ESI: Automated Scale-Up Crystallisation DataFactory for model-based pharmaceutical process development: A Bayesian case study

Thomas Pickles[a], Youcef Leghrib[a], Matt Weisshaar[b], Mikhail Goncharuk[b], Peter Timperman[b], Timothy Doherty[b], David D. Ford[b], Jonathan Moores[a], Alastair J. Florence[a, c], Cameron J. Brown[a, c*]

[a] Strathclyde Institute of Pharmacy and Biomedical Sciences, University of Strathclyde, Glasgow G4 0RE, UK

[b] Snapdragon Chemistry, a Cambrex company. 360 2nd Ave., Suite C, Waltham, MA 02451 USA

[c] EPSRC Future Manufacturing Hub in Continuous Manufacturing and Advanced Crystallisation, University of Strathclyde, Glasgow, G1 1RD, UK

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

S.1. The Automated Scale-Up Crystallisation DataFactory Platform

Sources of variation design of Experiments (DoE) were conducted to evaluate the performance of the crystallisation vessel chiller. Key process parameters—solvent type, stir rate, cooling rate, and volume—were systematically varied due to their influence on heat transfer during crystallisation. Each experiment was performed in triplicate, followed by uncertainty analysis. The percentage error, calculated as the deviation between the desired and actual cooling rates, was used to assess accuracy. Covariance analysis (Figure S.1) of the data presented in Table S.1 identified a strong correlation between the thermal conductivity of the solvent and the percentage error, with water (highest thermal conductivity) exhibiting the highest deviation from the target cooling rate. Additionally, a strong correlation was observed between cooling rate and percentage error. Despite these findings, regression-based uncertainty analysis showed a strong linear fit across all experiments, indicating reduced accuracy but high precision. Other correlations were found to be minimal or negligible.

Table S.1. Sources of variation experiments performed on the chillers as part of the hardware capability checks. R^2 is reported to two decimal places.

Chiller Solvent	Vessel Solvent	Thermal Conductivity (W/m K)	Cooling Rate (oc/min)	Stir Rate (rpm)	Reactor Volume (mL)	CR_1	CR_2	CR_3	CR_AVG	R2_1	R2_2	R2_3	% Error
Water	Water	0.609	0.5	150	500	0.48	0.48	0.48	0.48	1.00	1.00	1.00	-4.30%
Water	Water	0.609	0.5	350	500	0.48	0.48	0.48	0.48	1.00	1.00	1.00	-3.96%
Water	Water	0.609	0.1	150	500	0.10	0.10	0.10	0.10	1.00	1.00	1.00	-2.44%
Water	Water	0.609	0.1	350	500	0.10	0.10	0.10	0.10	1.00	1.00	1.00	-2.40%
Water	Water	0.609	0.5	150	1000	0.48	0.48	0.48	0.48	1.00	1.00	1.00	-3.75%
Water	Water	0.609	0.5	350	1000	0.49	0.48	0.48	0.48	1.00	1.00	1.00	-3.51%
Water	Water	0.609	0.1	150	1000	0.10	0.10	0.10	0.10	1.00	1.00	1.00	-1.63%
Water	Water	0.609	0.1	350	1000	0.10	0.10	0.10	0.10	1.00	1.00	1.00	-1.69%
Water	Ethyl acetate	0.137	0.5	150	500	0.49	0.49	0.49	0.49	1.00	1.00	1.00	-1.75%
Water	Ethyl acetate	0.137	0.5	350	500	0.49	0.49	0.49	0.49	1.00	1.00	1.00	-2.01%
Water	Ethyl acetate	0.137	0.1	150	500	0.10	0.10	0.10	0.10	1.00	1.00	1.00	-1.45%
Water	Ethyl acetate	0.137	0.1	350	500	0.10	0.10	/	0.10	1.00	1.00	/	-1.26%
Water	Ethyl acetate	0.137	0.5	150	1000	0.49	0.49	0.49	0.49	1.00	1.00	1.00	-1.70%
Water	Ethyl acetate	0.137	0.5	350	1000	0.49	0.49	0.49	0.49	1.00	1.00	1.00	-1.42%
Water	Ethyl acetate	0.137	0.1	150	1000	0.10	0.10	0.10	0.10	1.00	1.00	1.00	-1.38%

Water	Ethyl acetate	0.137	0.1	350	1000	0.10	0.10	0.10	0.10	1.00	1.00	1.00	-0.87%

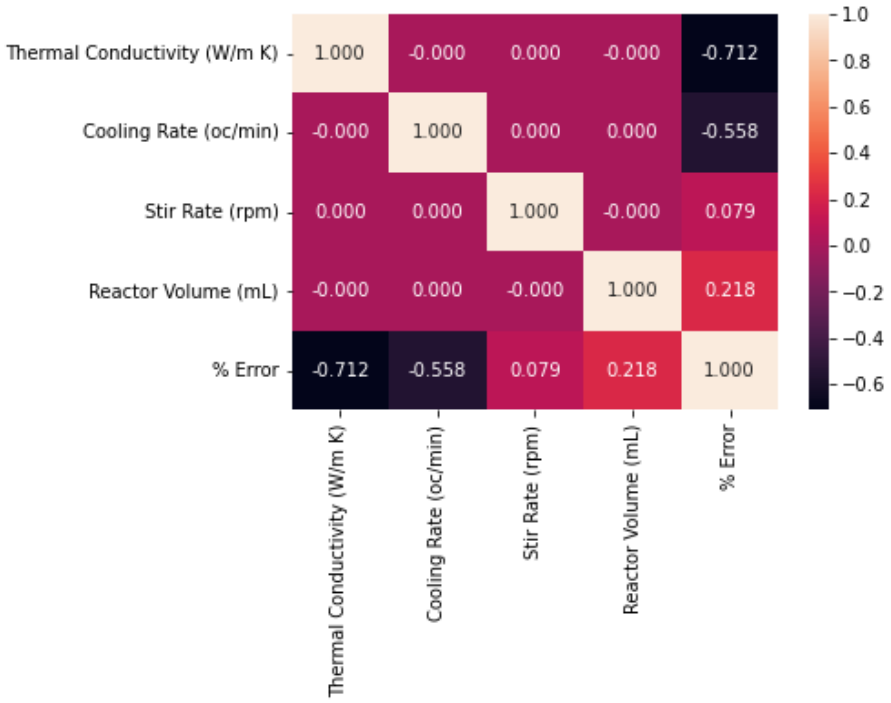

Figure S.1.Covariance matrix of the sources of variation experiments performed on the chillers as part of the hardware capability checks plotted using Seaborn.

Sources of variation design of Experiments (DoE) were conducted to evaluate the performance of the pumps and transfer system. Variables such as solvent type, flow rate, and transfer time were systematically varied, with the transferred volume as the response variable. Covariance analysis (Figure S.2) of the data in Table S.2 revealed a strong positive correlation between solvent surface tension and percentage error, with water—having the highest surface tension—showing the greatest deviation from the target transfer volume. However, the observed percentage error remained within ~3%, which is considered acceptable. Other correlations were found to be minimal or negligible.

Table S.2. Sources of variation experiments performed on the transfer/ pumps as part of the hardware capability checks.

Vessel Solvent	Viscosity (mPa.s)	Surface Tension (N/m)	Flow Rate (mL/min)	Transfer time (mins)	Volume transferred (mL)	V1	V2	V3	V_AVG	% Error	% Error_Abs
Water	0.89	0.072	5	5	25	24.5	24	24	24.2	-3%	3%
Water	0.89	0.072	25	5	125	130	127.5	130	129.2	3%	3%
Water	0.89	0.072	5	10	50	48	45	50	47.7	-5%	5%
Water	0.89	0.072	25	10	250	260	250	255	255.0	2%	2%
Water	0.89	0.072	50	5	250	255	270	265	263.3	5%	5%
Isopropyl acetate	0.49	0.022	5	5	25	25	24.5	25	24.8	-1%	1%
Isopropyl acetate	0.49	0.022	25	5	125	125	125	125	125.0	0%	0%
Isopropyl acetate	0.49	0.022	5	10	50	50.5	48	50	49.5	-1%	1%
Isopropyl acetate	0.49	0.022	25	10	250	250	250	250	250.0	0%	0%
Isopropyl acetate	0.49	0.022	50	5	250	255	250	250	251.7	1%	1%
Ethanol	1.04	0.022	5	5	25	26	25	25.5	25.5	2%	2%
Ethanol	1.04	0.022	25	5	125	135	125	130	130.0	4%	4%

IPA	2.3703	0.023	5	5	25	25	24.5	25.5	25.0	0%	0%
IPA	2.3703	0.023	25	5	125	120	130	130	126.7	1%	1%

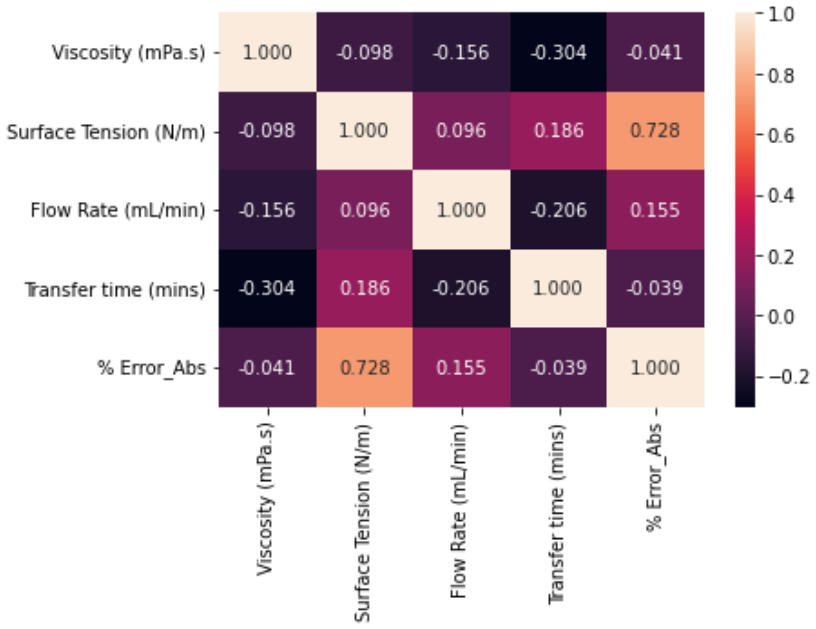

Figure S.2. Correlation matrix of the sources of variation experiments performed on the transfer/ pumps as part of the hardware capability checks plotted using Seaborn.

Sources of variation design of Experiments (DoE) were conducted to evaluate the performance of the pumps and transfer system in regards to the transfer of slurry. The result in Table S.3 shows a 6% error in the transfer of a 20 wt% slurry however all recorded repeats were of the same value therefore an offset was applied in determining volumes for seed transfer.

Table S.3. Sources of variation experiments performed on the transfer/ pumps of seed slurry as part of the hardware capability checks.

Reactor Solvent	**Temperature (°C)**	**Flow Rate (mL/min)**	**Transfer time (mins)**	**Volume transferred (mL)**	**%M1**	**%M2**	**%M3**	**%M_AVG**	**% Error**
20% wt slurry LAMV/ EtOH	25	100	0.5	50	19	19	19	19	6%

S.2. Methods

S.2.1. Reaction Procedure for E

A standard batch cooling crystallisation rea
components: the crystallisation vessel ch
controller (TCU-006-Ramp-Rate) and peris
with the seed slurry temperature held at 25
phases of interest include:

- Phase 4 – controls the temperature
- Phase 5 – controls the slurry transfe
- Phase 7 & 8 – controls the cooling

Figure S.3. Step-by-step reaction procedure wrote as inputted in LabOS.

Visualisation of the 3 key workers involved in the control of the crystallisation for the Latin-hypercube sampling (LHS) can be seen in Figure S.4.

Figure S.4. Visualisation of the reaction procedure with key process steps highlighted to show how the reaction procedure can be altered to fit to the Latin-hypercube plan, screenshotted from LabOS Data History tab.

S.2.2. Experimental Screening Design

The reaction procedure outlined in Section S.2.1 was applied across a range of independent variables. A Latin-hypercube sampling (LHS) method was employed to explore the design space, minimising bias and preventing localised searches. Variable bounds were established based on typical values for batch cooling crystallisations within the hardware's operational limits. Table S.4 presents the full experimental variable set-up with focus on the LHS (Figure S.5) values used as experimental points for cooling rate, seed mass and SS.

Table S.4.Experimental variable set-up and Latin hypercube sampling experimental design. The first row using square brackets defines the bounds of each independent variable.

eLN	Solute	Solvent	i_API Mass (g)	i_Solv Vol (L)	i_Conc (g/L solvent)*	Seed Slurry wt%	Cooling Rate (°C/min)	Seed Mass (%)	SS	Slurry Vol (mL)	Seed Temp (°C)
PICKL48	lamivudine	ethanol	35.51	0.75	47.34	20.00	0.27	2.34	1.48	5.30	59
PICKL44	lamivudine	ethanol	35.51	0.75	47.34	20.00	0.45	1.00	1.23	2.27	64
PICKL45	lamivudine	ethanol	35.51	0.75	47.34	20.00	0.25	4.84	1.38	10.93	61
PICKL46	lamivudine	ethanol	35.51	0.75	47.34	20.00	0.41	3.04	1.30	6.88	63
PICKL47	lamivudine	ethanol	35.51	0.75	47.34	20.00	0.14	3.80	1.33	8.59	62

*An initial concentration of 47.34 g/L solvent was fixed for the 'optimum experiment' also.

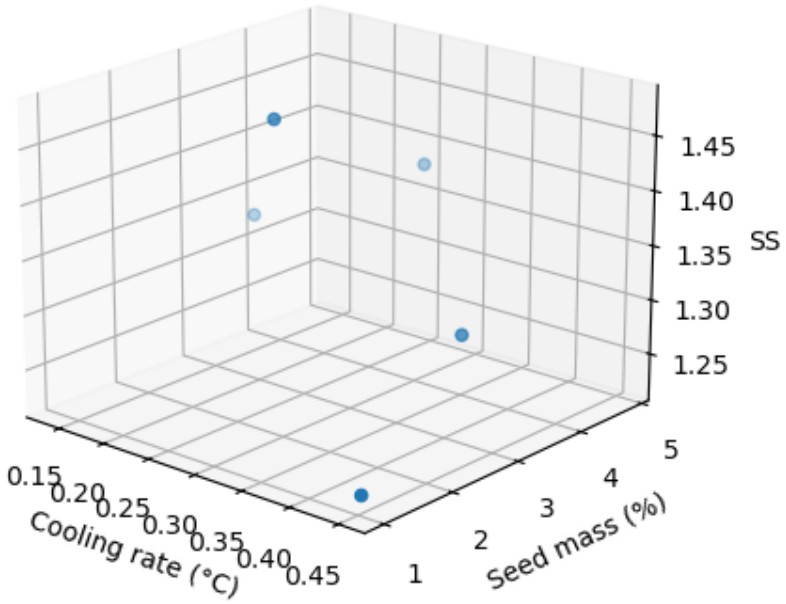

Figure S.5. 3-D plot visualisation of the Latin hypercube sampling experimental design.

To enable comparison in the main manuscript, the design space can be discretised into a grid of 480,000 experimental points, based on the defined bounds of the input parameters and a precision level of two decimal places.

S.2.3. Bayesian (Data-Driven) Optimisation

These three equations constitute the objective function used for Bayesian optimisation. The coefficient of determination (R^2) for Equation 2 and Equation 3 was calculated to assess the uncertainty in the kinetic estimates.

$$Yield(\%) = \frac{Conc._{init} - Conc._{end}}{Conc._{init}} \times 100$$

Equation 1.

where the concentration is liquid phase and measured via HPLC sampling with an 80-fold dilution with ethanol and has units in mg/mL solvent.

$$R_{growth} = \frac{particle\ size}{t}$$

Equation 2.

where the particle size is the Blaze-SW-mean-1-880 particle size and t is time in minutes during the entire cooling ramp post seeding. This measurement is a square-weighted mean average of particles measured via chord length distribution (CLD) in the range of 1 to 880 µm.

$$R_{nucleation} = \frac{particle\ count}{t}$$

Equation 3.

where the mean particle size is the Blaze-LW-counts-5-120 particle count and t is time in minutes during the entire cooling ramp post seeding. This measurement is a length-weighted count of particles in the range of 5 to 120 µm. This bin size was chosen due to the observed size of the particles of interest.

The data preprocessing and Bayesian optimisation code was developed in Python and executed using Visual Studio Code (Microsoft). Independent variables were imported via Pandas and converted to a NumPy-compatible format. Dependent variables, obtained from the five LHS experiments, were also imported, normalised, and used to calculate the objective function value before conversion to a NumPy-readable format. Bayesian optimisation was performed using the GPyOpt module with the following specifications:

- A simple one-dimensional function as the input data.
- The domain set to the bounds of the initial LHS for the independent variables.
- A Gaussian process probabilistic model.
- An expected improvement acquisition function.
- Exploitation-focused acquisition jitter.
- Emphasis on maximising the objective function value.
- Experimental inputs included X (independent variables) and Y (objective function value) calculated from measured data.
- Default settings were maintained for parameters such as normalisation of Y, evaluator type, and the number of cores.

These details are implemented in Python, as shown in Item S.1.

Item S.1. Bayesian optimisation code wrote in Python.

```
#BAYESIAN
seed(123)
def f(x):
    return x

bounds = [{'name': 'Cooling Rate (°C/min)', 'type': 'continuous', 'domain': (0.1, 0.5)},
          {'name': 'Seed Mass (%)', 'type': 'continuous', 'domain': (1, 5)},
          {'name': 'SS', 'type': 'continuous', 'domain': (1.2, 1.5)}]

bo_step = GPyOpt.methods.BayesianOptimization(f = f,
                                              domain=bounds,
                                              model_type='GP',
                                              acquisition_type ='EI',
                                              acquisition_jitter = 0.1,
                                              maximize = True,
                                              X=var,
                                              Y=obj)

x_next = bo_step.suggest_next_locations()
```

S.3. Results

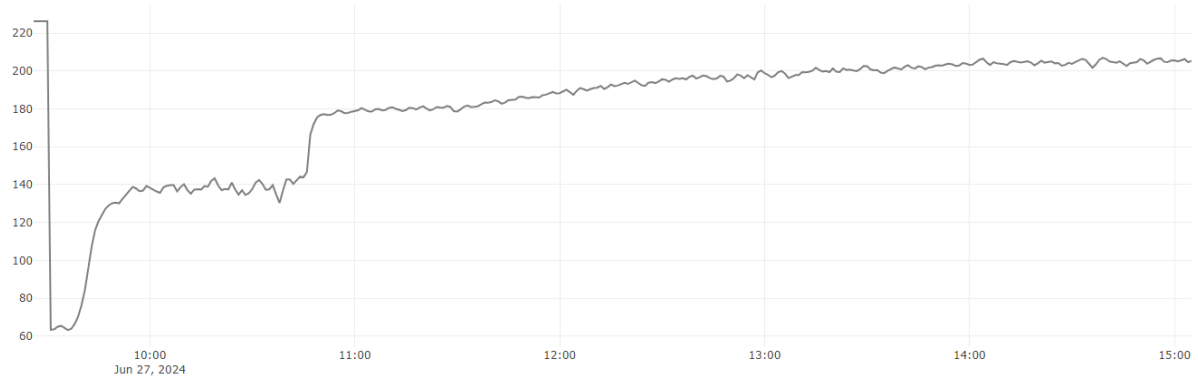

Figure S.6. Example trend for the Blaze Square Weighted mean particle size from the first experiment in the LHS.

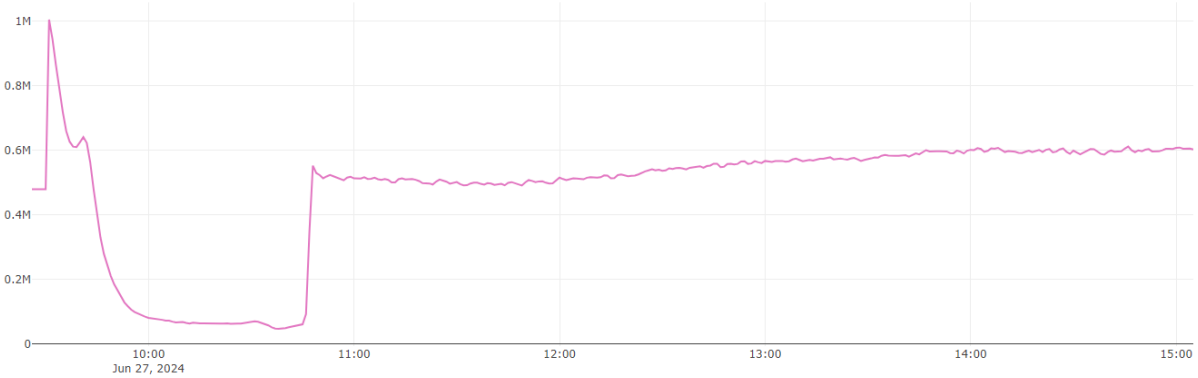

Figure S.7. Example trend for the Blaze Length Weighted counts 5-120 micron bin from the first experiment in the LHS.

The results of the five LHS experiments, particularly the effects of the process parameters on the measured outcomes, are visualised using a covariance matrix in Figure S.8. A strong positive correlation is observed between supersaturation (SS) at the point of seeding and nucleation rate, aligning with known crystallisation kinetics. However, unexpected correlations between SS and growth rate, cooling rate and growth rate, and seed mass and growth rate suggest limitations in using a small experimental sample size for multivariate analysis. Nonetheless, given the study's focus being on kinetic estimation, hardware application, and Bayesian optimisation, the sample size is considered adequate.

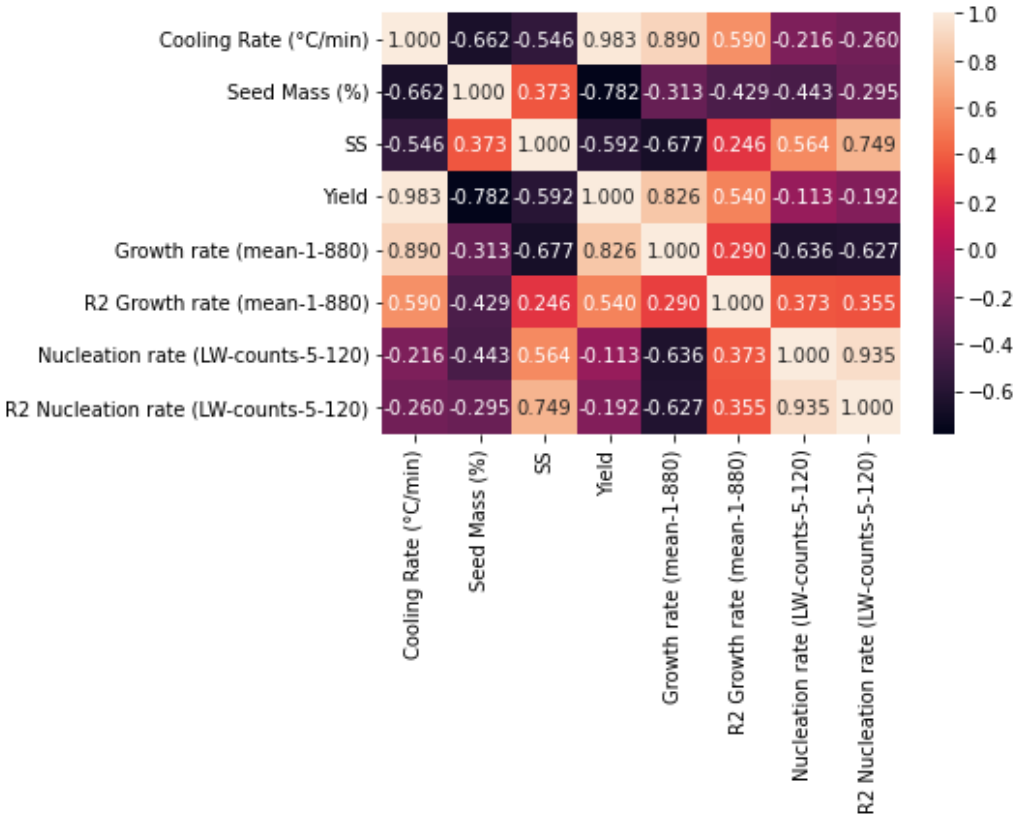

Figure S.8. Covariance matrix of the Latin-hypercube sampling experiments plotted using Seaborn.

Compared to the LHS experiments, the 'optimum experiment' achieved a 7% higher objective function than the best LHS result, a 46% increase over the LHS average, and a 107% increase over the lowest LHS outcome (Figure S.9). This improvement is attributed to a Pareto balance of yield, growth and nucleation rates and also due to an increased confidence in the kinetic parameters and thus the model.

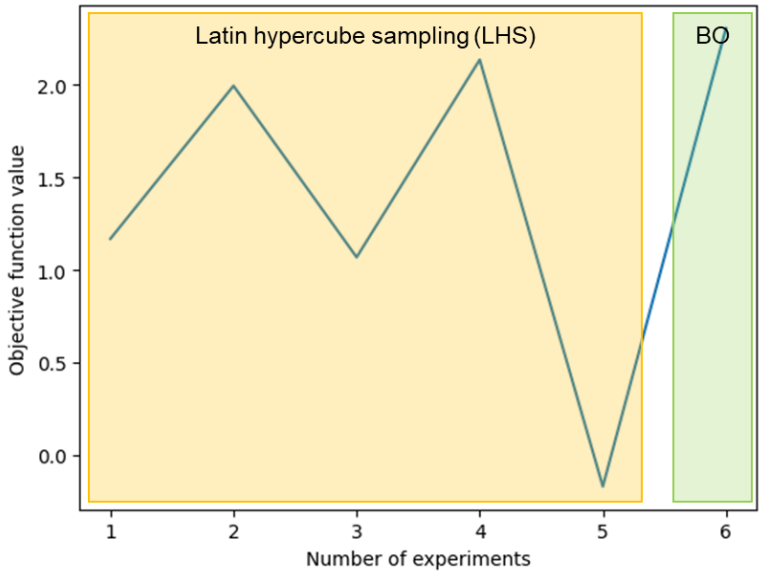

Figure S.9. Objective function value vs number of experiments, highlighted for the 5 Latin-hypercube sampling (LHS) experiments and the optimum next-best experiment predicted via Bayesian optimisation (BO).