Computers and Chemical Engineering 195 (2025) 109007

Contents lists available at ScienceDirect

Computers and Chemical Engineering

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

Machine learning-guided space-filling designs for high throughput liquid formulation development

Aniket Chitre [a,b,c], Daria Semochkina [d], David C. Woods [d], Alexei A. Lapkin [a,c,*]

[a] Department of Chemical Engineering and Biotechnology, University of Cambridge, Philippa Fawcett Drive, Cambridge CB3 0AS, United Kingdom
[b] Cambridge Centre for Advanced Research and Education in Singapore, CARES Ltd. 1 CREATE Way, CREATE Tower #05-05, Singapore 138602, Singapore
[c] Institute of Materials Research and Engineering, Agency for Science, Technology and Research (A*STAR), Singapore 138634, Singapore
[d] Southampton Statistical Sciences Research Institute and School of Mathematical Sciences, University of Southampton, Southampton SO17 1BJ, UK

ARTICLE INFO

Keywords:
Design of experiments
Machine learning
Liquid formulations
Phase stability

ABSTRACT

Liquid formulation design involves using a relatively limited experimental budget to search a high-dimensional space, owing to the combinatorial selection of ingredients and their concentrations from a larger subset of available ingredients. This work investigates alternative shampoo formulations. A space-filling design is desired for screening relatively unexplored formulation chemistries. One of the few computationally efficient solutions for this mixed nominal-continuous design of experiments problem is the adoption of maximum projection designs with quantitative and qualitative factors (MaxProQQ). However, such purely space-filling designs can select experiments in infeasible regions of the design space. Here, stable products are considered feasible. We develop and apply weighted-space filling designs, where predictive phase stability classifiers are trained for difficult-to-formulate (predominantly unstable) sub-systems, to guide these experiments to regions of feasibility, whilst simultaneously optimising for chemical diversity by building on MaxProQQ. This approach is extendable to other mixed-variable design problems, particularly those with sequential design objectives.

1. Introduction

Liquid formulations are complex multi-component mixtures where the ingredients have been selected, processed, and combined in a specific way to obtain well-defined functions (Conte et al., 2011). Liquid formulation design involves both quantitative and qualitative factors, with factors being controlled independent variables in the experiment. In this study, we address a mixed nominal-continuous experimental design problem, where the nominal factors are ingredient choices, and concentration selection represents the continuous factors. Typically, these formulated products, which are produced across several industries (e.g., consumer care, agrochemical, pharmaceutical; Bagajewicz et al., 2011; Bernardo and Saraiva, 2005; Gani, 2004; Narayanan et al., 2021; Taifouris et al., 2020), are developed through trial and error by specialists with extensive experience in the given domain. Industry seeks a more data-driven and predictive methodology to develop formulated products, particularly, as we wish to formulate novel products, either for enhanced performance and functionality (Gani and Ng, 2015; Martín and Martínez, 2013) or environmental reasons (Jessop et al., 2015; Kelly, 2023).

This work forms part of a broader study on machine learning for liquid formulation design where we aimed to collect a dataset for training predictive surrogate models for a set of industrially important property targets in this space: phase stability, turbidity, and viscosity. The experimental work and collected dataset are published separately (Chitre et al., 2024b). Here we focus specifically on the design of experiments (DoE) problem. Key aspects of this were as follows: (i) limited prior knowledge or models for the chemical space explored; (ii) a high-dimensional, mixed variable design space; (iii) a sequential design problem. Each of these is addressed in turn.

Formulation chassis (core ingredients) have remained relatively unchanged in personal care products such as shampoo for two or more decades. Within the broader study, we investigated a range of new surfactant ingredients in response to regulatory and sustainability pressures (Chitre et al., 2024b), therefore, we had limited *a priori* knowledge or models available for our formulation system. Had such knowledge been available, it could have been used to define a model-based experimental scheme which, for example, may have selected experiments in regions of predicted nonlinearities (Huang et al., 2019) and been integrated within a model-based design of experiments

* Corresponding author.
E-mail address: aal35@cam.ac.uk (A.A. Lapkin).

https://doi.org/10.1016/j.compchemeng.2025.109007
Received 6 August 2024; Received in revised form 15 January 2025; Accepted 17 January 2025
Available online 18 January 2025

(MBDoE) framework (Galvanin et al., 2012, 2007). Although we expect the response surface to be non-linear due to the complex ingredient interactions which are typical in liquid-formulated products, we do not know *a priori* where these non-linearities will be in the exploration of new chemical moieties.

Our goal was to generate a dataset of the most diverse set of formulations because we aimed to train representative property (phase stability, viscosity, turbidity) prediction models across the *entire* chemical design space, as defined by the ingredients and concentration ranges selected with our industrial partner. Given the anticipated complexity and limited knowledge at the start of the experimental campaign, we intended to use flexible non-parametric methods for property prediction. Contrast this to parametric models of a fixed functional form, which would be useful in later stages of product development, i.e., when a suitable form of the model is known. For non-parametric models, space-filling designs improve the smoothing of model predictions. Space-filling ensures the experimental design points, representing various combinations of ingredient choices and concentrations are spread as evenly as possible.

Our initial approach considered using a recently developed "bridge DoE" for liquid formulation design (Cao et al., 2023); bridge design refers to a combination of a parametric model-based and space-filling component. In this earlier study, a parametric method was viable as a simpler design space with only 10 ingredient combinations, unlike the 500+ in this work, was used. Furthermore, the bridge DoE method scales poorly to higher-dimensional problems as the model-based component relies on expensive Monte Carlo integration and the space-filling optimisation relies on an inefficient comparison between all possible pairs of rows in a generated design matrix.

Since formulation design is a combinatorial problem with different types of factors, we often have a very large design space to explore, yet we have a limited experimental budget. This is generally true for chemical (engineering) problems as experiments are time-, resource-, and labour-intensive, but particularly for formulation design, as developing a fully automated, high-throughput liquid formulation workflow is very challenging (Cao et al., 2021b; Chitre et al., 2024b). Therefore, we needed an efficient design of experiments (DoE) methodology. A mixed nominal-continuous design problem, as faced in this study, is particularly challenging for space filling because for the nominal factors, it is difficult to interpret the distance between points. Either the value of a nominal factor is the same between two experiments, or it is not – there is a lack of quantification of how different the two experiments are, unlike with continuous, discrete numeric or even ordinal factors. Furthermore, the mixed variable design space can become prohibitively expensive to explore using traditional methods, such as (fractional) factorial designs, as the number of factors increases because the design space grows exponentially due to the combinatorial nature of ingredient selection.

With the increasing adoption of ML-driven methods in the chemical sciences, a general consensus has formed that a well-distributed training set will lead to better predictions across the entirety of the chemical space of interest (Ahneman et al., 2018; Glavatskikh et al., 2019; Schrader et al., 2024; Strieth-Kalthoff et al., 2022). It is only practically possible to cover a limited fraction of the design space experimentally and a space-filling design ensures optimal spread or coverage such that when a prediction is made for a new formulation, a representative experimental sample is nearby (Johnson et al., 1990; Joseph, 2016; McKay et al., 1979). For brevity, we will focus the discussion on Latin hypercube designs (LHD) as these have been best extended to a mixed-variable design space. An LHD attempts to address the curse of dimensionality in space-filling by ensuring by construction uniform coverage in each one-dimensional projection of the design. Other types of space-filling approaches are also available such as Sobol sampling (Sobol, 1967), maximum entropy (Shewry and Wynn, 1987) or minimum energy (Joseph et al., 2015) designs. A comprehensive review is presented elsewhere (Garud et al., 2017).

In practice, the Maximin LHD (Mm LHD; Morris and Mitchell, 1995) is the most commonly used space-filling design due to its simplicity and availability in software packages. However, LHDs are only available for continuous factors. Therefore, the sliced LHD (SLHD) was introduced (Qian, 2012), which is a type of LHD that can be further partitioned into t smaller LHDs called slices where t is the number of all possible combinations of the qualitative factors. Despite a more computationally efficient construction of the SLHD proposed by Ba et al., 2015, the method remains limited for formulation design with numerous ingredient choices, as the number of required slices, t, increases exponentially with the number of qualitative factors. As an alternative, faster method, marginally coupled designs (MCDs) were proposed by Deng et al., 2015 which combine orthogonal arrays (OAs) for the qualitative factors with LHDs for the quantitative factors. The trade-off, however, is sub-optimal space-filling in higher dimensions as only certain groups of factors are optimised independently and this is an active research area (Zhou et al., 2021).

Maximin (S)LHDs have optimal space-filling properties in the full p-dimensions of a design problem and provide uniform 1-D projections. However, their space-filling properties can be poor in lower-dimensional projections (from 2 to $p-1$ dimensions), which can be relevant to formulation design. Commercial formulations contain many ingredients, some of which may not have an active effect on a particular property of interest, e.g., viscosity, in which case we are interested in the lower dimensional subspaces. In this reduced space, (S)LHDs may no longer ensure adequate space-filling properties, which is crucial for training ML models.

To address the limitation of (S)LHD only accommodating a small number of nominal factors and the space-filling limitations of MCD, Joseph et al. (2020) extended the Maximum Projection (MaxPro) design criterion for continuous variables, introduced by Joseph et al. (2015b), to accommodate continuous, nominal, discrete numeric, and ordinal types of factors in one criterion (MaxProQQ). These authors demonstrated the performance advantages, especially in space-filling requirements, of the MaxPro designs over alternative space-filling approaches (particularly MCD and LHD). A brief discussion of the MaxProQQ criterion is included as Supplementary Eqn. 1 in the SI along with a brief discussion of the algorithm used to optimise the design. This criterion ensures good projections in all the subspaces of the factors while having a computational cost comparable to the widely used Mm LHD criterion (Morris and Mitchell, 1995). For more details on the MaxProQQ design construction process, readers are directed to (Joseph et al., 2020) and the related R package (Ba and Joseph, 2018).

Returning to the specific formulation design problem, we illustrate the liquid formulation workflow used in Fig. 1. The formulations, a mixture of surfactants, polymer, and thickener in a base of water, are prepared and then characterised for their phase stability and, if stable, their turbidity, and viscosity. Unstable formulations are not characterised further as they have non-uniform turbidity/viscosity across the different phases of the formulation. This therefore results in a sequential or hierarchical design problem: having a stable formulation is a prerequisite to the collection of turbidity or viscosity data. MaxProQQ could have been directly used to generate space-filling liquid formulation designs. However, this could have resulted in the selection of points in areas of little relevance, where it is known no stable formulation can occur (Bowman and Woods, 2013). Therefore, we developed and used a weighted space-filling (MaxProQQ) design to guide experiments to regions of phase stability. We do not know *a priori* which regions of the formulation design space will be stable. Therefore, we used an active learning approach to train a predictive phase stability classifier across the design space, which was used to guide difficult-to-formulate (predominantly unstable) sub-systems to regions of stability, as part of a machine learning-guided DoE (ML-DoE).

The remainder of the paper is structured as follows. First, we present the methodology for the phase-stability-guided MaxProQQ designs, including details for featurising liquid formulations and training

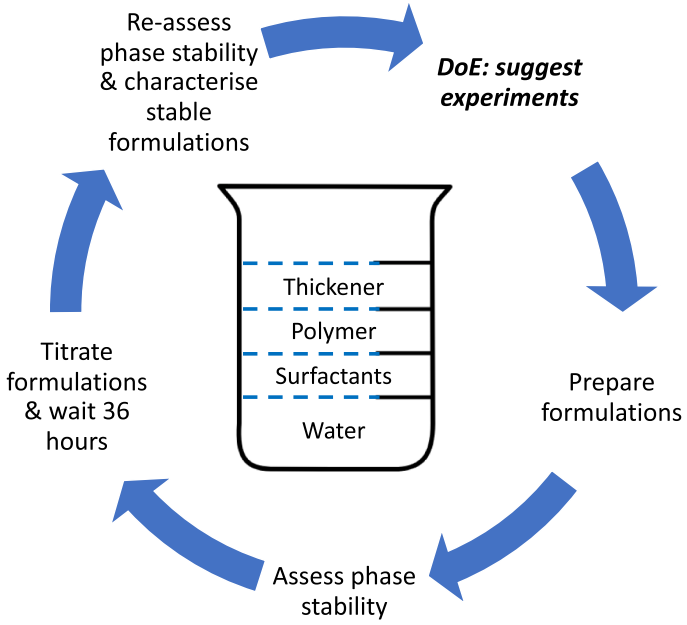

Fig. 1. An overview of the liquid formulation workflow driven by an ML-guided DoE method for a (weighted-) space-filling design towards phase stability and chemical diversity.

predictive stability classifiers. We then show in the Results and Discussion section, that we have been able to optimise formulations towards phase stability and demonstrate the spread and coverage of our designs. Finally, we present the performance of ML stability classifiers and discuss chemical interpretability of the obtained results.

2. Materials and methods

2.1. Definition of the experimental design problem

We used commercial formulation ingredients as received from our industrial partner, BASF. Materials are fully detailed elsewhere (Chitre et al., 2024a). Formulations were chosen from a set of 12 surfactants, four conditioning polymers (P_1 = Luviquat® Excellence, P_2 = Dehyquart® CC6, P_3 = Dehyquart® CC7 Benz, P_4 = Salcare® Super 7), and two thickeners (T_1 = Arlypon® TT, T_2 = Arlypon® F). Of these, we note that P_1 and P_2 were relatively highly charged cationic polyelectrolytes, whilst P_3 and P_4 had a lower charge density. This information will be used later to interpret the results.

In order to develop a phase stability classifier for a weighted-space filling design, the formulation samples need to be featurised (Wigh et al., 2022). Features are individual measurable properties or characteristics of the data used by models to make predictions. In this context, they serve as inputs representing the relevant aspects of the formulation samples. Featurisation is the process of transforming raw data into a structured set of features suitable for use by an ML model and is also synonymous with molecular representation (Pattanaik and Coley, 2020). Unfortunately, effectively featurising macromolecules, such as conditioning polymers and thickeners is an open research question, with many promising recent studies (Kim et al., 2018; Kuenneth and Ramprasad, 2023; Lin et al., 2019) but no general solution to date. Mixtures of such molecules – i.e., formulations, are even more difficult to represent. Currently, the simplest approach is to use concentrations of polymer and thickener added, and a one-hot encoding of the ingredients, as shown in a previous study from our group (Cao et al., 2021a).

We chose to split the design space into eight distinct sub-systems for each possible polymer, thickener combination: $(P_1,\ T_1)$, $(P_1,\ T_2)$, $(P_2,\ T_1)$ ..... $(P_4,\ T_2)$. Intuitively, we expect fixed combinations of polymer and thickener to exhibit some chemically similar behaviours with the different classes of surfactant molecules (anionic/non-ionic/amphoteric/cationic) tested. This step of fixing the polymer and thickener reduced the DoE problem by two dimensions to a 5-D problem: four continuous factors for concentrations of the surfactants, polymer and thickener ($C_{S1},\ C_{S2},\ C_P,\ C_T$), and one nominal factor (S_{pair}) representing the choice of a surfactant pair. A binary surfactant mixture, polymer and thickener is a typical shampoo formulation chassis. Mathematically, the experimental design vector for a formulation sample ($\boldsymbol{\varphi}$), could be represented as shown in Eqn. (1). The continuous factors will be sampled from concentration ranges pre-agreed with our industrial partner, while the nominal factor consists of combinations of two surfactant ingredients chosen from a set of 12, resulting in 66 possible levels.

$$\boldsymbol{\varphi} = \left[C_{S1},\ C_{S2},\ C_P,\ C_T,\ S_{pair}\right] \tag{1}$$

$$\begin{aligned} \text{with}\quad & 8.0 \le C_{S1}, C_{S2} \le 13.0\,\frac{w}{w}\% \\ & 1.0 \le C_P \le 3.0\ \frac{w}{w}\% \\ & 1.0 \le C_T \le 5.0\ \frac{w}{w}\% \\ & S_{pair}\ \in \left\{\left(S_i, S_j\right)\right\} : i,j = 1,\ \ldots,\ 12; i \neq j \end{aligned}$$

An important step for space-filling designs is to scale the different continuous factors to the same range, typically the unit interval, to ensure that the varying ranges of the factors do not unduly influence the design. We applied a conversion shown in Eqn. (2), based on the lower and the upper concentration bounds, as shown above, to convert between the experimental ($\boldsymbol{\varphi}$) and computational ($\boldsymbol{\varphi}'$) design vectors. Each design vector represents an individual formulation sample, and vectors can be combined into an experimental ($\mathbf{D}$) or computational ($\mathbf{D}'$) design matrix of dimension n, equal to the number of samples, as shown in Eqn. (3).

$$\boldsymbol{\varphi}' = \left[C'_{S1},\ C'_{S2},\ C'_P,\ C'_T,\ S_{pair}\right] \text{ where } C'_i = \frac{C_i - C_{i,\ LB}}{C_{i,\ UB} - C_{i,\ LB}},\ C'_i \in [0,\ 1] \tag{2}$$

$$\mathbf{D}' = \left[\boldsymbol{\varphi}'_1,\ \boldsymbol{\varphi}'_2,\ \ldots,\ \boldsymbol{\varphi}'_{\mathbf{n}}\right]^{\mathrm{T}} \tag{3}$$

In the broader work, it was of interest to characterise the phase

stability, turbidity, and viscosity of each formulation in the design matrix with methods as detailed in Chitre et al., 2024b. However, for the purposes of this ML-DoE the sole response variable considered is phase stability, as it is a pre-requisite to have stable formulations to be able to measure the other two properties of interest. Phase stability is reported in Chitre et al., 2024a as a binary result, stable or unstable, characterised by visual inspection as detailed in Chitre et al., 2024b. Consequently, there is no statistical confidence to associate with the measurement. The binary result is used to develop our weighted space-filling designs.

2.2. Featurising surfactants

As formulations are complex mixtures of ingredients, most studies to date, including prior work of Cao et al. (2023, 2021a) have directly trained models on the ingredient concentrations used, as presented in Fig. 2a. This, however, has limited interpretability and cannot be generalised to ingredients not in the training set. For this work, as there is no general solution to featurise polymers, we retained one-hot encoding for the polymers and thickeners, but we were able to featurise surfactants, which we can treat as small molecules. This featurisation is important as it forms the basis of the inputs for training the ML models for phase stability, which are used to bias proposed experiments to regions of feasibility, i.e., stability.

There are many methods for featurising small molecules, such as: (i) string-based representations, e.g., SMILES (Öztürk et al., 2016; Schwartz et al., 2013; Vidal et al., 2005), (ii) molecular graphs (Qin et al., 2021; Yang et al., 2019), and (iii) molecular features from 0-D to 3-D descriptors (Abooali and Soleimani, 2023; Consonni and Todeschini, 2010; Ghiringhelli et al., 2015; Seddon et al., 2022). This list is not exhaustive. There are many cheminformatics packages that can enumerate large numbers of descriptors, which should be feature-engineered down to a more sensible subset relative to the dataset size available for training (Bray et al., 2020; Moriwaki et al., 2018; O'Boyle et al., 2011; Yap, 2011). However, many of these featurisations require large training datasets or are not directly interpretable. With formulation design, we are constrained to generating hundreds, not thousands, of samples, even with state-of-the-art lab facilities, and so we developed a more chemically meaningful featurisation based on the surfactant functional groups, as shown in Fig. 2b. This was hypothesised to improve model performance and explainability, as illustrated in the Results and Discussion section.

A surfactant's behaviour is primarily governed by its head group and chain length (Kronberg et al., 2014). For each of the 12 surfactants used in the study, the unique functional groups (FG) were enumerated (Ertl, 2017) and tallied for each ingredient into a surfactant FG matrix, $\mathbf{S_{FG}}$. In Fig. 2a, the data frame presented is the experimental design matrix, $\mathbf{D}$, rows of formulation samples with ingredient concentrations. Taking the surfactant ingredient columns from the design matrix gives $\mathbf{D_S}$ and taking the matrix dot product of this with $\mathbf{S_{FG}}$, as shown by Eqn. (4), gives the concentration of surfactant functional groups used in each formulation, γ_{FG}, as shown in Fig. 2b. The experimental dataset and $\mathbf{S_{FG}}$ are summarised elsewhere (Chitre et al., 2024a, 2024b). The main advantage of our approach is that ML models trained over the surfactant functional group features can generalise to new surfactants with these functional groups.

$$\gamma_{FG} = \mathbf{D_S} \cdot \mathbf{S_{FG}} \tag{4}$$

2.3. Phase stability-guided MaxProQQ designs

Scheme 1 outlines the algorithm for the ML-guided DoE method. The DoE was conducted offline with all steps programmed in R, except for step three, which was performed in a Jupyter Notebook (Python). The output from the DoE was a CSV file readable by an Opentrons liquid handling robot, which dispensed formulation ingredients. As stated in Section 2.1, the design matrix is represented in a computational ($\mathbf{D}'$) and an experimental ($\mathbf{D}$) form, with the conversion for an individual sample from this matrix shown in Eqn. (2). Furthermore, a look-up table of surfactant pairs was used to equate each of the 66 levels of the S_{pair}

a)

ID	Texapon SB 3 KC	Plantapon ACG 50	Plantapon LC 7	Plantacare 818	Plantacare 2000	Dehyton MC	Dehyton PK 45	Dehyton ML	Dehyton AB 30	Plantapon Amino SCG-L	Plantapon Amino KG-L	Dehyquart A-CA	Luviquat Excellence
1	6.52	0.00	0.00	0.00	0.00	0.00	0.00	8.63	0.00	0.00	0.00	0.00	0.98
2	7.70	0.00	0.00	0.00	0.00	0.00	0.00	8.55	0.00	0.00	0.00	0.00	1.88
3	12.23	0.00	0.00	0.00	0.00	0.00	0.00	10.13	0.00	0.00	0.00	0.00	1.00
4	12.20	0.00	0.00	0.00	0.00	0.00	0.00	13.54	0.00	0.00	0.00	0.00	1.43
5	8.71	0.00	0.00	0.00	0.00	0.00	0.00	9.46	0.00	0.00	0.00	0.00	1.12
6	9.80	0.00	0.00	0.00	0.00	0.00	0.00	8.91	0.00	0.00	0.00	0.00	1.54
7	0.00	0.00	0.00	11.53	0.00	0.00	0.00	0.00	0.00	0.00	0.00	9.53	1.52
8	0.00	0.00	0.00	11.64	0.00	0.00	0.00	0.00	0.00	0.00	0.00	11.79	2.66
9	0.00	0.00	0.00	8.96	0.00	0.00	0.00	0.00	0.00	0.00	0.00	11.93	1.76
10	0.00	0.00	0.00	8.40	0.00	0.00	0.00	0.00	0.00	0.00	0.00	11.21	2.33

vs.

b)

	CS(=O)(=O)[O-]	[Na+]	COC	COC(C)=O	CC(=O)[O-]	[H]N(C)C(C)=O	[H]OC(C)=O	[H]OC	[H]C(C)(OC)OC	CN(C)C	C[N+](C)(C)C	[K+]	[Cl-]	(CH2)x	P	T	y
1	0.53	0.46	0.19	0.30	0.32	0.34	0.00	0.08	0.00	0.37	0.00	0.00	0.00	0.11	0.24	0.53	False
2	0.63	0.50	0.23	0.36	0.34	0.34	0.00	0.08	0.00	0.36	0.00	0.00	0.00	0.18	0.57	0.67	False
3	1.00	0.73	0.36	0.56	0.47	0.40	0.00	0.10	0.00	0.43	0.00	0.00	0.00	0.57	0.25	0.37	False
4	1.00	0.80	0.36	0.56	0.54	0.53	0.00	0.13	0.00	0.57	0.00	0.00	0.00	0.77	0.41	0.81	False
5	0.71	0.57	0.25	0.40	0.38	0.37	0.00	0.09	0.00	0.40	0.00	0.00	0.00	0.30	0.29	0.80	False
6	0.80	0.60	0.29	0.45	0.39	0.35	0.00	0.08	0.00	0.38	0.00	0.00	0.00	0.34	0.45	0.92	False
7	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.43	0.43	0.00	0.38	0.00	0.73	0.62	0.44	0.38	True
8	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.44	0.44	0.00	0.47	0.00	0.90	0.81	0.85	0.40	True
9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.34	0.34	0.00	0.48	0.00	0.91	0.67	0.53	0.54	False
10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.32	0.32	0.00	0.45	0.00	0.85	0.58	0.73	0.79	False

Fig. 2. A comparison of molecular representations of liquid formulations. Rows represent experiment samples and column headers are feature variables. (a) A direct representation of samples by their ingredient concentrations; (b) Formulations are represented by calculated surfactant functional group concentrations, plus polymer and thickener ingredient concentrations as before.

(0) Generate a large random set of N candidate samples, $\mathbf{C}'$ using MaxPro's **`CandPoints`**.

(1) Generate a random starting design of n_{init} samples, $\mathbf{D}'_{\mathbf{init}}$, using MaxPro's **`CandPoints`**. Convert this from the computational to experimental design space, $\mathbf{D}_{\mathbf{init}}$.

(2) Perform the formulation experiments and record the compositions of the prepared samples, $\mathbf{D}$. Convert this from the experimental to computational design space, $\mathbf{D}'$.

(3) ***If*** $\chi < \chi^{\dagger}$ where χ is the proportion of stable formulations prepared and $\chi^{\dagger}$ is a minimum user-defined stability target.

- Train a phase stability classifier on the experimental dataset ($\mathbf{D}$) with the:
  i. Surfactant concentrations converted to functional group concentrations.
  ii. Polymer and thickener concentrations used as is, with both (i) and (ii) pre-processed with min-max scaling.
  iii. Phase stability used as the response, y, to train the classifier.
- Test a variety of machine learning models and select the best performing one to make *in silico* predictions for the phase stability of samples in $\mathbf{C}'$.
- Filter $\mathbf{C}'$ by a phase stability criterion (ϕ) to a restricted candidate set $\mathbf{C}^{*}$ which is a subset of samples with a higher probability of stability. Use this for step 4.

Else:

Pass and use $\mathbf{C}'$as $\mathbf{C}^{*}$for step 4.

(4) Use **`MaxProAugment`** to suggest n_{add} additional experiments ($\mathbf{D}'_{\mathbf{add}}$) to augment the design matrix ($\mathbf{D}'$) by selecting points from $\mathbf{C}^{*}$. Return to step 2 and repeat until sufficient samples are prepared. Then move onto the next (P, T) sub-system.

Scheme 1. An ML-guided DoE algorithm for a formulation sub-system of a fixed polymer, thickener combination to generate a (weighted-) space-filling design using MaxProQQ.

factors with the ingredients to be used. A key **Cand_to_Dataset** function was written to interconvert between the design matrix as compatible with the MaxProQQ package (computational, $\mathbf{X}'$) and Opentrons robot (experimental, $\mathbf{X}$). Here ' denotes the form of a generic matrix compatible with MaxProQQ.

Each of the computational design matrices ($\mathbf{C}'$, $\mathbf{D}'_{\mathbf{init}}$, $\mathbf{D}'$) in Scheme 1 are collections of design vectors, $\boldsymbol{\varphi}'$, with five factors as shown in Eqn. (1). They are generated through two key functions from the MaxPro package, as detailed in Table 1.

As shown in Scheme 1, the algorithm begins by initialising a candidate set, $\mathbf{C}'$, which represents all possible experiments. This set is generated by randomly sampling $N = 360{,}000$ points within the design space. This value of N was chosen to approximate the total number of combinations of surfactants and concentration choices if the concentration ranges provided in Eqn. (1) are discretised by 0.5 w/w% intervals. This was determined to be an appropriate discretisation size that could be comfortably resolved experimentally on an Opentrons robot (Chitre et al., 2024b). Alternatively, a grid-based method could have been employed for a fixed discretisation of the concentration ranges to generate $\mathbf{C}'$. However, random sampling allows for the inclusion of intermediate concentrations, not limited to the 0.5 w/w% intervals. We note that we could accurately determine compositions of the prepared formulations; however, we could not always accurately dispense the target amounts specified from our DoE, especially for the highly viscous formulation ingredients. Therefore, on each iteration of the DoE, we suggested the next batch of experiments based on the actual, recorded compositions.

Table 1
A summary of key MaxPro functions (Ba and Joseph, 2018) utilised in this work.

Function	Description
CandPoints	Generates uniform random numbers for each continuous factor and randomly sampled levels for each nominal factor.
MaxProAugment	Select the best set of design points to augment a given design matrix by optimising the MaxPro criterion sequentially.

A fixed random seed was used to always generate the same candidate set, $\mathbf{C}'$, independent of the polymer, thickener (P, T) sub-system being investigated. The DoE method would, therefore, select points from the same available pool of samples, restricted only according to phase stability predictions (if required) which were sub-system specific.

We started with an initial design for a fixed sub-system of (P, T), $\mathbf{D}'_{\mathbf{init}}$, with n_{init} number of points, where n_{init} was typically set to 36 samples, corresponding to the maximum weekly throughput of the specific experimental workflow. Different starting designs, including a MaxProQQ were possible, and the effect of design initialisation was not investigated further in this work. We analysed the proportion of stable formulations, χ, in the initial design: if $\chi < \chi^{\dagger}$, these sub-systems were

defined as "difficult-to-formulate", in which case we applied a phase-stability guided DoE strategy, as shown in Fig. 3. This ensured many experiments would not be wasted without generating any turbidity or rheology data. Otherwise, we preferred a purely space-filling design for other sub-systems, as this imposes no restriction, allowing better modelling of the entire design space. For this study, $\chi^{\dagger} = 40$ %. This threshold was selected based on the following initial observations and judgement: (i) preliminary results for a couple of sub-systems showed around one-in-three stable formulations, and (ii) if half the samples for a particular sub-system were stable, this was considered a satisfactory formulation. Thus, the threshold was selected as some intermediate value to demonstrate the utility of the weighted part of the algorithm towards regions of feasibility, i.e., phase stability.

For the difficult-to-formulate sub-systems, we would train a predictive phase stability classifier, using a featurisation of the experimental data as explained in Section 2.2. Details for how we trained this classifier are given in the SI. We would then use this classifier to predict the phase stability of each point in our candidate set, $\mathbf{C}'$, and applied a phase stability cut-off ($0 \leq \phi \leq 1$) to drop any points without a minimum probability of stability, to generate a restricted candidate set, $\mathbf{C}^{*}$. This phase-stability cut-off would be modified on each iteration of the DoE as explained in the Results and Discussion section. For the other, more stable sub-systems, we kept the original candidate set.

Finally, given our current experimental dataset, **D**, which was converted to a MaxProQQ-compatible design, $\mathbf{D}'$, we used **MaxProAugment** to select the next batch of experiments from the (restricted) candidate designs. **MaxProAugment** performs a greedy search, optimising the MaxPro criterion by making the locally optimal choice at each stage of selecting $1, \ldots, n$ ($n = 36$) points, with the goal of finding the global optimum. In particular, **MaxProAugment** sequentially evaluates each candidate point, calculating the MaxProQQ criterion (Eqn. SI 1) for the current design augmented with that point. The point that minimises the criterion (or maximises the space-filling property) is added to the design, forming a new, updated design for the next iteration. This is as opposed to performing an expensive optimisation to simultaneously calculate placements of the n best points in the design space. This method gives us a small trade-off in optimality for substantially increased computational efficiency, as we wish to be able to generate designs on-the-fly in a high-dimensional design space for high-throughput experimentation. We would then prepare and characterise the newly suggested batch of experiments and continue in a cycle till we collected a satisfactory number of samples for our dataset generation efforts, typically, around 100+ samples per sub-system.

3. Results and discussion

3.1. Formulating stable products

The method outlined above was for a particular polymer, thickener sub-system. This was applied to all eight sub-systems; however, for the formulations with P_2 (Dehyquart® CC6) too few formulations were stable, so we could not train an accurate stability classifier. We prepared 174 samples over two months with this polymer, but <15 % of these formulations were stable. Since we are not in a truly 'big data' domain, this is too large a class imbalance to train a predictive model. Henceforth, we exclude sub-systems (P_2, T_1) and (P_2, T_2) from results for difficult-to-formulate sub-systems, as we could not apply the weighted-space filling design without a predictive stability classifier. We, therefore, had three sub-systems for which our ML-guided DoE was used, as shown in Fig. 4. For the rest, a purely space-filling MaxProQQ design was used throughout. We highlight, except for (P_4, T_1), the difficult-to-formulate sub-systems were primarily those prepared with the highly charged cationic polyelectrolytes, P_1 and P_2, because these would often form coacervates with the anionic surfactants in the used set of ingredients. For future work, the limitation of a highly class-imbalanced (predominantly unstable) dataset for the P_2 sub-systems could be overcome by training a phase stability classifier on the other sub-systems, with transfer learning utilised to make better stability predictions on the P_2 sub-systems.

Fig. 4 shows that with the phase stability-guided MaxProQQ designs, we could tune experiments to stable regions of the design space across all three sub-systems. The proportion of stable formulations in a batch is coloured with a hue to represent the phase stability cut-off (ϕ). This threshold was progressively increased over each round. Initially, a low threshold is desired to favour unrestricted exploration of the design space. As more experimental data becomes available to train a better stability classifier, we can exploit this model to strongly bias formulations to regions of stability; compare the first and the last points across all three sub-systems in Fig. 4. Note, for the first point, the percentages of 'stable in round' and 'overall stable' are equivalent.

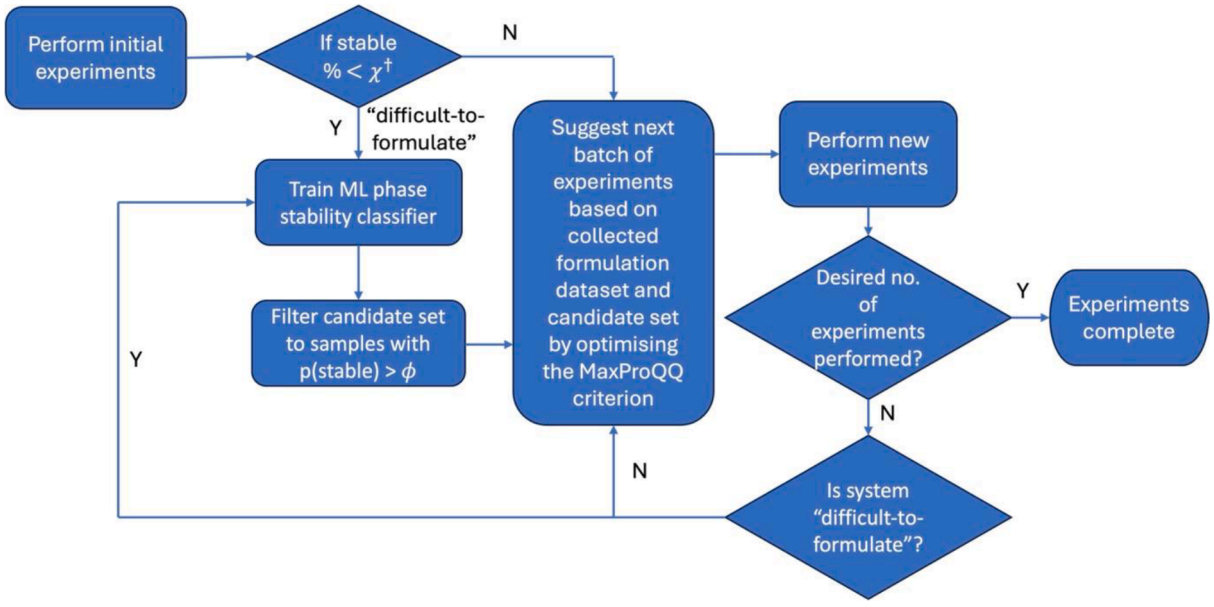

Fig. 3. Flowchart of the design of experiments process. An ML-guided approach is employed for "difficult-to-formulate" (primarily unstable) systems, while a space-filling method is applied directly for more stable systems.

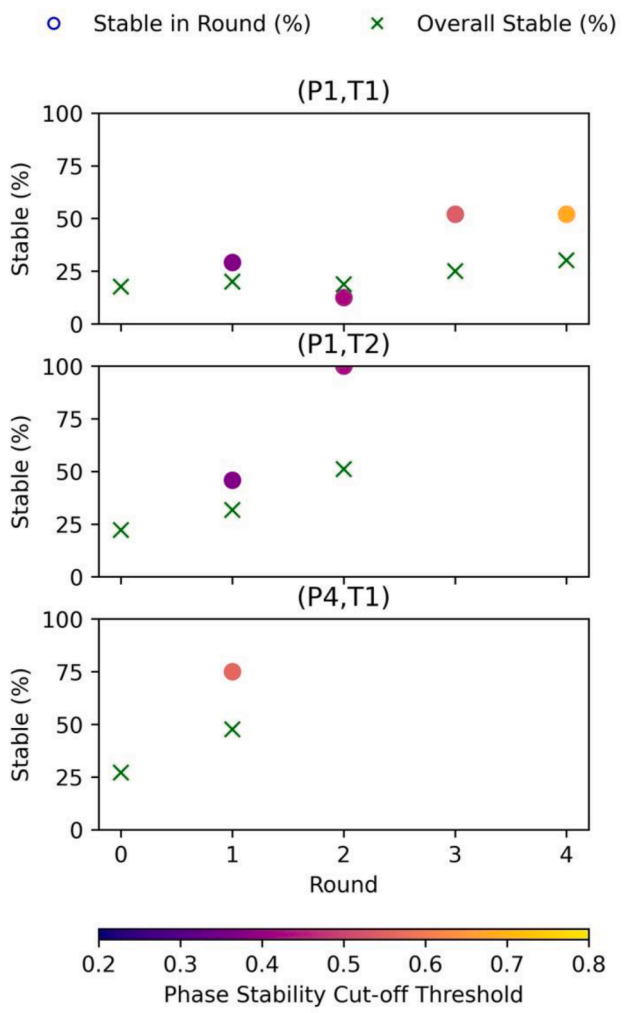

Fig. 4. The ML-guided DoE was used to bias three predominantly unstable sub-systems, where a sub-system is a fixed polymer, thickener combination used in the formulation base. 'Round' corresponds to experimental round, step 2 in Scheme 1. 'Stable in Round' refers to the percentage of stable formulations prepared in a particular round, as opposed to 'Overall Stable' which is the cumulative proportion of stable formulations made. Between experimental rounds, the phase stability cut-off (ϕ) was increased to progressively restrict the design space towards stable predicted regions. Successful application is demonstrated by an increased proportion of stable formulations across rounds.

The phase stability cut-off is used to tune how the design space is restricted. A higher threshold limits the design space to only those regions with a greater probability of phase stability predicted from the classification model. We progressively increased the phase stability cut-off on each iteration of the design cycle, as seen in Table S1 – S3, such that we could illustrate exploiting the predictions of the classifier to increase the proportion of stable samples formulated over each round. Beyond this aim, the cut-off selection was arbitrary.

The only example from Fig. 4 where the ML-guided DoE fails to increase the stable in a round (%) is after the first round of experiments with $(P_1,\ T_1)$. We go from 29 to 13 % of formulations being stable in round 1 vs. 2. However, this can be clearly explained as shown by Table S1 in the SI. Tables S1 - S3 give the full set of phase stability-guided DoE results for all three sub-systems, complementary to Fig. 4. Initially, we chose to apply a cut-off as some top x % of experiments. In round one of $(P_1,\ T_1)$, we restricted the candidate set to the top 20 % of stable predicted experiments; however, this cut-off was equivalent to a 0.29 phase stability threshold, which would still include a majority of unstable formulations, as seen in round two. Hence, we switched to only defining the phase stability cut-off as a predicted probability of stability between zero and one, so we could more clearly control the degree of stability tuning. Additionally, as seen in Table S1, the best classifier at round one had ROC AUC and F_1 scores of 0.62 and 0.73, respectively, which corresponds to a moderately predictive classifier. As highlighted above, it is essential to be able to develop a highly predictive classifier, otherwise, we cannot apply this weighted search strategy effectively. By contrast, the initial classifiers trained for $(P_1,\ T_2)$ and $(P_4,\ T_1)$ are excellent (see Tables S2 and S3), and so we could successfully guide our difficult-to-formulate sub-systems to regions of stability in just one or two iterations.

3.2. Design coverage and spread

The other objective of the ML-guided DoE was to optimally space-fill for future work on developing predictive models over the entire formulation design space. We have fixed the polymer and thickener for a particular sub-system and explored all these sub-systems. Therefore, we look at the spread of surfactants used in Fig. 5. We prepared a total of 384 formulations for the three difficult-to-formulate sub-systems, as identified earlier, and 438 further samples for the remaining five sub-systems. The dashed lines in Fig. 5 show the expected number of samples per surfactant if we had uniformly sampled the ingredients. We observe that for the purely space-filling designs, our surfactants' distribution is very close to this expected value, showing excellent space-filling properties. By comparison, and as we would expect, we have a non-uniform distribution for the stability-guided experiments as our classifier learned that certain surfactant(s) would lead to unstable results with a particular polymer, thickener, or indeed, another surfactant. For example, P_1 and P_2 are highly charged cationic polyelectrolytes, so Texapon® SB 3 KC, Plantapon® ACG 50, and Plantapon® LC7, which are anionic surfactants would often form coacervates with these ingredients and, therefore, they are under-sampled for the stability-guided designs. Following this argument, Dehyquart® A-CA, the only cationic surfactant in the set, was particularly favoured for the stability-guided experiments. Despite this, all the ingredients were relatively well sampled, which was achieved by modifying the phase stability cut-off to start 'relaxed' (low threshold) and successively restricting our experiments to feasible regions of the design space.

Fig. 6 shows coverage of quantitative design variables – ingredient concentrations (w/w%). As stated in Eqn. (1), we aimed for the surfactants to have a concentration distribution between 8 – 13, conditioning polymers 1 – 3, and thickeners 1 – 5, all values in w/w%. We observe a good distribution of concentrations across all ingredients, where the median and interquartile ranges (IQR) are given by the dashed lines on the violin plots in Fig. 6. The shape of a violin plot represents a probability density function (PDF) with a wider section of the PDF showing that the value occurs more frequently, and vice versa. Each violin in Fig. 6 has the same overall area. We can conclude that the full formulation design space has been represented in the generated dataset. We only note that for some ingredients, namely the very viscous ones, we exceeded the suggested concentration bounds for experimental reasons. This is acceptable and still informative towards developing property prediction models.

3.3. Phase stability classifiers and chemical interpretability

We now assess in Fig. 7 the quality of the phase stability classifiers trained over the complete set of experimental data for the three difficult-to-formulate sub-systems. The receiver operating characteristic (ROC) curves in Fig. 7 show performances of the classification models at all different classification thresholds and the area under this curve (ROC AUC) provides an aggregate measure of the classifier's performance. Please see the SI section on evaluating classifier performance for further information on how to read ROC plots, such as Fig. 7. Additionally, we have the class-weighted F_1 scores for the three classifiers. Both metrics range from 0 to 1. While what constitutes a good score may be subject-

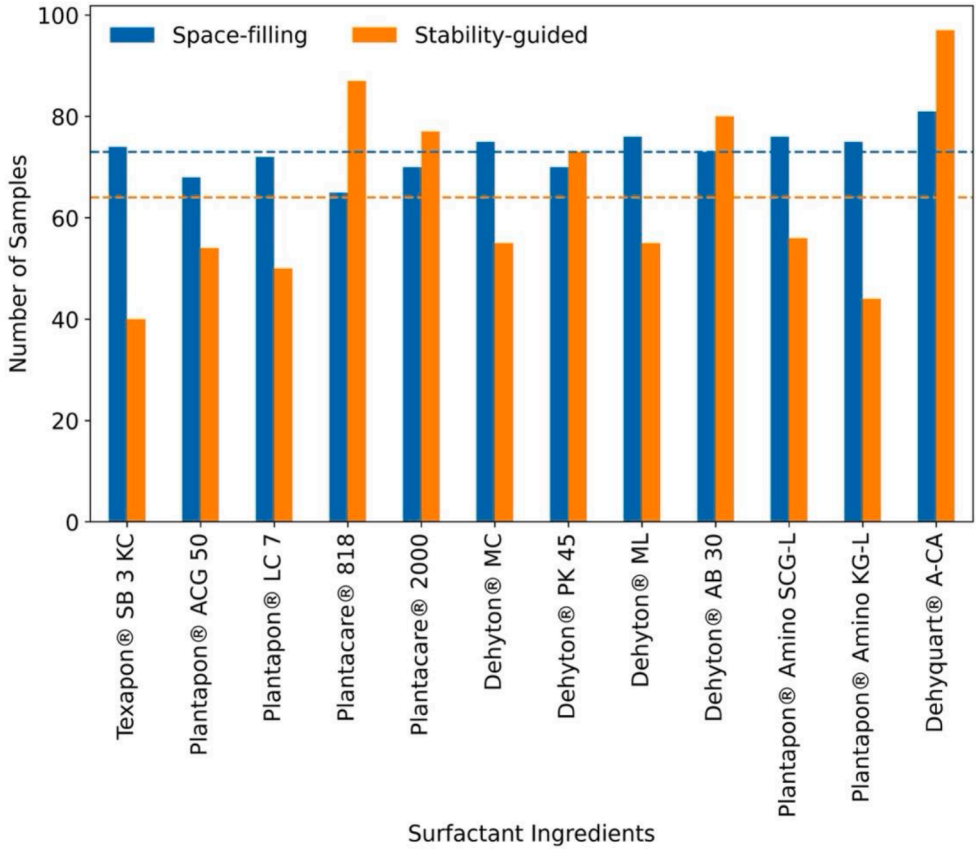

Fig. 5. An illustration of the spread of surfactants across the formulations dataset, subdivided by the purely space-filling designs and stability-guided designs for the difficult-to-formulate sub-systems.

dependent, typically, anything above 0.8 is considered good, and above 0.9 is considered an excellent classifier (Geron, 2019). Given the relatively limited experimental budget and the high dimensionality of this formulations case study, we consider that the trained stability classifiers are highly predictive.

Since we have trained strong phase stability classifiers and used a chemically interpretable representation for the surfactants, we can now draw reliable scientific insights from the results shown in Fig. 8, illustrating feature importances and explanations for the $(P_1,\ T_1)$ sub-system. Similar results for the $(P_1,\ T_2)$ and the $(P_4,\ T_1)$ sub-systems are given in the SI, in Figure S4. The results in Figure S3 show that across all three sub-systems the best performing stability classifier was a random forest. Tree-based classifiers have the beneficial property that feature importances can be computed directly (Breiman, 2001), as shown in Fig. 8a and Figure S4. These results are computed based on the decrease in model performance if a particular feature is removed.

Another popular method in the field of ML interpretability is the use of SHAP (SHapley Additive exPlanations) (Lundberg and Lee, 2017). These values show how each feature affects the final prediction (Lundberg et al., 2020). SHAP is based on the magnitude of feature attributions. Feature importances and SHAP values are different measures, but it is interesting to note that the order of features in both Figs. 8a and 8b are very similar. In both cases, the concentration of thickener is the most important factor governing phase stability, with Fig. 8b showing less thickener is better for preparing stable formulations. These results are directly interpretable for a chemist or a formulator, as we have attributed stability (or instability) to surfactant functional groups, or polymer and thickener concentrations. Furthermore, as shown for an illustrated sample in Figure S5, SHAP can also provide feature attributions on a sample-by-sample basis for a deeper investigation of a formulation's properties. These *a posteriori* analyses can aid in developing novel formulated products. Further discussion linking the chemistry of the functional groups to the phase stability results is out of the scope of this work and will be treated elsewhere.

3.4. Extension to other design problems

While this work has focused entirely on the context of liquid formulation design, the problem discussed is generic. The 'curse of dimensionality' from working in high-dimensional design spaces is a well-known problem in the chemical sciences (Probst and Reymond, 2018; Schrader et al., 2024; Strieth-Kalthoff et al., 2022). Furthermore, many problems are of a mixed-variable nature, e.g., battery material optimisation, catalyst design, pharmaceutical development etc. For a preliminary exploration of complex design spaces, particularly to train non-parametric ML models, space-filling designs are suitable for screening experiments, as discussed in the Introduction. The MaxPro method (in its MaxProQQ form) can generate space-filling designs with an optimal spread in all subspaces of factors for mixed-variable problems (Joseph et al., 2020). Here nominal and continuous variables are considered; however, the method can also handle discrete numeric and ordinal factors.

The key contribution of this study is combining the concept of a weighted design with the application of the MaxPro method to bias designs to feasible regions, which are iteratively predicted based on an ML model trained in an active learning cycle. Feasibility refers to phase stability in this work, and there are examples of other types of stability being of paramount importance for chemical/material development, e.g., oxidative stability in battery electrolytes (Kasnatscheew et al., 2017); crystal stability for inorganic material design (Zhu et al., 2024); thermal stability for thermoelectric materials (Aminorroaya Yamini et al., 2015) etc. Aside from stability, there are other applications, e.g., the

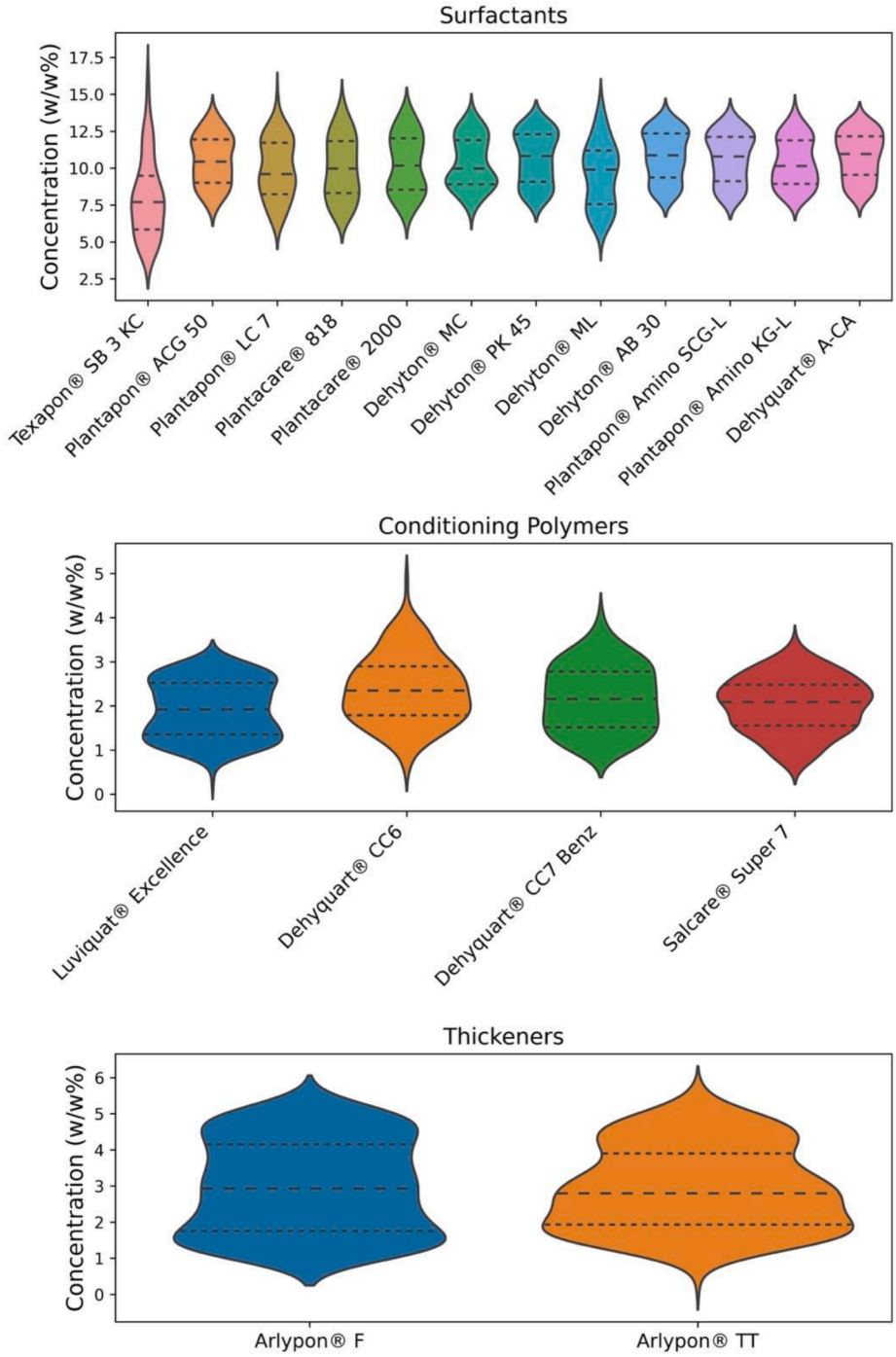

Fig. 6. Distribution of formulation ingredient concentrations for the study's surfactants, conditioning polymers, and thickeners, with target ranges of 8 – 13, 1 – 3, and 1 – 5 w/w%, respectively. In each violin plot, the height represents the concentration range, while the width indicates the frequency of observations. Dashed lines mark the Q_1, Q_2 (median), Q_3 quartiles. Instances where target ranges are exceeded are due to viscous liquid handling challenges in the experimental protocol.

biocompatibility of drug delivery systems (Kohane and Langer, 2010). Without biocompatibility, the rest of the properties are irrelevant, as the drug cannot be safely administered. For such problems, a weighted space-filling design as presented in this work, can be a useful framework for the design of experiments.

4. Conclusions

We developed a weighted-space filling design for liquid formulation based on restricting MaxProQQ designs to stable predicted regions, trained iteratively within an active learning cycle. The liquid formulation problem was decomposed into sub-systems defined by fixed polymer, thickener combinations. Sub-systems yielding fewer than 40 % stable formulations in an initial experimental round (typically a week's experiments) were defined as "difficult-to-formulate" for this study. Out of eight sub-systems, five were predominantly unstable. For three of these challenging sub-systems, we successfully trained highly predictive phase stability classifiers based on a chemically interpretable featurisation of surfactant functional group concentrations. These classifiers helped bias subsequent experiments towards feasible, i.e., stable regions within the design space. The best models trained on the final datasets for these sub-systems achieved ROC AUC scores of 0.85, 0.94, and 0.86,

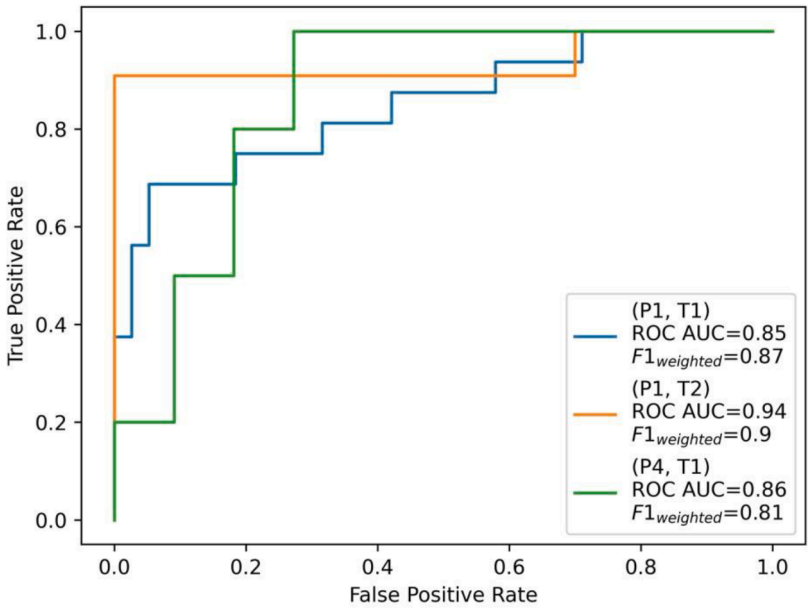

Fig. 7. Receiver operating characteristic (ROC) curves for the best phase stability classifiers for each of the three difficult-to-formulate sub-systems.

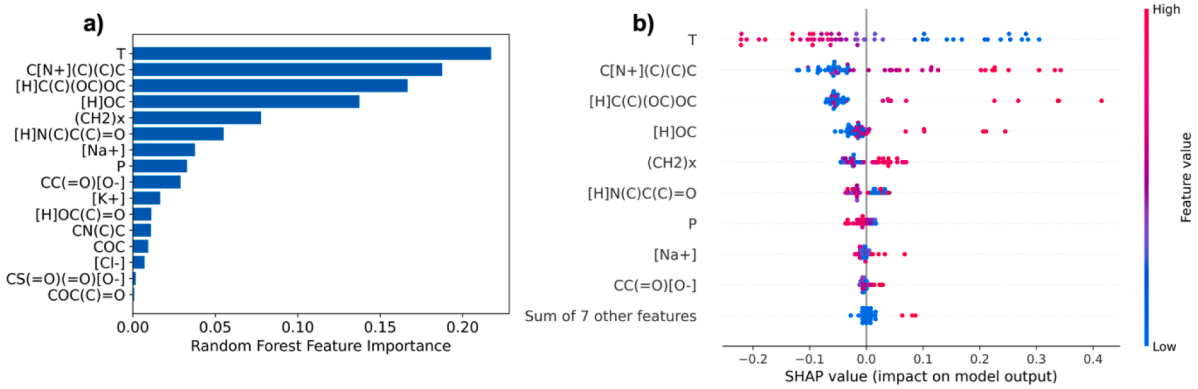

Fig. 8. Two illustrations of physical interpretability of the phase stability classifier for sub-system (P_1, T_1) via (a) random forest feature importances, and (b) SHAP feature explanations.

respectively.

In two other "difficult-to-formulate" sub-systems, both associated with a specific polymer, a significant imbalance existed between stable and unstable formulations (< 15 % stable), precluding the development of a sufficiently predictive model to guide experiments. For these cases, we suggest leveraging transfer learning from other sub-systems/model building on the full dataset to enhance predictive capability in the most challenging cases.

For the remaining sub-systems, a purely space-filling design was adopted to ensure optimal spread across the design space, as guiding stability optimisation was unnecessary. We discuss that even for the weighted designs, good space-filling properties are maintained through a user-defined phase stability cut-off in the method. This is progressively increased over iterations to maintain a balance between initially exploring the complete design space and later narrowing the sampling to feasible regions. The satisfactory coverage of the stability-guided designs is evidenced by each ingredient being sampled to at least 50 % of its expected frequency from uniform sampling. At the same time, the model learns which are unfavourable ingredient interactions for formulating stable products and under-samples these purposefully.

Overall, all the selected ingredients were tested and explored within the complete range of target concentrations specified by our industrial partner. The presented approach drove the collection of a dataset of over 800 formulations, including nearly 300 stable samples for which additional turbidity and viscosity characterisation was performed, as detailed in the published dataset, elsewhere. The presented methodology has been effective in a high-dimensional, mixed-variable design space where there is a principal property of interest and as such can be applied to other design problems with these characteristics in the chemical/material sciences.

CRediT authorship contribution statement

Aniket Chitre: Writing – original draft, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Daria Semochkina:** Writing – review & editing. **David C. Woods:** Writing – review & editing, Supervision, Conceptualization. **Alexei A. Lapkin:** Writing – review & editing, Supervision, Project administration, Funding acquisition, Conceptualization.

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.

Acknowledgements

The project was co-funded by UKRI Program Grant Chembots: Digital-Chemical-Robotics to Convert Code to Molecules and Complex Systems (EP/S019472). AC is grateful to BASF for co-funding his PhD studentship. The project was co-funded by 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). Experiments for this study were performed at IMRE, A*STAR in the group of Prof. Kedar Hippalgaonkar; experimental work is described in detail elsewhere.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.compchemeng.2025.109007.

Data availability

Data will be made available on request.

References

Abooali, D., Soleimani, R., 2023. Structure-based modeling of critical micelle concentration (CMC) of anionic surfactants in brine using intelligent methods. Sci Rep 13, 13361. https://doi.org/10.1038/s41598-023-40466-1.

Ahneman, D.T., Estrada, J.G., Lin, S., Dreher, S.D., Doyle, A.G., 2018. Predicting reaction performance in C–N cross-coupling using machine learning. Science 360, 186–190. https://doi.org/10.1126/science.aar5169.

Aminorroaya Yamini, S., Brewis, M., Byrnes, J., Santos, R., Manettas, A., Pei, Y.Z., 2015. Fabrication of thermoelectric materials – thermal stability and repeatability of achieved efficiencies. J. Mater. Chem. C 3, 10610–10615. https://doi.org/10.1039/C5TC02210J.

Ba, S., Joseph, V.R., 2018. MaxPro: maximum Projection Designs.

Ba, S., Myers, W.R., Brenneman, W.A., 2015. Optimal Sliced Latin Hypercube Designs. Technometrics 57, 479–487. https://doi.org/10.1080/00401706.2014.957867.

Bagajewicz, M., Hill, S., Robben, A., Lopez, H., Sanders, M., Sposato, E., Baade, C., Manora, S., Hey Coradin, J., 2011. Product design in price-competitive markets: a case study of a skin moisturizing lotion. AIChE J 57, 160–177. https://doi.org/10.1002/aic.12242.

Bernardo, F.P., Saraiva, P.M., 2005. Integrated process and product design optimization: a cosmetic emulsion application. Computer Aided Chemical Engineering. Elsevier, pp. 1507–1512.

Bowman, V.E., Woods, D.C., 2013. Weighted space-filling designs. J. Simul. 7, 249–263. https://doi.org/10.1057/jos.2013.8.

Bray, S.A., Lucas, X., Kumar, A., Grüning, B.A., 2020. The ChemicalToolbox: reproducible, user-friendly cheminformatics analysis on the Galaxy platform. J Cheminform 12, 40. https://doi.org/10.1186/s13321-020-00442-7.

Breiman, L., 2001. Random Forests. Mach Learn 45, 5–32. https://doi.org/10.1023/A:1010933404324.

Cao, L., Russo, D., Felton, K., Salley, D., Sharma, A., Keenan, G., Mauer, W., Gao, H., Cronin, L., Lapkin, A.A., 2021a. Optimization of formulations using robotic experiments driven by machine learning DoE. Cell Rep. Phys. Sci. 2, 100295.

Cao, L., Russo, D., Lapkin, A.A., 2021b. Automated robotic platforms in design and development of formulations. AIChE J 67, e17248.

Cao, L., Russo, D., Matthews, E., Lapkin, A., Woods, D., 2023. Computer-aided design of formulated products: a bridge design of experiments for ingredient selection. Comput. Chem. Eng. 169, 108083. https://doi.org/10.1016/j.compchemeng.2022.108083.

Chitre, A., Querimit, R.C.M., Rihm, S.D., Dogancan, K., Zhu, B., Wang, K., Wang, L., Hippalgaonkar, K., Lapkin, A.A., 2024a. Accelerating formulation design via machine learning: generating a high-throughput shampoo formulations dataset. https://doi.org/10.6084/m9.figshare.c.7132624.v1.

Chitre, A., Querimit, R.C.M., Rihm, S.D., Karan, D., Zhu, B., Wang, K., Wang, L., Hippalgaonkar, K., Lapkin, A.A., 2024b. Accelerating formulation design via machine learning: generating a high-throughput shampoo formulations dataset. Sci Data 11, 1–10. https://doi.org/10.1038/s41597-024-03573-w.

Consonni, V., Todeschini, R., 2010. Molecular Descriptors. In: Puzyn, T., Leszczynski, J., Cronin, M.T. (Eds.), Recent Advances in QSAR Studies. Springer, Netherlands, Dordrecht, pp. 29–102.

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.

Deng, X., Hung, Y., Lin, C.D., 2015. Design for computer experiments with qualitative and quantitative factors. STAT SINICA. https://doi.org/10.5705/ss.2013.388.

Ertl, P., 2017. An algorithm to identify functional groups in organic molecules. J Cheminform 9, 1–7. https://doi.org/10.1186/s13321-017-0225-z.

Galvanin, F., Barolo, M., Pannocchia, G., Bezzo, F., 2012. Online model-based redesign of experiments with erratic models: a disturbance estimation approach. Comput. Chem. Eng. 42, 138–151.

Galvanin, F., Macchietto, S., Bezzo, F., 2007. Model-based design of parallel experiments. Ind. Eng. Chem. Research 46, 871–882.

Gani, R., 2004. Chemical product design: challenges and opportunities. Comput. Chem. Eng. 28, 2441–2457. https://doi.org/10.1016/j.compchemeng.2004.08.010.

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.

Garud, S.S., Karimi, I.A., Kraft, M., 2017. Design of computer experiments: a review. Comput Chem Eng 106, 71–95. https://doi.org/10.1016/j.compchemeng.2017.05.010.

Geron, A., 2019. Hands-On Machine Learning With Scikit-Learn, 2nd ed. Keras & TensorFlow, O'Reilly.

Ghiringhelli, L.M., Vybiral, J., Levchenko, S.V., Draxl, C., Scheffler, M., 2015. Big data of materials science: critical role of the descriptor. Phys. Rev. Lett. 114, 105503. https://doi.org/10.1103/PhysRevLett.114.105503.

Glavatskikh, M., Leguy, J., Hunault, G., Cauchy, T., Da Mota, B., 2019. Dataset's chemical diversity limits the generalizability of machine learning predictions. J Cheminform 11, 69. https://doi.org/10.1186/s13321-019-0391-2.

Huang, Y., Gilmour, S.G., Mylona, K., Goos, P., 2019. Optimal design of experiments for non-linear response surface models. J. Royal Statist. Soc. Series C 68, 623–640. https://doi.org/10.1111/rssc.12313.

Jessop, P.G., Ahmadpour, F., Buczynski, M.A., Burns, T.J., Green II, N.B., Korwin, R., Long, D., Massad, S.K., Manley, J.B., Omidbakhsh, N., Pearl, R., Pereira, S., Predale, R.A., Sliva, P.G., VanderBilt, H., Weller, S., Wolf, M.H., 2015. Opportunities for greener alternatives in chemical formulations. Green Chem 17, 2664–2678. https://doi.org/10.1039/C4GC02261K.

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.

Joseph, V.R., 2016. Space-filling designs for computer experiments: a review. Qual. Eng. 28, 28–35. https://doi.org/10.1080/08982112.2015.1100447.

Joseph, V.R., Dasgupta, T., Tuo, R., Wu, C.F.J., 2015. Sequential exploration of complex surfaces using minimum energy designs. Technometrics 57, 64–74. https://doi.org/10.1080/00401706.2014.881749.

Joseph, V.R., Gul, E., Ba, S., 2020. Designing computer experiments with multiple types of factors: the MaxPro approach. J. Qual. Technol. 52, 343–354. https://doi.org/10.1080/00224065.2019.1611351.

Kasnatscheew, J., Streipert, B., Röser, S., Wagner, R., Cekic Laskovic, I., Winter, M., 2017. Determining oxidative stability of battery electrolytes: validity of common electrochemical stability window (ESW) data and alternative strategies. PCCP 19, 16078–16086. https://doi.org/10.1039/C7CP03072J.

Kelly, C.L., 2023. Addressing the sustainability challenges for polymers in liquid formulations. Chem. Sci. 14, 6820–6825.

Kim, C., Chandrasekaran, A., Huan, T.D., Das, D., Ramprasad, R., 2018. Polymer genome: a data-powered polymer informatics platform for property predictions. J. Phys. Chem. C 122, 17575–17585. https://doi.org/10.1021/acs.jpcc.8b02913.

Kohane, D.S., Langer, R., 2010. Biocompatibility and drug delivery systems. Chem. Sci. 1, 441–446. https://doi.org/10.1039/C0SC00203H.

Kronberg, B., Holmberg, K., Lindman, B., 2014. Surface Chemistry of Surfactants and Polymers. John Wiley & Sons, Ltd, Chichester, UK.

Kuenneth, C., Ramprasad, R., 2023. polyBERT: a chemical language model to enable fully machine-driven ultrafast polymer informatics. Nat Commun 14, 4099. https://doi.org/10.1038/s41467-023-39868-6.

Lin, T.-S., Coley, C.W., Mochigase, H., Beech, H.K., Wang, W., Wang, Z., Woods, E., Craig, S.L., Johnson, J.A., Kalow, J.A., Jensen, K.F., Olsen, B.D., 2019. BigSMILES: a structurally-based line notation for describing macromolecules. ACS Cent. Sci. 5, 1523–1531. https://doi.org/10.1021/acscentsci.9b00476.

Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., Lee, S.-I., 2020. From local explanations to global understanding with explainable AI for trees. Nat Mach Intell 2, 56–67. https://doi.org/10.1038/s42256-019-0138-9.

Lundberg, S.M., Lee, S.-I., 2017. A Unified Approach to Interpreting Model Predictions. Presented at the 31st Conference on Neural Information Processing Systems, Long Beach CA, USA, pp. 1–10.

Martín, M., Martínez, A., 2013. A methodology for simultaneous process and product design in the formulated consumer products industry: the case study of the detergent business. Chem. Eng. Res. Des. 91, 795–809. https://doi.org/10.1016/j.cherd.2012.08.012.

McKay, M.D., Beckman, R.J., Conover, W.J., 1979. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. Technometrics 21, 239–245.

Moriwaki, H., Tian, Y.-S., Kawashita, N., Takagi, T., 2018. Mordred: a molecular descriptor calculator. J Cheminform 10, 1–14.

Morris, M.D., Mitchell, T.J., 1995. Exploratory designs for computational experiments. J. Stat. Plan. Inference 43, 381–402.

Narayanan, H., Dingfelder, F., Condado Morales, I., Patel, B., Heding, K.E., Bjelke, J.R., Egebjerg, T., Butté, A., Sokolov, M., Lorenzen, N., Arosio, P., 2021. Design of biopharmaceutical formulations accelerated by machine learning. Mol. Pharmaceutics 18, 3843–3853. https://doi.org/10.1021/acs.molpharmaceut.1c00469.

O'Boyle, N.M., Banck, M., James, C., Vandermeersch, T., Hutchnison, G., 2011. Open Babel: an open chemical toolbox. J Cheminform 3, 14.

Öztürk, H., Ozkirimli, E., Özgür, A., 2016. A comparative study of SMILES-based compound similarity functions for drug-target interaction prediction. BMC Bioinform 17, 128. https://doi.org/10.1186/s12859-016-0977-x.
Pattanaik, L., Coley, C.W., 2020. Molecular representation: going long on fingerprints. Chem 6, 1204–1207.
Probst, D., Reymond, J.-L., 2018. A probabilistic molecular fingerprint for big data settings. J Cheminform 10, 66. https://doi.org/10.1186/s13321-018-0321-8.
Qian, P.Z.G., 2012. Sliced Latin Hypercube Designs. J. Am. Stat. Assoc. 107, 393–399. https://doi.org/10.1080/01621459.2011.644132.
Qin, S., Jin, T., Van Lehn, R.C., Zavala, V.M., 2021. Predicting critical micelle concentrations for surfactants using graph convolutional neural networks. J. Phys. Chem. B 125, 10610–10620. https://doi.org/10.1021/acs.jpcb.1c05264.
Schrader, M.L., Schäfer, F.R., Schäfers, F., Glorius, F., 2024. Bridging the information gap in organic chemical reactions. Nat. Chem. 16, 491–498. https://doi.org/10.1038/s41557-024-01470-8.
Schwartz, J., Awale, M., Reymond, J.-L., 2013. SMIfp (SMILES fingerprint) chemical space for virtual screening and visualization of large databases of organic molecules. J. Chem. Inf. Model. 53, 1979–1989. https://doi.org/10.1021/ci400206h.
Seddon, D., Müller, E.A., Cabral, J.T., 2022. Machine learning hybrid approach for the prediction of surface tension profiles of hydrocarbon surfactants in aqueous solution. Colloid J 625, 328–339. https://doi.org/10.1016/j.jcis.2022.06.034.
Shewry, M.C., Wynn, H.P., 1987. Maximum entropy sampling. J Appl Stat 14, 165–170. https://doi.org/10.1080/02664768700000020.
Sobol, I.M., 1967. On the distribution of points in a cube and the approximate evaluation of integrals. Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki 7, 784–802. https://doi.org/10.1016/0041-5553(67)90144-9.
Strieth-Kalthoff, F., Sandfort, F., Kühnemund, M., Schäfer, F.R., Kuchen, H., Glorius, F., 2022. Machine learning for chemical reactivity: the importance of failed experiments. Angew Chem Int Ed 61, e202204647. https://doi.org/10.1002/anie.202204647.
Taifouris, M., Martín, M., Martínez, A., Esquejo, N., 2020. Challenges in the design of formulated products: multiscale process and product design. Curr. Opin. Chem. Eng. 27, 1–9. https://doi.org/10.1016/j.coche.2019.10.001.
Vidal, D., Thormann, M., Pons, M., 2005. LINGO, an efficient holographic text based method to calculate biophysical properties and intermolecular similarities. J. Chem. Inf. Model. 45, 386–393. https://doi.org/10.1021/ci0496797.
Wigh, D.S., Goodman, J.M., Lapkin, A.A., 2022. A review of molecular representation in the age of machine learning. WIREs Comput Mol Sci 12, 1–19.
Yang, K., Swanson, K., Jin, W., Coley, C., Eiden, P., Gao, H., Guzman-Perez, A., Hopper, T., Kelley, B., Mathea, M., Palmer, A., Settels, V., Jaakkola, T., Jensen, K., Barzilay, R., 2019. Analyzing Learned Molecular Representations for Property Prediction. J. Chem. Inf. Model. 59, 3370–3388. https://doi.org/10.1021/acs.jcim.9b00237.
Yap, C.W., 2011. PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. J. Comput. Chem. 32, 1466–1474. https://doi.org/10.1002/jcc.21707.
Zhou, W., Yang, J., Liu, M.-Q., 2021. Construction of orthogonal marginally coupled designs. Stat Papers 62, 1795–1820. https://doi.org/10.1007/s00362-019-01156-1.
Zhu, R., Nong, W., Yamazaki, S., Hippalgaonkar, K., 2024. WyCryst: wyckoff inorganic crystal generator framework. Matter 7, 3469–3488.