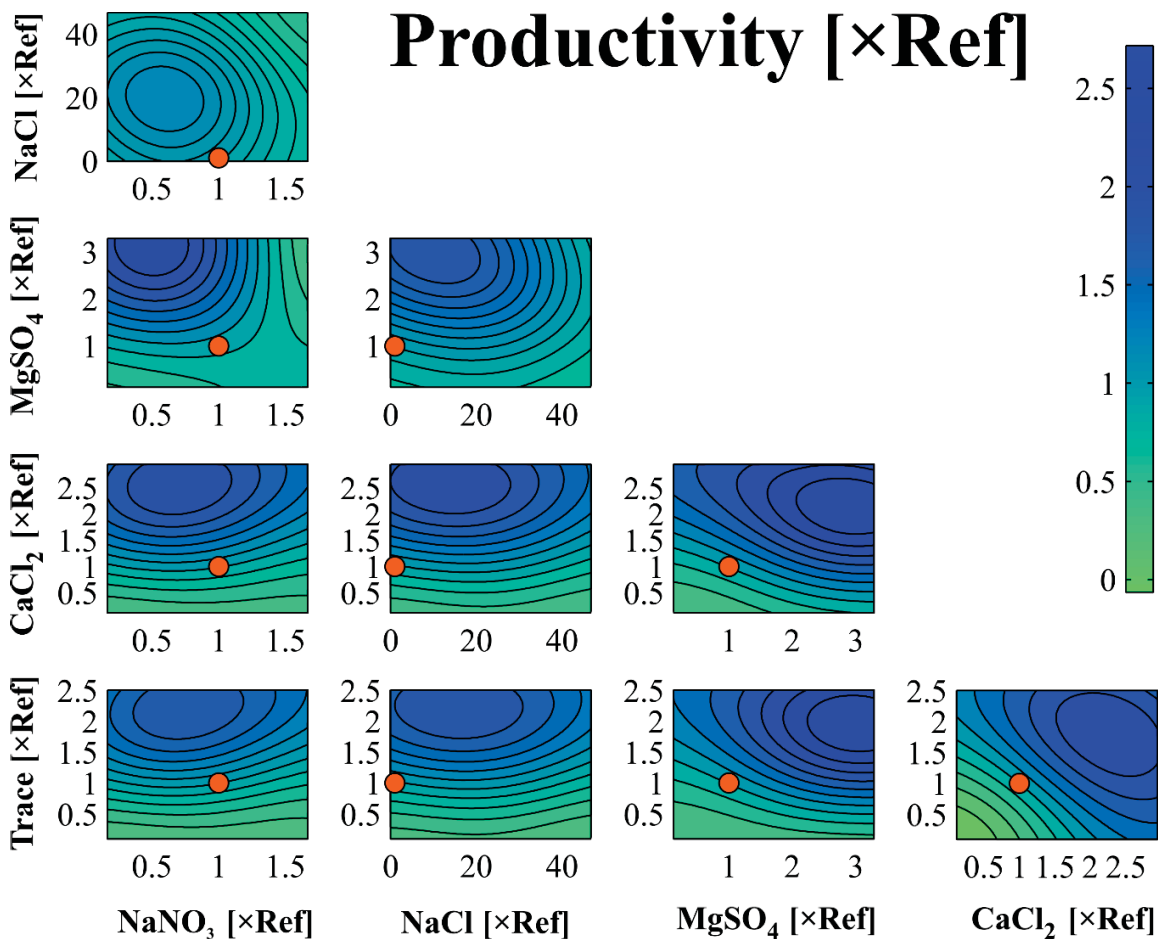


Figure 2: Screening plot for full factorial design

Additional File 6

Table 6: Experimental design and results, locating optimum – Part 1

#experiment	MES [mmol L-1]	NaNO3 [g L-1]	K2HPO4 / KH2PO4 [g L-1]	NaCl [mmol L-1]	MgSO4 [mmol L-1]
1	50	1.5	2	0.107	0.761
2	50	1.5	2	0.107	0.761
3	50	1.5	2	0.107	0.761
4	50	1.5	2	0.107	0.761
5	50	1.5	2	0.107	2.5
6	50	1.5	2	0.107	0.1
7	50	0.5625	2	0.107	1.3
8	50	0.875	2	0.107	1.3
9	50	1.1875	2	0.107	1.3
10	50	0.875	2	0.107	2.5
11	50	0.5625	2	0.107	1.3
12	50	1.1875	2	0.107	1.3
13	50	1.5	2	0.107	2.5
14	50	0.25	2	0.107	2.5
15	50	1.5	2	0.107	0.1
16	50	0.25	2	0.107	2.5
17	50	0.5625	2	0.107	1.3
18	50	0.875	2	0.107	0.1
19	50	0.5625	2	0.107	1.3
20	50	0.25	2	0.107	2.5
21	50	1.5	2	0.107	2.5
22	50	1.1875	2	0.107	1.3
23	50	0.25	2	0.107	0.1
24	50	0.25	2	0.107	0.1
25	50	1.5	2	0.107	2.5
26	50	0.25	2	0.107	0.1
27	50	1.5	2	0.107	0.1
28	50	1.1875	2	0.107	1.3
29	50	1.5	2	0.107	2.5
30	50	0.25	2	0.107	2.5
31	50	0.875	2	0.107	0.1
32	50	0.875	2	0.107	2.5
33	50	0.5625	2	0.107	1.3
34	50	1.1875	2	0.107	1.3
35	50	0.25	2	0.107	0.1
36	50	0.25	2	0.107	0.1
37	50	1.5	2	0.107	0.1
38	50	0.25	2	0.107	2.5
39	50	1.5	2	0.107	0.1

Table 7: Experimental design and results, locating optimum – Part 2

#experiment	CaCl2 [mmol L-1]	trace [x fold]	FeSO4 [mmol L-1]	EDTA [mmol L-1]	Lipid Productivity [a.u.]
1	0.85	1	0.004	0.297	340.9035096
2	0.85	1	0.004	0.297	333.6144351
3	0.85	1	0.004	0.297	312.5471286
4	0.85	1	0.004	0.297	290.1184726
5	1.3	0.1	0.004	0.297	121.2865784
6	0.1	1.925	0.004	0.297	123.8941969
7	1.9	1.0125	0.004	0.297	1433.279089
8	0.7	1.0125	0.004	0.297	374.5055619
9	0.7	2.8375	0.004	0.297	436.9745285
10	0.1	3.75	0.004	0.297	922.6879307
11	0.7	2.8375	0.004	0.297	1008.90762
12	1.9	1.0125	0.004	0.297	678.6264889
13	0.1	3.75	0.004	0.297	413.6182137
14	0.1	0.1	0.004	0.297	4.871580398
15	0.1	0.1	0.004	0.297	0.25569757
16	2.5	1.925	0.004	0.297	708.2690884
17	1.9	2.8375	0.004	0.297	83.36907215
18	2.5	0.1	0.004	0.297	410.6805861
19	1.3	2.8375	0.004	0.297	1395.134038
20	0.1	3.75	0.004	0.297	759.980445
21	0.1	0.1	0.004	0.297	4.175337844
22	1.9	1.925	0.004	0.297	829.3965231
23	0.1	0.1	0.004	0.297	0.293767254
24	2.5	3.75	0.004	0.297	127.710933
25	2.5	0.1	0.004	0.297	306.3594785
26	1.3	3.75	0.004	0.297	426.2240551
27	0.1	3.75	0.004	0.297	229.9717074
28	0.7	1.0125	0.004	0.297	289.6067478
29	2.5	3.75	0.004	0.297	4.151178205
30	2.5	3.75	0.004	0.297	243.1858845
31	1.06	1.56	0.004	0.297	555.6468172
32	1.06	1.56	0.004	0.297	1248.779151
33	0.7	1.0125	0.004	0.297	1022.233687
34	1.9	2.8375	0.004	0.297	575.8948131
35	0.1	3.75	0.004	0.297	379.4746197
36	2.5	0.1	0.004	0.297	288.251663
37	2.5	3.75	0.004	0.297	5.262414031
38	2.5	0.1	0.004	0.297	338.7270819
39	2.5	0.1	0.004	0.297	458.8031935

Additional File 7

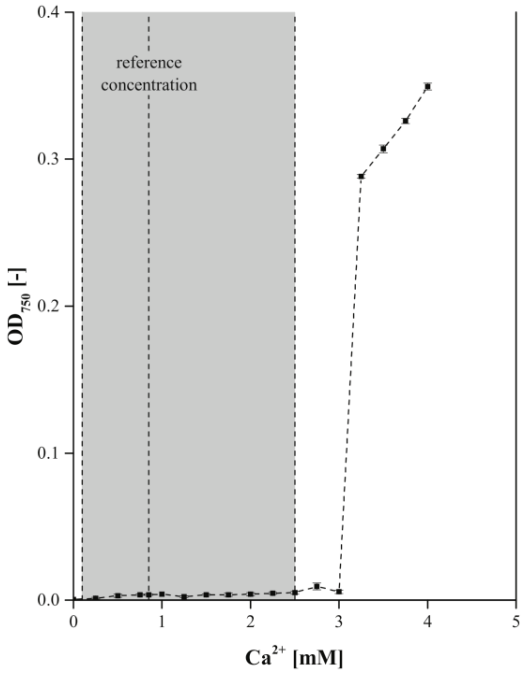


Figure 3: Precipitation of Ca2+

Additional File 8

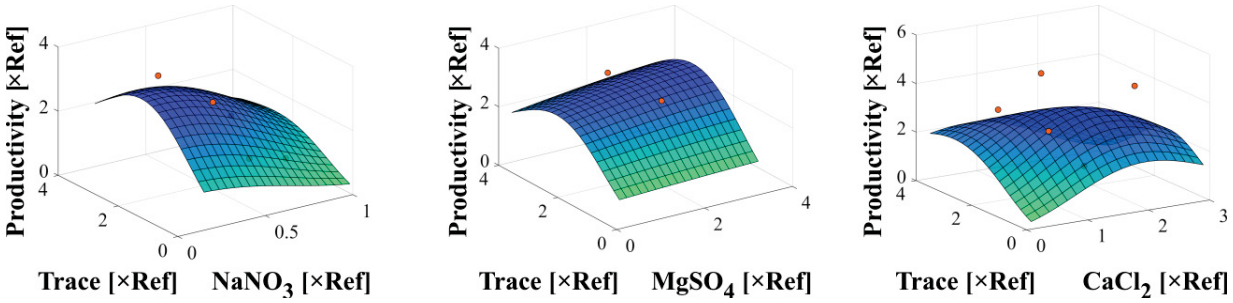


Figure 4: A three-dimensional plot of the Kriging model together with the measured data can be obtained. Analog to contour plots in Fig. 3b in original publication

Additional File 9

Table 8: Experimental design and results, refining the of optimum – Part 1

#experiment	MES [mmol L-1]	NaNO3 [g L-1]	K2HPO4 / KH2PO4 [g L-1]	NaCl [mmol L-1]	MgSO4 [mmol L-1]
1	50	1.5	2	0.107	0.761
2	50	1.5	2	0.107	0.761
3	50	1.5	2	0.107	0.761
4	50	1.5	2	0.107	0.761
5	50	0.375	2	0.107	0.7494
6	50	0.625	2	0.107	0.6245
7	50	0.75	2	0.107	2.3731
8	50	0.5	2	0.107	0.4996
9	50	1.125	2	0.107	2.1233
10	50	0.375	2	0.107	1.249
11	50	0.625	2	0.107	2.498
12	50	1.125	2	0.107	1.249
13	50	0.375	2	0.107	1.249
14	50	0.25	2	0.107	1.1241
15	50	0.5	2	0.107	2.2482
16	50	1.5	2	0.107	0.8743
17	50	0.625	2	0.107	0.7494
18	50	1.375	2	0.107	0.6245
19	50	1.125	2	0.107	1.8735
20	50	1.375	2	0.107	0.4996
21	50	0.375	2	0.107	1.7486
22	50	0.375	2	0.107	1.1241
23	50	0.75	2	0.107	0.7494
24	50	0.375	2	0.107	2.1233
25	50	0.5	2	0.107	0.2498
26	50	0.5	2	0.107	1.1241
27	50	0.75	2	0.107	1.6237
28	50	0.25	2	0.107	1.4988
29	50	1.125	2	0.107	2.2482
30	50	0.625	2	0.107	1.4988
31	50	0.625	2	0.107	2.498
32	50	0.875	2	0.107	1.9984
33	50	1	2	0.107	0.4996
34	50	0.25	2	0.107	2.498
35	50	0.5	2	0.107	0.3747
36	50	0.375	2	0.107	2.1233
37	50	0.5	2	0.107	0.6245
38	50	1	2	0.107	0.9992
39	50	1.125	2	0.107	1.8735

Table 9: Experimental design and results, refining the of optimum – Part 2

number of experiment	CaCl2 [mmol L-1]	trace [x fold]	FeSO4 [mmol L-1]	EDTA [mmol L-1]	Lipid Productivity [a.u.]
1	0.85	1	0.004	0.297	322.1342808
2	0.85	1	0.004	0.297	408.70142
3	0.85	1	0.004	0.297	338.9989834
4	0.85	1	0.004	0.297	377.4232757
5	0.2499	3.375	0.004	0.297	1319.709365
6	2.12415	1.875	0.004	0.297	1281.414526
7	1.12455	1.625	0.004	0.297	739.1991486
8	1.87425	3.5	0.004	0.297	792.794156
9	0.7497	1.875	0.004	0.297	1324.696489
10	1.62435	1.875	0.004	0.297	968.6335413
11	1.62435	1.25	0.004	0.297	649.3189801
12	1.37445	1.625	0.004	0.297	796.9939824
13	0.9996	1.875	0.004	0.297	881.7006164
14	1.62435	1.5	0.004	0.297	294.8120007
15	1.4994	1.625	0.004	0.297	728.5409248
16	1.9992	0.875	0.004	0.297	509.3817877
17	1.37445	2	0.004	0.297	651.2138754
18	1.2495	3.125	0.004	0.297	604.4186875
19	0.4998	2	0.004	0.297	1140.265032
20	0.37485	2.375	0.004	0.297	1153.231708
21	1.4994	2.375	0.004	0.297	699.8037576
22	1.7493	1.875	0.004	0.297	628.0610996
23	0.87465	2.375	0.004	0.297	1174.484775
24	1.37445	1.625	0.004	0.297	1154.945702
25	1.7493	3.625	0.004	0.297	706.8882145
26	2.2491	1.25	0.004	0.297	804.281105
27	0.9996	0.75	0.004	0.297	468.5243295
28	0.7497	3.375	0.004	0.297	1069.491537
29	2.12415	3	0.004	0.297	755.2045658
30	1.12455	1.625	0.004	0.297	696.8222886
31	1.7493	0.375	0.004	0.297	1335.275897
32	0.4998	0.875	0.004	0.297	972.5113385
33	0.9996	3.625	0.004	0.297	694.3640094
34	1.2495	1.75	0.004	0.297	1218.536374
35	2.12415	2.875	0.004	0.297	928.0238298
36	1.12455	1.25	0.004	0.297	809.5954026
37	0.12495	2.125	0.004	0.297	709.5202223
38	0.2499	1.75	0.004	0.297	967.9693233
39	2.12415	1.375	0.004	0.297	814.265332

Additional File 10

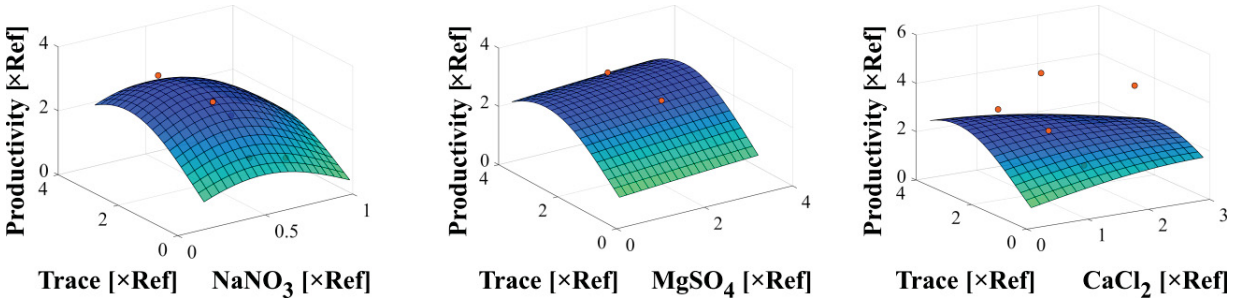


Figure 5: A three-dimensional plot of the Kriging model together with the measured data can be obtained. Analog to contour plots in Fig. 4 in original publication

Additional File 11

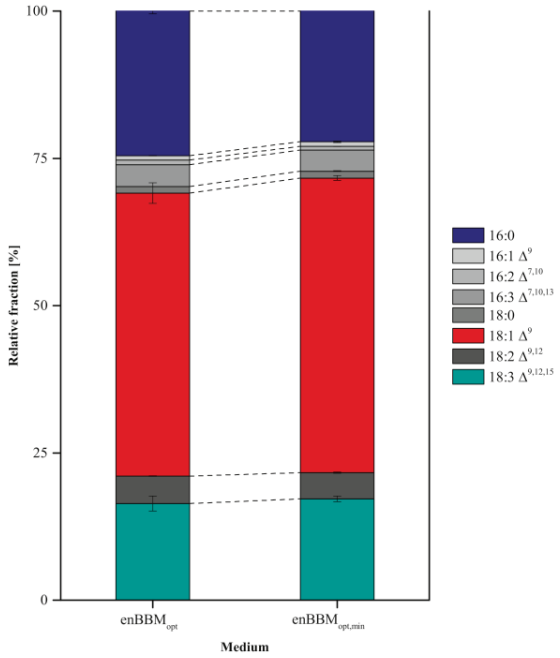


Figure 6: Relative composition of the fatty acids from the neutral lipid product fractions for enBBM$_{opt,min}$ and enBBM$_{opt}$. Error bars represent min/max from biological replicates (n=2)

Supplement to "Multi-objective global optimization (MOGO): Algorithm and case study in gradient elution chromatography"

Supplement

1 Model Parameters

Table 1: Model parameters for the different components

Parameter	**Component**	**Value**
Desorption rate $\left[\frac{\left(m^3_{Mobil}\right)^{\nu}}{mol^{-\nu}s^{-1}}\right]$	Lysozyme	1000
	Cytochrome C	1000
	Ribonuclease A	1000
Adsorption rate $\left[\frac{\left(m^3_{Mobil}\right)^{(\nu-1)}}{mol^{-\nu}s^{-1}}\right]$	Lysozyme	35.5
	Cytochrome C	1.59
	Ribonuclease A	7.7
Characteristic Charge ν $[-]$	Lysozyme	4.7
	Cytochrome C	5.29
	Ribonuclease A	3.7
Shielding Factors $[-]$	Lysozyme	11.83
	Cytochrome C	10.6
	Ribonuclease A	10
Film diffusion $\left[\frac{m}{s}\right]$	Lysozyme	6.9×10^{-6}
	Cytochrome C	6.9×10^{-6}
	Ribonuclease A	6.9×10^{-6}
Diffusion coefficient	All components	6.07×10^{-6}
Film diffusion $\left[\frac{m}{s}\right]$	All components	6.9×10^{-6}

Table 2: Column parameters

Parameter	Value
Axial Dispersion Coefficient (Column) $\left[\frac{m^2}{s}\right]$	5.75×10^{-8}
Interstitial Velocity $\left[\frac{m}{s}\right]$	5.75×10^{-4}
Column length $[m]$	0.014
Particle Radius $[m]$	4.5×10^{-5}
Column Porosity	0.37
Particle Porosity	0.75
Ionic capacity $\left[\frac{mol}{m_{Solid}^3}\right]$	1200
Initial bound salt concentration $\left[\frac{mol}{m_{Solid}^3}\right]$	1200

2 Pareto Density

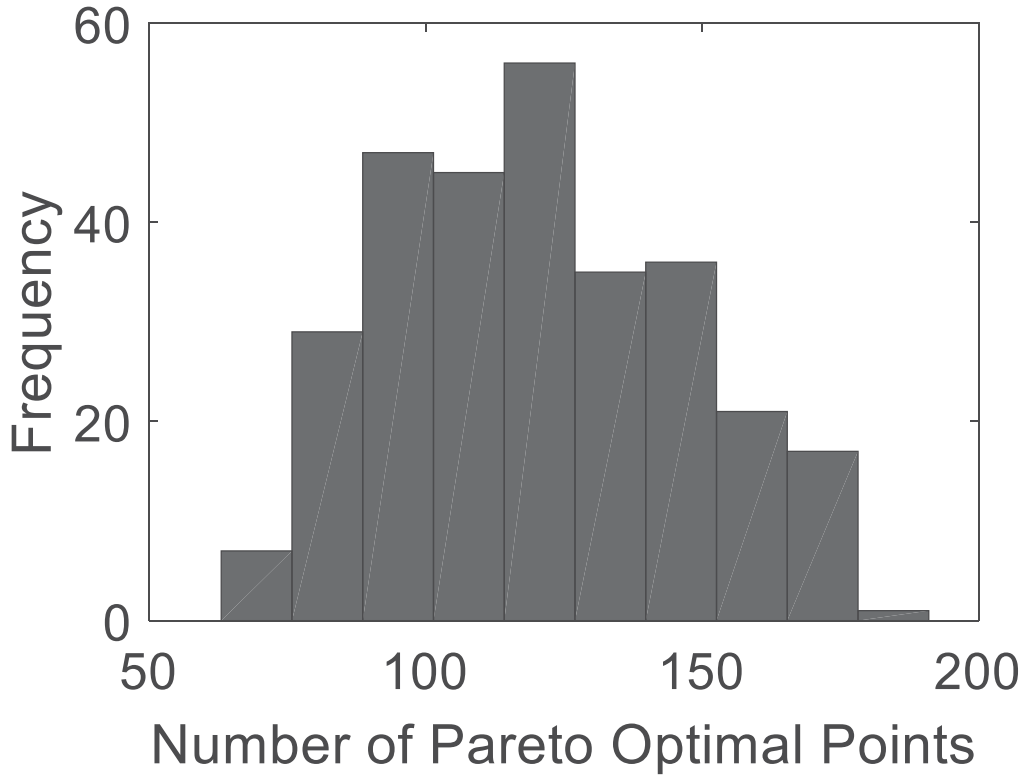


Figure 1: Distribution of amount of Pareto optimal points for 260 optimization runs applying the cooling down strategy. For further details see section 4.4 (main document).