Tuning Evolutionary Multiobjective Optimization for Closed-Loop Estimation of Chromatographic Operating Conditions

Richard Allmendinger, Spyridon Gerontas,
Nigel J. Titchener-Hooker, and Suzanne S. Farid

Department of Biochemical Engineering, University College London,
Torrington Place, London WC1E 7JE, UK
{r.allmendinger,s.gerontas,nigelth,s.farid}@ucl.ac.uk

Abstract. Purification is an essential step in the production of biopharmaceuticals. Resources are usually limited during development to make a full assessment of operating conditions for a given purification process commonly consisting of two or more chromatographic steps. This study proposes the optimization of all operating conditions simultaneously using an evolutionary multiobjective optimization algorithm (EMOA). After formulating the closed-loop optimization problem, which is subject to constraints and resourcing issues, four state-of-the-art EMOAs — NSGAII, MOEA/D, SMS-EMOA, and ParEGO — were tuned and evaluated on test problems created from real-world data available in the literature. The simulation results revealed that the performance of an EMOA depends on the setting of the population size, and constraint and resourcing issue-handling strategies adopted. Tuning these algorithm parameters revealed that the EMOAs, in particular SMS-EMOA and ParEGO, are able to discover reliably within 100 evaluations operating conditions that lead to high levels of yield and product purity.

1 Introduction

Manufacturing costs of therapeutic proteins are driven by costs associated with the purification of a protein of interest from impurities, such as host cell proteins and DNA, arising during the fermentation and harvest process, and by the need to achieve strictly controlled levels of key impurities. Chromatography is a commonly-used technique for purifying proteins and has been identified as a key cost driver [1]. The overall goal of this study is to optimize the operating conditions of a chromatography platform so as to improve multiple criteria, such as recovery yield and final product purity, contributing to a reduction in manufacturing costs.

Approaches for optimizing a chromatographic process can be classified broadly into two classes: *experimentally-validated simulation approaches* or *direct experimental approaches* [2]. The former includes approaches that describe a chromatographic process using a (predictive) linear or non-linear model based on mass transfer and thermodynamics [3]. Although simulation-based, this approach relies on physical experiments being performed to calibrate and validate the model. Various optimization methods have been used to estimate the parameters of a chromatographic model (see e.g. [3,4]).

T. Bartz-Beielstein et al. (Eds.): PPSN XIII 2014, LNCS 8672, pp. 741–750, 2014.

If the process modeled is not well-understood or cost prohibitive to define in terms of a simulation model, then a *direct experimental optimization approach* can be adopted. Such approaches optimize a chromatographic process by performing physical experiments guided, for example, by *design of experiments* (DoE) in combination with a response surface analysis [5,6] or an evolutionary algorithm (EA) [2,7]. An experimental optimization approach might incur high experimental costs, whilst a simulation-based approach relies heavily on the computational resources available.

Recently, multiobjective methods have found application in chromatography processes optimization. For example, in [8] an evolutionary multiobjective optimization algorithm (EMOA) was used within a simulation-based approach to optimize purity, productivity, and/or yield of a single chromatography step. EMOAs have found application in various experimental optimization problems [9], and are easily adaptable to problems featuring constrained, non-linear, non-convex, noisy, dynamically changing, and/or multiple objective functions. EMOAs have also been extended to cope with resourcing issues in experimental optimization leading e.g. to delayed/missing objective values [10] and temporary non-availability of certain solutions for evaluation [9]. The issue around missing objective values can also be encountered in chromatography process optimization and is investigated in more detail in this study.

Although a chromatographic purification process consists of multiple steps, the work cited above focuses on the optimization of a single step only. The goal of this study is to optimize multiple chromatography steps simultaneously so as to maximize recovery yield and final product purity (or equivalently minimize impurities). Optimizing multiple steps means that ideally interactions, technical limitations and/or resourcing issues between steps can be accounted for in the optimization. The lack of models capable of capturing interactions and constraints between multiple chromatography steps accurately, means however that a direct experimental approach needs to be adopted. To realize this experimental optimization platform, a sophisticated and precise laboratory setup is required as well as an optimization method capable of dealing efficiently with the enlarged search space and additional constraints (arising due to the presence of multiple chromatography steps). This study focuses on the design of an effective optimizer to guide the selection of conditions for physical experiments.

2 Problem Definition

This section describes the problem formulation for the multiobjective optimization of chromatographic operating conditions (MOCOC) subject to resourcing issues. The experimental platform adopted for the optimization of operating conditions across multiple chromatography steps is visualized in Figure 1, and can be formulated mathematically as follows:

$$\begin{aligned} &\text{maximize} \quad \mathbf{f}(\mathbf{x},\sigma) = (f_1(\mathbf{x},\sigma),...,f_m(\mathbf{x},\sigma)) \\ &\text{subject to} \quad \mathbf{x} \in X, \end{aligned}$$

where $\mathbf{x} = (x_1,...,x_l)$ is a *solution vector* (here a set of operating conditions), and X a *feasible search space* (here the set of all possible operating conditions). The *objective*

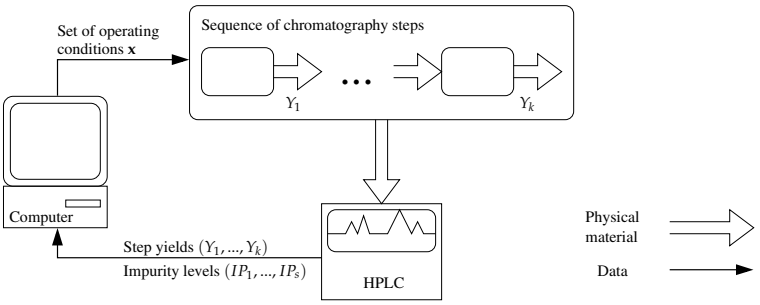

Fig. 1. Schematic of a typical experimental setup for the closed-loop optimization of chromatographic operating conditions. Following the set up of the operating conditions, defined by **x**, the sample is passed through a sequence of chromatography steps $i = 1, ..., k$. An HPLC device is used to obtain the step yields Y_i and the final levels of individual impurities $IP_j, j = 1, ..., s$. Based on this quality measure, an optimizer running on the computer then selects the next set of operating conditions for testing.

vector function f is a *black box* and represents a time-consuming and costly physical experiment on **x**, which is characterized by $m > 0$ noisy measurements $f_1, ..., f_m$. The functions f_i are known as *objectives* and are typically in conflict. The vector σ represents environmental factors that cannot be controlled, e.g. imprecision in the experimental equipment. In the following, these problem features are described in more detail.

Decision variables $x_1, ..., x_l$**:** A solution vector **x** represents a set of relevant operating conditions, such as pH and salt concentration, for a set of chromatography steps $i = 1, ..., k$. Figure 2 shows the solution encoding used in this work: each step i is associated with a pre-defined resin_i and a variable number of operating conditions $c_{i,j}, 1 \leq j \leq d_i$, resulting in $l = \sum_{i=1}^{k} d_i$ decision variables in total. Typically, the values $c_{i,j}$ are represented by discretized real values.

Objective functions $f_1, ..., f_m$**:** Two ($m = 2$) commonly-used metrics were considered in order to characterize the quality of a chromatographic process: the overall recovery yield Y and the final product impurities $\sum IP_j$:

$$\text{maximize } f_1 = Y = Y_1 \times ... \times Y_k$$

$$\text{minimize } f_2 = \sum IP_j = \sum_{j=1}^{s} IP_j,$$

where Y_i is the yield of chromatography step i, and $IP_j, j = 1, ..., s$, the levels of different impurity types, such as host cell proteins and DNA; note, the objective of minimizing product impurities is equivalent to maximizing purity and used here due to the structure of the test problems considered (see Section 3.1). Both objectives, yield and (im)purity, are obtained by analyzing the sample using an HPLC (high-performance liquid chromatography) device. Measuring the yield and levels of different impurities takes around 4 min and 30 min per sample, respectively, whilst the robot takes around 1

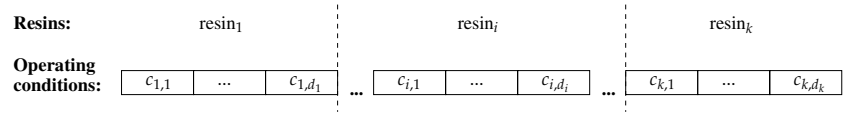

Fig. 2. Representation of a solution $\mathbf{x}$ for a process with $i = 1, ..., k$ chromatography steps. Each step i is linked to a fixed resin$_i$ and a set of operating conditions $c_{i,j}, 1 \leq j \leq d_i$.

hour per chromatography step to prepare a sample (this robotic step can be parallelized to up to 8 samples).

Feasible Search Space X**:** In addition to standard constraints on the decision variable value ranges, the search space may be defined by *dependency constraints* between chromatography steps. For instance, setting the salt concentration of a step i dictates the lowest possible salt concentration of the successive step, or $c_{i,j} \leq c_{i+1,t}$, assuming indices j and t point to the salt concentration at chromatography step i and $i+1$, respectively.

In addition to constraints defining the search space X, there may also be constraints on the objective values $f_1, ..., f_m$. For example, for antibodies, regulatory requirements specify that the final product impurities needs to be $\sum IP_j < 5\%$. Moreoever, in the presence of limited resources, there may be a threshold $Y_{\min}$ on the minimum recovery yield. This limitation can be seen as a *resourcing issue* and shall prevent the waste of resources dedicated to the evaluation of inefficient purification processes. Mathematically, this resourcing issue can be expressed by a Boolean clause as follows

$$\textbf{if } Y_i \times ... \times Y_p < Y_{\min} \textbf{ then } \textit{terminate experiment and return } Y_1, ..., Y_p \qquad (1)$$

where $p \leq k$ denotes the chromatography step after which the cumulative yield is below the threshold $Y_{\min}$. That is, if the resourcing issue is 'activated', then the objectives Y and $\sum IP_j$ are missing. However, the measurements $Y_1, ..., Y_p$ are available and might be used to estimate Y and/or $\sum IP_j$.

Uncertainties: The decision variables $x_1, ..., x_l$ and the measurements f_1 and f_2 might be subject to some level of uncertainty (noise) given the experimental nature of the problem. This level is typically low if the experimental platform is set up accurately, and thus neglected here.

3 Experimental Setup

This section describes the case study, extensions augmented on the EMOAs for coping with the challenges of the MOCOC problem, and algorithm parameter settings as used in the subsequent experimental analysis.

3.1 Case Study

Ultimately, the goal is to tackle MOCOC problems with $k \approx 3$ chromatography steps and $l \approx 8$ operating conditions in total. The purification sequence considered in [6] falls into this problem domain and was used in this study as the "test problem" to tune and validate different state-of-the-art EMOAs (using computational experiments).

Table 1. MOCOC problem characteristics

Chromat. Step i	*Operating condition*	*Value range*	*Step size δ*
$i = 1$, affinity	pH_{wash}	[4.5; 5.5]	0.1
	$pH_{elution}$	[2.5; 3.5]	0.1
	$NaCl_{wash}$	[50; 500]	25
$i = 2$, cation exchange	$pH_{load,2}$	[4.5; 5.5]	0.1
	Grad. length	[10; 20]	2
	Load	[45;55]	1
$i = 3$, anion exchange	$pH_{load,3}$	[7; 8]	0.1
	Load	[100; 200]	5

Table 2. EMOA default parameter settings

EMOA	*Parameter*	*Setting*
All	Max evaluations G	100
	Crossover probability p_c	0.6
	Per-variable mutation probability p_m	$1/l$
NSGAII	Population size n	10
MOEA/D	Population size n	20
	#Weight vectors T	20
SMS-EMOA	Population size μ	4
ParEGO	Initial population size n	50
	#Scalar vectors s	10

Table 1 lists the operating conditions to be optimized for each of the $k = 3$ steps (affinity, cation and anion exchange); the operating condition values were discretized as specified by the step size δ. The heatmap data published in [6] was used to construct two interpolated fitness landscapes for each of the $k = 3$ steps, one for the step yield Y_i and one for the step's impurity level IP_i, using the Kriging approach.[1] The experimental study considered the problem with $k = 2$ steps (the first two steps) and $l = 6$ operating conditions, and the complete problem with all $k = 3$ steps and $l = 8$ operating conditions.

The independent optimization of each chromatography step has been studied before [6]. In our work the problem was extended by dependency constraints and resourcing issues (mimicking real limitations of the problems of interest): the dependency constraint was defined by $pH_{wash} \leq pH_{load,2}$ (for the sake of this constraint, the value ranges of both variables were set identically), and the resourcing issue was represented by Equation (1).

3.2 Tuning Evolutionary Search for the MOCOC Problem

To run an EMOA on the MOCOC problem, strategies for coping with the constraints and resourcing issues need to be defined.

Handling Dependency Constraints: Four strategies — *random, copy, swap,* and *regenerate* — were investigated for coping with the dependency constraint defined above. Upon encountering an infeasible solution, the strategy *random* sets its $pH_{load,2}$ value to a random value selected from the range [pH_{wash}, 5.5], whilst the strategy *copy* sets $pH_{load,2} = pH_{wash}$. The strategy, *swap*, swaps the values of $pH_{load,2}$ and pH_{wash}, which results in a feasible solution due to the identical value range of the two variables. Finally, the strategy *regenerate* iteratively generates new solutions until it generates one that is feasible.

Handling Resourcing Issues: Three strategies — *strict penalizing, relaxed penalizing,* and *fitness-inheritance* — were investigated for coping with the resourcing issue defined in Equation (1). The aim of these strategies is to substitute missing objective

[1] A Kriging function, Krig(), was used from the *fields* package of the statistical software R.

values with some surrogate. The strategy, *strict penalizing*, sets the objectives of a solution violating Equation (1) to the worst possible values; i.e. $f_1 = 0 = 1$ (assuming a normalized objective space). The strategy, *relaxed penalizing*, uses the available yield measurements to set the objective values to $f_1 = Y_i \times ... \times Y_p$ and $f_2 = 0$. Finally, the strategy, *fitness-inheritance*, selects for a solution with missing objectives a solution from the set of all solutions evaluated so far that is both closest to it in the decision space (in terms of normalized Euclidean distance) and has no missing objectives, and then simply copies the solution's values of f_2 and $Y_{p+1}, ..., Y_k$ to allow the computation of f_1.

3.3 Algorithm Parameter Settings

Four state-of-the-art EMOAs were considered in the experimental analysis: NSGAII [11], MOEA/D [12], SMS-EMOA [13], and ParEGO [14]. All EMOAs avoided the evaluation of duplicate solutions (solutions were regenerated until a unique one is created), each used a latin hypercube initialization procedure, uniform crossover, binary tournament selection (with replacement), and a mutation operator that selects a value at random from the feasible variable value range (see Table 1). Both NSGAII and MOEA/D employ a generational reproduction scheme (using a fixed population size of μ), whilst SMS-EMOA uses a steady-state scheme, and ParEGO considers all solutions evaluated to create a single solution.

The aim of the experimental study was to understand how the performance of the EMOAs is affected by different algorithm parameter settings, and constraint and resourcing issue-handling strategies. The default settings of the EMOAs are given in Table 2. The total number of evaluations $G = 100$ represents the estimated budget available for the problems of interest. Results shown are the average across 30 independent runs. Hypervolume and attainment surface results were obtained by considering all solutions found during a run that had no missing objective values. For the hypervolume calculation, the objective values were normalized to lie in the range [0,1], and the reference point was set to a value of 2 for all objectives.

4 Experimental Analysis

The first two sets of experiments investigate the performance of the constraint-handling strategies and sensitivity of algorithm parameter settings in the absence of the resourcing issue, which is the focus of the last set of experiments.

Investigation of Constraint-Handling Strategies: Figure 3 shows the performance of the different constraint-handling strategies when augmented on NSGAII. From Figure 3(a) it is apparent there is a trade-off between the population size n and the number of generations G/n available for optimization: the performance increases until a population of $n \approx 10$ after which any further increases in n lead to a performance reduction (due to the smaller number of generations available). The *random* strategy performs most robustly, followed closely by the *copy* and *regenerate* strategy. The *swap* strategy performs significantly worse than the other strategies, in particular around the sweet spot of $n \approx 10$, as it deteriorates the original solution's string most. The attainment

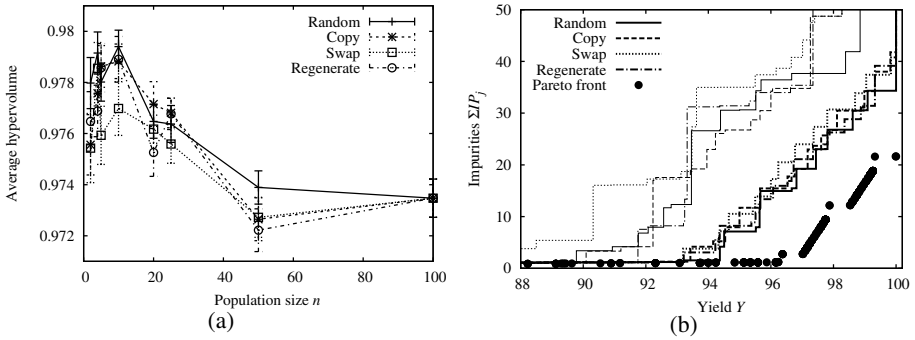

Fig. 3. a)Average hypervolume and its standard error as a function of the population size n, and b) worst (thin lines) and median (bold lines) attainment surfaces for $n = 10$ obtained by different constraint-handling strategies augmented on NSGAII for a MOCOC problem with $k = 2$ steps and $l = 6$ variables. For every setting marked by a point in a), a Kruskall-Wallis test (significance level of 5%) has been carried out. *Random*, *copy*, and *regenerate* perform best for $n = 10$. There is no clear winner for the other settings.

surface plot, Figure 3(b), confirms this ranking as well as the significant gap of the *random* strategy to the estimated true Pareto front (which has been obtained by taking the non-dominated front discovered across multiple and long runs of NSGAII). The performance patterns were similar for both test problems, i.e. $k = 2$ and 3 steps, and the other EMOAs. For ParEGO, the choice of the constraint-handling strategy is not as crucial as the search for a new solution is performed over an interpolated landscape (instead of the actual search space), which is cheap to evaluate.

Investigation of Crucial Algorithm Parameter Settings: Figure 4 analyses the sensitivity in performance of the different EMOAs as a function of algorithm parameter settings, in particular the population size n. From Figure 4(a), it is obvious that the performance of all EMOAs improves until a certain n is reached, and then degrades for further increases in n. SMS-EMOA is able to achieve the highest average hypervolume, when used in combination with small population size of $n \approx 4$. ParEGO performed slightly worse than SMS-EMOA in terms of the average hypervolume but is more robust to variations in n. Note, for $n = 100$, no optimization was performed as all EMOAs sample the search space using a latin hypercube design, which can be seen as the default performance obtained with a DoE approach. As can be seen from Figure 4(a), this performance is clearly beaten by the EMOAs for most settings of n. The performance ranking of EMOAs with respect to worst and median attainment surfaces obtained, which are shown in Figure 4(b), was in alignment with the hypervolume results. It is also apparent from the plot that SMS-EMOA and ParEGO were able to get significantly closer to the Pareto front than NSGAII.

Investigation of Resourcing Issue-Handling Strategies: Finally, Figure 5 analyzes the performance of the resourcing issue-handling strategies when augmented on SMS-EMOA. In Figures 5(a) and 5(b), the resourcing issue was present from the very beginning of the optimization, whilst, in Figures 5(c) and 5(d), it was ignored and the evaluation completed for the first 10 solutions with a cumulative yield below $Y_{\min}$.

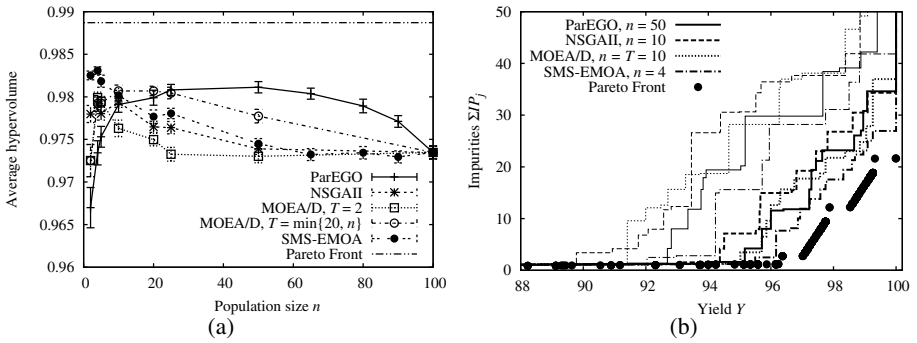

Fig. 4. a) Average hypervolume and its standard error as a function of the population size n, and b) worst (thin lines) and median (bold lines) attainment surfaces obtained for optimal settings of n by several EMOAs for a MOCOC problem with $k = 2$ steps and $l = 6$ variables. For every setting marked by a point in a), a Kruskall-Wallis test (significance level of 5%) has been carried out. SMS-EMOA performs best for $n < 10$, and ParEGO for $25 < n < 100$. There is no clear winner for the other settings.

From Figures 5(a) and 5(c) it can be observed that the presence of the resourcing issue has a significant negative impact on performance for $Y_{\min} > 90\%$. Relaxing the resourcing issue reduces the impact on performance but it is an expensive approach. Preventing the optimizer from entering certain regions of the objective space introduces a search bias towards other parts of the Pareto front, as evident from the attainment surfaces shown in Figures 5(b) and 5(d), especially in Figure 5(b), for $Y_{\min} = 94\%$. Comparing the different resource issue-handling strategies, it is apparent that a penalizing strategy performs best in Figures 5(a) and 5(b). A fitness-inheritance strategy performs better in the relaxed scenario (Figures 5(c) and 5(d)) because, once the resourcing issue is switched on, it allows the optimizer to enter more quickly a feasible region in the objective space than the penalizing strategies (as indicated by the number of experiments below $Y_{\min}$). Note, although relaxing the resourcing issue leads to more distributed attainment surfaces (see Figure 5(d)), the surfaces are further away from the Pareto front than in the unrelaxed case due to the lower level of exploitation. The other EMOAs are affected in a similar way by the resourcing issue.

5 Conclusion and Future Work

This paper has considered a real-world problem concerned with the optimization of operating conditions for chromatographic purification processes so as to maximize recovery yield and product purity. The problem has been formulated as a multiobjective closed-loop optimization problem subject to dependency constraints, resourcing issues, uncertainties, and a limited number of evaluations. Several strategies were proposed for dealing with the constraints and resourcing issues, and subsequently augmented and validated on four state-of-the-art EMOAs — ParEGO, NSGAII, MOEA/D, and SMS-EMOA — for two test problems created from published real-world data. The experimental study revealed that EMOAs can achieve a better performance within 100

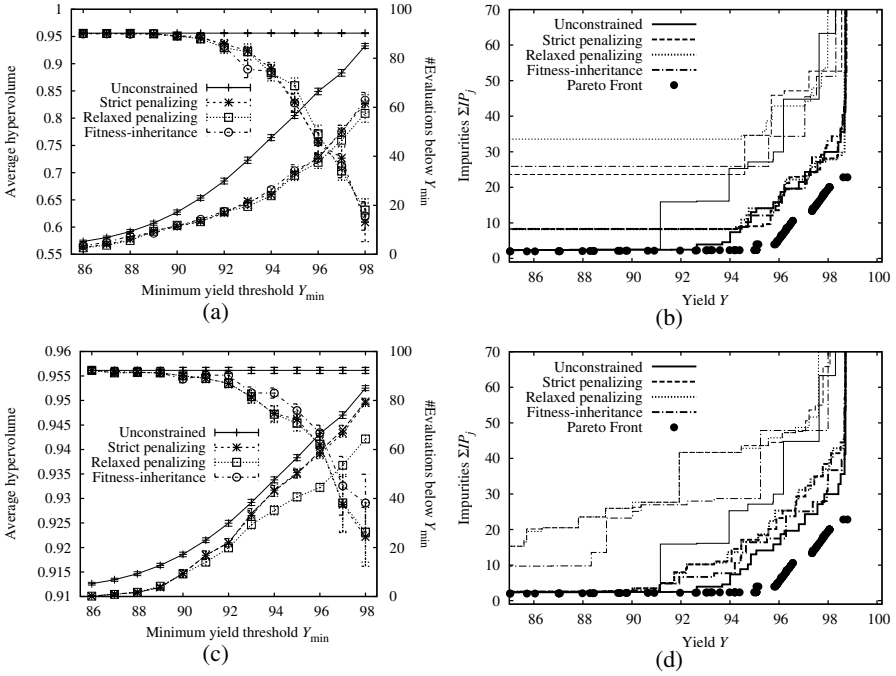

Fig. 5. a) and c) Average hypervolume, its standard error, and #evaluations below $Y_{\min}$ as a function of the threshold $Y_{\min}$, and b) and d) worst (thin lines) and median (bold lines) attainment surfaces obtained for $Y_{\min} = 94\%$ by SMS-EMOA for a MOCOC problem with $k = 3$ steps and $l = 8$ variables. In a) and b), the resourcing issue was present throughout the search, whilst, in c) and d), it was ignored for the first 10 solutions with $f_1 < Y_{\min}$. For every setting marked by a point in a) and c), a Kruskall-Wallis test (significance level of 5%) has been carried out. Relaxed penalizing performs best in a) for $Y_{\min} = 95\%$, whilst fitness-inheritance performs best in c) for $Y_{\min} = 92\%$ and 94%. There is no clear winner for the other settings.

evaluations than a standard DoE approach, such as a latin hypercube design. The best performance was achieved by SMS-EMOA when used in combination with a small population of size $n \approx 4$ and a random sampling-based constraint-handling strategy. The performance of an EMOA depended on the resourcing issue-handling strategy: A penalizing strategy performed best if a resourcing issue is present throughout the search, whilst a fitness-inheritance approach performs better if the resourcing issue is relaxed. Future research will focus on applying SMS-EMOA to guide real physical chromatographic experiments.

References

1. Pollock, J., Bolton, G., Coffman, J., Ho, S.V., Bracewell, D.G., Farid, S.S.: Optimising the design and operation of semi-continuous affinity chromatography for clinical and commercial manufacture. Journal of Chromatography A 1284, 17–27 (2013)

2. Susanto, A., Treier, K., Knieps-Grünhagen, E., von Lieres, E., Hubbuch, J.: High throughput screening for the design and optimization of chromatographic processes: automated optimization of chromatographic phase systems. Chemical Engineering & Technology 32(1), 140–154 (2009)
3. Guiochon, G.: Preparative liquid chromatography. Journal of Chromatography A 965(1), 129–161 (2002)
4. Irizar Mesa, M., Llanes-Santiago, O., Herrera Fernández, F., Curbelo Rodríguez, C., Da Silva Neto, A.J., Câmara, L.D.T.: An approach to parameters estimation of a chromatography model using a clustering genetic algorithm based inverse model. Soft Computing 15(5), 963–973 (2011)
5. Ferreira, S.L.C., Bruns, R.E., Ferreira, H.S., Matos, G.D., David, J.M., Brandao, G.C., da Silva, E.G.P., Portugal, L.A., dos Reis, P.S., Souza, A.S., dos Santos, W.N.L.: Box-behnken design: An alternative for the optimization of analytical methods. Analytica Chimica Acta 597(2), 179–186 (2007)
6. GE Healthcare Life Sciences. A platform approach for the purification of antibody fragments (fabs). Application note 29-0320-66 AA (2012)
7. Treier, K., Berg, A., Diederich, P., Lang, K., Osberghaus, A., Dismer, F., Hubbuch, J.: Examination of a genetic algorithm for the application in high-throughput downstream process development. Biotechnology Journal 7, 1203–1215 (2012)
8. Nfor, B.K., Zuluaga, D.S., Verheijen, P.J.T., Verhaert, P.D.E.M., van der Wielen, L.A.M., Ottens, M.: Model-based rational strategy for chromatographic resin selection. Biotechnology Progress 27(6), 1629–1643 (2001)
9. Allmendinger, R., Knowles, J.: On handling ephemeral resource constraints in evolutionary search. Evolutionary Computation 21(3), 497–531 (2013)
10. Allmendinger, R., Knowles, J.: 'Hang On a Minute': Investigations on the Effects of Delayed Objective Functions in Multiobjective Optimization. In: Purshouse, R.C., Fleming, P.J., Fonseca, C.M., Greco, S., Shaw, J. (eds.) EMO 2013. LNCS, vol. 7811, pp. 6–20. Springer, Heidelberg (2013)
11. Deb, K., Agrawal, S., Pratap, A., Meyarivan, T.: A fast and elitist multiobjective genetic algorithm: NSGA-II. IEEE Transactions on Evolutionary Computation 6(2), 182–197 (2002)
12. Zhang, Q., Hui, L.: MOEA/D: a multiobjective evolutionary algorithm based on decomposition. IEEE Transactions on Evolutionary Computation 11(6), 712–731 (2007)
13. Beume, N., Naujoks, B., Emmerich, M.: SMS-EMOA: multiobjective selection based on dominated hypervolume. European Journal of Operational Research 181(3), 1653–1669 (2007)
14. Knowles, J.: ParEGO: A hybrid algorithm with on-line landscape approximation for expensive multiobjective optimization problems. IEEE Transactions on Evolutionary Computation 10(1), 50–66 (2006)