Computers and Chemical Engineering 169 (2023) 108083

Contents lists available at ScienceDirect

Computers and Chemical Engineering

journal homepage: www.elsevier.com/locate/compchemeng

Computer-aided design of formulated products: A bridge design of experiments for ingredient selection

Liwei Cao [a,b,1], Danilo Russo [a,1], Emily Matthews [c,1], Alexei Lapkin [a,b,*], David Woods [c,*]

[a] Department of Chemical Engineering and Biotechnology, University of Cambridge, Cambridge CB3 0AS, UK
[b] Cambridge Centre for Advanced Research and Education in Singapore (CARES Ltd), #05-05 CREATE Tower, 1 Create Way, Singapore 138602
[c] Southampton Statistical Sciences Research Institute, Department of Mathematical Sciences, University of Southampton, Southampton SO17 1BJ, UK

ARTICLE INFO

Keywords:
Design of Experiments (DoE)
Bayesian optimization
Product design
Gaussian processes
Machine Learning (ML)

ABSTRACT

Formulations are ubiquitous in many industries. As formulations are being modified and re-developed to include more renewable and recyclable ingredients, the speed of formulations development becomes important. This study expands on the previous work demonstrating successful application of multi-objective Bayesian optimization to design of formulations within a restricted set of the available ingredients. Here we develop an approach that resolves the un-solved to date problem in algorithmic formulations development, when a subset of ingredients should be chosen from a larger available pool of suitable ingredients. The new DoE algorithm was demonstrated in a workflow making use of a 'make and test' formulation robots. The developed new DoE procedure demonstrated an efficient selection of a subset of ingredients from a larger number of the available ones, optimizing their concentration and allowing assignment of differential priorities to the optimization objectives.

1. Introduction

Liquid formulated products are a type of liquid blends, tailored to meet specific customer-defined functionalities and properties, such as colour, fragrance, viscosity and functional performance, for example, cleaning performance in the case of domestic and personal care cleaning products (Uhlemann et al., 2020; Yunus et al., 2014). Liquid formulations are ubiquitous in many traditional applications, including pharmaceuticals, paints, food, cosmetics, detergents, pesticides and so on (Zhang et al., 2018). At the same time, there is an increasing demand for new formulations in which (*i*) components offer better functional performance at reduced concentrations, and (*ii*) some urgency in developing formulations with a significantly improved environmental profile (lower toxicity to environment, lower life cycle emissions, lower overall footprint, etc.) (Heintz et al., 2014; Ten et al., 2017).

Development of formulations is, however, a rather complex process. The characteristic feature of formulations is the relatively large number of components (typically called 'ingredients') in any single formulation: several functional components deliver the desired set of target performance features, typically as a result of synergistic emergent interactions; other (non-functional) ingredients are added to achieve the desired combination of properties, that could be fairly customizable for specific (e.g., regional or cultural) consumer preferences (Gani and Ng, 2015; Peremezhney et al., 2012). As a result, the development of new formulations or studies of substitution of ingredients in existing formulations are necessarily experimentally heavy and require expensive trial-and-error campaigns (Fung et al., 2016).

The most common approaches for systematic development of formulated products can be categorized into two types (Kontogeorgis et al., 2019; Zhang et al., 2020): (*i*) the conventional experimental trial-and-error approach (Wesselingh et al., 2007), which is time and resource demanding, and (*ii*) integrated approaches combining data and models, which can give inaccurate solutions, as first principles predictions of emergent properties of multi-component mixtures of functional molecules (polymers, surfactants, etc.) are not yet accurate enough for physical properties prediction (Conte et al., 2011). Moreover, in many cases, formulation design is based on the choice of a certain subset of m components from a large number n of available chemicals ($m < n$). One typical example is the choice of a certain number of surfactants, which are used as stabilizers of emulsions and also influencing their final properties (Kontogeorgis et al., 2019). The number of possible combinations is very large when many components

* Corresponding authors.
E-mail addresses: aal35@cam.ac.uk (A. Lapkin), D.Woods@soton.ac.uk (D. Woods).
[1] These authors have made equal contribution to the study.

https://doi.org/10.1016/j.compchemeng.2022.108083
Received 21 July 2021; Received in revised form 15 November 2022; Accepted 25 November 2022
Available online 28 November 2022

are available. Due to manufacturing constraints or regulation issues, binary and ternary mixtures are often used in practice (Li et al., 2015). As a result, finding suitable binary or ternary mixtures with desired properties from all possible binary and ternary combinations is challenging (Jouyban et al., 2011, 2006). For example, in a ternary mixture design ($m = 3$), if there are $n = 10$, 20, or 50 possible components to choose from, there are 120, 1140, and 1960 combinations, respectively. As n gets larger, finding an optimal design for each possible combination and comparing the obtained results becomes increasingly prohibitive. The situation is complicated by the fact that each component can be used at different concentrations, corresponding to different final properties, further expanding the search space.

Here we ask a question - *could we use an algorithmic design of experiments (DoE) approach, coupling statistical models with robotic experiments, to guide design of functional emulsions and to efficiently select optimal ingredients?*

To tackle the overall problem of complexity, cost and duration of formulations development, the first plausible solution is to make use of statistical methods that would maximize the amount of useful information derived from the available experimental data and could guide experimental programme in a Design of Experiments process. Recently, we have illustrated this approach with the high-throughput robotic systems guided by advanced machine-learning sampling algorithm, which led to a marked acceleration of the overall product development cycle (Cao et al., 2021, 2020).

In this study we extend previous work by introducing a so-called "bridge" design of experiments methodology (Jones et al., 2015), to enable the earlier developed Bayesian algorithms to select a sub-set of ingredients from the available pool. Specifically, this was done for selection of surfactants in a cleaning liquid formulation model, in the absence of available physical models. The implemented bridge design approach allowed a trade-off between choosing formulations to explore the overall available experimental space, and estimation of flexible non-parametric models, and optimal choice of combinations to estimate a known parametric statistical model. Such methodology is particularly relevant for experiments measuring multiple responses, with differing modelling approaches being adopted for each response.

The developed methodology was applied to a commercial formulation to simultaneously meet specific customer-defined binary and continuous targets: stability and viscosity. The aim was to demonstrate that the developed methodology can optimize two responses for a real detergent, allowing for the choice of a subset of ingredients from the available pool, while using a relatively small number of experiments, generated by robotic a robotic platform.

2. Materials and methods

2.1. Case study and materials

The case study under consideration in this work is a commercial formulation which contains a polymer (P1 = Dehyquart CC7), a thickener (T1 = Arlyon TT), and three different surfactants. To develop and demonstrate the methodology, five available surfactants were used: Dehyton PK 45 (S1), Dehyton AB 30 (S2), Plantacare 818 (S3), Plantacare 2000 (S4), and Texapon SB 3 (S5). pH was adjusted using citric acid (ACS reagent, $\geq$ 99.5%, Sigma-Aldrich), used as received.

The constraints of the input variables are given below:

1) the sum of the concentrations of S1 to S5, of which three at most can be non-zero (or active), must be in the range 13.00 – 15.00 g L^{-1}.
2) P1 concentration must be in the range 0.00 – 2.10 g L^{-1}
3) T1 concentration must be in the range 0.00 – 2.10 g L^{-1}.

Once the formulated product has been manufactured, it is tested for stability, which has a pass (1) or fail (0) outcome, and viscosity at a shear rate of 10 s^{-1}, which must be between 2.00 and 4.00 Pa·s.

2.2. Experimental set-up

The experimental samples were generated using the previously developed semi-automated robotic platform (Cao et al., 2021, 2020). Briefly, the platform consists of two stations: (*i*) for preparation and (*ii*) analysis of the prepared samples. The algorithmic procedure developed in this work (Section 3) generated a .csv file containing the experimental design to be tested. This triggered the first station, consisting of 8 syringe pumps separately feeding ingredients to the dispensing element. This was used to fill a batch of up to 24 sampling vials (V_{sample} = 10 mL) located on a rotating wheel. All surfactants were previously diluted in water to achieve a concentration of 20 g L^{-1} in the feeding bottles and their pH was adjusted to 5.5. The generated samples were transferred into an incubator (Corning LSE 71 L shaking incubator) and processed at 50 °C, 300 rpm for 2 h. The processed samples were cooled to ambient temperature and placed on the rotating wheel of the second station where the samples were automatically processed through a camera to distinguish between stable (homogeneous) and unstable (presenting phase separation) formulations. Automatic pH tests were carried out and no pH variations were observed in any sample after the processing. Finally, viscosity of the samples at a shear rate of 10 s^{-1} was measured off-line in a non-automated fashion using a rotational viscometer (ARES Rheometric Scientific, strain controlled, Couette configuration).

3. Theory

The workflow used in this work is shown in Fig. 1. The experimental workflow described in Section 2 dispenses the required amounts of the selected ingredients, processes the formulations, tests their stability, and measures their viscosity. The algorithmic workflow is detailed in the following subsections. Briefly, initial sampling was used to efficiently explore the input space and to design appropriate experiments to maximize the information gain. This was successively used to trigger an iterative search of the optima, i.e. stable sample with a viscosity as close as possible to the 3.00 Pa·s, based on the trained surrogate models for the responses.

3.1. Initial sampling

The initial 230 experiments were performed using a maximin space filling design (Johnson et al., 1990) with the aim of efficiently exploring the entire chemical design space. The entire initial sampling is reported in the Supplementary material, Table S1.

3.2. Design of experiments algorithm

3.2.1. Objective function

The objective function was inspired by the bridge design reported in Jones et al. (Jones et al., 1998), which has the dual objectives of optimality with respect to parameter estimation from a parametric model (D-optimality (Atkinson et al., 2007)), and space filling. However, the objective function in this work differs in three ways: 1) it does not use the same space filling criteria; 2) it is a weighted objective function; 3) it uses the Bayesian D-optimality objective function (Chaloner and Verdinelli, 1995; Ryan et al., 2016) which is focussed on estimation of a logistic regression model, approximated using Monte Carlo integration (Overstall and Woods, 2017). A bespoke objective function and algorithm have been written to find an optimal design for these experiments, where the objective function considers two different types of outputs: a binary response from the stability test, and a continuous response from the viscosity test.

For the binary output, there exists a variety of models suitable to model the discrete response, with the logistic and probit regressions being the most commonly used. The above-mentioned initial sampling was used to identify a suitable model by using forward variable selection with the following steps, modelled on Sure Independence Screening

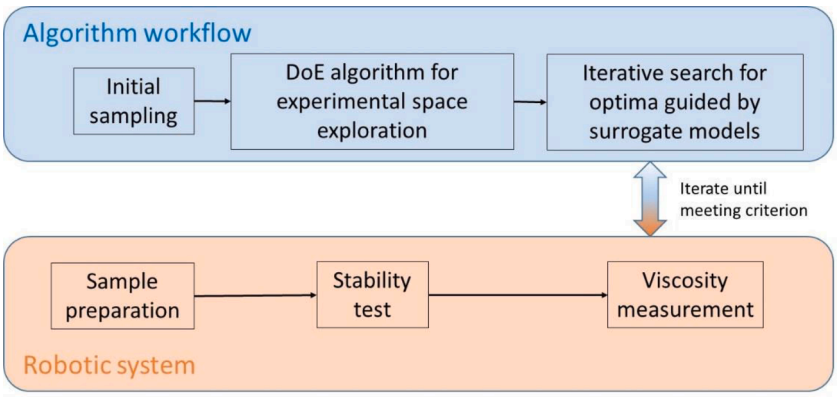

Fig. 1. Schematic diagram of the algorithmic-experimental workflow.

(Fan and Lv, 2008):

1) Fit models to each variable individually and identify the effects with p-values less than a given level of significance (set to 5% in this code).
2) Fit models which contain at least two of the effects identified in Step 1, and calculate AIC, BIC and deviance for these models.
3) Identify models which minimise AIC, BIC and deviance, and from these select the model with the most terms (to avoid erroneously deleting variables at this stage).

A logistic regression model with active parameters for the individual effect of S1, S4, P1 and P2 was found to be the best fitting model for these data. The initial sampling data were used to construct a prior distribution for this model, which was subsequently updated as new experimental data became available. To facilitate data collection, a Bayesian decision-theoretic methodology was incorporated into the bridge design approach described below.

As mentioned in Section 1, there are no physical models available for the viscosity test. Analysis of the past experimental data for this test using polynomial regression did not find any suitable models. A nonparametric Gaussian Process (GP) model was selected as a flexible, data-driven approach that works well with small-to-moderate sized experiments (Rasmussen and Williams, 2006). GPs are typically fit to data derived from designs with good space-filling properties, allowing accurate prediction across the design space.

The objective function used in this work is given by Eq. (1), and has a component relating to estimation of a logistic regression model and a component relating to a Euclidean distance between design points:

$$\varnothing(\mathbf{D}) = \omega\log(\widetilde{E}[\mathrm{U}(\mathbf{D})]) + (1-\omega)\log(\mathrm{d}(\mathbf{D})) \tag{1}$$

where: $\omega \in [0,1]$ is the weight placed on the part of the objective function relating to the stability test response; $(1-\omega)$ is the weight placed on the part of the objective function relating to the viscosity test response; $\mathbf{D}$ is the design scaled to be between 0 and 1; $[\mathrm{U}(\mathbf{D})]$ is the estimate of the expected utility for $\mathbf{D}$, which is the part of the objective function related to the logistic regression for stability test responses; and $\mathrm{d}(\mathbf{D})$ is the average Euclidean distance between all possible pairs of rows in $\mathbf{D}$, which is the space-filling part of the objective function related to the viscosity test responses. The expected utility is:

$$\mathrm{E}[\mathrm{U}(\mathbf{D})] = \int u(\theta,\ y,\ \boldsymbol{D})\pi(y|\theta,\ \boldsymbol{D})\pi(\theta|\boldsymbol{D})d\theta dy = \int u(\theta,\ y,\ \boldsymbol{D})\pi(\theta|\boldsymbol{D})d\theta dy \tag{2}$$

where θ are the parameters in a logit model for the stability test response y, $\pi(y|\theta,\ D)$ is the posterior distribution of the response, $\pi(\theta|D)$ is the prior for θ and $\pi(y,\ \theta|D)$ is the joint distribution of y and θ. The utility function, $\mathrm{u}(\theta,\ \mathrm{y},\ \mathrm{D})$, can be chosen based on the aims of the experiments. In this case, Shannon information gain, which maximizes the expected divergence between the posterior and prior distributions, is used as the utility function. The prior for θ is adapted from the initial experimental results.

Under the assumptions made in this work, $\mathrm{E}[\mathrm{U}(\mathrm{D})]$ is not analytically tractable, and it is estimated using Monte Carlo integration as

$$\widetilde{E}[\mathrm{U}(\mathbf{D})] = \frac{1}{B}\sum_{b=1}^{B} u(\theta_b,\ y_b,\ D) \tag{3}$$

where $\mathrm{y_b}$ and θ_b are sampled from $\pi(y,\ \theta|D)$, and B is the number of samples. This estimate is found using *utilityglm* function in the *acebayes* package in R given by Overstall and Woods (Overstall and Woods, 2017). Here, we let $B = 1000$.

Space filling designs can be used when a GP model is assumed for the response. Space filling designs impose restriction on the space of, or distance between, points in the design space. In this case, we use the average Euclidean distance as a space filling criterion for the viscosity test response. This distance is calculated using the pairwise distance between rows in the unscaled design, so D is converted from 0 to 1 scaling back to the original scale in the function $\mathrm{d}(\mathbf{D})$ in Eq. 1.

The weight on each of these two components can be adjusted based on the experimenter's aims. For example, if it is assumed that the outcome of the stability test is more important than that of the viscosity test, $\omega > 0.5$ would be appropriate, and vice versa. In this case, we set $\omega = 0.5$ as we treat the two responses as equally important.

3.2.2. *Point exchange algorithm*

Point exchange algorithms find an optimal design by optimizing each row of the design with respect to a certain objective function, whilst assuming the other rows are fixed (Fedorov, 1972). These algorithms perform multiple loops through the design and continue to optimize rows until stopping criteria are met. In order to avoid any issues with local optima, such algorithms are run for multiple random starting designs. The optimal design is the design found using the algorithm from these random starts which maximizes the objective function.

The estimated expected utility, Eq. (2), is computationally expensive to calculate, and hence also Eq. (1). Hence, we require a computationally efficient method of optimization. Also, we want to consider samples of possible values of Eq. (1) when choosing whether to accept or reject a proposed new row, as Eq. (2) is dependant on random samples from the joint distribution of θ and y. We therefore optimize the rows using the Efficient Global Optimization (EGO) algorithm (Jones et al., 1998), and accept or reject a proposed row using the Kolmogorov-Smirnov (KS) test (Smirnov, 1939).

The EGO algorithm is a type of Bayesian optimization algorithm which can be used to optimize computationally expensive functions (Shahriari et al., 2016). Bayesian optimization algorithms fit a Gaussian process model to the observed function values, and then choose the next

location at which to evaluate the function using an acquisition function that relies on this Gaussian process model. Points are iteratively added until a stopping criterion is met.

The acquisition function is chosen to balance the objectives of exploring the space where little is known about the function and exploiting the information we have gained by observing the function at given points. Bayesian optimization is demonstrated for a function with a single controllable variable in Fig. 2.

The EGO algorithm uses Expected Improvement (EI) (Mockus et al., 1978) as the acquisition function, and continues to add points to the algorithm until the current maximum EI value is less than or equal to 1% of the current maximum objective function value. In this algorithm, we also add a restriction on the number of new points that can be added.

We choose to accept or reject a proposed new row based on a comparison of samples of Eq. (1) for a design with and without this new row, which are found by generating R ($R = 1000$ in this work) samples from the joint distribution of θ and y. The KS test compares two samples to assess whether they come from the same distribution, where the null hypothesis is that these samples come from the same distribution. If the p-value of the KS test is less than α (set to 0.05 in this work), then there is evidence to reject this null at a 100α% significance. Hence, such a p-value for a KS test between two samples of Eq. (1) gives evidence to suggest that the objective function distribution after the swap differs from that before the swap and, therefore, gives evidence to accept the proposed new row. We also add the condition that the objective function itself must have increased, as we want to find the design which maximizes Eq. (1). An example of the estimated densities for these samples is given in Fig. 3. The Point Exchange Efficient Global Optimization (PEEGO) algorithm, summarized in Scheme 1, has been packed as an R package which is available online (https://github.com/sustainable-processes/PEEGO; https://doi.org/10.5281/zenodo.5908388).

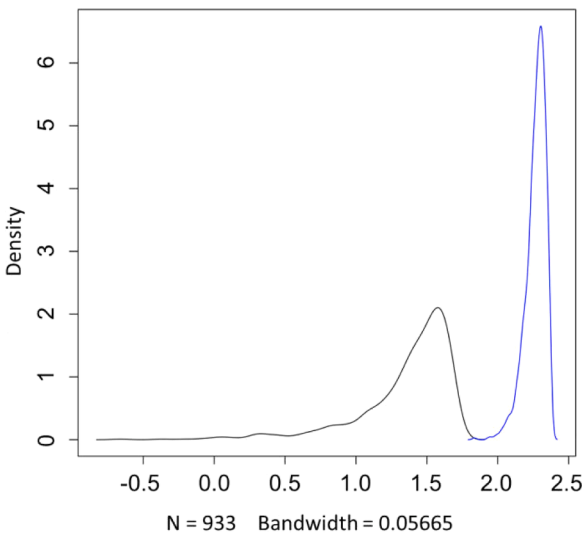

Fig. 3. The estimated densities of two objective function samples, one before a swap (black line) and one after a swap (blue line). The KS test for the comparison of these two samples has a p-value of less than 0.05, hence the null hypothesis that these two samples are drawn from the same distribution can be rejected at a 5% level.

4. Results and discussion

As explained in Section 3.1, the algorithm was trained using a maximin space filling design consisting of 230 data points. 52% of the samples in the training data set were stable, 12.61% met the viscosity target and only 3.48% passed both criteria. The algorithm was run in

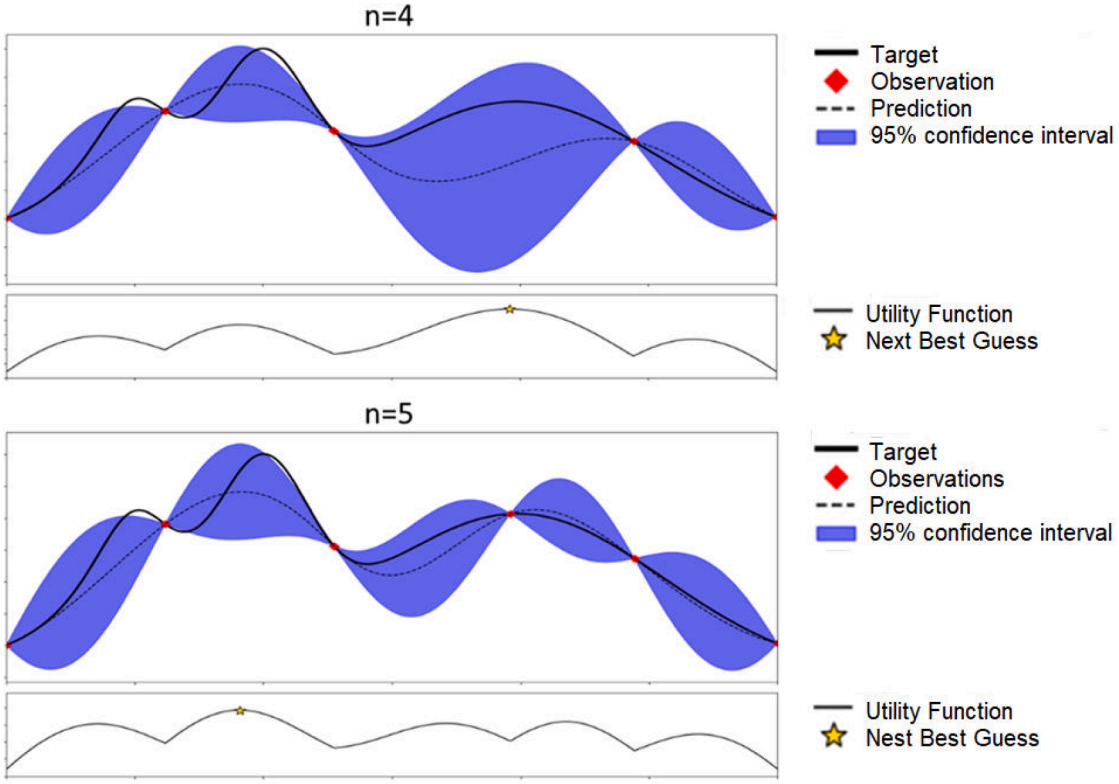

Fig. 2. Illustration of a Bayesian optimization approach (adapted from Shahriari et al. (Shahriari et al., 2016)). The red points are the current evaluations of the function $f(x)$, the solid black line is the Gaussian process estimate of the objective function, the dotted black line is the unknown objective function. The purple shaded area gives the uncertainty in the prediction of the objective function. Note that new points, given by red points, are added where the acquisition function is maximized.

(1) Generate a random starting design **D**.

(2) For each row, x_i, in **D** $i = 1, \ldots, n$:

(a) Let p be a vector of length 10. For each of the 10 possible combinations of three of S1-S5 being active. $j = 1, \ldots, 10$:

(i) Generate $\chi_j = \{x^1, \ldots x^T\}$ ($r = 50$ in this case) random rows to be swapped for x_j, where the jth combination of S1-S5 is active in all elements of χ_j.

(ii) Fit a Gaussian process to $\phi(\chi_j) = (\phi(\mathbf{D}(x^1)), \ldots, \phi(\mathbf{D}(x^T)))$ where **D**(x) is the design **D** with i-th row x.

(iii) Find the row x_{r+1} which has the j-th combination of S1-D7 active and maximizes EI when swapped with x_i (this is done using a constrained optimization algorithm in the package *DiceOptim* in R)

(iv) If the maximum EI is less then $0.01 \max(\phi(\chi_j))$ and $its_{EGO} \geq maxits_{EGO}$ ($maxits_{EGO} = 100$ in this work), do not amend χ_j and repeat from step (ii)

(v) Let x_j^* be the element of χ_j for which $\phi(\chi_j)$ is maximized. Add this to a set χ^*.

(vi) Let the j-th element of $\boldsymbol{p}$, p_j, be the p-value of the KS test for the comparison of samples of $\phi(\boldsymbol{D}(x_j^*))$.

(b) Find the $x^* \in \chi^*$ which relates to $p^* = \min(\boldsymbol{p})$.

(c) If $p^* < \alpha$ ($\alpha = 0.05$ in this work) and $\phi(\boldsymbol{D}(x^*)) > \phi(\boldsymbol{D}(x_i))$, then swap x_i for x^*. Otherwise, do not change x_i.

3 Let D^* be the design after all rows have been optimized using step 2.

4 If:

- The p-value of the KS test for comparison of the samples of $\phi(\boldsymbol{D})$ and $\phi(\boldsymbol{D}^*)$ is less than α,
- $\phi(\boldsymbol{D}^*) > \phi(\boldsymbol{D})$,
- The average absolute difference between all pairwise values in $\boldsymbol{D}$ and $\boldsymbol{D}^*$ is less than a tolerance t ($t = 0.01$ in this work), and
- $its_{PE} \geq maxits_{PE}$ ($maxits_{PE} = 100$ in this work),

Then stop the algorithm and return $\boldsymbol{D}^*$ as the optimal design. Otherwise, let $\boldsymbol{D}^*$ be $\boldsymbol{D}$ and $its_{PE} = its_{PE} + 1$ and return to step 2.

Scheme 1. The Point Exchange Algorithm for the Specific Problem.

order to suggest an experimental bridge design of 48 samples. The entire DoE is reported in the Supplementary Information (Table S2). In Fig. 3, it is shown that the blue line (the new design) in the plot shifted to the right of the black line (the initial design). Based on the p-value for the KS test, we infer that by applying the point exchange algorithm, the new design can give more information about the system than the original design.

In Table 1, we show a comparison between the percentage of samples passing stability and viscosity criteria in the training data set and in the suggested DoE. In the latter, 91.67% of the samples were stable, and 12.5% passed both criteria. This may be ascribed to the fact that in the initial design we select experimental points according to a space-filling criterium, without taking into account the information gain of the models. Therefore, in order to gain more information of the system, the algorithm seeks to explore the part never seen in the initial dataset, more likely to consist of stable samples in the right viscosity range.

Table 1
Comparison between the training data set and the suggested DoE: percentages of sample that passed and/or failed stability and viscosity tests.

	Maximin DoE (%)	Bridge-Design DoE (%)
Stability test: passed Viscosity test: passed	3.48	12.50
Stability test: failed Viscosity test: passed	9.13	0.00
Stability test: passed Viscosity test: failed	48.70	79.17
Stability test: failed Viscosity test: failed	38.69	8.33

The resulting dataset was used to train a GP for the prediction of viscosity of samples and to guide optimization of the formulation within three experimental iterations. To mimic a common situation in product

development, some prior knowledge was included in the data set. In fact, it was known that mixtures of the surfactants without any added polymer and thickener show a water-like viscosity. The trained GP was used to predict the viscosity response over the entire input variable space. At each iteration the candidates predicted to be closer to the midpoint of the desired target range (2.0 – 4.0 Pa•s) were selected. Using the trained classifier, solutions predicted to be unstable were discarded, and the resulting 20 best experiments were tested experimentally. The 60 experiments carried out are reported in the Supplementary Information, Table S3. In Table 2 the formulations passing both criteria are reported. Interestingly, all 60 experiments resulted in clear, stable formulations, 20% of which passed the viscosity test.

At this point, it is worth pointing out that a good number of candidates was obtained within a total number of 338 experiments carried out in 17 working days, without any need for a physical model for properties prediction. 80% of the time was needed for the non-automated viscosity tests, 20% of the time for the automated sample preparation and stability tests, whereas the computational time was negligible. As a reference, we can use the results recently published by some of the authors for a similar system using a combination of Latin hypercube sampling and Thompson-sampling efficient multi-objective optimization algorithm (TSEMO) coupled with a Bayesian classifier.[19] In that case, the recipes of the formulations were optimized to obtain stable and clear formulations with the same target viscosity, but without allowing for a free choice of a surfactant. Only three surfactants were available, corresponding to S2, S3, and S5 used in this work. The optimization procedure was started using 96 LHC experiments and 128 further experiments were collected in 16 iterations of the optimization algorithm, with a total of 224 experiments performed. Although in the reported earlier paper the formulations were also optimized with respect to price, it is worth noting that in the current work the choice of three surfactants from the five available makes the input space variable one order of magnitude larger. However, thanks to the adoption of a maximin design coupled with a bridge-design approach, the total number of the required experiments increased only to 338.

The examination of the solutions gives an insight into the role of the different ingredients in the processed formulations. The lowest occurrence of surfactants S2, S3, and S4 suggests that interactions of these compounds with the other components have a higher probability to form unstable, turbid mixtures. As expected, the thickener T1 is responsible for higher viscosity of the samples and its concentration tends to be close to the upper limit of the adopted constraints; however, contrary to suggestions of human experts, the algorithm was able to find good solutions also using a concentration of T1 lower than 2 g L^{-1}, which significantly decreases price of the final product. Interestingly, although polymer P1 was considered by human experts to be responsible for the increase in viscosity, when certain combinations of surfactants are adopted, the polymer concentration can be reduced, suggesting that interactions between these ingredients are contributing to the increase in viscosity.

As shown, this preliminary analysis gave some qualitative insight about the physics of the system. Current research is stressing the need of using the results of black-box optimizations and robotic experimental campaigns to derive some physical knowledge about the investigated systems: some examples can be found in the analysis of the hyperparameters (Schweidtmann et al., 2018), the automated identification of physical laws from bare data (Neumann et al., 2020), and the analysis of the Pareto front (Cao et al., 2021). In this regard, future research will need to rationalize and combine these different approaches to maximize the amount of physical information derived from automated procedures.

5. Conclusions

In this work, the Point Exchange Efficient Global Optimization (PEEGO) algorithm was used to find a bridge-design of experiments to maximize the information gain, in order to find suitable solutions for a commercial formulated product. The corresponding R package is available on Github through the link (https://github.com/sustainable-processes/PEEGO). The proposed methodology was tested with the design of a commercial liquid formulated product, where only three surfactants can be chosen from a library of ingredients. A logistic model and a Gaussian process model was selected to describe a discrete and a continuous target of the product, i.e. stability and viscosity. The PEEGO algorithm was then applied to simultaneously optimize the information gain for the two responses.

A cheap-to-evaluate GP was trained using the experimental results and used to predict the viscosity response over the entire input variable space. This triggered an iterative process that allowed to increase the percentage of samples passing both quality criteria from 3.68% (maximin DoE) and 12.50% (bridge-design DoE), to 20.00% over 60 samples obtained in three iterations. This outperformed the results previously obtained for a similar case study, using a Latin hypercube sampling approach coupled with an iterative procedure, in the absence of a bridge-design approach.

In addition to the good number of candidates obtained in a short time in the absence of physical predictive models, the *a posteriori* analysis of the obtained solutions gives some qualitative physical insight to the role of the different ingredients and their non-trivial complex interactions. Further research will be needed to rationalize this information using systematic approaches for the generation of physical knowledge from fast automated development of formulated products.

CRediT authorship contribution statement

Liwei Cao: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Visualization. **Danilo Russo:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Visualization. **Emily Matthews:** Conceptualization, Methodology, Software. **Alexei Lapkin:** Conceptualization, Funding acquisition, Project administration, Resources, Supervision. **David Woods:**

Table 2
Experimentally tested formulations passing the stability and viscosity criteria at the same time. S1-S5: surfactants; P1: polymer; T1: thickener".

S1 (g/L)	S2 (g/L)	S3 (g/L)	S4 (g/L)	S5 (g/L)	P1 (g/L)	T1 (g/L)	Viscosity (Pa•s)	Iteration
1.66	0.00	8.34	0.00	5.00	2.00	1.30	3.60	1
1.66	0.00	0.00	6.66	6.66	1.80	1.60	2.52	1
5.00	0.00	0.00	5.00	5.00	2.00	0.90	2.38	1
5.35	0.00	6.43	3.21	0.00	1.29	1.71	2.86	2
0.00	5.36	4.29	0.00	4.29	0.29	1.57	2.75	2
0.00	3.21	0.00	6.43	5.36	1.43	1.86	3.36	3
2.14	6.43	0.00	0.00	5.36	0.00	1.86	3.13	3
0.00	6.43	0.00	3.21	5.36	2.00	1.86	2.72	3
7.50	0.00	1.07	5.36	0.00	1.00	1.71	2.69	3
2.14	0.00	7.50	0.00	4.29	1.57	1.29	3.62	3
5.36	3.21	6.43	0.00	0.00	0.00	1.71	3.28	3
10.71	0.00	0.00	0.00	4.29	0.00	1.57	2.49	3

Conceptualization, Funding acquisition, Resources, Software, Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This project was co-funded by the UKRI project "Combining Chemical Robotics and Statistical Methods to Discover Complex Functional Products" (EP/R009902/1), and National Research Foundation (NRF), Prime Minister's Office, Singapore under its Campus for Research Excellence and Technological Enterprise (CREATE) program as a part of the Cambridge Centre for Advanced Research and Education in Singapore Ltd (CARES). LC is grateful to BASF for co-funding her PhD.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.compchemeng.2022.108083.

References

Atkinson, A.C., Donev, A.N., Tobias, R., 2007. Optimum Experimental Designs, with SAS, 2nd. ed. Oxford University Press, Oxford, UK.

Cao, L., Russo, D., Felton, K., Salley, D., Sharma, A., Keenan, G., Mauer, W., Gao, H., Cronin, L., Lapkin, A.A., 2021. Optimization of formulations using robotic experiments driven by machine learning DoE. Cell Reports Phys. Sci. 2, 100295 https://doi.org/10.1016/j.xcrp.2020.100295.

Cao, L., Russo, D., Mauer, W., Gao, H.H., Lapkin, A.A., 2020. Machine learning-aided process design for formulated products. Comp. Aided Chem. Engng. 48, 1789–1794. https://doi.org/10.1016/B978-0-12-823377-1.50299-8.

Chaloner, K., Verdinelli, I., 1995. Bayesian Experimental Design: A Review. School of Statistics Technical Reports, University of Minnesota. Technical Report 607. https://hdl.handle.net/11299/199630.

Conte, E., Gani, R., Ng, K.M., 2011. Design of formulated products: a systematic methodology. AIChE J 57, 2431–2449. https://doi.org/10.1002/aic.12458.

Fan, J., Lv, J., 2008. Sure independence screening for ultrahigh dimensional feature space. J. R. Stat. Soc. Ser. B (Statistical Methodol. 70, 849–911. https://doi.org/10.1111/j.1467-9868.2008.00674.x.

Fedorov, V.V., 1972. Theory of Optimal Experiments. Academic Press, New York.

Fung, K.Y., Ng, K.M., Zhang, L., Gani, R., 2016. A grand model for chemical product design. Comput. Chem. Eng. 91, 15–27. https://doi.org/10.1016/j.compchemeng.2016.03.009.

Gani, R., Ng, K.M., 2015. Product design – molecules, devices, functional products, and formulated products. Comput. Chem. Eng. 81, 70–79. https://doi.org/10.1016/j.compchemeng.2015.04.013.

Heintz, J., Belaud, J.-.P., Pandya, N., Teles Dos Santos, M., Gerbaud, V., 2014. Computer aided product design tool for sustainable product development. Comput. Chem. Eng. 71, 362–376. https://doi.org/10.1016/j.compchemeng.2014.09.009.

Johnson, M.E., Moore, L.M., Ylvisaker, D., 1990. Minimax and maximin distance designs. J. Stat. Plan. Inference 26, 131–148. https://doi.org/10.1016/0378-3758(90)90122-B.

Jones, B., Silvestrini, R.T., Montgomery, D.C., Steinberg, D.M., 2015. Bridge designs for modeling systems with low noise. Technometrics 57, 155–163. https://doi.org/10.1080/00401706.2014.923788.

Jones, D.R., Schonlau, M., Welch, W.J., 1998. Efficient global optimization of expensive black-box functions. J. Glob. Optim. 13, 455–492. https://doi.org/10.1023/A:1008306431147.

Jouyban, A., Chan, H.-.K., Chew, N.Y.K., Khoubnasabjafari, M., Acree Jr, W.E., 2006. Solubility prediction of paracetamol in binary and ternary solvent mixtures using Jouyban-Acree model. Chem. Pharm. Bull. 54, 428–431. https://doi.org/10.1248/cpb.54.428.

Jouyban, A., Shayanfar, A., Panahi-Azar, V., Soleymani, J., Yousefi, B., Acree, W., York, P., 2011. Solubility prediction of drugs in mixed solvents using partial solubility parameters. J. Pharm. Sci. 100, 4368–4382. https://doi.org/10.1002/jps.22589.

Kontogeorgis, G.M., Mattei, M., Ng, K.M., Gani, R., 2019. An integrated approach for the design of emulsified products. AIChE J 65, 75–86. https://doi.org/10.1002/aic.16363.

Li, G., Bastian, C., Welsh, W., Rabitz, H., 2015. Experimental design of formulations utilizing high dimensional model representation. J. Phys. Chem. A 119, 8237–8249. https://doi.org/10.1021/acs.jpca.5b04911.

Mockus, J., Tiesis, V., Zilinskas, A., 1978. The application of Bayesian methods for seeking the extremum. In: Dixon, L.C., Szego, G. (Eds.), Towards Global Optimisation. North Holland, Amsterdam, pp. 117–129.

Neumann, P., Cao, L., Russo, D., Vassiliadis, V.S., Lapkin, A.A., 2020. A new formulation for symbolic regression to identify physico-chemical laws from experimental data. Chem. Eng. J. 387, 123412 https://doi.org/10.1016/j.cej.2019.123412.

Overstall, A.M., Woods, D.C., 2017. Bayesian design of experiments using approximate coordinate exchange. Technometrics 59, 458–470. https://doi.org/10.1080/00401706.2016.1251495.

Peremezhney, N., Connaughton, C., Unali, G., Hines, E., Lapkin, A.A., 2012. Application of dimensionality reduction to visualisation of high-throughput data and building of a classification model in formulated consumer product design. Chem. Eng. Res. Des. 90, 2179–2185. https://doi.org/10.1016/j.cherd.2012.05.010.

Rasmussen, C.E., Williams, C.K.I., 2006. Gaussian Processes for Machine Learning. MIT Press, Cambridge, MA.

Ryan, E.G., Drovandi, C.C., McGree, J.M., Pettitt, A.N., 2016. A review of modern computational algorithms for bayesian optimal design. Int. Stat. Rev. 84, 128–154. https://doi.org/10.1111/insr.12107.

Schweidtmann, A.M., Clayton, A.D., Holmes, N., Bradford, E., Bourne, R.A., Lapkin, A.A., 2018. Machine learning meets continuous flow chemistry: automated optimization towards the Pareto front of multiple objectives. Chem. Eng. J. 352, 277–282. https://doi.org/10.1016/j.cej.2018.07.031.

Shahriari, B., Swersky, K., Wang, Z., Adams, R.P., De Freitas, N., 2016. Taking the human out of the loop: a review of Bayesian optimization. In: Proc. IEEE. https://doi.org/10.1109/JPROC.2015.2494218.

Smirnov, N.V., 1939. Estimate of deviation between empirical distribution functions in two independent samples. Bull. Moscow Univ. 2, 3–16.

Ten, J.Y., Hassim, M.H., Ng, D.K.S., Chemmangattuvalappil, N.G., 2017. A molecular design methodology by the simultaneous optimisation of performance, safety and health aspects. Chem. Eng. Sci. 159, 140–153. https://doi.org/10.1016/j.ces.2016.03.026.

Uhlemann, J., Costa, R., Charpentier, J.-.C., 2020. Product design and engineering — past, present, future trends in teaching, research and practices: academic and industry points of view. Curr. Opin. Chem. Eng. 27, 10–21. https://doi.org/10.1016/j.coche.2019.10.003.

Wesselingh, J.A., Kiil, S., Vigild, M.E., 2007. Design & Development of Biological, Chemical, Food and Pharmaceutical Products. John Wiley & Sons, West Sussex, England.

Yunus, N.A., Gernaey, K.V., Woodley, J.M., Gani, R., 2014. A systematic methodology for design of tailor-made blended products. Comput. Chem. Eng. 66, 201–213. https://doi.org/10.1016/j.compchemeng.2013.12.011.

Zhang, L., Fung, K.Y., Wibowo, C., Gani, R., 2018. Advances in chemical product design. Rev. Chem. Eng. 34, 319–340. https://doi.org/10.1515/revce-2016-0067.

Zhang, L., Mao, H., Liu, Q., Gani, R., 2020. Chemical product design – recent advances and perspectives. Curr. Opin. Chem. Eng. 27, 22–34. https://doi.org/10.1016/j.coche.2019.10.005.