

Benchmarking model-based design of experiment approaches with a pharmaceutical crystallisation emulator

Thomas Pickles

University of Strathclyde

Kakasaheb Nandiwale

Pfizer Worldwide Research and Development

Bhavik A. Mehta

Siemens Industry Software Limited

Sammy Sambu

UCB Pharma (Belgium)

Jonathan Moores

University of Strathclyde

Alastair J. Florence

University of Strathclyde

Cameron J. Brown

`cameron.brown.100@strath.ac.uk`

University of Strathclyde

Article

Keywords:

Posted Date: September 18th, 2025

DOI: <https://doi.org/10.21203/rs.3.rs-7045911/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Title: Benchmarking model-based design of experiment approaches with a pharmaceutical crystallisation emulator

Journal: “Self-Driving Laboratories for Chemistry and Materials Science” collection in *npj Computational Materials*

Authors: Thomas Pickles^a, Kakasaheb Nandiwale^b, Bhavik A. Mehta^c, Sammy Sambu^d, Jonathan Moores^a, Alastair J. Florence^{a, e}, Cameron J. Brown^{a, e*}

Affiliations: ^a Strathclyde Institute of Pharmacy and Biomedical Sciences, University of Strathclyde, Glasgow G4 0RE, UK, ^b Chemical Research and Development, Pfizer Worldwide Research and Development, Groton, Connecticut 06340, USA, ^c Siemens Industry Software Limited, Process Automation Software, 26-28 Hammersmith Grove, London, U.K, ^d UCB Biopharma SRL, Allée de la Recherche 60, B – 1070 Brussels, Belgium, ^e EPSRC MediForge Hub, CMAC, University of Strathclyde, Glasgow, G1 1RD, UK.

Email: *cameron.brown.100@strath.ac.uk

Abstract

This study presents a benchmarking framework for model-based design of experiments (MB-DoE) in pharmaceutical crystallisation using an in-silico emulator. The emulator simulates a generic cooling-antisolvent crystallisation underpinned by a mechanistic population balance model (PBM) calibrated with experimental data from the self-driving Scale-Up CMC DataFactory. The PBM generated over 20,000 synthetic experiments to train a random forest emulator with an error < 1% which served as the benchmark function for evaluating optimisation strategies. Twelve initial design strategies were assessed, combining Latin hypercube sampling (LHS), Sobol sequences, and random sampling with varying sample sizes. These were followed by Bayesian optimisation using a Gaussian process (GP) and six acquisition functions: Mean, Random Sampling, Space Filling, Expected Hypervolume Improvement (EHVI), Noisy EHVI (NEHVI), and Pareto efficient global optimisation (NParEGO) leading to 360 emulated experimental campaigns. Performance was assessed using the hypervolume metric. Sobol outperformed other methods at low sample counts, while LHS showed consistent improvement with larger sample sizes. EHVI consistently delivered the highest hypervolume values, indicating effective identification and convergence towards optimal crystallisation conditions. The emulator was also demonstrated on a non-standard optimisation method using gradient boosting. However, the GP with expected improvement outperformed gradient boosting by between 6 and 12 % across differing initial conditions. These results highlight the emulator’s utility in revealing performance differences between optimisation approaches. The benchmark function demonstrated

in this work represents a novel, adaptable and scalable framework for systematically evaluating new and existing MB-DoE strategies under realistic noisy pharmaceutical conditions.

1. Introduction

Crystallisation is a cornerstone process in the pharmaceutical industry, playing a pivotal role in defining the critical quality attributes (CQAs) of active pharmaceutical ingredients (APIs), such as polymorphism, purity, and particle size distribution^{1, 2}. These attributes directly influence bioavailability, stability, and manufacturability, making crystallisation a critical target for both scientific investigation and process optimisation³. However, crystallisation remains a notoriously complex and poorly understood phenomenon, governed by a rich interplay of thermodynamic and kinetic parameters that span multiple length and time scales⁴.

Recent advances in automation, high-throughput experimentation, and machine learning are poised to revolutionise the discovery and optimisation of chemical processes, including crystallisation⁵⁻⁸. Self-driving laboratories (SDLs)⁹⁻¹² are experimental platforms that integrate robotics, real-time analytics, machine learning and computational orchestration and control. They offer a promising route to explore high-dimensional design spaces more efficiently than traditional trial-and-error approaches¹³. These autonomous systems can learn from experimental feedback to adaptively propose new conditions, accelerating the discovery of optimal processing regimes and unveiling novel structure–property relationships.

A crucial component underpinning the utility of self-driving laboratories is the optimisation algorithms responsible for navigating complex experimental design spaces¹². These algorithms, including Bayesian optimisation, genetic algorithms, reinforcement learning, and surrogate modelling techniques, can efficiently balance exploration and exploitation to identify optimal experimental conditions within constrained resource and time budgets^{14, 15}. To ensure robustness, scalability, and generalisability, such algorithms are often evaluated against synthetic benchmark functions that provide known optima and controlled complexity¹⁶. Common test suites, such as the Branin, Hartmann, and Rosenbrock functions, enable systematic comparisons across different optimisation strategies and offer insights into algorithmic strengths and limitations in high-dimensional or multimodal landscapes^{17, 18}.

While SDLs have shown impressive results in fields such as materials discovery¹⁹, organic synthesis²⁰, and perovskite catalyst design^{21, 22}, their application to pharmaceutical crystallisation is comparatively underexplored²³. This is partly due to the unique challenges presented by crystallisation systems, such as the stochasticity of nucleation, the complex thermodynamics of solubility, and the need for tight integration

of process modelling with experimentation²⁴. Furthermore, whilst standard benchmark functions are invaluable for initial testing and development, translating algorithmic performance from idealised benchmarks to noisy, dynamic, and experimentally constrained systems such as pharmaceutical crystallisation remains a significant challenge²⁵. This highlights the need for domain-specific benchmarking²⁶ platforms that more faithfully reflect real-world process characteristics, variability, and constraints.

To address these challenges, this work introduces an emulator demonstrated through a series of in-silico model-based design of experiment (MB-DoE) campaigns for the pharmaceutical crystallisation of lovastatin²⁷. This emulator integrates mechanistic process modelling and simulation with data-driven prediction and optimisation in a SDL context. While mathematical modelling of crystallisation systems remains challenging, population balance models (PBMs) provide a mechanistic framework for predicting particle CQAs by explicitly accounting for key crystallisation mechanisms such as nucleation, growth, and dissolution kinetics²⁸. Using gPROMS FormulatedProducts (Siemens), an advanced modelling suite with dedicated pharmaceutical modules²⁹, we develop and calibrate a PBM to simulate lovastatin crystallisation across a wide design space. These virtual experiments enable the generation of synthetic datasets that expand the available training data, reduce the reliance on physical experimentation, and serve as a benchmark for evaluating MB-DoE strategies within the SDL community. This approach offers a scalable and realistic platform for testing optimisation algorithms in complex and experimentally constrained environments.

2. Methods

The methods employed in this study are summarised in Figure 1, which outlines the overall workflow. The process begins with a set of crystallisation experiments used to calibrate a population balance model (PBM). This calibrated PBM is then used to generate a large synthetic dataset, which serves as the training set for a random forest model that acts as an emulator of the underlying process. This emulator functions as a benchmark tool and forms a central focus of the study. To demonstrate its application, we present a comparative design of experiment case study that explores the influence of different initial conditions, modelling approaches and acquisition functions. This section is intended to be flexible, allowing users to apply their own models while leveraging the benchmark function for consistent comparison.

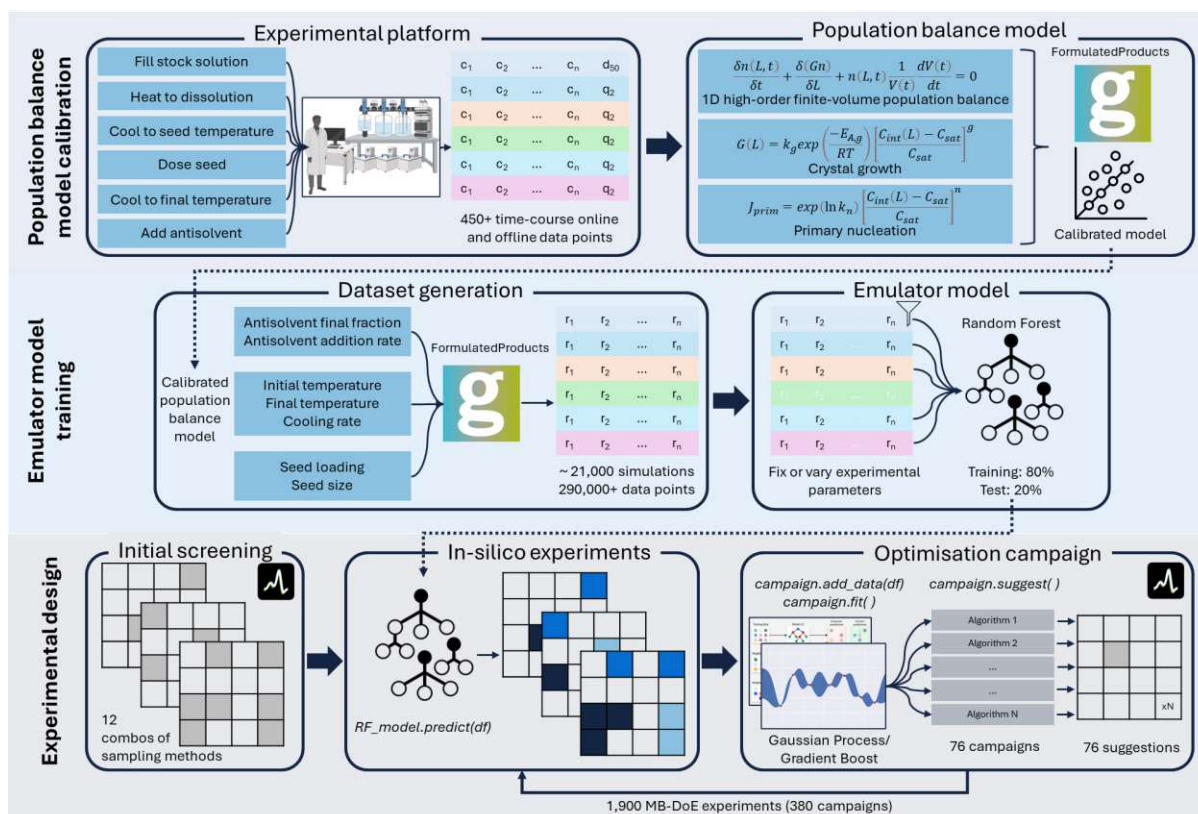


Figure 1. Overview of the logical flow of the methods.

2.1. Self-Driving Crystallisation Lab

Experimental Platform

All experiments were conducted using the Scale-up CMC DataFactory™³⁰, a modular multi-vessel platform which in this configuration comprised of seed slurry, antisolvent and 1 L crystallisation vessels. Each vessel was integrated with pumps, transfer systems and heater-chiller units and instrumented with single-board IoT devices to enable bidirectional data communication for control and data logging. Unit operations were orchestrated using LabOS™, a digital laboratory environment providing electronic lab notebook (eLN) capabilities, real-time control, live data visualisation dashboards and 2D data storage. Process analytical technology (PAT) instrumentation included in situ temperature probes, a Blaze Micro probe (Blaze Metrics) and high-performance liquid chromatography (HPLC) (Shimadzu). Offline particle characterisation was performed using a Morphologi 4 (Malvern Panalytical).

Crystallisation Experiments

A model-based design of experiments (ESI Section 2.1.) was applied to the crystallisation of lovastatin (Molekula) from an acetone (VWR): water system, using thermodynamic

solubility data obtained from the literature³¹. A total of 16 experiments were executed following a standardised reaction procedure:

1. Initialise crystallisation vessel with stock solution at 250 RPM
2. Heat to dissolution at 45 °C
3. Cool to the target Seed Supersaturation (SS) (*continuous: 1.2 to 1.5*)
4. Introduce seed at the specified Seed Load (*ordinal: 0, 1 or 5%*) and Seed Size (*categorical: small or large*)
5. Cool to 10 °C at the assigned Cooling Rate (*ordinal: 0, 0.25, or 0.5 °C/min*)
6. Add antisolvent at the specified Flow Rate (*ordinal: 3, 6, or 10 mL/min*)

In experiments with Seed Load = 0, the process involved direct cooling from dissolution without seeding. For Cooling Rate = 0 °C/min, the crystallisation vessel was held isothermally for 1 hour at the seeding temperature without antisolvent addition, as the process never reached 10 °C.

Data Extraction

Process data included in situ liquor concentration profiles (via HPLC), real-time particle size percentiles from the Blaze probe and full particle size distributions (PSDs) of isolated solids obtained post-process using Morphologi analysis. This dataset served as the input in parameter estimation and fitting of the population balance model in gPROMS.

2.2. Population Balance Model & Simulation

Mathematical Model

To capture the behaviour of the lovastatin crystallisation process, a fit for purpose 1D population balance model (PBM) was developed in FormulatedProducts 2025.1.0 using the well mixed MSMPR Crystalliser model, modified for a fed-batch process. The population balance equation describing the evolution of particle size distribution is given by:

$$\frac{\delta n(L, t)}{\delta t} + \frac{\delta(Gn)}{\delta L} + n(L, t) \frac{1}{V(t)} \frac{dV(t)}{dt} = 0 \quad \text{Eq 1}$$

Crystal growth was described as a two-step phenomenon comprising diffusion and integration. The involved mass transfer and growth kinetics were modelled as follows:

$$G(L) = k_d(L) \left[\frac{C_{bulk} - C_{int}(L)}{\rho_{crys}} \right] \quad \text{Eq 2}$$

where k_d is the mass-transfer coefficient, ρ_{crys} is the lovastatin density, and C_{bulk} and C_{int} are the concentrations of lovastatin in the bulk solution phase and at the solid–liquid interface of the crystal and solution, respectively. k_d is determined based on the crystal size and mixing conditions as well as the dimensionless Sherwood number as follows:

$$k_d(L) = \frac{D_{AB}}{L} \left[2 + 0.8 \left(\frac{\bar{\epsilon} L^4}{\nu_L^3} \right)^{1/5} \left(\frac{\nu_L}{D_{AB}} \right)^{1/3} \right] \quad \text{Eq 3}$$

where D_{AB} is the diffusion coefficient of lovastatin in the bulk solution phase, $\bar{\epsilon}$ is the average energy dissipation in the reactor system, and ν_L is the kinematic viscosity of the bulk liquid phase. The diffusion coefficient can be calculated using the modified Stokes–Einstein equation as follows:

$$D_{AB} = \alpha \frac{kT}{3\pi\eta d_m} \quad \text{Eq 4}$$

where α is the correction coefficient, fitted during model calibration, k is the Boltzmann constant, T is the solution temperature, η is the dynamic viscosity of the solution, and d_m is the diameter of the lovastatin molecule.

The surface integration step was modelled as:

$$G(L) = k_g \exp\left(\frac{-E_{A,g}}{RT}\right) \left[\frac{C_{int}(L) - C_{sat}}{C_{sat}} \right]^g \quad \text{Eq 5}$$

where, k_g , $E_{A,g}$ and g , are the growth rate coefficient, activation energy and supersaturation dependency, respectively, and are fitted during model calibration. R is the gas constant, and C_{sat} is the saturation concentration of lovastatin in the bulk solution phase.

Primary nucleation was described through an empirical power law expression:

$$J_{prim} = \exp(\ln k_n) \left[\frac{C_{int}(L) - C_{sat}}{C_{sat}} \right]^n \quad \text{Eq 6}$$

where, k_n , and n , are the nucleation rate coefficient, and supersaturation dependency, respectively, and are fitted during model calibration.

To complete the mass balance, the solubility of lovastatin in the bulk solution was modelled through the Van't-Hoff Jouyban-Acree model³² for binary systems:

$$\ln x_{sat} = \sum_{p \in \{1,2\}} \left(\psi_{0,p} w_p + \psi_{1,p} \frac{w_p}{T^*} \right) + \sum_{p \in \{1,2\}; p \neq q} \frac{w_p w_q}{T^*} \left[\psi_{2,p,q} + \psi_{3,p,q} (w_p - w_q) + \psi_{4,p,q} (w_p - w_q)^2 \right] \quad \text{Eq 7}$$

where x_{sat} is the mole fraction solubility at temperature T^* [K/1000], w is the weight fraction (solute-free basis) of the solvent component p or q , and ψ are model parameters.

The system of equations was solved in FormulatedProducts 2025.1.0 using a high-order finite-volume method. Particle size distributions were segmented into 50 bins between 0.1 and 1500 μm .

Model calibration

Solubility:

Solubility curve calibration was performed using MATLAB's (R2024B) Curve Fitting Toolbox with a non-linear least-squares regression method. A custom surface fit equation was implemented of equation 7. The fitting process employed the trust-region reflective algorithms. Goodness-of-fit was evaluated using the coefficient of determination (R^2), root mean squared error (RMSE), and residual plots. Confidence intervals for the model parameters were also obtained

Population balance model:

The model was calibrated by using the data generated from the experimental platform (lovastatin concentration in bulk solution and offline particle sizes), in a sequential manner³³:

1. Seeded experiments with a high seed loading were used to calibrate the model parameters assuming only growth as a mechanism.
2. All seeded experiments (high and low loading) were used to calibrate the model parameters assuming secondary nucleation and growth as active mechanisms. Growth model parameters were bounded by the results in step 1.
3. All experiments (seeded and unseeded) were used to calibrate the model parameters assuming primary nucleation, secondary nucleation and growth as active mechanisms. Growth and secondary nucleation parameters were bounded by the results in step 2.

In each step only the lovastatin concentration data with a linear variance (relative term = 0.05 and constant term = 0.0005 kg/kg) was used as measured data to calibrate the model.

Model calibration was carried out by minimizing an overall maximum likelihood objective function, Φ , which is defined as:

$$\Phi = \frac{NT}{2} \ln(2\pi) + \frac{1}{2} \min_{\theta} \left\{ \sum_{i=1}^{NE} \sum_{j=1}^{NV_i} \sum_{k=1}^{NM_{ij}} \left[\ln(\sigma_{ijk}^2) + \frac{(\tilde{z}_{ijk} - z_{ijk})^2}{\sigma_{ijk}^2} \right] \right\} \quad \text{Eq 8}$$

where NT is the total number of measurements, θ the set of parameters to be calibrated, NE the number of experiments performed, NV_i the number of variables measured in the i^{th} experiment, NM_{ij} the number of measurements of the j^{th} variable, NM_{ij} the number of measurements of the j^{th} variable in the i^{th} experiments, σ_{ijk}^2 the variance of the k^{th} measurement of variable j in experiment i , $\tilde{z}_{ijk}$ the k^{th} measured value of variable j in experiment i , and z_{ijk} the k^{th} predicted value by the model of variable j in experiment i .

Particle size distribution correction:

To improve the accuracy of predicted PSD metrics generated by PBM, a machine learning-based correction model was developed. The input dataset comprised PBM predicted values for the PSD quantiles d10, d50 and d90 and the corresponding ground-truth values from the offline particle characterisation of the initial 16 experiments. A MultiOutput Regressor was employed using Ridge Regression as the base estimator to jointly model all three PSD quantiles. Ridge regression was chosen due to its regularization capability, which helps mitigate overfitting and multicollinearity. The model was implemented using scikit-learn's³⁴ MultiOutputRegressor with a fixed random state to ensure reproducibility. To assess model performance and ensure generalizability to unseen data, RSME averaged uniformly across all output dimensions was used along with Leave-One-Out Cross-Validation (LOOCV).

Simulation for dataset generation

To generate a wide dataset for training the emulator model, a parametric study using the calibrated model was performed in FormulatedProducts 2025.1.0. The crystallisation process followed the same reaction procedure as the experimental work, described previously. Factors varied, bounds and distribution type are shown in Table 1. To ensure there was representative generation of seeded and unseeded conditions, these were treated as 2 separate parametric studies and collated post simulation. Seed size was treated as 2 discrete cases (small and large seed) in all cases. For unseeded conditions, seed loading was fixed at 0 g and 3,500 samples (7,000 total) of the remaining factors were generated using a Sobol approach. For seeded conditions, the seed loading was specified as a uniform distribution and 7,000 samples (14,000 total) of all factors were generated also using a Sobol approach. Response factors from the parametric study were crystal mass, final solute concentration, final solubility, and the PSD quantiles d10,

d50, and d90. The initial output from the parametric study underwent pre-processing to calculate the yield and apply the PSD correction model. All simulations were performed on a workstation with AMD Ryzen Threadripper 3990X @ 2.9 GHz processor and 256 GB RAM using 128 workers and took 3 hr to complete.

Table 1. PBM parametric study factors.

Factor	Lower bound	Upper bound	Unit	Distribution
Antisolvent final fraction	0.45	0.9	-	Uniform
Antisolvent addition rate	3	10	mL/min	Uniform
Initial (seeding) temperature	34	40.1	°C	Uniform
Final temperature	10	25	°C	Uniform
Cooling rate	-0.5	0	°C/min	Uniform
Seed loading	0 (unseeded), 0.525 (seeded)	2.625	g	Enumerated, Uniform
Seed size	46.9	55.9	µm	Enumerated

2.3. Benchmark Function

Parameter Configuration for Benchmarking

To construct a relevant and flexible benchmark function for model-based experimental design, a dataset comprising over 20,000 simulated pharmaceutical crystallisation experiments was used (see Section 2.2) of various parameters and objectives. The parameter space was defined by selecting a subset of input variables, e.g., cooling rate, end temperature, and antisolvent flow rate. These variables could either be explicitly varied or held fixed by filtering the dataset to isolate relevant process regimes, e.g., final temperature = 10 °C if wanting to cool to same thermodynamic point each experiment. This configurability allows the benchmark to reflect practical design constraints and to be tailored for user's specific case studies. Target outputs, such as yield and selected percentiles of the particle size distribution (PSD), were similarly filtered. While the full dataset includes a broad array of possible outputs, only those metrics aligned with the goals of Section 2.4 were retained. This step ensured that the constructed benchmark function aligned with computational tractability and gave appropriate results for realistic optimisation.

Random Forest Emulator

A random forest (RF) regression model (Python implementation using scikit-learn³⁴) was developed to act as an emulator for the crystallisation process. The model was trained to predict selected crystallisation outcomes from the chosen process parameters, enabling rapid evaluation of experimental campaigns in silico. Input features and target outputs were extracted from the filtered dataset (parameter configuration). The data were split into training and test sets using an 80/20 ratio. Prior to model fitting, all input features were standardised using z-score normalisation to ensure uniform scaling across variables. A random forest regressor with 100 decision trees and a fixed random seed was used to ensure reproducibility. After training, model performance was assessed on the test set using two standard error metrics: mean squared error (MSE) and mean absolute error (MAE). The resulting model and preprocessing pipeline were saved for integration with the design of experiment framework.

2.4. Experimental Design & Optimisation

The following three sections were implemented in Python using Obsidian³⁵.

Initial Screening Designs

Twelve initial datasets were generated using combinations of design of experiment (DoE) methods (Latin Hypercube Sampling (LHS), Sobol sequence, Random sampling) and number of experiments (5, 10, 15, 20). Results were emulated for each dataset using the random forest surrogate model described in Section 2.3. Input parameters were scaled using the previously fitted scaler and passed to the RF model to obtain predicted outcomes. Predictions were stored in the corresponding dataframes.

Optimisation Approach 1: Bayesian Optimisation

Each of the 12 initial datasets was used to start multiple BO campaigns. Firstly, input parameter values were normalised before a GP model was fitted. A single new design point (batch size = 1) was proposed using one of six acquisition functions: Mean, Random Sampling (RS), Space Filling (SF), Expected Hypervolume Improvement (EHVI), Noisy EHVI (NEHVI) and Pareto efficient global optimisation (NParEGO), now resulting in 72 dataframes. Predicted results for each new point were emulated using the random forest model. The updated dataset was then used to refit the GP model, and another point was proposed using the same acquisition function as before. This procedure (fit, suggest, emulate) was repeated five times total for each dataset, resulting in 5 additional design iterations per acquisition function for each of the initial screening datasets.

Campaign Repeats

The full procedure, including initial screening and five rounds of Bayesian optimisation, was repeated five times. This resulted in a total of 360 emulated experimental

campaigns. The logic of the for loops from the code relating to the above sections can be visualised in Figure 2.

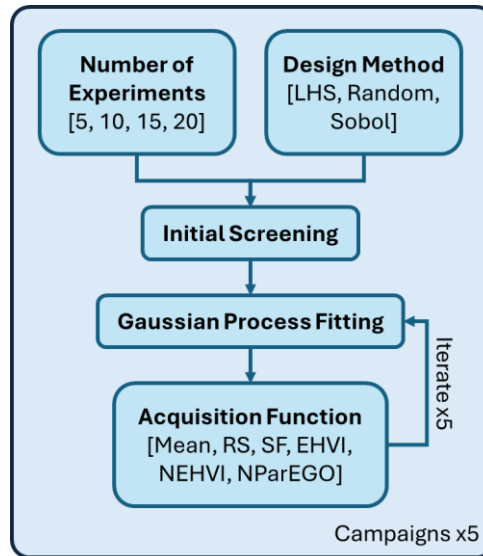


Figure 2. Schematic showing the logic of the experimental campaigns.

Optimisation Approach 2: Gradient Boost

A separate scientist at a different site conducted further emulator-based in-silico experimental campaigns, following the same workflow as shown in Figure 2 but choosing a single sampling method and single acquisition function applied to a different model fit than GP. The campaigns began with the same 20 initial LHS designs as used in the previous section, spanning 5, 10, 15 and 20 starting experiments, each repeated five times. A gradient boosting decision tree (GBDT)(xgboost³⁶) model was fitted to each initial dataset, and an expected improvement acquisition function was applied to propose the next single experiment. The outcome for each suggested point was predicted using the random forest emulator, after which the GBDT model was refitted. This procedure was repeated to complete five MB-DoE iterations per campaign.

Hypervolume Calculation

The performance of each campaign was quantitatively assessed using the hypervolume (HV) metric, implemented via the pymoo Python package³⁷. Prior to HV calculation, all campaign data were normalised to a unit cube defined by the bounds of the simulated dataset: PSD ranging from 78 to 174 μm and yield from 17% to 99%. The reference point for hypervolume computation was set at 0.1 beyond the normalised bounds in all objectives, representing the worst-case performance scenario. This ensured consistent and comparable evaluation across campaigns regardless of scale or initial condition variations.

3. Results

3.1. Population Balance Model & Simulation

Model calibration

The results obtained after model calibration of the solubility and PBM are shown in Table 2 with the nominal values of the parameters and the 95% confidence interval bounds. Goodness of fit across all experiments is shown in the parity plots in Figure 3.

Table 2. Estimated values for the population balance model

	Parameter	Value	95% Confidence Interval	Units
Solubility	$\psi_{0,1}$	6.346	0.44435	-
	$\psi_{1,1}$	12.033	1.29155	-
	$\psi_{0,2}$	-3.1914	0.13355	-
	$\psi_{1,2}$	-9.1843	1.9563	-
	$\psi_{2,1,2}$	4.3185	3.88125	-
	$\psi_{3,1,2}$	-0.8031	3.6425	-
	$\psi_{4,1,2}$	1.3245	1.96455	-
Growth: Surface integration	$E_{A,g}$	31895	25207	J/mol
	k_g	0.00292123	0.0303728	m/s
	g	0.500110	0.231870	-
Growth: Mass transfer	α	0.248754	11.7377	-
Primary nucleation	$\ln k_n$	22.6190	0.212818	$\ln(\#/m^3.s)$
	n	0.884977	0.265064	-

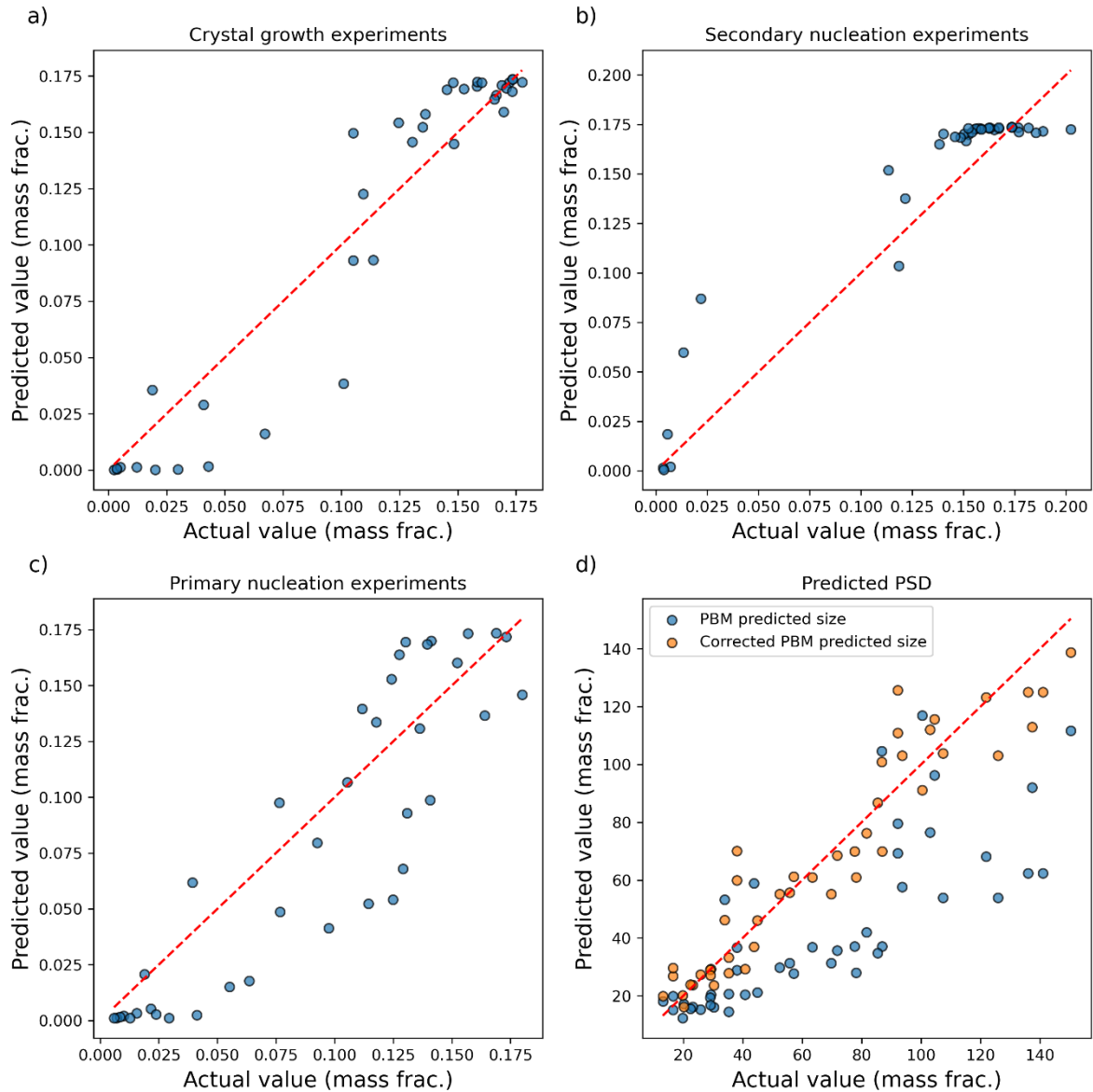


Figure 3. Comparison of actual and predicted values for a) concentration measurements to calibrate crystal growth parameters, b) concentration measurements for validation of secondary nucleation experiments, using a growth only model, c) concentration measure to calibrate primary nucleation parameters, and d) PBM particle sizes with and without correction model.

As can be seen from Table 2, parameters relating to secondary nucleation are absent. During the sequential parameter estimation approach, low seed loading experiments (expected to have secondary nucleation present) were simulated using the growth only model calibrated from the high seed loading experiments (Figure 3b). It was found that the data did not significantly deviate from the growth only model predictions, suggesting that secondary nucleation was either not present or to a degree that no measurable difference was observed. Therefore, to reduce the PBM complexity, secondary nucleation was not included as a mechanism. Figure 2d shows the improvement in PSD

prediction with the machine learning based correction model. Across all experiments used for the PBM calibration, the RMSE of the PBM predictions was 32.9 μm . However, with the correction this was reduced to 12.4 μm .

Overall, for the PBM, although the goodness of fit could have been improved further and confidence intervals reduced through further experimental work and/or targeted experiments, it was determined that the PBM reasonably captured the behaviour of lovastatin crystallisation, as shown in ESI Figures 1, 2 and 3.

Dataset generation:

Figure 4 shows the results of exploratory data analysis on the dataset generated from the PBM parametric study. For the product particle size distributions (Figure 4a, b, and c), the d10, d50 and 90 all show a roughly symmetric distribution centred around the mean. In contrast, the yield (Figure 4d) shows a strongly skewed distribution towards higher values (80% +), suggesting that a large fraction of the simulated crystallisation processes are efficient. Additionally, the correlation analysis (Figure 4e) shows that all particle size quantiles have a high degree of positive correlation with the cooling rate, followed by a moderate degree of positive correlation with the seed loading. In contrast, several factors correlate with yield with different behaviours. Both antisolvent addition rate and cooling rate have low to moderate positive correlations with the yield. Whereas the antisolvent end fraction and final temperature have moderate negative correlations.

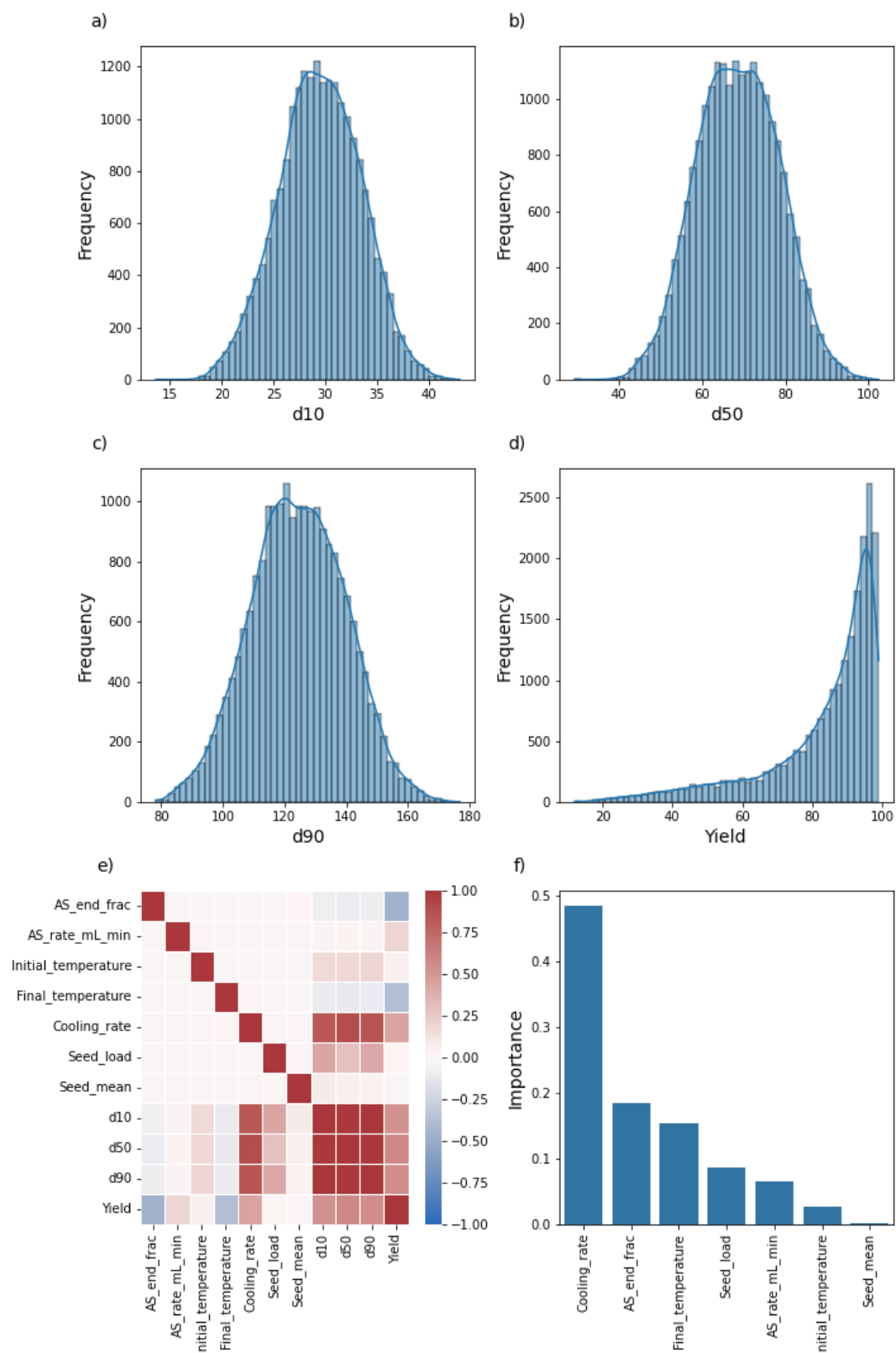


Figure 4. Exploratory data analysis of generated dataset for emulator training showing histograms of responses a) particle size quantile d10, b) particle size quantile d50, c) particle size quantile d90, and d) yield. e) Pearson correlation heatmap for input factors and responses. f) Features importance plot for the random forest crystallisation emulator.

3.2. Benchmark Function

Parameter Configuration for Benchmarking

To demonstrate the most complex crystallisation case study, all seven input parameters were defined for optimisation, including six continuous variables and one ordinal variable. The continuous parameters were initial temperature (34–40 °C), seed load (0–2.625%), final temperature (10–25 °C), cooling rate (0.1–0.5 °C/min), antisolvent endpoint fraction (0.45–0.9), and antisolvent addition rate (3–10 mL/min). Seed size was treated as an ordinal parameter with two discrete levels (mean sizes: 46.94 µm and 55.94 µm), reflecting the two available seed batches.

Two process and product performance metrics were selected as optimisation targets: yield (to be maximised) and particle size distribution D90 (to be minimised). These metrics were extracted from the simulated crystallisation dataset to represent realistic objectives consistent with typical pharmaceutical product profiles.

Random Forest Emulator

A random forest regression model was trained on the selected dataset to serve as a fast surrogate for the crystallisation process. The model was trained on a workstation with an Intel® Core™ i7-8700K CPU @ 3.70 GHz and 32 GB RAM. Training completed in under 5 seconds.

Model performance on the test set yielded a mean squared error (MSE) of 2.71 and a mean absolute error (MAE) of 1.05. Given that the maximum observed value of D90 in the dataset was 173 µm, the absolute error corresponds to less than 1% of the dynamic range, indicating adequate surrogate accuracy for downstream optimisation.

The feature importance plot generated for the random forest (RF) model (Figure 4f) indicates a diverse distribution of feature relevance, which will make for complex optimisation scenarios that mirror real-world systems. Cooling rate, a kinetic parameter, emerged as the most influential feature with an importance score of 0.483. This was followed by two thermodynamic variables that largely influence process yield: antisolvent end fraction (0.185) and final temperature (0.153). In contrast, seed size had the lowest importance score (0.002), which is consistent with the minimal variation in this parameter across the dataset, which was only a 9 µm difference between ‘small’ and ‘big’.

3.3. Experimental Design & Optimisation

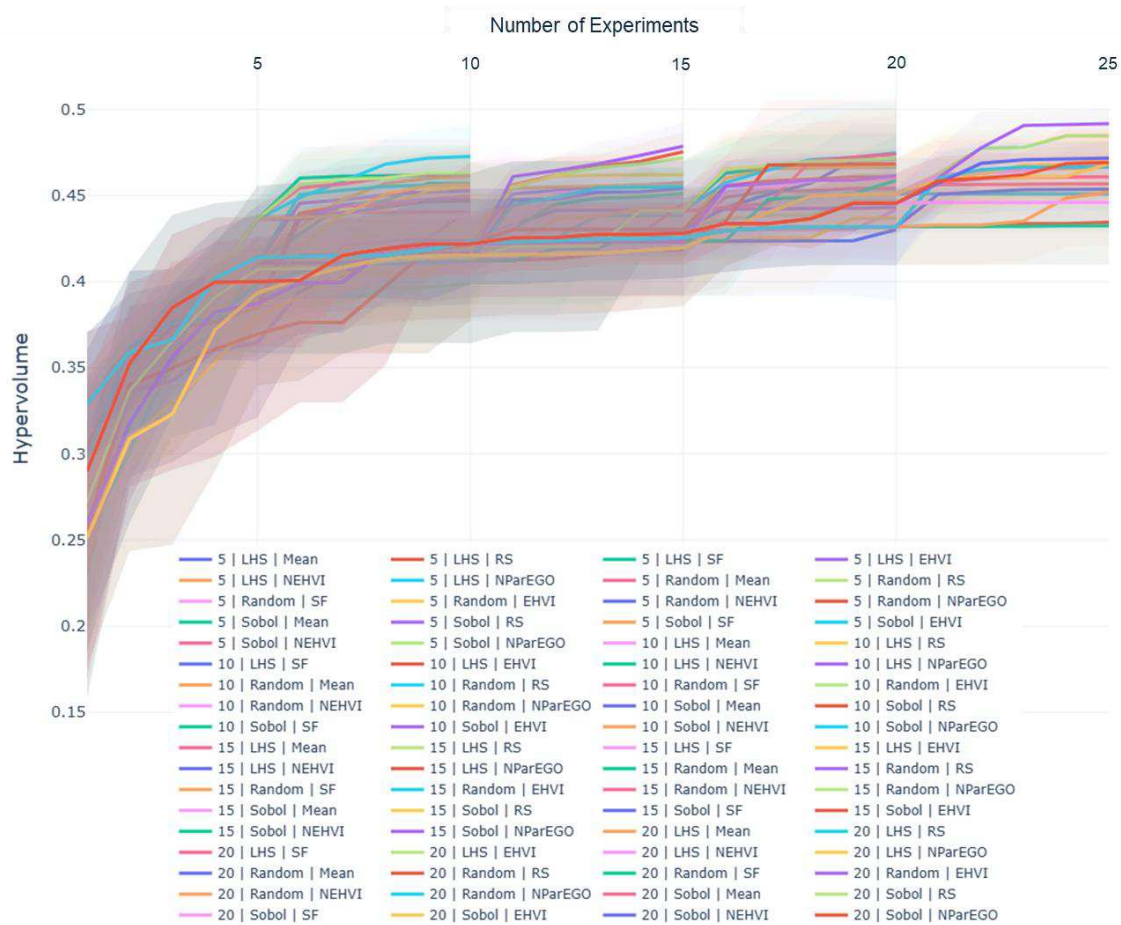


Figure 5. Tracking of hypervolume improvements over number of experiments performed for all 72 combinations of initial points, sampling methods and acquisition function. The mean and confidence bound across the 5 full campaigns are given by the line and fills respectively. An interactive version of this plot can be found at <https://tpicks95.github.io/npjComputationalMaterials/plot.html>.

A general increasing trend (Figure 5) is observed, with steady gains as more experiments are added, and slight tailing as the number of data points approaches the upper bound of the campaign. The most pronounced improvements in hypervolume occur within the first 6 to 7 experiments, corresponding to the transition from sparse to moderately populated design spaces. A secondary increase is consistently observed following the application of MB-DoE strategies after the initial screening phase.

The highest final hypervolume value achieved was 0.492 for the campaign consisting of 20 initial random experiments followed by 5 experiments guided by the EHVI acquisition function. A comparable outcome, but in far fewer experiments, was observed for the configuration using 10 initial Sobol-sequence experiments followed by 5 EHVI-guided experiments, yielding a final hypervolume of 0.479. The lowest final hypervolume result

was 0.399, obtained from a campaign initiated with 5 random experiments and extended with 5 additional experiments also selected randomly. It is also important to note that the confidence intervals around the mean hypervolume metrics narrowed as MB-DoE led experiments were introduced as the campaigns moved away from random or quasi-random initial sampling methods.

Initial Screening Designs

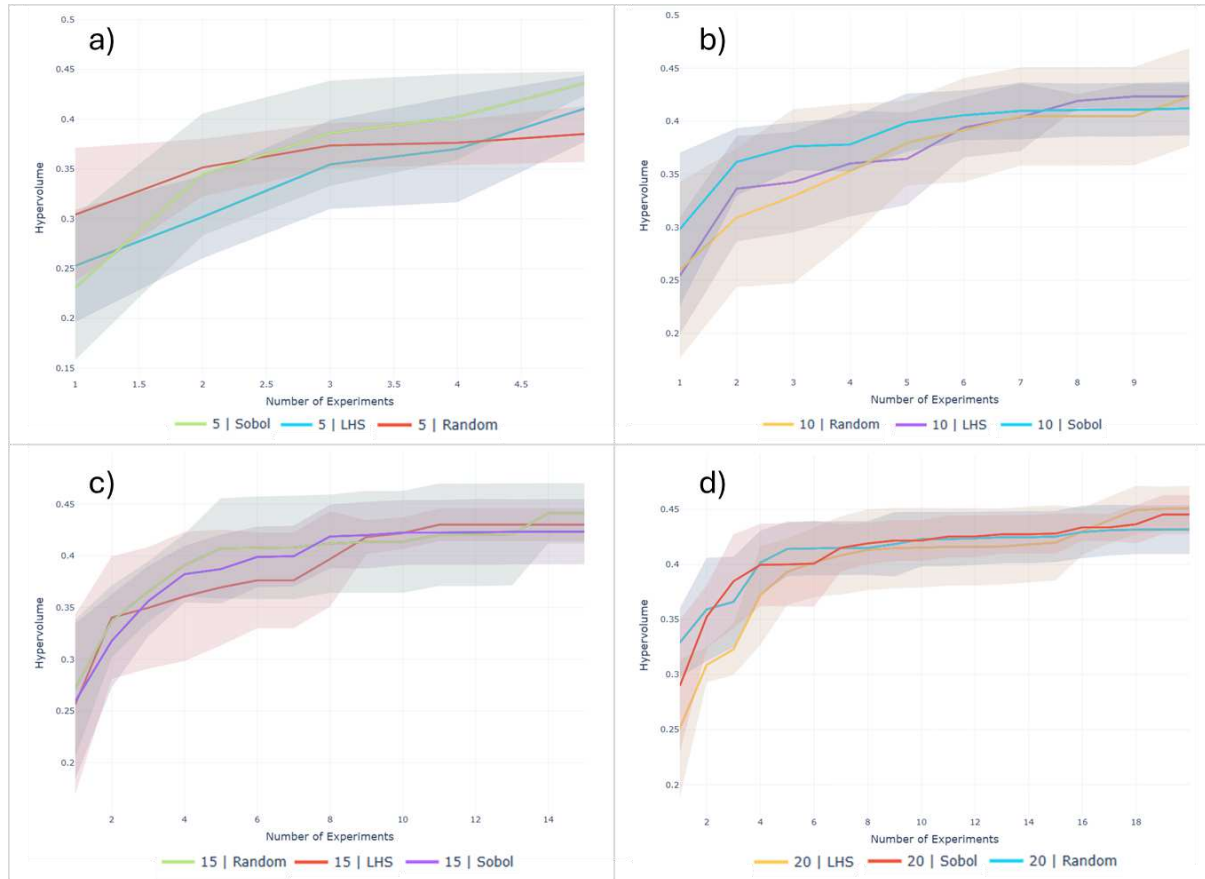


Figure 6. Tracking of hypervolume improvements over number of experiments performed for all combinations of sampling methods and acquisition functions for a) 5 initial points, b) 10 initial points, c) 15 initial points and d) 20 initial points.

The performance of three initial screening sampling strategies, Sobol, LHS, and Random, was compared using the HV metric after 5, 10, 15, or 20 initial experiments.

With 5 initial experiments (Figure 6a), Sobol achieved the highest HV (0.436), followed by LHS (0.411) and Random (0.385). With 10 initial experiments (Figure 6b), Random and LHS performed comparably, both achieving an HV of 0.423. Sobol yielded a slightly lower HV of 0.412. With 15 initial experiments (Figure 6c), Random led with an HV of 0.441. LHS and Sobol followed with HV values of 0.430 and 0.423, respectively. Finally, with 20 initial experiments (Figure 6d), LHS achieved the highest HV (0.451), outperforming Sobol (0.445) and Random (0.432). Sobol showed the strongest performance at the lowest

sample count. LHS improved consistently with sample size. Random sampling occasionally matched or exceeded the other strategies at intermediate stages, reflecting higher variability.

Optimisation Approach 1: Bayesian Optimisation

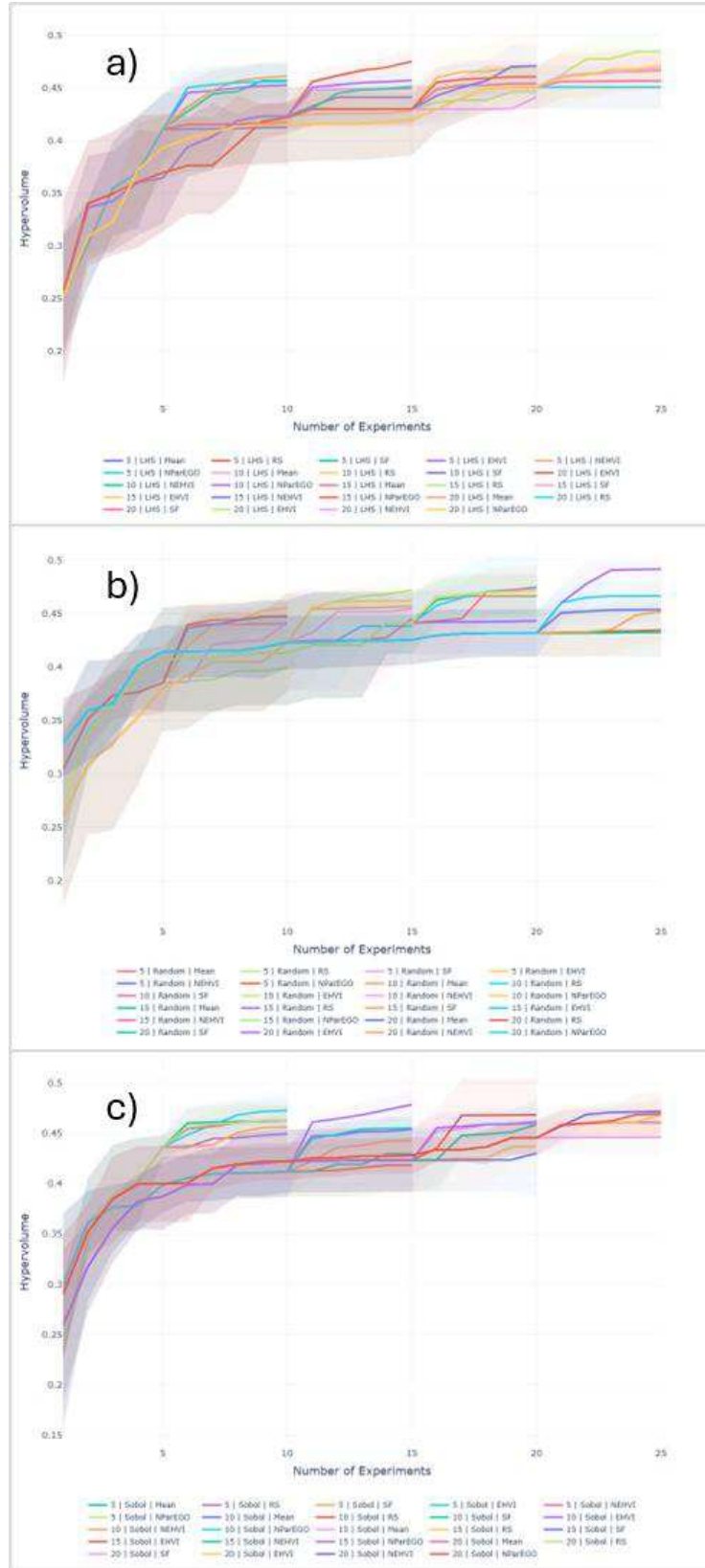


Figure 7. Tracking of hypervolume improvements over number of experiments performed for all combinations of initial points and acquisition functions for a) LHS, b) Random and c) Sobol sampling methods.

Next, the performance of MB-DoE led experiments, following the three different initial design approaches, were compared. For each scenario, MB-DoE algorithms were compared based on their final HV after subsequent optimisation, starting from an initial 5, 10, 15, or 20 experiments.

With LHS as the initial design (Figure 7a), EHVI and NEHVI were consistently among the top performers, while RS was the poorest. With 5 initial experiments, NEHVI achieved the highest HV (0.461), narrowly surpassing SF (0.457), while RS lagged at 0.417. With 10 and 15 points, EHVI became dominant (0.475 and 0.471), with NParEGO and NEHVI close behind. Notably, with 15 experiments, EHVI and NEHVI performed identically (0.471). With 20 initial points, EHVI again led (0.485), followed by NParEGO (0.471), while RS remained the weakest (0.451). Across this scenario, EHVI's advantage grew with more initial data, whereas RS consistently underperformed. In the initial random sampling scenario (Figure 7b), EHVI maintained a clear lead across all initial sizes. NParEGO was the next strongest, particularly with 10, 15, and 20 initial experiments (0.462, 0.471, and 0.467, respectively). RS was again the weakest method overall, with the lowest HV values at each stage, dropping to as low as 0.399 with 5 initial points. The HV spread was most pronounced at 20 experiments, where EHVI reached 0.492 compared to RS at 0.434, indicating the strongest differentiation in algorithm performance under this initialisation. With Sobol sequence sampling (Figure 7c), EHVI again achieved the highest HV values with 5, 10, and 15 initial experiments (0.473, 0.479, and 0.468, respectively), followed closely by NParEGO. However, with 20 initial experiments, NEHVI slightly surpassed EHVI (0.472 vs. 0.468), suggesting its potential to improve with larger initial datasets. The lowest performance in this scenario was shared between SF and RS, both reaching only 0.446 at 20 experiments. Overall, Sobol sampling resulted in relatively higher baseline performance across methods, yet the ranking among MB-DoE strategies remained consistent with other scenarios.

Optimisation Approach 2: Gradient Boosting

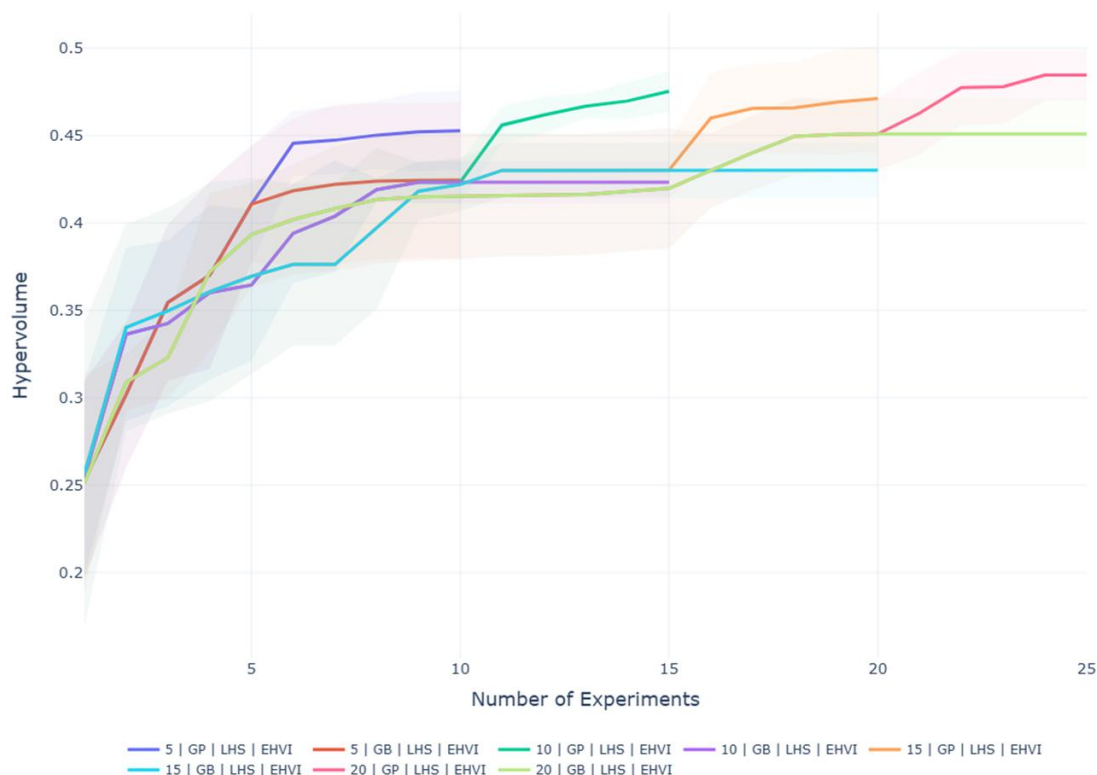


Figure 8. Tracking hypervolume improvements over the number of experiments for all initial design sizes, using LHS sampling and either a GP or gradient boosting model with expected improvement as the acquisition function.

To evaluate the applicability of the crystallisation emulator as a benchmark when using non-standard optimisation strategies, we demonstrated a gradient boosting model with expected improvement acquisition trained on a range of initial LHS designs. These were the best performing sampling methods and acquisition functions from the previous experimental campaigns. For the case with five initial experiments, the hypervolume increased from 0.411 to 0.424 (Figure 8) over five iterations, indicating consistent but modest improvements. However, for larger initial designs (10, 15 and 20 initial experiments), no further gains were observed, with hypervolume values plateauing immediately after the initial sampling phase. In contrast to the GP model paired, shown in the above section, with the same acquisition function, outperformed the gradient boosting approach in all cases. Relative to the gradient boosting model, the GP achieved hypervolume improvements of 6.8%, 12.3%, 9.5%, and 7.5% for 5, 10, 15 and 20 initial experiments, respectively. These results highlight the emulator's ability to reveal performance disparities between surrogate models in data-driven experimental design.

4. Discussion

This work introduces a novel, domain-specific benchmark framework tailored to pharmaceutical crystallisation, addressing a longstanding gap in the evaluation of MB-

DoE strategies within the pharmaceutical industry. Conventional benchmark functions used in optimisation research are typically abstract and poorly suited to noisy, constrained, and kinetically complex chemical systems. In contrast, the emulator developed here, based on a mechanistic PBM of lovastatin crystallisation, captures realistic process dynamics, including particle nucleation, growth and dissolution kinetics. Trained on over 20,000 simulated experiments, the RF emulator enables rapid, reproducible, and scalable evaluation of MB-DoE algorithms under conditions reflective of industrial practice⁸. Calibration using experimental data, along with a machine learning correction for PSD metrics, improved PBM fidelity and ensured the emulator remains both mechanistically grounded and statistically robust.

The configurability of the benchmark, including its capacity to simulate seeded and unseeded crystallisation, ordinal and continuous inputs, and multiple output objectives, also ensures its future extensibility. This flexibility ensures that the framework can evolve with emerging experimental needs and supports broader benchmarking efforts aligned with domain-specific process characteristics.

The benchmarking framework ran 360 in-silico emulated optimisation campaigns, covering various combinations of initial sampling strategies and acquisition functions applied to GP models. While EHVI consistently delivered the best performance across all configurations^{38, 39}, the greater contribution lies in establishing a crystallisation-specific testbed for algorithm comparison for all model and acquisition types. The emulator revealed performance differences between acquisition functions and allowed these distinctions to be quantified at a level of resolution and repeatability not feasible through physical experimentation alone. Beyond acquisition strategy, the emulator also distinguished between surrogate modelling approaches demonstrated by a further 20 optimisation campaigns using gradient boost. This use of the emulator was performed by a separate scientist at a different site showing the ease of implementation. Gradient boost, although this method is common in data-driven modelling⁴⁰, underperformed relative to GP across all LHS design sizes and sometimes even showed no improvements from the initial sampling. This result reinforces the value of uncertainty-aware models in BO⁴¹ and highlights the potential pitfalls of adopting alternative surrogates without rigorous validation. The emulator benchmark provides a controlled environment for stress-testing optimisation strategies before laboratory deployment.

In summary, although the comparison highlights that GP with EHVI combined with LHS initialisation proved the most effective configuration for experimental design in this study, the principal innovation is the development of a scalable, realistic, and mechanistic crystallisation benchmark. The emulator enables rigorous, reproducible assessment of MB-DoE strategies, and we invite the community to benchmark their optimisation approaches and compare against the GP campaigns using the provided emulator (<https://github.com/tpicks95/npjComputationalMaterials>). By facilitating transparent

comparison under realistic conditions, the emulator can support more informed algorithm selection and contribute to the wider integration of AI-driven methods in pharmaceutical development.

Author Contributions

Conceptualization: TP, CJB. Methodology: TP, KN, BAM, CJB. Investigation: TP, KN, CJB. Writing – Original Draft: TP, CJB, JM. Writing – Review & Editing: CJB. Supervision: SS, CJB. Project administration: JM. Funding acquisition: AJF.

Data Availability

All data underpinning this publication are openly available from the University of Strathclyde KnowledgeBase at: <https://doi.org/10.15129/39b0828f-082a-4e71-88a1-f0d67aab4375>.

Code Availability

The code and models supporting this article can be found at:
<https://github.com/tpicks95/npjComputationalMaterials>.

Acknowledgements

The authors would like to acknowledge E.Hadjittofis (UCB), N. Nazemifard (Takeda), J. Merritt (Eli Lilly), O. Watson (AstraZeneca), Y. Jangjou (Sanofi), D. Ford (Snapdragon/Cambrex), K. Girard (Pfizer), C. Burcham (Eli Lilly), D. Remick (Eli Lilly), A. Piccinini (Chiesi) for discussion and guidance. This work was funded jointly by Astra Zeneca, Chiesi, Eli Lilly, Pfizer, Roche, Sanofi, Takeda, and UCB. It was carried out within the CMAC Future Manufacturing Research Hub (EPSRC Grant ref: EP/P006965/1) using equipment bought through the UKRPIF Net Zero Medicines Manufacturing Research Pilot funded by Research England and the Scottish Funding Council.

References

1. A. Myerson, *Handbook of industrial crystallization*, Butterworth-Heinemann, 2002.
2. R. J. Davey and J. Garside, *From molecules to crystallizers*, 2000.

3. Y. Qiu, Y. Chen, G. G. Zhang, L. Yu and R. V. Mantri, *Developing solid oral dosage forms: pharmaceutical theory and practice*, Academic press, 2016.
4. D. Erdemir, A. Y. Lee and A. S. Myerson, *Accounts of chemical research*, 2009, **42**, 621-629.
5. T. Pickles, C. Mustoe, C. Boyle, J. Cardona, C. J. Brown and A. J. Florence, *CrystEngComm*, 2024, **26**, 822-834.
6. M. A. Soldatov, V. V. Butova, D. Pashkov, M. A. Butakova, P. V. Medvedev, A. V. Chernov and A. V. Soldatov, *Nanomaterials*, 2021, **11**, 619.
7. T. Pickles, C. Mustoe, C. Brown and A. Florence, *2022 APS Special Issue*, 2022, **7**.
8. T. Pickles, V. Svoboda, I. Marziano, C. J. Brown and A. J. Florence, *CrystEngComm*, 2024.
9. B. P. MacLeod, F. G. Parlane, T. D. Morrissey, F. Häse, L. M. Roch, K. E. Dettelbach, R. Moreira, L. P. Yunker, M. B. Rooney and J. R. Deeth, *Science Advances*, 2020, **6**, eaaz8867.
10. G. Tom, S. P. Schmid, S. G. Baird, Y. Cao, K. Darvish, H. Hao, S. Lo, S. Pablo-García, E. M. Rajaonson and M. Skreta, *Chemical Reviews*, 2024, **124**, 9633-9732.
11. B. Burger, P. M. Maffettone, V. V. Gusev, C. M. Aitchison, Y. Bai, X. Wang, X. Li, B. M. Alston, B. Li, R. Clowes, N. Rankin, B. Harris, R. S. Sprick and A. I. Cooper, *Nature*, 2020, **583**, 237-241.
12. F. Häse, L. M. Roch and A. Aspuru-Guzik, *Trends in Chemistry*, 2019, **1**, 282-291.
13. C. W. Coley, N. S. Eyke and K. F. Jensen, *Angewandte Chemie International Edition*, 2020, **59**, 22858-22893.
14. D. R. Jones, M. Schonlau and W. J. Welch, *Journal of Global optimization*, 1998, **13**, 455-492.
15. B. Shahriari, K. Swersky, Z. Wang, R. P. Adams and N. De Freitas, *Proceedings of the IEEE*, 2015, **104**, 148-175.
16. M. Jamil and X.-S. Yang, *International Journal of Mathematical Modelling and Numerical Optimisation*, 2013, **4**, 150-194.
17. S. Surjanovic and D. Bingham, *Journal*, 2013.
18. E. Brochu, V. M. Cora and N. De Freitas, *arXiv preprint arXiv:1012.2599*, 2010.
19. A. E. Gongora, B. Xu, W. Perry, C. Okoye, P. Riley, K. G. Reyes, E. F. Morgan and K. A. Brown, *Science advances*, 2020, **6**, eaaz1708.

20. J. M. Granda, L. Donina, V. Dragone, D.-L. Long and L. Cronin, *Nature*, 2018, **559**, 377-381.
21. J. Liang, T. Wu, Z. Wang, Y. Yu, L. Hu, H. Li, X. Zhang, X. Zhu and Y. Zhao, *Energy Mater*, 2022, **2**, 200016.
22. Y. Zhao, J. Zhang, Z. Xu, S. Sun, S. Langner, N. T. P. Hartono, T. Heumueller, Y. Hou, J. Elia and N. Li, *Nature communications*, 2021, **12**, 2191.
23. T. Pickles, C. Mustoe, C. J. Brown and A. J. Florence, *Crystal Growth & Design*, 2024, **24**, 1245-1253.
24. Z. K. Nagy and R. D. Braatz, *Annual review of chemical and biomolecular engineering*, 2012, **3**, 55-75.
25. F. Mahmood, *Nature Medicine*, 2025, 1-1.
26. J. Riebesell, R. E. Goodall, P. Benner, Y. Chiang, B. Deng, A. A. Lee, A. Jain and K. A. Persson, *arXiv preprint arXiv:2308.14920*, 2023.
27. J. McGinty, M. W. Chong, A. Manson, C. J. Brown, A. Nordon and J. Sefcik, *Crystals*, 2020, **10**, 925.
28. A. A. Lahiq and S. M. Alshahrani, *Arabian Journal of Chemistry*, 2023, **16**, 104929.
29. R. Leeming, T. Mahmud, K. J. Roberts, N. George, J. Webb, E. Simone and C. J. Brown, *Industrial & Engineering Chemistry Research*, 2023, **62**, 11067-11081.
30. T. Pickles, Y. Leghrib, M. Weisshaar, M. Goncharuk, P. Timperman, T. Doherty, D. D. Ford, J. Moores, A. J. Florence and C. J. Brown, *Digital Discovery*, 2025.
31. H. Sun and J. Wang, *Journal of Chemical & Engineering Data*, 2008, **53**, 1335-1337.
32. A. Jouyban and W. E. Acree Jr, *Journal of Molecular Liquids*, 2018, **256**, 541-547.
33. J. F. Pérez-Calvo, S. S. Kadam and H. J. Kramer, *AIChE Journal*, 2016, **62**, 3992-4012.
34. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss and V. Dubourg, *the Journal of machine Learning research*, 2011, **12**, 2825-2830.
35. K. Stone, Y. Xu and Merck, Obsidian, <https://github.com/MSDLLCpapers/obsidian/tree/main>).
36. T. Chen and C. Guestrin, 2016.
37. J. Blank and K. Deb, *Access*, 2020, **8**, 89497-89509.

38. U. Noè and D. Husmeier, *arXiv preprint arXiv:1808.06918*, 2018.
39. F. Rehbach, M. Zaefferer, B. Naujoks and T. Bartz-Beielstein, 2020.
40. T. R. Noviandy, G. M. Idroes and I. Hardi, *Journal of Soft Computing and Data Mining*, 2024, **5**, 46-56.
41. Y. Wu, A. Walsh and A. M. Ganose, *Digital Discovery*, 2024, **3**, 1086-1100.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [esi.docx](#)