UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Head Office: Università degli Studi di Padova

Department of Industrial Engineering

---

Ph.D. COURSE IN: Industrial Engineering

CURRICULUM: Chemical and Environmental Engineering

SERIES: XXXVI

# DEVELOPMENT OF MODEL-BASED METHODS TO STREAMLINE RESEARCH AND DEVELOPMENT IN THE PHARMACEUTICAL INDUSTRY

**Coordinator**: Prof. Giulio Rosati

**Supervisor**: Prof. Pierantonio Facco

**Ph.D. student**: Francesca Cenci

ACADEMIC YEAR 2022 – 2023

# Foreword

The fulfillment of the research results included in this Dissertation involved the intellectual and financial support of many people and institutions, to whom the author is very grateful.

Most of the research activity that led to the results reported in this Dissertation has been carried out at CAPE-Lab, Computer-Aided Process Engineering Laboratory, at the Department of Industrial Engineering of the University of Padova (Italy), under the supervision of Prof. Pierantonio Facco, with the help of Prof. Massimialino Barolo and Prof. Fabrizio Bezzo.

Part of the work has been carried out in collaboration with Dr. Simeone Zomer, Dr. Gabriele Bano, Dr. Charalampos Christodoulou, Dr. Yuliya Vueva, Dr. Samir Diab, Dr. Paola Ferrini, Dr. Konstantinos Stamatopoulos and Ms Katy Harabajiu from GSK, Ware and Stevenage (UK).

Part of the work has been carried out at University College London (UK) during a 6-month stay under the supervision of Prof. Federico Galvanin. Others collaborators were: Prof. Asterios Gavriilidis, Dr. Arun Pankajakshan and Mr. Solomon Gajere Bawa.

Financial support has been provided by the University of Padova and by *Fondazione Ing. Aldo Gini*, Padova (Italy).

DISCLOSURE STATEMENT

All the material reported in this Dissertation is original, unless explicit references to studies carried out by other people are indicated. In the following, a list of publications stemmed from this project is reported.

CONTRIBUTIONS IN INTERNATIONAL JOURNALS

Cenci, F., Pankajakshan, A., Bawa, G., Facco, P. and Galvanin, F. (2023). An exploratory model-based design of experiments technique to aid parameters identification and reduce prediction uncertainty. *Computers and Chemical Engineering*, **177**, 108353.

Cenci, F., Bano, G., Christodoulou, C., Vueva, Y., Zomer, S., Barolo, M., Bezzo, F. and Facco, P. (2022). Streamlining tablet lubrication design via model-based design of experiments. *International Journal of Pharmaceutics*, **614**, 121435.

CONTRIBUTIONS IN INTERNATIONAL JOURNALS (in preparation)

Cenci, F., Pankajakshan, A., Bawa, S. G., Gavriilidis, A., Facco, P. and Galvanin, F.. Explorative optimal experimental design for the identification of total methane oxidation kinetics in automated microreactor platforms.

Cenci, F., Pankajakshan, A., Bawa, S. G., Gavriilidis, A., Facco, P. and Galvanin, F.. Novel algorithm for the autonomous execution of G-map eMBDoE experiments by means of automated chemical platforms.

Cenci, F., Diab, S., Harabajiu, K., Ferrini, P., Barolo, M., Bezzo, F. and Facco, P.. Machine-Learning approach based on group contributions for the prediction of solubility of drug and drug-like molecules in organic solvent mixtures.

Cenci, F., Stamatopoulos, K., Diab, S., Ferrini, P., Barolo, M., Bezzo, F. and Facco, P.. Machine-Learning approach to represent food effect and inter- and intra-subject variability of intestinal solubility.

CONTRIBUTIONS IN PEER-REVIEWED CONFERENCE PROCEEDINGS

Cenci, F., Pankajakshan, A., Bawa, S. G., Gavriilidis, A., Facco, P., Galvanin, F., 2023. An exploratory model-based design of experiments technique to aid parameters identification and reduce prediction uncertainty. In 33 European Symposium on Computer Aided Process Engineering, Kokossis, A. C., Georgiadis, M. C., Pistikopoulos, E. Eds., Comput. Aided Chem. Eng., Elsevier, **52**, 1-6.

Cenci, F., Bano, G., Christodoulou, C., Vueva, Y., Zomer, S., Barolo, M., Bezzo, F., Facco, P., 2022. Reducing the experimental effort to design pharmaceutical tablet lubrication by model-based design of experiments. In 32 European Symposium on Computer Aided Process Engineering, Montastruc, L., Negny, S., Eds., Comput. Aided Chem. Eng., Elsevier, **51**, 25-30.

CONFERENCE PRESENTATIONS

Cenci, F., Pankajakshan, A., Bawa, S. G., Gavriilidis, A., Facco, P., Galvanin, F., 2023. An exploratory model-based design of experiments technique to aid parameters identification and reduce prediction uncertainty [poster presentation]. 33 European Symposium on Computer Aided Process Engineering, Athens, Greece, June 18-21

Cenci, F., Pankajakshan, A., Galvanin, F., Facco, P., 2023. Trade-off between space exploration and information maximization in experimental design [oral presentation]. Colloquium Chemiometricum Mediterraneum, Padova, Italy, June 27-30

Cenci, F., Bano, G., Christodoulou, C., Vueva, Y., Zomer, S. Barolo, M., Bezzo, F., Facco, P., 2022. Development of model based strategies to accelerate the experimental campaign for the production of oral solid dosage through direct compression [oral presentation]. GRICU conference, Ischia, Italy, Jul 3-6

Cenci, F., Bano, G., Christodoulou, C., Vueva, Y., Zomer, S. Barolo, M., Bezzo, F., Facco, P., 2022. Reducing the experimental effort to design pharmaceutical tablet lubrication by model-based design of experiments [oral presentation]. 32 Symposium on Computer Aided Process Engineering (ESCAPE-32), Toulouse, France, Jun 12-15

# Abstract

The pharmaceutical industry has contributed positively to global longevity and well-being, and as well to the economy. However, several challenges must be faced by the pharmaceutical R&D, like the increasing difficulty in developing new treatments with significant advantages over the existing ones and the need to accelerate the entire R&D process, which often lasts more than half of the patent life and requires substantial investments of resources, labor and money. In this Dissertation, novel model-based methods are developed in order to streamline pharmaceutical R&D, while improving product and process understanding and ensuring product quality and process robustness. Specifically, the methods proposed in this Dissertation allow to: (*i*) streamline the design of tablets lubrication; (*ii*) minimise model prediction variance in the whole design space by means of a trade-off between space exploration and information maximisation, (*iii*) adapt the novel procedure for the minimisation of model prediction variance to the operation of an automated platform in order to streamline kinetic studies; (*iv*) implement a fully autonomous operation of a chemical platform with the aim of collecting experiments to estimate model parameters and minimise model prediction variance; (*v*) accurately predict drug solubility in mixtures of organic solvents; (*vi*) predict drug solubility in intestinal fluids, considering the effects of food and physiological factors.

**Steamlining the design of tablets lubrication**: Tablets lubrication is an important step in direct compression processes, for instance to prevent the powder from adhering on metal surfaces, however it cannot be excessive otherwise it degrades tablets manufacturability and properties (such as dissolution). An extended version of the Kushner and Moore model proposed by Nassar et al. (2021) is usually employed to predict tablets properties (tensile strength) based on tablets solid fraction and lubrication extent. Experiments to estimate model parameters are generated through a tablet press, where the same blend (i.e., the powder characterised by the same lubrication extent) is compressed multiple times at different compression pressures. However, too many blends (typically 7-9) are used in the experimentation to calibrate the model, thus causing an excessive waste of Active Pharmaceutical Ingredient (API). In this Dissertation, a novel model-based design of experiments (MBDoE) method is proposed to determine the most informative lubrication extent for the blend used in the experimentation,

namely, the lubrication extent of the blend that allows to maximise the precision of the estimated model parameters. The use of this method leads to a significant reduction of the experimental burden: only 3-4 blends are sufficient to obtain statistically precise parameters and satisfactory prediction accuracy for industrial applications. Therefore, the experimental burden is reduced by 60-70%.

**Minimise model prediction variance in the whole design space by means of a trade-off between space exploration and information maximisation:** state-of-the-art MBDoE methods for parameters precision select experimental conditions that minimise parameters uncertainty, but this does not necessarily lead to a minimisation of model prediction variance in the whole design space. Moreover, they tend to localise the experiments in regions of high information content (namely, leading to maximum parameters precision), but this may translate into a scarce exploration of the design space. Therefore, a novel exploratory MBDoE (*eMBDoE*) method is developed in order to minimise model prediction variance (G-optimality, $J_G$) in the whole design space with the minimum experimental burden, while ensuring parameters precision. This is achieved by selecting a first set of candidate design points based on a user-defined threshold $J_{G,\mathrm{thr}}$ of G-optimality, among which the information content is maximised to find the best experimental condition. This is practically implemented through maps of G-optimality. Therefore, the proposed method is named *G-map eMBDoE*. The performance of G-map eMBDoE is tested with two simulated systems (i.e.: an algebraic model with two inputs and one output; a differential equation model with two constant inputs and two dynamic outputs sampled at three sampling points) and compared to an information-based method, MBDoE, and to two exploration-based methods, namely Latin Hypercube sampling and statistical design of experiments. The results suggest that G-map eMBDoE, compared to the abovementioned state-of-the-art methodologies, is able to enhance space exploration with respect to MBDoE and to minimise model prediction variance in the whole design space in the most efficient way, while ensuring parameters precision.

**Adaptation of the novel procedure for the minimisation of model prediction variance to the operation of an automated platform in order to streamline kinetic studies:** to further confirm the advantages of the G-map eMBDoE, this method is applied to an automated platform for kinetic studies. The reaction considered is the total methane oxidation, but the procedure is applicable to any other model of interest in the (bio)pharmaceutical industry. Based on preliminary simulations on the system under study, a G-optimality constraint is selected; then,

MBDoE and two G-map eMBDoE designs (with two different thresholds $J_{G,\mathrm{thr}}$) are compared. The experimental results show that the most explorative G-map eMBDoE design (i.e., the one with the lower $J_{G,\mathrm{thr}}$ of the two) also leads to the greatest reduction of model prediction variance in the whole design space and to a minimisation of 78% of the experimental burden to obtain statistically precise model parameters.

**Fully autonomous operation of a chemical platform with the aim of collecting experiments to estimate model parameters and minimise model prediction variance**: based on the results of G-map eMBDoE with both simulated and experimental data, this method can be useful to be implemented in fully autonomous chemical platforms, with no human intervention, with the aim of generating highly informative experiments with the minimum experimental burden. However, to do so a method to automatically select the most suitable G-optimality constraint is needed. Therefore, an adaptative G-map eMBDoE is proposed, which re-determines the best G-optimality constraint as soon as a new experiment is measured, without human intervention. This is done by analysing the overlap between maps of information content and maps of model prediction variance: if the most informative points have the highest G-optimality, space exploration is enhanced if points with lower G-optimality are favored; if the most informative points have the lowest G-optimality, space exploration is enhanced if points with higher G-optimality are favored. For a fair comparison with the original G-map eMBDoE method, the adaptative G-map eMBDoE is applied to the two simulated systems (namely, both the algebraic and the differential equations model) previously used with G-map eMBDoE. The results show that the novel method is able to find a satisfactory trade-off between space exploration and information maximisation. Even if its results may be less efficient than the ones obtained by changing manually $J_{G,\mathrm{thr}}$ and selecting the most favorable threshold $J_{G,\mathrm{thr}}$ at the end of the simulated experimental campaign, it is still significantly better than the worst scenario encountered when an unfavorable $J_{G,\mathrm{thr}}$ is manually selected. Considering that the results of the adaptative G-map eMBDoE are achieved without requiring any human intervention, they suggest that the proposed method is ready to be implemented in a fully autonomous chemical platform, thus fully exploiting Industry 4.0 technologies.

**Accurately predicting drug solubility in mixtures of organic solvents** : a data-driven model is proposed to improve the prediction of drug solubility in mixtures of organic solvents. Little information is required to predict solubility: (*i*) temperature; (*ii*) mixture composition before API dissolution; (*iii*) UNIFAC subgroups associated to the solvents in the mixture. Such input

variables can be correlated among each other, but correlation is handled by the use of Partial Least Square (PLS). To test the proposed model with experimental data, a real drug substance and 14 organic solvents commonly used in crystallisation units are used and a high-throughput technology with 96-vials plates is employed. The PLS model is calibrated with solubility measurements in the single solvents and in a few binary mixtures at two temperatures (20 and 40°C) and it is validated with: binary mixtures with the same pair of solvents used in calibration, but at a different composition; binary mixtures with different pairs of solvents; ternary mixtures; data at a higher temperature (50°C). The PLS method provides satisfactory predictions both in calibration and validation, as shown by the coefficient of determination equal to 0.92 and 0.90, respectively. Finally, the same modelling approach is applied to 9 datasets from the Literature involving drug and drug-like molecules and the majority of the calibration and validation datasets lead to a coefficient of determination between 0.95 and 0.99.

**Predict drug solubility in intestinal fluids, considering the effects of food and physiological factors**: the safety and efficacy of solid oral dosage forms depend on their bioavailability, namely on the fraction of drug that reaches the systemic circulation. Only the drug that is dissolved in the intestinal fluids can pass through the gut wall and reach the bloodstream. Therefore, solubility in intestinal fluids is a key property that must be assessed from the early stages of drug development. However, this assessment is complicated by the fact that a high level of variability is present due to differences in physiological factors within one individual (e.g., variations in pH, bile salts concentration and food digestion products concentrations due to fasted and fed conditions) and among different subjects (e.g., due to age, sex and ethnicity). Moreover, human intestinal fluids of treated patients cannot be sampled extensively, therefore in-vitro biorelevant media must be employed to study drug solubility. The in-vitro data can then be used to calibrate pharmacokinetic (PK) models describing the interaction between the drug and the human body. In fact, model-based PK simulations with commercial software are extremely useful because they allow to simulate a variety of scenarios and to generate virtual populations with intra- and inter-subject variability. In this Dissertation, a Gaussian Process model is developed to support PK studies: (*i*) it is able to represent food effects, improving the accuracy of the prediction of in-vitro data with respect to standard models available in commercial software; (*ii*) it has a structure that allows to integrate it within a commercial PK software that simulates multiple individuals with different physiological factors. Thus the

improved prediction of drug solubility will allow to improve the overall simulation of physiologically-based pharmacokinetics, taking into account inter- and intra-subject variability.

# List of Symbols

## Acronyms

| | |
|---|---|
| ADME | Absorption, distribution, metabolism and excretion |
| AE | absolute error |
| API | Active Pharmaceutical Ingredient |
| BS | Bile salts |
| cGMP | current Good Manufacturing Practices |
| CH | cholesterol |
| CI | confidence interval |
| $CI_{GP}$ | confidence interval of the GP model |
| $CI_{PLS}$ | confidence interval of the PLS model |
| CMA | Critical material attribute |
| CMC | critical micelles concentration |
| CPP | Critical process parameter |
| CQA | Critical quality attribute |
| CTA | clinical trial application |
| DoE | design of experiments |
| DS | Design space |
| EFPIA | European Federation of Pharmaceutical Industries Associations |
| EMA | European Medicines Agency |
| eMBDoE | exploratory MBDoE |
| FaSSIFs | fasted SIFs |
| FDA | Food & Drug Administration |
| FIM | Fisher Information Matrix |
| GC | Gas chromatography |
| GC-MSD | Gas chromatography with mass detector |
| GDC | sodium glycodeoxy-cholate |
| GLC | sodium glycocholate hydrate |
| G-map | mapping of G-optimality |
| GP | Gaussian Process |
| HIF | Human intestinal fluids |
| H-map | mapping of FIM-based information |
| HPLC | High-performance liquid chromatography |
| HTE | High-throughput experimentation |
| ICH | International Council on Harmonization |
| IND | investigational new drug application |
| IoT | Internet of Things |
| IRR | Internal Rate of Return |
| IVGTT | Intravenous glucose tolerance tests |
| IVIV | In vitro in vivo |
| LB | lower bound |
| LH | Latin Hypercube |
| LV | latent variables |
| MBDoE | model-based design of experiments |

| | |
|---|---|
| MCC | microcrystalline cellulose |
| ML | Machine Learning |
| MLR | Multivariate Linear Regression |
| MS | Mass spectrometers |
| NDA | New Drug Application |
| OA | oleic acid |
| OFAT | One-factor-at-a-time |
| OSD | oral solid dosage |
| PAT | Process analytical technology |
| PBPK | physiologically-based pharmacokinetics |
| PC | lecithin, namely L-alpha-phosphatidylcholine |
| PD | pharmacodynamics |
| PK | pharmacokinetics |
| PLS | Partial Least Squares |
| QbD | Quality-by-Design |
| QbT | Quality-by-Testing |
| QSAR | Quantitative structure–activity relationship |
| QSPR | Quantitative structure–property relationship |
| QTPP | Quality target product profile |
| R&D | research and development |
| RMG | Reaction Mechanism Generator |
| RMSE | root mean squared error |
| SE | squared exponential function |
| $SE_{PLS}$ | standard error of the PLS model |
| SF | solid fraction |
| SIF | Simulated intestinal fluids |
| SMILES | Simplified Molecular Input Line Entry System |
| SPE | squared prediction error |
| $SPE_{lim}$ | confidence limit of the squared prediction error |
| TC | sodium taurocholate hydrate |
| TCDC | sodium taurochenodeoxycholate |
| TS | tensile strength |
| UB | upper bound |
| UPLC | Ultra- high-performance liquid chromatographic systems |
| VIP | variable importance in projection |

## Symbols

| | |
|---|---|
| A | number of latent variables |
| $a_1$, $a_2$ | extended Kushner and Moore parameters |
| $b_1$, $b_2$ | extended Kushner and Moore parameters |
| $[BS]_{app}$ | apparent bile salts concentration |
| $\mathrm{cov}(\mathbf{f}_*)$ | covariance function of the noise-free predictions |
| det(·) | matrix determinant |
| $d$ | degree of freedom of the PLS model |
| $d_{ij}$ | distance metrics between two designs, $i$ and $j$ |
| $D$ | tablet diameter |
| $\mathbb{E}[\cdot]$ | expected value |
| **E** | matrix of residuals for the input matrix of a PLS model |

| | |
|---|---|
| $E_{a,1(/2/3)}$ | parameters of the kinetic model of total methane oxidation |
| $\mathbf{f}$ | set of model equations |
| $\bar{\mathbf{f}}_*$ | mean function at test conditions |
| $f_i^{\mathrm{S(/L)}}$ | fugacity of the $i$-th compound in the solid (or liquid) phase |
| $\mathbf{F}$ | matrix of residuals for the output matrix of a PLS model |
| $F$ | breaking force |
| $F_h$ | fraction occupied by the headspace |
| $F_{G,\mathrm{min(/mean/max)}}^{\psi\gg}$ | minimum (or mean or maximum) fraction of G-optimality for the design points having the highest information content, with respect to the maximum G-optimality calculated in the whole design space |
| $g_k$ | effect of the $k$-th subgroup in the mixture |
| $\mathcal{GP}$ | Gaussian Process model |
| $\mathbf{h}$ | set of equations of measurable responses |
| $h_n$ | leverage of the $n$-th observation of the PLS model |
| $\mathbf{H}_{\widehat{\boldsymbol{\theta}}}$ | Fisher Information Matrix of a general model with estimated parameters $\widehat{\boldsymbol{\theta}}$ |
| $J_G$ | scalar value summarising the G-optimality values calculated for all responses and all time points at a specific point in the design space |
| $J_{G,\mathrm{min(/mean/max)}}$ | minimum (or mean or maximum) values of $J_G$ calculate in the whole design space |
| $J_{G,\mathrm{thr}}$ | thresholds of G-optimality |
| $J_{G,\mathrm{thr,prior}}$ | G-optimality threshold based on prior knowledge on the system |
| $J_{G,\mathrm{thr,mean}}$ | G-optimality threshold based on measured preliminary data |
| $J_{G,\mathrm{min(/mean/max)}}^{\psi\gg}$ | minimum (or mean or maximum) scalar measure of G-optimality calculated for the most informative points |
| $k_{1(/2/3),\mathrm{ref}}$ | parameters of the kinetic model of total methane oxidation |
| $K(\cdot,\cdot)$ | matrix of covariances between two specified matrices |
| $K_{m:w,u(/i)}$ | water-to-micelle partition coefficients of unionised (or ionised) species |
| $\ell$ | hyperparameter of the squared exponential function |
| $L(\widehat{\boldsymbol{\theta}})$ | negative log-likelihood function |
| $m(\cdot)$ | prior mean function of the GP model |
| $m_T$ | tablet weight |
| $N$ | total number of experimental measurements |
| $\mathcal{N}$ | normal distribution |
| $N_e$ | number of performed experiments |
| $N_k$ | number of UNIFAC subgroups in the mixture |
| $N_{\mathrm{sp}}$ | number of sampling points considering all the performed experiments |
| $N_{\mathrm{sp}_i}$ | number of sampling points in the $i$-th experiment |
| $N_u$ | number of control variables |
| $N_x$ | number of state variables |
| $N_y$ | number of response variables |
| $N_\theta$ | number of model parameters |

| | |
|---|---|
| $N_{\varphi}$ | number of possible experimental conditions within the discretised design space |
| $N_K$ | number of lubrication extents |
| $N_L$ | number of liquid organic solvents in the mixture |
| $N_{\mathrm{reg}}$ | number of fictitious data points |
| $N_{SF}$ | number of solid fraction values |
| $N_*$ | number of test (or validation) data |
| $\mathbf{P}$ | matrix of loadings of the input matrix |
| $\mathrm{pK}_{a,1(/2)}$ | dissociation constant of the acid (or base) |
| $\mathbf{Q}$ | matrix of loadings of the output matrix |
| $r$ | kinetic rate |
| $R^2$ | coefficient of determination |
| $R_a^2$ | amount of $y$ variance explained by the $a$-th latent variable |
| $s$ | standard deviation of the PLS model |
| $\mathbf{s}$ | row-vector of sensitivity indices |
| $\mathbf{S}$ | sensitivity matrix |
| $S_0$ | intrinsic solubility |
| $S_{\mathrm{BS,u(/i)}}$ | bile-salt mediated enhancement of solubility of the unionized (or ionized) species |
| $S_{\mathrm{pH}}$ | overall pH-dependent solubility |
| $S_{\mathrm{pH,i}}$ | pH-dependent solubility of ionised species |
| $S_T$ | total solubility |
| $t$ | time |
| $\mathbf{T}$ | matrix of scores |
| $t_{\mathrm{blend}}$ | blending time |
| $t_i$ | *t*-value for the $i$-th parameter |
| $\mathbf{t}_{\mathrm{sp}}$ | vector of sampling points |
| $t_{\mathrm{ref}}$ | reference *t*-value |
| $T^2$ | Hotelling statistics |
| $T^2_{\mathrm{lim}}$ | confidence limit of the Hotelling statistics |
| $ts_{sf=0.85,0}$ | initial tensile strength at 0.85 solid fraction |
| $\mathbf{u}, \mathbf{u}_{\mathrm{LB}}, \mathbf{u}_{\mathrm{UB}}$ | vector of control variables, lower bounds and upper bounds of control variables |
| $\mathbf{U}$ | matrix of input observations |
| $V$ | number of regressors of PLS models |
| $\mathbf{V}_y$ | matrix of model prediction variances calculated by means of the G-optimality definition |
| $\mathbf{V}_{\hat{\boldsymbol{\theta}}}$ | variance-covariance matrix |
| $\mathbf{V}^0_{\hat{\boldsymbol{\theta}}}$ | prior variance-covariance matrix of model parameters |
| $V_b$ | blender volume |
| $V_c$ | cup volume |
| $W$ | wall height of the tablet |
| $\mathbf{W}$ | matrix of weights |
| $\mathbf{x}$ | vector of state variables |
| $x_{\mathrm{mol},i}$ | molar fraction of the *i*-th compound in the PLS model |
| $\dot{\mathbf{x}}$ | vector of first derivatives of state variables |
| $\mathbf{y}$ | vector of measurable model responses |

| | |
|---|---|
| $\mathbf{Y}$ | matrix of observed response variables |
| $\hat{\mathbf{y}}$ | vector of measurable model responses predicted by the model |
| $\hat{\mathbf{y}}_{\mathrm{GP}}$ | prediction of the Gaussian Process model |
| $\hat{\mathbf{y}}_{\mathrm{LR}}$ | prediction of the linear regression model |

## Greek letters

| | |
|---|---|
| $\alpha$ | significance level |
| $\alpha_{\mathrm{equip}}$ | equipment dependent factor |
| $\gamma$ | lubrication rate constant of the blend |
| $\gamma_i$ | activity coefficient of the $i$-th compound |
| $\gamma_i^{C(/R)}$ | combinatorial (or residual) contribution to the activity coefficient of the $i$-th compound |
| $\boldsymbol{\beta}$ | parameters of a linear regression model |
| $\varepsilon$ | error of the mathematical model |
| $\boldsymbol{\theta}$ | vector of model parameters |
| $\widehat{\boldsymbol{\theta}}$ | vector of estimated model parameters |
| $\boldsymbol{\theta}_0, \widehat{\boldsymbol{\theta}}_{\mathrm{true}}, \boldsymbol{\theta}_{\mathrm{LB}}, \boldsymbol{\theta}_{\mathrm{UB}}$ | vector of initial parameters values, true parameters values, lower bounds (LB) and upper bounds (UB) for parameters estimation |
| $\boldsymbol{\Theta}$ | vector of model parameters when variables are scaled |
| $\kappa(\cdot,\cdot)$ | kernel function |
| $\kappa_{\mathrm{cond}}, \kappa_{\mathrm{cond,max}}$ | condition number, maximum condition number |
| $\lambda_a, \lambda_{\mathrm{min(/max)}}$ | $a$-th eigenvalue, maximum (or maximum) eigenvalue |
| $\boldsymbol{\Lambda}^{-1}$ | matrix with the inverse of the eigenvalues of the PLS model |
| $\mu$ | mean |
| $\mu_{\mathrm{cal}}$ | mean of the residuals of the calibration dataset of the PLS model |
| $\upsilon_{ik}$ | number of occurrences of the $k$-th subgroup in the $i$-th solvent |
| $\rho_t$ | true density of the powder blend |
| $\boldsymbol{\sigma}_y$ | standard deviation of the response measurement error |
| $\sigma_{\mathrm{cal}}$ | variance of the residuals of the calibration dataset of the PLS model |
| $\sigma_{\mathrm{SE}}^2$ | hyperparameter of the squared exponential function |
| $\boldsymbol{\Sigma}_y$ | response variance-covariance matrix |
| $\sigma_y^2$ | response variance |
| $\boldsymbol{\varphi}$ | design vector |
| $\boldsymbol{\varphi}_{\mathrm{cand}}$ | candidate design points |
| $\boldsymbol{\varphi}_{\mathrm{opt}}$ | design vector solving the MBDoE optimisation problem |
| $\boldsymbol{\varphi}_{\mathrm{opt}}^*$ | reference design (e.g., MBDoE design) used to determine $J_{G,\mathrm{thr,meas}}$ |
| $\boldsymbol{\varphi}^{\psi\gg}$ | design points with the highest information content |
| $\chi_y^2$ | statistics for the $\chi^2$-test calculated for a given dataset |
| $\chi_{\mathrm{ref}}^2$ | reference value for the $\chi^2$-test |
| $\psi, \psi_{\mathrm{min(/max)}}$ | scalar measure of the FIM, minimum (or maximum) scalar measure of the FIM |
| $\omega_{\mathrm{blend}}$ | mixer rotational speed |

# Table of contents

# Chapter 1
# Motivation and state of the art

This Chapter provides an overview of the main features of research and development (R&D) in the pharmaceutical industry. First, the socio-economic framework is described, highlighting strengths and difficulties of this sector. Then, the main steps from drug discovery to the launch of the final product are illustrated, focusing the attention on the main challenges concerning both product and process development. The regulatory framework is then described, and its impact on R&D and manufacturing explained. Moreover, two major contributions to the innovation of the pharmaceutical industry are discussed: Quality by Design (QbD) initiatives and Industry 4.0 new technologies. Opportunities and challenges of both are highlighted, together with some relevant examples from the literature. Finally, the objectives of this work are presented and the roadmap of this Dissertation is provided.

## 1.1 Motivation of the State-of-the-Art

The research activities presented this Dissertation (Chapters 3-8) focus on the reduction of time, labor and resources for the pharmaceutical R&D, while ensuring product quality and process robustness. Chapter 1 explains why this work provides a valuable contribution to the pharmaceutical sector:

- First, the socio-economic framework is illustrated in order to explain why the pharmaceutical industry deserves attention, thanks to the positive contribution to the well-being and economic wealth worldwide, and why it requires reduced costs to maintain a satisfactory return on investment.
- Then, the different steps from the discovery of a new molecular entity to the launch of the final product are explained, in order to highlight the necessity to: (*i*) accelerate the R&D timeline for a better patent exploitation; (*ii*) propose novel methods for a consistent delivery of high-quality products; (*iii*) propose novel methods to improve process robustness and flexibility.

- However, the pharmaceutical R&D can be innovated only considering the boundaries set by the regulatory framework. Therefore, the role of regulatory agencies and the consequences to the pharmaceutical sector are explained.
- Afterwards, recent incentives (i.e., Quality-by-Design) to favour innovation are explained, focusing on the two aspects that are most relevant for this Dissertation: mathematical models and statistical design of experiments (DoE).
- Statistical DoE is established in the pharmaceutical industry nowadays, with the advantage of modernising the experimental approach with respect to previous trial and error approaches, but it may be suboptimal for the development of models different from regression ones. Therefore, another method to design experiments is presented: model-based design of experiments (MBDoE). State-of-the-art applications of MBDoE to the pharmaceutical R&D are reviewed in detail to better explain the innovations of the MBDoE methods presented in this Dissertation.
- In addition, Industry 4.0 has provided useful tools to streamline pharmaceutical R&D, therefore the overall framework of Industry 4.0 technologies is described. Then, high-throughput and/or automated technologies are explained in detail since they are used in this Dissertation: (*i*) in fact, partially automated high-throughput technologies are used to generate the experimental data of Chapters 7-8; (*ii*) an automated microreactor platform is used to experimentally validate the new MBDoE method proposed in Chapter 5; (*iii*) a new MBDoE algorithm is developed in Chapter 6 with the purpose of allowing a fully autonomous decision-making in the platform.

Finally, the objectives of this Dissertation are explained and the roadmap illustrated.

## 1.2 Socio-economic framework of the pharmaceutical industry

The pharmaceutical industry is concerned with the research, development, manufacturing and distribution of life-saving drugs and medical treatments. This sector gives a valuable contribution to the global economy on many levels: pharmaceutical companies pay salaries, support suppliers and pay taxes (www.efpia.eu/more-than-medicine). Different professionals are employed in these companies for a variety of roles: (*i*) in product development, for instance in laboratory-based R&D, clinical trials and regulatory affairs; (*ii*) in product manufacturing, such as process development and quality control; (*iii*) in the commercialisation of the final product, involving marketing, sales and distribution (Getreskilled, 2022). Moreover, revenues

of the pharmaceutical companies have had a positive trend in recent years: in 2014, total pharmaceutical revenues at global level exceeded 1 trillion United States dollars (USD) for the first time (González Peña, 2021), while the total global pharmaceutical market was estimated at 1.48 trillion US dollars in 2022 (Statista, 2023). One of the largest pharmaceutical markets is US and it is expected to grow from $567 billions in 2022 to $903 billions by 2030 (Insights10, 2022). Another important pharmaceutical market is United Kingdom, which is among the 10 top markets at global level according to Statista (2022). In fact, UK holds 2.6 percent of the global pharmaceutical sector and its two main companies, namely GSK and AstraZeneca, had a market capitalisation of 57 and 170 billion British pounds, respectively, at late 2022 (with the latter one being boosted by the development of the COVID-19 vaccine in collaboration with Oxford University). Positive economic trends are also encountered in the so-called *pharmerging* countries, namely countries with a low position in the rank of pharmaceutical markets, but having a rapid growth, such as India, China, South Africa, Brazil, Russia, Indonesia and Turkey. According to IMARC Group (2022), their global market size reached US$ 1.1 billions in 2022 and it is expected to grow up to US$ 2.2 billions by 2028.

Besides the benefits for the global economy, the innovative products of the pharmaceutical industry have greatly contributed to the well-being and increase of longevity worldwide. For instance, Buxbaum et al. (2020) considered the increase of 3.9 years of life expectancy between 1990 and 2015 in the US and isolated the factors contributing to at least 0.1 years of variation in the life expectancy. They obtained 12 categories overall (namely, diseases related to the circulating system, malignant neoplasms, traumas, neurological neoplasms and others), which explained the increase of longevity of 2.9 years; in turn, this increase in longevity was attributable to public health by 44%, to pharmaceuticals by 35% and to other medical care by 13%. This positive result is a further incentive to promote and improve the R&D activities in the pharmaceutical sector. Coherently, the 2022 report by the European Federation of Pharmaceutical Industries Associations (EFPIA) highlighted that the European citizens have an expectancy of life that is 30 years longer than a century ago and their quality of life has greatly improved thanks to innovative treatments against a variety of diseases, such as some cancer types, HIV and cardiovascular diseases. To further reduce mortality and improve quality of life, it is important to keep investing on pharmaceutical R&D and, also, to align the research with unmet patients' needs, especially treatments against Alzheimer's, Multiple Sclerosis, many cancers, and rare diseases (EFPIA, 2022; Panteli and Edwards, 2018).

Despite the positive economic trend and the importance of innovative pharmaceutical treatments for the life of patients worldwide, there are several challenges to be faced in the pharmaceutical R&D. Deloitte has published annual reports titled "Measuring the return from pharmaceutical innovation" from 2010 onwards; the trend of Internal Rate of Return (IRR) per year is decreasing in the period analysed: besides the positive peak in 2021, likely due to COVID-19 vaccines, the IRR decreased to 1.2% in 2022, reaching the lowest value in the time lapse considered (Figure 1.1a). A contribution to the decreasing IRR is given by the increasing costs and duration of pharmaceutical R&D: Deloitte's report 2022 highlights that the total R&D expenditure increased from \$85.5 billions in 2013 to \$139.2 billions in 2022 (Figure 1.1b); moreover, 12-13 years can elapse between the discovery of a new active molecule and the launch of the final product in the market (Destro and Barolo, 2022). More details on product and process development and on their challenges are provided in Section 1.2.


**Figure 1.1.** *Economic trend of the pharmaceutical industry in terms of: a) IRR [%] from 2013 to 2022; b) total R&D expenditure from 2013 to 2022 (adapted from Deloitte, 2023).*

## 1.3 New drug: from R&D to marketing

The launch of a new drug is made of three main steps, as illustrated in Figure 1.2: (*i*) research, (*ii*) development and (*iii*) manufacturing. It is a long process, typically lasting 12-13 years,

therefore more than half of the patent life that starts when a new active substance is synthesised and expires 20 years later (EFPIA, 2020). In fact, pharmaceutical companies need to streamline every step of this long procedure in order to exploit the patent to the fullest and to repay the investments made on the successful drug, as well as on the candidates that have not passed the pre-clinical and clinical trials.

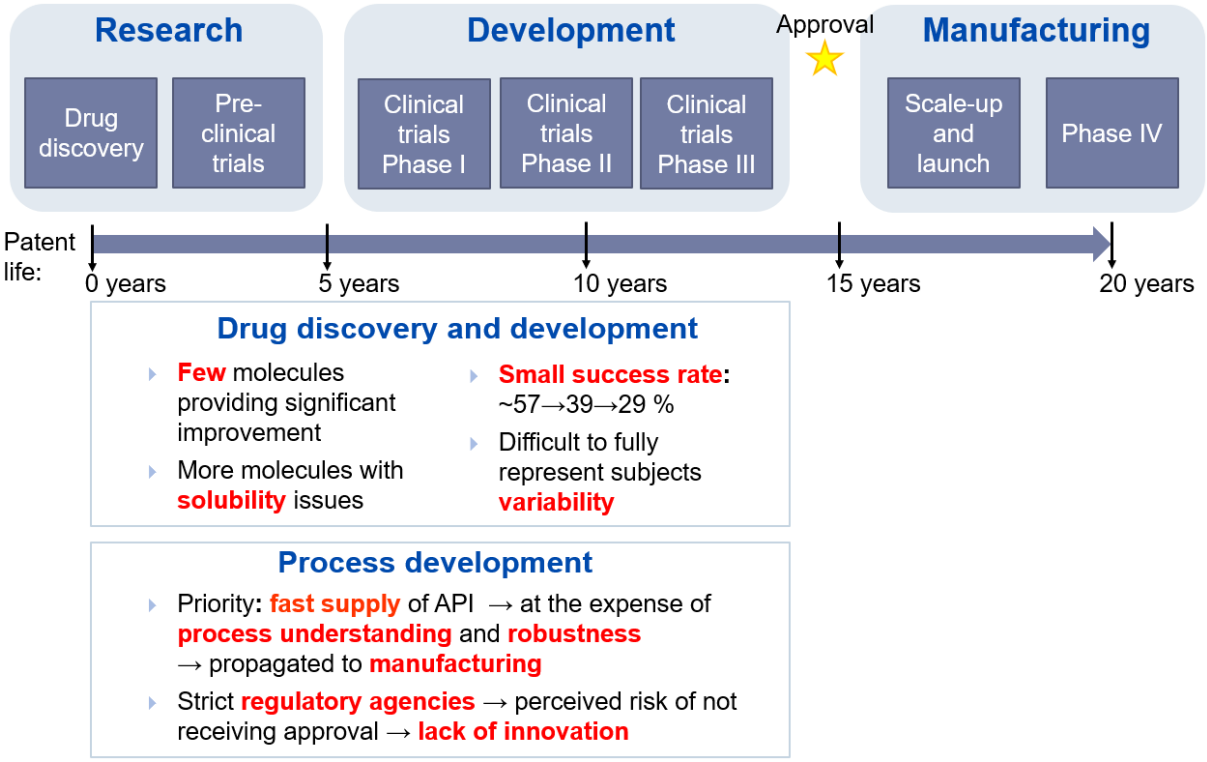



**Figure 1.2.** *Illustration of the main steps of pharmaceutical R&D and manufacturing (adapted from Destro and Barolo, 2022).*

After patent application, pre-clinical studies are carried out to obtain detailed information on dosing and toxicity levels. To this purpose, both in-vivo and in-vitro experiments are performed and they must comply with the good laboratory practices regarding personnel, equipment and operating procedures (FDA, 2018). Usually, 1 out of 5000-10000 entities that enter pre-clinical studies is able to reach the final approval (Lipsky and Sharp, 2001); the research phase involves 3-6 years of work and about 15% of the overall budget to launch a new drug (EFPIA, 2020; Destro and Barolo, 2022).

While pre-clinical trials focus mainly on drug safety, clinical trials of the development phase aim at characterizing the interaction of the drug with the human body (FDA, 2018). However, tests in humans cannot be performed without official authorisation by competent authorities, such as the investigational new drug application (IND) by the FDA or the clinical trial application (CTA) by the EMA (Destro and Barolo, 2022). After that, clinical trials must be

properly designed: first of all, research questions and objectives (typically related to drug safety and efficacy) must be defined; then, duration, selection criteria for participants, administration route and schedule and type of data analysis must be defined. Moreover, there are three clinical trials phases (as shown in Figure 1.2), which involve an increasing number of volunteers:

- clinical trial Phase I: from 20 to 100 volunteers;
- clinical trial Phase II: from 100 to 500 volunteers;
- clinical trial Phase III: from 1000 to 5000 volunteers.

In parallel with the product development, the process is developed in order to produce the Active Pharmaceutical Ingredient (API) for clinical trials and to have a manufacturing process ready when the drug passes all the tests and is approved. Overall, the development phase lasts 6-7 years and absorbs approximately 50% of the budget, whose 30% is used in phase III, to launch a new drug (Destro and Barolo, 2022).

When the drug passes all clinical trials proving its safety and efficacy, the drug developer can file an application to FDA (Food & Drug Administration) or EMA (European Medicines Agency) to market the drug. The review from regulators can take up to two years and the drug can be commercialised after the approval. During drug manufacturing, tests must be carried out periodically to ensure safety and efficacy of the drug; this is considered as the Phase IV of clinical trials (Destro and Barolo, 2022).

### *1.3.1 Main challenges of pharmaceutical R&D*

As shown in Figure 1.2, one of the major difficulties of drug research is the decreasing availability of candidate drugs to be tested. This trend started from the 80s and is related to the difficulty of proposing new treatments having evident advantages over the ones already existing, such as increased efficacy, higher potency, reduced toxicity, ease of administration or affordability (Kiriiri et al., 2020). To favour the synthesis of new molecular entities, high-technology platforms and combinatorial chemistry have been employed, but they led to highly lipophilic molecules displaying poor water solubility (Kiriiri et al., 2020). In turn, scarce solubility can degrade drug manufacturability and pharmacokinetic properties (i.e., properties related to the interaction between drugs and human body). For this reason, drug solubility in a variety of conditions of relevance for the pharmaceutical industry is studied in the works presented in this Dissertation (see Section 1.9).

As regards drug development, one of the main issues is the small success rate: 57% of the drugs tested in Phase I pass to Phase II; 39% of the drugs tested in Phase II are admitted to Phase III; 29% of the candidate drugs pass Phase III. Moreover, the recruitment of volunteers is an arduous task for researches, contributing to the increase of duration and costs of clinical trials: it is estimated that only 55% of trials is able to recruit the original specified number of participants (Allison et al., 2022). The inclusion of proper percentages of minorities is even more difficult, with African Americans and Hispanics still covering small percentages of participants despite the efforts to be more inclusive (Allison et al., 2022; Fisher and Kalbaugh, 2011). It is important to improve this aspect because individuals from different ethnic groups may react differently to the same drug. For instance, Tamargo et al. (2022) observed that after treating ethnically diverse patients with antithrombotics or lipid-lowering drugs, the differences of drug-metabolising enzymes and drug transporters attributable to patients ethnicity caused different responses to the drugs. A more detailed characterisation of physiological factors and of their impact on drug efficacy will be important also to pave the way for personalised medicine (Tamargo et al., 2022).

On the other side, the main issues of process development are related to the necessity of rapidly producing the API for product development, passing from few grams for pre-clinical trials to hundreds of kilograms for phase III of the clinical trials. Rapidity is often achieved at the expense of process understanding and robustness, which translates into lack of process robustness at the manufacturing scale (Destro and Barolo, 2022).

The need to reach the market in reduced timelines also favours the employment of proven technology that might be suboptimal in terms of costs and efficacy in the long term (Peters, 2019). Moreover, the adoption of innovative solutions is hampered by the strict regulatory environment in which pharmaceutical companies operate, due to the perceived risk of not receiving approval from regulatory agencies (Peters, 2019). More details on the regulatory framework are provided in Section 1.4.

## 1.4 Regulatory framework and Quality-by-design initiatives

Regulatory agencies have an important role in the protection of patients' health and were created as a response to some incidents, such as the fatalities caused by Elixir Sulfanilamide in the US in 1937 and the birth defects caused by Thalidomide in the 60s (Destro and Barolo, 2022; Collins, 2018). While the initial aim was mainly to control toxicologic effects of drugs, the role of regulatory agencies nowadays includes several aspects; for instance: the review of

submissions concerning the approval of new drugs; the approval of variations to products and/or processes in the drug lifecycle; the visit of manufacturing sites and the assessment of compliance with regulations (Collins, 2018). Two benchmark regulatory agencies are the Food and Drug Administration (FDA) in the US and the European Medicines Agency (EMA) in Europe, but several others exist at national and international level and they can be found in Global Regulatory Authority Websites (2023). Most of them have specific regulations, even though some basics principles are shared and a further alignment is promoted by the International Council on Harmonization (ICH; Collinds, 2018)

Besides the important role in ensuring safety and efficacy of the drugs, this strict regulatory environment discouraged innovation in the pharmaceutical processes, as highlighted also by the Wall Street Journal in 2003 (Abboud and Hensley, 2003). For this reason, FDA started a series of initiatives at early 2000s to promote innovation in the pharmaceutical industry. For instance, "*Pharmaceutical current Good Manufacturing Practices (cGMPs) for the 21st century – A risk-based approach"* was introduced in 2004 (FDA, 2004b) in order to promote the following: adoption of new technologies; implementation of modern quality management techniques; implementation of risk-based approaches. This and other documents published by FDA (FDA, 2004c; FDA, 2006) and ICH (ICH 2006, 2009a, 2009b, 2011, 2012) promoting Process Analytical Technology (PAT) and Quality-by-Design (QbD) principles contributed to modernising R&D and manufacturing industry, focusing on the adoption of science-driven methods and risk-based approaches.

### *1.4.1 Quality by Design (QbD)*

One of the main issues highlighted by cGMPs for the 21st Century was the need for a common definition of pharmaceutical quality for regulatory purposes. Nowadays, a widely recognised definition of drug quality is the absence of contamination and a consistent correspondence between the drug effects and the drug performance declared in the label (Woodcock, 2004; Sivaraman and Banga, 2015; Destro and Barolo, 2022). The main innovation given by Quality-by-Design principles relies in the way pharmaceutical quality is attained. In fact, prior to the QbD principles, a Quality-by-Testing (QbT) approach was adopted (Sivaraman and Banga, 2015; Yu, 2008): in a QbT context, the raw material is tested and if it is below the approved standards it is discarded; the approved raw material is processed through a strictly controlled manufacturing process, with the necessity of asking FDA for approval of any changes; the end-product is tested a posteriori and if the results do not meet the approved performance it must be discarded. However, the QbT approach

lacks flexibility and proper understanding of root causes for failures, with the risk of repeatedly waste time and material until the root causes are well understood.

On the other side, QbD focuses on the idea that product quality should be built in at the design stage (Sivaraman and Banga, 2015, Destro and Barolo, 2022). The main steps to implement the QbD paradigm are (Djuris and Djuric, 2017):

- step 1: the definition of the Quality target product profile (QTPP), namely the drug's safety, efficacy and performance characteristics;
- step 2: based on the QTPP, the definition of the Critical quality attributes (CQAs), namely the physical, chemical, biological, or microbiological properties that contribute to the attainment of the desired quality (ICH, 2009);
- step 3: the risk assessment in order to identify critical material attributes (CMAs) and critical process parameters (CPPs), namely the properties of the raw materials and the operating variables, respectively, having an impact on CQAs;
- step 4: the definition of the Design Space (DS), namely the multivariate space of CMAs and CPPs that ensures the attainment of the desired product quality. The DS is based on the scientific understanding of products and processes and allows for much greater flexibility with respect to QbT: once the DS is approved, pharmaceutical companies can make variations within the DS without filing for a new approval;
- step 5: the definition of a control strategy, i.e. a set of actions and controls that ensures process performance and product quality (Destro and Barolo, 2022);
- step 6: continual improvement and verification. The data collected from the process are used to optimise the current process, to validate the chemometric models of PAT, to review mathematical models. Also, the performance of the process can be improved by increasing process capability, defined as the number of standard deviations between the process mean and the nearest specification limit (Yu and Kopcha, 2017).

Both industry and regulators support the benefits of adopting QbD principles in the pharmaceutical industry. QbD is expected to give valuable contribution to the reduction of time and costs of pharmaceutical R&D, as well as to improve quality and profitability of both branded and generic drugs (Grangeia et al., 2020).

Different methodologies can be used to efficiently implement QbD in the pharmaceutical industry; Sections 1.5 and 1.6 describe two methodologies that are used in this Dissertation: mathematical modelling and design of experiments.

## 1.5 Implementation of quality by design through mathematical modelling

A mathematical model is the representation of a system or phenomenon through mathematical equations and parameters. Mathematical modelling has a crucial role in the implementation of key QbD principles: for instance, to identify CMAs and CPPs and to assess their level of criticality; to derive a functional relationship between CQAs and CMAs and CPPs; to improve product and process understanding (Djuris and Djuric, 2017).

Different categories of models are available: *first-principles* (or *mechanistic* or *white-box*) models, based on the scientific understanding of the system/phenomenon; *data-driven* (or *empirical* or *black-box*) models, based on relations (e.g., correlations) among and patterns in data; *hybrid* models, which combine the two types. Some examples of applications, together with advantages and limitations are shown in Figure 1.3 and explained in Sections 1.5.1-1.5.3

| | Mechanistic | Hybrid | Data-driven |
|---|---|---|---|
| Examples | ▸ Mass, energy, momentum balances<br>▸ Thermodynamics<br>▸ Transport phenomena<br>▸ PD, PK, **PBPK**<br>▸ **Kinetic laws and mechanisms**<br>▸ … | ▸ **Property estimation**<br>▸ (Bio) reactors<br>▸ Process simulation<br>▸ … | ▸ Statistics-based methods<br>  ▸ Statistical DoE → linear regression models<br>  ▸ Bayesian inference methods<br>▸ Machine Learning (ML)<br>  ▸ Unsupervised learning<br>  ▸ **Supervised learning**<br>  ▸ Reinforcement learning<br>▸ … |
| Pros | ▸ Science-driven system understanding<br>▸ Smaller calibration dataset<br>▸ Better extrapolation | ▸ Advantages of both | ▸ Correlation structure described<br>▸ Faster development<br>▸ Reduced computational burden |
| Cons | ▸ Difficult identification of equations and parameters<br>▸ Higher computational burden | ▸ Potential not fully exploited | ▸ Larger calibration dataset<br>▸ Poor extrapolation |

**Figure 1.3.** *Mathematical modelling: mechanistic, data-driven and hybrid approaches. In the description of each approach, three aspects are included: models examples; advantages; limitations.*

After the description of first-principles, data-driven and hybrid models, the guidelines for modellers in the pharmaceutical industry are described in Section 1.5.4.

### *1.5.1 First-principles models*

One of the main advantages of first-principles models is that they allow to deepen products and process understanding by describing the physical, chemical and biological properties and/or

phenomena of the system under study. For instance, this can be done by considering: mass, energy and momentum balances; thermodynamic or transport phenomena models; kinetic laws and reaction mechanisms. Moreover, modelling approaches widely used in the pharmaceutical industry since the early stages of drug R&D are the ones describing the interaction between the drug and the human body: pharmacodynamic (PD) models, representing the action of the drug on the human body; pharmacokinetic models (PK), representing the effect of the organism on the drug (Marino et al., 2023). In fact, PK models of different levels of complexity can be found, from simple empirical correlations to complex physiologically based pharmacokinetic (PBPK) models (Stamatopoulos, 2022). Specifically, a PBPK model is a compartmental model where compartments represent physiological entities like organs and tissues and the flows linking different compartments represent blood streams (Stamatopoulos, 2022). PBPK models allows to study all the main phenomena occurring after administering the drug, namely absorption, distribution, metabolism and excretion (ADME), and to assess the factors responsible for intra- and inter-subject variability in the PK of drugs.

The advantages of first-principles models is that they allow to improve the science-driven understanding of pharmaceutical products and processes, they require a relatively small number of calibration experiments (if parameters do not have identifiability issues, see Section 1.7) and they have a better performance in extrapolation with respect to data-driven models. On the other side, they have some limitations: the derivation of their equations and parameters is more complex and requires more time, resources and modelling efforts; their computational burden is usually higher (Destro and Barolo, 2022; Djuris and Djuric, 2017).

As an example, PBPK models are employed since pre-clinical trials and the first stages of clinical trials, where data on the specific drug of interest are not abundant; with a combination of data retrieved from in vitro experiments, in silico experiments and, eventually, clinical trials it is possible to obtain mathematical models that allow to extrapolate outside the studied population and experimental conditions (Tsamandouras et al., 2015). Moreover, at early stages of drug development is also important to develop reaction kinetic models, because they allow to understand the effect of reaction conditions (e.g., temperature, initial conditions) on the formation of both products and impurities. In turn, this information can be used to develop the reaction chemistry and determine process operating ranges (Sen et al., 2021).

Although DS can be obtained with any model type, first-principles models are preferred because they provide a more comprehensive description of the relationships between CMAs, CPPs and

CQAs; however, data-driven models are still common for systems that are not well understood (Destro and Barolo, 2022).

### 1.5.2 Data-driven models

A detailed review on data-driven applications in the (bio)pharmaceutical industry is out of the scope of this Dissertation, but can be found in the recent work of Dong et al. (2023), who considered:

- statistics-based methods, such as linear regression models built with design of experiments techniques (see Section 1.5) or Bayesian inference methods that also provide an estimation of uncertainty;
- machine learning (ML) methods, in turn divided into unsupervised, supervised (like the Partial Least Square and Gaussian Process models developed in this Dissertation in Chapters 7-8and reinforcement learning.

For instance, during drug development, Quantitative structure–activity relationship (QSAR) and Quantitative structure–property relationship (QSPR) models can be used to relate biological activity and molecule properties of the drug with the chemical structure of the molecules involved. Such relation can be built with any type of modelling approach, from linear regression models to neural networks (Borhani et al., 2019). Another method that is commonly used during drug R&D is *in vitro in vivo* (IVIV) correlation, namely a predictive model used to predict the in vivo drug response (e.g., drug concentration in plasma) based on its in vitro properties (e.g., rate or extent of drug dissolution or release; Lu et al., 2011).

The main advantages of data-driven models are: ability to represent correlation structure among data; faster development; reduced computational burden in case of simplified representations of high complex mechanistic models (Destro and Barolo, 2022). Due to these strengths, they have been applied at different stages of pharmaceutical R&D and manufacturing, such as drug discovery, reaction modelling, optimisation of reaction conditions, process design and optimisation and dynamic operation control (Dong et al., 2023).

However, open challenges must be faced in the implementation of data-driven approaches in the pharmaceutical industry. In fact, data-driven models usually require a larger calibration dataset with respect to mechanistic models, but data can be scarce or variable in terms of structure and quality. Moreover, model generalisation and interpretability is more difficult with this type of models and the performance may be poor in extrapolation. Furthermore, complex

data-driven models such as representations of complex (bio)-chemical reaction networks could be difficult to solve, thus reducing the usefulness of those approaches for monitoring and control purposes (Dong et al., 2023; Destro and Barolo, 2022).

### *1.5.3 Hybrid models*

There is not a univocal definition of *hybrid* (or *semi-empirical* or *grey-box* models; Sansana et al., 2021), even though they always refer to a combination of system understanding and empirical data. For example, Djuris and Djuric (2017) include in the definition of hybrid models the ones used to predict properties such as compression and compaction parameters, like the Kawakita and Lüdde (Kawakita and Lüdde, 1971) model relating tablet porosity to the applied pressure. In fact, the prediction of mechanical and structural properties of tablets using raw material properties and process parameters is still a challenge due a limited understanding of the powder compaction process (Wünsch et al., 2019). Therefore, semi-empirical models are available and commonly used instead of mechanistic ones. Even though those semi-empirical models do not describe the phenomenon in detail, their parameters can still have a physical interpretation; for instance, the two parameters of the Kawakita and Lüdde model are related to: (*i*) failure stress in the case of piston compression (Mani et al., 2004); (*ii*) initial porosity of the tablet.

Other definitions of *hybrid* or *grey-box* models refer to the combination of first-principles and data-driven models within the same modelling framework. As illustrated in Figure 1.4, this can be done by building different configurations (Sansana et al., 2021):

- serial configuration, preferred when the first-principles model is accurate enough. Two alternatives are available: (Figure 1.4a) a data driven model is used to predict a property/phenomenon for which a first-principles model is not available and its output is used as one of the inputs of a first-principles model; (Figure 1.4b) in case of data scarcity, both measured variables and outputs of a first-principles model are used as inputs of a data-driven model;
- parallel configuration (Figure 1.4c), where the data-driven model is trained to compensate for the mismatch between experimental data and predictions of the first-principles model, which is typically caused by the inability of the first-principles model to represent some effects, non-linearities or dynamic behaviour.

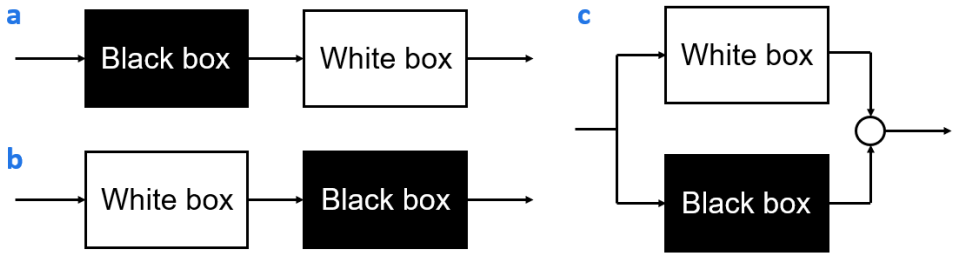



**Figure 1.4** *Hybrid models: a-b) two alternative series configurations; c) parallel configuration. White boxes refer to first-principles models, black boxes to data-driven models. Adapted from Sansana et al. (2021).*

Combinations of first-principles and data-driven models are often used to model (bio)reactors used in the pharmaceutical industry, where several unwanted impurities are produced through side reactions that are not well-understood. In this case, the known part can be described through first-principles models, like material balances, while the unknown stoichiometry and kinetics can be described though a data-driven model (Bonvin et al., 2016). For instance, one of the first applications of hybrid modelling was the one of Psichogios and Ungar (1992), who developed a hybrid approach like the one of Figure 1.5b: an artificial neural network model was used to describe the unknown kinetics and its output was sent as in input of a first-principles model representing mass and energy balances. This is configuration allowed to improve model extrapolation and interpretability and the presence of the first-principles model allowed to simplify the structure of the artificial neural networks.

Moreover, hybrid modelling approaches are advantageous for the simulation of entire processes, since rigorous models of some process units (e.g., reactors and decantors) may be absent or too cumbersome to derive (Asprion et al., 2019).

### *1.5.4 Guidelines for modellers in the pharmaceutical industry*

Regardless of the type of model selected, modelling is not just describing the system through mathematical equations, but it is a broader activity that includes several aspects, for instance the statement of modelling hypothesis and assumptions and the validation of the model itself (Sansana et al., 2021). In fact, models to be included in QbD-based submissions should be developed following a rigorous procedure, as highlighted in ICH Points to Consider (R2; see Figure 1.6):

- step 1: define the purpose of the model;

- step 2: select the modelling approach, namely first-principles, data-driven or hybrid; also the variables to be included in the model and the experimental/sampling strategy should be defined at this stage;
- step 3: understand the limitations of the modelling assumptions. In turn, this is useful to appropriately design experiments, to interpret the results and to develop proper risk-mitigation strategies;
- step 4: collect experimental data at laboratory, pilot or commercial scale. The variable ranges used in the experimentation should be representative of the real operating conditions;
- step 5: identify model structure and parameters, namely determine the most suitable model equations and estimate model parameters. This is done considering both the available knowledge on the system and the collected data;
- step 6: validate the model;
- step 7: assess the uncertainty of model predictions and elaborate a risk-mitigation strategy based on the level of impact of the model;
- step 8: document the outcomes of model development, from the modelling assumptions to the plans for further model update and improvement. Also the level of documentation depends on the impact of the model;
- step 9: implement the obtained model in the cGMP quality system.

The level of rigor depends on the  the impact of the model. For instance, a more detailed documentation must be produced for high-impact models, namely the ones whose predictions are the unique indicators of product quality  (ICH Points to Consider, R2).

The rigorous procedures to develop mathematical models in the pharmaceutical industry can benefit from the novel methods presented in this Dissertation to identify precise model parameters and to assess and reduce model prediction uncertainty, as explained in Section 1.9. Finally, one of the key aspects in model development is data collection. Due to the impact of the experimental plans on the time, labor and resources employed in pharmaceutical R&D, state-of-the-art design of experiments techniques are explained in Section 1.6.

## 1.6 Implementation of quality by design through design of experiments

For decades, the pharmaceutical industry has employed trial and error approaches to perform experiments, mainly based on the knowledge and experience of the experimenter. Usually, one-factor-at-a-time (OFAT) methods were adopted, meaning that one factor was changed within its plausible range while the others were kept fixed. However, OFAT experiments do not allow to detect interactions among factors, nor to find the optimum (Singh et al.,2005). To overcome inefficient experimental approaches, a science-based method has been promoted by QbD initiatives: statistical design of experiments (DoE, Montgomery, 2013). DoE is a systematic approach that allows to implement statistical thinking at the beginning of product development and it allows to build relationships between factors (e.g., CMAa and CPPs) and responses (e.g., CQAs). Therefore, it is useful to implement QbD principles. In fact, the use of factorial DoE has brought many advantages to the pharmaceutical R&D with respect to the commonly used OFAT approaches, allowing to: (*i*) select the factors that are more influential on the response; (*ii*) evaluate the effects of factors interaction on the response; (*iii*) reduce the experimental burden for experimental design exploration; (*iv*) identify regression models to be employed for the design, analysis and improvement of products and processes; (*v*) aid the definition of the design space (Grangeia et al., 2020; Fukuda et al., 2018; Politis et al., 2017; Singh et al., 2005a; Singh et al., 2005b). Nowadays FDA expects DoE to be part of New Drug Applications (NDAs) (Weissman and Anderson, 2015).

DoE does not require a detailed mathematical model in order to select the experimental conditions. In fact, it requires: response variables that are indicative of process conditions, experimental factors that may influence the responses and reasonable ranges for the input factors. Then, DoE makes purposeful changes of the factors within their ranges in order to verify their effects on the response (Montgomery, 2013). Three main types of DoE are available (Figure 1.5, Politis et al., 2017):

- mixture design (Figure 1.5a), where the overall amount is fixed and different proportions of the constituents are varied. Notice that the varied quantities are not independent to each other, because the increase of one variable determines the decrease of other variables in order to keep the total amount fixed;
- factorial (or process) designs (Figure 1.5b), where every factor can be changed independently from the others;

- mixture-process designs (Figure 1.5c), where mixture designs are performed at every level of the process factors.

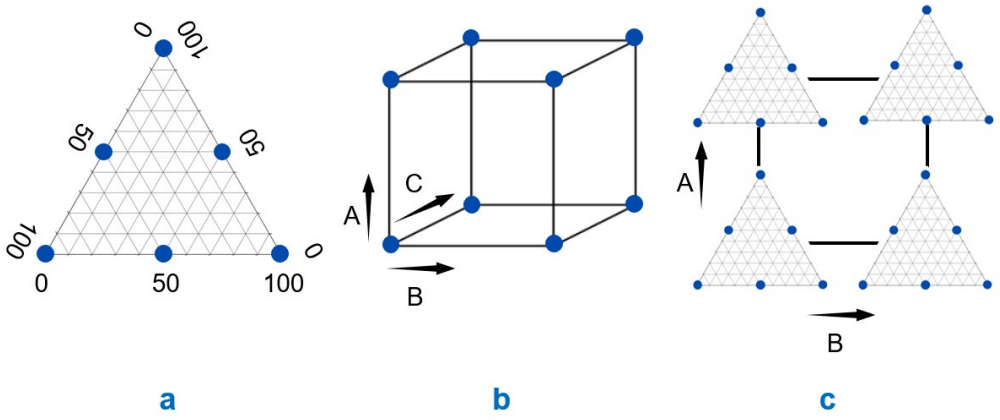



**Figure 1.5** *Categories of DoE: (a) mixture design; (b) factorial (or process) design; (c) mixture-process design. Capital letters A, B, C indicate different factors. Adapted from Politis et al. (2017).*

Several sub-types of DoE techniques exist, especially for factorial designs; a detailed description can be found in Montgomery (2013), Politis et al. (2017), Grangeia et al. (2020).

Moreover, DoE is an iterative procedure of increasing complexity. To study new systems, a screening design is usually performed with the aim of collecting data about a variety of factors that are supposed to have an influence on the response. Then, interactions are analysed in depth; finally, optimisation is carried out. Every DoE technique assumes a certain mathematical relationship between factors and response variables. The mathematical relationship is usually a linear regression model, but it can have different levels of complexity based on the scope of the experimentation. In fact, screening designs typically assume simple models, made of main effects only or main effects and interactions, and they consider two levels for every factor. Based on this data, only the few critical factors are retained and they are used for optimisation. In turn, optimisation DoE designs assume quadratic models, thus they require at least three levels for every factor in order to characterise quadratic terms (Beg, 2021; Grangeia et al., 2020).

Mixture designs are typically employed in the pharmaceutical industry for applications such as: the characterisation and optimisation of tablets formulation at a fixed tablet mass; the selection of proper diluent proportions in solid formulations; the selection of appropriate solvent–cosolvent combinations in liquid forms (Politis et al., 2017). Instead, mixture-process designs are not frequently used due to the high number of experiments designed (Politis et al., 2017). For instance, it was used by Dunn et al. (2019) in order to study intestinal drug solubility in a

variety of biorelevant conditions: they performed a mixture design with 4 biorelevant amphiphiles (bile salt, phospholipid, oleate, and monoglyceride) at 3 pH values (5, 6, and 7) and at 3 total amphiphile concentrations (11.7, 30.6, and 77.5 mM). The overall design was made of 351 experimental points, which may be excessive especially at the early stages of drug development where scarce API is available for experimentation.

In general, DoE techniques have been employed at all stages of product and process development. Several applications of DoE for product development were reviewed in the book of Sarwar Beg (2021), including DoE for the development of solid oral dosage forms, topical drug products, transdermal drug products, injectable drug products, inhalational products, ophthalmic and vesicular drug products. Moreover, applications of DoE process scale-up and process optimisation were reviewed by Weissman and Anderson (2015). Their study showed that DoE was used to optimise a variety of reaction types encountered in the pharmaceutical industry, such as hydrolysis, acylation, oxidation, halogenation, nitration, reduction, C−N bond formation, metal-catalysed cross-couplings and metal-mediated reactions, aryl alkylation, O-alkylation and miscellaneous reactions. In these applications, DoE allowed to maximise yield, to achieve the desired quality for scale-up, to minimise side-reactions occurring after scale-up, to gain scientific insights on the side-reactions leading to those impurities and to understand and counterbalance the causes of a dropped yield after scale-up.

However, DoE methods have some limitations too: 1) experimental conditions are designed all at once, without progressively updating the process knowledge as soon as data from a new experiment are available; thus, the experimenter is not taking advantage of the deeper process knowledge to improve the quality of the designed experiments; 2) conventional DoE approaches are typically used to calibrate linear regression models, which are reliable in interpolation, but not always in extrapolation as they lack mechanistic process knowledge (Garud et al., 2017; Duarte et al., 2004). For these reasons, a different method to design experiments will be used in Chapters 3-6 of this Dissertation, namely model-based design of experiments (explained in following section).

## 1.7 Model-based design of experiments (MBDoE) in the pharmaceutical industry

As explained in Sections 1.3-1.4, one of the objectives of QbD is to improve the understanding of pharmaceutical products and processes and this can be supported by mathematical modelling. For the model to be representative of the system under study, both its equations and parameters

must be identified through proper experimental data. However, not all experiments are equally informative for a given model and/or for a given modelling purpose (e.g., selection of the best model structure or estimation of precise model parameters). In fact, the statistical DoE approach presented in Section 1.5 is suitable for regression models, but it may be uninformative for the development of first-principles ones.

On the other side, model-based design of experiments (Espie and Macchietto, 1989) allows to define the most informative experimental conditions for the specific model and/or modelling purpose of interest. The generalisability of MBDoE to any mathematical model derives from the fact that experiments information content is estimated through the Fisher Information Matrix (FIM; Fisher, 1950), which in turn is calculated using model equations and current parameters values (see Chapter 2). Moreover, MBDoE can be applied to different purposes of model identification (Figure 1.6):

1. to discriminate among candidate model structures (*MBDoE for model discrimination*; Hunter and Reiner, 1965; Box and Hill, 1967; Buzzi-Ferraris and Forzatti, 1983; Buzzi-Ferraris et al., 1984; Buzzi-Ferraris et al., 1990);
2. to maximise the prediction of parameters estimates (*MBDoE for parameters precision*; Pukelsheim, 1993; Franceschini and Macchietto, 2008a; Franceschini and Macchietto, 2008b;);
3. to minimise the uncertainty of model predictions (*MBDoE for minimisation of model prediction variance*; Kiefer and Wolfowitz, 1959; Kiefer and Wolfowitz, 1960; Wong, 1995).

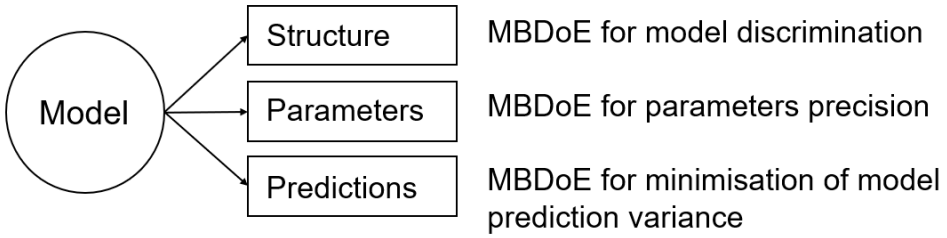



**Figure 1.6** *Elements of a mathematical model and corresponding MBDoE designs.*

MBDoE for parameters precision is the most frequent MBDoE method in pharmaceutical applications and it is involved in all MBDoE case studies of the first part of this Dissertation (Chapters 1-4). Therefore, it is explained with more details in the following Section. Then, applications of MBDoE in the (bio)pharmaceutical industry are reviewed.

### 1.7.1 MBDoE for parameters precision

Assuming that the most suitable model structure has already been defined for the system under study (e.g., through MBDoE for model discrimination), parameters identification must be carried out, namely the values of the parameters must be determined. Two types of analysis can be performed to assess parameters identifiability (McLean and McAuley, 2012):

- structural identifiability, which assesses if it is possible to obtain unique parameters estimates using perfect noise-free data. In other terms, the parameters are structurally identifiable if there are not two (or more) different sets of parameters values giving the same input-output behaviour of the model;
- practical identifiability, which assesses if it is possible to obtain precise parameters values with the available data.

In turn, the practical identifiability of a given parameter may be hampered by two main factors (McLean and McAuley, 2012): *i*) the response variable of the model has little sensitivity with respect to the parameter; in other terms, the response is scarcely influenced by variations of the parameter; *ii*) the parameter is correlated to other parameters of the model, meaning that its effects on the response can be correlated to the effects of other parameters.

Practical identifiability can be preliminarily analysed by visualising the profiles of sensitivity indices calculated for all model parameters, at all relevant experimental conditions. An illustrative example is shown in Figure 1.7: parameters $\hat{\theta}_1$ and $\hat{\theta}_2$ have a specular profile, meaning that they are correlated and it is impossible to obtaine unique parameters estimates for them; instead, $\hat{\theta}_3$ and $\hat{\theta}_4$ have profiles of different shape, therefore they can be uniquely estimated; finally, the sensitivity of the response variable with respect to $\hat{\theta}_4$ is negligible until the 10$^{th}$ experimental condition, suggesting that the experiments from no. 1 to 10 are not useful to precisely estimate this parameter.

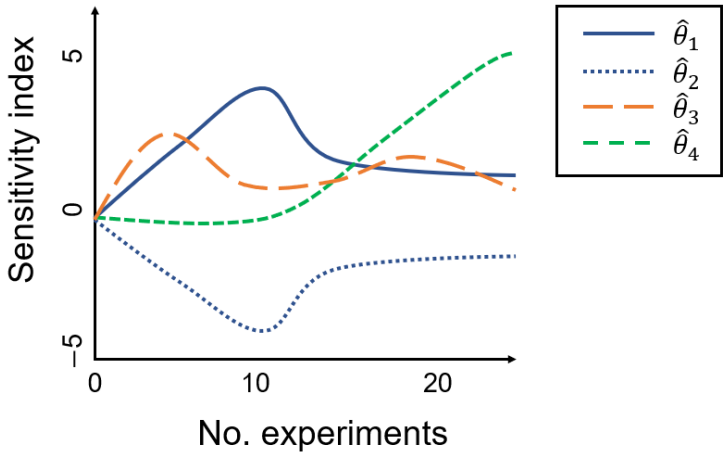



**Figure 1.7** *Profiles of sensitivity indices calculated at different experimental conditions (adapted from McLean and McAuley, 2012). Values in x- and y-axis are chosen for illustrative purposes only.*

Practical identifiability can be studied in a more rigorous way using the FIM, which is built using sensitivity indices of the response variable with respect to every model parameter and becomes singular when parameters are correlated. Therefore, a rank-deficient FIM or a high condition number (ratio between maximum and minimum eigenvalues of the FIM) suggest that some parameters are not practically identifiable (Petersen et al., 2001; Dochain and Vanrolleghem, 2001; McLean and McAuley, 2012; López C. et al., 2015).

Moreover, based on the Cramer-Rao Theorem, the inverse of the FIM is a minimum bound for the variance-covariance matrix ($\mathbf{V}_{\hat{\boldsymbol{\theta}}}$) of the parameters (Bard, 1974), meaning that a small information content of the experiments corresponds to a high uncertainty of the parameters estimates obtained with those data.

Therefore, to improve practical identifiability of the parameters (e.g., to reduce the condition number of the FIM) and, at the same time, to maximise parameters precision, highly informative data must be collected. This is the objective of MBDoE for parameters identification and it can be achieved by maximising scalar indices of the FIM or, equivalently, by minimising corresponding scalar indices of $\mathbf{V}_{\hat{\boldsymbol{\theta}}}$. To do so, the state-of-the-art optimality criteria (also called alphabetical criteria) and be considered (Figure 1.8; Kiefer, 1959; Pukelsheim, 1993; Franceschini and Macchietto, 2008a):

- D-optimality, namely the maximisation of the determinant of the FIM, or the minimization of the determinant of $\mathbf{V}_{\hat{\boldsymbol{\theta}}}$. From a geometrical point of view, it corresponds to the minimisation of the volume of the confidence region;

- A-optimality, namely the maximisation of the trace of the FIM, or the minimisation of the trace of $\mathbf{V}_{\hat{\boldsymbol{\theta}}}$. It corresponds to a minimisation of the dimension of the enclosing box around the confidence region;
- E-optimality, namely the maximisation of the smallest eigenvalue of the FIM, or the minimisation of the largest eigenvalue of $\mathbf{V}_{\hat{\boldsymbol{\theta}}}$. It corresponds to the minimisation of the major axis of the confidence region.

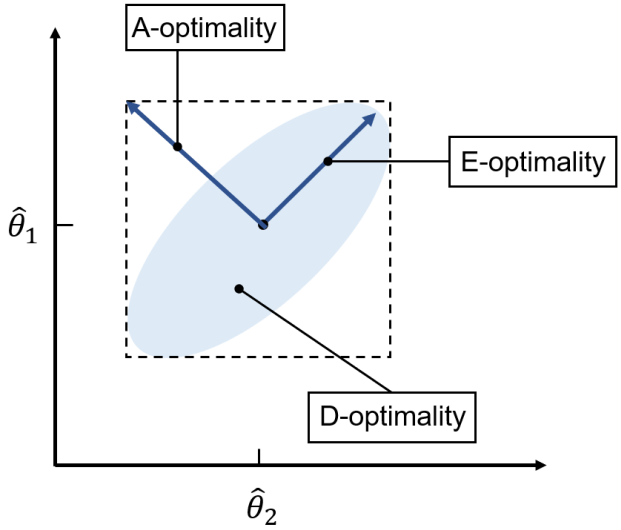



**Figure 1.8** *Graphical visualisation of the state-of-the-art optimality criteria of MBDoE for parameters identification (adapted from Franceschini and Macchietto, 2008a).*

Finally, the variance-covariance matrix of parameters $\mathbf{V}_{\hat{\boldsymbol{\theta}}}$ is also representative of parameters correlation, since it can be used to calculate correlation indices for every pair of model parameters (Franceschini and Macchietto, 2008b). From a geometrical point of view, parameters correlation can be visualised in Figure 1.9 by looking at the angle between the major axis of the confidence ellipse and the x-axis. Parameters correlation would be absent if the major axis of the ellipse was perpendicular to the x-axis.

### *1.7.2 MBDoE applications in the (bio)pharmaceutical industry*

Since MBDoE maximises the information content of experiments, the experimental burden required to achieve a specific modelling purpose is minimised. This is beneficial in the (bio)pharmaceutical industry, due to the urgent need to reduce time and costs of product and process development. Moreover, the high information content provided by a relatively small number of experiments allows to improve the experimentation outcomes at the early stages of drug development, when a limited amount of the API of interest is available.

In Table 1.1, different applications of MBDoE in the (bio)pharmaceutical industry are shown. They are divided into two main categories, namely physiological systems and process development/optimisation, with the former one showing a higher number of applications from the 80s onwards.

**Table 1.1.** *MBDoE applications in the (bio)pharmaceutical sector.*

| Category | Application | Topic | Reference |
|---|---|---|---|
| Physiological systems | Optimal blood sampling protocols | Parameters precision | Mori and DiStefano (1979) |
| | Optimal blood sampling protocols | Parameters precision | D'Argenio (1981) |
| | Optimal blood sampling protocols | Parameters precision | DiStefano (1981) |
| | Optimal blood sampling protocols | Parameters precision | DiStefano (1982) |
| | Optimal blood sampling protocols | Parameters precision | Kalicka and Bochen (2006) |
| | Dose response studies | Parameters precision | Fedorov and Leonov (2001) |
| | Dose response studies | Parameters precision | Dragalin and Fedorov (2006) |
| | Dose response studies | Minimisation of prediction variance | Dette et al., (2008) |
| | PK, PD models | Parameters precision | Nyberg et al., (2009) |
| | PK, PD models | Parameters precision | Galvanin et al. (2013) |
| | Population kinetic studies | Parameters precision | Foracchia et al. (2004) |
| | Population PK and PD models | Review | Ogungbenro (2009) |
| | Transdermal drug delivery | Parameters precision | Schittkowski (2008) |
| | Glucose-insulin system | Parameters precision | Silber (2009) |
| | Dose-response, PK, PD models | Review | Sverdlov et al. (2020) |
| Process development/ optimisation | Batch crystallisation | Parameters precision | Chung et al. (2000) |
| | Freeze-Drying Operations | Parameters precision | De-Luca et al. (2020) |
| | Freeze-Drying Operations | Parameters precision | Geremia et al. (2022) |
| | Batch reactor | Parameters precision and minimisation of prediction variance | Shahmohammadi and McAuley (2019) |
| | Batch reactor | Parameters precision | Shahmohammadi and McAuley (2020) |
| | System models | Parameters precision and prediction fidelity | Geremia et al. (2023) |
| | Bioprocess engineering | Review | Abt et al. (2018) |

## 1.6.2.1 MBDoE applied to physiological systems

The main applications of MBDoE to physiological systems concern:

- optimisation of blood sampling protocols (DiStefano, 1981; DiStefano, 1982; D'Argenio, 1981; Kalicka and Bochen, 2006);
- dose-response studies (Fedorov and Leonov, 2001; Dragalin and Fedorov, 2006; Dette et al., 2008);
- PK, PD studies (Nyberg et al., 2009, Galvanin et al., 2013) and population studies, namely kinetic studies that consider inter-subject variability (Foracchia et al., 2004; Ogungbenro, 2009);
- development of specific drugs/drug delivery systems (Schittkowski, 2008; Silber, 2009);
- review of dose-response, PK, PD studies (Sverdlov et al., 2020).

Such applications will be described with more detail in the following. **Optimisation of blood sampling protocols**. Different works published from the 80s focused mainly in the optimisation of the blood sampling schedule, while the remaining input variables (such as administration protocol for the drug/tracer) were selected by the experimenter. For instance, DiStefano (1981) considered physiological systems where only one input port is accessible (e.g., blood) and one output port is accessible for sampling (e.g., blood), from which one response variables can be measured (e.g., tracer concentration), even though the procedure could be generalised to systems with more input/output variables. An "impulse-response" experimental procedure was selected: the species of interest was introduced into the blood through an approximated impulse and the blood was sampled at discrete sampling points. While the administration protocol was selected a priori, the objective of the MBDoE was to determine the optimal sampling points in order to maximise the precision of the parameters of the differential equations representing the physiological system. The procedure was successfully applied to study thyroid hormone kinetics and it allowed to maximise parameters precision with a number of blood samples equal to the number of model parameters to be identified. Moreover, the work of D'Argenio (1981) aimed at comparing the precision of parameters estimates obtained with conventional sampling schedules and with optimal ones, assuming there is no prior knowledge on parameters values at the beginning of the experimentation. They developed a sequential MBDoE procedure involving a group of subjects, where the data retrieved from one subject were used to update model parameters and refine the optimisation of the sampling schedule for the next subject. In the two in silico examples considered, the dose regimen (e.g., amount of therapeutic, duration of infusion) was selected a priori.. The simulated results showed that the best performance was achieved with the optimised sampling schedule, besides the biological variability existing among different subjects that could potentially affect the values of the pharmacokinetic parameters.

**Dose-response studies**. Dose-response studies typically involve experiments where binary responses are analysed, such as success-failure and dead-alive responses in toxicological studies (Fedorov and Leonov, 2001). Usually, Phase I clinical trials determine the maximum tolerated dose based on toxicity only; once the dose range corresponding to acceptable toxicity levels is determined, Phase II clinical trials determine the dose within that range that ensures efficacy (Dragalin and Fedorov, 2006). In this context, Dragalin and Fedorov (2006) developed a MBDoE procedure to study both safety and efficacy responses in one experimental campaign, thanks to the use of models having two dependent binary outcomes, one for efficacy and one

for toxicity. Since the model was non-linear, the FIM (thus, the estimation of information content for every experimental condition) was dependent on the current estimates of model parameters; therefore, the authors developed an adaptive MBDoE procedure: the procedure was initialised with preliminary experiments performed in a first cohort of patients in order to have preliminary parameters estimates; then, the range of doses satisfying the constraints on efficacy and toxicity was assessed and the dose level providing the highest information content was selected for the new cohort of patients; the new data were used to update model parameters and the new optimal dose satisfying constraints on efficacy and toxicity was determined; and so on. The successful results suggested that this approach can aid the acceleration of drug development thanks to the combination of Phase I and Phase II objectives in one single experimentation.

**PK, PD, population studies**. MBDoE has been applied to PK, PD and population models to support pharmaceutical product development, especially to maximise parameters precision. For instance, Nyberg et al. (2009) observed that commonly used optimal design software allowed to optimise sampling points only, but also other input variables (e.g., infusion duration of a drug, titration schemes, etc.) are crucial to select properly. Therefore, they applied MBDoE for parameters precision with the aim of optimising the information content of both dose and sampling points. The successful results suggested that this method was suitable to support both early and late stages of drug development and could be extended to a larger set of variables to be optimised, for instance: time to change treatment in titration or disease progression studies; dose schedule in oncology trials; duration of the different steps of drug-drug interaction studies.

Furthermore, Galvanin et al. (2013) developed a method to combine structural identifiability and MBDoE for parameters precision and validated it using a model representing the relationship between microbial burden and antimicrobial agent concentrations, that is commonly studied through in vitro time-kill experiments.

Finally, the study of inter-subject variability through population model was considered by Foracchia et al. (2004), who developed a software for the optimisation of number and location of sampling points for each subject, and by Ogungbenro et al. (2009), who reviewed different MBDoE applications for the development of mixed-effect modelling techniques (that are part of population analysis).

**Development of specific drugs/drug delivery systems**. In the work of Schittkowski (2008), the transdermal diffusion of drugs was modelled through a differential equation model representing the effects of different variables (e.g., thickness of tissue, thickness of membrane, initial mass of substrate, etc.) on diffusion and metabolism. MBDoE was successfully applied

to improve practical identifiability of model parameters. Moreover, Silber at al. (2009) considered Intravenous glucose tolerance tests (IVGTT), which are useful to study glucose-insuline systems, but are complex and laborious. Therefore, they applied MBDoE to optimally modify IVGTT for type 2 diabetic patients and they were able to minimise parameters uncertainty with a reduced experimental burden.

**Review**. Finally, different applications of MBDoE techniques to aid product development from phase I to Phase III clinical trials were reviewed by Sverdlov et al. (2020), including: dose-response studies; simultaneous studies of efficacy and toxicity; comparisons between different treatment groups; population PD-PK studies.

### 1.6.2.2 MBDoE applied to (bio)pharmaceutical process development

MBDoE is a promising technique also to support process development, even though it has not been applied systematically to all steps of the development of (bio)pharmaceutical processes.

Chung et al. (2000) used MBDoE to a crystallisation unit, aiming at precisely estimating nucleation and growth parameters with the minimum number of experiments. The successful results suggested that this method can aid to accelerate the launch into the market of crystal products. De-Luca et al. (2020) implemented a MBDoE approach to calibrate a model of the primary drying phase of a freeze-drying process and were able to precisely estimate heat- and mass-transfer parameters with only one optimally designed experiment. Moreover, Geremia et al. (2023) developed an optimal protocol to calibrate primary drying models using only pressure measurements and gravimetric tests, thus eliminating the need for temperature measurements that require invasive experiments and/or sensors. Moreover, Shahmohammadi and McAuley (2019) developed two methods to handle singularity issues of the FIM and applied them to a fed-batch reactor model, namely Michaelis Menten reaction model, typically used to describe the production of therapeutic agents. The two methods allowed to obtain precise parameters estimates and reduced model prediction variance for the system under study and they are promising also for processes involving a higher number of parameters or for automated experiments in case of a non-invertible FIM. The issue of non-invertible FIM in the context of pharmaceutical applications was tackled also by Shahmohammadi and McAuley (2020), who proposed a Bayesian approach accounting for prior information and tested it with a Michaelis-Menten batch reaction system for the production of a pharmaceutical agent.

Geremia et al. (2023) considered a system model for direct compression of tablets made of sub-models representing: (*i*) tablet press unit, (*ii*) tablet disintegration test unit, and (*iii*) in-vitro

dissolution test unit. They combined MBDoE for parameters precision with multivariate statistical methods to aid the interpretation of the results. The procedure was tested in silico and the satisfactory results suggested that it can be used to aid model-based development of pharmaceutical processes.

Finally, Abt et al. (2018) reviewed MBDoE applications in order to promote its employment in the biopharmaceutical sector, where empirical strategies are predominant for process development and optimisation. In fact, statistical DoE experiments are commonly used to implement QbD principles, but biopharmaceutical applications require numerous and time-consuming DoE experiments. Therefore model-based methods were suggested as a way to minimise time and costs.

## 1.8 Industry 4.0 technologies in the pharmaceutical industry

Industry 4.0, also referred to as Fourth industrial revolution, is a new industrial stage in which emerging technologies are used to provide digital solutions (Frank et al., 2019). In this section, general concepts of Industry 4.0 technologies applicable to any type of industries are illustrated; then, the applications that are relevant for this Dissertation are explained, especially high-throughput and automated experimental platforms used in the pharmaceutical sector.

### *1.8.1 Industry 4.0 technologies*

Industry 4.0 technologies can be grouped into two main categories, as shown in Figure 1.9 (Frank et al., 2019): (*i*) front-end technologies, related to manufacturing and market needs; (*ii*) base technologies, that provide connectivity and intelligence for front-end technologies.

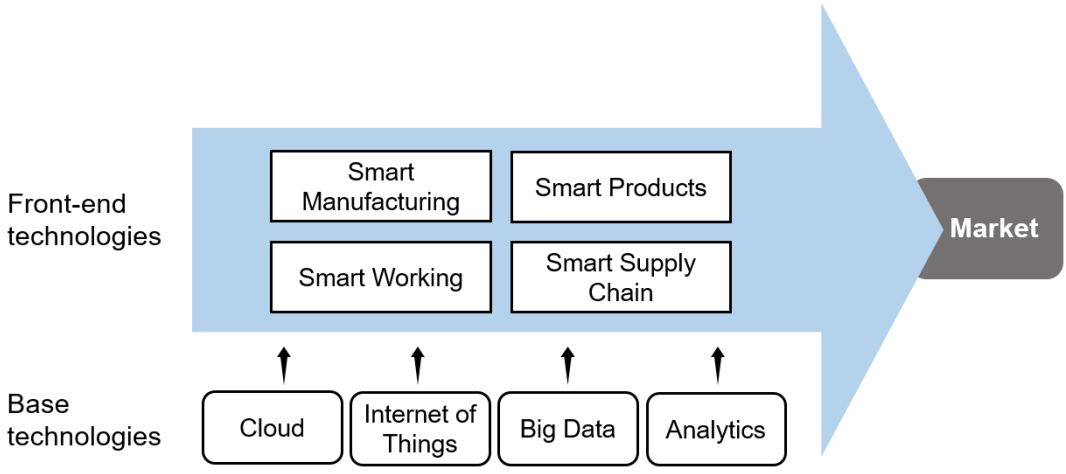



**Figure 1.9** *Industry 4.0 technologies (adapted from Frank et al., 2019).*

Front-end technologies include:

- **Smart Manufacturing**. It can be achieved by digitalising all physical objects with sensors, actuators and Programmable Logic Controllers and allowing a flow of information across the hierarchical levels of the company, in order to help decision-making and to reduce human intervention. Moreover, a key role in Smart Manufacturing is played by automation, since the use of robots allows to improve accuracy, reliability and efficiency and to let workers dedicate to more demanding tasks. Another aim is to have flexible production lines, although this is the most difficult level of Smart Manufacturing to implement: for instance, traceable materials (e.g., materials with sensors applied on them) are sent to manufacturing machines, which can read products requirements in the sensors and consequently execute the actions needed for manufacturing.
- **Smart Products**. Technologies for Smart Products include sensors applied to the product that allows to connect the product itself with other objects or systems. With artificial intelligence, products can autonomously optimise themselves (Autonomous Smart Products) and this is the most complex implementation of Smart Product principles.
- **Smart Working**. Smart working technologies aim at improving working conditions and workers' productivity. For instance, they include: mobile devices for the remote control of operation activities; virtual reality to train workers in manufacturing maintenance; augmented reality to provide workers with an interactive guidance on the steps to be made.
- **Smart Supply Chains**. Smart Supply Chains technologies mainly consists in digital platforms for the real-time exchange of information with suppliers and distribution centres, in order to reduce operational costs and delivery time.

All the abovementioned Smart technologies benefits from the base technologies, namely: Cloud, Internet of Things (IoT), Big Data and Analytics. Cloud services allow to integrate different devices and to access a shared pool of computing resources, while IoT refers to the network of physical objects ("things") embedded with sensors and software to exchange data in an internet environment through wireless communication. In this context, a huge amount of data is collected from systems and objects, generating the so-called Big Data, that must be analysed through Analytics techniques in order to extract valuable information. In turn, Big Data and Analytics are considered as the key drivers of the fourth industrial revolution.

Overall, the main objective of Industry 4.0 is to achieve autonomous decision-making, namely to implement algorithms that use the data generated by the plant in order to autonomously plan and execute the needed activities (Destro and Barolo, 2022). However, this level of autonomous

operation is still an open challenge: for instance, in the 92 companies from the machinery and equipment sector analysed by Frank et al. (2019), the adoption of flexible production lines was very limited, probably due to a higher interest of the companies on productivity rather than flexibility and to the need of changing the layout and production methods, thus requiring financial investments and interruption of operations routines.

### *1.8.2 Applications of Industry 4.0 technologies in the pharmaceutical industry*

Applied to pharmaceutical companies, Industry 4.0 technologies can bring several benefits (Bhattamisra et al., 2023; Borkar et al., 2023; Stasevych and Zvarych, 2023; Destro and Barolo, 2022; McKinsey & Company, 2021):

- reduction of costs from drug discovery, to product and process development, to manufacturing;
- decrease of the time-to-market, allowing a better exploitation of the patent;
- faster and more robust scale-up from laboratory to commercial scale;
- more consistent quality assurance;
- increase of productivity and yield;
- decrease of deviations and non-conformances;
- faster approval of new products and/or changes;
- higher employee satisfaction from user-friendly processes/tools;
- possibility for patients to receive precise and rapid diagnosis and customised treatments.

In fact, flexible production lines and the possibility to quickly integrate data coming from patients paves the way to personalised medicine, which is expected to grow in the future. Some relevant examples of personalised medicine concern the production of customised drug dosage and release profile based on patients' physiological characteristics and the production of customised medications for the treatment of neurological disorders, such as Alzheimer's disease (Stasevych and Zvarych, 2023).

Relevant examples of applications of Industry 4.0 technologies for this Dissertation are high-throughput and automated experimental platforms, since they have been used to carry out different case studies (see Chapters 4-8), therefore they are described more in detail.

### 1.8.2.1 High-throughput experimentation (HTE)

High-throughput experimentation (HTE) is defined as the workflow of performing multiple experiments in parallel. HTE is spreading in pharmaceutical companies, because they allow to fasten the experimentation without degrading the quality of the results. In fact, the review of Mennen et al. (2019) showed that high-throughput technologies have been employed to study more than two dozen bio- and chemocatalytic reaction types, ranging from common organic synthesis to novel reaction methods (such as photoredox catalysis and C−H activation) and including the analysis of both continuous and discrete variables. Moreover, it has been used to support process development, for instance by assessing salt formation and metal scavenging useful to study downstream unit operations. In the academic and industrial laboratories analysed in the review, high-throughput technologies were frequently made of the components shown in Table 1.1.

**Table 1.2.** *Equipment found in THE laboratories (adapted from Mennen et al., 2019)*

| Manual | Automatic | Analytical |
|---|---|---|
| vials-plates | liquid handler | UPLC |
| pipettors | solid handler | HPLC |
| gloveboxes | | MS |
| evaporators | | GC-MSD |
| stirrers | | |
| gas delivery | | |

As shown in Table 1.2, high-throughput technologies are often a mixture of manual and automated components, where different degrees of automation may be found in different laboratories. The parallelisation is possible thanks to the employment of plates with multiple vials of varying volumes (typically, 1, 2 or 4 mL), namely 24-, 48- and 96-vials plates. Other frequent components used are: single and multichannel pipettors; nitrogen filled gloveboxes for screen setup; centrifugal evaporators for solvents removal; stirrers for agitation and heating; gas delivery systems usually for hydrogenations and carbonylations and allowing to work at atmospheric or elevated pressures. Usually, the automated components are limited to liquid or solid handling robots that dispense material into the vials, even though higher level of automation can be found in some platforms thanks to the presence of automated heating/cooling, stirring and sampling at schedules sampling times.

Moreover, the rapidity of the screen is influenced by the technology used to analyse the samples. Some possibilities are: high-performance liquid chromatography (HPLC); mass spectrometers (MS); gas chromatography with mass detector (GC-MSD). However, the best solution in terms of rapidity is ultra- high-performance liquid chromatographic systems

(UPLC), which in some cases takes 20−30 seconds against the common 16−18 hours of analysis required for a single 96-vials plate.

Overall, high-throughput technologies can support the activities from preclinical development to manufacturing, allowing to reduce experiments duration, to give insights for reaction discovery and process development and to enhance the exploration of chemical space and process variables (Mennen et al., 2019).

### 1.8.2.2 Automated (bio)reactor platforms in flow conditions

The hardware of automated flow platforms requires mainly four elements (Figure 1.10): 1) a delivery system for the reagents; 2) a flow reactor; 3) an online or inline measurement system, e.g. gas chromatography (GC) or high-pressure liquid chromatography (HPLC); 4) a digital device to connect the various platform components, to impose changes to the manipulated variables (e.g., reagents flow rates, reactor temperature, reactor pressure) and to receive and elaborate the information from the measurement system (Barz et al., 2022). Another crucial aspect is the software used to integrate all the components and to implement user-defined algorithms. To this purpose, one of the preferred software is LabView, where customised Python or MATLAB algorithms can be implemented (Cherkasov et al., 2018).

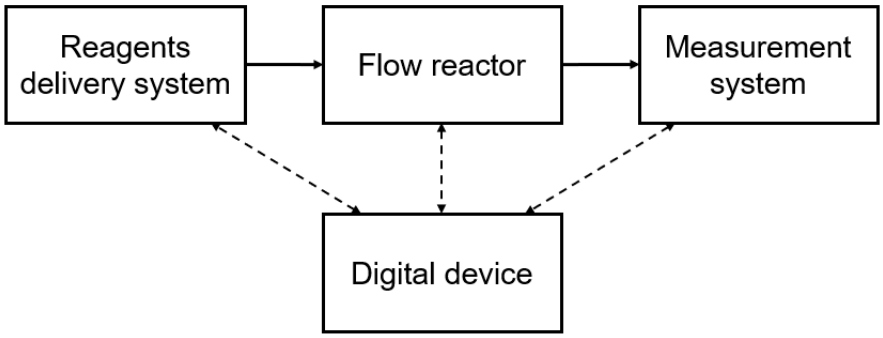



**Figure 1.10** *Industry 4.0 technologies (adapted from Frank et al., 2019).*

The realisation of flow conditions has several benefits for kinetic studies with respect to the traditional experiments in batch reactors: in fact, flow systems have a more precise and reproducible control over reagents additions, higher accuracy in temperature and reaction time and the possibility to explore conditions that are difficult to study in batch systems, such as very fast reactions or the presence of unstable intermediates (Taylor et al., 2021).

Moreover, continuous platforms are usually operated at steady-state due to the easier control and operation and to the consistency of data obtained at stationary conditions (Barz et al., 2022).

This may be a limitation due to the need of waiting until steady-state conditions are attained. However, recent works have developed experimental strategies to perform sampling at transient flow conditions, allowing to increase the number of data retrieved in one experiment and to shorten the duration of the experimental campaign (Waldron et al., 2020).

Different applications of flow reactor platforms have been reviewed by Barz et al. (2022) and they show the advantages in terms of improved system understanding, for instance allowing to derive suitable kinetic model and precise kinetic parameters. In turn, this type of information reduces the time for process development and the overall scale-up costs (Taylor et al., 2021). Moreover, besides platforms to perform chemical reactions, also bioreactor platforms have been used to optimise cultivation and production media and the cultivation conditions (Barz et al., 2022).

In the pharmaceutical industry, continuous-flow (bio)reactor platforms are present, but they are not always exploited to the fullest due to the lack of a proper experimental plan. Consequently, expensive data are produced, leading to a waste of time and resources due to uninformative experimental conditions (Barz et al., 2022). For this reason, the attention of recent research is focusing on the development of algorithms that allow to improve the selection of the experimental conditions and to fully automate this selection, thus reducing human intervention.

## 1.9 Objectives of the research

As described in the previous sections, several challenges must still be faced by the pharmaceutical industry to streamline R&D. Specifically, open challenges for modellers that are tackled in this Dissertation are related to:

- **Development of first-principles models**. First-principles models are beneficial to improve robustness, to enhance product and process understanding, to have reliable predictions in extrapolation. However, the bottleneck is the derivations of equations and parameters that accurately represent the system (Chatterjee et al., 2017). For instance, mechanistic models are useful to improve the fundamental scientific understanding of a new process and to optimise known processes (Mortier et al., 2011), but their derivation maybe hampered by the lack of fundamental equations due to the high complexity (Djurisand Djuric, 2017). Moreover, even when a model structure is available, parameters identifiability (thus, the possibility to obtain unique parameters estimates) should be assessed. For instance, this is a crucial aspect of PD and PK models where parameters must be properly estimated in order to describe the exposure-response relationship, but identifiability issues become more

difficult to be analysed in case of complex mechanistic models. Therefore, this Dissertation presents new MBDoE methods to streamline the identification of model parameters for semi-empirical and first-principles models.

- **Evaluation of model reliability**. International guidelines promoted by regulatory agencies recognise the important of the assessment of model reliability. In fact, the usefulness of the model in making conclusions and decisions depend on model reliability (Mortier et al., 2011). For instance, ICH Points to Consider (R2) state that model predictions uncertainty must be considered to set operational boundaries. However, this problem is not addressed systematically during product and process development: in fact, once the most suitable model structure is identified, experimental data are designed and collected primarily with the aim of precisely estimating model parameters, with less efforts to assess and/or improve the level of model prediction uncertainty in a rigorous way. Moreover, the minimisation of uncertainty of model parameters, for example through MBDoE for parameters precision, does not lead necessarily to a minimisation of model prediction uncertainty in the whole design space. Therefore, new methods based on MBDoE are proposed in this Dissertation aiming at assessing model prediction variance and minimising it in the whole design space.
- **Adaptation of modelling strategies to new technologies**. Industry 4.0 has brought several innovations that can be useful to the pharmaceutical industry, such as: (*i*) the possibility to increase the experimentation outcomes through high-throughput technologies; (*ii*) the possibility to execute repetitive tasks by means of automated robots, leaving more time to experimenters for more demanding tasks; (*iii*) at the highest level of automation, the possibility to set up an automated platform that autonomously initialises, adapt and stop a complete experimental campaign and that automatically calculated key indicators of system performance; (*iv*) combination of science-based understanding of intra- and inter-subject variability and of automated flexible production lines to realise personalised medicines. However, this potential has not been fully exploited yet and some of the reasons are related to the lack of suitable models to accurately represent the system under study and the lack of an adaptable and science-based strategy to design experiments through automated and/or high-throughput platforms. For these reasons, novel model-based methods are developed in order to guide the use of high-throughput and/or automated technology in a science-based way to produce highly informative experiments with a reduced experimental effort and to better represent the obtained data through mathematical modelling. Also an autonomous

MBDoE method is developed to find the most suitable trade-off between space exploration and information maximisation with an automated chemical platform, without requiring human intervention.

- **Assessm of drug solubility to support product and process development**. Solubility is one of the key properties for drugs because it has an impact on drug manufacturability and on its safety and efficacy. In fact, process units widely employed in the pharmaceutical industry, e.g. crystallisation units, require the knowledge of drug solubility in a variety of organic solvents and mixtures of organic solvents in order to be designed, operated and optimised (Ruether and Sadowski, 2009; Papadakis et al., 2016; Ye and Ouyang, 2021). Moreover, the effect of a drug on the human body depends on the fraction of drug that reaches the bloodstream, which in turn relies on the possibility for a drug to be absorbed at intestinal level. Only drug particles that are solubilised in human intestinal fluids can be absorbed, therefore intestinal solubility is a critical property that must be characterised from early stages of drug development. However, there is a need for mathematical models able to predict drug solubility in human intestinal fluids or in mixtures of organic solvents with satisfactory accuracy, therefore new data-driven approaches are proposed in this Disssertation with this purpose.

- **Reduction of time, labor and resources in R&D**. Together with the abovementioned improvements, there is an urgent need of reducing time and costs to launch a new drug into the market. This will be beneficial on several levels: for pharmaceutical industries, thanks to an increased Return On Investments of R&D; for patients, because the access to new high quality treatments will be favored, with less shortages and recalls and more investments on currently unmet patients' needs; for national health systems, for the possibility of retrieving adequate amounts of therapeutics at more affordable prices. Therefore, MBDoE methods are developed in this Dissertation, aiming at providing highly experimental data with a reduced experimental effort.

In this context, the main objective of this Dissertation is to develop model-based methods to streamline pharmaceutical R&D, while at the same time improving product and process understanding and ensuring product quality and process robustness. This is done by proposing different methods to solve the following case studies:

- **Tablets lubrication** (Case study 1). Considering a semi-empirical model describing the lubrication of tablets produced by direct compression processes (Nassar et al., 2021), a

novel MBDoE method to calibrate the model with minimum experimental burden is developed. The method must be adapted to the features of the tablet press used to perform experiments. The calibrated model must meet the industrial standards of model prediction accuracy in order to be used to design and scale-up lubrication units during process development.

- **Minimisation of model prediction variance** (Case study 2). A novel explorative MBDoE (*eMBDoE*) method is developed in order to minimise model prediction variance in the whole design space, while ensuring high information content (namely, ensuring statistically sound parameters estimates) and minimising the number of experiments required. Model prediction variance is estimated in terms of G-optimality and the most suitable trade-off between space exploration and information maximisation is determined based on a user-defined threshold on G-optimality. The proposed method, named *G-map eMBDoE* , is validated in silico with two models of increasing complexity: an algebraic model with one output and two inputs; a differential equation with two outputs sampled at different sampling points and two constant inputs.
- **Minimisation of model prediction variance with an automated chemical platform** (Case study 3). The G-map eMBDoE method is upgraded and made suitable to be applied to an automated chemical platform for total methane oxidation operating in flow conditions. An algorithm to select the best threshold on G-optimality based on preliminary experimental data is proposed (previously, that threshold was selected by the user). The upgraded G-map eMBDoE method is validated with experimental data generated by an automated platform for total methane oxidation.
- **Autonomous operation of the chemical platform to reduce model prediction variance** (Case study 4). A general framework that includes both the results of Case study 2 and 3 is developed: an autonomous algorithm that selects the G-optimality thresholds and updates its calculation as soon as new experiments are available. Such method increases the generalisability of the G-map eMBDoE and reduces the need of human intervention. Therefore, ongoing work consists in its integration into the software of the automated chemical platform in order to achieve full autonomous operation from the beginning to the end of the experimental campaign.
- **Machine Learning model to predict drug solubility in organic mixtures** (Case study 5). A machine learning supervised model (inspired by QSPR models)for the prediction of

solubility of complex drug and drug-like compounds in organic solvent mixtures is developed. The model requires a relatively limited amount of information for predictions: (*i*) temperature; (*ii*) composition of the solvents mixture before solid dissolution; (*iii*) UNIFAC (UNIQUAC Functional-group Activity Coefficients) subgroups. Differently from the state-of-the-art solubility models, able to predict drug solubility only in water or in single organic solvents (or few binary mixtures), the proposed PLS model is able to accurately predict solubility in a variety of conditions: single solvents, binary mixtures, ternary mixtures at different compositions and temperatures. The model is developed with calibration and validation experimental data of a real drug generated with a high-throughput technology. The proposed modelling approach is further validated with 9 literature of drug-like compounds retrieved in the literature.

- **Machine Learning model to predict drug solubility in intestinal fluids** (Case study 6). A machine learning model based on a Gaussian Process regression model, for the prediction of drug solubility in Simulated Intestinal Fluids is developed. To have a better representation of the complexity of the interactions in Human Intestinal Fluids, a recently published dataset by Stamatopoulos et al. (2023) on a real API is used. Instead of typical oversimplified biorelevant media made of one pH value and one bile salts type and, sometimes, lecithin to form micelles with the bile salts, the dataset used in this work considered a range of pH values, 4 different types of bile salts in different proportions, lecithin and oleic acid and cholesterol to mimic food effects. Also the experimental data of the published dataset were generated by means of a high-throughput technology. The structure of the GP model is selected in order to make it suitable for an integration in commercial software like Simcyp (Certara UK limited, Sheffield), where dynamic simulations of virtual populations can be performed, thus allowing the study on intra- and inter-subject variability. Ongoing work consists in the final implementation of the GP models proposed in this Dissertation.

Case studies 1, 5, 6 have been carried out in collaboration with a multinational pharmaceutical company, GSK (Ware, U.K.; Stevenage, U.K.), while Case studies 2,3,4 have been carried out in collaboration with University College of London (London, U.K.).

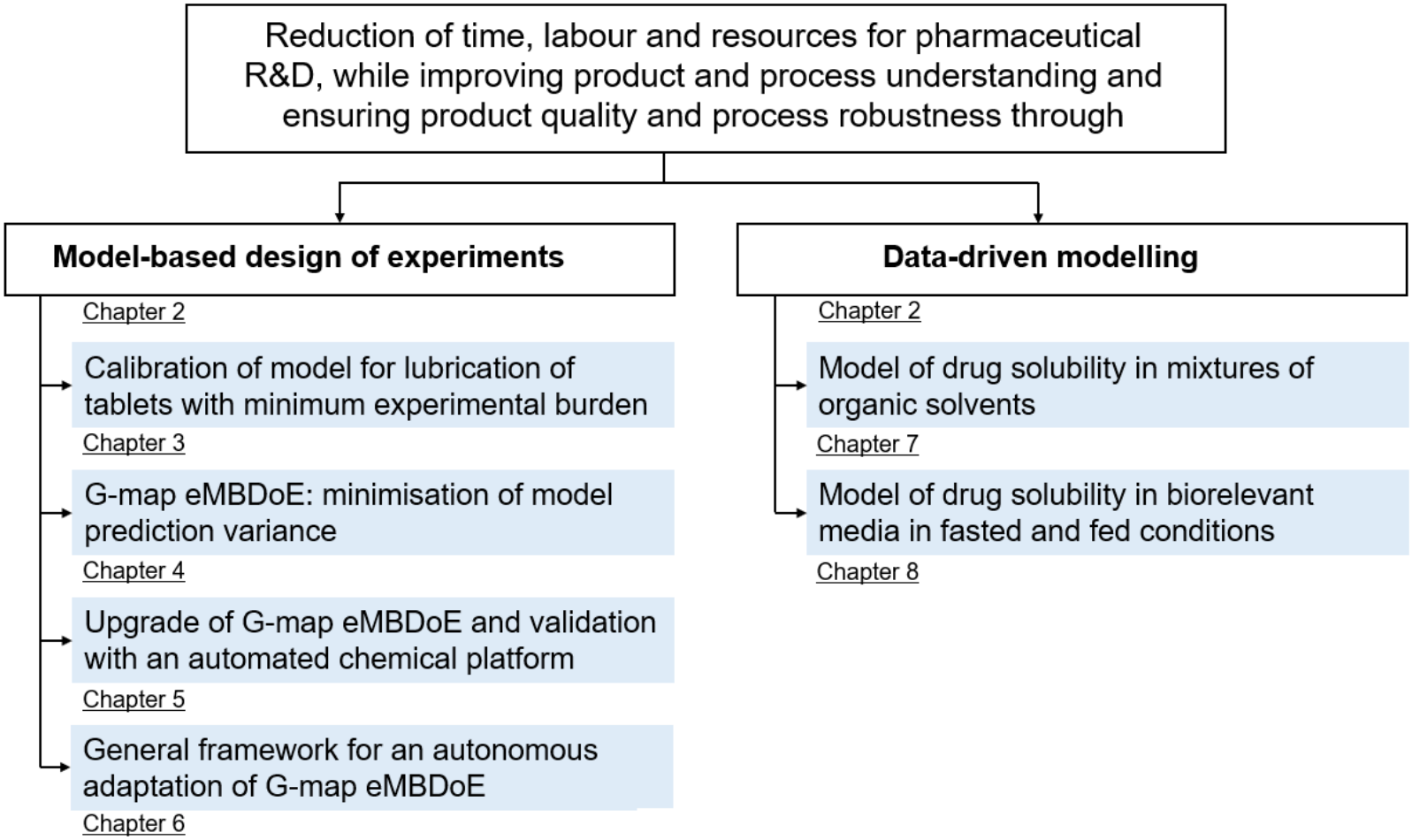
### *1.9.1 Dissertation roadmap*

As explained in the previous section, two main types of model-based techniques are employed to solve the six case studies: (*i*) model-based design of experiments for Case studies 1-4; (*ii*) data-driven modelling for Case studies 5-6.

Chapter 2 describes the two abovementioned techniques, providing more details for the methods used in this Dissertation: MBDoE for parameters precision; MBDoE to minimise model prediction variance; PLS; GP regression; statistical indices to assess parameters precision and prediction accuracy. Chapters 3-8 shows the results of Case studies 1-6, respectively. Specifically, Chapters 3-6 concern the development of new MBDoE methods that allow to support process and product development in the pharmaceutical industry: Chapter 3 proposes a novel MBDoE method that allows to design highly informative experiments to design tablets lubrication, thus to support the development of direct compression processes; Chapters 4-6 develop an exploratory MBDoE method that is validated experimentally using a chemical platform, therefore it can support product and process development through an improved understanding of main and side reactions (although the method is general, therefore it can be useful for any application involving the use of a mathematical model). Moreover, Chapters 7-8 propose modelling approaches to better characterise drug solubility for both process and product development: in Chapter 7, drug solubility is studied in mixtures of 14 organic solvents typically employed in cristallisation units, thus supporting the development and/or optimisation of crystallisation processes; in Chapter 8, solubility is assessed in vitro by using biorelevant media representing human intestinal fluids, therefore they can support product development during clinical trials.

Finally, Appendix A provides further explanation of the method used to solve Case study 1. Appendices B-D show additional results of Case study 2, while Appendices E-F shows additional results of Case study 3. Finally, Appendix G-I provide additional explanations and results for Case study 5.

**Figure 1.9** *Roadmap of this Dissertation. The mathematical methods used are explained in Chapter 2, while the six case studies in Chapters 3-8.*



# Chapter 2
# Mathematical methods

This Chapter provides an overview of the techniques employed in this Dissertation, which can be grouped into two main categories: model-based design of experiments and data-driven modelling. Model based design for parameters identification and minimisation of model prediction uncertainty is introduced. Furthermore, data-driven modelling approaches employed in this work, such as Partial Least Square and Gaussian Process regression, are then presented.

## 2.1 Model-based design of experiments (MBDoE)

Physical systems can be represented through mathematical models by identifying two main entities: model equations and parameters. In both cases, experimental data are required to develop a model that reliably represents the system under study. Therefore, different design of experiments techniques have been proposed in literature to guide experimental campaigns in a science-driven way. As explained in Chapter 1, statistical DoE has improved the use of resources in the pharmaceutical industry by replacing the ineffective trial-and-error or one-factor-at-a-time approaches, but it has still some limitations, such as the fact that multiple experiments are designed at once, without being updated when new knowledge is available, or the fact that it is suitable to identify linear regression models, but not necessarily non-linear mechanistic models.

To overcome these limitations, model-based design of experiments (MBDoE; Espie and Macchietto, 1989) techniques can be employed. MBDoE defines experiments information content based on the purpose of the experimentation, which may be: (*i*) identification of model structure, (*ii*) identification of model parameters, or (*iii*) minimisation of model prediction uncertainty. In the former case, MBDoE designs experimental conditions that are most useful to discriminate among model candidates; different MBDoE techniques for model discrimination have been proposed in literature and some examples can be found in Hunter and Reiner (1965), Buzzi-Ferraris and Forzatti (1983), Galvanin et al. (2016), Waldron et al., (2019). In this Dissertation, the second and third types of MBDoE techniques are employed. The basic assumption is that the most representative model structure for the system under study

is available and that experiments must be collected to get parameters estimates and/or to quantify model prediction uncertainty. More details are provided in the following Sub-sections.

### 2.1.1 MBDoE to maximise parameters precision

Model-based design of experiments methods for parameters precision requires the model structure in order to quantify the expected information content of an experiment. The structure of a general differential and algebraic model can be represented as:

$$\begin{aligned} &\mathbf{f}(\dot{\mathbf{x}}, \mathbf{x}, \mathbf{u}, t, \boldsymbol{\theta}) = \mathbf{0} \\ &\hat{\mathbf{y}} = \mathbf{h}(\mathbf{x}) \,, \end{aligned} \tag{2.1}$$

where $\mathbf{f}$ is a set of model equations, $\mathbf{x}$ and $\dot{\mathbf{x}}$ are $N_x$-dimensional vectors of state variables and their first derivatives respectively, $\mathbf{u}$ is a $N_u$-dimensional vector of control variables, $t$ is time, $\boldsymbol{\theta}$ is a $N_\theta$-dimensional vector of model parameters, $\mathbf{y}$ is a $N_y$-dimensional vector of response variables that are measurable.

Parameter estimates (indicated as $\widehat{\boldsymbol{\theta}}$) from experimental data are computed by minimising the difference between measured responses ($\mathbf{y}$) and predicted responses ($\hat{\mathbf{y}}$) through the negative log-likelihood function $L(\widehat{\boldsymbol{\theta}})$ (Bard, 1974):

$$\begin{aligned} L(\widehat{\boldsymbol{\theta}}) = \tfrac{N}{2}\log(2\pi) + \tfrac{N_{\mathrm{sp}}}{2}\log\left(\det\left|\boldsymbol{\Sigma}_y\right|\right) + \\ + \tfrac{1}{2}\textstyle\sum_{i=1}^{N_{\mathrm{sp}}}\left[\mathbf{y}_i - \hat{\mathbf{y}}_i(\widehat{\boldsymbol{\theta}})\right]^T \boldsymbol{\Sigma}_y{}^{-1}\left[\mathbf{y}_i - \hat{\mathbf{y}}_i(\widehat{\boldsymbol{\theta}})\right], \end{aligned} \tag{2.2}$$

where $\boldsymbol{\Sigma}_y$ is the variance-covariance matrix of measurement error, $N_{\mathrm{sp}}$ is the number of sampling points considering all the $N_e$ performed experiments, namely $N_{\mathrm{sp}} = \sum_{i=1}^{N_e} N_{\mathrm{sp}_i}$ ($N_{\mathrm{sp}_i}$ is the number of sampling points in the $i$-th experiment), $N$ is the total number of experimental measurements calculated as $N = \sum_{i=1}^{N_e} N_{\mathrm{sp}_i} N_y$. When experimental data are collected, the variance terms in $\boldsymbol{\Sigma}_y$ can be calculated as the square of the pooled standard deviations (Killeen, 2005). However, not all the experiments are equally able to provide estimates $\widehat{\boldsymbol{\theta}}$ with enough statistical precision, because this depends on their information content. The information content of the experiments is evaluated through the Fisher Information Matrix (FIM) $\mathbf{H}_{\hat{\boldsymbol{\theta}}}$ which, for dynamic systems, is expressed as (Zullo, 1991):

$$\mathbf{H}_{\hat{\boldsymbol{\theta}}}(\widehat{\boldsymbol{\theta}},\ \boldsymbol{\varphi}) = \left[\mathbf{V}_{\hat{\boldsymbol{\theta}}}^{0}\right]^{-1} + \textstyle\sum_{i=1}^{N_{\mathrm{sp}}} \left(\frac{\mathrm{d}\hat{\mathbf{y}}}{\mathrm{d}\widehat{\boldsymbol{\theta}}}\right)_i^T \boldsymbol{\Sigma}_y^{-1} \left(\frac{\mathrm{d}\hat{\mathbf{y}}}{\mathrm{d}\widehat{\boldsymbol{\theta}}}\right)_i , \tag{2.3}$$

where $\mathbf{V}_{\hat{\boldsymbol{\theta}}}^{0}$ is the $N_\theta \times N_\theta$ prior variance-covariance matrix of model parameters, while $\left(\frac{\mathrm{d}\hat{\mathbf{y}}}{\mathrm{d}\hat{\boldsymbol{\theta}}}\right)_i$ is the $N_y \times N_\theta$ matrix with first-order derivatives of model responses with respect to the parameters at time point $i$.

Based on Cramer-Rao Theorem, the inverse of the FIM represents a lower limit for the variance-covariance matrix ($\mathbf{V}_{\hat{\boldsymbol{\theta}}}$) of the parameters (Bard, 1974):

$$\mathbf{V}_{\hat{\boldsymbol{\theta}}}(\hat{\boldsymbol{\theta}}, \boldsymbol{\varphi}) \geq \left[\mathbf{H}_{\hat{\boldsymbol{\theta}}}(\hat{\boldsymbol{\theta}}, \boldsymbol{\varphi})\right]^{-1} . \tag{2.4}$$

In other terms, Eq. (2.4) provides an upper limit to parameters precision and, when the equality holds, parameters are defined efficient (Bard, 1974). Finally, the variance-covariance matrix ($\mathbf{V}_{\hat{\boldsymbol{\theta}}}$) can be approximated as the inverse of the FIM by using the first term Taylor expansion (Bard, 1974).

MBDoE for parameter identification is an optimisation problem that aims at minimizing the parametric uncertainty (represented by $\mathbf{V}_{\hat{\boldsymbol{\theta}}}$) by maximizing a scalar measure ($\psi(\mathbf{H}_{\hat{\boldsymbol{\theta}}})$) of the FIM. To this purpose, the so-called alphabetical criteria (Pukelsheim, 1993) are widely used: (*i*) maximisation of the FIM determinant or minimisation of $\mathbf{V}_{\hat{\boldsymbol{\theta}}}$ determinant (D-optimal criterion); (*ii*) maximisation of FIM trace or minimisation of $\mathbf{V}_{\hat{\boldsymbol{\theta}}}$ trace (A-optimal criterion); (*iii*) maximisation of the FIM minimum eigenvalue or minimisation of the $\mathbf{V}_{\hat{\boldsymbol{\theta}}}$ maximum eigenvalue (E-optimal criterion); (*iv*) minimisation of the ratio between maximum and minimum FIM eigenvalue (modified E-optimal). The optimisation problem is formulated as:

$$\boldsymbol{\varphi}_{\text{opt}} = \arg\min_{\boldsymbol{\varphi}} \psi(\mathbf{V}_{\hat{\boldsymbol{\theta}}}) , \tag{2.5}$$

where the outcome $\boldsymbol{\varphi}_{\text{opt}}$ is made of the values of all control variables leading to a maximum information content; in other terms, $\boldsymbol{\varphi}_{\text{opt}}$ represents the most informative experimental conditions to be measured.

For instance, $\psi(\mathbf{V}_{\hat{\boldsymbol{\theta}}}) = \psi_{\text{E}}(\mathbf{V}_{\hat{\boldsymbol{\theta}}})$ in the case of E-optimal design:

$$\psi_{\text{E}}(\mathbf{V}_{\hat{\boldsymbol{\theta}}}) = \lambda_{\max}(\mathbf{V}_{\hat{\boldsymbol{\theta}}}) , \tag{2.6}$$

where $\boldsymbol{\varphi}$ is the "design vector", which contains the set of control variables that define the experimental conditions, while $\lambda_{\max}$ refers to the maximum eigenvalue of the variance-covariance matrix $\mathbf{V}_{\hat{\boldsymbol{\theta}}}$.

### 2.1.2 MBDoE to minimise model prediction uncertainty

Once a model structure such as Eq. (2.1) is available, experiments can be collected to minimise model prediction uncertainty. Specifically, the conventional G-optimal criterion allows to minimise the maximum prediction variance through the following optimisation (Kiefer and Wolfowitz, 1959):

$$\boldsymbol{\varphi}_{\text{opt}} = \arg\min_{\boldsymbol{\varphi}} \psi(\mathbf{V}_y), \tag{2.7}$$

If all $N_y$ responses can be characterised through $N_{\text{sp}_i}$ sampling points, $\mathbf{V}_y$ is a $N_y N_{\text{sp}_i} \times N_y N_{\text{sp}_i}$ matrix containing the estimated variance of each response at each time point. Its $ji$-th element $\mathbf{V}_y(\widehat{\boldsymbol{\theta}}, \boldsymbol{\varphi})\big|_{j,i}$ is calculated as:

$$\mathbf{V}_y(\widehat{\boldsymbol{\theta}}, \boldsymbol{\varphi})\big|_{j,i} = \left(\frac{\mathrm{d}\hat{y}_j}{\mathrm{d}\widehat{\boldsymbol{\theta}}}\right)_i^T [\mathbf{H}_{\widehat{\boldsymbol{\theta}}}]^{-1} \left(\frac{\mathrm{d}\hat{y}_j}{\mathrm{d}\widehat{\boldsymbol{\theta}}}\right)_i, \quad \text{for } j = 1, \dots, N_y;\ i = 1, \dots, N_{\text{sp}_i} \tag{2.8}$$

where $\left(\frac{\mathrm{d}\hat{y}_j}{\mathrm{d}\widehat{\boldsymbol{\theta}}}\right)_i$ is the $N_\theta \times 1$ vector of first derivatives of $\hat{y}_j$ with respect to the full set of model parameters at sampling point $i$, while $\mathbf{H}_{\widehat{\boldsymbol{\theta}}}$ is the Fisher information matrix of Eq. (2.3). The scalar index $\psi(\mathbf{V}_y)$ in Eq. (2.7) is usually the largest diagonal element of $\mathbf{V}_y$.

### 2.1.3 Sequential and parallel MBDoE procedure

Estimation of model parameters and/or quantification of model prediction variance takes place after collecting the experiments designed though the optimisation in Eq. (2.5) or (2.7). There are two main procedures in which model calibration and experiments design can be performed: sequential (Espie and Macchietto, 1989; Asprey and Macchietto, 2000) and parallel (Galvanin et al., 2007) MBDoE procedures.

As shown in Figure 2.1, the sequential MBDoE procedure is made of the following steps:

- step 1: the prior knowledge on the system is defined; it is typically made of: model equations; preliminary experiments to provide initial parameters values $\widehat{\boldsymbol{\theta}}_0$; lower $\widehat{\boldsymbol{\theta}}_{\text{LB}}$ and upper $\widehat{\boldsymbol{\theta}}_{\text{UB}}$ bounds for parameters estimates; response measurement errors $\sigma_y^2$;
- step 2: MBDoE is performed, namely the optimisation of Eq. (2.5) or (2.7) is solved to obtain the optimal experimental condition $\boldsymbol{\varphi}_{\text{opt}}$;
- step 3: the experiment is performed. In physical systems, the experiment is carried out with real equipment. In case of simulated systems, the experiment can be generated in silico using model equations and the parameters assumed as "true" values $\widehat{\boldsymbol{\theta}}_{\text{true}}$; then, gaussian

noise with zero mean and $\sigma_y^2$ variance is usually added to the predicted response variable in order to mimic experimental errors;

- step 4: once data of the new experiment are available, they are used to re-estimate model parameters $\widehat{\boldsymbol{\theta}}$;
- step 5: a user-defined criterion to stop the experimental campaign is evaluated. This criterion can be based on the maximum number of experiments allowed in the experimentation or on the model performance assessed by means of the indices explained in Subsection 2.1.4. Steps 2-5 are iterated until the chosen criterion is satisfied, for instance until the experimental budget is reached.

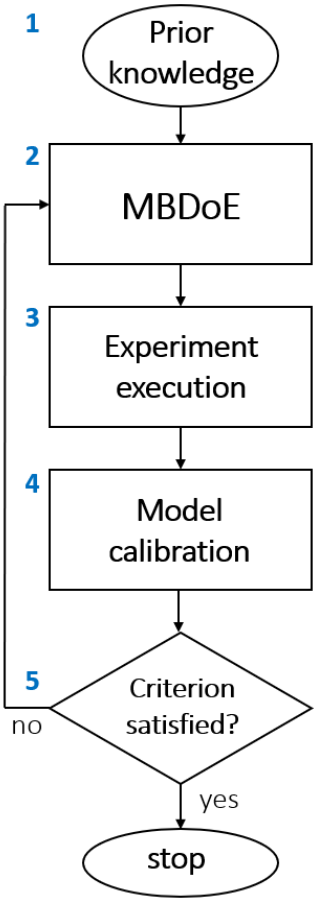



**Figure 2.1.** *Schematic of the sequential MBDoE procedure.*

Parallel MBDoE (Figure 2.2) is different from the sequential one beacuse the MBDoE optimisation in step 2 provides multiple optimal experimental conditions that can be executed in parallel in step 3. Afterwards, the model can be calibrated and analysed as explained in section 2.1.4.

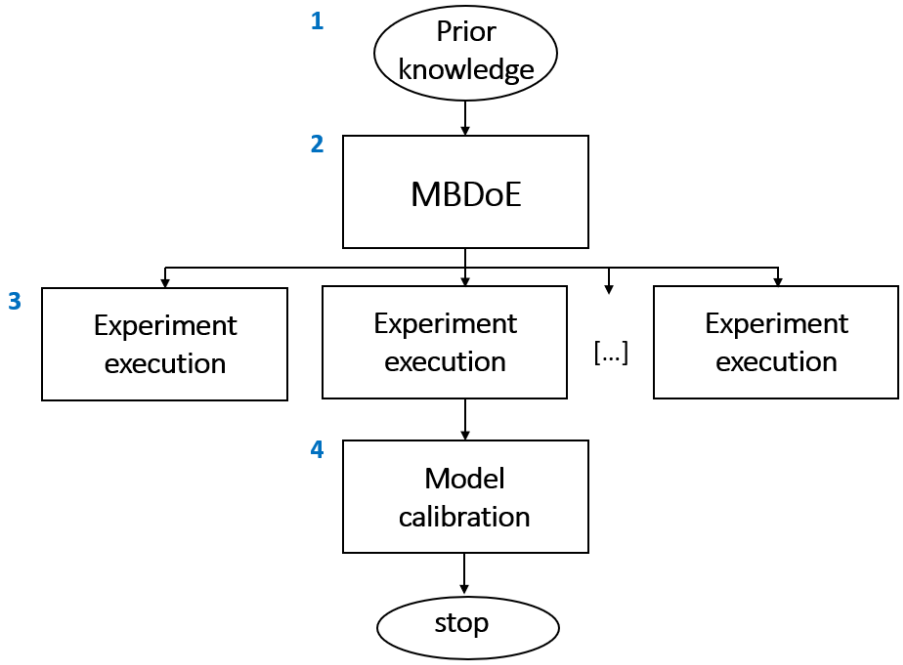



**Figure 2.2.** *Schematic of the parallel MBDoE procedure.*

The main advantage of the sequential approach over the parallel one is that it allows to improve parameters estimates as soon as new data are available. In turn, the FIM is updated since it is a function of model parameters, meaning that the estimation of information content improves as the experimentation progresses.

On the other side, parallel MBDoE is often preferred when multiple pieces of equipment are available or when it is preferable to design all experiments simultaneously for practical reasons (e.g. the need to set up or book experimental equipment in advance).

### *2.1.4 Analysis of model performance*

After model calibration, the performance of the model can be assessed in terms of parameters precision, model adequacy and prediction accuracy, as explained in the following Sub-subsections.

#### 2.1.4.1 Parameters precision

The precision of parameter estimates can be assessed through 100(1-α)% confidence intervals (CIs) and statistical tests, such as $t$-tests (Bernaerts et al., 2001; Asprey and Naka, 1999).

Confidence intervals are calculated from the parameter variance-covariance matrix $\mathbf{V}_{\hat{\boldsymbol{\theta}}}$:

$$\mathrm{CI} = \; t_{1-\alpha/2}\left(N_y - N_\theta\right)\sqrt{\mathbf{V}_{\hat{\boldsymbol{\theta}}ii}}, \quad \forall\, i = 1, \dots, N_\theta \tag{2.9}$$

where $t_{1-\alpha/2}\left(N_y - N_\theta\right)$ is the Student $t$-value with $(N_y - N_\theta)$ degrees of freedom and a significance level $\alpha$, while $\mathbf{V}_{\hat{\boldsymbol{\theta}}ii}$ is the $i$-th diagonal element of $\mathbf{V}_{\hat{\boldsymbol{\theta}}}$. In this work, 95% confidence intervals are considered.

Moreover, a $t$-test is performed for every model parameter. To have sufficient parameters precision, the following condition has to be satisfied:

$$\frac{\hat{\theta}_i}{t_{1-\alpha/2}\left(N_y - N_\theta\right)\sqrt{\mathbf{V}_{\hat{\boldsymbol{\theta}}\,ii}}} > t_{\mathrm{ref}}\,, \qquad \forall\, i = 1, \dots, N_\theta \tag{2.10}$$

where $t_{\mathrm{ref}} = t_{1-\alpha}\left(N_y - N_\theta\right)$.

## 2.1.4.2 Model adequacy

Statistical tests based on residuals are performed in order to assess the fitting quality of the model. Assuming a correct model structure, residuals should be due to errors in the observations only, and have a Gaussian (i.e., random) distribution with zero mean and standard deviation $\sigma_y$. Instead, if residuals are large and/or nonrandom they suggest that the model structure is not adequate (Bard, 1974). This is assessed through a $\chi^2$-test on the sum of the squares of residuals, in which the $\chi^2_y$ statistics is calculated as:

$$\chi^2_y = \sum_{i=1}^{N_y}\left(\frac{(y_i - \hat{y}_i)}{\sigma_y}\right)^2 \tag{2.11}$$

where $\hat{y}_i$ indicates the $i$-th predicted response; $\chi^2_y$ is compared against a reference value at a $100(1-\alpha)\%$ confidence level and $N_y - N_\theta$ degrees of freedom ( $\chi^2_{\mathrm{ref}} = \chi^2_{100(1-\alpha)\%, N_y - N_\theta}$) and the test is passed if $\chi^2_y < \chi^2_{\mathrm{ref}}$.

## 2.1.4.3 Model predictive power

The model predictive power is analysed by means of root mean squared error (RMSE), absolute error (AE) and coefficient of determination ($R^2$). The first two measure the difference between measured and predicted responses, while $R^2$ represents the quality of fitting of the calibrated model. They are calculated as:

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{N_y}(y_i - \hat{y}_i)^2}{N_y}} \tag{2.12}$$

$$\mathrm{AE} = |y_i - \hat{y}_i| \tag{2.13}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N_y}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N_y}(y_i - \bar{y})^2} \tag{2.13}$$

where $|\cdot|$ indicates the absolute value and $\bar{y}$ is the dependent variable mean.

## 2.2 Data-driven modelling

As explained in Section 1.5, a variety of data-driven approaches have been applied in the pharmaceutical industry, with many benefits at all stages of drug discovery and development. In this Dissertation, two Machine Learning (ML) techniques are used: Partial-Least Squares (PLS) and Gaussian Process (GP) regression, which are introduced in the following Subsections.

### *2.2.1 Partial Least-Squares (PLS)*

Partial-Least Squares (PLS; Wold et al., 1983; Geladi and Kowalski,1986) is a multivariate regression technique that allows to: (*i*) explain the joint correlation structure of the input matrix $\mathbf{U}$ and the response matrix $\mathbf{Y}$ and (*ii*) predict the response variable $\widehat{\mathbf{Y}}$ at new experimental conditions.

After pre-treating $\mathbf{U}$ and $\mathbf{Y}$ through auto-scaling (i.e., mean-centering and scaling to unit variance of the variables, Eriksson et al., 2006), PLS identifies a subspace of $A \ll \min(N, V)$ directions of maximum variability of the input data $\mathbf{U}$, also called latent variables (LVs), that are most correlated, and accordingly predictive, for the $N_y$ response variables $\mathbf{Y}$. Considering $\mathbf{U}$ and $\mathbf{Y}$ having dimensions $[\mathrm{N} \times V]$ and $[\mathrm{N} \times N_y]$, respectively, the model structure is given by the following equations:

$$\mathbf{U} = \mathbf{T}\mathbf{P}^{\mathrm{T}} + \mathbf{E}\,, \tag{2.15}$$

$$\mathbf{Y} = \mathbf{T}\mathbf{Q}^{\mathrm{T}} + \mathbf{f}\,, \tag{2.16}$$

$$\mathbf{T} = \frac{\mathbf{U}\mathbf{W}}{\mathbf{P}^{\mathrm{T}}\mathbf{W}}\,, \tag{2.17}$$

where $\mathbf{T}$ is the $[N \times A]$ matrix of scores, $\mathbf{P}^{\mathrm{T}}$ is the $[\mathrm{A} \times V]$ matrix of loadings of $\mathbf{U}$, $\mathbf{Q}^{\mathrm{T}}$ is the $[\mathrm{A} \times N_y]$ matrix of loadings of $\mathbf{Y}$, $\mathbf{W}$ is the $[\mathrm{V} \times A]$ matrix of weights, $\mathbf{E}$ and $\mathbf{F}$ are the $[\mathrm{N} \times V]$ and $[\mathrm{N} \times N_y]$ residuals of the $\mathbf{U}$ and $\mathbf{Y}$ spaces, respectively.

Given a new observation $\mathbf{u}_n$, the predicted response $\hat{\mathbf{y}}_n$ can be calculated as:

$$\hat{\mathbf{y}}_n = \mathbf{t}_n\mathbf{P}^{\mathrm{T}}\,, \tag{2.18}$$

where $\mathbf{t}_n$ is the $[1 \times A]$ vector of scores calculated with $\mathbf{u}_n$ as indicated in Eq. (2.17).

Moreover, a PLS model can be interpreted as a linear regression model (the derivation can be found in Wold et al., 1983):

$$\hat{y}_n = \mathbf{u}_n^{\mathrm{T}}\boldsymbol{\beta} \,. \tag{2.19}$$

. The form of Eq. (2.19) is useful also because it gives an indication of the type and relevance of the effect of every regressor through parameters $\boldsymbol{\beta}$; moreover, this is the form usually employed to define model prediction uncertainty.

More details on the analysis and interpretation of the PLS model are provided in the following Sub-Subsection.

### 2.2.1.1 Analysis of PLS model performance

To evaluate the performance of the PLS model, the following analyses are performed:

- test on sample diagnostics, to identify potential outliers and/or observations with a high impact on the model;
- calculation of model prediction uncertainty;
- study of variable importance, to identify the input variables with a considerable impact on the model;
- test on model diagnostics, to calculate the amount of data variability captured by the PLS model, both in calibration and for new unknown validation/test data.

Tests on sample diagnostics is performed considering the so-called Hotelling $T_n^2$ statistic and Squared prediction error $\mathrm{SPE}_n$ statistics calculated as:

$$T_n^2 = \mathbf{t}_n\boldsymbol{\Lambda}^{-1}\mathbf{t}_n^{\mathrm{T}} \,, \tag{2.20}$$

$$\mathrm{SPE}_n = \mathbf{e}_n\mathbf{e}_n^{\mathrm{T}}, \tag{2.21}$$

where $\mathbf{t}_n$ is the score of the $n$-th observation, $\boldsymbol{\Lambda}^{-1}$ is a matrix whose diagonal elements are the inverse of the eigenvalues $\lambda_a$, $\mathbf{e}_n$ is the residual of the $n$-th observation. Hotelling $T_n^2$ statistic represents the distance of the observation from the average conditions of the calibration dataset, while $\mathrm{SPE}_n$ statistic represents the distance of the observation from the latent space identified by the $A$ latent variables. Therefore, observations with high $T_n^2$ have a high leverage on the PLS model, while observations with high $\mathrm{SPE}_n$ have a correlation structure that differs from the one captured by the model.

Under the assumption of multinormally distributed observations, both statistics can be compared against their confidence limits $T_{\mathrm{lim}}^2$ (Eq. 16) and $\mathrm{SPE}_{\mathrm{lim}}$:

$$T^2_{\text{lim}} = \frac{(N-1)A}{N-A} F(V, N-V, \alpha) \,, \tag{2.22}$$

$$\text{SPE}_{\text{lim}} = \frac{\sigma_{\text{cal}}}{2\mu_{\text{cal}}} \chi^2_{2\mu_{\text{cal}}/\sigma_{\text{cal}},\alpha} \,, \tag{2.23}$$

In Eq. (18), $F(V, N-V, \alpha)$ is a Fisher's distribution with $V$ and $N-V$ degrees of freedom and significance level $\alpha$, namely a confidence limit of $100(1-\alpha)\%$. In Eq. (19), $\mu_{\text{cal}}$ and $\sigma_{\text{cal}}$ are the mean and the variance of the residuals of the calibration dataset, respectively, while $\chi^2_{2\mu_{\text{cal}}/\sigma_{\text{cal}},\alpha}$ is the $\chi^2$ distribution with $2\mu_{\text{cal}}/\sigma_{\text{cal}}$ degrees of freedom. In this work, 95% confidence limits are considered.

Once the prediction $\hat{y}_n$ of a given observation is calculated as in Eq. (2.19), its uncertainty can be characterised in terms of confidence interval as in Faber and Kowalski (1997) and in Facco et al. (2015). The wider the confidence interval, the larger prediction uncertainty. If the prediction error follows a t-distribution, the $100(1-\delta)\%$ confidence interval ($\text{CI}_{\text{PLS}}$) of $\hat{y}_n$ is calculated as:

$$\text{CI}_{\text{PLS}} = \hat{y}_n \pm t_{\frac{\alpha}{2},N-d} \cdot s \,, \tag{2.24}$$

where $t$ indicates a t-statistic, $\alpha$ is the significance level of $\text{CI}_{\text{PLS}}$ and $d$ is the degree of freedom of the PLS model, in this case $d = A$. Moreover, $s$ is the standard deviation calculated as:

$$\text{s} = \text{SE}_{\text{PLS}} \sqrt{1 + h_n + \frac{1}{N}} \,. \tag{2.25}$$

In Eq. (2.25), $\text{SE}_{\text{PLS}}$represents the standard error of calibration calculated for the PLS model and $h_n$ represents the leverage of the observation. They are calculated as:

$$\text{SE}_{\text{PLS}} = \sqrt{\frac{\sum_{n=1}^{N}(y_n - \hat{y}_n)^2}{N-d}} \,, \tag{2.26}$$

$$h_n = \frac{\mathbf{t}_n \boldsymbol{\Lambda}^{-1} \mathbf{t}_n^{\text{T}}}{N-1} \,, \tag{2.27}$$

where $y_n$ and $\hat{y}_n$ are measured and predicted outputs of the calibration dataset.

The PLS model is also analysed in order to identify the input variables having a higher impact on the model itself. This can be done considering the PLS regression coefficients $\boldsymbol{\beta}$ of Eq. (2.19): the sign of a given parameter indicates whether an increase of the corresponding regressor determines an increase or decrease of the response variable, while the absolute value of the parameters indicates the relevance of the impact of the regressor on the response. Moreover, the variable importance in the projection, namely the VIP index (Chong and Jun, 2005), can be calculated for the *v*-th input variable as:

$$\mathrm{VIP}_v = \sqrt{V \frac{\sum_{a=1}^{A} R_a^2 (w_{v,a})^2}{\sum_{a=1}^{A} R_a^2}}, \tag{2.28}$$

where $R_a^2$ is the amount of $y$ variance explained by the $a$-th latent variable and $w_{v,a}$ weight of the $v$-th input variable on the $a$-th LV. The higher $\mathrm{VIP}_v$, the higher the influence of the corresponding input variable on the PLS model. Usually, a threshold of 1 is employed (Chong and Jun, 2005): if $\mathrm{VIP}_v > 1$, the *v*-th variable is deemed highly influential on the model.

Finally, model diagnostics can be performed by calculating RMSE and $R^2$ as in Eq.s (2.12) and (2.14), respectively.

### *2.2.2 Gaussian Process (GP) regression*

Consider a response variable $y$ expressed by a general regression model:

$$y = f(\mathbf{u}) + \varepsilon, \tag{2.29}$$

where $\mathbf{u}$ is the vector of $V$ regressors and $f$ is the unknown function relating $y$ to $\mathbf{u}$. Assume that the difference between measured y and the calculated $f(\mathbf{u})$, namely $\varepsilon$, is only due to measurement errors following an independent, identically distributed Gaussian distribution with mean equal to 0 and variance $\sigma_n^2$, namely:

$$\varepsilon \sim \mathcal{N}(0, \sigma_n^2)\ . \tag{2.30}$$

A Gaussian Process model defines a Gaussian distribution over $f$, therefore $f$ is completely characterised by a mean function $\mu(\mathbf{u})$ and a covariance function $\kappa(\mathbf{u})$ (Rasmussen and Williams, 2006):

$$\mu(\mathbf{u}) = \mathbb{E}[f(\mathbf{u})]\ , \tag{2.31}$$

$$\kappa(\mathbf{u}, \mathbf{u}') = \mathbb{E}[(f(\mathbf{u}) - \mu(\mathbf{u}))(f(\mathbf{u}') - \mu(\mathbf{u}'))]\ , \tag{2.32}$$

where $\mathbb{E}[\cdot]$ indicates the expected value.

Therefore, the GP model can be indicated as (Rasmussen and Williams, 2006):

$$f(\mathbf{u}) \sim \mathcal{GP}\big(\mu(\mathbf{u}), \kappa(\mathbf{u}, \mathbf{u}')\big)\ . \tag{2.33}$$

Prior distributions of $\mu(\mathbf{u})$ and $\kappa(\mathbf{u}, \mathbf{u}')$ are defined before collecting experiments and they are updated with observed data. The prior distribution of $\mu(\mathbf{u})$, namely the mean of data points before observing the actual measurements, is typically assumed to be equal to zero (Rasmussen and Williams, 2006). Different covariance functions $\kappa(\cdot,\cdot)$ are available in the literature; a common choice is the squared exponential function (SE, Wang et al., 2020; Petsagkourakis and Galvanin, 2021):

$$\kappa(\mathbf{u}, \mathbf{u}') = \sigma_{\mathrm{SE}}^2 \exp\left(-\frac{\|\mathbf{u}-\mathbf{u}'\|^2}{\ell^2}\right), \tag{2.34}$$

where $\sigma_{\mathrm{SE}}^2$ and $\ell$ are two hyperparameters.

If $N$ observations are available ("training" or "calibration" data, $\mathbf{U}$) and the GP model is used to make predictions at $N_*$ new conditions ("test" or "validation" data, $\mathbf{U}_*$) and if measurements have Gaussian errors with variance $\sigma_y^2$, the joint distribution of the response variables $\mathbf{y}$ at the training data and of the function values $\mathbf{f}_*$ at the test data under the prior is expressed as:

$$\begin{bmatrix}\mathbf{y}\\ \mathbf{f}_*\end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix}K(\mathbf{U},\mathbf{U}) + \sigma_y^2\mathbf{I} & K(\mathbf{U},\mathbf{U}_*)\\ K(\mathbf{U}_*,\mathbf{U}) & K(\mathbf{U}_*,\mathbf{U}_*)\end{bmatrix}\right), \tag{2.35}$$

where the output of the function $K(\cdot,\cdot)$ is a matrix of covariances calculated with the elements of the two input matrices: for instance, $K(\mathbf{U},\mathbf{U}_*)$ provides the $[N \times N_*]$ matrix containing the covariances calculated for every pair of training ($\mathbf{U}$) and test ($\mathbf{U}_*$) point.

Then, the joint Gaussian prior distribution is conditioned on the observed data (more details can be found in Rasmussen and Williams, 2006) and the GP model is completely defined by the posterior mean $\bar{\mathbf{f}}_*$ and covariance $\mathrm{cov}(\mathbf{f}_*)$ functions:

$$\bar{\mathbf{f}}_* = K(\mathbf{U}_*,\mathbf{U})[K(\mathbf{U},\mathbf{U}) + \sigma_y^2\mathbf{I}]^{-1}\mathbf{y}\,, \tag{2.36}$$

$$\mathrm{cov}(\mathbf{f}_*) = K(\mathbf{U}_*,\mathbf{U}_*) - K(\mathbf{U}_*,\mathbf{U})[K(\mathbf{U},\mathbf{U}) + \sigma_y^2\mathbf{I}]^{-1}K(\mathbf{U},\mathbf{U}_*)\,; \tag{2.37}$$

where $\mathrm{cov}(\mathbf{f}_*)$ is the covariance of the noise-free predictions and it becomes $(\mathrm{cov}(\mathbf{f}_*) + \sigma_y^2\mathbf{I})$ if also noise is taken into account; $\sigma_y^2$ is often considered as a hyperparameter of the GP model (Petsagkourakis and Galvanin, 2021). The covariance functions allow also to estimate model prediction uncertainty through the calculation of 95% confidence intervals ($\mathrm{CI_{GP}}$):

$$\mathrm{CI_{GP}} = \bar{\mathbf{f}}_* \pm 1.96\sqrt{\mathrm{cov}(\mathbf{f}_*)}\,. \tag{2.38}$$

A common approach to estimate the hyperparameters $\boldsymbol{\theta}_{\mathbf{GP}} = \{\sigma_{\mathrm{SE}}^2, \ell, \sigma_y^2\}$ of the GP model is to maximise the log-marginal likelihood of the observations $\mathbf{y}$ (Rasmussen and Williams, 2006; Wang et al., 2020). The marginal likelihood (or evidence) $p(\mathbf{y}|\mathbf{U})$ is defined as the integral of the product between the likelihood $p(\mathbf{y}|\mathbf{f},\mathbf{U})$ and the prior $p(\mathbf{f}|\mathbf{U})$:

$$p(\mathbf{y}|\mathbf{U}) = \int p(\mathbf{y}|\mathbf{f},\mathbf{U})p(\mathbf{f}|\mathbf{U})\mathrm{d}\mathbf{f}\,; \tag{2.39}$$

therefore, it is a marginalisation over the function values $\mathbf{f}$. The derivation of the integral in Eq. (2.39) takes advantage of the fact that the prior is Gaussian and the likelihood is a factorised Gaussian; more details can be found in Rasmussen and Williams (2006). The final log-marginal likelihood to be maximised results to be:

$$\log p(\mathbf{y}|\mathbf{U}) = -\frac{1}{2}\mathbf{y}^T\left(\mathbf{K}+\sigma_y^2\mathbf{I}\right)^{-1}\mathbf{y} - \frac{1}{2}\log\left|\mathbf{K}+\sigma_y^2\mathbf{I}\right| - \frac{n}{2}\log 2\pi, \tag{2.40}$$

where $\mathbf{K}$ is used to simplify the notation of $K(\mathbf{U},\mathbf{U})$. In Eq. (2.40), three main terms are present: $-\frac{1}{2}\mathbf{y}^T\left(\mathbf{K}+\sigma_y^2\mathbf{I}\right)^{-1}\mathbf{y}$, which is the only one containing observed values of the response variable $\mathbf{y}$; $-\frac{1}{2}\log\left|\mathbf{K}+\sigma_y^2\mathbf{I}\right|$, which is a penalisation term depending only on the input variables and covariance function; $-\frac{n}{2}\log 2\pi$ is a normalisation term. Typically, the best hyperparameters $\boldsymbol{\theta}^*$ are obtained by maximising the log-likelihood with a gradient-based method (Basak et al., 2022).

Finally, the GP model does not necessarily need the assumption of zero mean; it may be useful to indicate the mean function for several reasons, for example to express prior knowledge on the system or to improve the interpretability of the model (Rasmussen and Williams, 2006). If a prior mean function $m(\mathbf{U})$ for the system of interest is known, the zero-mean GP model can be applied to the difference between observations $\mathbf{y}$ and the mean function $m(\mathbf{U})$, namely to $(\mathbf{y}-m(\mathbf{U}))$. Therefore, the predictive mean for test data $\mathbf{U}_*$ is given by the contribution of the prior mean function, namely $m(\mathbf{U}_*)$, and of the GP regression:

$$\bar{\mathbf{f}}_* = m(\mathbf{U}_*) + K(\mathbf{U}_*,\mathbf{U})[\mathbf{K}+\sigma_y^2\mathbf{I}]^{-1}(\mathbf{y}-m(\mathbf{U})), \tag{2.41}$$

where $K(\mathbf{U}_*,\mathbf{U})$ and $\mathbf{K}$ are the same as in Eq. (2.36).

# Chapter 3
# Streamlining tablet lubrication design via model-based design of experiments[1]

In this Chapter, a novel MBDoE method is proposed to calibrate a lubrication model (Nassar et al., 2021) with minimum experimental effort, while ensuring a satisfactory prediction accuracy for industrial applications. Both sequential and par[1]allel MBDoE configurations are compared. Experimental results involving two placebo blends with different lubrication sensitivity show that this methodology is able to reduce the experimental effort by 60-70% with respect to the standard industrial practice independently of the formulation considered and configuration (i.e. parallel vs. sequential) adopted.

## 3.1 Introduction

In oral solid dosage forms manufacturing, lubrication is a processing step used to enhance the ejection of pharmaceutical tablets from a tablet press by reducing the wall friction between the tablet and the die walls (Wang et al., 2010). The process consists of mixing a lubricant, usually Magnesium Stearate (MgSt), with the remaining pre-blended components of the product formulation, and it is typically performed prior to the compaction in the press.

Besides enhancing tablet ejection during compaction, the addition of the lubricant to the powder blend contributes to improving powder flowability (Podczeck and Miah, 1994), increasing

[1] Cenci, F., Bano, G., Christodoulou, C., Vueva, Y., Zomer, S., Barolo, M., Bezzo, F. and Facco, P. (2022). Streamlining tablet lubrication design via model-based design of experiments. *International Journal of Pharmaceutics*, **614**, 121435.

Cenci, F., Bano, G., Christodoulou,C., Vueva, Y., Zomer, S.Barolo, M., Bezzo, F., Facco, P., 2022. Reducing the experimental effort to design pharmaceutical tablet lubrication by model-based design of experiments. In 32 European Symposium on Computer Aided Process Engineering, Montastruc, L., Negny, S., Eds., Comput. Aided Chem. Eng., Elsevier, 51, 25-30.

Cenci, F., Bano, G., Christodoulou,C., Vueva, Y., Zomer, S. Barolo, M., Bezzo, F., Facco, P., 2022. Development of model based strategies to accelerate the experimental campaign for the production of oral solid dosage through direct compression [oral presentation]. GRICU conference, Ischia, Italy, Jul 3-6

Cenci, F., Bano, G., Christodoulou,C., Vueva, Y., Zomer, S. Barolo, M., Bezzo, F., Facco, P., 2022. Reducing the experimental effort to design pharmaceutical tablet lubrication by model-based design of experiments [oral presentation]. 32 Symposium on Computer Aided Process Engineering (ESCAPE-32), Toulouse, France, Jun 12-15

powder bulk density (Dansereau and Peck, 1987), and reducing the risk that the powder might adhere to metal surfaces during tablet compression (Sabir et al., 2001; Yamamura et al., 2009). However, common lubricants such as MgSt can have a negative impact on product intermediate and critical quality attributes such as tablet hardness (Kikuta and Kitamori, 1994; Sheskey et al., 1995), disintegration time (Kikuta and Kitamori, 1994), and dissolution (Sheskey et al., 1995). Therefore, finding the right balance between the risks and benefits of lubrication is critical to avoid manufacturability issues during process operation while guaranteeing consistent product quality. This balance can be found by changing the formulation, the process conditions, or a combination of the two.

Formulation-related factors affecting lubrication are the choice of lubricant (Wang et al., 2010) and the amount of lubricant to be added to the formulation (Uchimoto et al., 2013). Process-related factors are the type of blender, the batch size and the blending time (for a given blending speed, or vice versa). In an industrial context, the standard practice is to manipulate blending time (or speed) before altering any of the other factors. The choice of blending time can be aided by the use of mathematical models that link the blending operating conditions across different scales with the compaction performance (e.g., quantified by tablet tensile strength). A semi-empirical model that serves to this purpose and has found a broad adoption in industry is the one presented by Kushner and Moore (2010). This model can be used to identify the best operating conditions for a commercial scale blender from laboratory scale data, with the caveat that its applicability is limited to tablets manufactured at a given solid fraction (i.e., 0.85). An extension to this model that accounts for the effect of solid fraction has recently been proposed by Nassar et al. (2021). This semi-empirical model proved useful in industrial environments, but several blending runs at different lubrication extents are required to calibrate it: as many as nine blending runs may be required depending on the sensitivity to lubrication of the formulation under development.

This has two main drawbacks. First, a significant amount of active pharmaceutical ingredient (API) is required, which may not be available during early drug development, and/or may significantly impact costs and resources during late-phase development. Second, a considerable amount of time (and related labor) is needed to prepare the blends and to carry out the calibration experiments. Overall, the above drawbacks translate into experimental campaigns whose cost can easily ramp to tens of thousands of dollars.

In this study, we tackled this issue by employing a model-based design of experiments (MBDoE; Espie and Macchietto,1989; Franceschini and Macchietto, 2008) approach to

optimally design the experimental campaign. As explained in section 2.1, MBDoE can be seen as an optimisation framework where experiments are selected in order to maximise their information content for the purpose of parameters estimation. As a consequence, compared to trial-and-error or statistical DoE approaches, fewer highly informative data are typically sufficient to obtain statistically significant parameters estimates (Galvanin et al., 2009; Akkermans et al., 2018). Given the advantages in terms of time and resources savings, MBDoE has found broad applications in a variety of scientific areas, including the (bio)pharmaceutical one. Abt et al. (2018) reviewed state-of-the-art techniques to design experiments in bioprocess engineering, and highlighted the potential of MBDoE for parameter identification from product development to manufacturing. Kroll et al. (2017) provided an overview of MBDoE applications in the biopharmaceutical process life cycle, discussing both past applications and future challenges. De-Luca et al. (2020) implemented a MBDoE approach to calibrate a model of the primary drying phase of a freeze-drying process, obtaining a significant reduction of the required experimental time. MBDoE was applied to estimate the parameters of pharmacokinetics and pharmacodynamics models in drug development from preclinical tests to phases I-III clinical (Ogungbenro et al., 2009; Galvanin et al., 2013). Violet et al. (2016) applied MBDoE to discriminate among stoichio-kinetic models used to predict side reactions and mass or heat transfer properties in continuous microreactors used in the pharmaceutical industry. Bogacka et al. (2011) applied MBDoE for parameters precision to an enzyme inhibition kinetic model used to evaluate the inhibitory potential of a drug when is co-administered with other drugs. Shahmohammadi and McAuley (2019, 2020) discussed different strategies to mitigate FIM ill-conditioning issues for MBDoE applications to a nonlinear kinetic model based on a Michaelis-Menten batch reaction for the production of a pharmaceutical agent.

In this study, a novel MBDoE procedure for the estimation of the parameters of the semi-empirical model of Nassar et al. (2021) relating tablet tensile strength to lubrication and solid fraction is proposed. Results demonstrate that the optimized procedure dramatically reduces the experimental effort and related costs, adopting either a parallel or sequential MBDoE approach.

This chapter is structured as follows. Section 3.2 briefly presents the direct compression process and describes the experimental strategy to produce calibration and validation data. Section 3.3 presents the mathematical model used in this study and explains the proposed MBDoE procedure. Finally, results are presented in section 3.4 and critically discussed within section 3.5. Some final remarks conclude the work.

## 3.2 Materials and experimental methods

Direct compression is usually the preferred tablet manufacturing method when it is not necessary to improve the materials compaction and flow properties, e.g. by converting fine powders into agglomerates through wet or dry granulation (Šantl et al., 2011). It also allows removing heat and moisture effects on materials (Alpizar-Ramos and González-de la Parra, 2017; Cox Gad, 2008). Moreover, direct compression comprises a limited number of unit operations which include blending, lubrication, compression and coating. This study focuses on the lubrication step, and more specifically, on quantifying compression performance in terms of tablet tensile strength as a function of the tablet solid fraction (therefore, compression pressure) and lubrication extent (thus, powder blending time), as discussed by Nassar et al. (2021). The experimental protocols used in this study are described next.

### *3.2.1 Materials*

Experiments are performed with blends composed of the following materials:

- microcrystalline cellulose (MCC) as Avicel PH102 (FMC Corporation, Philadelphia, USA);
- anhydrous lactose as lactose Supertab 21AN (DFE pharma, Goch, Germany) ;
- mannitol Pearlitol SD200 (Roquette, Lestrem, France);;
- croscarmellose-Na as Ac-Di-Sol (FMC Corporation, Philadelphia, USA);
- magnesium stearate (MgSt) as LIGAMED MF-2-V (Peter Greven, Bad Münstereifel, Germany).

All materials are used as received by the vendors.

### *3.2.2 Blend preparation*

Two types of placebo blends with the following compositions are used:

- formulation A: 2 parts MCC: 1 part lactose, 5% Ac-Di-Sol, 1% MgSt;
- formulation B: 1 part MCC: 2 parts mannitol, 1% MgSt.

Formulation A is prepared using a 300L binin blender (Pharmatech, Coleshill, United Kingdom) at 60% fill level. All excipients are transferred into the blender and mixed at 17 rpm for 20 min.

Formulation B is prepared in a 500 ml HDPE plastic bottle at 75% fill level using a model T2F, Glen mills Turbula (Wab Group, Muttenz, Switzerland) blender. Mannitol and MCC are

screened through 1 mm mesh sieve before transferring into the bottle. Formulation B is mixed at 46 cycles/min for 20 min.

### 3.2.3 Blend lubrication

The pre-mixed blends are weighted out to the required amount corresponding to 35% head space in 500 ml HPDE plastic bottles with diameter/high ratio of 0.5. Magnesium stearate is added on the top of the blends through a 500 microns mesh sieve and blended at 46 cycles/min for appropriate lubrication time to achieve the targeted extent of lubrication. In particular, the lubrication extent ($k$, dm) is calculated considering different blender parameters (Nassar et al., 2021) as follows:

$$k = \alpha_{\text{equip}} V_b^{\frac{1}{3}} F_h \omega_{\text{blend}} t_{\text{blend}} \quad (3.1)$$

where $\alpha_{\text{equip}}$ is an equipment dependent factor; this is 1.5 for the Turbula blender which has dual axes of rotation (Nassar et al., 2021). In Eq. (3.1), $V_b$ is the blender volume [$dm^3$], $F_h$ is the fraction [%] occupied by the headspace, which was fixed to 0.35% in these experiments, $\omega_{\text{blend}}$ is the mixer rotational speed [$min^{-1}$] and $t_{\text{blend}}$ is the blending time [min].
The extent of lubrication tested for each formulation ranges between 90 and 2000 dm.

### 3.2.4 Lubricated blend compression

A Phoenix compaction simulator (Phoenix, West Midlands, USA) is used to manufacture tablets from both blends simulating a Fette 1200 press profile with turret speed of 35 rpm. An 8 mm round concave tooling is utilized to compress tablets with a targeted weigh of 200 mg.
The pre-compression force used is 0.75 MPa and the tablets are compressed in the pressure range 50 - 245 MPa. At each compression pressure three tablets are produced and the weight, thickness and hardness of the tablets are recorded. This information is used to calculate solid fraction ($sf$, -) and tensile strength ($ts$, MPa), which, for round concave tablets, are calculated as (Pitt et al., 1988; Nassar et al, 2021):

$$sf = \frac{m_T}{\rho_t\left((2\times V_c)+W\times\left(\pi\times\left(\frac{D}{2}\right)^2\right)\right)} \quad (3.2)$$

$$ts = \frac{10\,F}{\pi D^2\left(2.84\frac{t}{D}-0.126\frac{t}{W}+3.15\frac{W}{D}+0.01\right)} \quad (3.3)$$

where $m_T$ is the tablet weight [kg], $V_c$ is the cup volume [m$^3$], $W$ is the wall height of the tablet [m], $D$ is the tablet diameter [m], $\rho_t$ is the true density of the powder blend [kg/m$^3$], $F$ is breaking force [N], $t$ is the overall tablet thickness [m].

## 3.3 Mathematical modelling

The original Kushner and Moore equation (Kushner and Moore, 2010) relates tensile strength $ts$ [MPa] to lubrication extent $k$ [dm], and is valid only for solid fraction $sf = 0.85$:

$$\frac{ts_{sf=0.85}}{ts_{sf=0.85,0}} = (1-\beta) + \beta\exp(-\gamma k) \tag{3.4}$$

where $ts_{sf=0.85,0}$ [MPa] is the initial tensile strength at 0.85 solid fraction, $\gamma$ [dm$^{-1}$] is the lubrication rate constant of the blend, and $\beta$ [-] is the total fraction of tensile strength that can be lost due to lubrication.

The empirical model proposed by Nassar et al. (2021) makes the dependence of Kushner and Moore parameters on solid fraction explicit by including the following relations into (3.4):

$$ts_{sf=0.85,0} = a_1 \exp(b_1(1-sf)) \tag{3.5}$$

$$\beta = a_2(1-sf) + b_2. \tag{3.6}$$

Therefore, to estimate model parameters of the resulting extended Kushner and Moore model, i.e. $a_1$ [MPa], $b_1$ [-], $a_2$ [-], $b_2$ [-] and $\gamma$ [dm$^{-1}$], two input variables can be manipulated: lubrication extent ($k$), related to the blending time, and tablet solid fraction ($sf$), related to the compression pressure.

### *3.3.1 Model-based design of experiments*

To represent the extended Kushner and Moore model (Eq.s 3.4, 3.6), that is a nonlinear algebraic model, the general model equations of Eq. (2.1) (section 2) can be simplified to the following form:

$$y = f(\mathbf{u}, \boldsymbol{\theta}) \tag{3.7}$$

where $y$ is tensile strength (namely, $ts$), $\mathbf{u}$ is made of solid fraction and lubrication ($\mathbf{u} = [sf, k]^{\mathrm{T}}$), and $\boldsymbol{\theta}$ is $\boldsymbol{\theta} = [a_1, b_1, a_2, b_2, \gamma]^{\mathrm{T}}$**.** In addition, symbol ^ indicates estimated variables: for example, $\hat{\mathbf{y}}$ indicates tensile strength estimated by the model, and $\widehat{\boldsymbol{\theta}}$ is a set of parameters estimated either as initial guesses (based on prior process knowledge) or using calibration data. The objective is calibrating the model (i.e., identifying the model parameters) with the minimum experimental effort. MBDoE aims at maximizing the amount of information provided

by the experiments evaluated through the $N_\theta \times N_\theta$ Fisher Information Matrix (FIM), as explained in section 2.1. In the specific case of the extended Kushner and Moore model, where the response ($ts$) depends on two inputs ($sf$ and $k$) and on parameter estimates $\widehat{\boldsymbol{\theta}} = [\hat{a}_1, \hat{b}_1, \hat{a}_2, \hat{b}_2, \hat{\gamma}]^{\mathrm{T}}$, the FIM (Eq. 2.3 of section 2.1) can be rewritten in the following form (Box and Lucas, 1959):

$$\mathbf{H}_{\widehat{\boldsymbol{\theta}}}\left(\widehat{\boldsymbol{\theta}}, \boldsymbol{\varphi}\right) = \frac{1}{\sigma_y^2}\mathbf{S}(\widehat{\boldsymbol{\theta}}, \boldsymbol{\varphi})^{\mathrm{T}}\mathbf{S}(\widehat{\boldsymbol{\theta}}, \boldsymbol{\varphi}) \tag{3.8}$$

where $\boldsymbol{\varphi}$ is the design vector (namely, $\boldsymbol{\varphi} = \mathbf{u} = [sf, k]^T$), $\sigma_y^2$ is the response variance, while $\mathbf{S}(\widehat{\boldsymbol{\theta}}, \boldsymbol{\varphi})$ is the $N \times N_\theta$-dimensional matrix of first-order sensitivity indices of the response with respect to each parameter calculated at each experimental point (namely, $\boldsymbol{\varphi}$) and at a given set of parameters guesses $\widehat{\boldsymbol{\theta}}$. A given row of the sensitivity matrix is a $1 \times N_\theta$-dimensional vector ($\mathbf{s}_{sf,k}\left(\widehat{\boldsymbol{\theta}}, \boldsymbol{\varphi}\right)$) calculated at the experimental point $[sf, k]$:

$$\mathbf{s}_{sf,k}\left(\widehat{\boldsymbol{\theta}}, \boldsymbol{\varphi}\right) = \left[\frac{\partial ts}{\partial a_1}\left(\widehat{\boldsymbol{\theta}}, \boldsymbol{\varphi}\right), \frac{\partial ts}{\partial b_1}\left(\widehat{\boldsymbol{\theta}}, \boldsymbol{\varphi}\right), \frac{\partial ts}{\partial a_2}\left(\widehat{\boldsymbol{\theta}}, \boldsymbol{\varphi}\right), \frac{\partial ts}{\partial b_2}\left(\widehat{\boldsymbol{\theta}}, \boldsymbol{\varphi}\right), \frac{\partial ts}{\partial \gamma}\left(\widehat{\boldsymbol{\theta}}, \boldsymbol{\varphi}\right)\right]_{sf,k} \tag{3.9}$$

where the subscript $sf, k$ indicates that $\mathbf{s}_{sf,k}$ is calculated for one specific experiment $[sf, k]^{\mathrm{T}}$. Then, the FIM scalar index $\psi$ can be optimized as in Eq. (2.5) to determine the most informative experimental conditions. In this work, experiments are designed through the D-optimal criterion and once they are performed, new parameters estimates $\widehat{\boldsymbol{\theta}}$ are obtained through maximum likelihood estimation with constant variance (Eq. 2.2, Section 2.1.1).

### *3.3.2 Proposed MBDoE procedure*

For the specific application considered in this study, the MBDoE problem cannot be formulated in a standard way as in Eq. (2.5) due to operational constraints. This is because, when a model of the form (3.7) is employed, the optimal design vector $\boldsymbol{\varphi}_{\mathrm{opt}}$ returns one optimal value of solid fraction and one optimal value of lubrication extent. When a new experiment is designed, the optimal value of lubrication extent may be different from the previous one. However, from a practical point of view, this would require filling the blender with API and excipients; achieving the desired lubrication extent after the optimal blending time, and generating a single compression point (hence, producing only one tablet) at the desired solid fraction value using the powder blend lubricated at the desired lubrication extent. Notwithstanding the fact that obtaining a single target solid fraction value in a compaction simulator is operationally infeasible, this approach would also require the preparation of several different powder blends, i.e. one for each optimal solid fraction value. Consequently, even if the number of data points

for model calibration is minimised through MBDoE, such measurements would determine a significant number of blends to be prepared. However, one should consider that the major contribution to the experimental effort is not given by the change of solid fraction for the same powder blend (thus, the same $k$), but rather by the change of powder lubrication.

Therefore, the experimental procedure proposed in this study aims at minimizing the number of blends to be prepared in order to achieve model calibration. First, the experimental domain must be defined in terms of range and discretization for each input variable (namely, $sf$ and $k$). Then, a two-step MBDoE procedure is performed (Figure 3.1):

- step 1: calculation of a set of $N_{SF}$ optimal values of $sf$ for every admissible value of lubrication; these sets will be denoted as "profiles" in the following;
- step 2: selection of the optimal profile, i.e. the one maximizing the objective function in Eq.(2.5).

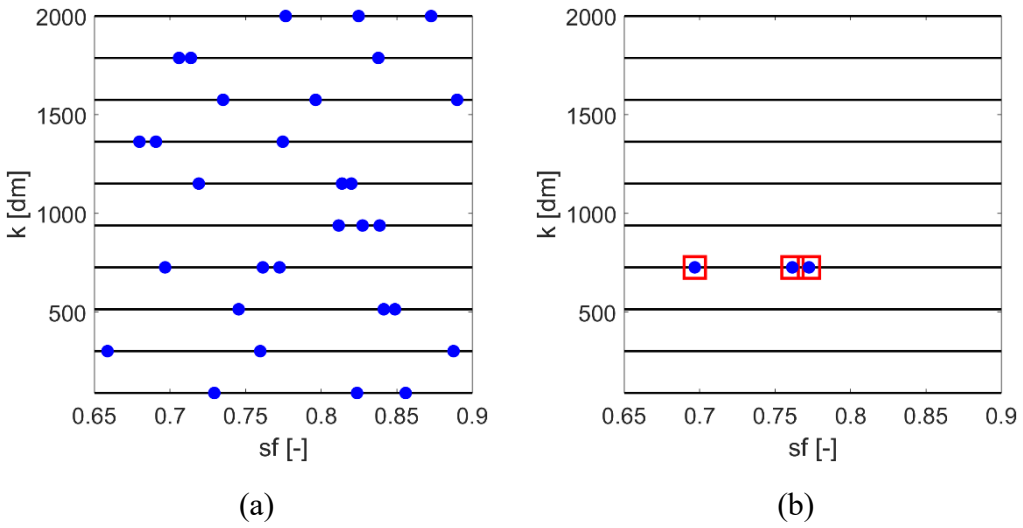



**Figure 3.1.** *Illustration of the proposed MBDoE procedure. (a) After defining the range and discretization of sf and k, a set of $N_{SF}$ optimal values of sf (dots), each of which constitutes a "profile", is calculated for every possible value of k. (b) Finally, the profile that optimizes the MBDoE objective function (dots with squares) is selected as the optimal experiment.*

The result of each iteration of the two-step MBDoE procedure is a profile (namely, $\boldsymbol{\varphi}_{\text{opt}} = [sf_{\text{opt},1}, \dots, sf_{\text{opt},N_{SF}}, k_{\text{opt}}]^{\text{T}}$) characterizing one specific blend.

To obtain a calibration dataset, steps 1 and 2 of the MBDoE procedure are iterated $N_K$ times. Specifically, two datasets of $N_K$ optimal profiles are designed: one with a parallel approach, i.e. performing model calibration at the end of the experimental campaign; one with a sequential approach, i.e. performing model calibration after measuring each single profile. The former

approach would be preferred from a practical perspective, because it allows accelerating the experimentation by preparing all optimal blends in advance and compressing them in the compaction simulator without interruptions.

From a mathematical standpoint, the two-step MBDoE procedure (both sequential and parallel) requires redefining the sensitivity matrix $\mathbf{S}(\widehat{\boldsymbol{\theta}}, \boldsymbol{\varphi})$, where rows corresponding to data points of the same profile are stacked together. Therefore, considering $N_K$ optimal profiles with $N_{SF}$ optimal solid fractions each, the sensitivity matrix becomes:

$$\mathbf{S}(\widehat{\boldsymbol{\theta}}, \boldsymbol{\varphi}) = \begin{pmatrix} \mathbf{s}_{11}(\widehat{\boldsymbol{\theta}}, \boldsymbol{\varphi}) \\ \mathbf{s}_{21}(\widehat{\boldsymbol{\theta}}, \boldsymbol{\varphi}) \\ \vdots \\ \mathbf{s}_{N_{SF}1}(\widehat{\boldsymbol{\theta}}, \boldsymbol{\varphi}) \\ \mathbf{s}_{12}(\widehat{\boldsymbol{\theta}}, \boldsymbol{\varphi}) \\ \mathbf{s}_{22}(\widehat{\boldsymbol{\theta}}, \boldsymbol{\varphi}) \\ \vdots \\ \mathbf{s}_{N_{SF,}2}(\widehat{\boldsymbol{\theta}}, \boldsymbol{\varphi}) \\ \vdots \\ \mathbf{s}_{1N_K}(\widehat{\boldsymbol{\theta}}, \boldsymbol{\varphi}) \\ \mathbf{s}_{2N_K}(\widehat{\boldsymbol{\theta}}, \boldsymbol{\varphi}) \\ \vdots \\ \mathbf{s}_{N_{SF}N_K}(\widehat{\boldsymbol{\theta}}, \boldsymbol{\varphi}) \end{pmatrix} \tag{3.10}$$

where $\mathbf{s}_{ij}(\widehat{\boldsymbol{\theta}}, \boldsymbol{\varphi})$ is the sensitivity vector calculated with the $i$-th solid fraction and the $j$-th lubrication value, $i = 1, \ldots, N_{SF}$, $j = 1, \ldots, N_K$, as in Eq. (3.9).

### 3.3.2.1 Numerical issues

When sensitivity coefficients are calculated, an appropriate scaling should always be applied in order to avoid that some variables and/or parameters dominate due to larger numerical values (Thompson et al., 2009). To this aim, the following scaling that involves all input and all output variables is adopted:

$$SF = \frac{sf}{sf_{\text{ref}}} \tag{3.11}$$

$$K = \frac{k}{k_{\text{ref}}} \tag{3.12}$$

$$TS = \frac{ts}{ts_{\text{ref}}} \tag{3.13}$$

where $sf$, $k$ and $ts$ can be measured or simulated values, while $sf_{\text{ref}}$, $k_{\text{ref}}$ and $ts_{\text{ref}}$ are reference values within the typical range of each variable. Similarly, parameters estimated with scaled variables ($SF$, $K$, $TS$) are indicated with upper-case letters ($\widehat{\boldsymbol{\Theta}} = [\hat{A}_1, \hat{B}_1, \hat{A}_2, \hat{B}_2, \hat{\Gamma}]^{\mathrm{T}}$).

Another typical numerical problem is related to the (potential) ill-conditioning of the sensitivity matrix, which translates into a solution that does not exist, or that exists, but is not unique, or that is subjected to large perturbations when experimental data have small perturbations (López C. et al., 2015). Ill-conditioned matrices (e.g., the sensitivity matrix $\mathbf{S}(\widehat{\boldsymbol{\theta}}, \boldsymbol{\varphi})$ and the Fisher information matrix $\mathbf{H}_{\widehat{\boldsymbol{\theta}}}(\widehat{\boldsymbol{\theta}}, \boldsymbol{\varphi})$) are characterised by a very high condition number $\kappa$:

$$\kappa = \frac{\lambda_{\max}}{\lambda_{\min}} \tag{3.14}$$

where $\lambda_{\max}$ and $\lambda_{\min}$ are, respectively, the maximum and minimum matrix eigenvalues; usually, the empirical upper bound $\kappa_{\max} = 1000$ is used (Grah, 2004).

In our application, we observe that $\mathbf{S}(\widehat{\boldsymbol{\theta}}, \boldsymbol{\varphi})$ is ill-conditioned at the beginning of the MBDoE procedure ($\kappa$ up to the order of $10^{17}$), when experimental data is scarce, but becomes well-conditioned (i.e., $\kappa < 1000$) when at least three profiles are used to calculate $\mathbf{S}(\widehat{\boldsymbol{\theta}}, \boldsymbol{\varphi})$. Based on this empirical evidence, we define a strategy to tackle $\mathbf{S}(\widehat{\boldsymbol{\theta}}, \boldsymbol{\varphi})$ ill-conditioning based on the temporary addition of fictitious data, named *ghost* data, to $\mathbf{S}(\widehat{\boldsymbol{\theta}}, \boldsymbol{\varphi})$. Details on this technique are provided in Appendix A.

## 3.4 Results and discussion

Experiments are performed in order to demonstrate the effectiveness of the proposed MBDoE procedure in obtaining statistically sound parameters estimates for the extended Kushner and Moore model. This is achieved through the following steps:

- optimal lubrication values are calculated by solving the MBDoE problem (Eq. 2.5);
- validation lubrication values are selected a priori in the range of interest;
- multiple tablets are produced with powder blends having the assigned (e.g., optimal and validation) lubrication extents and solid fractions; then, tablets tensile strength is experimentally measured;
- results are analyzed to assess parameters precision, model adequacy and model predictive power as described in section 2.1.4. In industry, an empirical upper threshold of 0.25 MPa as tensile strength absolute error ($ts$ AE) is often employed; the $ts$ AE acceptance criterion is satisfied when a percentage of at most 5% of data points exceeds this upper threshold (Nassar et al., 2021).

To perform experiments, two placebo formulations with different sensitivities to lubrication are employed: formulation A, characterised by high lubrication sensitivity, and formulation B, characterised by low lubrication sensitivity.

In general, the following decisions must be made to solve the optimisation in Eq. (2.5):

- initial guesses ($\widehat{\boldsymbol{\Theta}}_0$) on parameters to initialize the calculations;
- input variables domains, namely range and discretization;
- optimality criterion (i.e., the optimisation objective function);
- MBDoE approach, namely parallel or sequential MBDoE.

In this study, the mean of parameters estimates ($\widehat{\boldsymbol{\Theta}}_{\mathrm{A,hist}}$, $\widehat{\boldsymbol{\Theta}}_{\mathrm{B,hist}}$, $\widehat{\boldsymbol{\Theta}}_{\mathrm{C,hist}}$, $\widehat{\boldsymbol{\Theta}}_{\mathrm{D,hist}}$ and $\widehat{\boldsymbol{\Theta}}_{\mathrm{E,hist}}$) obtained with five historical datasets (respectively, $A_{hist}$, $B_{hist}$, $C_{hist}$, $D_{hist}$ and $E_{hist}$; Nassar et al., 2021) are used as reasonable initial parameters guesses ($\widehat{\boldsymbol{\Theta}}_0 = \widehat{\boldsymbol{\Theta}}_{\mathrm{mean}}$) .

Moreover, the input variables are optimized within the following domains:

- $sf \in [0.65, 0.90]$ (continuous interval);
- $k \in [90, 2000]$ dm; integer values only are considered.

For every optimal lubrication extent, $N_{SF} = 3$ optimal solid fractions are calculated. During experiments execution in the compaction simulator, three replicates are measured for every optimal solid fraction. Moreover, since it is possible to set up to five main compression levels in the equipment, two additional $sf$ levels are measured. They are randomly chosen in order to explore the whole range [0.65, 0.90], and they are used in the validation step only.

Moreover, in order to better explore the experimental domain, replications of similar optimisation results are avoided by imposing that any two optimal values of $sf$ and $k$ differ by at least 0.04 and 150 dm, respectively.

Overall, three sets of optimal experiments are designed and executed experimentally:

- four optimal lubrication extents calculated with a parallel two-step MBDoE procedure for formulation A;
- four optimal lubrication extents calculated with a parallel two-step MBDoE procedure for formulation B (identical to the previous optimal set since the calculation is formulation-independent);
- four optimal lubrication extents calculated with a sequential two-step MBDoE procedure for formulation A.

In addition, validation profiles are measured at low lubrication extents, where tablet tensile strength is more affected by small changes in compression pressure and/or lubrication extent.

Both optimal blends calculated with a parallel approach and validation blends are prepared in advance before being compressed in the compaction simulator.

The two-step procedure is implemented in MATLAB R2020a v. 9.8, using the 'active-set' optimisation algorithm (Gill et al., 1984). All calculations are performed with an Intel® Core™ i7-10875H CPU, @ 2.30 GHz processor with 64.0 GB RAM. The calculation of one optimal profile with five optimal solid fractions takes approximately 1 s.

### *3.4.1 Parallel MBDoE*

The optimal lubrication values calculated through the parallel two-step MBDoE procedure are shown in Table 3.1. Also validation profiles are measured at 100 dm and 400 dm, which are in a range interesting for practical applications ($k \leq 400$ dm).

**Table 3.1.** *Optimal lubrication extents designed through a parallel two-step MBDoE procedure, together with validation ones, that are experimentally measured for both formulation A and B.*

| **Lubrication extent $k$ (dm)** | **Purpose** |
|---|---|
| 90 | calibration and validation |
| 2000 | calibration and validation |
| 718 | calibration and validation |
| 1849 | calibration and validation |
| 100 | validation |
| 400 | validation |

Even though optimal lubrication values in Table 3.1 are designed in parallel, i.e. without alternating experiments design and execution, model calibration is performed in an iterative way considering an increasing number of optimal profiles from one to four.

The results in terms of parameters estimates, 95% confidence intervals and t-tests are shown in Table 3.2 for formulation A and in Table 3.3 for formulation B. For both formulations, parameter precision is not statistically sufficient when the parameters are estimated using only two optimal profiles. For formulation A, all parameters passed successfully the t-test when three optimal profiles are employed, with the only exception of parameter $A_2$ that needs at least four optimal profiles. On the other side, parameter $A_2$ is never estimated in a satisfactory way for formulation B. This may be due to the specific characteristics of formulation B: since parameter $A_2$ accounts for lubrication effects and since formulation B has low lubrication sensitivity, additional lubrication extents add little useful information for the identification of that parameter.

**Table 3.2**. *Formulation A: identification of the extended Kushner and Moore model parameters by parallel two-step MBDoE procedure at increasing number of lubrication extents. An asterisk (*) denotes a parameter not passing the statistical significance test.*

| No. of profiles | Parameters | Estimate ± 95% CI | *t*-value 95% | Reference *t*-value 95% |
|---|---|---|---|---|
| **2** | $A_1$ | 0.31±3.65 | 0.084* | 1.771 |
| | $B_1$ | −4.67±7.14 | 0.654* | |
| | $A_2$ | 0.55±3.74 | 0.146* | |
| | $B_2$ | 0.90±1.21 | 0.747* | |
| | Γ | −3.50±148.52 | 0.024* | |
| **3** | $A_1$ | 0.27±0.04 | 7.308 | 1.717 |
| | $B_1$ | −4.74±0.32 | 14.631 | |
| | $A_2$ | 0.47±0.35 | 1.370* | |
| | $B_2$ | 0.86±0.14 | 6.276 | |
| | Γ | −2.09±0.34 | 6.078 | |
| **4** | $A_1$ | 0.27±0.04 | 7.547 | 1.696 |
| | $B_1$ | −4.74±0.31 | 15.187 | |
| | $A_2$ | 0.45±0.25 | 1.825 | |
| | $B_2$ | 0.84±0.09 | 8.757 | |
| | Γ | −2.14±0.33 | 4.491 | |

**Table 3.3.** *Formulation B: identification of the extended Kushner and Moore model parameters by parallel two-step MBDoE procedure at increasing number of lubrication extents. An asterisk (*) denotes a parameter not passing the statistical significance test.*

| No. of profiles | Parameter [-] | Estimate ± 95% CI | t-value 95% | Reference t-value 95% |
|---|---|---|---|---|
| **2** | $A_1$ | 0.12±7.05 | 0.017* | 1.771 |
| | $B_1$ | −5.95±66.50 | 0.090* | |
| | $A_2$ | 0.74±4.71 | 0.157* | |
| | $B_2$ | 0.66±22.65 | 0.029* | |
| | Γ | −3.48±1016.96 | 0.003* | |
| **3** | $A_1$ | 0.112±0.04 | 2.898 | 1.717 |
| | $B_1$ | −6.01±0.97 | 6.220 | |
| | $A_2$ | 0.56±0.96 | 0.582* | |
| | $B_2$ | 0.59±0.34 | 1.712* | |
| | Γ | −2.83±1.63 | 1.735 | |
| **4** | $A_1$ | 0.11±0.03 | 3.220 | 1.696 |
| | $B_1$ | −6.14±0.89 | 6.898 | |
| | $A_2$ | 0.25±0.74 | 0.339* | |
| | $B_2$ | 0.49±0.27 | 1.851 | |
| | Γ | −2.29±0.96 | 2.385 | |

Validation results in terms of model predictive power (RMSE, $R^2$, $ts$ AE, $ts$ parity plots) and model adequacy ($\chi^2$ test) are shown in Table 3.4 and in Figure 3.2. With concern to formulation A, two optimal experiments are not enough for model calibration: RMSE is high, $R^2$ is not sufficiently close to 1 and the $ts$ AE acceptance criterion is not satisfied (almost 30% of data points have a $ts$ absolute error greater than 0.25 MPa). In addition, the model is inadequate because the $\chi^2$ test is not passed. However, the addition of the third optimal profile leads to dramatic improvement (Table 3.4): both the $ts$ AE acceptance criterion and the $\chi^2$ test are

satisfied. Negligible improvements are obtained by adding the fourth optimal profile. On the other side, formulation B exhibits a rather different behavior: model adequacy is statistically satisfactory (i.e., the $\chi^2$ test is passed) even when only two optimal profiles are employed (Table 3.4), while adding the third and fourth profiles does not improve significantly the results. Also RMSE and $R^2$ reveal a good model predictive power when two optimal profiles are employed and they do not differ significantly by adding more profiles. This is consistent with what observed during calibration: additional profiles bring in information on the effects of lubrication, but since formulation B is less influenced by lubrication with respect to formulation A, the information content is mostly related to the solid fraction. Thus, even if parameters are not estimated satisfactorily, the model capability of representing solid fraction effects is immediately attained. New lubrication profiles can improve parameter estimation, but there is no dramatic improvement in the model capability of predicting tensile strength.

Similarly, the parity plot confirm that a satisfactory model predictive power is achieved with three optimal profiles for formulation A (Figures 3.2a-3.2c) and with two optimal profiles with formulation B (Figures 3.2d-3.2f).

**Table 3.4.** *Formulations A and B: validation of the extended Kushner and Moore model calibrated with optimal data obtained through parallel two-step MBDoE procedure at increasing number of lubrication levels. "% exceeding" indicates the percent of data points for which the absolute error on tensile strength is greater than 0.25 MPa. An asterisk (*) denotes a statistical significance test that is not passed.*

| Formulation | No. of profiles used in calibration | RMSE | $R^2$ | % exceeding | $\chi^2_y$ | $\chi^2_{\text{ref}}$ |
|---|---|---|---|---|---|---|
| **A** | 2 | 0.230 | 0.97 | 29.76 | 221.692* | 100.75 |
| | 3 | 0.101 | 0.99 | 0 | 42.733 | |
| | 4 | 0.099 | 0.99 | 0 | 40.986 | |
| **B** | 2 | 0.113 | 0.98 | 5 | 38.256 | 73.312 |
| | 3 | 0.104 | 0.99 | 5 | 32.377 | |
| | 4 | 0.098 | 0.99 | 3 | 29.013 | |

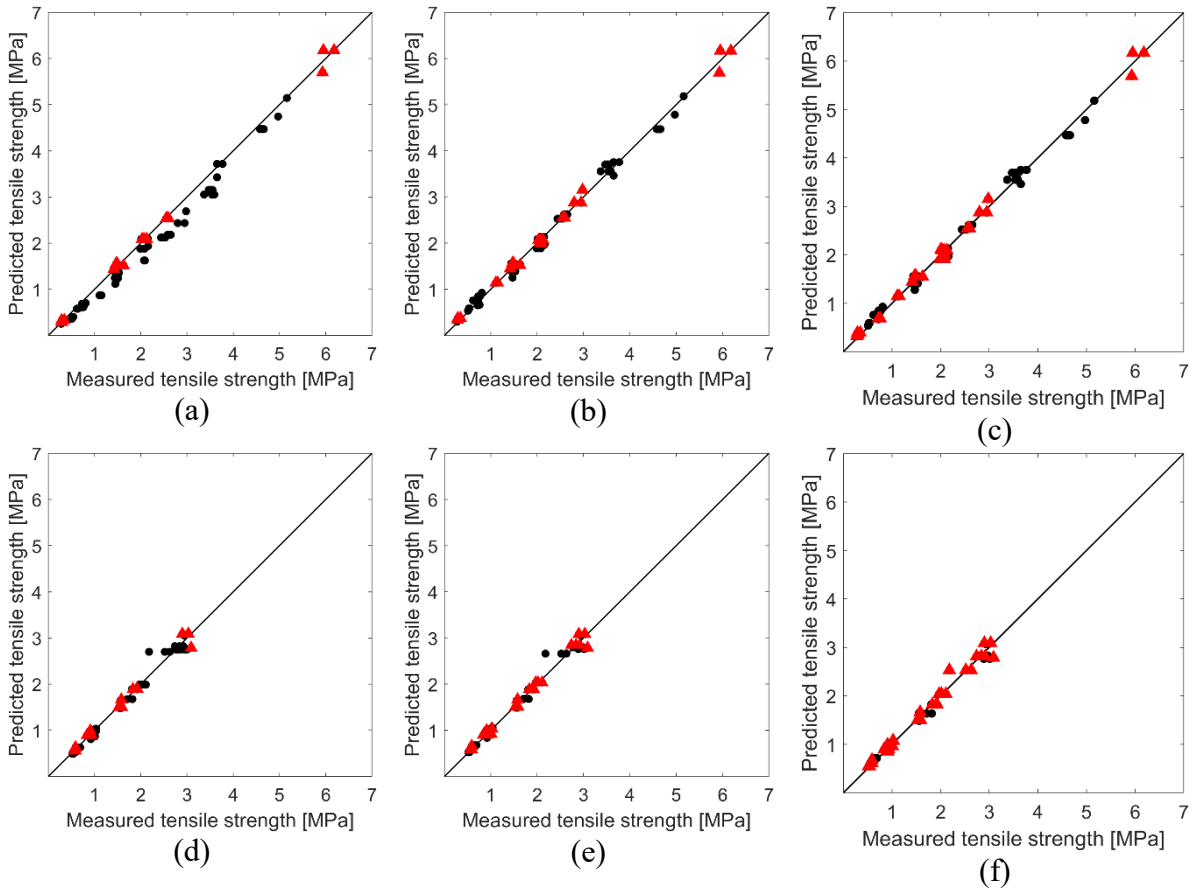



**Figure 3.2.** *Predicted and measured tensile strength, including both calibration and validation experiments. Triangles denote data used in calibration and validation, dots denote data used in validation only. Tensile strength is predicted with the model calibrated using the following optimal data: (a) two optimal profiles of formulation A; (b) three optimal profiles of formulation A; (c) four optimal profiles of formulation A; (d) two optimal profiles of formulation B; (e) three optimal profiles of formulation B; (f) four optimal profiles of formulation B. Optimal profiles are obtained through a parallel two-step MBDoE procedure.*

## 3.4.2 Sequential MBDoE

In this section, optimal experiments for formulation A are calculated by means of sequential two-step MBDoE procedure (Table 3.5). Model calibration is performed after measuring every new optimal profile; model validation included also blends at 100 dm, 400 dm, 718 dm (Table 3.1, Section 3.4.1).

**Table 3.5.** *Formulation A: validation of the extended Kushner and Moore model calibrated with optimal data obtained through the sequential two-step MBDoE procedure at increasing number of lubrication levels.*

| Lubrication extent $k$ (dm) | Purpose |
|---|---|
| 90 | calibration and validation |
| 2000 | calibration and validation |
| 354 | calibration and validation |
| 1849 | calibration and validation |

Results in terms of parameters precision are shown in Table 3.6; they are almost identical to those obtained with the parallel method. Indeed, also in the sequential case two optimal profiles are not sufficient to get statistically satisfactory parameter estimates, while addition of the third optimal profile allows estimating all parameters precisely. The only exception is again parameter $A_2$, which passes the t-test when four optimal profiles are employed.

**Table 3.6.** *Formulation A: identification of the extended Kushner and Moore model parameters by sequential two-step MBDoE procedure at increasing number of lubrication levels. An asterisk (*) denotes a parameter not passing the statistical significance test.*

| No. of profiles | Parameter [-] | Estimate ± 95% CI | *t*-value 95% | Reference *t*-value 95% |
|---|---|---|---|---|
| **2** | $A_1$ | 0.30±5.54 | 0.055* | 1.771 |
| | $B_1$ | −4.69±9.97 | 0.471* | |
| | $A_2$ | 0.48±4.89 | 0.146* | |
| | $B_2$ | 0.88±2.27 | 0.388* | |
| | $\Gamma$ | −3.5±233.35 | 0.015* | |
| **3** | $A_1$ | 0.28±0.04 | 7.213 | 1.717 |
| | $B_1$ | −4.75±0.32 | 14.816 | |
| | $A_2$ | 0.51±0.33 | 1.565* | |
| | $B_2$ | 0.87±0.12 | 6.970 | |
| | $\Gamma$ | −2.23±0.35 | 6.428 | |
| **4** | $A_1$ | 0.28±0.037 | 7.413 | 1.696 |
| | $B_1$ | −4.75±0.31 | 15.258 | |
| | $A_2$ | 0.48±0.24 | 1.983 | |
| | $B_2$ | 0.85±0.09 | 9.305 | |
| | $\Gamma$ | −2.27±0.33 | 6.781 | |

Results in terms of model prediction performance and model adequacy are shown in Table 3.7 and Figure 3.3. Using two optimal profiles for model calibration does not allow achieving the desired model predictive power, as demonstrated by the relatively low $R^2$ and high RMSE and $ts$ AE, as well as by the considerable number of points not lying on the diagonal of the parity plot (Figure 3.3.a). Moreover, two optimal profiles are not enough to satisfy the $\chi^2$ test, nor the $ts$ AE acceptance criterion.

However, addition of the third optimal profile is sufficient to improve the model performance significantly: the model is adequate (i.e., the $\chi^2$ test is passed); the $ts$ AE acceptance criterion is satisfied; the RMSE is sufficiently low; the $R^2$ is sufficiently close to 1; the majority of data points lie on the diagonal of the parity plot (Figure 3.b). Finally, adding the fourth optimal profile brings negligible improvement in both model adequacy and predictive power.

**Table 3.7.** *Formulation A: validation of the extended Kushner and Moore model calibrated with optimal data obtained through the sequential two-step MBDoE procedure at increasing number of lubrication levels. "% exceeding" indicates the percent of data points for which the absolute error on tensile strength is greater than 0.25 MPa. An asterisk (*) denotes a statistical significance test that is not passed.*

| Formulation | No. profiles used in calibration | RMSE | $R^2$ | % exceeding | $\chi^2_y$ | $\chi^2_{ref}$ |
|---|---|---|---|---|---|---|
| A | 2 | 0.224 | 0.97 | 25 | 210.535* | 100.749 |
| | 3 | 0.096 | 0.99 | 0 | 39.105 | |
| | 4 | 0.094 | 0.99 | 0 | 37.334 | |

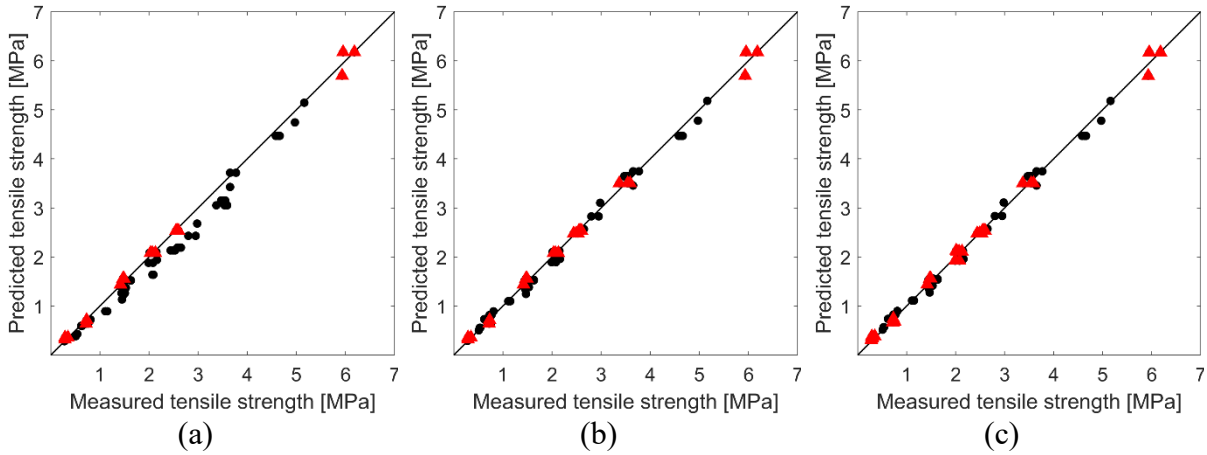



**Figure 3.3.** *Predicted and measured tensile strength, including both calibration and validation data. Triangles denote data used in calibration and validation, dots denote data used in validation only. In particular, ts was predicted with the model calibrated using the following optimal data of formulation A: (a) two optimal profiles; (b) three optimal profiles; (c) four optimal profiles. Optimal profiles were obtained through the sequential two-step MBDoE procedure*

## 3.5 Discussion

Results of section 3.4 show that the proposed methodology allows to reduce the experimental effort required to calibrate the extended Kushner and Moore model: indeed, the desired model predictive power and model adequacy are achieved; with concern to parameter precision, a higher uncertainty must be accepted for parameter $A_2$ in some formulations. In case of formulation B, results indicate that parameter $A_2$ cannot be estimated with sufficient precision. This appears related to the compactability properties of the formulation itself. However, this does not degrade the capability of the model to predict the process behavior, since model adequacy and predictive power are still satisfactory.

Therefore, considering that in a typical experimental campaign 7-9 powder blends are needed, the proposed MBDoE procedure leads to a reduction of the experimental burden by 60-70%, with great savings in terms of cost, labor and time.

Also note that the proposed experimental procedure not only cuts down the number of lubrication extents, thus the API usage, but also the time and labor needed for the experimentation. The fact that sequential and experimental designs provide very similar results means that it is possible to prepare all optimally lubricated blends in advance and compress them in the compactor simulator without interruptions, namely without recalculating optimal experimental conditions once new data are available.

## 3.6 Conclusions

Given a model linking blending operating conditions with the compaction performance in a tablet lubrication process, a model-based design of experiments approach based on a two-step optimisation procedure is proposed to design an experimental campaign for parameters estimation. Results demonstrated that it is possible to reduce the experimental effort by 60-70% with respect to standard industrial practice, with significant benefits in terms of API and labor savings. In addition, the comparison between parallel and sequential optimal designs reveal that model performance is not penalized if all optimal blends are designed with the same parameters guesses and are prepared in advance before being compressed. This allows streamlining the experimentation and better organizing the scheduling of lubrication experiments when multiple formulations must be compacted with the same equipment.

Future work will focus on the analysis of the robustness of the proposed approach, when sub-optimal experiments are collected due to practical and/or economical constraints, and on the systematic integration of structural identifiability analysis in the parameters precision task in order to mitigate (or, at least, identify) the effects of process-model mismatch.

# Chapter 4
# An exploratory model-based design of experiments approach to aid parameters identification and reduce model prediction uncertainty[2]

This Chapter deals with the study of the trade-off between experimental design space exploration and information maximisation, which is still an open question in the field of optimal experimental design. Moreover, in state-of-the-art optimal experimental design methods, the uncertainty of model prediction throughout the design space is not always assessed after parameter identification, although the maximisation of parameters precision does not guarantee that the model prediction variance is minimised in the whole domain of model utilisation. To tackle these issues, a novel MBDoE method is proposed: a mapping of model prediction variance (G-optimality mapping) is used in order to enhance space exploration and reduce model prediction uncertainty. This explorative MBDoE (eMBDoE) named *G-map eMBDoE* is tested on two simulated systems and compared against classical methods: (*i*) factorial design of experiments, (*ii*) Latin Hypercube (LH) sampling and (*iii*) MBDoE for parameters precision. The results show that G-map eMBDoE is more efficient in exploring the experimental design space when compared to a standard MBDoE and outperforms classical design of experiments

[2] Cenci, F., Pankajakshan, A., Bawa, G., Facco, P. and Galvanin, F. (2023). An exploratory model-based design of experiments technique to aid parameters identification and reduce prediction uncertainty. *Computers and Chemical Engineering*, **177**, 108353.

Cenci,F., Pankajakshan,A., Bawa, S. G., Gavriilidis, A., Facco, P.,Galvanin, F., 2023. An exploratory model-based design of experiments technique to aid parameters identification and reduce prediction uncertainty. In 33 European Symposium on Computer Aided Process Engineering, Kokossis, A. C., Georgiadis, M. C., Pistikopoulos, E. Eds., Comput. Aided Chem. Eng., Elsevier, **52**, 1-6.

Cenci, F., Pankajakshan,A., Bawa, S. G., Gavriilidis, A., Facco, P.,Galvanin, F., 2023. An exploratory model-based design of experiments technique to aid parameters identification and reduce prediction uncertainty [poster presentation]. 33 European Symposium on Computer Aided Process Engineering, Athens, Greece, June 18-21.

Cenci, F., Pankajakshan, A., Galvanin, F., Facco, P., 2023. Trade-off between space exploration and information maximization in experimental design [oral presentation]. Colloquium Chemiometricum Mediterraneum, Padova, Italy, June 27-30.

methods in terms of reduction of model prediction uncertainty and maximisation of parameters precision.

## 4.1 Introduction

Models are widespread in process industries for a variety of applications, from process understanding to product and process optimisation. For instance, simulations of a process model allow to evaluate the influence of the process conditions and/or disturbances (Prada et al., 2019) on the variables of interest, while the use of modelling at all stages of product development enables the compliance of the product to the clients' need and the selection of the most convenient manufacturing route (Mihaluta et al., 2008). The development of predictive models to be used in model-based activities requires the identification of model structure, i.e. the set of model equations, and model calibration, i.e. the precise estimation of model parameters from experimental data.

Improving the quality of data and information generation has a direct impact on model identification activities, thus several design of experiments (DoE) techniques have been developed in the last century at the purpose. One of the first is factorial DoE (Montgomery, 2013), which requires a limited preliminary knowledge on the system (Chapter 1). The data thereby generated are used to calibrate an empirical model, usually including first or second-order terms, which is eventually refined in order to exclude uninfluential factors and/or to add higher order terms. Factorial DoE can be beneficial to many industrial sectors, e.g. pharmaceutical industries (Singh et al., 2005 Part I and II), food science and technology (Granato and de Araújo Calado, 2013), manufacturing industries (Czitrom, 1999). Statistical DoE has brought many advantages in the experimentation with respect to the commonly used OFAT, but it has some limitations like the fact that process knowledge is not incorporated into the design as soon as new data are available and the scarce information obtained for the identification of first-principles models (see Chapter 1 for more details).

To overcome DoE limitations, MBDoE methods have been proposed that are centered on process knowledge (Espie and Macchietto, 1989), since physics-based models are employed in the calculation of the optimal experimental conditions for model discrimination, parameters estimation or minimisation of model prediction variance (see Chapters 1-2).

MBDoE for parameter precision select the most informative experimental conditions, namely the ones minimising the dimension of the uncertainty region of model parameters. Thus, fewer experiments are sufficient to identify statistically sound parameters, with great benefits in terms

of time, labor and resources. Consequently, MBDoE has been successfully applied to both industry and research: in chemical processes, like the production of aziridine through the C-H activation with a Pd-catalysis (Echtermeyer et al., 2017) or the execution of transient flow experiments to study the esterification of benzoic acid with ethanol (Waldron et al, 2020); in the production of renewably-sourced polymers like Cerenol, which has several applications in automotive, cosmetics and polymer specialties (Vo et al., 2021); in the pharmaceutical industry, e.g. the study of Michaelis-Menten kinetics for the production of a pharmaceutical agent (Shahmohammadi and McAuley, 2019); in civil engineering , e.g. for the determination of the optimal sensor locations to obtain the Young's moduli of tall structures (Reichert et al., 2021).

MBDoE methods determine experimental conditions that optimise a specific objective function and hence these optimal conditions are usually restricted in small regions of high information content. For instance, the optimal experiments to identify parameters of a kinetic model with two unknown parameters is made of a set of two distinct points; if several experiments are designed, they should fall in one of the two optimal conditions to avoid information loss (Box, 1968). However, this feature of MBDoE may lead to a scarce exploration of the design space, which in turn may result in poor predictive capability of the model, particularly in unexplored regions of the experimental design space.

The minimisation of parameters uncertainty ensured by MBDoE data does not necessarily imply that the entire design space is characterised by a minimisation of model prediction uncertainty. In literature, approaches have been proposed to evaluate the regions of model reliability within the design space based on different criteria to evaluate the prediction error. Dasgupta et al. (2021) built a map of supremum of the mean squared prediction error (SMSPE) using a kriging interpolating technique, while Quaglio et al. (2018) mapped the design space through a reliability function that depends mainly on the difference between predicted and measured responses. As an alternative, model prediction uncertainty can be quantified in terms of model prediction variance using the so-called "G-optimality" (Smith, 1918; Kiefer and Wolfowitz, 1959), a metric which can be evaluated in the whole design space to detect regions of model reliability without increasing the experimental burden. G-optimality has been explored in an MBDoE context; for instance, it has been used as an objective function in order to determine the optimal experimental conditions that minimise the response prediction variance (Smith, 1918; Kiefer and Wolfowitz, 1959). Moreover, the relationship between D-optimality and G-optimality (namely, between maximisation of the FIM determinant and minimisation of G-optimality, respectively) has been analysed for different types of models in

order to define the specific conditions under which the equivalence of the two criteria holds. For example, the equivalence of D-optimality and G-optimality is demonstrated by Kiefer and Wolfowitz (1960) for linear models with homoscedastic errors. Instead, Wong (1995) demonstrates that the equivalence between D- and G- optimality rarely holds in case of heteroscedastic models. Prus (2019) discusses the features of G-optimal designs with random coefficient regression models and states that the equivalence with D-optimal designs does not hold in general with this type of models. In addition, the classical G-optimal criterion for MBDoE is modified by Stigler (1971) in order to allow for a few experiments suitable to check the adequacy of model structure (namely, to assess process-model mismatch). However, to the author's knowledge G-optimality has never been used to enhance space exploration of MBDoE designs, to precisely estimate parameters and reduce model prediction variance in the whole design space with the minimum experimental effort. Furthermore, a formal description of the relation between G-optimality and other MBDoE criteria for non-linear systems with general variance models is still lacking in the scientific community. Therefore, a numerical approach that is not strictly related to a specific type of model is adopted in this work and validated with simulated data. In this Chapter, these issues are tackled by proposing a general new technique that integrates G-optimality maps into the conventional MBDoE optimisation framework for parameter estimation.

Section 4.2 describes the proposed MBDoE method, highlighting the novelties with respect to the conventional MBDoE explained in Chapter 2. Methods to analyse model performance are explained, too. Section 4.3 shows the results of the application of G-map eMBDoE to two simulated systems, including a comparison with exploitation-based methods, i.e. MBDoE, and exploration-based ones, i.e. factorial DoE and LH. Finally, in Section 4.4 conclusions are drawn and future works are proposed.

## 4.2 Mathematical modelling

### *4.2.1 Explorative MBDoE (eMBDoE) based on G-optimality maps*

The novel MBDoE method proposed in this work aims at enhancing space exploration, precisely estimating model parameters and minimising model prediction variance across the whole design space with a minimum experimental effort. To this aim, the calculation of model prediction variance is included within the MBDoE optimisation framework. More specifically,

mapping of G-optimality values is performed to obtain an explorative MBDoE (*eMBDoE*) method; therefore, the novel method is named *G-map eMBDoE*.

Figure 4.1a shows the standard sequential procedure for MBDoE (Espie and Macchietto, 1989; Asprey and Macchietto, 2000), where optimal experimental design, experiment execution and model calibration are carried out sequentially in the design of $N_e$ experiments.

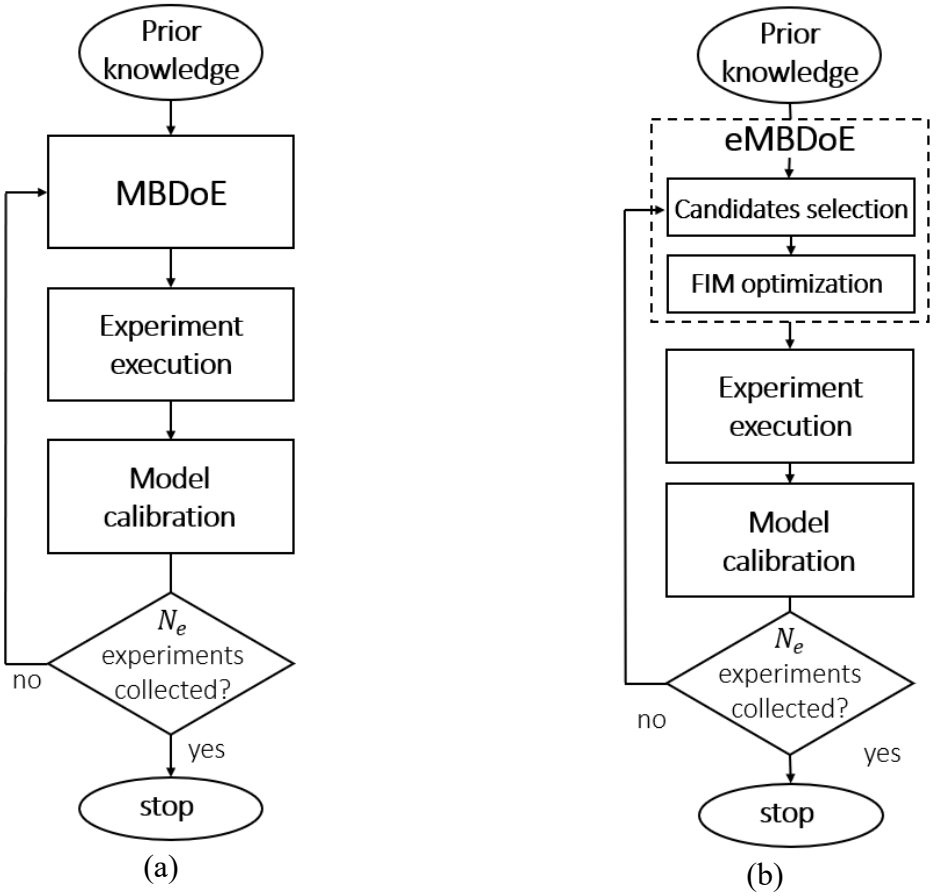



**Figure 4.1.** *Workflow of a conventional (a) sequential MBDoE and of a (b) sequential G-map eMBDoE.*

Similarly, Figure 4.1b shows the sequential procedure of the proposed G-map eMBDoE, in which optimal experimental design is carried out using the novel G-map eMBDoE method. The following Subsections 4.2.1.1-4.2.1.3 provide more details on each step of the proposed G-map eMBDoE procedure.

## 4.2.1.1 Prior knowledge

The prior knowledge is defined as the information needed to initialise the MBDoE or eMBDoE procedure. It includes:

- structurally identifiable models **f**;

- a set of preliminary experiments to obtain initial parameter estimates to initialise the MBDoE or G-map eMBDoE procedure. Preliminary experiments are usually designed by means of factorial DoE or LH;
- upper and lower bounds for each control variable included in the design vector $\boldsymbol{\varphi}$;
- variance-covariance matrix of measurement error $\boldsymbol{\Sigma}_y$.

This information allows to calculate both FIM $\mathbf{H}_{\hat{\boldsymbol{\theta}}}$ and G-optimality $\mathbf{V}_y$ as defined in Chapter 2.

### 4.2.1.2 G-map eMBDoE design

In the sequential MBDoE procedure shown in Figure 4.1a, the optimal design $\boldsymbol{\varphi}_{\text{opt}}$ is given by the solution to the optimisation problem presented in Chapter 2, Section 2.1. Whereas, in the G-map eMBDoE shown in Figure 4.1b, the most informative experiment is evaluated using an additional step as described below and illustrated in Figure 4.2:

- Step 1: design space characterisation. Each experimental condition in the design space is characterised in terms of model prediction variance, represented by scalar indices $J_G$, which leads to a map of G-optimality named *G-maps*. Similarly, every point in the design space is characterised in terms of information content, represented by the scalar measure $\psi$, generating a map of FIM-based information named *H-map*;
- Step 2: candidates selection. Experiments that satisfy a threshold $J_{G,thr}$ on model prediction variance represented by $J_G$ are retained for the subsequent optimisation (blue points in Figure 4.2);
- Step 3: information maximisation. Among these candidate design points, the experimental condition maximising information is chosen as the optimal experiment (red square in Figure 4.2).

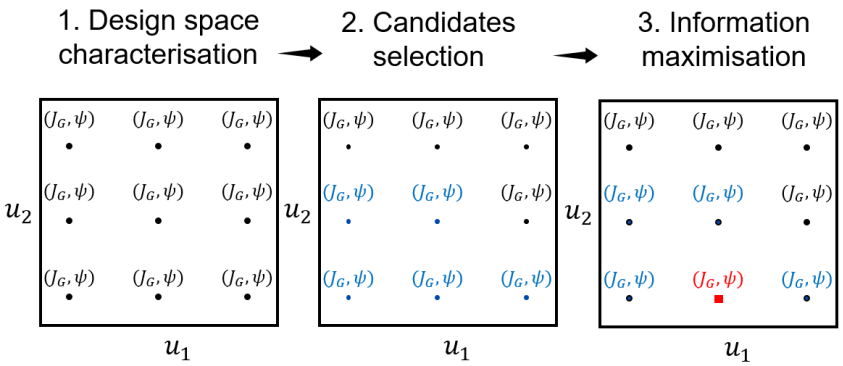



**Figure 4.2**. *Schematic representation of the novel eMBDoE method based on G-maps. Grids refer to two general control variables $u_1$ and $u_2$; $J_G$ and $\psi$ indicate the G-optimality index and the FIM scalar measure selected.*

This procedure has two main degrees of freedom:

1. the definition of the scalar index $J_G$;
2. the definition of the G-optimality-based requirement to be satisfied.

In this work, the following settings are used:

a) for a given point in the grid (i.e., for a given $\boldsymbol{\varphi}$), the prediction variance $\mathbf{V}_y(\hat{\boldsymbol{\theta}}, \boldsymbol{\varphi})\big|_{j,i}$ of each response at every time point is calculated and then summed to obtain a single scalar $J_G$:

$$J_G = \sum_{j=1}^{N_y} \sum_{i=1}^{N_{sp}} \boldsymbol{V}_y(\hat{\boldsymbol{\theta}}, \boldsymbol{\varphi})\big|_{j,i}; \tag{4.1}$$

b) the G-optimality-based constraint to be satisfied by the candidate design points is:

$$J_G \geq J_{G,thr} J_{G,max}, \tag{4.2}$$

where $J_{G,\max}$ is the maximum value of G-optimality in the grid, while $J_{G,\text{thr}}$ is a threshold chosen by the user such that: $0 \leq J_{G,\text{thr}} \leq 1$. More specifically, $J_{G,\text{thr}} = 0$ means that all points in the grid are candidates to solve the MBDoE optimisation (Chapter 2), therefore the design becomes equivalent to a standard MBDoE. The closer $J_{\text{G,thr}}$ gets to 1, the fewer are the remaining design candidate design points, since only points having the highest model prediction variance are accepted.

The above optimal experimental design procedure of G-map eMBDoE can be translated into a constrained optimisation problem described by:

$$\begin{aligned} &\boldsymbol{\varphi}_{opt} = arg\min_{\boldsymbol{\varphi}} \psi(\boldsymbol{V}_{\hat{\boldsymbol{\theta}}}) \\ &\boldsymbol{\varphi} \text{ s.t. } J_G \geq J_{G,thr} J_{G,max}. \end{aligned} \tag{4.3}$$

Therefore, information is maximised considering only the candidate design points having $J_G \geq J_{G,\text{thr}} J_{G,\max}$. In this paper, the grid-search approach employed to solve Eq. (4.3) does not impact on the final result, since the grid is so fine that an optimisation over continuous variables would provide almost identical results (as shown by ad hoc simulations, omitted here for sake of conciseness). If the system dimensionality increases, the computational burden to generate the grids also increases. Therefore, it may be convenient to build grids to initialise the procedure and to set the result obtained as an initial guess for further optimisation over continuous variables.

### 4.2.1.3 Experiment execution and iterative model calibration

Once the new experiment $\boldsymbol{\varphi}_{\text{opt}}$ is designed, it can be executed either in the physical process or in the simulated one. The new acquired measurement is added to the calibration dataset

collected up to the previous iteration and model parameters are estimated using the maximum likelihood method (Chapter 2). Then, a criterion is assessed in order to decide whether to continue the experimental campaign or not. This criterion is user-defined and can be based on model performance (such as parameter precision or model prediction accuracy) or on the maximum allowed experimental budget. The latter is used in this paper, which implies the experimental campaign is terminated when the experimental budget of $N_e$ experiments is reached. The performance of the model at every iteration of the sequential procedure is assessed after the model calibration step. The types of analyses used for this performance evaluation are described in section 4.2.2.

### *4.2.2 Model calibration analysis*

After calibrating the model with the new experiments designed and executed at the $i$-th iteration, the performance of the optimal experimental design procedure is assessed considering:

- space exploration;
- parameter precision;
- model prediction variance throughout the design space;
- FIM-based optimality metrics $\psi$ (Chapter 2) throughout the design space. Asprey and Macchietto (2000) suggested E-optimality ($\psi_{\mathrm{E}}$, Chapter 2) as the most effective criterion to use for the Model 2 (Section 4.3.2), being particularly effective on reducing parametric uncertainty when a sequential experimental design approach (as in this work) is adopted (Galvanin et al., 2007). Therefore, this criterion is used for both Model 1 and 2 for comparison purposes. However, ongoing work is showing that the advantages of G-map eMBDoE over conventional design techniques hold true also when different optimality criteria are used.

The corresponding analysis methods will be detailed in the following sub-sections.

#### **4.2.2.1 Space exploration**

The profile of each control variable is visualised to qualitatively compare the level of space exploration of the proposed experimental design techniques (eMBDoE, MBDoE or LH). As illustrated in Figure 4.3, two control variables $u_1$ and $u_2$ are represented in the x-axis and y-axis, respectively; red squares indicate the preliminary experiments used to initialise the

procedure; the blue squares represent the subsequent (eMBDoE, MBDoE or LH) experiments added iteratively in the sequential procedure. Figure 4.3a shows an example of an *exploratory design*, i.e. a design that covers the entire design space, typical of space-filling design methods like LH, while Figure 4.3b shows an example of an *exploitative design* where new experiments (blue squares) are designed in a limited region of high information content, frequently encountered in MBDoE applications.

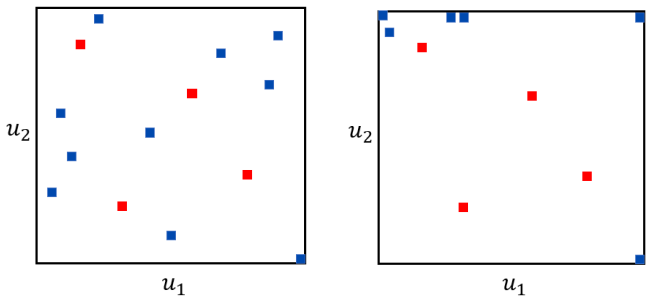



**Figure 4.3**. *Graphical representation of design space exploration for (a) an explorative, space-filling design; (b) an exploitative design (MBDoE).*

Since optimal experimental design methods tend to select replicated design points, i.e. optimal experiments with the same design vector $\boldsymbol{\varphi}_{\text{opt}}$, the different methods are compared in terms of number of distinct design points $\boldsymbol{\varphi}_{\text{opt}}$ (see Tables 4.2 and 4.4 of Section 4.3) to give an indication of space exploration.

### 4.2.2.2 Parameters precision

Parameter precision is assessed through *t*-tests, as explained in Chapter 2.

### 4.2.2.3 Maps of G-optimality and information content

At every iteration of the eMBDoE sequential procedure, two maps are built and compared:

- a G-optimality map (*G-map*), where $J_G$ is calculated at every point of the grid and displayed as a contour plot.
- an information map (*H-map*), where $\psi$ is calculated at every point of the grid and displayed as a contour plot.

The comparison between the two maps is useful to better understand which are the regions that would be selected based only on $\psi$ (i.e. regions of maximum information) and how much the method will move the design points from those regions by changing the threshold on $J_G$.

In addition to building G-maps using $J_G$ values, the distribution of G-optimality values ($J_G$ values) in the entire experimental design space at each iteration of the G-map eMBDoE procedure are represented by the following scalar indices:

$$J_{G,min} = min(J_{G,i}) , \qquad i = 1, \dots, N_\varphi \tag{4.4}$$

$$J_{G,mean} = mean(J_{G,i}) = \frac{\sum_{i=1}^{N_\varphi} J_{G,i}}{N_\varphi} , \qquad i = 1, \dots, N_\varphi \tag{4.5}$$

$$J_{G,max} = max(J_{G,i}) , \qquad i = 1, \dots, N_\varphi \tag{4.6}$$

where $J_{G,\text{min}}$, $J_{G,\text{mean}}$ and $J_{G,\text{max}}$ are the minimum, mean and maximum values of $J_G$, considering all the $N_\varphi$ points in the grid.

#### 4.2.2.4 Implementation of G-maps and H-maps

The two models of Section 4.3.1 and 4.3.2 are implemented in Python 3.9 (Rossum and Drake, 2009) and simulated in an Intel® Core™ i7-10875H CPU, @ 2.30 GHz processor with 64.0 GB RAM. The grid of points used to select candidate design points based on their G-optimality value is obtained by discretising the ranges of control variables into equal intervals. Maps of information content across the design space (*H-maps*) are also built using the same discretisation of the design space used with G-maps. The procedures of the two optimisation-based methods, namely MBDoE and G-map eMBDoE, differ mainly for the additional step of selection of candidate design points based on their $J_G$ values. To compare the methods in terms of required computational time, the time to design one experiment in each case (MBDoE or G-map eMBDoE) are reported in Section 4.3.1 and 4.3.2.

## 4.3 Results and discussion

Two simulated systems are used to compare the performance of the following design of experiments techniques in terms of space exploration, precise parameter estimation and minimisation of model prediction uncertainty:

- MBDoE: this is an exploitative design (Chapter 2);
- full-factorial DoE and LH: these designs are exploratory designs, i.e. they guarantee an exploration of the whole experimental design space;
- G-map eMBDoE: this design seeks a trade-off between information maximisation and space exploration through the definition of a threshold $J_{G,\text{thr}}$. Different values of $J_{G,\text{thr}}$ are considered.

The space-filling LH design is generated with the doepy package for Python (https://doepy.readthedocs.io/en/latest/). Preliminary simulations revealed that the results in terms of parameters precision and model prediction variance are similar regardless of random variations of the selected LH samples.

In both simulated systems, in silico data is generated according to the following procedure:

- model equations and true parameter vector $\boldsymbol{\theta}_{\text{true}}$ (see Tables 4.1 and 4.3) are used to generate the exact value of the model responses $y_{\text{exact}}$ at the selected experimental condition;
- a gaussian error with zero mean and a user-defined standard deviation $\sigma_y$ is then added to $y_{\text{exact}}$ to obtain a "noisy" measurement $y_{\text{noisy}}$.

The user-defined standard deviation $\sigma_y$ is chosen by the user to mimic the precision of the measurements in the physical system and can be typically evaluated from a set of preliminary replicated experiments. From now on, the models used to simulate the systems of Section 4.3.1 and 4.3.2 will be indicated as *Model 1* and *Model 2*, respectively.

To make the results comparable, the same initial settings are used for all design methods applied to a given simulated system: true model parameter vector ($\boldsymbol{\theta}_{\text{true}}$) for the in silico data generation; initial parameters values and lower and upper bounds ($\boldsymbol{\theta}_0$, $\boldsymbol{\theta}_{\text{LB}}$, $\boldsymbol{\theta}_{\text{UB}}$ respectively) for parameter estimation; standard deviation of the response measurement error ($\boldsymbol{\sigma}_y$); ranges for the control variables; set of preliminary experiments and total (maximum) number of experiments ($N_e$) to be executed. Finally, the selection of the proper threshold for G-optimality is case-dependent, therefore different sets of $J_{G,\text{thr}}$ will be considered for Model 1 and 2.

Finally, robustness of results to the selection of sampling points for Model 2 and different realisations of random noise for the response variables of Models 1 and 2 are shown in Appendix B.

### *4.3.1 Model 1*

Model 1 is made of the following two-inputs, single response algebraic model:

$$y = \theta_1 u_1 + \theta_2 u_1 u_2 + \theta_3 u_1^2 + \theta_4 u_2^2 + \theta_5 \, sin(u_1) \, . \tag{4.7}$$

Preliminary analysis showed that this model is structurally identifiable (by using the structural identifiability technique of Asprey and Macchietto, 2000), therefore MBDoE for parameters precision can be applied. Moreover, E-optimality criterion is used to optimise information content, since Asprey and Macchietto (2000) showed its efficacy for the Model 2 (in Section

4.3.2) and the same optimality criterion is used with both case studies for comparison purposes. Initial settings on variables and parameters are provided in Table 4.1. In this case, the design vector $\boldsymbol{\varphi}$ is equal to $\mathbf{u} = [u_1, u_2]$.

Different thresholds of G-optimality are employed for the explorative MBDoE: $J_{G,\text{thr}} \in \{0, 0.25, 0.50, 0.65, 0.75, 0.85\}$ to evaluate the impact of threshold choice on design performance. Notice that $J_{G,\text{thr}} = 0$ corresponds to a state-of-the-art E-optimal MBDoE, since the constraint in Eq. (4.3) becomes $J_G \geq 0$ and $J_G$ is always non-negative. All eMBDoE scenarios employ the E-optimal criterion.

**Table 4.1.** *Model 1: initial settings.*

| **Model features** | **Values** |
|---|---|
| | $u_1, u_2$ |
| Inputs | $u_1 \in [-10,10]$ |
| | $u_2 \in [-10,10]$ |
| Output | $y$ |
| standard deviation of measurement errors | $\sigma_y = 5$ |
| Parameters | |
| true values | $\boldsymbol{\theta}_{\text{true}} = [\theta_1, \theta_2, \theta_3, \theta_4, \theta_5] = [3.5, -2, 1.7, 1.1, 8]$ |
| initial value | $\boldsymbol{\theta}_0 = [1,1,1,1,1]$ |
| lower bounds | $\boldsymbol{\theta}_{\text{LB}} = [-10, -10, -10, -10, -10]$ |
| upper bounds | $\boldsymbol{\theta}_{\text{UB}} = [10,10,10,10,10]$ |

Even though LH and $4^2$ full factorial DoE experiments are designed at once and should be evaluated at the end of the experimental campaign, in this Section intermediate results are included in order to understand and compare the evolution of different design methods.

For all the scenarios, the same preliminary dataset is used: 5 experiments selected through a LH sampling. These preliminary experiments are used to achieve a first parameter estimation and to initialise $\mathbf{H}_{\hat{\boldsymbol{\theta}}}$ calculations to avoid potential singularity issues in the information matrix. Moreover, parameter estimates from each iteration become initial parameters values for the Maximum Likelihood estimation in the subsequent iteration. A maximum budget of 16 designed experiments is considered for each method; therefore, $N_e = 21$ experiments are obtained at the end of the experimental campaign.

All the methods are compared in terms of space exploration: Figure 4.4 shows the location of the $N_u = 2$ control variables within the entire design space, while Table 4.2 shows the number of distinct experimental conditions (i.e., different $\boldsymbol{\varphi}$) selected by different methods.

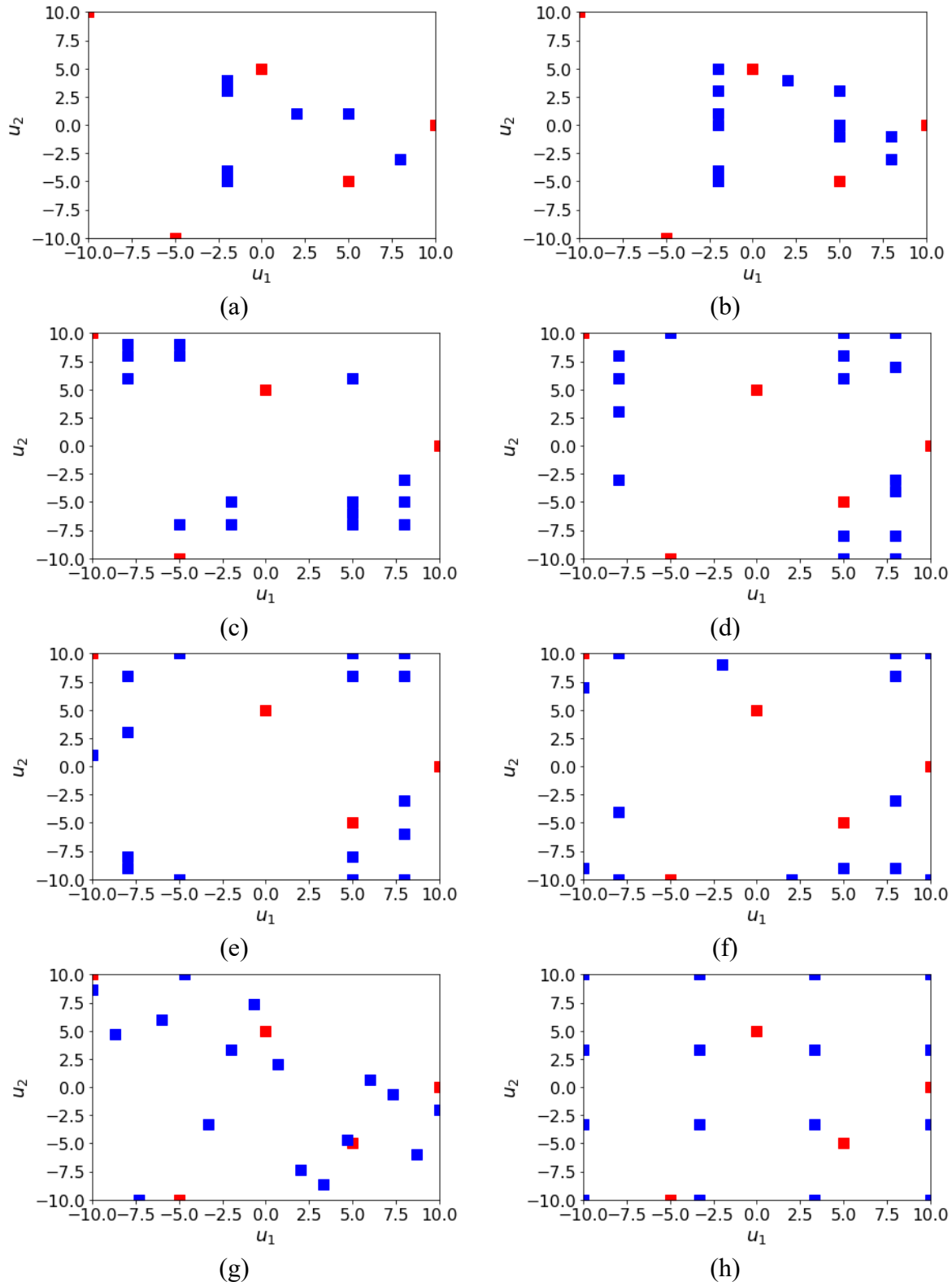



**Figure 4.4**. *Design space with the experiments selected by: (a) MBDoE ; (b) G-map eMBDoE $J_{G,thr}$=0.25; (c) G-map eMBDoE $J_{G,thr}$=0.50; (d) G-map eMBDoE $J_{G,thr}$=0.65; (e) G-map eMBDoE $J_{G,thr}$=0.75; (f) G-map eMBDoE $J_{G,thr}$=0.85; (g) Latin Hypercube; (h) $4^2$ full factorial DoE. Red squares indicate the 5 preliminary experiments*

**Table 4.2.** *Number of distinct design points for each scenario compared in the study.*

| Scenario | No. distinct design points |
|---|---|
| MBDoE | 7 |
| eMBDoE, thr:0.25 | 12 |
| eMBDoE, thr:0.50 | 15 |
| eMBDoE, thr:0.65 | 16 |
| eMBDoE, thr:0.75 | 16 |
| eMBDoE, thr:0.85 | 14 |
| LH | 16 |
| DoE | 16 |

These results show that the novel explorative MBDoE has the best performance in terms of space exploration when a threshold of 0.65-0.75 is selected (Figure 4.4d-e; 16 different design points indicated in Table 4.2). Similarly, LH (Figure 4.4g) and $4^2$ full factorial DoE (Figure 4.4h), which are inherently explorative, select 16 distinct points that cover all regions of the design space. Moreover, the smaller the threshold $J_{G,\text{thr}}$, the less eMBDoE experiments are spread in the design space (see Figures 4.4b-d): with $J_{G,\text{thr}}$=0.50, 15 different optimal experiments are selected; with $J_{G,\text{thr}}$=0.25, 12 different are selected; with $J_{G,\text{thr}}$=0, namely a conventional E-optimal MBDoE, only 7 distinct points are selected (Table 4.2).

Although G-map eMBDoE with $J_{G,\text{thr}}$=0.85 is one of the most explorative methods (Figure 4.4f) and has the greatest reduction of G-optimality throughout the entire design space (results are provided in Appendix C), it is the eMBDoE scenario that requires more experiments to precisely estimate parameter $\hat{\theta}_5$ (Appendix C). Hence, to find a better trade-off between space exploration and information maximisation, eMBDoE with $J_{G,\text{thr}}$=0.75 is considered in the analysis of precise parameter estimation. For this purpose, the *t*-values calculated at every iteration for the full set of model parameters are shown in Figure 4.5. As shown in Figure 4.5, the most critical parameter which requires a higher number of calibration experiments to be precisely estimated is $\theta_5$ (Figure 4.5e). MBDoE and eMBDoE with $J_{G,\text{thr}}$=0.75 require 10 experiments (5 preliminary and 5 optimal) to pass the *t*-test, LH requires 17 experiments, while DoE is not able to pass the *t*-test. Finally, details on parameters accuracy (i.e. distance from the assumed true value) can be found in Appendix C; moreover, reproducibility of the LH results despite random variations of different LH designs is shown in Appendix C.3.

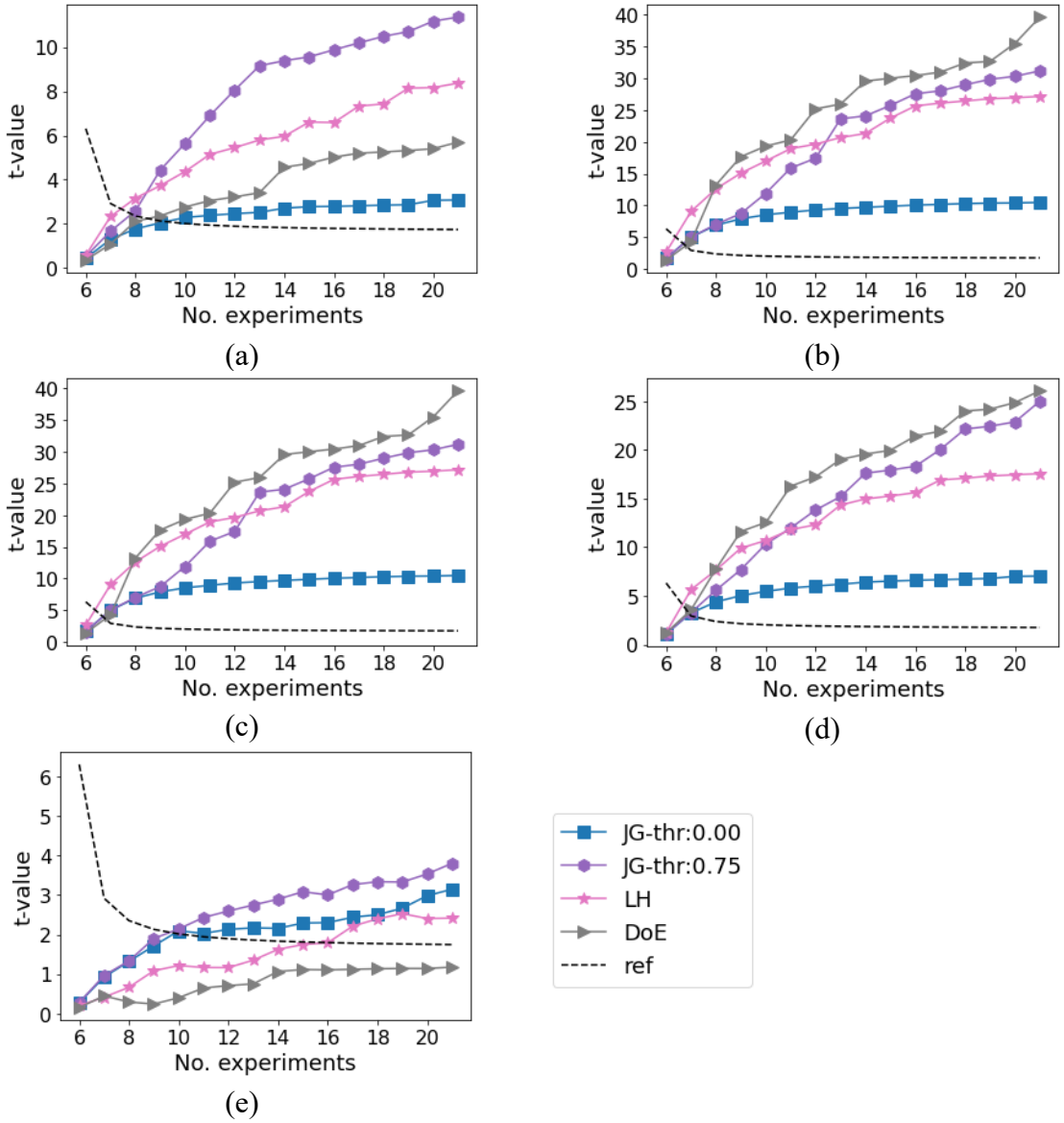


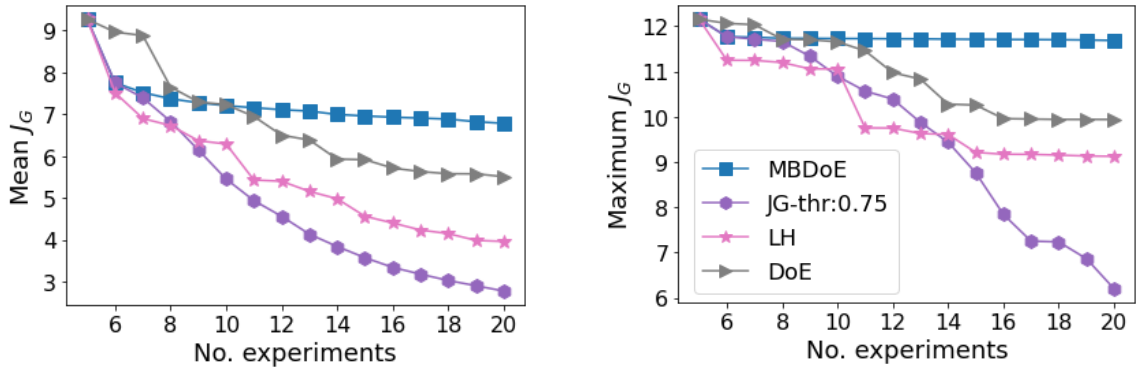

**Figure 4.5**. *Profiles of t-values calculated with: MBDoE ($J_{G,thr}$=0.00); G-map eMBDoE ($J_{G,thr}$=0.75); Latin Hypercube (LH); $4^2$ full factorial DoE. Figures (a)-(e) show results of parameters 1-5, respectively. t-values are compared against the reference t-value ('ref' in the legend). Only t-values referred to the 16 optimal/explorative data are shown.*

In Figure 4.6 the different scenarios are compared in terms of model prediction variance across the whole experimental design space. Results are shown for MBDoE, G-map eMBDoE with $J_{G,\text{thr}}$=0.75 threshold, factorial DoE and LH. The smaller the scalar index of $J_G$, the better the performance in terms of reduction of model prediction uncertainty. Scalar indices are calculated for the G-maps generated during experiment design step: therefore, the iterative generation of G-maps starts with a calibration dataset of 5 preliminary experiments and terminates with a calibration dataset of 20 experiment, which is the map used to calculate the last optimal experiment since $N_e$=21. The following ranking is obtained in terms of scalar measures of model prediction variances:

- mean G-optimality $J_{G,\text{mean}}$, from the 11$^{\text{th}}$ experiment onwards (Figure 4.6a): MBDoE > DoE > LH > eMBDoE ($J_{G,\text{thr}}$=0.75)
- maximum G-optimality $J_{G,\text{max}}$, from the 14$^{\text{th}}$ experiment onwards (Figure 4.6b): MBDoE > DoE > LH > eMBDoE ($J_{G,\text{thr}}$=0.75)

Instead, the minimum G-optimality is equal to zero in all scenarios throughout the experimental campaign, therefore it is omitted here.

To conclude, scalar indices of G-optimality prove that eMBDoE with $J_{G,\text{thr}}$=0.75 has the best performance in terms of reduction of model prediction variance. Moreover, compared to MBDoE, both mean and maximum values of G-optimality are smaller in explorative design methods such as DoE and LH. This suggests that space exploration promotes the reduction of prediction uncertainty, but the best overall result (i.e. minimum prediction variance and maximum parameter precision) is only achieved when a trade-off between space exploration and information maximisation is realised.




(a) (b)

**Figure 4.6.** *Profiles of scalar indices of G-optimality including (a) mean G-optimality; (b) maximum G-optimality calculated for: MBDoE ($J_{G,thr}$=0.00); G-map eMBDoE ($J_{G,thr}$=0.75); Latin Hypercube (LH); $4^2$ full factorial DoE.*

G-maps are shown in Figures 4.7 and 4.8 to visualise the regions of higher prediction variance within the design space for MBDoE, G-map eMBDoE with $J_{G,\text{thr}}$=0.75, LH and DoE. This result is compared to the maps of information content (H-maps, Figures 4.9, 4.10) for every scenario. For sake of conciseness, only a subset of G-maps and H-maps are included: *i*) calibration dataset of 6 experiments (5 preliminary and 1 optimal): maps obtained after the first iteration; *ii*) calibration dataset of 20 experiments (5 preliminary and 15 optimal): maps generated in the last iteration.

In these maps, different experimental conditions are highlighted: data from experiments already performed, which are used to calibrate the model (orange squares); candidate design points based on the G-optimality threshold (black dots); experiment selected at the current iteration

(red dot). Notice that the discretisation of the design space and the selection of candidates (black dots) are not performed in LH and DoE, therefore black points are not present in these figures (see Figures 4.7c-d and 4.8c-d). Finally, it must be noticed that:

- G-maps represent model prediction variance, which must be minimised; therefore, the best performance is found in blue regions of Figures 4.7 and 4.8 (i.e., the smallest G-optimality values) and the worst one is found in yellow regions (i.e., the highest G-optimality values);
- H-maps represent the amount of information, based on the E-optimal criterion, which must be maximised; in this case, the best performance is found in dark green regions of Figures 4.9 and 4.10 (i.e., high information values) and the worst performance is found at the red regions (small information values).

After measuring the first optimal experiment (Figure 4.7a-4.7d), the G-maps of MBDoE, eMBDoE and LH are quite similar in terms extension of regions with small model prediction variance (blue regions). Moreover, in all of them the G-optimality is smaller in central regions, while it increases towards extreme values of $u_1$ and $u_2$. The G-map of DoE (Figure 4.7d) is similar, but slightly worse due to a larger extension of regions with high model prediction variance (i.e., yellow regions).

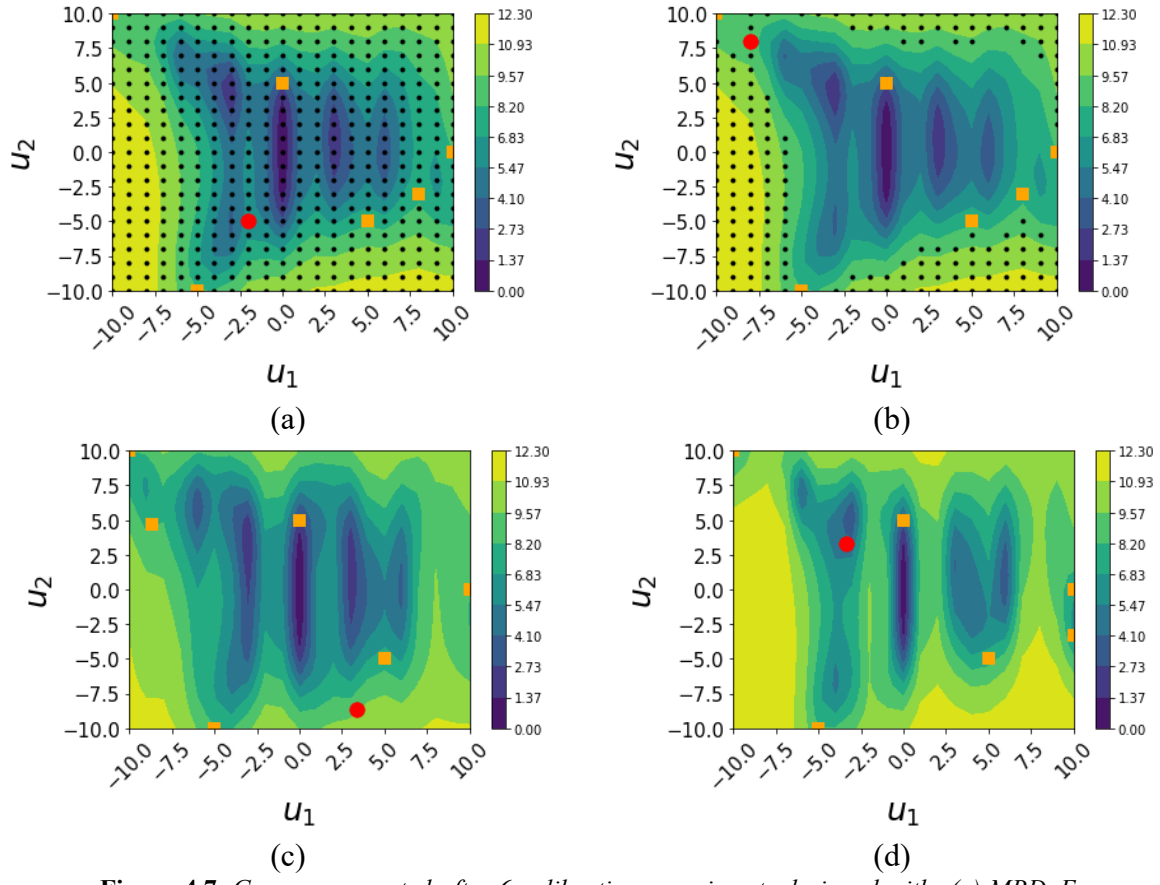



**Figure 4.7**. *G-maps generated after 6 calibration experiments designed with: (a) MBDoE, (b) eMBDoE and $J_{G,thr}$=0.75, (c) LH, (d) $4^2$ full factorial DoE.*

The differences in the distribution of G-optimality becomes more evident after measuring the 20th experiment: indeed, the best performance is achieved with eMBDoE using $J_{G,\text{thr}}$=0.75 (Figure 4.8b) since it has the largest extension of the blue region; moreover, the yellow regions at extreme values of the two control variables disappear. The second-best performance in terms of reduction of model prediction variance is found with LH (Figure 4.8c); instead, MBDoE and DoE (Figure 4.8a and 4.8d, respectively) still have regions with larger model prediction variance (yellow regions). This suggests that an explorative strategy as LH can improve the prediction precision with respect to an optimal design, but the best solution is found when a good trade-off between space exploration and information maximisation is achieved.

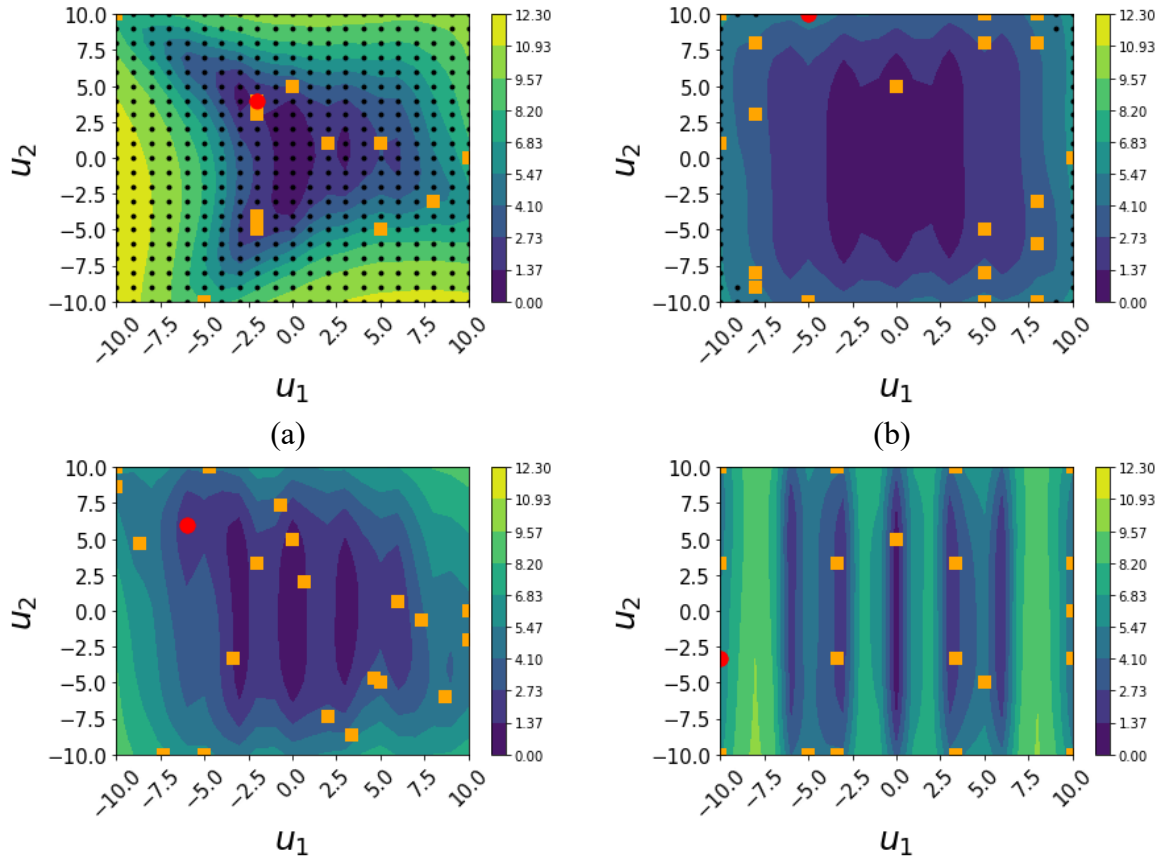



**Figure 4.8**. *G-maps generated after 20 calibration experiments designed with: (a) MBDoE, (b) eMBDoE and $J_{G,thr}$=0.75, (c) LH, (d) $4^2$ full factorial DoE.*

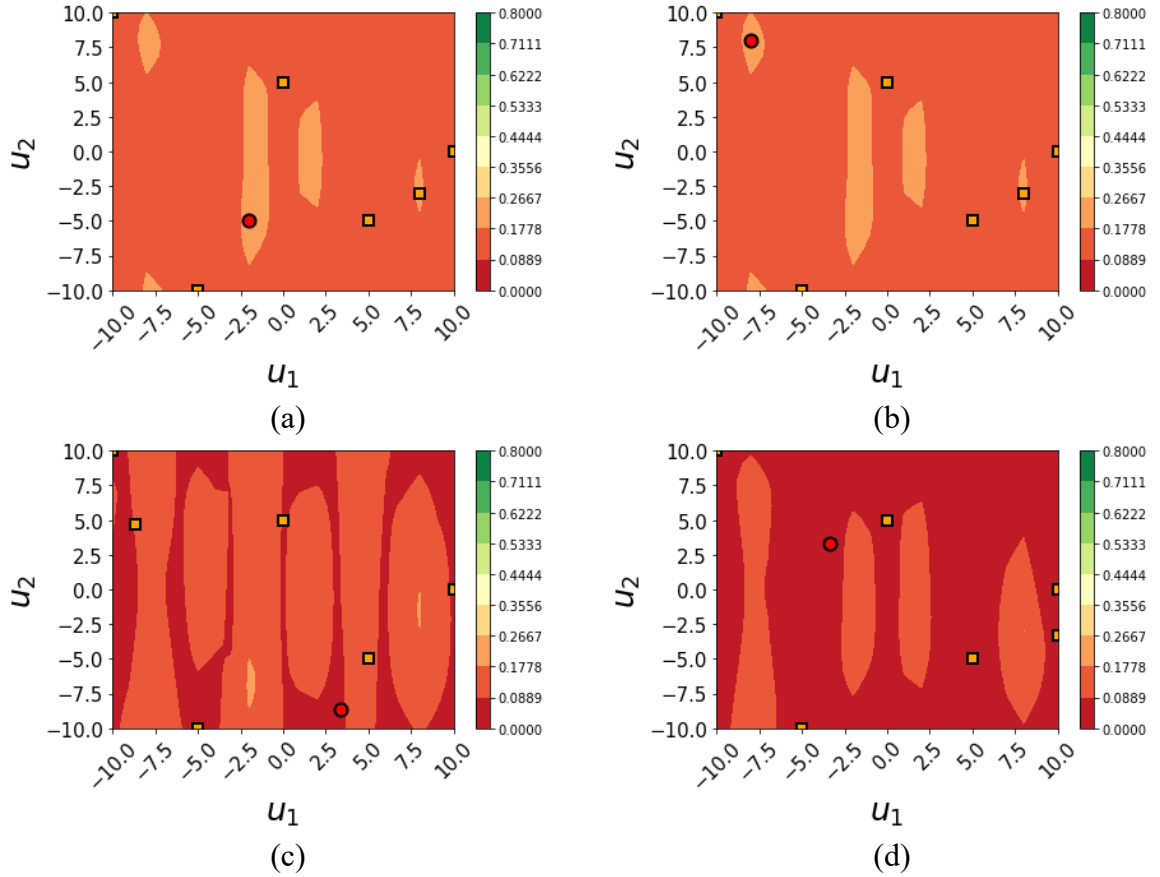



**Figure 4.9**. *H-maps after 6 experiments with: (a)MBDoE; (b)eMBDoE ($J_{G,thr}$=0.75); (c) LH; (d) DoE.*

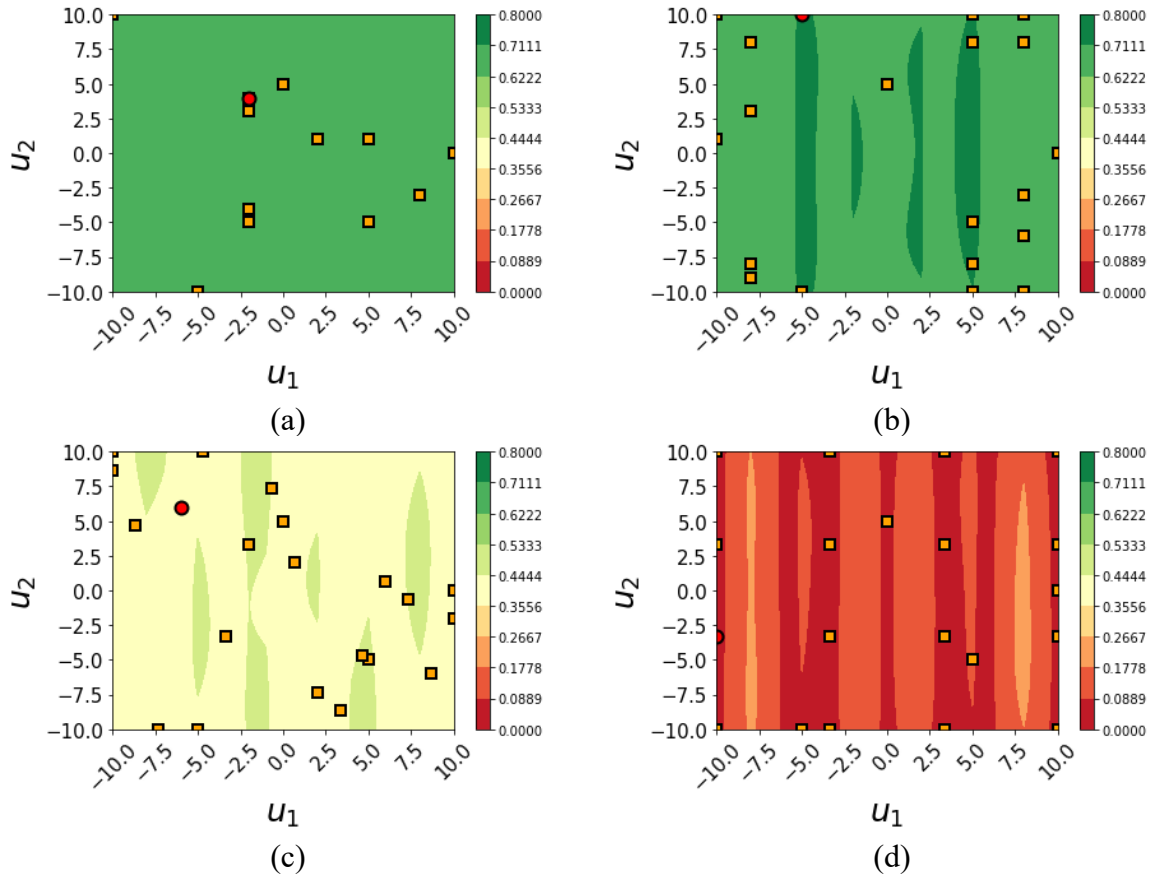



**Figure 4.10**. *H-maps after 6 experiments with: (a)MBDoE; (b)eMBDoE ($J_{G,thr}$=0.75); (c) LH; (d) DoE*

By looking at the H-maps generated with 6 measured experiment (Figure 4.9) and with 20 measured experiments (Figure 4.10), it is clear that MBDoE and G-map eMBDoE outperforms LH and DoE: indeed, their H-maps are red in the first iteration (Figures 4.9a-b) and become green in the subsequent iterations (Figures 4.10a-b), while LH and DoE end up with a light green (Figure 4.10c) and a red maps (Figure 4.10d), respectively. This further confirms that the enhancement of space exploration of the proposed explorative MBDoE does not entail a significant loss of information content with respect to a completely optimal design.

Finally, the computational times to build G-maps and H-maps (with E-optimal criterion) and to design one experiment with Python 3.9 in an Intel® Core™ i7-10875H CPU, @ 2.30 GHz processor with 64.0 GB RAM are:

- 0.12 seconds with G-map eMBDoE and $J_{G,\mathrm{thr}}$=0.75;
- 0.12 seconds with MBDoE.

### *4.3.2 Model 2*

The G-map eMBDoE is applied to the Canoid-type kinetic model describing the material balances of the fermentation of baker's yeast in a fed-batch reactor. Original model and proof of structural identifiability can be found in Asprey and Macchietto (2000).

The two-response dynamic model is represented by the following set of differential and algebraic equations:

$$r = \frac{\theta_1 x_2}{\theta_2 x_1 + x_2}, \tag{4.8}$$

$$\frac{dx_1}{dt} = (r - u_1 - \theta_4)x_1\,, \tag{4.9}$$

$$\frac{dx_2}{dt} = -\frac{r x_1}{\theta_3} + u_1(u_2 - x_1)\,. \tag{4.10}$$

Some simplifying assumptions are made: the inputs $u_1$ and $u_2$ are constant in time, and the number of sampling points is fixed ($N_{\mathrm{sp}} = 3$). Therefore, the design vector becomes: $\boldsymbol{\varphi} = \mathbf{u} = [u_1, u_2]$; $u_1$ is the dilution factor with range 0.05-0.20 $\mathrm{h}^{-1}$, while $u_2$ is the substrate concentration in the feed with range 5.0-35.0 g/L. The two measured concentrations are the biomass concentration $x_1$ [g/L] and the substrate concentration $x_2$ [g/L].

As in Section 4.3.1, model calibration data are obtained by using a simulated process: 'true' parameter vector $\boldsymbol{\theta}_{\mathrm{true}}$ is used in the model (Eqs. 4.8-4.10) to generate noise-free simulated responses for $x_1$ and $x_2$; then, gaussian noise with zero mean and a user-defined standard deviation $\boldsymbol{\sigma}_{\mathrm{y}}$ ($\sigma_y$ =[1.0, 1.0]$\mathrm{gL}^{-1}$, see Table 4.3) is added in order to mimic measurement errors.

Model parameters can be estimated using the in-silico calibration data starting from a set of initial parameters values $\boldsymbol{\theta}_0$ within the ranges $[\boldsymbol{\theta}_{\text{LB}}, \boldsymbol{\theta}_{\text{UB}}]$.

Details on settings and parameters and variables ranges are reported in Table 4.3.

**Table 4.3.** *Model 2: initial settings.*

| **Model features** | **Values** |
|---|---|
| *Inputs* | $\mathbf{u} = [u_1, u_2]$ |
| | $u_1 \in [0.05, 0.20]$ |
| | $u_2 \in [5.0, 35.0]$ |
| *Outputs* | $\mathbf{y} = [x_1, x_2]$ |
| *standard deviation of measurement errors* | $\boldsymbol{\sigma}_{\text{y}} = [1.0, 1.0]\text{gL}^{-1}$ |
| *initial conditions* | $\boldsymbol{y}_0 = [x_1(0), x_2(0)] = [1.0, 0.01]\ \text{gL}^{-1}$ |
| *sampling points* | $\mathbf{t}_{\text{sp}} = [7,14,21]\text{h}$ |
| *Parameters* | |
| *true values* | $\boldsymbol{\theta}_{\text{true}} = [\theta_1, \theta_2, \theta_3, \theta_4] = [0.31, 0.18, 0.55, 0.05]$ |
| *initial values* | $\boldsymbol{\theta}_0 = [5.0, 5.0, 5.0, 5.0]$ |
| *lower bounds* | $\boldsymbol{\theta}_{\text{LB}} = [-20, -20, -20, -20]$ |
| *upper bounds* | $\boldsymbol{\theta}_{\text{UB}} = [20,20,20,20]$ |

The same G-map based eMBDoE method applied to the algebraic model of Section 4.3.1 can be applied to this model, but a higher level of complexity is introduced here. In the system of Section 4.3.1, the algebraic model was simulated to get the value of a single response variable that corresponds to the single measurement in an experiment at a particular time instance, such as at steady state. However, in the system of this Section, the dynamic model is simulated to obtain output responses at different time points, which corresponds to a typical fed-batch experiment with multiple sampling points in time. Since $N_{\text{sp}}$=3 is set, 6 values of model prediction variance can be calculated: $V_y$ of $x_1$ at $\mathbf{t}_{\text{sp}} = [7,14, 21]$ h and $V_y$ of $x_2$ at $\mathbf{t}_{\text{sp}} = [7,14, 21]$ h. To summarise these results, the sum $J_G$ of all contributions $V_y$ is calculated as in Eq. (4.1), which is used in the G-map eMBDoE method to design the optimal and explorative experiments. More details on the single contributions $V_y$ can be found in Appendix D.

As in Model 1, 4 different design of experiments techniques are compared:

- an E-optimal MBDoE, which is Gmap eMBDoE with $J_{G,\text{thr}}$ =0.00;
- two explorative designs, namely LH and $4^2$ full factorial DoE;
- a G-map eMBDoE with E-optimal criterion and different threshold values $J_{G,\text{thr}} \in \{0.25, 0.50, 0.65, 0.75, 0.85\}$.

Three preliminary experiments are designed with LH in order to initialise all the four design of experiments methods. Then, the maximum number of experiments designed in each scenario is fixed to 20, providing a total number of experiments $N_e$=23. The extent of space exploration realised by each method can be deduced from Figure 4.11 and Table 4.4.

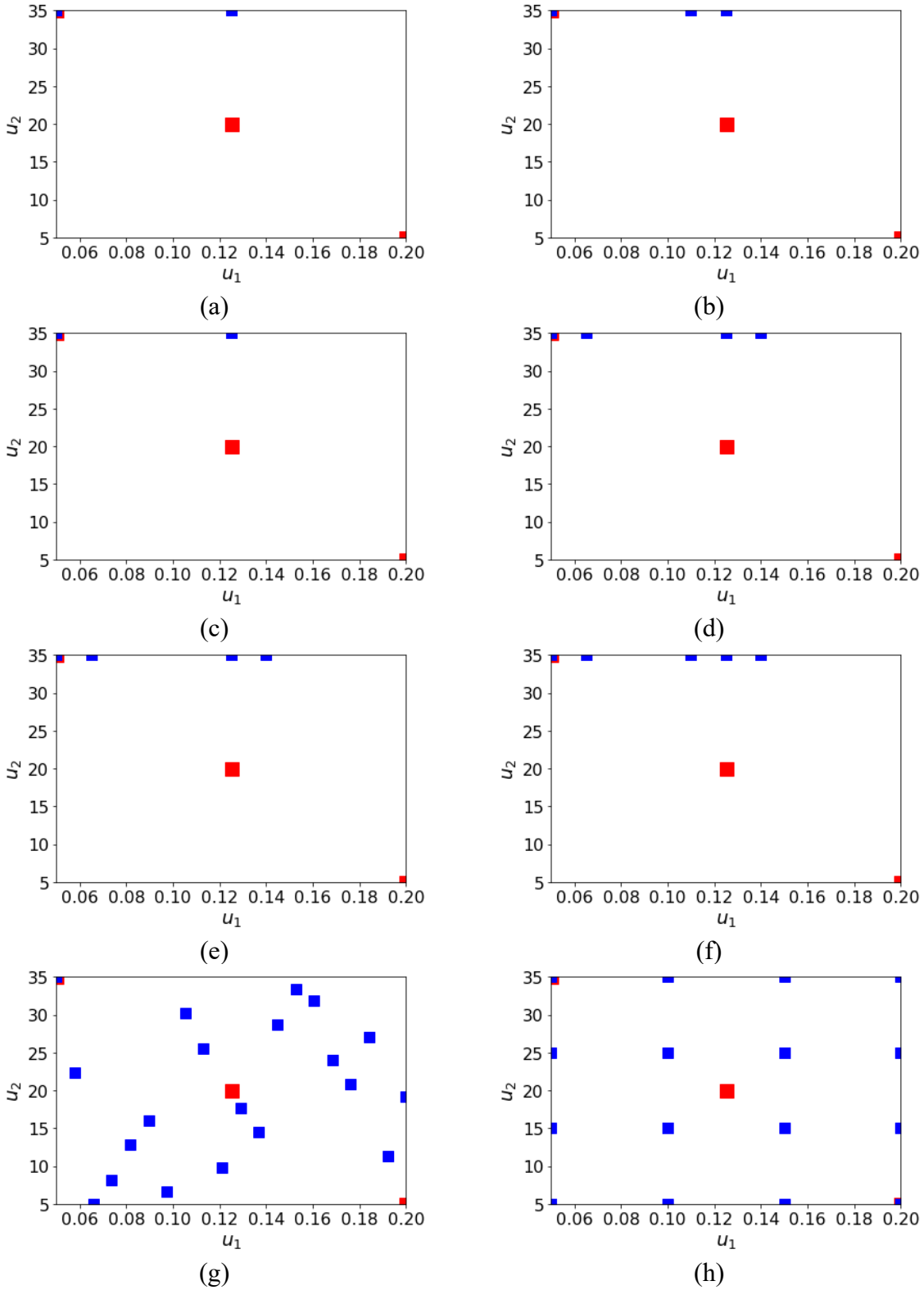



**Figure 4.11**. *Design space with the experiments selected by: (a) MBDoE ; (b) G-map eMBDoE $J_{G,thr}$=0.25; (c) G-map eMBDoE $J_{G,thr}$=0.50; (d) G-map eMBDoE $J_{G,thr}$=0.65; (e) G-map eMBDoE $J_{G,thr}$=0.75; (f) G-map eMBDoE $J_{G,thr}$=0.85; (g) Latin Hypercube; (h) $4^2$ full factorial DoE. Red squares indicate preliminary experiments, blue squares indicate experiments designed with the method considered.*

**Table 4.4.** *Number of distinct design points for every scenario compared in the study.*

| Scenario | No. distinct design points |
|---|---|
| MBDoE | 2 |
| eMBDoE, thr:0.25 | 3 |
| eMBDoE, thr:0.50 | 2 |
| eMBDoE, thr:0.65 | 4 |
| eMBDoE, thr:0.75 | 4 |
| eMBDoE, thr:0.85 | 5 |
| LH | 16 |
| DoE | 16 |

After 20 iterations of experiments design, MBDoE (Figure 4.11a) has selected replicates at 2 different experimental conditions and design space exploration is very limited. Instead, eMBDoE (Figure 4.11b-f) is able to increase space exploration in such a way as the number of replicated points is reduced, i.e., more distinct experimental conditions are obtained. This is evident with $J_{G,\text{thr}} \in \{0.65, 0.75, 0.85\}$ (Figure 4.11d-f). In these cases, the distinct points increase to 3 or 4 (Table 4.4) instead of the 2 selected by the conventional E-optimal MBDoE. The control variable which is affected by the $J_{G,\text{thr}}$ is $u_1$, whereas the different optimal designs select the same value for $u_2$ for all optimal experiments.

Parameters precision is assessed through *t*-tests; Figure 4.12 shows the results of MBDoE (i.e., $J_{G,\text{thr}} = 0$), eMBDoE with $J_{G,\text{thr}}$=0.75 (the others are omitted for sake of conciseness), LH and factorial DoE. The G-optimality threshold $J_{G,\text{thr}}$=0.75 is selected since it has the best performance in terms of precise parameters estimation and in reduction of model prediction variance (for more details, see Appendix D).

Parameters $\hat{\theta}_1$ (Figure 4.12a) and $\hat{\theta}_3$ (Figure 4.12c) pass the $t$-test with few experiments in every scenario, while the most critical parameters are $\hat{\theta}_2$ (Figure 4.12b) and, especially, $\hat{\theta}_4$ (Figure 4.12d). Conventional E-optimal MBDoE requires 2 optimal experiments to estimate parameter $\hat{\theta}_2$ (Figure 4.12b) and 12 experiments to estimate $\hat{\theta}_4$ (Figure 4.12d). The performance is improved by enhancing space exploration through G-map eMBDoE with $J_{G,\text{thr}}$=0.75, requiring 2 experiments to pass the $t$-test for $\hat{\theta}_2$ (Figure 4.12b) and 11 experiments to pass the t-test for $\hat{\theta}_4$ (Figure 4.12d). However, explorative designs such as factorial DoE and LH do not allow to precisely estimate $\hat{\theta}_4$ (Figure 4.12d) within the experimental budget of $N_e$=23 experiments. This suggests that a trade-off between space exploration and information maximisation provides the best results in terms of parameters precision.

Additional details on parameter estimation accuracy (i.e. distance from the assumed true value) can be found in Appendix D; reproducibility of the LH results is shown in Appendix D.

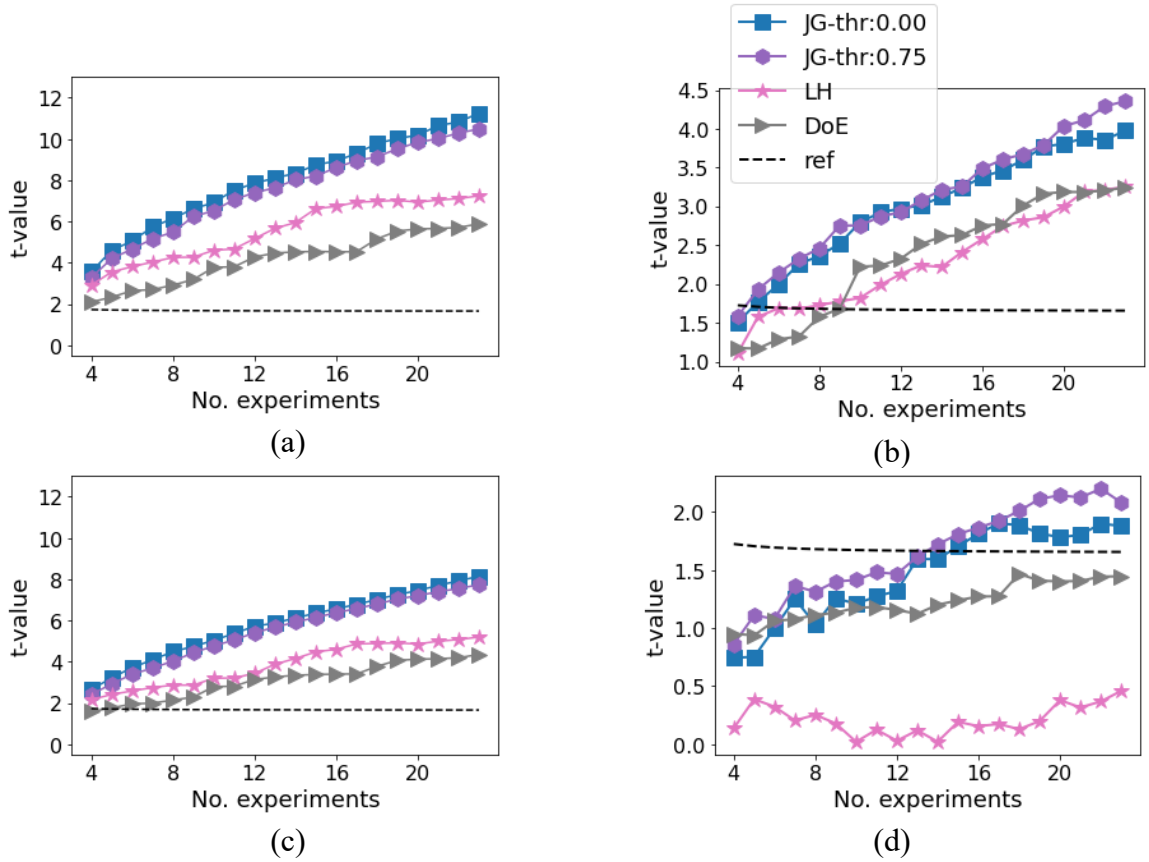



**Figure 4.12.** *Profiles of t-values calculated with: MBDoE ($J_{G,thr}$=0.00); G-map eMBDoE ($J_{G,thr}$=0.75); Latin Hypercube (LH); $4^2$ full factorial DoE. The t-values are compared against the reference t-value ('ref' in the legend) for parameters 1-4 in figures (a)-(d), respectively.*

As indicated in Eq. (4.1), the G-map for the selection of eMBDoE experiment is built by summing the prediction variance contributions from the two model responses. More details of the contributions to the overall $J_G$ can be found in Appendix D. To compare quantitively the grids of G-optimality values obtained at every iteration of MBDoE, eMBDoE, factorial DoE and LH, scalar measures of G-optimality are calculated: minimum, mean and maximum values of the $J_G$ calculated for each point in the grid (Eqs. 4.4-4.6).

Considering the minimum, mean and maximum values of G-optimality (Figures 4.13a-c), explorative designs such as LH and factorial DoE provide a slower reduction in model prediction variance at the beginning of the experimental campaign and they stabilise at higher values when the maximum experimental budget is reached. When $J_{G,\text{mean}}$ and $J_{G,\text{max}}$ are considered (Figure 4.13b-c), the explorative MBDoE has a better performance than conventional MBDoE and it is able to reduce the model prediction variance with the lowest number of experiments. This further suggests that the trade-off between space exploration and

information maximisation realised by eMBDoE leads to the best performance also in terms of prediction variance.

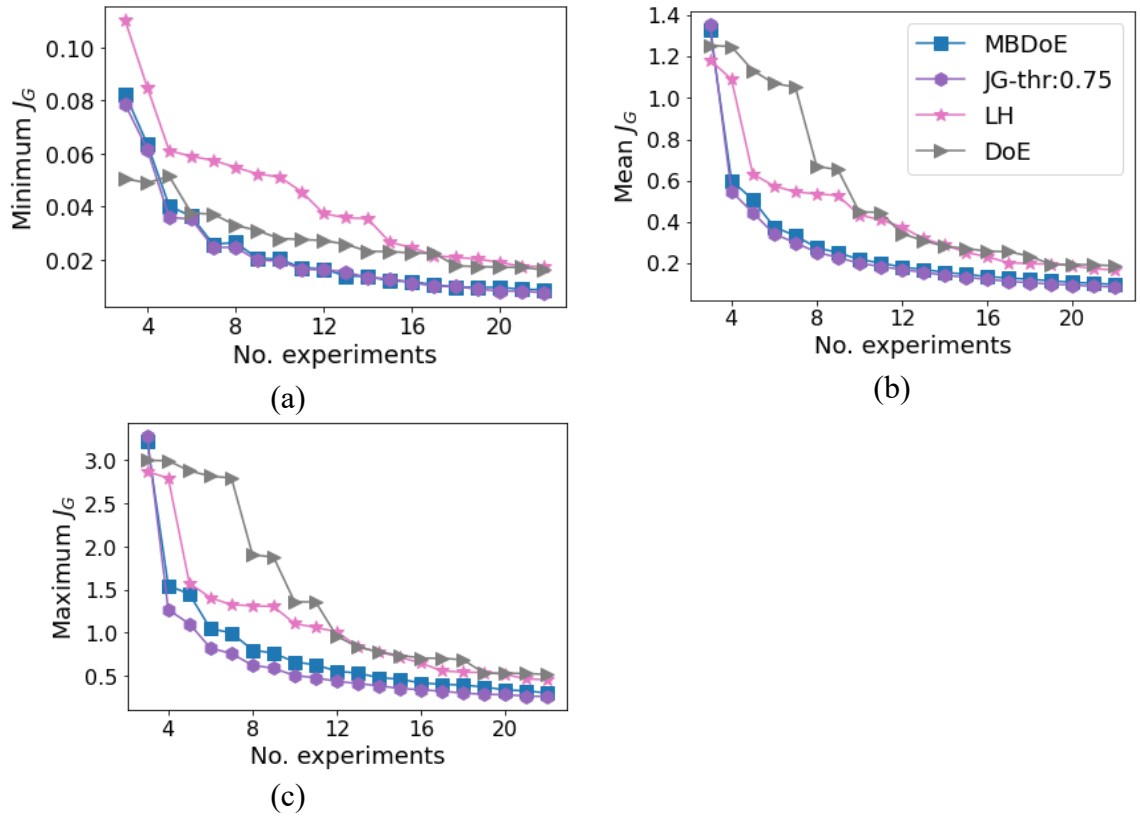



**Figure 4.13.** *Profiles of scalar indices of G-optimality calculated with: MBDoE ($J_{G,thr}$=0.00); G-map eMBDoE ($J_{G,thr}$=0.75); Latin Hypercube (LH); $4^2$ full factorial DoE. Results are obtained by summing up the variance calculated for every response at every time point. Three different scalar measures are considered: (a) minimum G-optimality; (b) mean G-optimality; (c) maximum G-optimality.*

The different experimental design methods are compared in terms of G-maps and H-maps:

- E-optimal MBDoE (Figure 4.14a, 4.15a, 4.16a and 4.17a);
- eMBDoE, with E-optimal criterion and a threshold of $J_{G,\text{thr}} = 0.75$ (Figure 4.14b, 4.15b, 4.16b and 4.17b);
- LH (Figure 4.14c, 4.15c, 4.16c and 4.17c);
- $4^2$ full-factorial DoE (Figure 4.14d, 4.15d, 4.16d and 4.17d).

Both maps are built at every iteration in order to calculate the optimal and explorative experiments as in Eq. (4.3). Here, only the results obtained in two iterations are shown for the sake of conciseness. This include results after 4 calibration experiments (Figures 4.14-4.16), to assess model prediction variance reduction and information gain after adding only one designed experiment besides the 3 preliminary ones; and results after 15 calibration experiments (Figure 4.15 and 4.17), being the first iteration where G-map eMBDoE is able to reduce completely

model prediction variance in the entire design space (darkest blue in the whole design space, as shown in Figure 4.15).

After calibrating the model with the data obtained from the fourth experiment (Figure 4.14), the G-map is characterised by high G-optimality values with $u_2$ between 30 and 35 g/L for MBDoE (Figure 4.14a) and eMBDoE (Figure 4.14b) and with $u_2$ between 15 and 35 g/L in case of LH (Figure 4.14c) and factorial DoE (Figure 14d). Therefore, designs that take into account information content, such as MBDoE and eMBDoE, provide a better performance at the first iteration with respect to completely explorative designs, such as LH and factorial DoE.

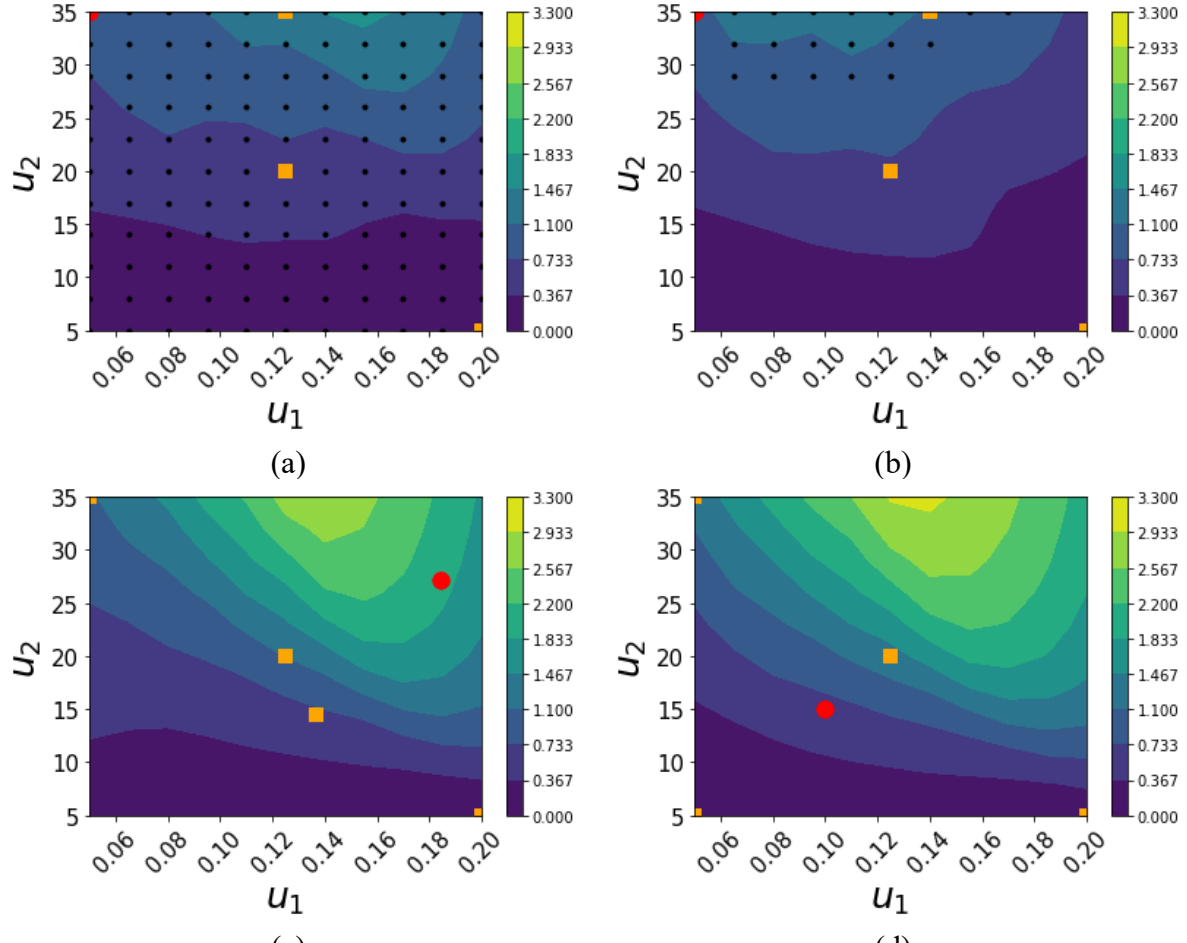



**Figure 4.14.** *G-maps generated after 4 calibration experiments. Four methods are compared: (a) MBDoE; (b) G-map eMBDoE with $J_{G,thr}$=0.75; (c) LH; (d) $4^2$ full factorial DoE. Orange squares indicate already measured data (namely, data used to calibrate the model); black dots indicate candidate design points; the red point indicates the experiment designed at the current iteration.*

When the number of experiments used in calibration increases to 15, G-map eMBDoE (Figure 4.15b) is the only scenario able to reduce completely model prediction variance in the entire design space, confirming the results obtained with scalar indices of G-optimality (Figure 4.13).

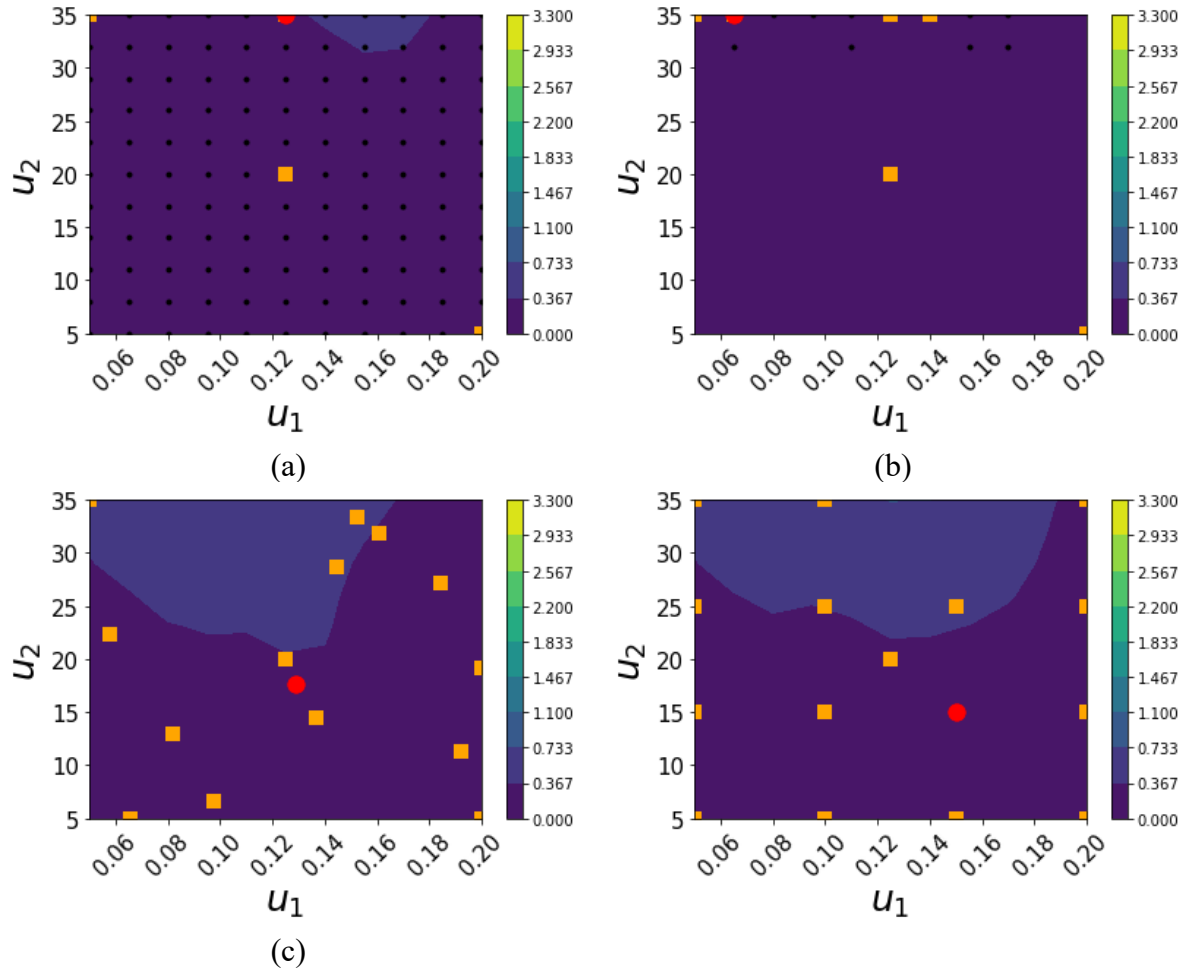



**Figure 4.15.** *G-maps generated after 15 calibration experiments. Four methods are compared: (a) MBDoE; (b) G-map eMBDoE with $J_{G,thr}$=0.75; (c) LH; (d) $4^2$ full factorial DoE. Orange squares indicate data already used to calibrate the model; black dots indicate candidate design points; the red point indicates the experiment designed at the current iteration.*

By analysing the distribution of information content throughout the design space with four calibration experiments, factorial DoE (Figure 4.16d) provides the smallest amount of information, while eMBDoE (Figure 4.16b) and MBDoE (Figure 4.16a) guarantee higher information levels. Unexpectedly, LH generates the highest values of information (Figure 4.16c); this may be caused by the initialisation of the MBDoE and eMBDoE procedure with parameters estimates that are still quite far from the true values. In fact, with 15 experiments, the following rank of information content is found: MBDoE (Figure 4.17a) > eMBDoE (Figure 4.17b) > LH (Figure 4.17c) > factorial DoE (Figure 4.17d). Therefore, conventional MBDoE generates the maximum amount of information content as expected, while eMBDoE provides an intermediate result between optimal (MBDoE) and explorative designs (LH, factorial DoE). Maps generated in the last experiment design iteration are shown in Appendix D.

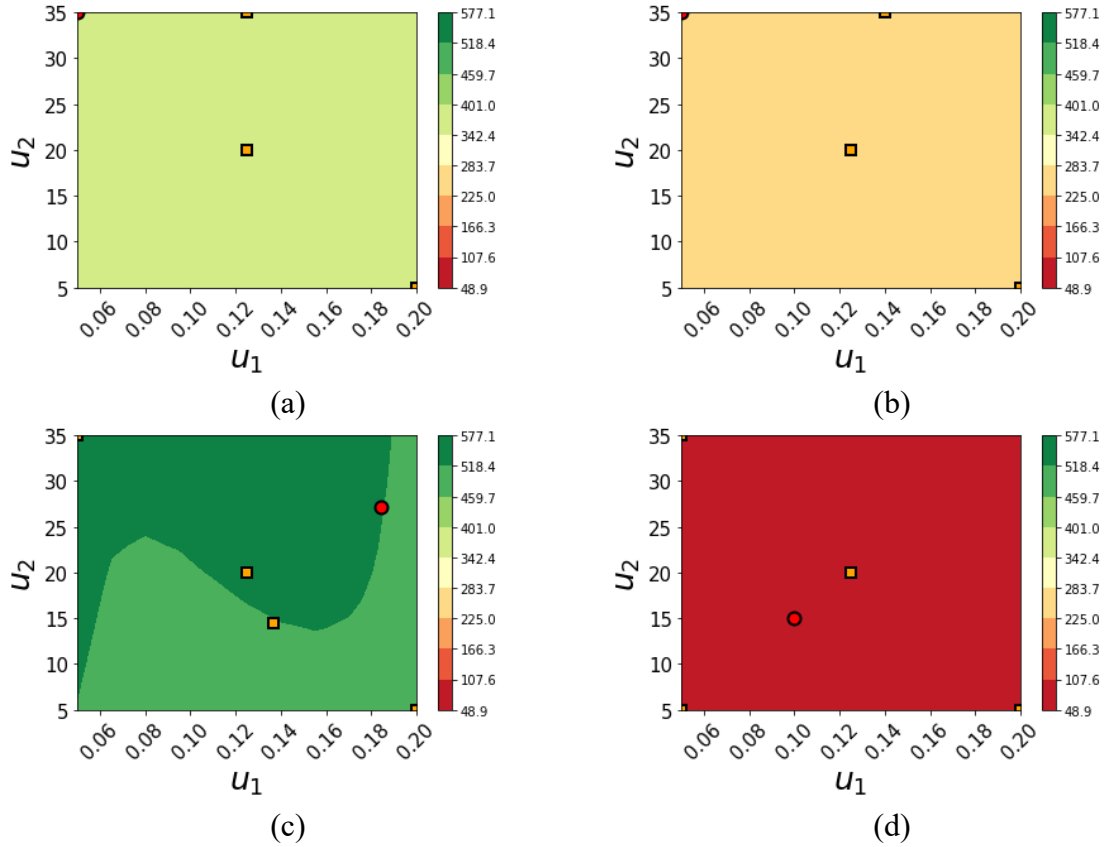



**Figure 4.17.** *H-maps after 4 experiments: (a) MBDoE; (b) G-map eMBDoE ($J_{G,thr}$=0.75); (c) LH; (d) DoE*

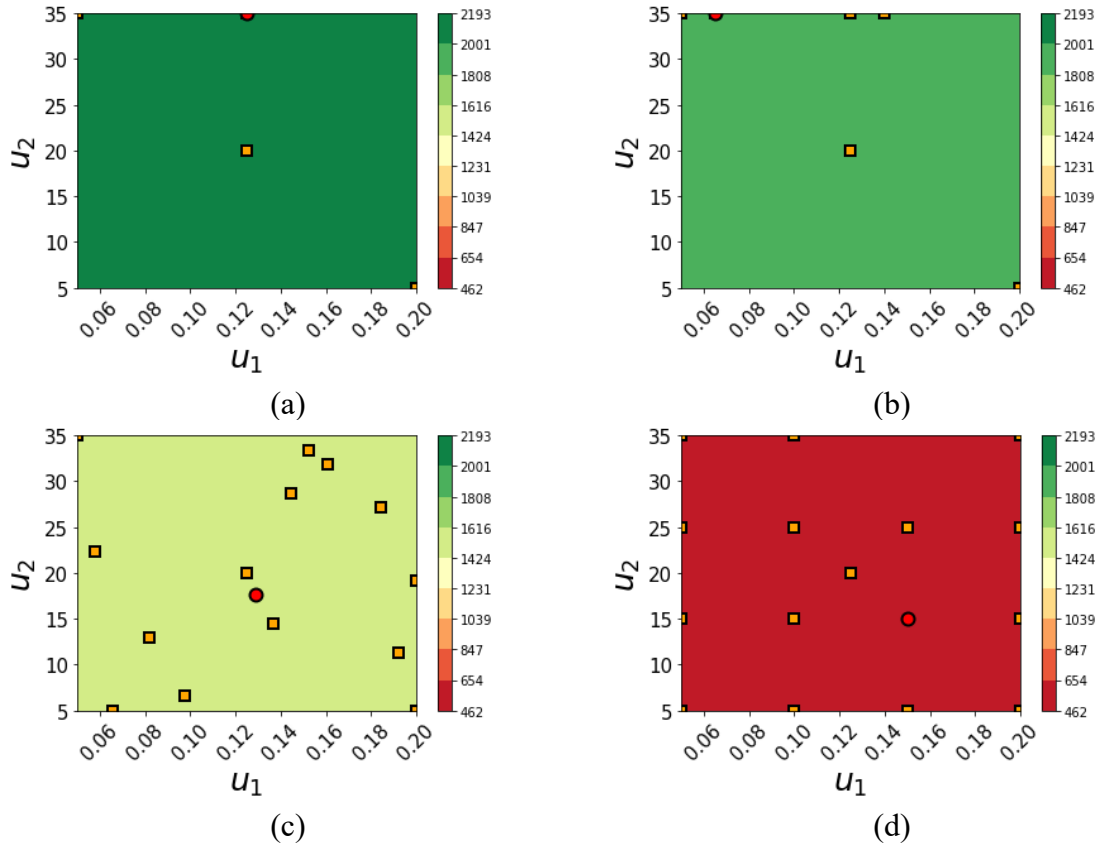



**Figure 4.17.** *H-maps after 15 experiments: (a) MBDoE; (b) G-map eMBDoE ($J_{G,thr}$=0.75); (c) LH; (d) DoE.*

The computational time required to discretise the design space, characterise it in terms of information content and model prediction variance (i.e., to build H-maps and G-maps) and to calculate the optimal and/or explorative experiment is approximately 0.65 seconds for both MBDoE and eMBDoE.

This suggests that building of both G-maps and H-maps does not lead to an excessive computational burden even though the model complexity is increased significantly with respect to Model 1.

## 4.4 Conclusions and future work

A novel exploratory MBDoE has been proposed in this paper with the objective of precisely estimating model parameters and minimising model prediction uncertainty in the whole domain of model utilisation, with the minimum experimental effort. The proposed method is based on a mapping of model prediction variance evaluated across the entire design space through a scalar measure of G-optimality (G-map), which is calculated based on the evaluation of Fisher information matrix and requires knowledge on model structure (set of equations) and estimated parameter values. Experimental conditions within the design space are then selected from the candidate design points, which are associated to a G-optimality value higher than a user-defined threshold. Therefore, the exploration of the space is pushed towards regions that are still not well-described by the model. The G-map based explorative MBDoE is compared against purely explorative methods, i.e. full factorial DoE and LH, and against a purely information-based exploitative method, i.e. MBDoE.

These experimental design techniques are applied to two models of increasing complexity: Model 1, an algebraic model with one response and two control variables; Model 2, a nonlinear differential equation model of baker's yeast fermentation in a fed-batch reactor with two response variables measured at three sampling points and two control variables. In both cases, the constraint on G-optimality enables an increase of space exploration: the larger the threshold on G-optimality, the more the designed experiments depart from the ones selected by MBDoE.

Results from both case studies suggest that the trade-off between space exploration and information maximisation achieved by G-map eMBDoE allows to minimise the number of experiments required to precisely estimate model parameters and to minimise model prediction variance in the whole design space. In fact, with Model 1, 14 calibration experiments designed through G-map eMBDoE with $J_{G,\mathrm{thr}}$=0.75 allow to estimate all model parameters with statistical precision and to reduce G-optimality to the minimum values among all considered

scenarios. As regards Model 2, the experimental burden is minimised by eMBDoE with $J_{G,\mathrm{thr}}$=0.75: indeed, 15 calibration experiments are enough to precisely estimate model parameters and reduce G-optimality to a minimum value throughout the design space. Different simulations of the systems under study suggest that a good trade-off is found with a G-optimality threshold of 0.65-0.85. Future work is focused on further validation of G-map eMBDoE through data generated by a physical system and on the development of a systematic method to determine the best G-optimality threshold.

Finally, the additional step of candidate design points selection required by G-map eMBDoE leads to a negligible increase in computational time with respect to the state of the art MBDoE. Moreover, the time required to design a single experiment increases with model complexity, namely with the increase of the number of control variables, response variables and/or sampling points, but it is still negligible in the case studies analysed in this work: 0.12 seconds in case of Model 1 and 0.65 seconds with Model 2. Thanks to the satisfactory performance of the G-map eMBDoE with the two simulated processes and the limited computational burden, this method will be implemented in automated platforms for online model identification, in order to integrate and test the proposed experimental design method in an actual experimental system (see Chapter 5).

# Chapter 5
# Exploratory optimal experimental design for the identification of total methane oxidation kinetics in automated microreactor platforms[3]

In this Chapter, an explorative model-based design of experiments method based on G-optimality mapping, *G-map eMBDoE*, is tested in an automated flow micropacked bed catalytic reactor platform for total methane oxidation on palladium-based catalyst. Differently from state-of-the-art MBDoE which selects the most informative experimental condition (namely, the one minimising parameters uncertainty), the proposed G-map eMBDoE seeks a trade-off between information maximisation and space exploration. This is done by restricting information maximisation to a subset of design points that satisfy a threshold on model prediction variance. A novel method to achieve the best trade-off is presented and tested in this work, but it can be easily extended to any process of interest. Thus, three scenarios are compared: purely optimal designs, namely MBDoE; explorative MBDoE with a user-defined threshold on model prediction variance; explorative MBDoE with the novel method to select the threshold.

## 5.1 Introduction

The development of automated platforms for chemical and biological applications is gaining attention both in academia and industry (Barz et al., 2022). These platforms are particularly appealing because of the possibility of assigning repetitive tasks to automated components, e.g. liquid handling robots, thus improving profitability, speed and reliability of a process and giving experimenters time for more demanding tasks (Cherkasov et al., 2018; Waldron et al., 2020).

[3] Cenci, F., Pankajakshan, A., Bawa, S. G., Gavriilidis, A., Facco, P. and Galvanin, F.. Explorative optimal experimental design for the identification of total methane oxidation kinetics in automated microreactor platforms [in preparation].

Both batch and flow reactors can be used in these automated systems. On one hand, batch reactors are suitable to kinetic studies thanks to the possibility of sampling at different time intervals, while flow reactors are typically operated at steady-state, therefore more time is needed to sequentially measure different experimental conditions (Waldron et al., 2020). Furthermore, setting up a fully automated flow chemistry platform is more challenging than a batch one (Cherkasov et al., 2018). However, flow reactors are promising since they can improve heat and mass transfer, ensure higher safety, reduce reagents consumption and increase productivity and product selectivity (Waldron et al., 2020; Cherkasov et al., 2018). Due to these advantages, recent research has tried to overcome some of the limitations of flow systems. For instance, if a rapid experimentation is required, transient experiments can be performed instead of waiting until steady-state (Hone et al., 2017; Waldron et al., 2019; Waldron et al., 2020; Taylor et al., 2021). Moreover, Cherkasov et al. (2018) developed an open access platform named OpenFlowChem to speed up the creation and/or modification of new automated flow systems, encouraging the spread of automated flow platforms.

In automated flow platforms, LabView is one of the preferred software for hardware/software integration and to integrate customised Python or MATLAB algorithms (Cherkasov et al., 2018). In turn, different algorithms can be implemented based on the purpose of experimentation; in literature, two major goals are found: 1) reaction systems self-optimisation; 2) development of kinetic models. Self-optimising reactors are referred to as "black-box" optimisation systems because they aim at improving a specific key performance indicator, like yield or conversion, without having a deep knowledge of the reaction system. In this context, factorial Design of Experiments (DoE; Montgomery, 2013) can be used to screen several input factors and understand their effect on the response variable; experimental results are then used to build a response surface (namely, a linear regression model that acts as surrogate model) useful for operating the platform around optimal conditions (Reizman et al., 2016; Reizman and Jensen, 2016). Black-box optimisation systems can be used to scale up to a plug flow reactor (PFR), but they may be unsatisfactory when the objective of the optimisation changes (e.g., from yield to purity) or when the process is scaled up to a reactor different from a PFR (McMullen and Jensen, 2011).

On the other hand, the availability of a kinetic model reduces scale up risks and costs and allows to simulate different optimisation scenarios and/or equipment configurations (Hone et al., 2017; Waldron et al., 2019; Reizman and Jensen, 2016). Kinetic models can be identified in automated chemical platforms with minimal consumption of time and resources through the

implementation of MBDoE (Espie and Macchietto, 1989), for example in form of Python or MATLAB codes integrated in LabView (Waldron et al., 2019; Pankajakshan et al., 2023). For instance, MBDoE can be used to calculate the optimal experimental conditions to compare different candidate structures and select the one that better fits the experimental data or to maximise parameters precision (see Chapter 2). This leads to a minimisation of the number of experiments needed to reduce parameters uncertainty. For this reason, MBDoE for parameters precision has been successfully implemented in automated chemical platforms. For instance, McMullen and Jensen (2011) applied the D-optimal criterion to maximise the parameters precision of the Diels-Alder reaction of isoprene and maleic anhydride in DMF. Reizman and Jensen (2012) studied the series- parallel nucleophilic aromatic substitution of morpholine onto 2,4-dichloropyrimidine: first, they optimised the yield of a specific product and isolated intermediates; then, they minimised parameters uncertainty by 50% by means of a D-optimal MBDoE. Moreover, Echtermeyer et al. (2017) streamlined the identification of the Pd catalysed C-H activation model through MBDoE and they used that mechanistic model to train a surrogated one for optimisation purposes. Instead, Waldron et al. (2019) studied the esterification of benzoic acid with ethanol in a heterogeneous catalyst through a multi-step procedure made of: initial factorial DoE screening at steady-state; model discrimination through identifiability analysis and MBDoE; parameters estimation through MBDoE. The same reaction in a sulfuric acid homogeneous catalyst was investigated by means of three different designs of experiments by Waldron et al. (2019): factorial DoE at steady state; MBDoE at steady state; user-defined experiments at transient conditions. The possibility to maximise parameters precision while reducing time and resources using measurement of ramp transient experiments was confirmed by Waldron et al. (2020). Finally, Pankajakshan et al. (2019) equipped an automated platform for the esterification of benzoic acid and ethanol with a Python code that automatically calculates the experimental conditions with the best trade-off between maximisation of parameters precision and minimisation of experimentation costs. More details on these and others applications in chemical and biochemical platforms can be found in Barz et al. (2022).

To the author's knowledge, optimal experimental designs in automated flow platforms have been implemented mainly with the aim of discriminating among candidate kinetic models, of estimating precise model parameters and/or reducing the resources employed. However, little efforts have been made to precisely estimate the parameters of a kinetic model with minimum model prediction variance across the whole design space, while keeping the experimental

burden at minimum. In fact, most works dealing with minimisation of model prediction variance focus on the reliability of the model in the restricted region of the optimal experimental conditions. For instance, Reizman et al. (2016) analysed and optimised Pd-catalyzed Suzuki–Miyaura cross-coupling reactions by means of a DoE-based algorithm and considering both continuous (i.e., temperature, time, and loading) and discrete (i.e., precatalysts and ligand) variables. After initialising the procedure with preliminary experiments, response surface models were built for every precatalyst, thus allowing a comparison among them. Then, experiments were designed through a G-optimal criterion with the aim of minimising model prediction uncertainty in correspondence of the predicted optima for every precatalyst. While model prediction uncertainty was minimised in these conditions, unsuccessful precatalysts were eliminated from the list of candidates. Therefore, the experiments gradually focused on regions of optimal reaction performance and on the most promising precatalysts. As expected, this lead to a greater model prediction uncertainty in the regions scarcely selected by the DoE-based algorithm. However, the development of a kinetic model with a minimised model prediction uncertainty across the entire design space improves model reliability in a wider set of experimental conditions, thus increasing the usability of the model for future applications, such as reactor design for technology transfer and scale-up, reaction system optimisation and model-based control.

In this work, an explorative MBDoE method is implemented in an automated flow micropacked bed catalytic reactor platform for the total methane oxidation reaction on a 5% $Pd/Al_2O_3$ catalyst, which is described by a Mars-van Krevelen mechanism (Pankajakshan et al., 2023). The explorative MBDoE method aims at minimising model prediction variance across the whole design space, while ensuring maximum parameters precision and reduced experimental burden. This is done through the use of G-optimality maps (*G-map eMBDoE*) as presented in Chapter 4 (Cenci et al., 2003): a trade-off between information maximisation and space exploration is found by the combined use of maps of G-optimality and maps of FIM-based scalar measures (namely, maps of model prediction uncertainty and maps of data information content, respectively). While Chapter 4 demonstrated the advantages of the G-map eMBDoE over conventional designs (DoE and standard MBDoE) by means of simulated case studies, here it is shown that this methodology can be beneficial to improve the efficiency of kinetic model identification in automated chemical platforms. Moreover, the original G-map eMBDoE optimal design problem is here reformulated to better handle the trade-off between experimental design space exploration and information maximisation for the system under study.

This Chapter is structured as follows: Section 5.2 describes experimental and modelling strategies, with details on: reaction and platform components; model calibration and results analysis; G-map eMBDoE description and practical implementation. Results in terms of experimental design, parameters precision and model prediction variance and accuracy are shown and discussed in Section 5.3, while conclusions are drawn in Section 5.4.

## 5.2 Materials and methods

A comparison between G-map eMBDoE and state-of-the-art MBDoE is performed in order to assess the advantages of a trade-off between space exploration and information maximisation over an approach (MBDoE) only targeting information optimisation. Figure 5.1 shows the experimental procedure adopted in this work:

- Step 1: preliminary experiments are designed by a factorial DoE to initialise the procedure;
- Step 2: the designed experiments are executed by the automated chemical platform;
- Step 3: the kinetic model is calibrated and analysed;
- Step 4: a user-defined criterion is assessed in order to decide whether to stop the experimental campaign or not. In this work, the experimental campaign is terminated when the budget of $N_e$=35 experiments overall is reached;
- Step 5: if new experiments must be collected, they are designed though G-map eMBDoE (or MBDoE if $J_{G,\mathrm{thr}}$=1).

In the following Sections, Steps 2,3 and 5 are explained in detail.

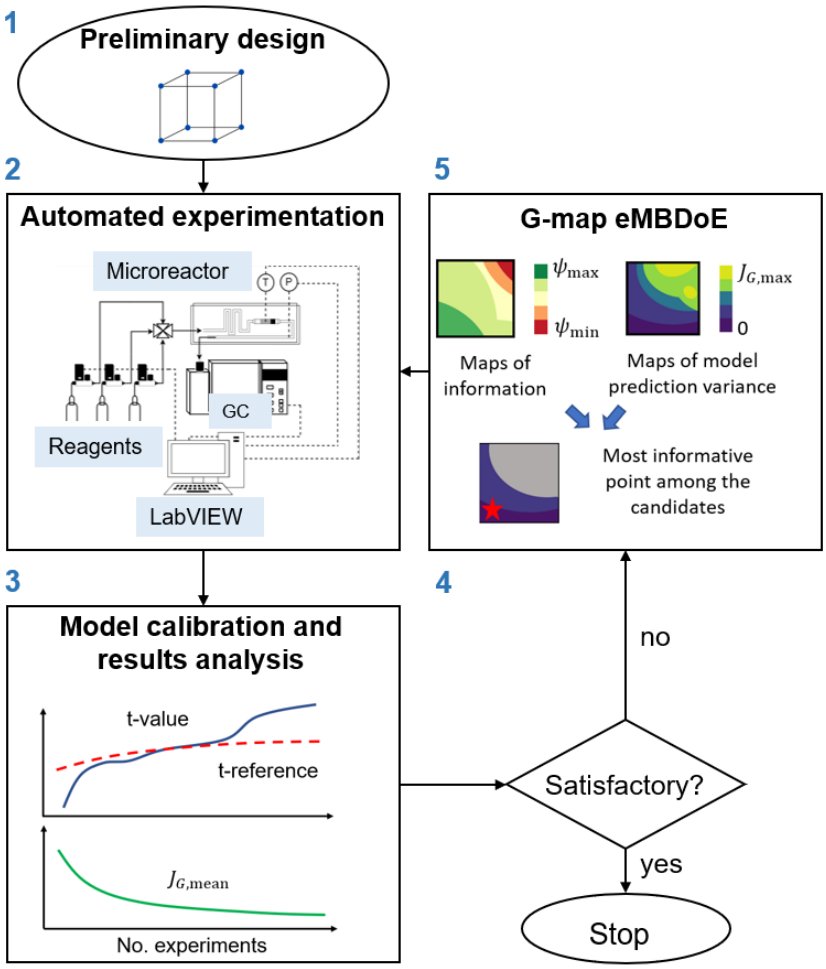



**Figure 5.1.** *Schematic representation of the proposed G-map eMBDoE-driven procedure for optimal experimental design in automated platforms. The representation of the automated platform is adapted from Bawa et al. (2022).*

## *5.2.1 Automated experimentation*

The main components of the platform (as shown in Figure 5.1) include: (*i*) mass flow controllers (Brooks, 5850TR; Hatfield, USA), which serve as the delivery system for the inlet streams; (*ii*) a packed bed microreactor made of silicon-glass; (*iii*) an online gas chromatography (GC) as measurement system; (*iv*) Lab View 2018, to interface between the hardware and Python programme for smooth integration process. The reaction is catalytic complete oxidation of methane over 5% $Pd/Al_2O_3$ catalyst. The average size of the catalysed reactor is 69 µm and 10 mg of the catalyst is used for the catalytic reaction. The composition of the stream is nitrogen as an internal standard, oxygen and methane as the main reactants, while helium is part of the methane stream (5% methane in helium). The silicon-glass microreactor is fabricated using photolithography and deep reactive ion etching (DRIE) process. The silicon wafer, inscribed with the desired reactor pattern is sealed with glass cover using anodic bonding process

performed at 400 °C and 500 V. The reaction channel dimensions are 2 mm and 420 μm in width and depth. A retainer for the catalyst bed is positioned at the end of the reaction channel. The pressure at both inlet and outlet of the micro packed bed are monitored by pressure sensors (Honeywell, 40PC, 100 psig), whereas the outlet pressure is held constant with the help of pressure controller (Brooks, 5866). The online GC (Agilent, 7890A) is equipped with sampling valve, sampling loop, GS-Carbon PLOT (Agilent) and HP-PLOT molecular sieve (Agilent) columns and thermal conductivity detector (TCD) and is used for effective online analysis of the reactor effluent. The computer contains the software that controls the GC as well as the Lab View and Python codes.

### *5.2.2 Model calibration and results analysis*

Once experiments are generated with the experimental setup described in Subsection 5.2.1, model parameters $\widehat{\boldsymbol{\theta}}$ are estimated by maximising Maximum Likelihood (Chapter 2); afterwards, FIM $\mathbf{H}_{\widehat{\boldsymbol{\theta}}}$ and parameters variance-covariance matrix $\mathbf{V}_{\widehat{\boldsymbol{\theta}}}$ can be recalculated as explained in Chapter 2. While the experimentation progresses, the different design methods employed are compared in order to assess the level of space exploration of eMBDoE with respect to state-of-the-art MBDoE (Subsection 5.2.2.1). Moreover, the performance of the calibrated model at the current iteration is assessed in terms of: parameters precision; model prediction variance; model prediction accuracy (Subsections 5.2.2.2-5.2.2.4).

#### **5.2.2.1 Designed experiments**

One of the objectives of G-map eMBDoE is to enhance space exploration with respect to conventional MBDoE. Therefore, the results of eMBDoE and MBDoE designs are compared in terms of calculated values of $\boldsymbol{\varphi}_{\text{opt}}$ at every iteration. To qualitatively compare the level of space exploration, profiles of a given control variable calculated by different methods are plotted within the same figure. Figure 5.2 provides and illustrative example where the profile of eMBDoE with $J_{G,\text{thr},1}$ is close to the profile of MBDoE, meaning that $J_{G,\text{thr},1}$ does not considerably improve space exploration with respect to MBDoE, while eMBDoE with $J_{G,\text{thr},2}$ calculates values of the $i$-th control variable that are far from the ones of MBDoE, thus enhancing space exploration.

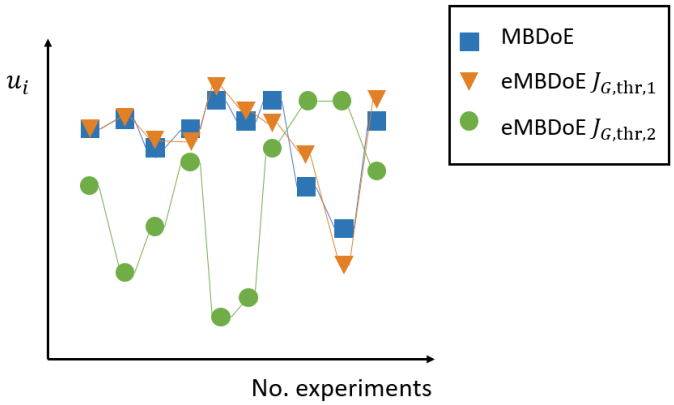



**Figure 5.2.** *Illustration of a possible set of profiles of the generic control variable $u_i$. One profile is made of the values of $u_i$ calculated at every iteration by the same MBDoE or eMBDoE method. Different profiles are obtained with different design methods (MBDoE vs G-map eMBDoE with different thresholds, namely $J_{G,thr,1}$ and $J_{G,thr,2}$).*

To better evaluate the departure of a given G-map eMBDoE design from a conventional MBDoE one, the following index is calculated at every iteration of the sequential procedure in Figure 5.1:

$$d_{ij} = \frac{|u_{i,j} - u_{i,\text{MBDoE}}|}{|u_{i,\text{MBDoE}}|}, \qquad i = 1, \dots, N_u \; ; j = J_{G,\text{thr}} \tag{5.1}$$

where $i$ refers to the specific control variable, $j$ refers to the threshold $J_{G,\text{thr}}$ used by the G-map eMBDoE method, $u_{i,\text{MBDoE}}$ indicates the value of the $i$-th control variable designed through MBDoE. Considering the example in Figure 5.2, $d_{ij}$ of $J_{G,\text{thr},2}$ is expected to be higher than the one of $J_{G,\text{thr},1}$ at every iteration.

### 5.2.2.2 Parameters precision

Parameters precision is assessed by means of 100(1-α) % confidence intervals and $t$-tests, as explained in Chapter 2.

### 5.2.2.3 Model prediction variance using G-maps

Every point of the discretisation of the design space is characterised in terms of G-optimality (namely, $J_G$ given by the sum of model prediction variance $V_y$ of every response at every time point), obtaining the G-maps (see Chapter 4 for more details). G-maps are then used to compare graphically and quantitatively the performance of different design methods in terms of reduction of model prediction variance in the whole design space.

The visualisation of G-maps has the advantage of showing the regions where model prediction variance can still be improved and of highlighting in a graphical way the reduction of model

prediction variance throughout the experimentation. An example of G-map is in Figure 5.3. G-maps are coloured such as high G-optimality values (i.e., high model prediction variance; worst performance) are found in yellow regions, while low G-optimality values (i.e., low model prediction variance; best performance) are in dark blue regions.


**Figure 5.3**. *Illustration of a G-map.*

A quantitative comparison is facilitated by the calculation of scalar measures of G-optimality $J_G$ . Therefore, minimum ($J_{G,\min}$), mean ($J_{G,\text{mean}}$) and maximum ($J_{G,\max}$) values of $J_G$ calculated across the entire design space are considered (as explained in Chapter 4).

#### 5.2.2.4 Model prediction accuracy

Model prediction accuracy, namely the difference between predicted responses and experimentally measured values, is evaluated in terms of absolute error $AE_i$ calculated as:

$$AE_i = |\hat{y}_i - y_i| \tag{5.2}$$

where $\hat{y}_i$ is the predicted value for the $i$-th response variable and $y_i$ is the corresponding measured value. To summarise such results through a scalar index at every iteration, the mean of $AE_i$, indicated as $\mu_{AE_i}$, is calculated for $i = \mathrm{CH_4}, \mathrm{O_2}, \mathrm{CO_2}$.

### 5.2.3 G-map eMBDoE

A possibility to obtain explorative MBDoE (*eMBDoE*) experiments is to select of the most informative experiment $\boldsymbol{\varphi}_{\text{opt}}$ among a subset $\boldsymbol{\varphi}_{\text{cand}}$ of candidate design points:

$$\boldsymbol{\varphi}_{\text{opt}} = \arg\min_{\boldsymbol{\varphi}_{\text{cand}}} \psi(\mathbf{V}_{\hat{\boldsymbol{\theta}}}) \tag{5.3}$$

where $\psi(\mathbf{V}_{\hat{\boldsymbol{\theta}}})$ can be one of the classical criteria described by Pukelsheim (1993). In the eMBDoE method based on mapping of G-optimality, named *G-map eMBDoE*, the candidate design points $\boldsymbol{\varphi}_{\text{cand}}$ satisfy a predefined condition on G-optimality $\mathbf{V}_y$. Since G-optimality $\mathbf{V}_y$ is calculated for all $N_y$ response variables at all $N_{\text{sp},i}$ sampling points, it is summarised by the

scalar $J_G$ calculated as the sum of every contribution ($J_G = \sum_{j=1}^{N_y}\sum_{i=1}^{N_{sp,i}} \mathbf{V}_y(\hat{\boldsymbol{\theta}}, \boldsymbol{\varphi})\big|_{j,i}$, as explained in Chapter 4). The calculation of information content $\psi(\mathbf{V}_{\hat{\boldsymbol{\theta}}})$ and model prediction variance $J_G$ for all experimental conditions in the design space leads to the so-called *H-maps* and *G-maps*, respectively. They can be used to solve Eq. (5.3) through a grid-search approach and to graphically visualise the results.

The G-optimality constraint that must be satisfied by candidate design points $\boldsymbol{\varphi}_{\text{cand}}$ of Eq. (5.3) should be selected in such a way as eMBDoE experiments are more explorative than MBDoE ones, but without losing too much information. In order to select the most suitable constraint for the system under study, preliminary in silico experiments have been performed by exploiting prior knowledge on model structure and plausible initial parameters values retrieved from Bawa et al. (2022) and Pankajakshan et al. (2023). Those simulations suggest that space exploration is enhanced by the constraint:

$$J_G \leq J_{G,\text{thr}} J_{G,\text{max}} \tag{5.4}$$

where $J_{G,\text{max}}$ is the maximum value of $J_G$ in the whole design space, while $J_{G,\text{thr}}$ is a user-defined fraction between 0 and 1. If $J_{G,\text{thr}}$ =1, the entire design space is retained for information maximisation, therefore eMBDoE corresponds to conventional MBDoE. Instead, the closer $J_{G,\text{thr}}$ is to 0, the smaller the number of candidate design points. Considering Figure 4.2 of Chapter 4, the constraint of Eq. (5.4) is used to define the candidate design points indicated in blue, among which the most informative experiment is selected.

The level of space exploration can be handled by the selection of the G-optimality threshold $J_{G,\text{thr}}$. This value is case-dependent and it impacts on the final outcome of G-map eMBDoE. Due to its importance, two different methods to select $J_{G,\text{thr}}$ are tested with the experimental data generated by the platform:

1. selection of a threshold $J_{G,\text{thr,prior}}$ based on prior knowledge on the system (namely, model equations and preliminary values for $\hat{\boldsymbol{\theta}}$);
2. determination of the threshold $J_{G,\text{thr,meas}}$ using the first set of experiments measured in the platform. The threshold value is selected in such a way as the first optimal design obtained through G-map eMBDoE is different from the one obtained through MBDoE.

With the former method, the following procedure is adopted: (*i*) using prior knowledge on the system, different experimental campaigns are simulated considering multiple $J_{G,\text{thr}}$ values (as was done in Chapter 4 with Model 1 and 2); (*ii*) the range of $J_{G,\text{thr}}$ providing the best results in terms of $t$-tests and reduction of model prediction variance is determined; (*iii*) a threshold

$J_{G,\text{thr,prior}}$ is selected within this range. In this work, preliminary simulations suggest that the range between 0.65-0.75 is suitable; therefore, a value within this range is selected for actual experiments in the platform (see Section 5.3).

The latter method to select the threshold is implemented following a new procedure illustrated in Figure 5.3:

- step 1: the set of preliminary experiments measured in the platform is used to estimate model parameters $\widehat{\boldsymbol{\theta}}$;
- step 2: $J_{G,\text{thr}} = 1$ is set; state-of-the-art MBDoE is performed obtaining the optimal experimental condition $\boldsymbol{\varphi}^*_{\text{opt}}$;
- step 3: $J_{G,\text{thr}} = J_{G,\text{thr}} - 0.05$ is set; G-map eMBDoE is performed obtaining the optimal and explorative condition $\boldsymbol{\varphi}_{\text{opt}}$ by solving Eq. (5.3);
- step 4: the design of MBDoE, $\boldsymbol{\varphi}^*_{\text{opt}}$, and the one of G-map eMBDoE, $\boldsymbol{\varphi}_{\text{opt}}$, are compared. If the two designs are different, the procedure stops and the threshold $J_{G,\text{thr}}$ of the current iteration is selected as $J_{G,\text{thr,meas}}$. If the two designs are equal, step 3 is iterated until a $\boldsymbol{\varphi}_{\text{opt}}$ such that $\boldsymbol{\varphi}_{\text{opt}} \neq \boldsymbol{\varphi}^*_{\text{opt}}$ is found.

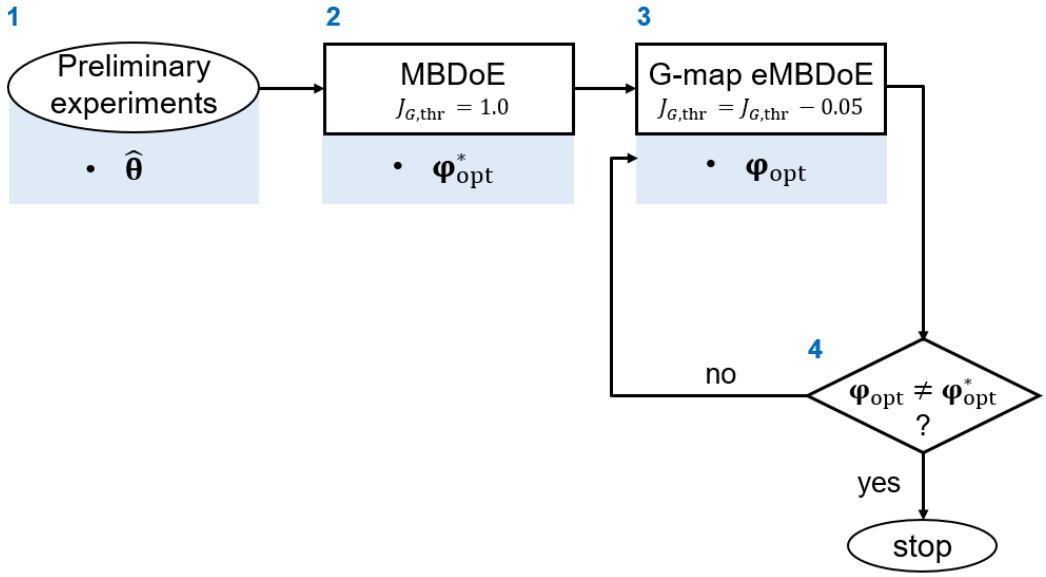



**Figure 5.4**. *Steps to select the G-optimality threshold $J_{G,thr}$ through a comparison between the optimal experimental conditions $\boldsymbol{\varphi}^*_{opt}$ and $\boldsymbol{\varphi}_{opt}$ calculated by, respectively, state-of-the-art MBDoE and G-map eMBDoE.*

When $J_{G,\text{thr,meas}}$ leading to $\boldsymbol{\varphi}_{\text{opt}} \neq \boldsymbol{\varphi}^*_{\text{opt}}$ is found, it is kept fixed until the end of the experimental campaign. The advantage of the proposed procedure to obtain $J_{G,\text{thr,meas}}$ is that it

does not require any prior knowledge on the system (differently from $J_{G,\mathrm{thr,prior}}$), therefore it can be calculated for completely new systems.

### 5.2.3.1 Generation of H-maps and G-maps with multiple control variables

To solve the optimisation in Eq. (5.3) through a grid-search approach, maps of G-optimality and of FIM-based can be built as described in Chapter 4. In fact, the procedure shown in Chapter 4 describes maps built with two control variables $u_1$ and $u_2$ only, but it can be extended to any number of control variables. In this work, $\boldsymbol{\varphi} = \mathbf{u} = [u_1, u_2, u_3, u_4]$ (see Section 5.3 for more details) therefore the steps to build G-maps and H-maps are as follows (Figure 5.5):

- initialise the procedure: (*i*) define the prior knowledge on the reaction system, namely model equations $\mathbf{f}$, current parameters values $\widehat{\boldsymbol{\theta}}$, variance-covariance matrix of measurement errors $\boldsymbol{\Sigma}_y$; control variables $\mathbf{u}$ and their admissible ranges; (*ii*) discretise the ranges of all control variables $\mathbf{u}$; (*iii*) naming $i_1$ and $i_2$ the indices of the current value within the discretisation of $u_1$ and $u_2$ ranges, respectively, set $i_1 = 1$ and $i_2 = 1$;
- Step 1: consider the $i_1$-th value of $u_1$ and keep it fixed;
- Step 2: consider the $i_2$-th value of $u_2$ and keep it fixed;
- Step 3: at the fixed values of $u_1$ and $u_2$, build a grid with every combination of values of $u_3$ and $u_4$;
- Step 4: for every point of the grid determined in Step 3, calculate the G-optimality $J_G$ and the FIM-based information content $\psi$, providing the G-maps and H-maps, respectively;
- Step 5: using the same $i_1$, set $i_2 = i_2 + 1$ and repeat Steps 2-4; repeat this step until all $u_2$ values (namely, all $i_2$) are used;
- Step 6: set $i_1 = i_1 + 1$ and repeat Steps 1-5; repeat this step until all $u_1$ values (namely, all $i_1$) are used.

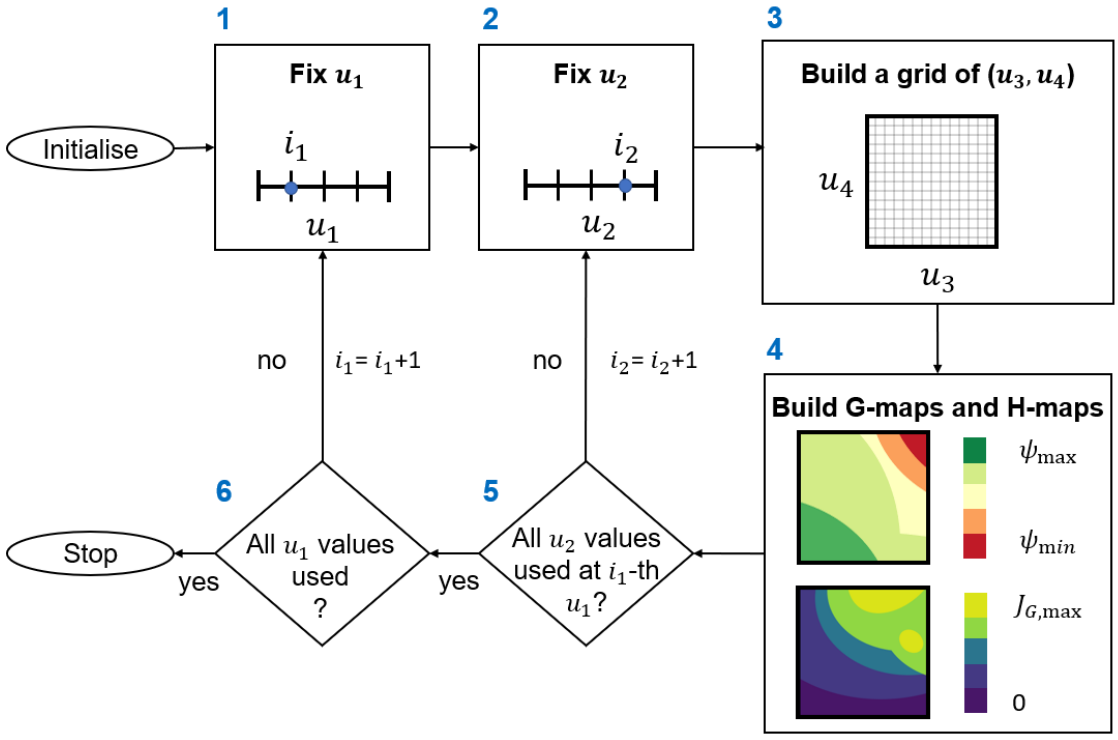



**Figure 5.5**. *Steps to build G-maps and H-maps with 4 control variables ($u_1$, …, $u_4$).*

Once the G-maps and H-maps have been built and recorded for all the possible values of the control variables **u**, the optimal and explorative experiment $\boldsymbol{\varphi}_{opt}$ is determined by solving Eq.(5.3). This method can be applied to any number of control variables.

### 5.2.4 Software implementation and experimental procedure

The reaction platform for total catalytic methane oxidation (Section 5.2.1) performs the desired experiments in an automated way. This means that the experimenter only needs to specify the design vector **φ**, while the platform sets the control variables accordingly and measures the corresponding output values. As regards the experimental design (i.e. the determination of the optimal design vector **φ**), this is implemented in a separate code in Python 3.9 (Rossum and Drake, 2009). To perform the G-map eMBDoE experiments presented in this work and discussed in Section 5.3, experiment design and execution are operated sequentially as shown in Figure 5.1: first, the platform provides the results of the preliminary experiments; then, these are used by the experimenter to calibrate the model and design a new experiment; the new design is fed to the platform, whose input/output data are used by the Python code for model calibration until the maximum experimental budget is reached (or any other stop criterion is satisfied).

## 5.3 Results and discussion

The experimental setup described in Section 5.2 was first used by Bawa et al. (2023), who experimentally assessed consistency in the control of input variables and good reproducibility of the results. Moreover, a subset of plausible candidate models to describe the system was selected and these models were deemed practically identifiable thanks to a full-rank FIM, even though not all model parameters were estimated with sufficient statistical precision. Afterwards, Pankajakshan et al. (2023) implemented a method to automatically discriminate candidate models and to improve parameters precision for the most probable model. The most promising kinetic models were Langmuir-Hinshelwood (Hurtado et al.,2004; Specchia et al., 2010) and Mars–van Krevelen (Hurtado et al.,2004; Specchia et al., 2010). The former considers a surface reaction between adsorbed methane and dissociatively chemisorbed oxygen, while the latter describes the surface reaction between adsorbed molecular oxygen and gas phase methane and slow desorption of the reaction products (Hurtado et al.,2004; Specchia et al., 2010; Pankajakshan et al., 2023). At the end of the MBDoE campaign, the most probable model was Mars–van Krevelen, consequently this kinetic model structure is considered in this work. Parameters $\boldsymbol{\theta} = [\theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6]$ to be estimated are shown in Table 5.1; parameters lower and upper bounds are, respectively, $\theta_{\mathrm{LB},i}$=0 and $\theta_{\mathrm{UB},i}$=15 for $i$=1,…,6 (here, the total number of parameters is $N_\theta$=6). More details on kinetic modelling and model reparametrisation can be found in Appendix E.

**Table 5.1.** *Parameters of Mars–van Krevelen kinetic model.*

| Original parameters | Units of measure | Parameters to be estimated |
|---|---|---|
| $k_{1,\mathrm{ref}}$ | [mol bar$^{-1}$ g$^{-1}$ min$^{-1}$] | $\theta_1 = -\log(k_{1,\mathrm{ref}})$ |
| $E_{a,1}$ | [J mol$^{-1}$] | $\theta_2 = E_{a,1}/10^4$ |
| $k_{2,\mathrm{ref}}$ | [mol bar$^{-1}$ g$^{-1}$ min$^{-1}$] | $\theta_3 = -\log(k_{2,\mathrm{ref}})$ |
| $E_{a,2}$ | [J mol$^{-1}$] | $\theta_4 = E_{a,2}/10^4$ |
| $k_{3,\mathrm{ref}}$ | [mol bar$^{-1}$ g$^{-1}$ min$^{-1}$] | $\theta_5 = -\log(k_{3,\mathrm{ref}})$ |
| $E_{a,3}$ | [J mol$^{-1}$] | $\theta_6 = E_{a,3}/10^4$ |

The response variables predicted by the model are mole fractions of $CH_4$, $O_2$ and $CO_2$ in the stream exiting from the reactor. Their standard deviations $\boldsymbol{\sigma}_y$ were calculated by Bawa et al. (2023) as pooled standard deviations: $\boldsymbol{\sigma}_y$ =[0.00043, 0.00202, 0.0005] [mol mol$^{-1}$]. Their square values form the diagonal of the variance-covariance matrix of the measurement errors ($\boldsymbol{\Sigma}_y$) used for parameters estimation through Maximum Likelihood and for FIM calculation (Chapter 2).

Moreover, the control variables of the model are: reaction temperature [°C], flow rate of the feed [Nml min$^{-1}$], inlet methane mole fraction [mol mol$^{-1}$], oxygen to methane mole ration in the feed [mol mol$^{-1}$]. They are indicated as $u_1, u_2, u_3$ and $u_4$, respectively, and they form the design vector $\boldsymbol{\varphi}$ to be optimised by G-map eMBDoE. The settings used to solve the constrained optimisation problem (Eq. 5.3) are shown in Table 5.2.

**Table 5.2.** *Constraints of the design vector for G-map eMBDoE.*

| Control variable | Constraints |
|---|---|
| $u_1$: temperature [°C] | [254, 355.5] |
| $u_2$: flow rate of the feed [Nml min$^{-1}$] | [20.0,30.0] |
| $u_3$: inlet methane mole fraction [mol mol$^{-1}$] | [0.005, 0.025] |
| $u_4$: oxygen to methane mole ration in the feed [mol mol$^{-1}$] | [2.0, 4.0] |

The constraints in Table 5.2 are used to constrain the design variables when building G-maps and H-maps as described in Section 5.2.

After the design and execution of 12 preliminary experiments through factorial DoE, three different model-based design methods are compared:

- state-of-the-art MBDoE, i.e. G-map eMBDoE with $J_{G,\mathrm{thr}}$ =1.00;
- G-map eMBDoE with a threshold selected in the 0.65-0.75 range that was suggested by preliminary simulations of the system under study; thus, $J_{G,\mathrm{thr,prior}}$ =0.70 is chosen;
- G-map eMBDoE with the threshold determined by using experimental data as described in Section 5.2: $J_{G,\mathrm{thr,meas}}$ =0.60.

The computational time required to solve Eq. (5.3) through the grid-search approach implemented in Python 3.9 and using an Intel® Core™ i7-10875H CPU, @ 2.30 GHz processor with 64.0 GB RAM is about 85 seconds.

### *5.3.1 Designed experiments*

The design of input variables made by MBDoE and G-map eMBDoE is qualitatively compared in Figure 5.6, where profiles of each control variable $u_i$, $i$=1,..,4 are shown.

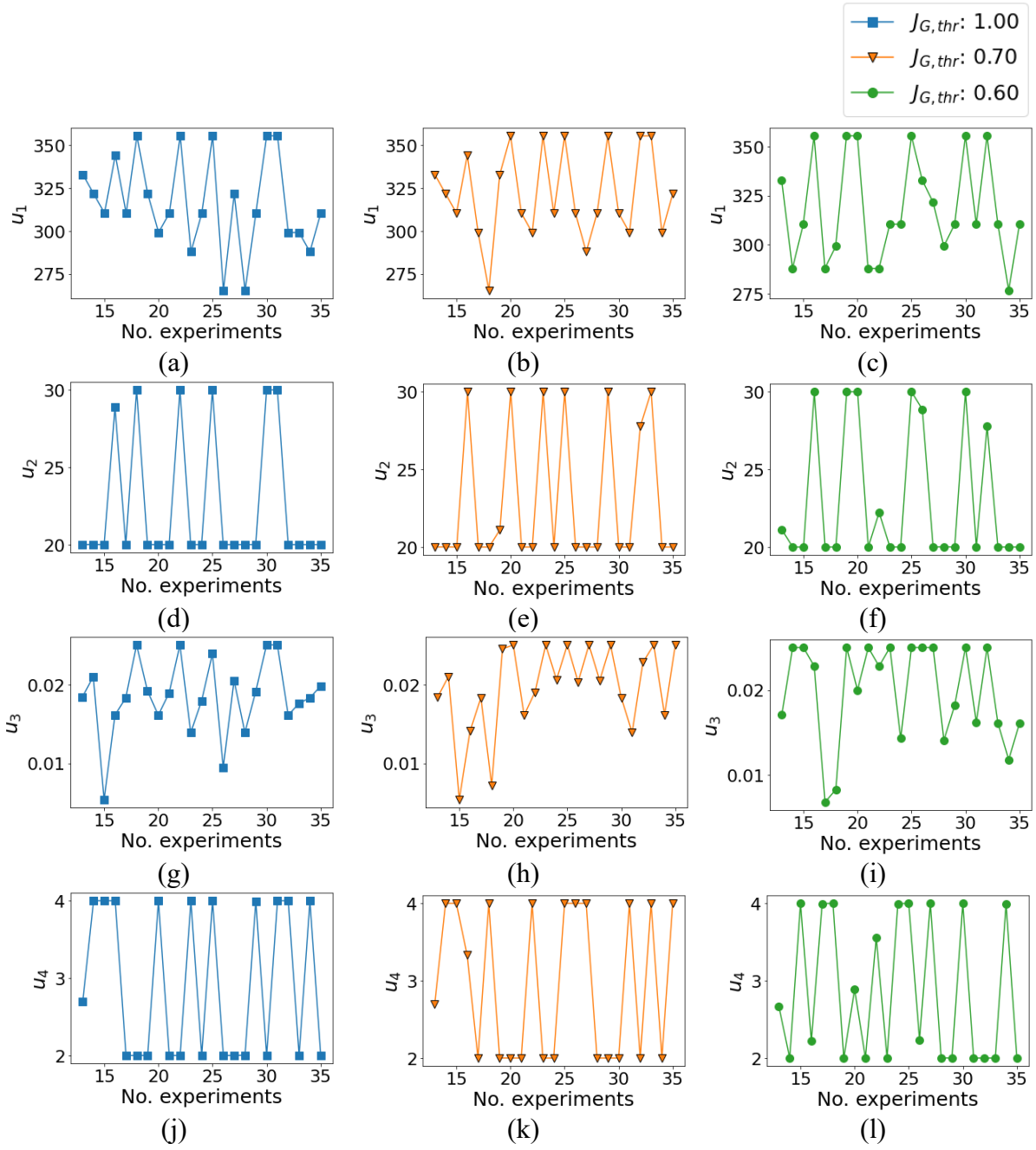



**Figure 5.6**. *Design of the four control variables $u_j$, j=1,…,4.*

The profiles of Figure 5.6 show that the major differences are found with $u_1$ (Figure 5.6a-c) and $u_4$ (Figure 5.6j-l), especially within the first 25 experiments: indeed, G-map eMBDoE with $J_{G,thr}$=0.60 explores more the values of $u_1$ below 300°C and the intermediates values of $u_4$ (namely, $u_4$ different from the two extreme values 2 and 4 molmol$^{-1}$) with respect to conventional MBDoE and G-map eMBDoE with $J_{G,thr}$=0.70.

To facilitate the comparison among the three methods, the difference between G-map eMBDoE and MBDoE designs are calculated as in Eq. (5.1). The results for $J_{G,thr}$=0.70, 0.60 calculated

et every iteration are summarised in Table 5.3 as number of occurrences of equal designs between conventional MBDoE and G-map eMBDoE.

**Table 5.3.** *Distance between the designs of the four control variables calculated by MBDoE and G-map eMBDoE with $J_{G,thr}$=0.70 and with $J_{G,thr}$=0.60. The number of iterations where a certain* $d_{i,j}$ *is equal to zero (*no. $d_{i,j} = 0$, *i=1,2,3,4 and j=0.70, 0.60) is calculated.*

| G-map eMBDoE | No. $\boldsymbol{d_{1,j} = 0}$ | No. $\boldsymbol{d_{2,j} = 0}$ | No. $\boldsymbol{d_{3,j} = 0}$ | No. $\boldsymbol{d_{4,j} = 0}$ | No. $\boldsymbol{d_{1,j} = d_{2,j} = d_{3,j} = d_{4,j} = 0}$ |
|---|---|---|---|---|---|
| $J_{G,\text{thr}}$=0.70 | 7 | 12 | 4 | 10 | 3 |
| $J_{G,\text{thr}}$=0.60 | 7 | 11 | 1 | 4 | 0 |

If the four control variables are considered singularly, the number of iterations in which the G-map eMBDoE and MBDoE values are equal is not negligible, neither with $J_{G,\text{thr}}$=0.70 neither with $J_{G,thr}$=0.60, even though the number of occurrencies is always smaller in the latter case. However, if all the four control variables are considered together, only $J_{G,\text{thr}}$=0.70 has some identical experiments with MBDoE: they are the first three experiments. The results of the following sections reveal that this is sufficient to get very different results in terms of parameters precision and model prediction variance.

### 5.3.2 Parameters precision and estimates

Figure 5.7 shows the results carried out at every iteration of the sequential procedure; the same $t$-values are explicitly shown in Appendix F. The most critical parameters to estimate are $\hat{\theta}_1$ and $\hat{\theta}_2$: state-of-the-art MBDoE requires 15 optimal experiments to precisely estimate $\hat{\theta}_1$ and 23 optimal experiments to estimate $\hat{\theta}_2$; instead, G-map eMBDoE with $J_{G,\text{thr}}$=0.70 requires 12 optimal experiments to estimate $\hat{\theta}_1$, while 23 eMBDoE experiments are not enough to estimate $\hat{\theta}_2$ precisely. The best performance is achieved through G-map eMBDoE with $J_{G,\text{thr}}$=0.60: using this approach, 5 optimal experiments are enough to obtain a statistically precise estimation of the full set of model parameters. This suggests that the increase of space exploration achieved through a threshold of $J_{G,\text{thr}}$=0.60 does not lead to a loss of information content for parameters estimation purposes.

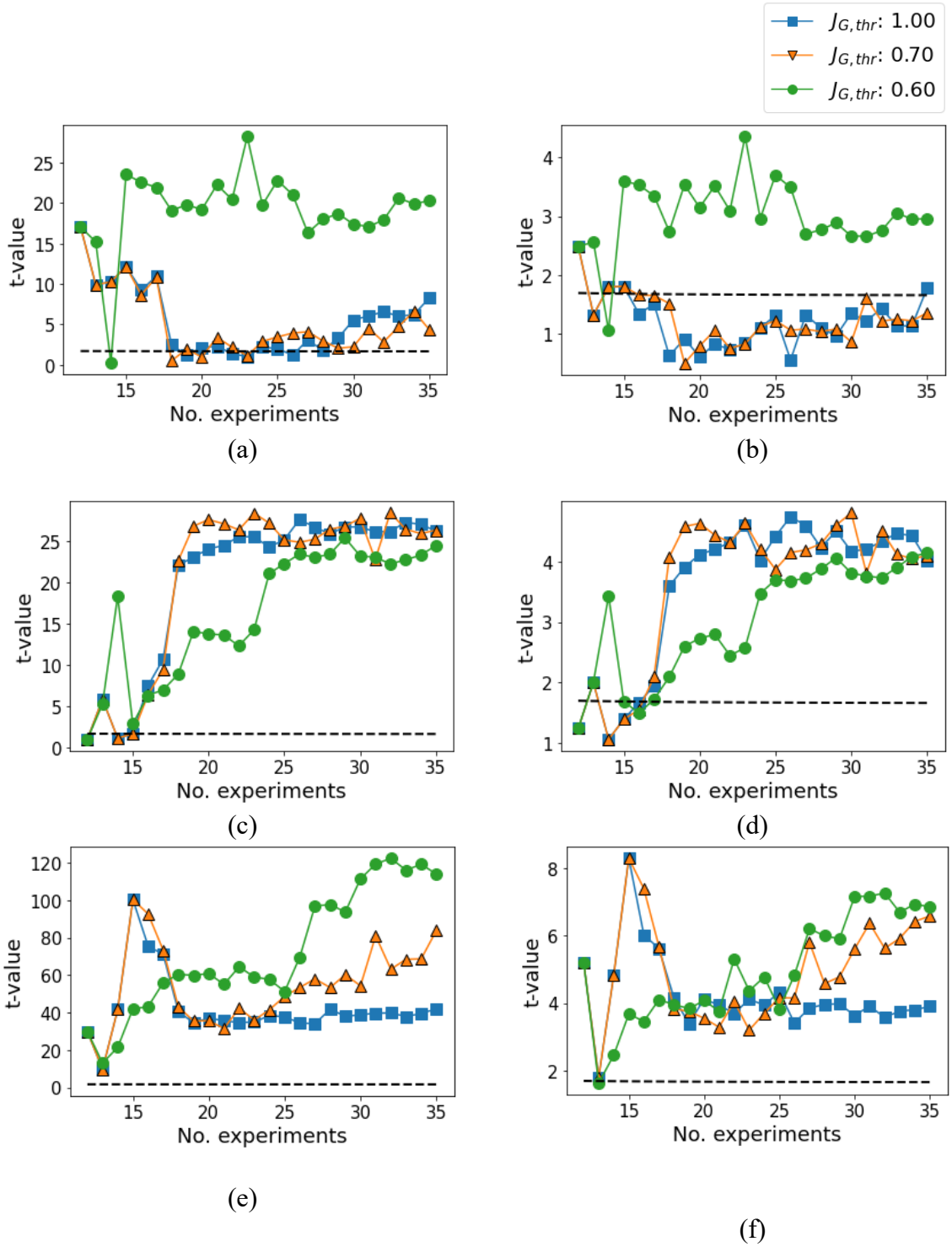



**Figure 5.7**. *Profiles of t-values calculated at every iteration of the sequential G-map eMBDoE experimentation, considering $J_{G,thr}$={1.00, 0.70, 0.60} as shown in the legend. All model parameters are considered: (a) t-values of $\hat{\theta}_1$; (b) t-values of $\hat{\theta}_2$; (c) t-values of $\hat{\theta}_3$; (d) t-values of $\hat{\theta}_4$; (e) t-values of $\hat{\theta}_5$; (f) t-values of $\hat{\theta}_6$.*

Since the G-map eMBDoE method with $J_{G,\text{thr}}$=0.60 provides the best performance in terms of maximisation of parameters precision, its parameters estimates and their 95% C.I. obtained at the last iteration are shown in Table 5.4.

**Table 5.4.** *Parameters estimates and 95%C.I. obtained with 23 optimal experiments obtained with G-map eMBDoE with $J_{G,thr}$=0.60.*

| Parameters to be estimated | Estimated value | 95% C.I. |
|---|---|---|
| $\theta_1$ | 5.51 | 0.27 |
| $\theta_2$ | 8.22 | 2.78 |
| $\theta_3$ | 5.64 | 0.23 |
| $\theta_4$ | 10.26 | 2.48 |
| $\theta_5$ | 10.41 | 0.09 |
| $\theta_6$ | 8.53 | 1.25 |

### 5.3.3 Scalar indices of model prediction variance

Scalar indices summarising $J_G$ in the whole design space are shown in Figure 5.8: minimum (Figure 5.8a), mean (Figure 5.8b) and maximum (Figure 5.8c) G-optimality.

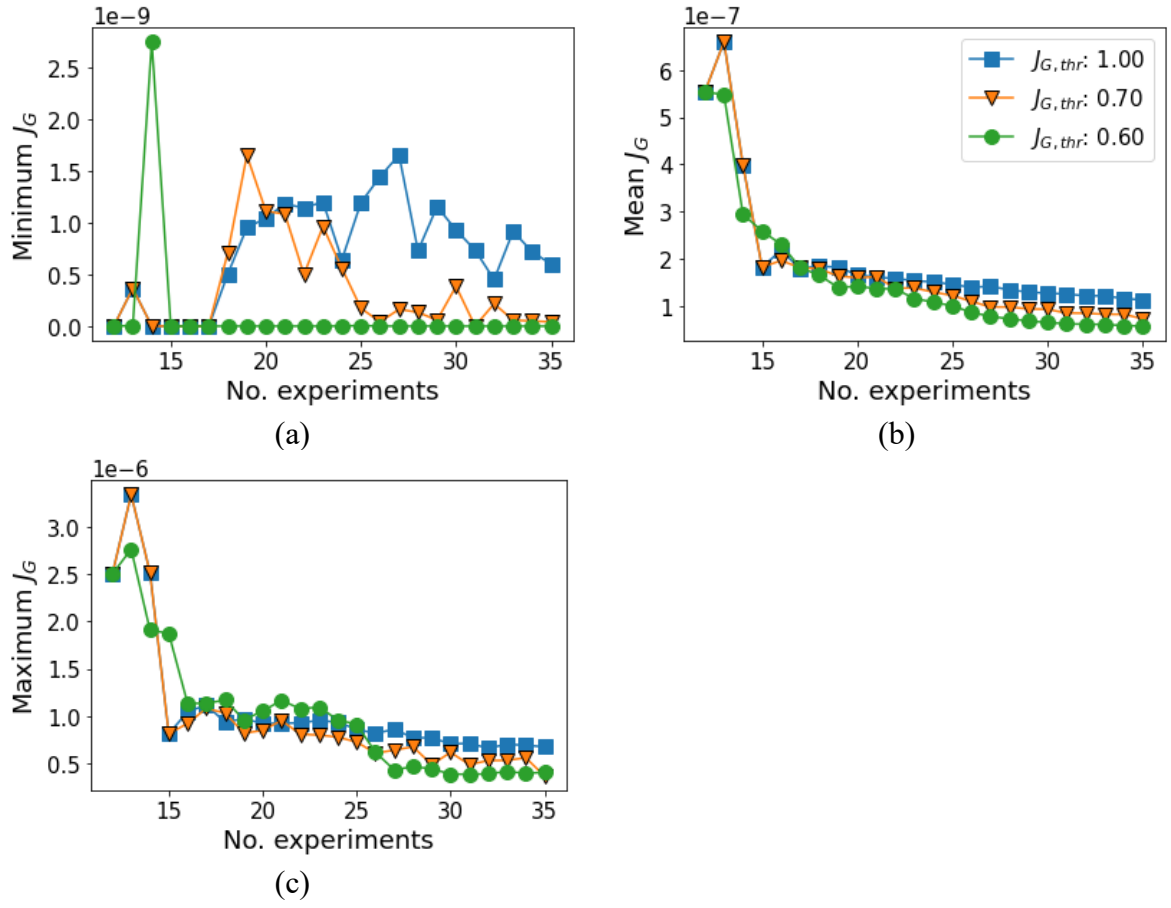



**Figure 5.8**. *Scalar indices of G-optimality at every iteration of the G-map eMBDoE experimentation with $J_{G,thr}$=1.00, $J_{G,thr}$=0.70 and $J_{G,thr}$=0.60. Minimum, mean and maximum G-optimality values calculated in the whole design space are shown in (a), (b) and (c), respectively.*

The minimum $J_G$ obtained by G-map eMBDoE with $J_{G,\text{thr}}$=0.60 is always smaller than the one calculated by $J_{G,\text{thr}}$=0.70, which in turn is smaller by the one of conventional MBDoE from the 21st experiment onward. The same rank of G-optimality reduction is found with $J_{G,\text{mean}}$ from

the 18th experiment onward and with $J_{G,\max}$ after the 25th experiment. Therefore, all of the three indices suggest that G-map eMBDoE with $J_{G,\text{thr}}$=0.60 has the best performance in terms of reduction of model prediction variance.

Since the scalar value of G-optimality $J_G$ is calculated by using the contributions of three response variables, namely $V_{y_1}$, $V_{y_2}$ and $V_{y_3}$, the profiles of their mean values are analysed in Figure 5.9 in order to check whether the rank of G-optimality reduction found in Figure 5.8 is confirmed by every single response or not.

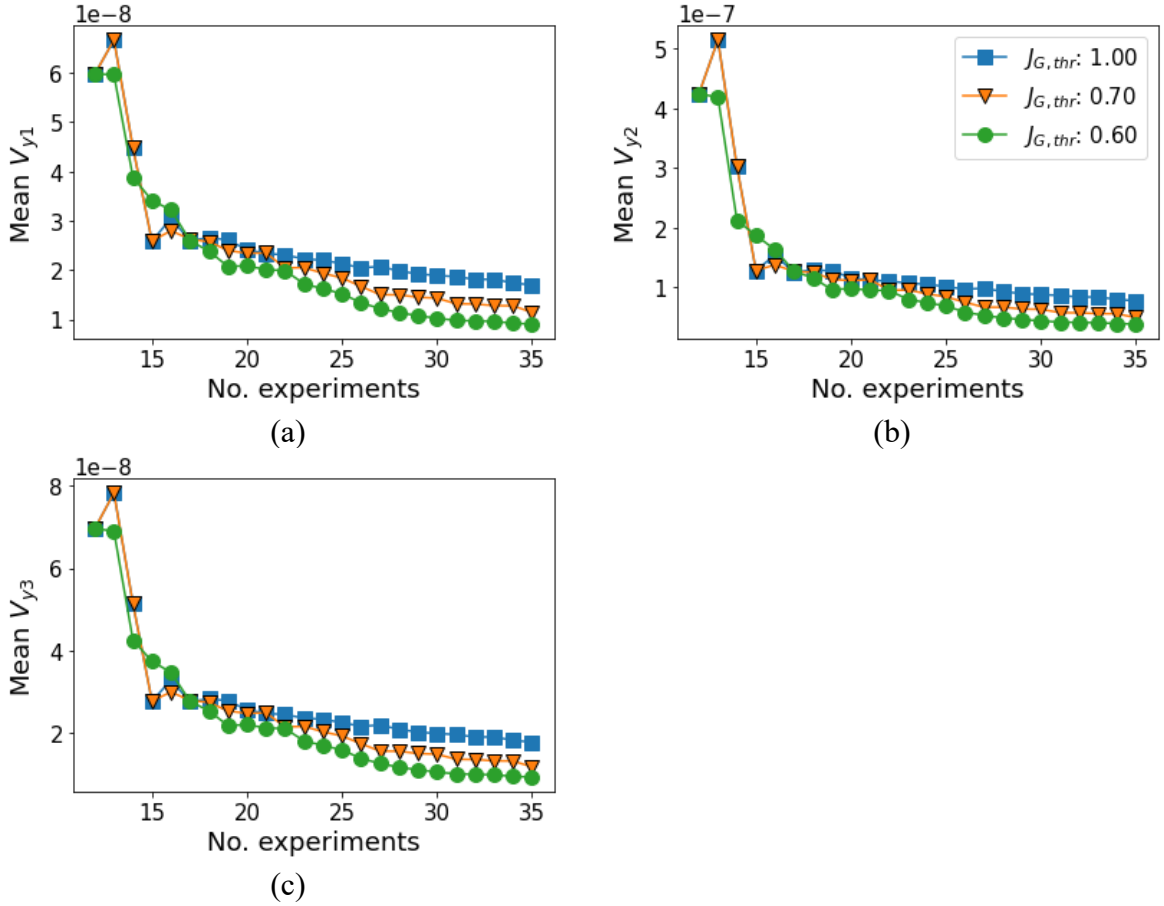



**Figure 5.9**. *Mean values of the single contributions to $J_G$ provided by the three response variables: (a) mean $V_{y_1}$; (b) mean $V_{y_2}$; (c) mean $V_{y_3}$.*

Figure 5.9a-c shows that all response variables are characterised by a greater G-optimality reduction with G-map eMBDoE. More specifically, the rank is:

MBDoE> G-map eMBDoE ($J_{G,\text{thr}} = 0.70$) > G-map eMBDoE ($J_{G,\text{thr}} = 0.60$).

Therefore, all the three responses confirm the behaviour found in Figure 5.8.

Finally, the y-axis of Figures 5.9a-c shows that the contribution to $J_G$ given by $V_{y_2}$ is one order of magnitude higher than the one of $V_{y_1}$ and $V_{y_3}$.

### 5.3.4 Maps of model prediction variance

Even though the design space is defined by four control variables, two-dimensional maps allow an easier visualisation of the results: for this reason, the G-maps of Figures 5.10 and 5.11 are built by fixing the values of two control variables, namely temperature ($u_1$) and total flowrate ($u_2$) (as explained in Section 5.2). Preliminary visualisations of several G-maps revealed that high model prediction variance is found at medium-high values of temperature and low values of total flowrate, therefore $u_1$=321.67°C and $u_2$=20.00 Nml min$^{-1}$ are chosen to be conservative. Two interesting iterations are considered: after 1 optimal experiment (Figure 5.10a-c), to see the difference in G-optimality caused by one eMBDoE experiment; after 15 optimal experiments (Figure 5.11a-c), because this is the first iteration where the maximum G-optimality obtained by eMBDoE with $J_{G,\mathrm{thr}}$=0.60 is the smallest among the three design methods in the whole design space (as shown in Figure 5.8c).

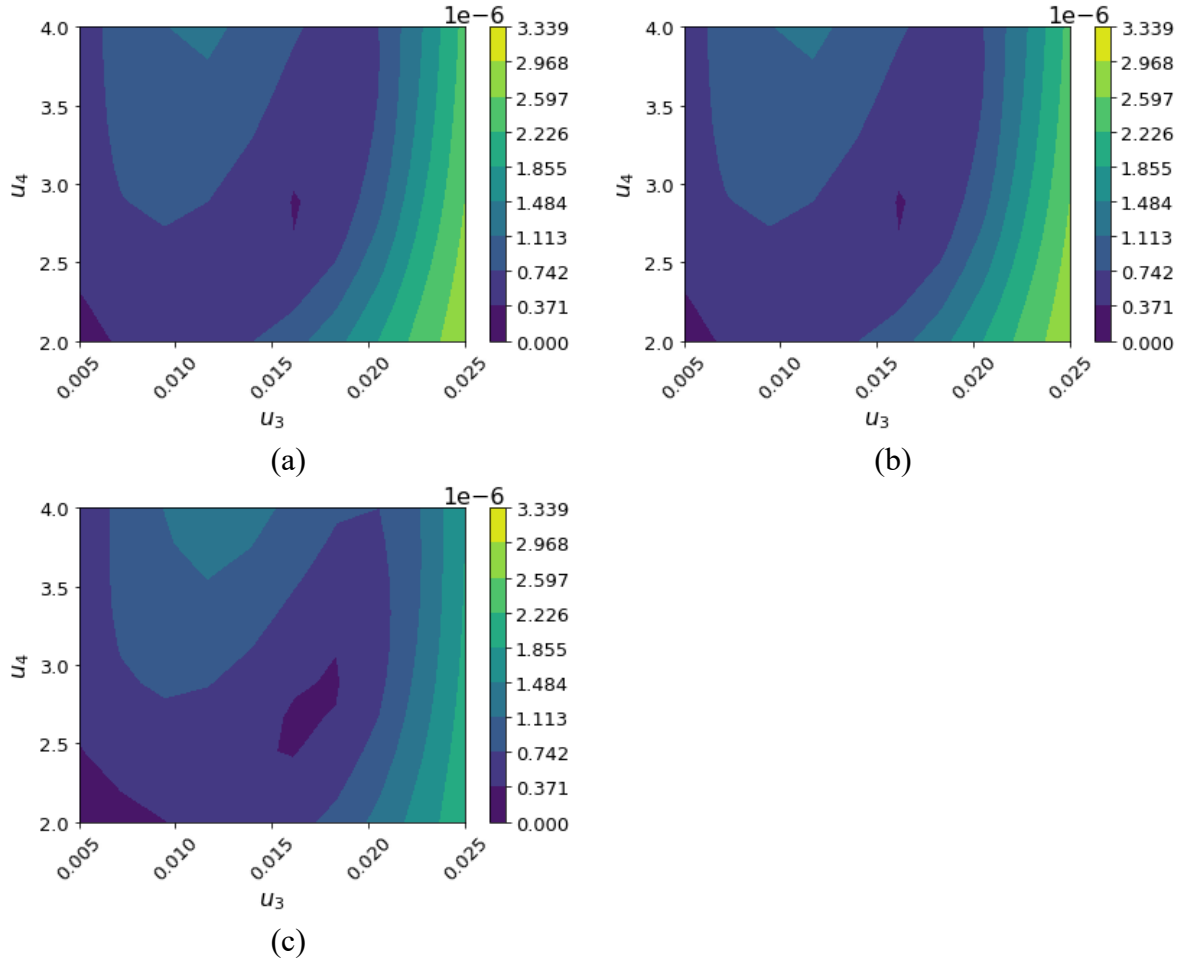



**Figure 5.10**. *G-maps obtained after 1 optimal experiment (besides the 12 preliminary ones) calculated by: (a) MBDoE; (b) G-map eMBDoE with $J_{G,thr}$=0.70; (c) G-map eMBDoE with $J_{G,thr}$=0.60. Fixed control variables: $u_1$=321.67°C and $u_2$=20.00 Nml min$^{-1}$.*

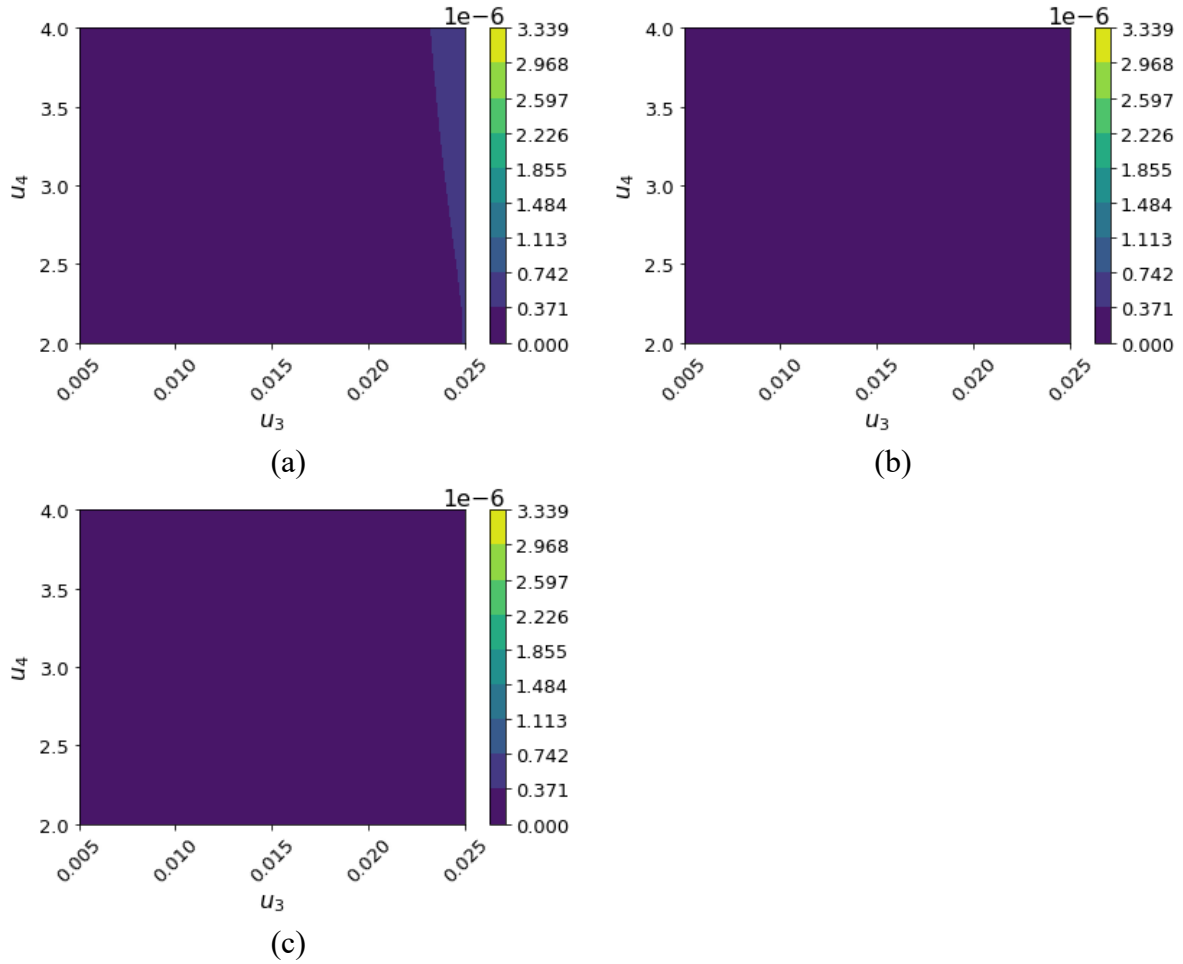



**Figure 5.11**. *G-maps obtained after 15 optimal experiments (besides the 12 preliminary ones) calculated by: (a) MBDoE; (b) G-map eMBDoE with $J_{G,thr}$=0.70; (c) G-map eMBDoE with $J_{G,thr}$=0.60. Fixed control variables: $u_1$=321.67°C and $u_2$=20.00 Nml min$^{-1}$.*

After measuring one optimal experiment, MBDoE (Figure 5.10a) and G-map eMBDoE with $J_{G,thr}$=0.70 (Figure 5.10b) have an identical distribution of model prediction variance, as expected by the fact that their design is the same at that iteration. Model prediction variance is slightly improved by G-map eMBDoE with $J_{G,thr}$=0.60 (Figure 5.10c), as indicated by a slightly wider region characterised by low model prediction variance where $u_3$ is between 0.015 and 0.018 molmol$^{-1}$ and $u_4$ is between 2.4 and 3 molmol$^{-1}$ (dark blue regions). After measuring 15 optimal experiments (Figures 5.11a-c), model prediction variance is greatly reduced by all of the three methods across the entire design space. Only a small region at $u_3$ close to 0.025molmol$^{-1}$ has still a slightly higher G-optimality with conventional MBDoE (Figures 5.11a).

Therefore, G-maps confirm the results obtained with scalar indices of G-optimality (Subsection 5.3.3): G-map eMBDoE with $J_{G,thr}$=0.60 has the best performance in terms of model prediction

variance reduction. The corresponding H-maps (namely, maps of information content) are shown in Appendix F.

### 5.3.5 Model prediction accuracy

The overall model performance in terms of model prediction accuracy is assessed by using:

- 12 preliminary DoE experiments;
- 23 MBDoE experiments, i.e., G-map eMBDoE with $J_{G,\text{thr}}$=1.00;
- 23 G-map eMBDoE experiments with $J_{G,\text{thr}}$=0.70;
- 23 G-map eMBDoE experiments with $J_{G,\text{thr}}$=0.60.

Therefore, 81 experiments are used to calculate the mean absolute error (as described in Section 5.2); results are shown in Figures 5.12 a-c for the predicted mole fractions of $CH_4$, $O_2$ and $CO_2$, respectively.

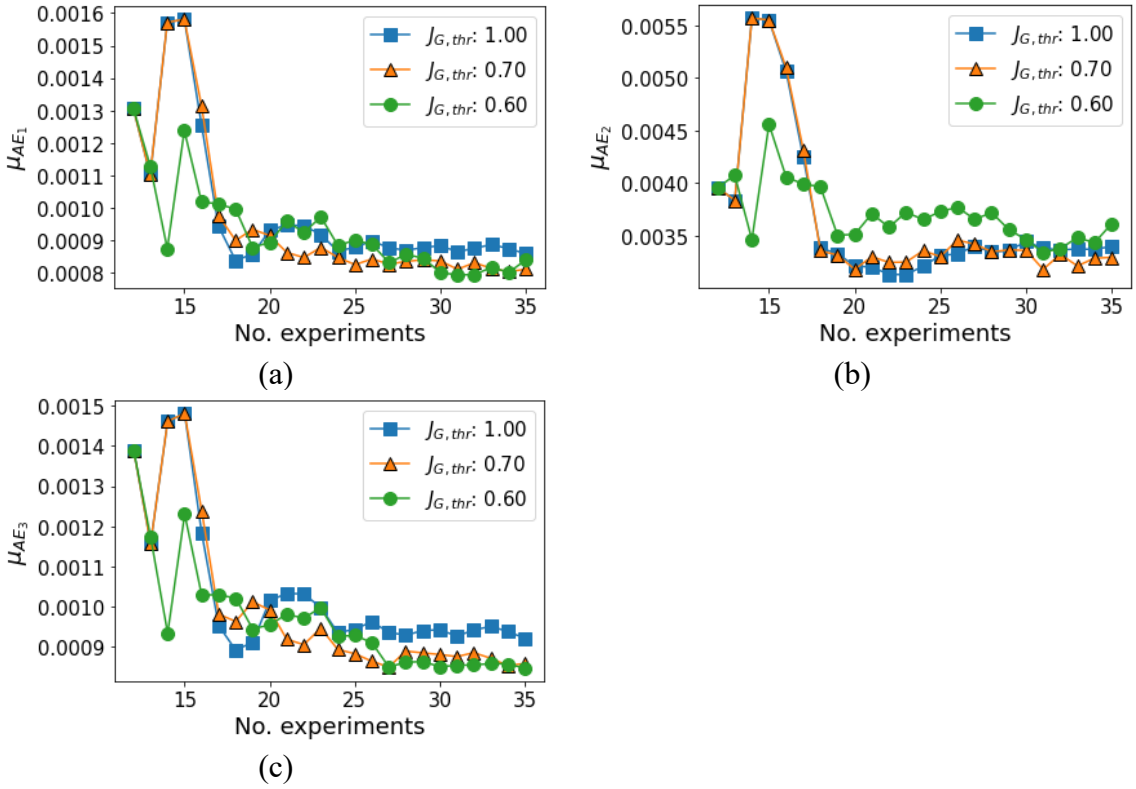



**Figure 5.12**. *Profiles of mean absolute error $\mu_{AE_i}$: (a) $i = CH_4$; (b) $i = O_2$; (c) $i = CO_2$.*

Considering the mean of the absolute error, eMBDoE with $J_{G,\text{thr}}$=0.60 has the best predictions of all three responses in the first five iterations. Then, it predicts $CH_4$ and $CO_2$ molar fractions better than MBDoE from the 26$^{\text{th}}$ and 20$^{\text{th}}$ experiment onwards, respectively. The mean absolute error of $O_2$ is higher than the ones with $J_{G,\text{thr}}$={1, 0.70} between the 18$^{\text{th}}$ and 29$^{\text{th}}$ experiments, then it reduces to values that are comparable to those of the other two design methods.

Therefore, the good results of eMBDoE in terms of model prediction variance (Subsections 5.3.3-5.3.4) are coherent with the good results in terms of prediction accuracy.

## 5.4 Conclusions

The explorative MBDoE method based on G-maps enhances the capability of the automated flow micropacked bed catalytic reactor platform to save time and resources to identify the kinetics of total methane oxidation. Indeed, conventional MBDoE is able to get statistically sound parameters estimates within the experimental budget of 23 optimal experiments, but 5 optimal experiments are sufficient with G-map eMBDoE with a proper G-optimality threshold, thus reducing the number of experiments of 78%. Moreover, the novel method proposed to select the G-optimality threshold based on the difference between eMBDoE and MBDoE designs has proved to be effective in achieving a good trade-off between space exploration and information maximisation: it provides both the highest parameters precision achieved and the minimum model prediction variance in the whole design space. Results are satisfactory also when measured and predicted values of the response variables are compared for all 81 collected experiments, thus confirming the adequacy of the identified kinetic model in predicting a variety of experimental conditions within the design space.

In addition, graphical visualisation of maps of model prediction variance is effective in showing regions of the design space where the kinetic model is more reliable and in showing the progressive reduction of model prediction uncertainty as soon as new experiments are collected by the platform. Ongoing work consists in the integration of the Python code for G-map eMBDoE into LabView software in order to achieve a fully autonomous execution of experiments design, measurement and analysis in the microreactor platform.

# Chapter 6
# Autonomous adaptation of the trade-off between space exploration and information maximisation for exploratory MBDoE[4]

In this Chapter, a novel method, named *adaptative G-map eMBDoE*, is developed in order to select the G-optimality constraint without human intervention, thanks to the analyses the overlap between maps of information content and maps of G-optimality. To make a fair comparison, the adaptative G-map eMBDoE is applied to Model 1 and 2 used in Cenci et al. (2023; Chapter 4), using the same simulation settings and comparing its results to the ones previously obtained by MBDoE, G-map eMBDoE with thresholds 0.25 and 0.75, LH and factorial DoE. The reduction of model prediction variance achieved by the novel method is intermediate between the one of eMBDoE with threshold 0.25 (worse performance) and with threshold 0.75 (best performance) and more marked than the one of MBDoE. Moreover, the novel method leads to precise parameters with an experimental burden that is comparable to the one of MBDoE and G-map eMBDoE with threshold 0.75. This suggests that the adaptative G-map eMBDoE allows to find a proper trade-off between space exploration and information maximisation, having the advantage with respect of G-map eMBDoE of requiring no prior information on the suitable G-optimality constraint for the system under study.

## 6.1 Introduction

Mathematical models are employed at all steps of pharmaceutical R&D and they should be built in a rigorous way for QbD-based submissions. In fact, some key aspects highlighted in ICH Points to Consider (R2) to guide modellers in the pharmaceutical industry are: *i*) the importance of data collection to identify model equations and parameters; *ii*) the necessity to assess model

[4] Cenci, F., Pankajakshan, A., Bawa, S. G., Gavriilidis, A., Facco, P. and Galvanin, F.. Novel algorithm for the autonomous execution of G-map eMBDoE experiments by means of automated chemical platforms [in preparation].

prediction uncertainty and to periodically update the model in order to always ensure model reliability. Both goals can be achieved with model-based design of experiments, since it is an optimisation problem where the objective function can be formulated in order to discriminate among candidate model structures, to precisely identify model parameters or to minimise model prediction uncertainty (Espie and Macchietto, 1989; Asprey and Macchietto, 2000; Kiefer and Wolfowitz, 1959). Specifically, the G-map eMBDoE method proposed in Chapter 4 (Cenci et al., 2023) and Chapter 5 have the advantage of minimising model prediction uncertainty in the whole design space with reduced experimental burden, while obtaining statistically precise parameters values. This is achieved through a trade-off between space exploration and information maximisation handled by means of a constraint on model prediction variance (calculated as G-optimality $J_G$; Kiefer and Wolfowitz, 1959; Kiefer and Wolfowitz, 1960; Wong, 1995) that must be satisfied by candidate design points. However, the best G-optimality constraint, namely the inequality type (i.e., $J_G \geq J_{G,\text{thr}} J_{G,\max}$ or $J_G \leq J_{G,\text{thr}} J_{G,\max}$ ) and the threshold value itself, i.e. $J_{G,\text{thr}}$, is case-dependent, meaning that it may vary based on the mathematical model and the preliminary parameters values used. In Chapter 4, different threshold values are compared for the constraint $J_G \geq J_{G,\text{thr}} J_{G,\max}$ and the most suitable one, $J_{G,\text{thr}}$=0.75, is identified at the end of the simulated experimental campaigns. In Chapter 5, the most suitable inequality type is selected based on preliminary simulations of the system, resulting to be $J_G \leq J_{G,\text{thr}} J_{G,\max}$, and two thresholds are used: a) 0.70, which provides satisfactory results in preliminary simulations performed with initial parameters values; b) 0.60, based on a method proposed in Chapter 5 that selects the first threshold (from $J_{G,\text{thr}}$=1 to $J_{G,\text{thr}}$=0, with steps of 0.05) leading to a designed experiments that is different from the one designed by MBDoE. In both cases, the threshold is kept fixed throughout the experimentation. However, it may be convenient to have a method that automatically selects the most suitable inequality type and threshold without requiring preliminary experiments and/or human intervention. In fact, it would be beneficial in case of completely new systems where the plausible range of parameters $[\widehat{\boldsymbol{\theta}}_{\text{LB}}, \widehat{\boldsymbol{\theta}}_{\text{UB}}]$ is large and little to no information is available on the most suitable initial parameters values $\widehat{\boldsymbol{\theta}}_0$. Moreover, it may aid the exploitation of the full potential of Industry 4.0 technologies in a variety of industrial sectors, including the (bio)pharmaceutical one. In fact, Frank et al., (2019) reviewed the application of Industry 4.0 technologies in 92 manufacturing companies from machinery and equipment sector and concluded that the highest complexity in the implementation is represented by the full flexibility

of Smart Manufacturing, for instance the adoption of flexible lines that automatically adjust the manufacturing to multiple product types and/or to changing conditions without human intervention. Similarly, Barz et al. (2022) reviewed the applications of automated continuous flow platforms and bioreactor platforms and concluded that automated technology is available in the (bio) pharmaceutical industry, but expensive laboratories providing scarcely informative experiments are still used due to a lack of proper experimental plan.

In this Chapter, a novel G-map eMBDoE method to automatically select the most suitable G-optimality constraint is presented. It is based on the analysis of the overlap between maps of information content, H-maps, and maps of model prediction variance (estimated in terms of G-optimality), G-maps: if the most informative experiments have the highest model prediction variance, space exploration is enhanced with respect to state-of-the-art MBDoE by selecting candidate design points having lower model prediction variance (namely, by setting a constraint of the type $J_G \leq J_{G,\mathrm{thr}}\, J_{G,\mathrm{max}}$); if they have the lowest model prediction variance, space exploration is enhanced by favoring points with higher model prediction variance (namely, by setting $J_G \geq J_{G,\mathrm{thr}}\, J_{G,\mathrm{max}}$). This procedure is completely general, since it does not depend on the specific model type and/or current parameters values, and the calculated G-optimality constraint can be adapted as soon as a new calibration experiment is measured. Therefore, the proposed method is referred to as *adaptative G-map eMBDoE*; more details are provided in Section 6.2. To be able to make direct comparisons with the performance of the original G-map eMBDoE method (Cenci et al., 2023; Chapter 4), the proposed method is applied to Model 1 and 2 used in Chapter 4, using the same simulation settings. Sections 6.3.1 and 6.3.2 show the results of the adaptative G-map eMBDoE applied to Model 1 and 2, respectively, comparing them to the ones of MBDoE, G-map eMBDoE, LH and factorial DoE. Finally, conclusions and future work are explained in Section 6.5.

## 6.2 Mathematical methods

The proposed method assumes that the most suitable model structure **f** has already been selected among a set of candidates and that experiments must be collected to to minimise model prediction variance in the whole design space, while ensuring statistically sound model parameters and a reduced experimental burden. As explained in Chapter 5, an explorative MBDoE can be obtained by defining a set of candidate design points $\boldsymbol{\varphi}_{\mathrm{cand}}$ over which the objective function of MBDoE can be optimised. In G-map eMBDoE (Cenci et al., 2023), the set of candidates $\boldsymbol{\varphi}_{\mathrm{cand}}$ is given by the points satisfying a constraint on G-optimality, while the

objective function can be one of the classical 'alphabetical' criteria for parameters precision (Pukelsheim, 1993). To determine $\boldsymbol{\varphi}_{\text{cand}}$, a scalar index $J_G$ of G-optimality is calculated for every possible experimental condition. In this Dissertation, $J_G$ is given by the sum of the contributions of model prediction variance given by all responses at all sampling points. Then, candidate design points are selected based on a specific constraint: in Chapter 4, the most suitable constraints for Model 1 and 2 were given by $J_G \geq J_{G,\text{thr}} J_{G,\text{max}}$; in Chapter 5, the most suitable constraints for the kinetic model of total methane oxidation was given by $J_G \leq J_{G,\text{thr}} J_{G,\text{max}}$. Also, the most suitable threshold $J_{G,\text{thr}}$ was case-dependent.

In fact, both inequality types ($\geq$ and $\leq$) may be suitable to enhance space exploration with respect to state-of-the-art MBDoE. The main rationale is illustrated in Figure 6.1 (consider that the state-of-the-art MBDoE would select the point with maximum information content $\psi_{\text{max}}$, namely the white point in Figure 6.1):

- if the region with the highest information content (namely, $\psi_{\text{max}}$) overlaps with the region having the maximum model prediction variance (Figure 6.1a), space exploration can be achieved by selecting candidate experimental conditions having lower values of G-optimality. This is equivalent of choosing candidate design points $\boldsymbol{\varphi}_{\text{cand}}$ satisfying the following condition:

$$J_G \leq J_{G,\text{thr}} J_{G,\text{max}}. \tag{6.1}$$

- if the region with the highest information content (namely, $\psi_{\text{max}}$) overlaps with the region having the minimum model prediction variance (Figure 6.1b), space exploration can be achieved by favoring the experiments with higher model prediction variance $J_G$. In other terms, the candidate design points $\boldsymbol{\varphi}_{\text{cand}}$ of the G-map eMBDoE problem should be of the form:

$$J_G \geq J_{G,\text{thr}} J_{G,\text{max}}; \tag{6.2}$$

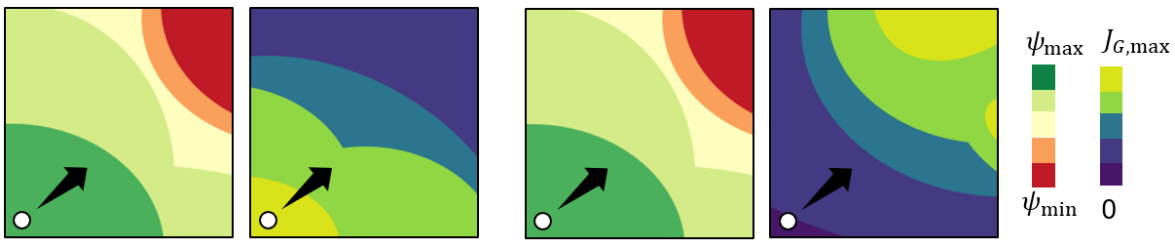



**Figure 6.1**. *Illustrative examples of H-maps (red-green maps with values from $\psi_{\text{min}}$ to $\psi_{\text{max}}$) and G-maps (blue-yellow maps with values from 0 to $J_{G,\text{max}}$).*

In this work, a method to automatically select the G-optimality constraint (both inequality type and threshold value) is developed. Moreover, the constraint is adapted as soon as a new

experiment is performed; therefore, the proposed method will be referred to as *adaptative G-map eMBDoE* (Figure 6.2) from now on.

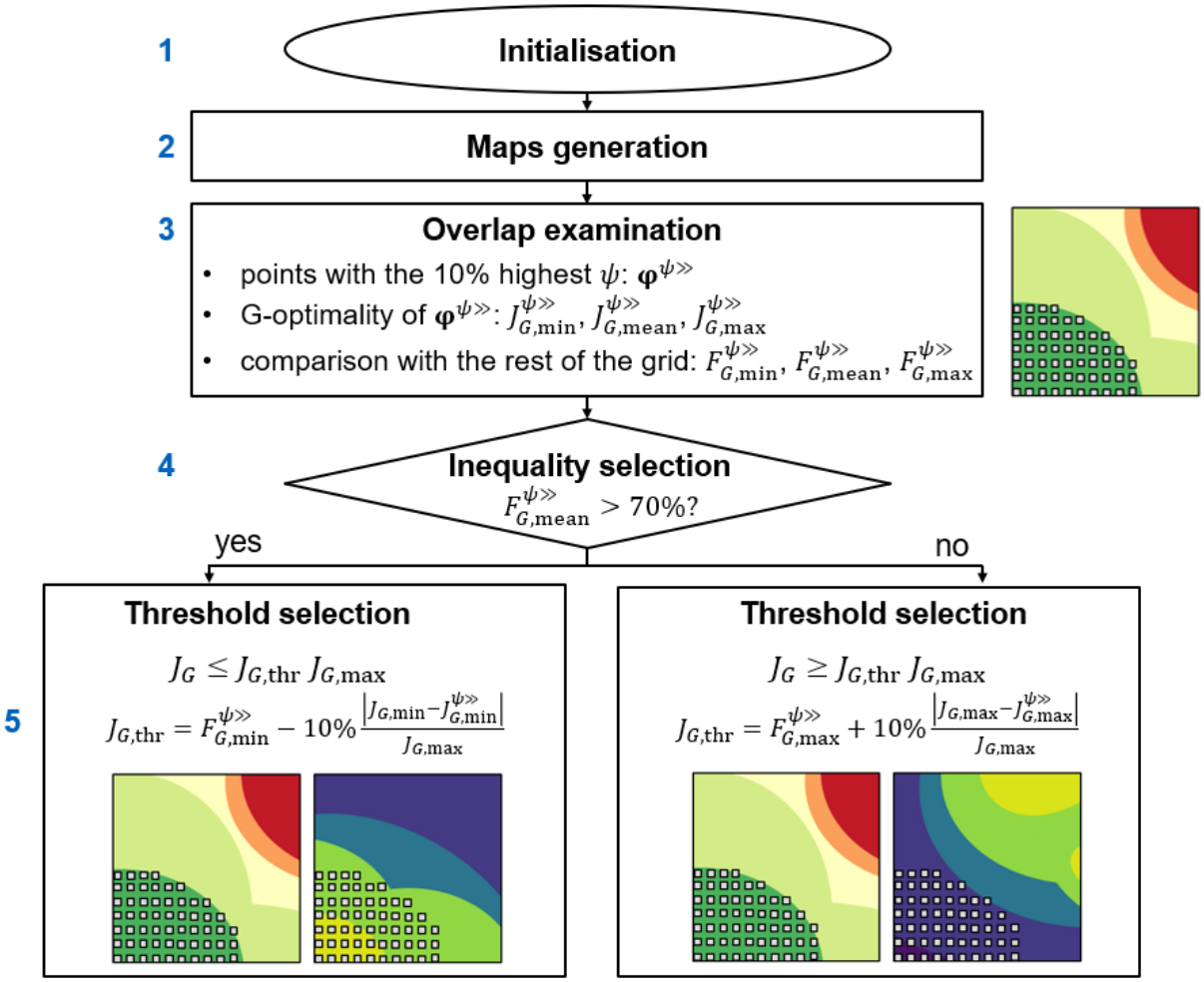



**Figure 6.2**. *Scheme of the adaptative G-map eMBDoE.*

As shown in Figure 6.2, the proposed method is made of the following steps:

- Step 1. Initialisation of the procedure, namely in the definition of: model equations **f**; parameters bounds and initial values $\widehat{\boldsymbol{\theta}}_{\mathrm{LB}}, \widehat{\boldsymbol{\theta}}_{\mathrm{UB}}, \widehat{\boldsymbol{\theta}}_0$; response measurement errors $\boldsymbol{\sigma}_y$; control variables **u** and their bounds $\mathbf{u}_{\mathrm{LB}}, \mathbf{u}_{\mathrm{UB}}$.
- Step 2. Generation of G-maps and H-maps with the current parameters values and considering the whole design space.
- Step 3. Analysis of the overlap of H-maps and G-maps. First, the range of values of information content calculated in the entire H-map, namely $[\psi_{\min}, \psi_{\max}]$ is divided into 10 equally spaced intervals, thus intervals with length $\frac{1}{10}(\psi_{\max} - \psi_{\min})$. Then, the points having the highest information content, i.e. the ones with $\psi \in [\frac{9}{10}(\psi_{\max} - \psi_{\min}), \psi_{\max}]$,

indicated as $\boldsymbol{\varphi}^{\psi\gg}$ and represented by grey squares in Figure 6.2, are characterised: *i*) the $J_G$ values of all highly informative points $\boldsymbol{\varphi}^{\psi\gg}$ are considered (considering Figure 6.2, the $J_G$ values of the grey squares in the G-map) and their minimum, mean and maximum values (respectively, $J_{G,\min}^{\psi\gg}, J_{G,\text{mean}}^{\psi\gg}, J_{G,\max}^{\psi\gg}$) are calculated; *ii*) minimum, mean and maximum values of G-optimality of $\boldsymbol{\varphi}^{\psi\gg}$ are compared to the $J_G$ values of the rest of the G-map by calculating the following fractions:

$$F_{G,\min}^{\psi\gg} = \frac{J_{G,\min}^{\psi\gg}}{J_{G,\max}}; \tag{6.3}$$

$$F_{G,\text{mean}}^{\psi\gg} = \frac{J_{G,\text{mean}}^{\psi\gg}}{J_{G,\max}}; \tag{6.4}$$

$$F_{G,\max}^{\psi\gg} = \frac{J_{G,\max}^{\psi\gg}}{J_{G,\max}}. \tag{6.5}$$

These indices are useful for the following steps.

- Step 4. Selection of the inequality type: a) if $F_{G,\text{mean}}^{\psi\gg} > 70\%$, it means that the regions with highest information content overlap with regions with the highest model prediction variance, therefore the inequality type $J_G \leq J_{G,\text{thr}} J_{G,\max}$ is used and the threshold is selected as in Step 5a; b) if $F_{G,\text{mean}}^{\psi\gg} \leq 70\%$, it means that the regions with highest information content overlap with regions with lower model prediction variance, therefore the inequality type $J_G \geq J_{G,\text{thr}} J_{G,\max}$ is used and the threshold is selected as in Step 5b.
- Step 5. Selection of the G-optimality threshold $J_{G,\text{thr}}$. If $F_{G,\text{mean}}^{\psi\gg} > 70\%$, space exploration is favoured by the following threshold (Step 5a):

$$J_{G,\,\text{thr}} = F_{G,\min}^{\psi\gg} - 10\% \frac{\left|J_{G,\,\min} - J_{G,\min}^{\psi\gg}\right|}{J_{G,\,\max}} \tag{6.6}$$

If $F_{G,\text{mean}}^{\psi\gg} \leq 70\%$, space exploration is favoured by the following threshold (Step 5b):

$$J_{G,\,\text{thr}} = F_{G,\max}^{\psi\gg} + 10\% \frac{\left|J_{G,\,\max} - J_{G,\max}^{\psi\gg}\right|}{J_{G,\,\max}} \tag{6.7}$$

In the procedure explained, the use of G-optimality values referring to the set of points $\boldsymbol{\varphi}^{\psi\gg}$ (namely, $J_{G,\min}^{\psi\gg}, J_{G,\text{mean}}^{\psi\gg}, J_{G,\max}^{\psi\gg}, F_{G,\min}^{\psi\gg}, F_{G,\text{mean}}^{\psi\gg}, F_{G,\max}^{\psi\gg}$ ) and to the whole H-map ($\psi_{\min}, \psi_{\max}$) and G-map ($J_{G,\,\max}$) are necessary in order to obtain a general procedure that can be applied at any model at any step of the experimental campaign. For instance, in Chapter 4 G-optimality values are of the order of 10 for Model 1 of and of the order of 1 for Model 2, therefore a constraint like, e.g., $J_G \geq J_{G,\text{thr}} J_{G,\max}$=8 would be meaningful for Model 1, but useless with Model 2.

The method illustrated in Figure 6.2 can be applied iteratively, as soon as a new experiment is used to update model parameters. Since it does not require user-defined settings, it is suitable for the integration in an automated chemical platform. Ongoing work is focusing on the integration of a Python 3.9 (Rossum and Drake, 2009) script into the LabView program of the platform for total methane oxidation presented in Chapter 5. In this Chapter, the adaptative G-map eMBDoE is evaluated in silico by comparing its results to the ones obtained in Chapter 4 with the original G-map eMBDoE method. Specifically, the results of G-map eMBDoE, MBDoE, factorial DoE and LH in terms of experiments design, parameters precision and model prediction variance (Section 4.3, Chapter 4) are directly compared to the ones of the new method applied to Model 1 and Model 2. For sake of comparison, two different G-optimality thresholds are used for the original G-map eMBDoE: $J_{G,\mathrm{thr}}$=0.25 and $J_{G,\mathrm{thr}}$=0.75.

## 6.3 Results

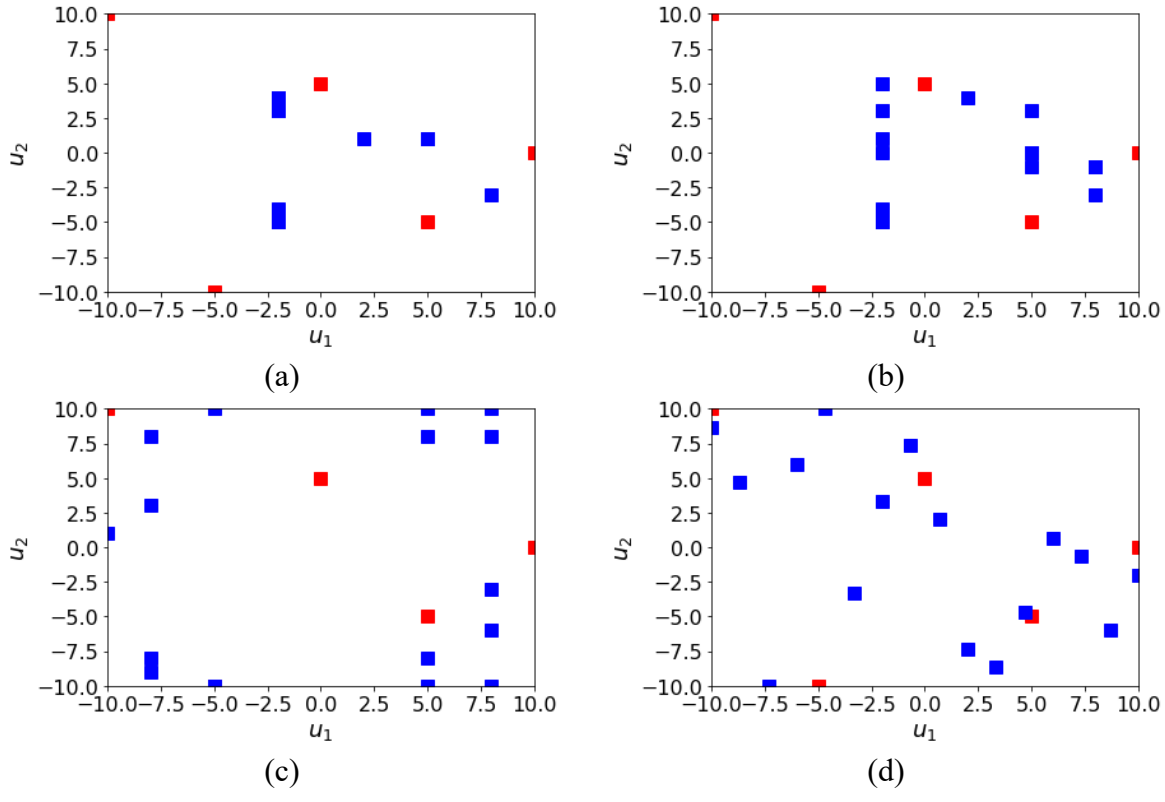
To test the performance of the adaptative G-map eMBDoE method proposed in this work, the same models used to test the original G-map eMBDoE are used: (*i*) Model 1, represented by the algebraic equation and simulation settings (e.g., control variables ranges, parameters bounds and initial values; preliminary experiments; E-optimal criterion) used in Section 4.3.1 of Chapter 4; (*ii*) Model 2, represented by the differential equations and simulation settings (e.g., control variables ranges, parameters bounds and initial values; sampling points; preliminary experiments; E-optimal criterion) used in Section 4.3.2 of Chapter 4. Similarly to Chapter 4, the experiments designed through the adaptative G-map eMBDoE are simulated according to the following procedure:

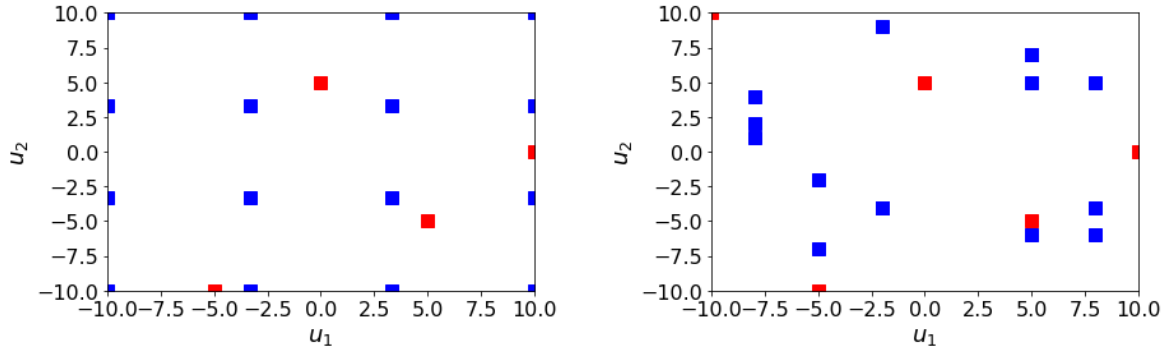
- model equations and true parameter vector $\boldsymbol{\theta}_{\mathrm{true}}$ (the same as in Tables 4.1 and 4.3 of Chapter 4) are used to generate the exact value of the model responses $y_{\mathrm{exact}}$ at the selected experimental condition;
- a gaussian error with zero mean and standard deviation $\sigma_y$ (the same $\sigma_y$ used in Chapter 4 for Model 1 or 2) is then added to $y_{\mathrm{exact}}$ to obtain a "noisy" measurement $y_{\mathrm{noisy}}$.

Finally, the results are analysed in terms of: (*i*) experiments design; (*ii*) parameters precision as $t$-tests; (*iii*) profiles of scalar indices of G-optimality; (*iv*) G-maps at specific iterations.

### *6.3.1 Model 1*

The experimental budget for Model 1 is made of $N_e$=21 experiments, with 5 preliminary LH experiments equal for all methods considered. Figure 6.3 shows the 16 experiments designed by means of the different methods. The experiments designed through MBDoE (Figure 6.3a) focus in the central region of the design space, namely around $u_1 = 0$ and $u_2 = 0$, while an increasing threshold such as $J_{G,\text{thr}}$=0.25 (Figure 6.3b) and $J_{G,\text{thr}}$=0.75 (Figure 6.3c) tend to favour experimental conditions towards the boundary of the design space, namely $u_1$ and $u_2$ around $-10$ and 10. The outcome of the adaptative G-map eMBDoE is an intermediate situation between the two G-map eMBDoE considered: in fact, it selects points that are more away from the center of the design space with respect to G-map eMBDoE with $J_{G,\text{thr}}$=0.25 (and, especially, with respect to MBDoE), but less points laying on the boundary of the design space with respect to G-map eMBDoE with $J_{G,\text{thr}}$=0.75.






(e) (f)

**Figure 6.3**. *Design space with the experiments selected by: (a) MBDoE; (b) G-map eMBDoE $J_{G,thr}$=0.25; (c) G-map eMBDoE $J_{G,thr}$=0.75; (d) Latin Hypercube; (e) $4^2$ full factorial DoE; (f) adaptative G-map eMBDoE. Red squares indicate the 5 preliminary experiments.*

Table 6.1 shows the number of distinct design points for every method used. The adaptative G-map eMBDoE selects 13 distinct points, therefore also in this case the outcome is an intermediate situation between G-map eMBDoE with $J_{G,\text{thr}}$=0.25 (12 distinct points) and G-map eMBDoE with $J_{G,\text{thr}}$=0.75 (16 distinct points).

**Table 6.1.** *Number of distinct design points for each scenario compared in the study.*

| Scenario | No. distinct design points |
|---|---|
| MBDoE | 7 |
| eMBDoE, thr:0.25 | 12 |
| eMBDoE, thr:0.75 | 16 |
| LH | 16 |
| DoE | 16 |
| Adaptative eMBDoE | 13 |

Table 6.2 shows the G-optimality constraints selected in the 16 optimal designs. The two inequality types ≥ and ≤ are used with the same frequency, while the thresholds $J_{G,\text{thr}}$ selected are all between 0.50 and 0.90.

**Table 6.2.** *Constraint selected by the adaptative G-map eMBDoE. "L" stands for "lower" and indicates the inequality ≤; "H" stands for "higher" and indicates the inequality ≥.*

| Inequality type | $J_{G,thr}$ |
|---|---|
| L | 0.89 |
| H | 0.76 |
| L | 0.50 |
| L | 0.76 |
| H | 0.89 |
| H | 0.72 |
| H | 0.82 |
| H | 0.72 |
| H | 0.65 |
| L | 0.50 |
| H | 0.90 |
| L | 0.57 |
| L | 0.84 |
| L | 0.83 |
| L | 0.81 |

| | |
|---|---|
| L | 0.81 |

Parameters precision is assessed in terms of t-tests, as shown in Figure 6.4. The parameter requiring a higher number of calibration experiments to be precisely estimated is $\theta_5$ (Figure 6.4e). MBDoE and G-map eMBDoE with $J_{G,thr}$={0.25, 0.75} require 10 experiments (5 preliminary and 5 optimal) to pass the *t*-test, LH requires 17 experiments, while DoE is not able to pass the *t*-test within the experimental budget. The adaptive G-map eMBDoE method is able to estimate the parameter with 11 experiments, therefore its performance is similar to the ones of state-of-the-art MBDoE for parameters precision.

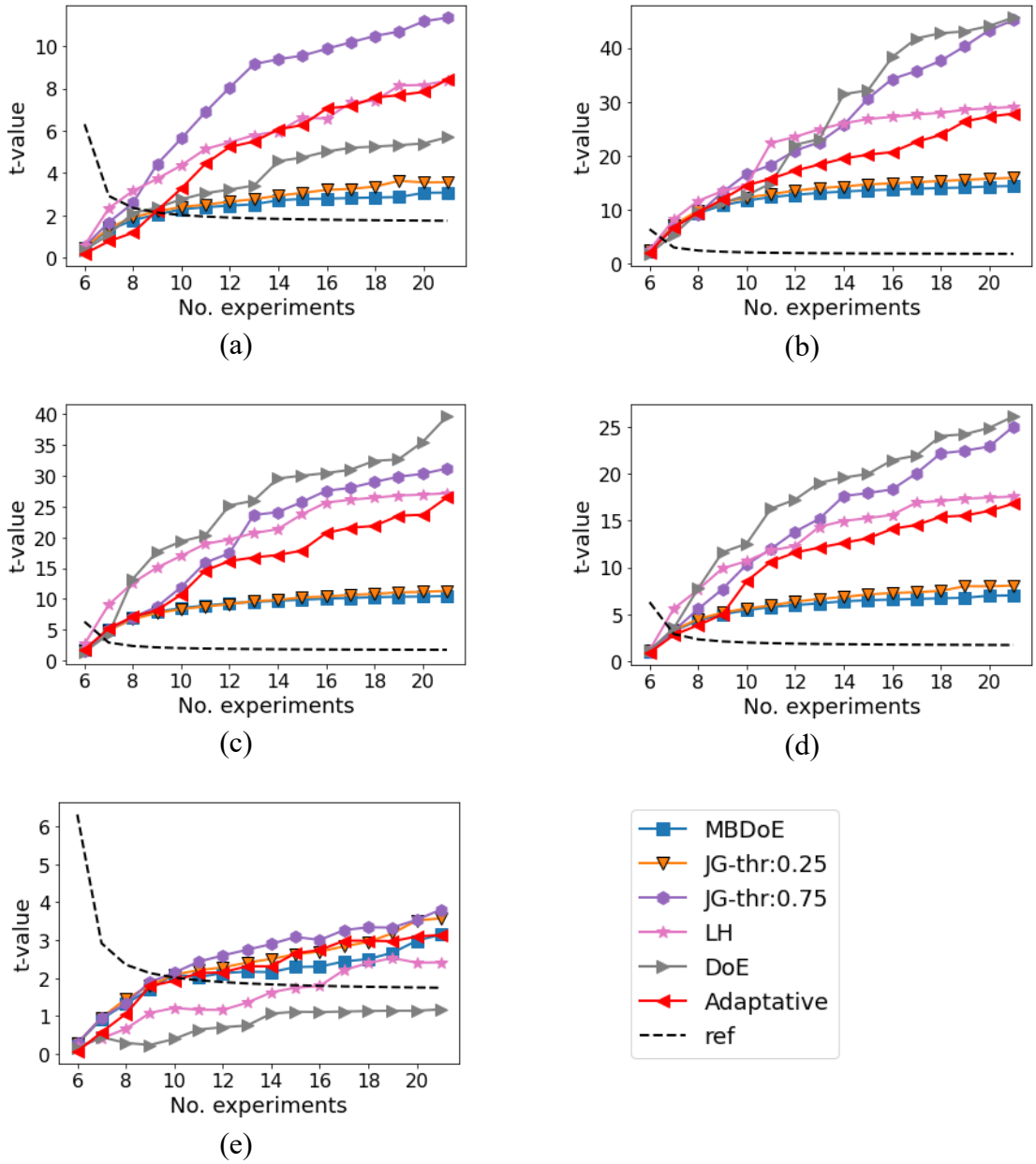



**Figure 6.4**. *Profiles of t-values calculated with: MBDoE; G-map eMBDoE ($J_{G,thr}$={0.25;0.75}); Latin Hypercube (LH); $4^2$ full factorial DoE; adaptative MBDoE. Figures (a)-(e) show results of parameters 1-5, respectively; t-values are compared against*

*the reference t-value ('ref' in the legend). Only t-values referred to the 16 optimal/explorative data are shown.*

The performance of the different methods are compared in terms of reduction of model prediction variance in the whole design space, represented by scalar measures of G-optimality (mean and maximum $J_G$ as explained in Chapter 4). The closer $J_{G,\text{mean}}$ and $J_{G,\text{max}}$ to zero, the better the performance. Therefore, the following ranking is obtained considering Figure 6.5:

- mean G-optimality $J_{G,\text{mean}}$, from the 11$^{\text{th}}$ experiment onwards (Figure 6.5a):
  MBDoE > eMBDoE ( $J_{G,\text{thr}}$=0.25) > DoE > LH, adaptative eMBDoE > eMBDoE ($J_{G,\text{thr}}$=0.75);
- maximum G-optimality $J_{G,\text{max}}$ (Figure 6.5b):
  MBDoE > eMBDoE ($J_{G,\text{thr}}$=0.25) > DoE, LH, adaptative eMBDoE > eMBDoE ($J_{G,\text{thr}}$=0.75).

The minimum G-optimality $J_{G,\text{min}}$ is equal to zero for all methods, therefore it is omitted. Therefore, profiles of both $J_{G,\text{mean}}$ and $J_{G,\text{max}}$ indicate that space exploration allows to reduce model prediction variance of Model 1 with respect to conventional MBDoE. The greatest reduction of model prediction variance with respect to MBDoE is achieved by G-map eMBDoE with $J_{G,\text{thr}}$=0.75, while the smallest one is achieved by G-map eMBDoE with $J_{G,\text{thr}}$=0.25. Although the performance of the adaptative G-map eMBDoE is not as good as the one of G-map eMBDoE with $J_{G,\text{thr}}$=0.75, it is still much better than the one of conventional MBDoE or of a G-map eMBDoE with an inappropriate threshold (here, $J_{G,\text{thr}}$=0.25). These results suggest that the use of the proposed adaptative eMBDoE is advantageous in cases where it is not possible to determine the most suitable $J_{G,\text{thr}}$ beforehand.

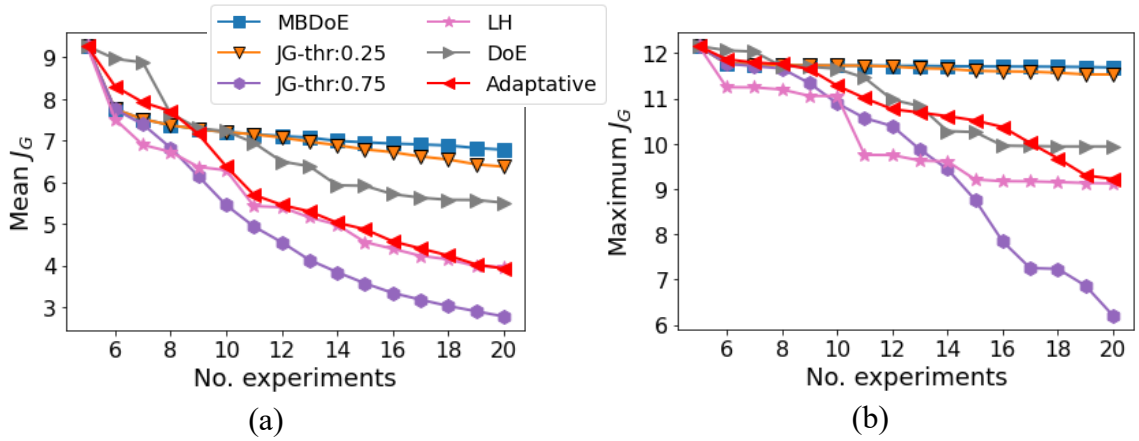
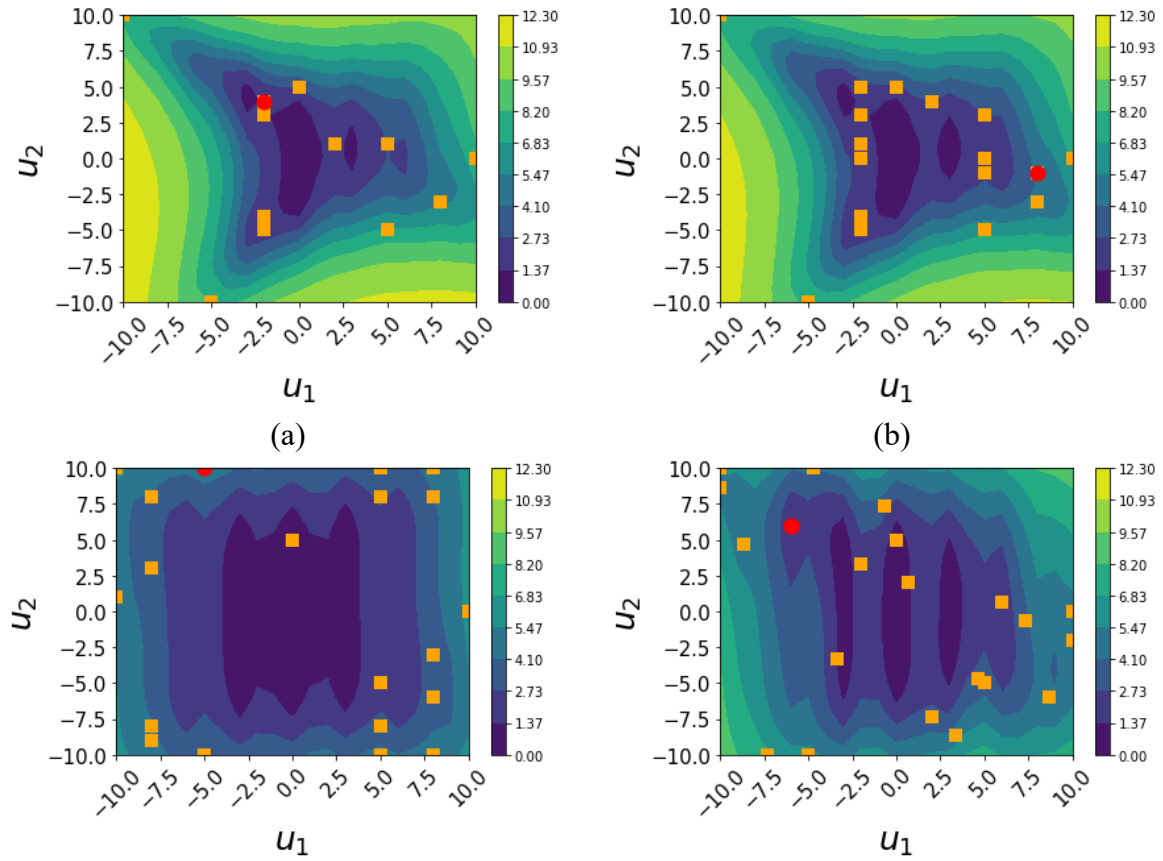
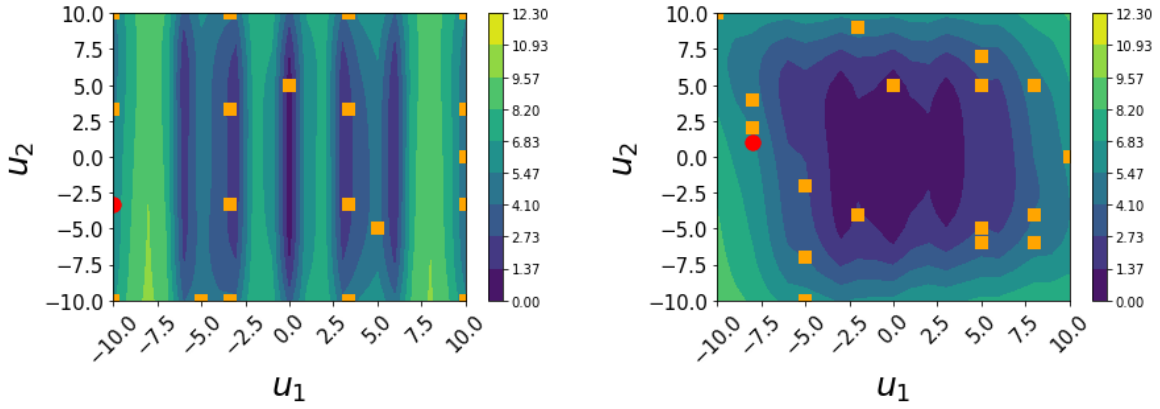



**Figure 6.5.** *Profiles of scalar indices of G-optimality including (a) mean G-optimality; (b) maximum G-optimality calculated for: MBDoE; G-map eMBDoE ($J_{G,thr}$={0.25;0.75}); Latin Hypercube (LH); full factorial DoE; adaptative G-map eMBDoE.*

Moreover, Figure 6.6 shows the G-maps obtained with 20 calibration experiments, namely the maps used to select the last point of the simulated experimental campaign. Already measured experiments are indicated as orange squares, while the red point indicates the experiment designed at the current iteration. MBDoE (Figures 6.6a) and G-map eMBDoE with $J_{G,\mathrm{thr}}$=0.25 (Figures 6.6b) have similar G-maps: the smallest model prediction variance (represented by blue regions) is found around the center of the design space, namely with $u_0$ and $u_1$ close to zero, while the highest model prediction variance (represented by yellow regions) are found at the boundary of the design space. A better performance is found with factorial DoE, although not uniformly in the design space: in fact, model prediction variance is smaller in correspondence of the levels of $u_1$ selected by the design and higher in between. A greater and more homogeneous reduction of model prediction variance is obtained with LH, adaptative G-map eMBDoE and G-map eMBDoE with $J_{G,\mathrm{thr}}$=0.75, further confirming that the adaptative eMBDoE method allows to considerably improve model predictions in the whole design space without requiring human intervention.





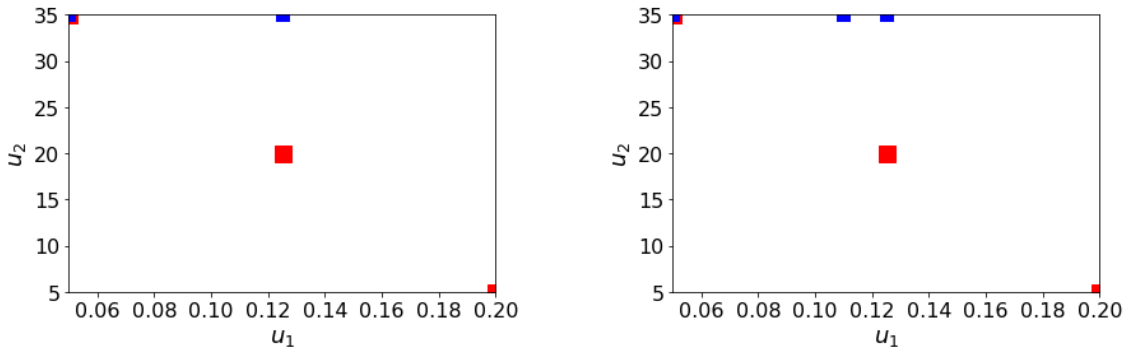

(e) (f)

**Figure 6.6**. *G-maps generated after 20 calibration experiments designed with: (a) MBDoE, (b) eMBDoE and $J_{G,thr}$=0.25; (c) eMBDoE and $J_{G,thr}$=0.75, (d) LH, (e) $4^2$ full factorial DoE; (f) adaptative G-map eMBDoE. The already measured experiments are indicated with orange squares, while the red point indicates the experiment selected at this iteration.*

The computational times to build G-maps and H-maps (with E-optimal criterion) and to design one experiment with Python 3.9 in an Intel® Core™ i7-10875H CPU, @ 2.30 GHz processor with 64.0 GB RAM are: 0.12 seconds with G-map eMBDoE; 0.12 seconds with MBDoE; 0.27 seconds with adaptative G-map eMBDoE. Therefore, the calculation of the best G-optimality constraint trough the overlap of G-maps and H-maps increases the computational time with respect to the original G-map eMBDoE method, but overall the time required is negligible.

### 6.3.2 Model 2

As in Section 4.3.2, the same 3 LH preliminary experiments are used for all methods and then 20 experiments are designed, thus reaching a total experimental burden of $N_e$=23 experiments.




(a) (b)

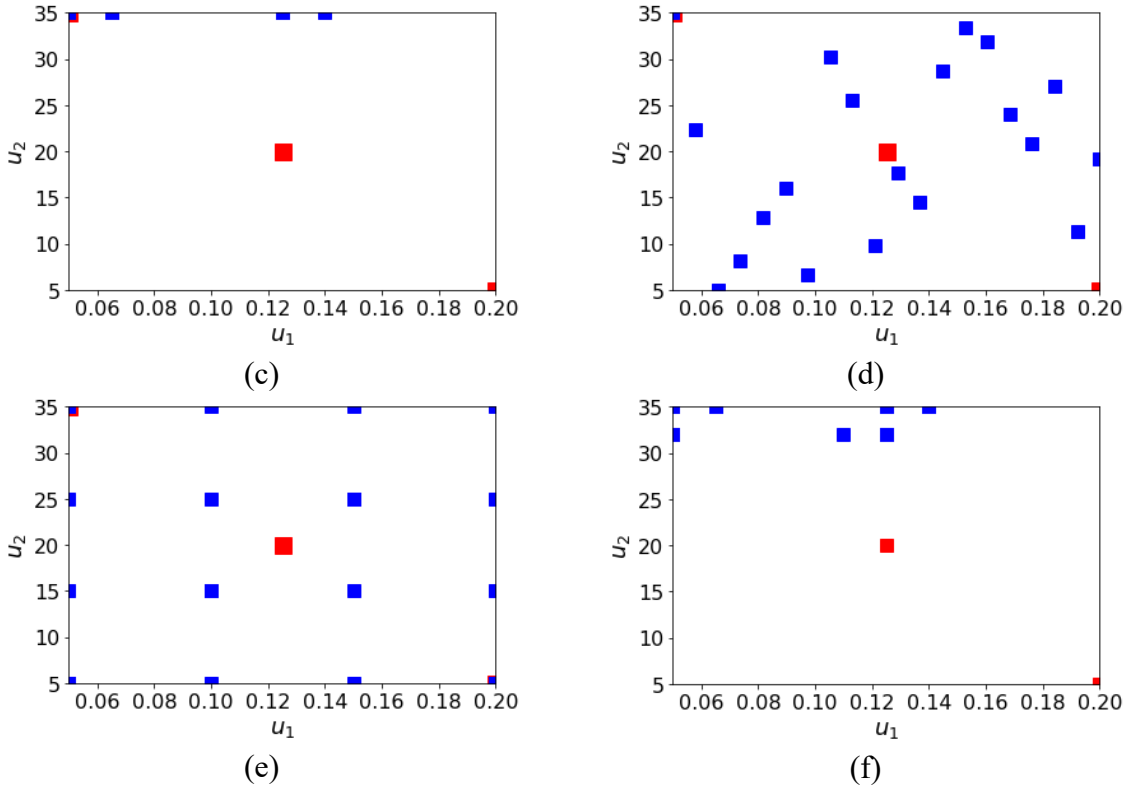



**Figure 6.7**. *Designs: (a) MBDoE; (b) eMBDoE, $J_{G,thr}$=0.25; (c) eMBDoE, $J_{G,thr}$=0.75; (d) LH; (e) factorial DoE; (f) adaptative.*

**Table 6.3.** *Number of distinct design points for every scenario compared in the study.*

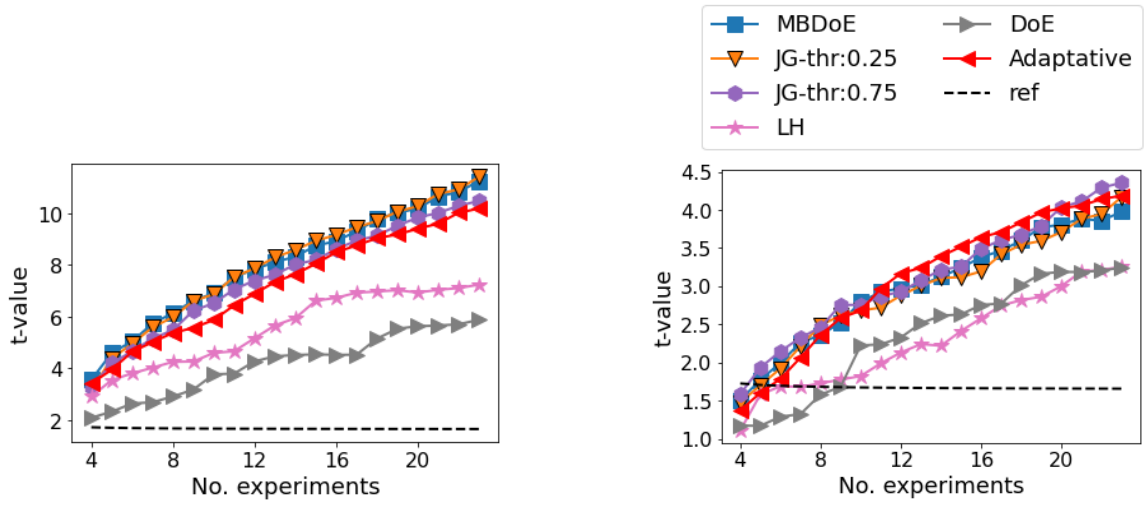
| Scenario | No. distinct design points |
|---|---|
| MBDoE | 2 |
| eMBDoE, thr:0.25 | 3 |
| eMBDoE, thr:0.75 | 4 |
| LH | 16 |
| DoE | 16 |
| Adaptative eMBDoE | 7 |

The experiments designed for Model 2 by MBDoE (Figure 6.7a) are concentrated in two distinct points: $(u_1, u_2)$=(0.05 h$^{-1}$, 35 g/L) and $(u_1, u_2)$=(0.125 h$^{-1}$, 35 g/L). The methods looking for a trade-off between space exploration and information maximisation, namely G-map eMBDoE with $J_{G,\text{thr}}$=0.25 (Figure 6.7b) and with $J_{G,\text{thr}}$=0.75 (Figure 6.7c) and adaptative G-map eMBDoE (Figure 6.7f), select experiments that are clustered around the two optimal conditions selected by MBDoE. Among the three G-map eMBDoE methods, the adaptative one selects the highest number of distinct design points (Table 6.3): 7 distinct points, thus more than 4 and 3 selected by G-map eMBDoE with $J_{G,\text{thr}}$=0.75 and $J_{G,\text{thr}}$=0.25, respectively. Moreover, as shown in Table 6.4, 13 out of 20 designed experiments have the "≤" inequality type and all thresholds are between 0.50 and 0.90.

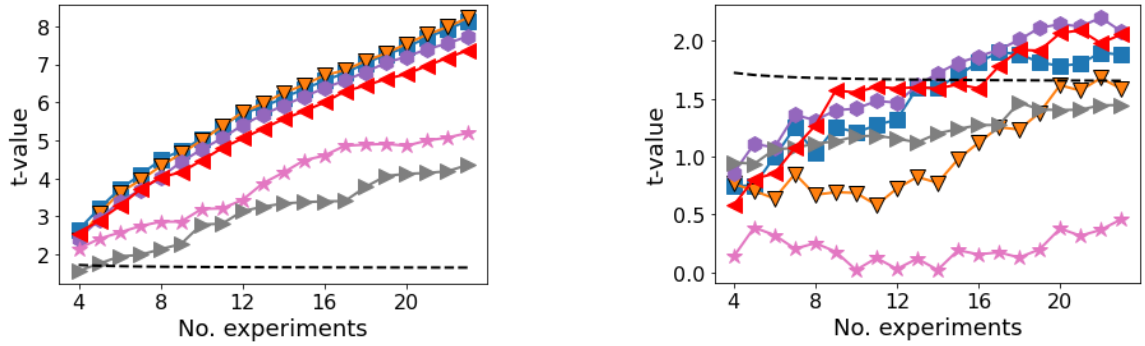
**Table 6.4.** *Constraint selected by the adaptative G-map eMBDoE. "L" stands for "lower" and indicates the inequality ≤; "H" stands for "higher" and indicates the inequality ≥.*

| Inequality type | $J_{G,thr}$ |
|---|---|
| L | 0.89 |
| H | 0.76 |
| L | 0.50 |
| L | 0.76 |
| H | 0.89 |
| H | 0.72 |
| H | 0.82 |
| H | 0.72 |
| H | 0.65 |
| L | 0.50 |
| H | 0.90 |
| L | 0.57 |
| L | 0.84 |
| L | 0.83 |
| L | 0.81 |
| L | 0.81 |
| L | 0.51 |
| L | 0.79 |
| L | 0.50 |
| L | 0.78 |

Parameters precision is evaluated in terms of *t*-tests in Figure 6.8. The most critical parameter is $\hat{\theta}_4$, since a higher number of experiments is required to precisely estimate it. In fact, completely exploratory procedures such as factorial DoE and LH are not able to estimate it within the experimental budget. On the other side, state-of-the art E-optimal MBDoE requires 12 optimal experiments to estimate $\hat{\theta}_4$ (Figure 6.7d), while G-map eMBDoE with $J_{G,\text{thr}}$=0.75, requires 11 experiments. The adaptative G-map eMBDoE is able to estimate $\hat{\theta}_4$ with 14 optimal experiments, therefore the experimental burden is not considerably increased with respect to the previous two methods and it is reduced with respect to G-map eMBDoE with an unfavourable threshold such as $J_{G,\text{thr}}$=0.25.



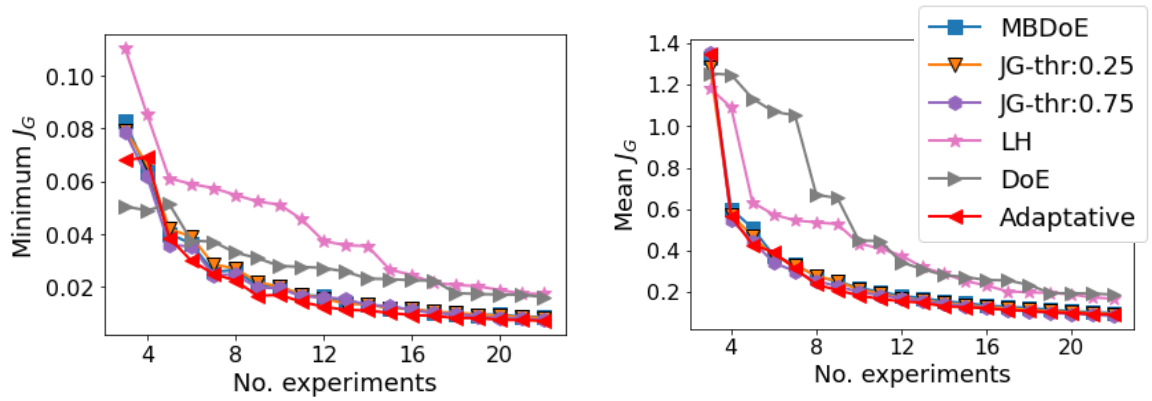

(a) (b)



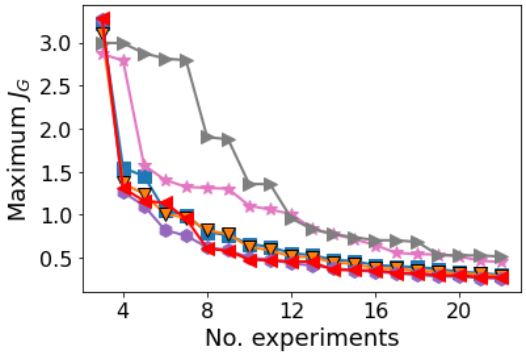

(c) (d)

**Figure 6.8.** *Profiles of t-values calculated with: MBDoE; G-map eMBDoE ($J_{G,thr}$={0.25; 0.75}); Latin Hypercube (LH); $4^2$ full factorial DoE; adaptative G-map eMBDoE. The t-values are compared against the reference t-value ('ref' in the legend) for parameters 1-4 in figures (a)-(d), respectively.*

Then, the different design methods are compared in terms of model prediction variance. As shown in Figure 6.9, minimum, mean and maximum G-optimality calculated in the whole design space decrease with MBDoE and G-map eMBDoE methods with respect to the values found with completely explorative methods like factorial DoE and LH. Moreover, $J_{G,\text{mean}}$ and $J_{G,\text{max}}$ are smaller with adaptative G-map eMBDoE and G-map eMBDoE with $J_{G,\text{thr}}$=0.75 with respect to the values obtained with MBDoE and G-map eMBDoE with $J_{G,\text{thr}}$=0.25. This suggests that the adaptative method proposed in this work is able to lead to a minimisation of model prediction variance in the whole design space without the need to manually select the most convenient threshold of G-optimality.



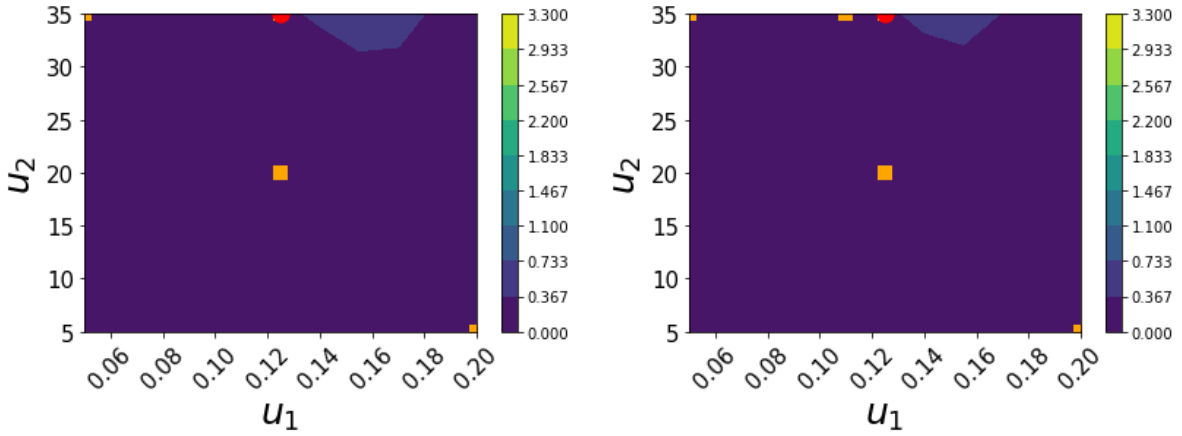

(a) (b)




(c)

**Figure 6.9.** *Profiles of scalar indices of G-optimality calculated with: MBDoE; G-map eMBDoE ($J_{G,thr}$={0.25;0.75}); Latin Hypercube (LH); $4^2$ full factorial DoE; adaptative G-map eMBDoE. Three different scalar measures are shown: (a) minimum G-optimality; (b) mean G-optimality; (c) maximum G-optimality.*

As in Chapter 4, G-maps at after 15 calibration experiments are visualised (Figure 6.10), since it is the first iteration where G-map eMBDoE with $J_{G,\text{thr}}$=0.75 is able to reduce completely model prediction variance in the entire design space (darkest blue in the whole design space, Figure 6.10a). The only other method that is able to reduce completely model prediction variance in the entire design space is the adaptative G-map eMBDoE proposed in this work (Figure 6.10f), while all the others have a higher model prediction variance at $u_2$ close to 35g/L. Therefore, the performance of the adaptable G-map eMBDoE in terms of model prediction variance is equivalent to the one of G-map eMBDoE with the most favorable threshold (namely, $J_{G,\text{thr}}$=0.75).




(a) (b)

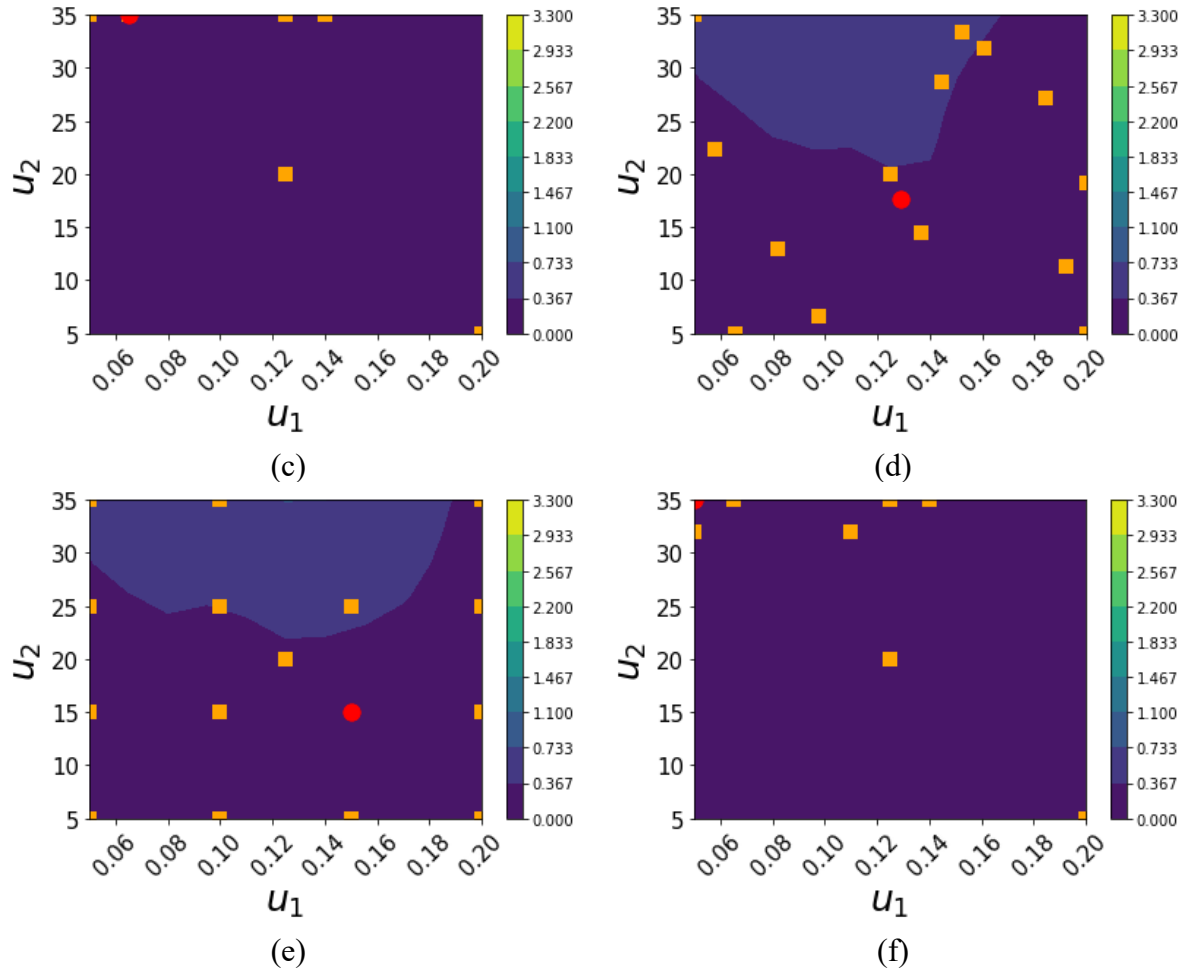



**Figure 6.10.** *G-maps generated after 15 calibration experiments. Four methods are compared: (a) MBDoE; (b) G-map eMBDoE with $J_{G,thr}$=0.25;(c) G-map eMBDoE with $J_{G,thr}$=0.75; (d) LH; (e) $4^2$ full factorial DoE; (f) adaptative G-map eMBDoE. Orange squares indicate data already used to calibrate the model, while the red point indicates the experiment designed at the current iteration.*

The computational times to build G-maps and H-maps (with E-optimal criterion) and to design one experiment with Python 3.9 in an Intel® Core™ i7-10875H CPU, @ 2.30 GHz processor with 64.0 GB RAM are: 0.65 seconds with G-map eMBDoE; 0.65 seconds with MBDoE; 0.85 seconds with adaptative G-map eMBDoE. Therefore, similarly to Model 1 (Section 6.3.1), the computational time required by the adaptative G-map eMBDoE is higher than the one required by MBDoE and G-map eMBDoE, but it is still negligible.

## 6.4 Conclusions and future work

Model-based design of experiments methods allow to maximise the information content of the experiments, therefore the modelling objectives such as discrimination among model equations or estimation of precise model parameters can be achieved with a reduced experimental burden. However, MBDoE aims only at optimising information content of the experiments and this may

lead to a scarce exploration of the design space; in turn, this may result in regions of the design space characterised by high model prediction variance. The G-map eMBDoE method proposed by Cenci et al. (2023; Chapter 4) allows to find a trade-off between space exploration and information maximisation thanks to a mapping of G-optimality (a measure of model prediction variance): first, G-optimality is calculated for every possible condition in the design space; then, candidate design points are selected based on a threshold on G-optimality that is selected by the user at the beginning and kept fixed throughout the experimental campaign; finally, the experimental condition having the highest information content among the candidates is selected as the experiment to be performed. The G-map eMBDoE method reduces model prediction variance and parameters uncertainty with respect to conventional information-based methods, such as MBDoE, and exploration-based methods, such as LH and factorial DoE, but the quality of the results depends on the selection of the G-optimality threshold selected by the user. In this work, a novel G-map eMBDoE method is proposed in order to automatically select the most suitable constraint on G-optimality for the system under study. The main rationale is that space exploration can be enhanced with respect to state-of-the-art MBDoE by considering the overlap between maps of information content, H-maps, and maps of model prediction variance, G-maps. For instance, if the regions of the design space having the highest information content correspond to the regions with the highest model prediction variance, experiments design should be moved toward regions with lower model prediction variance. On the other side, if the points having the highest information content have the lowest model prediction variance, space exploration can be favored by selecting points with a higher model prediction variance. The analysis of the overlap between H-maps and G-maps allows to adapt the G-optimality constraint at every iteration, therefore the overall design method is referred to as *adaptative G-map eMBDoE*. To directly compare the performance of the adaptative G-map eMBDoE with respect to the original G-map eMBDoE with fixed threshold, the new method is applied to Model 1 and Model 2 of Cenci et al. (2023) and compared to the results previously obtained with MBDoE, LH, factorial DoE and G-map eMBDoE with two different thresholds, namely $J_{G,\mathrm{thr}}$=0.25 and $J_{G,\mathrm{thr}}$=0.75. Results show that the proposed adaptative G-map eMBDoE provides a satisfactory level of space exploration, which may be less evident than the one obtained with $J_{G,\mathrm{thr}}$=0.75 but still enhanced with respect to the one with $J_{G,\mathrm{thr}}$=0.25. Moreover, both inequality types, namely $J_G \leq J_{G,\mathrm{thr}} J_{G,\mathrm{max}}$ and $J_G \geq J_{G,\mathrm{thr}} J_{G,\mathrm{max}}$, are selected by the adaptative G-map eMBDoE, proving that there was no contradiction in the use of different inequality types in Chapter 4 and 5 since both can contribute to enhance space exploration. Parameters precision is achieved with

a number of experiments which is close to the one required by state-of-the-art MBDoE, thus reducing the experimental burden with respect to $J_{G,\mathrm{thr}}$=0.25, LH and factorial DoE. Moreover, the proposed adaptative G-map eMBDoE has a smaller reduction of model prediction variance with respect to $J_{G,\mathrm{thr}}$=0.75 with Model 1, but the reduction is comparable to the one of $J_{G,\mathrm{thr}}$=0.75 with Model 2 and it is always more marked than the one of MBDoE or G-map eMBDoE with $J_{G,\mathrm{thr}}$=0.25. Considering that the advantage of fixing $J_{G,\mathrm{thr}}$=0.75 with respect to fixing $J_{G,\mathrm{thr}}$=0.25 is case-dependent, namely it depends on the system under study, and that it cannot be predicted beforehand, the proposed adaptative procedure provides satisfactory results in terms of trade-off between space exploration and information maximisation without requiring human intervention. For this reason, the adaptative G-map eMBDoE is suitable for integration into automated chemical platforms used in the pharmaceutical industry, regardless of the specific mathematical model used. To prove this statement, ongoing work is focusing on the implementation of this method in the automated chemical platform for total methane oxidation (the same used in Chapter 5), with the aim of achieving autonomous operation of the platform from the beginning to the end of the experimental campaign.

# Chapter 7
# Prediction of drug solubility in organic solvent mixtures through machine-learning on group contributions[5]

In this work, a PLS model is proposed to make solubility predictions knowing: temperature, mixture composition before API dissolution and solvents structure in the form of UNIFAC (UNIQUAC Functional-group Activity Coefficients) subgroups. The PLS model is tested with experimental data of a real drug substance and 14 organic solvents typically employed for crystallisation. Model predictions are accurate and precise with single solvents, binary mixtures and ternary mixtures at different compositions and temperature: $R^2$ is equal to 0.92 and 0.90 with calibration and validation data, respectively.The adequacy of the modelling approach proposed is confirmed by the satisfactory results obtained with 9 literature datasets on organic solubility of drug and drug-like compounds: the majority of validation and calibration data have $R^2$ between 0.95 and 0.99.

## 7.1 Introduction

Crystallisation is a critical operation in the pharmaceutical industry, since more than 90% of Active Pharmaceutical Ingredients (APIs) are synthesised as crystalline products (Orehek et al., 2021) and crystallisation is used to separate and purify them. Crystallisation has a direct or indirect effect on the manufacturability and quality of the product (Orehek et al., 2021; Lemmer and Liebenberg, 2023): for instance, drug crystalline properties have an impact on downstream processes, such as filtration, drying and dissolution testing (Lemmer and Liebenberg, 2023). Moreover, solution crystallisation determines the solid-state modification of the API (Miller et al., 2007), which impacts on the final product performance, such as solubility, dissolution rates

[5] Cenci, F., Diab, S., Harabajiu, K., Ferrini, P., Barolo, M., Bezzo, F. and Facco, P.. Machine-Learning approach based on group contributions for the prediction of solubility of drug and drug-like molecules in organic solvent mixtures [in preparation].

and tablet hardness (Gao et al., 2017). Therefore, proper design and operation of the crystallisation units is crucial.

One of the most important properties for the production and purification of APIs by crystallisation is solubility (Ruether and Sadowski, 2009; Bouillot et al., 2011). In fact, solvents selection relies on the knowledge of pharmaceuticals solubility in multiple organic solvents and/or solvents mixtures (Ruether and Sadowski, 2009; Papadakis et al., 2016; Ye and Ouyang, 2021). Moreover, solid solubility in the crystallisation solvent is a key parameter for process design purposes: it determines the crystallisation configuration (cooling, evaporation, antisolvent) and the amount of cooling for a desired product yield (Bouillot et al., 2011).

Due to the high number of potential solvents and their mixtures, it would be unfeasible to experimentally measure solubility in all of them (Cysewski et al., 2022). In addition, experimental solvents screening is hampered by the fact that solubility experiments are generally laborious and costly (Ruether and Sadowski, 2009; Ye and Ouyang, 2021) and API availability may be limited at the early stages of drug development (Bouillot et al., 2011). For this reason, the experimental screening should be supported by a theoretical screening (Cysewski et al., 2022); in other words, a model built with a small amount of experimental data should allow for solubility predictions in solvents systems not explored experimentally (Ruether and Sadowski, 2009).

However, solubility prediction is still an open challenge (Boobier et al., 2020; Ye and Ouyang, 2021). Different approaches have been proposed in literature to predict solid-liquid equilibrium (SLE) and they can be grouped into two main categories: thermodynamic models and data-driven (or "black-box" or "empirical") models (Ruether and Sadowski, 2009; Gharagheizi et al., 2011).

For instance, thermodynamic models based on activity coefficients define solid-liquid equilibrium as the equality of the chemical potential of the solute in both the solid phase and the liquid phase. Then, the chemical potential of the solute in the saturated liquid is expressed in terms of activity coefficients, which in turn are calculated in different ways based on the thermodynamic model employed. A detailed explanation of these methods can be found in Elliott and Lira (2012). In general, thermodynamic models have the advantage of requiring a smaller amount of data to identify model parameters than data-driven ones (Ruether and Sadowski, 2009), but the choice of the best model for SLE is not trivial since they have been developed and used mainly for vapor-liquid or liquid-liquid equilibria (Bouillot et al., 2011). Bouillot et al. (2011) compared predictions accuracy of different thermodynamic models,

namely UNIFAC, UNIFAC mod., COSMO-SAC and NRTL-SAC. They employed five drug or drug-like molecules containing common functional groups (alcohols, ketones, amines), Ibuprofen, Acetaminophen, Benzoic acid, Salicylic acid and 4-aminobenzoic acid, together with a simple molecule, anthracene. The best results were obtained with UNIFAC and NRTL-SAC, which were able to catch the correct order of magnitude of solubility. However, predictions were not accurate enough to obtain quantitative results. Moreover, the modified Apelblat equation (Apelblat et al., 1997; Apelblat et al., 1999) and the λh equation (Buchowski, 1980) have been employed to correlate solubility experiments of several drug and drug-like molecules in organic solvents (Li et al., 2020; Wu et al, 2020; Hu et al., 2021; Zhou et al., 2021; Zhang et al., 2019; Huang et al., 2015; Wang et al, 2021). They are semi-empirical thermodynamic models that allow to accurately represent the dependence of solubility on temperature. However, neither the modified Apelblat equation nor the λh equation include molecular descriptors among the input variables, therefore the effect of different solvents cannot be represented by means of the same model and the model must be re-calibrated for every different solvent type. Furthermore, such models do not explicitly represent the dependence of solubility on the composition of the organic solvents mixture, therefore the model must be re-calibrated if, for instance, the same two organic solvents are mixed in different proportions.

Data-driven models relate drug solubility to physico-chemical properties and/or molecular structure of drugs and solvents. The theory-based quantitative structure property relationship (QSPR) models are well-known empirical models, which relate the property of interest (e.g., solubility) to molecular descriptors like topological and geometric indices, quantum-mechanical and thermodynamic quantities or group contributions. Moreover, QSPR models can relate input and output variables through either multivariate linear regression models, such as multivariate linear regression (MLR) and partial least squares (PLS), or nonlinear models, such as artificial neural network (Borhani et al., 2019). For instance, Duchowicz and Castro (2009) reviewed different QSPR models developed for the prediction of aqueous solubility of drug-like compounds, with special attention on linear approaches which have the advantage of limiting the over-fitting of data and of facilitating the interpretation of possible cause/effect relationships with respect to non-linear approaches. Moreover, Enciso et al. (2016) developed an open-source software with linear QSPR models for the prediction of three key properties for drug development: solubility in water, Caco-2 cell permeability and brain-blood barrier permeation. However, there are still some limitations in the adoption of QSPR to model solid-

liquid equilibria. First of all, the majority of studies consider solubility in water (Boobier et al., 2020; Ye and Ouyang, 2021; Cysewski et al., 2022), while few models have been developed for solubility in organic solvents. For instance, Balakin et al. (2004) considered only one organic solvent, DMSO, widely used in the pharmaceutical industry for its high solvent power and relatively low chemical reactivity and toxicity. They considered solubility data of diverse drug-like compounds and developed a neural network model to classify new compounds as well soluble (+) or poorly soluble (-) in DMSO. Moreover, Boobier et al. (2020) aimed at developing a model for the prediction of organic solubility of new compounds; to do so, they compared different machine learning models calibrated with diverse compounds. However, due to the limited availability of organic solubility data in literature, they restricted the study to solubility data of neutral solutes in three single solvents: ethanol, benzene, acetone. Benzene is included besides its limited use in modern chemistry due to the adequate amount of available data. A larger amount of organic solvents was included in the work of Yu and Ouyang (2021), where different machine learning models, including light gradient boosting machine (lightGBM) and deep neural networks, were developed to predict the organic solubility at different temperatures. Literature data on 266 compounds in 123 organic solvents (single solvents) were employed for the purpose. This work contributed to the free web server FormulationAI (www.formulationai.computpharm.org), where a user-friendly interface allows to specify the compound of interest in the form of SMILES string or chemical formula and the software containing lightGBM and RandomForest algorithms predicts the compound solubility in 27 different organic solvents. Another relevant example is the modelling approach proposed by Vermeire et al. (2022), where thermodynamic and machine learning models are combined to predict solubility of neutral solutes in different organic solvents and at different temperatures. This model was calibrated with three large solubility databases and lead to the development of a web service for the prediction of solubility for new compounds. The user must specify only three pieces of information, namely temperature and SMILES string or InChI of solutes and solvent, and solubility is predicted by the software. Besides these remarkable advances, such approaches based on single solvents are not exhaustive for crystallisation design and operation, where binary or ternary mixtures of organic solvents are typically present either as the main solvent system or after antisolvent addition. Some improvements in this regard were made by Cysewski et al (2022), who considered drug solubility in aqueous binary mixtures of 4-formylmorpholine, DMSO and DMF, and by Przybyłek et al. (2021), who included aqueous binary mixtures of acetonitrile, 1,4-dioxane, DMF, DMSO, methanol, but more organic

mixtures should be included to have a comprehensive understanding of organic drug solubility. Additional limitations of QSPR are related to modelling aspects. In many cases, the dependence of solubility on temperature is not explicitly defined and the range of applicability of the model is not clearly specified, making generalisation more difficult (Cysewski et al., 2022). Moreover, a large number of descriptors (>100) is available to build QSPR models, thus complicating model interpretability and the selection and/or decorrelation of regressors for the system of interest (Chinta and Rengaswamy, 2019; Boobier et al., 2020; Cysewski et al., 2022). Moreover, complex modelling approaches, such as deep machine learning models, have been recently employed, but they require a considerable number of data thus limiting their applicability to solvents and solvents mixtures not already explored (Cysewski et al., 2022).

In this Chapter, we propose a machine-learning approach to predict the solubility of drug and drug-like molecules in complex mixtures of organic solvents commonly employed for crystallisation design and operation. The input variables (or regressors) of the model explicitly include temperature and do not rely on the experts' knowledge, but are inspired to the UNIFAC theory: the solvents composing the mixture are represented in terms of functional groups as in Gmehling et al. (1978). Model regressors include both functional groups taken individually and interactions between pairs of functional groups. A PLS model is used to automatically handle correlation among regressors; moreover, PLS is a linear model, therefore less prone to over-fitting (Duchowicz and Castro, 2009). The model is tested with solubility data of real drug substance measured in 14 organic solvents; single solvents, binary and ternary mixtures are measured at different temperatures with a high-throughput technology employing 96-vials plates. The adequacy of the modelling approach proposed is further tested with 9 literature datasets on solubility of drug and drug-like molecules in organic solvents. More details on materials, experimental setup, literature datasets and modelling approach can be found in section 2. Results obtained with experimental and literature data are shown in section 3, while conclusions and future work are explained in section 4.

## 7.2 Materials and methods

This section illustrates experimental and modelling strategies adopted to predict drug and drug-like molecules solubility. First, the experimental setup to generate solubility experiments is explained in Subsection 7.2.1, by specifying the materials used and the experimental protocol employed. Then, Subsection 7.2.2 focuses on mathematical modelling and explains the novel

modelling approach proposed, together with the indices calculated to evaluate model performance with calibration and validation data.

### *7.2.1 Experimental setup*

In this section, the materials employed are described (Subsection 7.2.1.1), as well as the experimental procedures and the equipment used to perform the experiments (Subsections 7.2.1.2-7.2.1.4).

#### 7.2.1.1 Materials

The solute for solubility experiments is a real drug substance produced in house with a molecular weight of 823.18 g/mol and a purity of 98.3% by w/w. It will be named API #1 from now on.

The following solvents (commercially available) are used in the solubility screens without additional purification or analysis: 1-butanol (99.8% purity, Sigma Aldrich, USA), n-propanol (≥99.9%, Honeywell), 2MeTHF (BioRenewable, anhydrous ≥99.0%, contains 250 ppm BHT as inhibitor, Sigma Aldrich, USA), isopropanol (ACS reagent, ≥99.8% (GC), Sigma Aldrich, France), 3-pentanone (99%, Alfa Aeser, Germany), acetonitrile (for HPLC, gradient grade, ≥99.9%, Sigma Aldrich, France), cyclohexane (Chromasolv® for HPLC, ≥99.7%, Sigma Aldrich, Israel), ethanol (absolute, ≥99.8% (GC), Sigma Aldrich, UK), ethyl acetate (≥99.7%, Sigma Aldrich), GVL (ReagentPlus® 99%, Sigma Aldrich, China), isopropyl acetate (≥99.6%, Sigma Aldrich, USA), methanol (LChrosolv® for LC-MS hypergrade, Sigma Aldrich, Germany), propionitrile (99%, Sigma Aldrich), n-butyl acetate (anhydrous ≥99%, Sigma Aldrich, USA). Two datasets, one for calibration and one for validation, are generated with a high-throughput technology employing 96-vials plates; more details on the experimental procedures are found in Subsections 7.2.1.2-7.2.1.4. All measurements are replicated, therefore 48 different experimental conditions are available in one plate and the two replicates should not be split into calibration and validation data. With the abovementioned 14 organic solvents, one plate is filled using all single solvents and 34 binary mixtures to have calibration experiments. Solvent types are chosen to form binary mixtures as explained in Subsection 7.2.2.3; as in the industrial practice, the same plate is measured at two temperatures, namely 20 and 40°C. Some vials do not provide valid measurements, therefore 176 calibration experiments are obtained. Instead, validation data are chosen in order to test the model in new experimental conditions, such as binary mixtures of different compositions and/or solvent types, ternary mixtures,

experiments at a higher temperature, 50°C. The main features of the two datasets are summarized in Table 7.1. Notice that measurements of single solvents at 20 and 40°C have been repeated in validation and this is useful to assess inter-plate variability; moreover, among the 14 solvents, only cyclohexane is not included in calibration as single solvent because of non-valid measurements, but binary mixtures containing this solvent are present. If the standard deviation $\sigma_y$ of the measured solubility (as molar fraction) is calculated for every pair of replicated measurements, an average $\sigma_y$ equal to $5.73 \cdot 10^{-5}$ is found; if the same experimental condition is considered (repeated experimental conditions including measurements in different plates), an average $\sigma_y$ of $7.20 \cdot 10^{-5}$ is found. Especially if compared to the overall range of measured solubility, namely 0.0077, the difference between the two $\sigma_y$ suggests that the inter-plate variability is just a very limited fraction of the overall variability.

**Table 7.1.** *Description of the datasets generated in this work using API #1 and 14 organic solvents. "No. cal." stands for "number of calibration experiments"; "No. val." stands for "number of validation experiments"*

| Solute | Solvents | T[°C] | No. cal. | No. val. |
|---|---|---|---|---|
| API #1 | 1-butanol, n-propanol, 2MeTHF, isopropanol, 3-pentanone, acetonitrile, cyclohexane, ethanol, ethyl acetate, GVL, isopropyl acetate, methanol, propionitrile, n-butyl acetate | 20, 40, 50 | **Single solvents**: 26 (at 20°C) + 26 (at 40°C) = 52<br>**Binary mixtures**: 64 (at 20°C) + 60 (at 50°C) = 124<br>**Ternary mixtures**: 0<br>**Total**: 176 | **Single solvents**: 56 (at 20°C)+24 (at 40°C)+28 (at 50°C) =108<br>**Binary mixtures**: 69 (at 20°C)+8 (at 40°C)+61 (at 50°C) =138<br>**Ternary mixtures**: 22 (at 20°C)+14 (at 40°C)+6 (at 50°C) =42<br>**Total**: 288 |

Further assessment of the adequacy of the modelling approach proposed is performed with literature data on solubility of drug and drug-like molecules in organic solvents commonly employed in crystallization units are considered. The retrieved datasets include the following solutes: 1) N,N-Dibenzylhydroxylamine (DBHA), which is mainly used in polymers to inhibit inhibiting aging and degradation (Li et al., 2020), but is included in this study since arylamines constitute the core structure of many therapeutic agents (Svejstrup et al., 2017); 2) Fenofibrate, which is a hypolipidemic medication administered to patients with hypertriglyceridemia or type 2 diabetes (Sadeghi Rasmuson, 2020; Jung et al., 2018); 3) Benorilate, which has anti-inflammatory, analgesic and antipyretic properties (Wu et al., 2020); 4) L-Arginine L-pyroglutamate, which has positive effects on the immune capacity of humans and animals (Hu et al., 2021); 5) 2-chloro-4-amino-6,7-dimethoxyquinazoline, which is one of the key intermediates in the production on some cardiovascular drugs (Zhou et al., 2021); 6) Tetramethylpyrazine, which has physiological activity on cardiovascular and cerebrovascular

diseases (Zhang et al., 2019) 7) Coumarin, whose derivatives have several therapeutic applications, from antitumor and anti-HIV therapies to the production of stimulants for central nervous system, antibacterials, anti-inflammatory and anti-coagulants (Huang et al., 2015; Musa et al., 2008); 8) 1,3,5-Tris(1-phenyl-1H-benzimidazol-2-yl)benzene (TPBi), which is used as organic semiconductor (Wang et al., 2021); 9) Nicotinamide, used to treat diabetes mellitus, stroke, bullous pemphigoid, and psoriasis vulgaris (Khajir et al., 2024).

**Table 7.2.** *Description of the datasets retrieved from Krasnov et al., (2022)*

| Solute | Solvents | T[°C] | No. exp. | No. cal. | No. val. |
|---|---|---|---|---|---|
| 1) Dibenzylhydroxylamine<br>2) $C_{14}H_{15}NO$<br>3) 621-07-8 | 1-Butanol, **Acetone,** Acetonitrile, **DCM,** Ethanol, Ethyl acetate, Isopropanol, Methanol, **Toluene,** n-Propanol | approx. 0, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50 | 108 | 21 | 3 |
| 1) Fenofibrate<br>2) $C_{20}H_{21}ClO_4$<br>3) 49562-28-9 | **Acetone,** Acetonitrile, Ethanol, Ethyl acetate, Methanol | −30,−25, −20,−15, −10,−5, 0, 5, 10, 15, 20, 25, 30, 35 | 42 | 14 | 5 |
| 1) Benorilate<br>2) $C_{17}H_{15}NO_5$<br>3) 5003-48-5 | 1-Butanol, Acetonitrile, Ethanol, Ethyl acetate, Isopropanol, Methanol, **Toluene, Water,** n-Propanol | 5, 10, 15, 20, 25, 30, 35, 40, 45 | 99 | 19 | 3 |
| 1) Arginine PCA<br>2) $C_{11}H_{21}N_5O_5$<br>3) 56265-06-6 | **Acetone,** Acetonitrile, **DCM,** Ethanol, Ethyl acetate, Isopropanol, Methanol, **Water** | 10, 15, 20, 25, 30, 35, 40, 45 | 40 | 19 | 3 |
| 1) Doxazosin Related Compound C<br>2) $C_{10}H_{10}ClN_3O_2$<br>3) 23680-84-4 | 1-Butanol, **DMF,** Ethanol, Ethyl acetate, Isopropanol, Methanol, **Toluene,** n-Propanol | 0, 5, 10, 15, 20, 25, 30, 35, 40 | 72 | 18 | 4 |
| 1) Tetramethylpyrazine<br>2) $C_8H_{12}N_2$<br>3) 1124-11-4 | 1-Butanol, **Acetone,** Acetonitrile, Ethanol, Ethyl acetate, Isopropanol, Methanol, n-Propanol | 0, 5, 10, 15, 20, 25, 30, 35, 40 | 72 | 17 | 3 |
| 1) Coumarin<br>2) $C_9H_6O_2$<br>3) 91-64-5 | Cyclohexane, Ethanol, Isopropanol, Methanol, n-Propanol | 5, 10, 15, 20, 25, 30, 35, 40, 45 | 44 | 11 | 3 |
| 1) TPBi<br>2) $C_{45}H_{30}N_6$<br>3) 192198-85-9 | 1-Butanol, Acetonitrile, Cyclohexane, **DMSO,** Ethanol, Ethyl acetate, Isopropanol, Methanol | 10, 20, 30, 40, 50 | 44 | 18 | 3 |

Notice that they are all related to pharmaceutical applications, except for TPBi which is nevertheless included because its complex molecular structure is useful to further test the

modelling approach proposed. The first 8 datasets are retrieved from BigSolDB solubility dataset of Krasnov et al. (2022), while the last one is retrieved from Khajir et al. (2024). Their chemical formula is shown in Figures G.2 and G.3, respectively, of Appendix G. The first 8 datasets (Table 7.2) are made of solubility values in single solvents only; the majority of data points considered employ the same solvents used in the new generated data, namely 1-butanol, acetonitrile, cyclohexane, ethanol, ethyl acetate, isopropanol, methanol, n-propanol. Additional solvents that may be used for crystallization design are considered: water, dimethyl sulfoxide (DMSO), toluene, acetone, dimethylformamide (DMF) and dichloromethane (DCM). In this work, literature datasets are divided into calibration and validation datasets (Table 7.2): the calibration dataset is made of one measurement for every solvent at minimum and maximum temperature. Only calibration dataset 7 (Coumarin) also includes other two temperature values foe every solvent in order to have smaller 95%CIs and this is likely needed due to the smaller set of regressors available with this dataset (results are shown in section 7.3).

As regards the ninth literature dataset (Table 7.2), it includes solubility measurements in ethanol and acetonitrile, both as single solvents and as binary mixtures. These solvent systems were selected by Khajir et al. (2024) because they are frequently employed in extractions and high-pressure liquid chromatography, thus they are involved in crystallization procedures. Different molar fractions of ethanol are considered in the binary mixtures before solute dissolution, namely $x_{\text{ethanol}} = \{0, 0.10, 0.20, 0.40, 0.50, 0.60, 0.7, 0.8, 0.9, 1.0\}$, and every level of $x_{\text{ethanol}}$ is measured at 5 temperature values, namely $T = \{20.05, 25.05, 30.05, 35.05, 40.05\}$°C. The standard deviation of such experimental measurements ranges between 0.01 and 0.38; the mean value of the replicates at a given experimental condition is used to test the prediction accuracy of the model proposed (see section 7.3). For comparison purposes, this dataset is split into calibration and validation data as in Khajir et al. (2024): data of single solvents and of binary mixtures with $x_{\text{ethanol}} = \{0.30, 0.50, 0.7\}$ measured at minimum and maximum temperatures are used to calibrate the model, the rest is used to validate it.

**Table 7.3.** *Description of the datasets retrieved from Khajir et al. (2024)*

| Solute | Solvents | T[°C] | No. exp. | No. cal. | no. val. |
|---|---|---|---|---|---|
| 1) Nicotinamide<br>2) $C_6H_6N_2O$<br>3) 98-92-0 | Ethanol, Acetonitrile, Ethanol + Acetonitrile | 20.05, 25.05, 30.05, 35.05, 40.05 | 55 | 10 | 45 |

## 7.2.1.2 Solvent Mixture Preparation

Water used for preparation of solvent mixtures is purified via the Milli-Q IQ Water System. The following mixtures are prepared by in-house laboratory support groups: backing solvent

(acetonitrile:tetrahydrofuran:water, 62.5:25:12.5% by v/v), mobile phase A (water + 0.05% TFA by v/v) and mobile phase A (acetonitrile + 0.05% TFA by v/v).

All other binary or ternary solvent mixtures are prepared using the commercial solvents listed above. Those solvent mixtures are prepared volumetrically by adding the desired volumetric solvent fraction in 8 mL glass vials to obtain 4 mL nominal volume. For example, 50/50 %v/v binary mixtures are prepared by mixing 2 mL of each solvent. After stirring the binary and ternary mixtures, images of the vials are taken using the Freeslate CM3 platform (now Big Kahuna, Unchained Labs, California – USA).

The actual volume of the mixtures (i.e., to account for any volume changes occurring due to non-ideal solvent mixing) is determined by using a modified version of a MATLAB® (The MathWorks, Inc, Massachusetts – USA) image analysis algorithm. A detailed explanation of the image analysis algorithm is described in Duffield et al. (2021). Volumes after mixing are employed in the analysis of Appendix I.

### 7.2.1.3 Solubility Screen

Solubility screens of API #1 are performed in 1 mL glass shell vials placed in a 96-well metal plate. In each vial, 50 mg of API are dispensed though a Labman MultiDose™ (Labman, UK) powder dispensing platform. After addition of micro stir bars and 500 µL of solvent or solvent mixture in each well, the plate is closed with a PTFE sheet to ensure chemical compatibility, two rubber mats to ensure gas-tight sealing, and a metal lid. The sealed plate is then stirred on a temperature-controlled tumble stirrer at 550 rpm at the desired temperature (20, 40, or 50 °C) until equilibrium is ensured (18 h at 20 °C; 6 h at high temperature). After that time, the plate is centrifuged (3500 rpm, 5 min, 20 or 40 °C, Sorvall Lynx 4000, Thermo Fisher Scientific, Inc, Massachusetts – USA) to compact any insoluble API on the bottom of the vials. The lid and covering mats are then removed, and each vial is visually inspected to determine if the solutions are saturated or the API is fully dissolved, and to annotate any peculiar characteristic (e.g., formation of a gel or phase splitting). Finally, the supernatant is sampled and diluted for UPLC analysis. Sampling and dilution are performed on an CyBio Felix (Analytik Jena AG, Germany) liquid handling platform, equipped with a 250/96 head. 10x and 100x serial dilution are performed by sampling 40 µL of solution and adding 360 µL of diluent (acetonitrile:THF:water 62.5:25:12.5 vol%).

#### 7.2.1.4 Ultra Performance Liquid Chromatography (UPLC) analysis

The diluted samples are analysed by UPLC on an 1290 Infinity II (Agilent Technologies, Inc, California – USA) system equipped with a Waters XSelect CSH C18 30 mm (ID 2.1 mm, particle size 2.5 um) column kept at 40 °C. The mobile phases are water + 0.05 % TFA (A) and acetonitrile + 0.05 % TFA (B). The following gradient is used: t = 0 min, 97 % A; t = 0.5 min, 2 % A; t = 0.6 min, 2 % A; t = 0.61 min, 97 % A; analysis time = 0.8 min. The flow rate is 2.2 mL/min, and the injection volume is 2 µL. The chromatograms are collected at 220 nm (80 Hz) and quantitative analysis of the API in solution is carried out by comparison with a calibration curve built between 15.6 and 1000 µg/mL.

### *7.2.2 Mathematical models*

A machine learning approach is developed in order to predict API solubility based on temperature, mixture composition and solvents structure. Several molecular descriptors have been proposed in literature to represent molecular structure, but there is no consensus on the best selection. To overcome this limitation, the consolidated UNIFAC theory is used as a reference. This section briefly summarises the main principles of the UNIFAC theory and explains the model proposed.

#### 7.2.2.1 UNIFAC theory for solid-liquid equilibrium

As in Gmehling et al. (1978), let's assume that a solid can dissolve into the liquid phase, while the liquid does not enter into the solid phase. Solvent and solute species are indicated by the subscripts 1 and 2, respectively. Solid and liquid phases are indicated by the superscripts S and L, respectively. The phase equilibrium for the solid solute can be expressed as equality of fugacity in the solid phase ($f_2^{\mathrm{S}}$) and in the liquid phase ($f_2^{\mathrm{L}}$):

$$f_2^{\mathrm{S}} = f_2^{\mathrm{L}} \,. \tag{7.1}$$

Since there is no solubility of compound 1 in the solid phase, the left-hand side of Eq. (7.1) becomes equal to the fugacity of the pure compound:

$$f_2^{\mathrm{S}} = f_{2,\mathrm{pure}}^{\mathrm{S}} \,. \tag{7.2}$$

Instead, the right-hand side can be expressed in terms of the activity coefficient ($\gamma_2$) as:

$$f_2^{\mathrm{L}} = \gamma_2 x_{\mathrm{mol},2} f_{2,\mathrm{pure}}^{\mathrm{L}} \,, \tag{7.3}$$

where $x_{\mathrm{mol},2}$ is the mole fraction and $f_{2,\mathrm{pure}}^{\mathrm{L}}$ is the fugacity of the pure subcooled liquid 2 at system temperature.

UNIFAC calculates the activity coefficient $\gamma_2$ in Eq. (7.3) by considering two contribution: a combinatorial one, related to differences in size and shape, and a residual one, related to differences in intermolecular forces of attraction. Moreover, being a group-contribution method, it considers the set of functional groups in the mixture instead of considering molecules as a whole. Thus, the activity coefficient of the $i$-th component in a multicomponent mixture is given by:

$$\ln(\gamma_i) = \ln\left(\gamma_i^C\right) + \ln(\gamma_i^R), \tag{7.4}$$

where $\gamma_i^C$ and $\gamma_i^R$ denote combinatorial and residual contributions, respectively. The former is a function of composition expressed as molar fractions, while the latter depends on both composition and temperature. Moreover, $\gamma_i^R$ is calculated with binary-interaction parameters $a_{lm}$ which must be estimated with experimental data for every pair of functional groups $l$ and $m$ in the mixture. Finally, the definition of main groups (e.g., Methyl) and sub-groups (e.g., $CH_3$, $CH_2$, $CH$ and $C$) and their corresponding parameters are available in the literature, for instance in Dortmund Data Bank (http://www.ddbst.com/).

This theory is employed as starting point to develop the machine learning model described in Subsection 7.2.2.2.

### 7.2.2.2 Machine learning model for solid-liquid equilibrium

An empirical model $\mathbf{f}$ for the prediction of API solubility in organic solvents at different temperatures and composition is developed. It can be expressed by the general form:

$$y = \mathbf{f}(\mathbf{u}), \tag{7.5}$$

where $y$ is the response variable, $\mathbf{u}$ is the vector of $V$ regressors and $\mathbf{f}$ is a set of algebraic equations (the model is not dynamic since solubility is measured at equilibrium).

In this case, the response variable is the logarithm of the API molar fraction $x_{\mathrm{API}}$ in the liquid phase at equilibrium:

$$y = \log(x_{\mathrm{API}}), \tag{7.6}$$

where $x_{\mathrm{API}}$ is the same as $x_2$ in Eq. (7.3).

The UNIFAC theory (Subsection 7.2.2.1) is employed to define the basic components of the vector of regressors $\mathbf{u}$: a) temperature $T$; b) solvent types; c) molar composition.

Solvent types are identified through their subgroups as defined by the UNIFAC theory (http://www.ddbst.com/). Therefore, given a pool of $N_k$ subgroups for all organic solvents, the

$i$-th solvent is identified by a vector $\boldsymbol{\upsilon}_i[1\times N_k]$ containing the number of occurrences $\upsilon_{ik}$ of each of the $k$ subgroup in the $i$-th solvent:

$$\boldsymbol{\upsilon}_i = [\upsilon_{i,1}, \dots, \upsilon_{i,k}, \dots, \upsilon_{i,N_k}] \,. \tag{7.7}$$

For instance, the 14 organic solvents employed in this work for solubility experiments have $N_k$=11 subgroups overall: $CH_3$, $CH_2$, CH, OH, $CH_3OH$, $CH_2CO$, $CH_3COO$, $CH_2COO$, THF, $CH_3CN$, $CH_2CN$ . Table 7.4 shows the numeric labels used to indicate all subgroups in the following sections. Moreover, Table 7.5 shows the $\boldsymbol{\upsilon}_i$ vectors used to represent every solvent type used in this work.

**Table 7.4.** *Numerical label for every subgroup. The main group of every subgroup is shown.*

| No. | Main group | Subgroup |
|---|---|---|
| 1 | Methyl | $CH_3$ |
| 2 | Methyl | $CH_2$ |
| 3 | Methyl | CH |
| 4 | Alcohol | OH |
| 5 | Methanol | $CH_3OH$ |
| 6 | Ketone | $CH_2CO$ |
| 7 | Acetate | $CH_3COO$ |
| 8 | Acetate | $CH_2COO$ |
| 9 | Ether | THF |
| 10 | Nitrile | $CH_3CN$ |
| 11 | Nitrile | $CH_2CN$ |

**Table 7.5.** *Number of occurrences of the 11 subgroups in the 14 organic solvents employed for solubility experiments.*

| Solvent | $\upsilon_{i,1}$ | $\upsilon_{i,2}$ | $\upsilon_{i,3}$ | $\upsilon_{i,4}$ | $\upsilon_{i,5}$ | $\upsilon_{i,6}$ | $\upsilon_{i,7}$ | $\upsilon_{i,8}$ | $\upsilon_{i,9}$ | $\upsilon_{i,10}$ | $\upsilon_{i,11}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1-Butanol | 1 | 3 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2MeTHF | 1 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| 3-Pentanone | 2 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| Acetonitrile | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| Cyclohexane | 0 | 6 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Ethanol | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Ethyl acetate | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| GVL | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| Isopropanol | 2 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Isopropyl acetate | 2 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| Methanol | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| Propionitrile | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| n-propanol | 1 | 2 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| n-butyl acetate | 1 | 3 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |

Considering a mixture with $N_L$ organic solvents, the presence of the $k$-th subgroup is due to the contributions of all solvents in the mixture; moreover, the contribution of every solvent is proportional to its molar fraction in the mixture. Therefore, the input factor $g_k$ representing the effect of the $k$-th subgroup in the mixture is obtained by the following weighted sum:

$$g_k = \sum_{i=1}^{N_L} \upsilon_{ik} x_{\mathrm{mol},i}. \tag{7.8}$$

The molar fractions $x_{\mathrm{mol},i}$, $i$=1,…, $N_L$, of Eq. (7.8) refer to the organic solvents in the mixture, before the dissolution of API and assuming no volumetric effects due to mixing of liquids. This type of molar fractions are useful in extrapolation, since mixing effects and API solubility cannot be known before performing experiments. The effect of these assumptions on the final model are analysed in Appendix I.

Similarly to the UNIFAC theory, also binary interactions $g_l g_m$ between subgroups are considered. Temperature, subgroups and interactions among subgroups for all observations $N$ are collected in the input matrix $\mathbf{U}$ $[N \times V]$:

$$\mathbf{U} = \begin{bmatrix} \mathbf{u}_1 \\ \vdots \\ \mathbf{u}_n \\ \vdots \\ \mathbf{u}_N \end{bmatrix}, \tag{7.9}$$

where the $n$-th observation is represented by the $\mathbf{u}_n$ vector $[1 \times V]$ of input variables (or regressors):

$$\mathbf{u}_n = \left[u_{n,1},, \dots, u_{n,V}\right] = [T, g_1, \dots, g_{N_k}, g_1 g_2, \dots, g_{N_{k-1}} g_{N_k}]\,\big|_n ; \tag{7.10}$$

in other terms, the first regressor is temperature, the next $N_k$ regressors represent the single subgroups $g_k$ (Eq. 7.8) and the remaining $(V - N_k - 1)$ regressors represent the binary interactions $g_l g_m$ between subgroups.

Finally, the PLS model explained in section 2.2.1 is applied to handle regressors correlations and to relate $\mathbf{U}$ to the corresponding vector of response variables $\mathbf{y}$ $[N \times 1]$. The performance of the PLS model in calibration and validation is assessed in terms of sample diagnostics, model prediction uncertainty, variable diagnostics and model diagnostics as explained in section 2.2.1.1.

### 7.2.2.3 Selection of binary mixtures

The aim of the experimental screening is to calibrate a solubility model with measurements of single solvents and a few binary mixtures in such a way as to use it for solubility predictions in different conditions, e.g. binary mixtures with different solvent types and/or composition and ternary mixtures, correctly representing also the temperature dependence. Binary mixtures should be selected in order to have non-zero entries for every subgroup $g_i$ and subgroup interaction $g_i g_j$ (as defined in section 7.2.2.2). To do so, the procedure illustrated in Figure 7.1 is adopted:

- step 1: the list of solvents of interest is defined;

- step 2a: all single solvents are selected for the experiments;
- step 2b: the list of possible binary mixtures is made using the solvents included in step 1;
- step 3: mixtures affected by miscibility issues are excluded;
- step 4: mixtures having a boiling temperature $T_b$ too close to the maximum temperature for the experiments are excluded;
- step 5: binary mixtures are selected from the available ones in such a way as the effect of all subgroups $g_i$ and subgroups interaction $g_i g_j$ in Eq. (7.10) can be represented by measured data.
- step 6: binary mixtures are added following the rationale of step 5 until the 96-vial plate is filled.

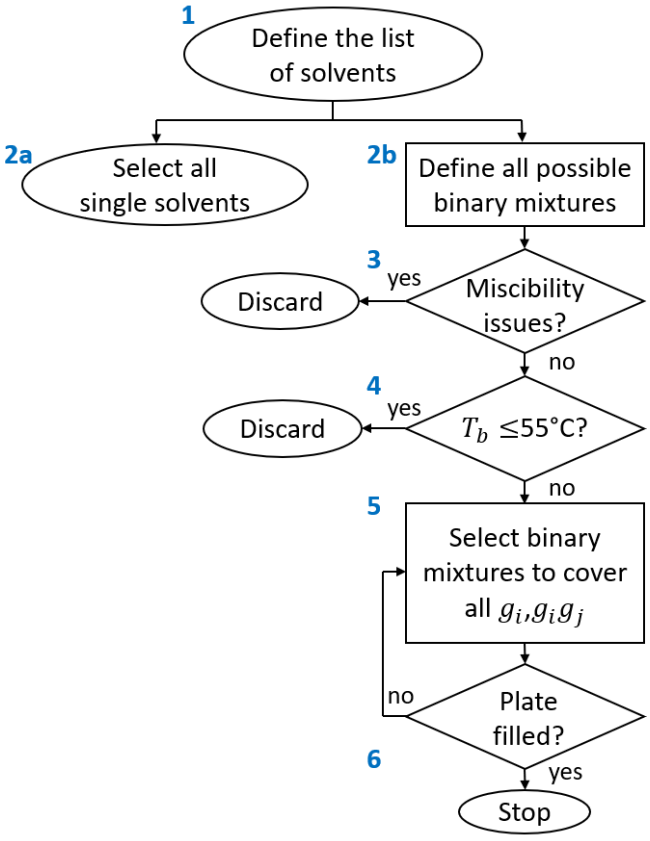



**Figure 7.1**. *Schematic representation of the method used to select experimental conditions to calibrate the PLS model.*

Selecting candidate solvent systems that are single phase liquid mixtures is required to facilitate the measurement of a reliable solubility value. Single phase systems are also generally more desirable for manufacturability purposes in reaction and crystallisation systems. Moreover, boiling temperatures too close to the maximum experimental temperature of 50°C should be

avoided in order to prevent solvent systems from evaporating excessively during the experiments; this improves the accuracy of the measured solubility values. Therefore, a minimum boiling temperature of 55°C is accepted. Explanation of the procedure to study miscibility and evaporation issues for steps 3 and 4 are provided in Appendix H. Once the procedure shown in Figure 7.1 is completed, the preparation of the 96-vial plate remains the same as in the common industrial practice: binary mixtures are prepared by adding specified volumes of each solvent; the 96 vials are measured at two temperatures (in this work, 20°C and 40°C or 20° and 50°C, see section 7.2.1).

## 7.3 Results and discussion

In this section, the performance of the proposed PLS model is analysed. Subsection 7.3.1 shows the results obtained with calibration data, while Subsection 7.3.2 shows the results with validation data, also distinguishing specific subsets of validation data (as defined in Subsection 7.2.1.1). The most influential regressors on the PLS model are analysed and discussed in Subsection 7.3.1.1, in order to provide a physical interpretation of the functional groups that may have a considerable effect on drug solubility.

Finally, the proposed methodology is validated using benchmark datasets from the literature in Subsection 7.3.3.

### *7.3.1 Calibration data*

The **U** and **Y** calibration data (Table 7.1, Subsection 7.2.1.1) are used to build the PLS model after being autoscaled. The selected number of LVs is chosen following the "eigenvalue-greater-than-one" rule (Mardia et al., 1979). The first 5 latent variables are retained for the model, thus greatly reducing the initial input dimensionality (i.e., 52 regressors $u_j$). Even if 5 LVs represent a limited amount of **U** variability, approximately 32%, they represent the 92% of variability of the response **Y**, which is absolutely adequate for predictive purposes.

Sample diagnostic for the obtained PLS model is performed by comparing Hotelling $T^2$ and SPE statistics with the corresponding 95% confidence limits (Figure 7.2).

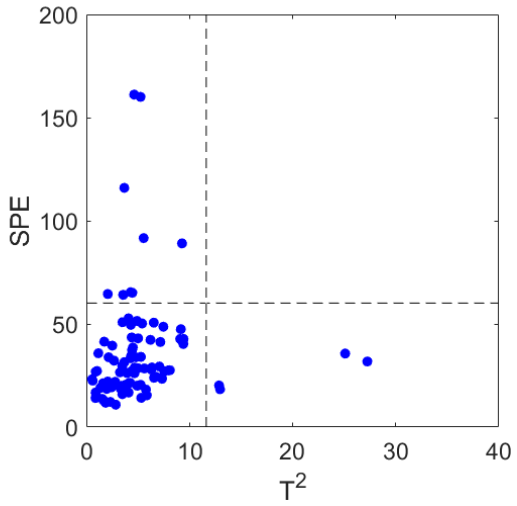



**Figure 7.2**. *Calibration results of the PLS model: SPE vs $T^2$ plot with 95% confidence limits (dotted lines).*

As shown in Figure 7.2, the majority of observations have a $T^2$ and SPE statistics below the 95% confidence limit, suggesting that the model is able to represent average behaviour and correlation structure of the data.

Predictions of API solubility with the respective 95% CI are compared to the corresponding measured values in the parity plot of Figure 3.

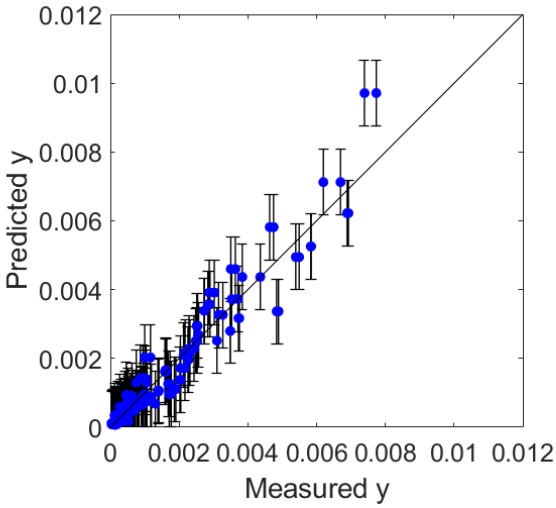



**Figure 7.3**. *Calibration results of the PLS model: parity plot of predicted and measured API solubility, in the original scale (mole fractions) and with 95% confidence intervals.*

The points in Figure 7.3 are close to the diagonal and have sufficiently small 95% CIs, suggesting that model predictions are in accordance with experimental data and their precision is satisfactory. Coherently, calibration measurements have $R^2$ close to 1, 0.92, and a small $RMSE$, $4.67 \cdot 10^{-4}$. Moreover, 94% of the observations have a 95%CI that crosses the diagonal, namely their prediction errors are statistically equivalent to zero.

### 7.3.1.1 Most impactful regressors

The most influential regressors on the PLS model are analysed through **β** regression coefficients (Eq. 19) and VIP indices (Eq. 2.28). In Figure 7.4, the length of the bars refer to the values of the VIPs, while filled and hatched bars indicate positive and negative values of **β**, respectively.

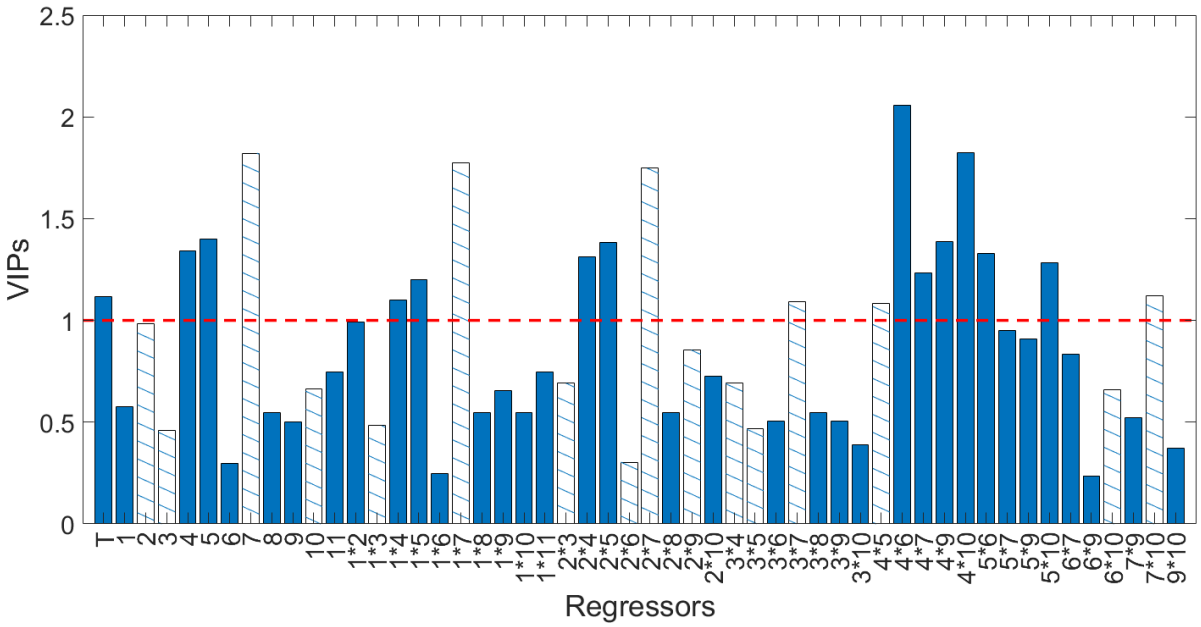



**Figure 7.4**. *VIP scores of the variables used to build the PLS model calibrated with API #1 experimental data. Filled bars represent positive* ***β****, hatched bars represent negative* ***β*** *. The red dotted line represents the threshold of 1 for the VIPs.*

Based on Figure 7.4 and considering also VIP>0.9 (to include regressors with high impact but slightly smaller than 1), the following regressors have a considerable impact on the PLS model:

- temperature;
- subgroups (sorted): 7($CH_3COO$) > 5 ($CH_3OH$) > 4 (OH)> 2($CH_2$);
- interactions among subgroups (sorted): 4*6 (OH *$CH_2CO$) >4*10 (OH *$CH_3CN$) > 1*7 ($CH_3$*$CH_3COO$) > 2*7 ($CH_2$*$CH_3COO$) > 4*9 (OH * THF) > 2*5 ($CH_2$*$CH_3OH$) >5*6 ($CH_3OH$* $CH_2CO$) >2*4 ($CH_2$*OH) > 5*10 ($CH_3OH$ *$CH_3CN$) > 4*7 (OH *$CH_3COO$) > 1*5($CH_3$*$CH_3OH$) > 7*10 ($CH_3COO$*$CH_3CN$) >1*4 ($CH_3$* OH) > 3*7 (CH*$CH_3COO$) > 4*5 (OH*$CH_3OH$) > 1*2 ($CH_3$*$CH_2$) > 5*7 ($CH_3OH$*$CH_3COO$) > 5*9($CH_3OH$* THF).

  Therefore, mainly interactions of 1 ($CH_3$), of 4 (OH) and of 5 ($CH_3OH$) with other subgroups.

The significant effect of temperature on solubility is expected, since solubility increase with temperature is evident also experimentally. As regards the relevant influence of $CH_3$ and $CH_2$, as single regressors and/or in binary interactions, it can be explained by the fact that they are very common functional groups, therefore they impact model parametrization. The influence of $OH$ and $CH_3OH$ is likely due to the possibility of forming hydrogen-bounds. Moreover, $CH_3COO$ has a polar oxygen due to the double bond $C = O$ and this polarity likely induces interactions with other groups in the mixture, hence impacting solubility.

Among the 6 regressors having VIP>1 and a negative $\beta$, 5 of them involve subgroup 7 ($CH_3COO$): $CH_3COO$ as single subgroup and in the binary interaction parameters 1*7 ($CH_3$*$CH_3COO$), 2*7 ($CH_2$*$CH_3COO$), 3*7 (CH *$CH_3COO$) and 7*10 ($CH_3COO$*$CH_3CN$). This suggests that the descrease of solubility of the tested drug is mainly correlated with the presence of subgroup 7 ($CH_3COO$).

Finally, subgroups 6 ($CH_2CO$), 9 (THF) and10 ($CH_3CN$), involved in significant interactions, likely owe their influence to polarity, as well: $CH_2CO$ has a polar oxygen due to the $C = O$ bond; ethers like THF have polar oxygens and nitriles like $CH_3CN$ have triple bond on CN, thus a polar nitrogen. Such polarity is likely responsible for the interaction with other functional groups, with an impact on solid solubility.

### *7.3.2 Validation of the model on new unknown binary and ternary mixtures*

After autoscaling, the validation dataset of 288 data points (see Table 7.1 in section 7.2.1.1) is projected onto the PLS model calibrated with the data of Subsection 7.3.1. Sample diagnostics are shown in Figure 7.5 in terms of Hotelling $T^2$ and SPE statistics.

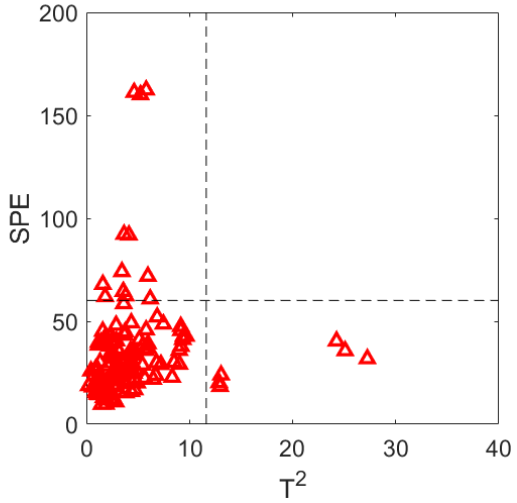



**Figure 7.5**. *Validation results of the PLS model: SPE vs $T^2$ plot with 95% confidence limits.*

As shown, Hotelling $T^2$ and SPE statistics are below their 95% confidence limits for the majority of the observations, with no observation with both statistics above the limits. These results are analogous to the ones obtained with calibration data (Figure 7.2).

Model predictions with 95% CI are compared against the corresponding measured values in Figure 7.6.

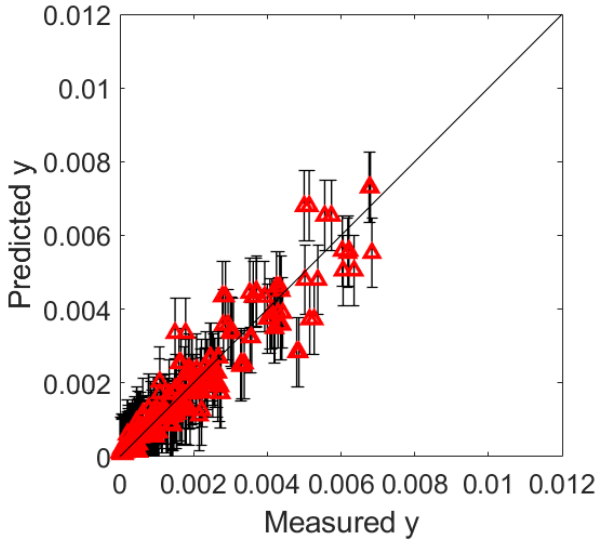



**Figure 7.6**. *Validation results of the PLS model: parity plot of measured versus predicted solubilities, in the original scale (mole fractions) and with 95% CI.*

Figure 7.6 shows that model predictions are precise and in accordance with the experimental measurements. This result is confirmed by $R^2 = 0.90$, thus close to 1, and by the small $RSME$ equal to $4.90 \cdot 10^{-4}$. Moreover, 93% of the predictions have a prediciton error that is statistically equivalent to 0, because their 95%CI crosses the diagonal. Therefore, model predictions of validation experiments have an accuracy and a precision that is comparable to the one of calibration experiments.

Validation experiments can be divided into subgroups having similar features, such as temperature and number and types of solvents employed. They are analysed in a separate way in Subsection 7.3.2.1.

### 7.3.2.1 Specific types of validation data

Validation data are divided into the following subgroups: a) single solvents; b) binary mixtures; c) subset of binary mixtures in b) that have a pair of organic solvents not employed in calibration; d) ternary mixtures; e) validation data (single solvents, binary mixtures, ternary mixtures) at 50°C only. Prediction and accuracy of model predictions for those subgroups are shown in Figure 7.7; the corresponding $R^2$ and RMSE are shown in Table 7.6.

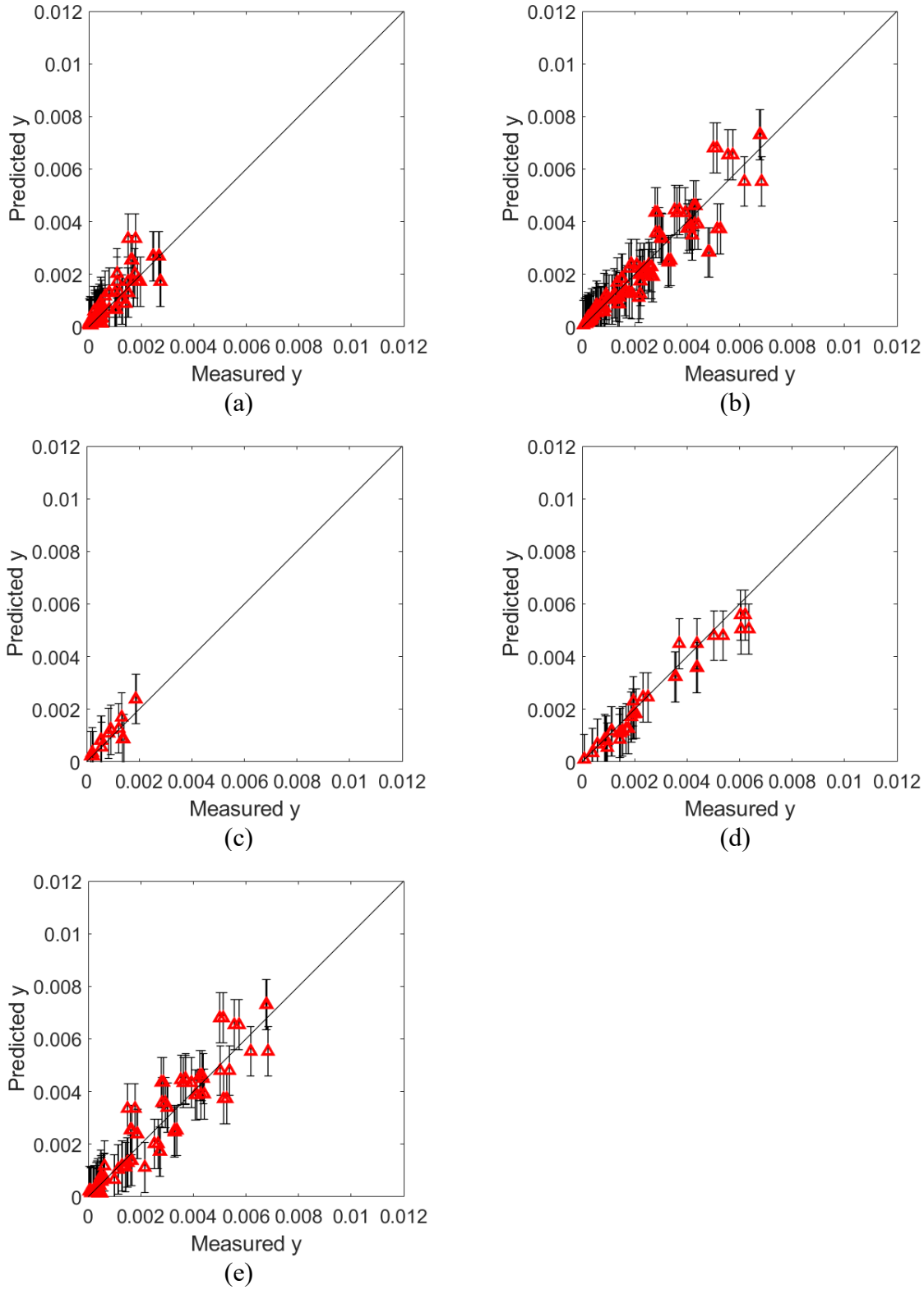



**Figure 7.7**. *Validation results of the PLS model: a) measurements of single solvents not used in calibration; b) all binary mixtures selected for validation; c) binary mixtures with mixing of solvent types not used in calibration; d) ternary mixtures; e) all validation measurements at 50°C. Predicted values of the response variable are shown with their 95% CI.*

**Table 7.6.** *Validation results of the PLS model: $R^2$ and RSME calculated with predicted and measured response values in the original scale (molar fractions).*

| Type of validation experiment | $R^2$ | $RMSE$ |
|---|---|---|
| single solvents | 0.71 | $3.80 \cdot 10^{-4}$ |
| all binary mixtures | 0.89 | $5.73 \cdot 10^{-4}$ |
| binary mixtures with solvent types not used in calibration | 0.66 | $3.02 \cdot 10^{-4}$ |
| ternary mixtures | 0.94 | $4.35 \cdot 10^{-4}$ |
| measurements at 50°C | 0.88 | $6.64 \cdot 10^{-4}$ |

By comparing the measured solubilities of single solvents (Figure 7.7a), binary mixtures (Figure 7.7b) and ternary mixtures (Figure 7.7d) it can be seen that mixtures allow to reach higher values of drug solubility with respect to single solvents. Moreover, there is not an increase of drug solubility by using such ternary mixtures rather than the binary mixtures.

Figure 7.7 shows also that model predictions are precise and accurate for every subgroup of validation experiments, even though the experimental conditions are outside the temperature range explored in calibration (Figure 7.7e) and/or mixtures with different number and types of organic solvents are used (Figures 7.7d and c, respectively). Satisfactory results are found with respect to the coefficient of determination (Table 7.6), since binary mixtures (Figure 7.7b), ternary mixtures (Figure 7.7d) and measurements at 50°C (Figure 7.7e) have a $R^2$ close to or higher than 0.90. Single solvents and binary mixtures not used in calibration (Figures 7.7a, and 7.7c, respectively) have a smaller $R^2$, approximately 0.70, but this is likely due to the smaller average solubility values that determine a smaller denominator in Eq. (2.14). In fact, the average values of the solubility data shown in Figure 7.7a-e are, respectively: a) $6.58 \cdot 10^{-4}$; b) 0.0019; c) $9.12 \cdot 10^{-4}$; d) 0.0024; e) 0.0021. This explanation is further supported by the fact that RMSE of single solvents and binary mixtures are small, $3.81 \cdot 10^{-4}$ and $3.02 \cdot 10^{-4}$ respectively, even smaller than the one obtained with calibration experiments. Overall, the RMSE in validation ranges between $3.02 \cdot 10^{-4}$ and $6.64 \cdot 10^{-4}$, therefore it is considerably smaller than the overall variation of measured solubility, namely 0.0077.

### *7.3.3 Literature data*

Some free web services for organic solubility predictions are available, such as Formulation AI and RMG - Reaction Mechanism Generator, thanks to the contributions of Ye and Ouyang (2021) and Vermeire et al. (2022), respectively. However, a direct quantitative comparison between the results of sections 7.3.1-7.3.2 and the predictions of these online models cannot be made: first, the online models allow to predict solid solubility only in single solvents (no binary or ternary mixtures); moreover, SMILES strings are required as inputs, but SMILES strings

typically refer to the freebase molecules (such as the API of the drug substance indicated as API#1), while in this work the mesylate salt of the freebase is employed.

Therefore, the adequacy of the proposed PLS method is assessed in a different way, namely by using 9 benchmark datasets from the Literature, as introduced in Section 7.2.1.1. Table 7.7 and 7.8 refer to, respectively, the 8 datasets of organic solubility in single solvents retrieved from Zenodo (Krasnov et al., 2022) and the dataset of organic solubility in single solvents and binary mixtures retrieved from Khajiret al. (2024). Tables 7.7 and 7.8 include also information on the PLS model: the number of original variables $u_j$ defined as in Eq. (7.10); the number of latent variables of the PLS model; the accuracy of models predictions calculated as $R^2$ and RMSE. The corresponding parity plots are shown in Figures 7.8 and 7.9.

**Table 7.7.** *Description of the datasets retrieved from Krasnov et al., (2022) and used to build PLS models to predict SLE. Solvent types different from the ones employed in the experimentation of section 3.1-3.2 are indicated in bold.*

| Dataset | No. exp. cal. | No. exp. val. | No. $u_j$ | No. LV | $R^2$ calib. | $RMSE$ cal. | $R^2$ val. | $RMSE$ val. |
|---|---|---|---|---|---|---|---|---|
| 1 | 20 | 88 | 21 | 4 | 0.98 | 0.0035 | 0.99 | 0.0014 |
| 2 | 10 | 32 | 14 | 4 | 0.99 | 0.0067 | 0.99 | 0.0041 |
| 3 | 22 | 77 | 19 | 3 | 0.82 | $3.49\cdot 10^{-4}$ | 0.87 | $1.99\cdot 10^{-4}$ |
| 4 | 16 | 24 | 19 | 3 | 0.94 | 0.0172 | 0.99 | 0.0056 |
| 5 | 16 | 56 | 18 | 4 | 0.995 | $1.91\cdot 10^{-4}$ | 0.998 | $8.20\cdot 10^{-5}$ |
| 6 | 16 | 56 | 17 | 3 | 0.97 | 0.0205 | 0.97 | 0.0121 |
| 7 | 20 | 24 | 11 | 3 | 0.95 | 0.0355 | 0.95 | 0.0109 |
| 8 | 16 | 23 | 16 | 3 | 0.92 | $2.4\cdot 10^{-4}$ | 0.96 | $1.33\cdot 10^{-4}$ |

As shown in Figure 7.8, the 95% CIs of model predictions are always smaller than the overall variation of the measured solubility values, suggesting that model predictions are sufficiently precise. Besides being close to the diagonal, the majority of prediction have a 95% CI that crosses the bisector, meaning that the prediction error is statistically negligible; more specifically, the percentages of predictions whose 95%CIs contain the corresponding measured value are: 99% (Figure 7.8a); 100% (Figure 7.8b); 98% (Figure 7.8c); 98% (Figure 7.8d); 99% (Figure 7.8e); 100% (Figure 7.8f); 100% (Figure 7.8g); 97% (Figure 7.8h). Coherently, $R^2$ is $\geq$0.90, thus very close to 1, for all the considered datasets, except for borylate (Figure 7.8c) which nonetheless has $R^2$ equal to 0.82 and 0.87 in calibration and validation, respectively, therefore still indicating a good prediction accuracy. Notice that the model performance in validation is not degraded with respect to the calibration one, with $R^2$ $\geq$0.99 for DBHA (Figure 7.8a), Fenofibrate (Figure 7.8b), L-Arginine L-pyroglutamate (Figure 7.8d) and 2-chloro-4-amino-6,7-dimethoxyquinazoline (Figure 7.8e).

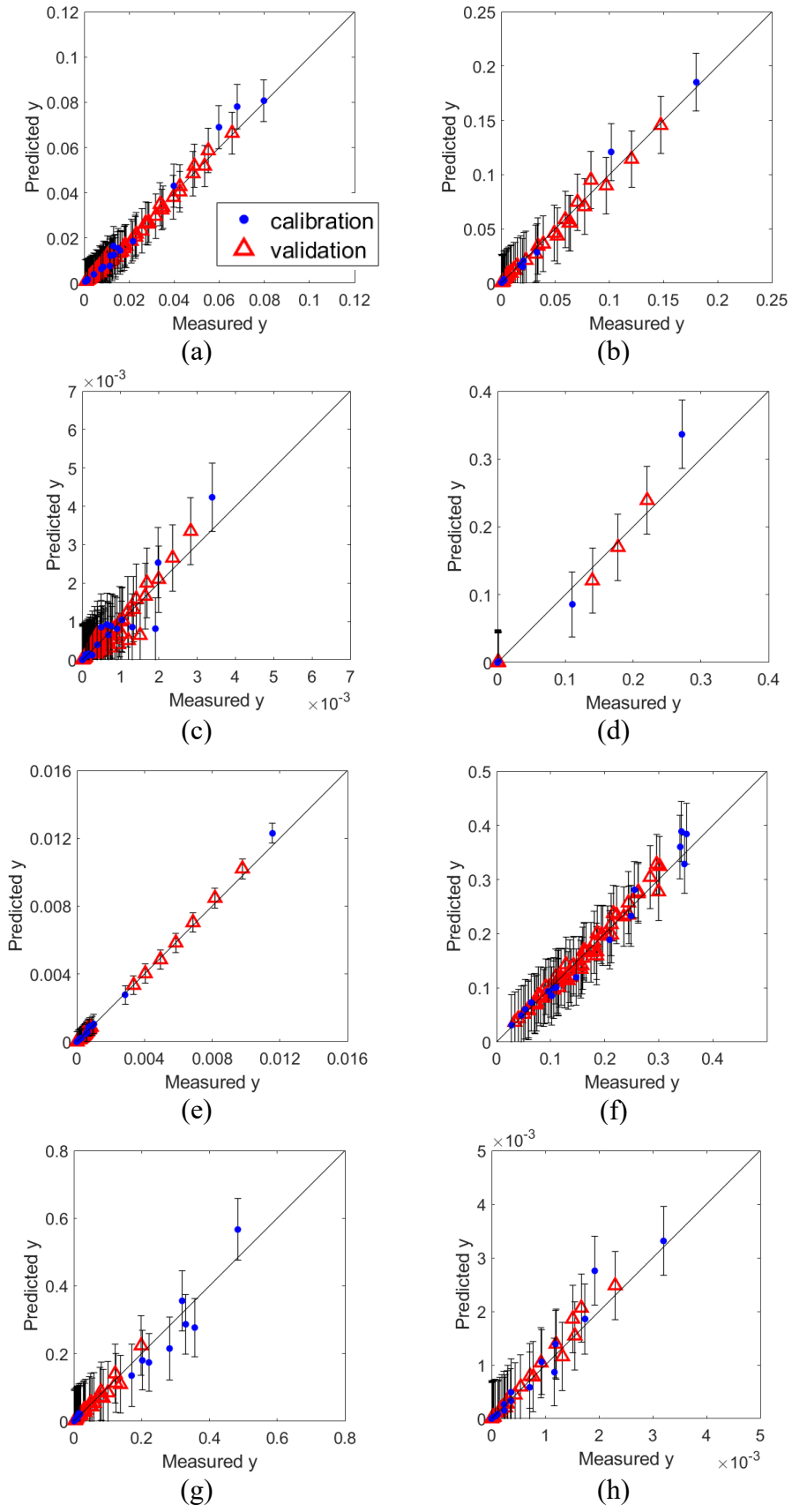



**Figure 7.8**. *Parity plots of data from Krasnov et al. (2022) versus the corresponding predictions. Model predictions are displayed with their 95% confidence intervals.*

The PLS model structure is appropriate also at temperature values very different from the range of 20-50°C explored with API #1 (Subsections 7.3.1-7.3.2): for instance, Fenofibrate solubility (Figures 7.8b) is measured at much lower temperatures, namely -30,-25, -20,-15, -10,-5 °C, but the determination coefficient is still very high in both calibration and validation, namely $R^2$=0.99.

**Table 7.8.** *Description of the datasets retrieved from Khajir et al. (2024)*

| No. exp. cal. | No. exp. val. | No. $u_j$ | no. LV | $R^2$ cal. | *RMSE* cal. | $R^2$ val. | *RMSE* val. |
|---|---|---|---|---|---|---|---|
| 10 | 45 | 11 | 3 | 0.99 | 0.0023 | 0.99 | 0.0018 |

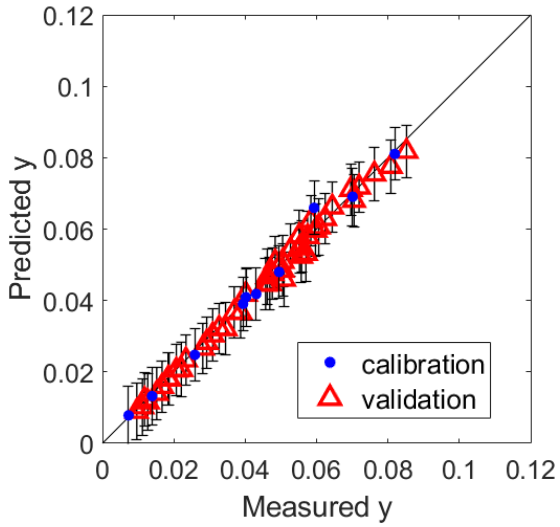



**Figure 7.9**. *Parity plots of measured literature data from Khajir et al. (2024) versus the corresponding predictions obtained with the PLS model. Model predictions are displayed with their 95% confidence intervals.*

Even though only 10 experimental points over 55 measured overall are employed in calibration (Table 7.8), the PLS model is able to accurately and precisely predict solubility: predictions are very close to the diagonal (Figure 7.9) and their 95%CI are much smaller than the overall range of variation of solubility and they always (100% of the times) cross the diagonal itself (Figure 7.9). This holds with both calibration and validation data, with $R^2$=0.99 (Table 7.8).

Overall, the 9 literature datasets further prove that the proposed modelling approach is adequate with a variety of solid molecules and organic solvents commonly employed in crystallization units. Notice that the 9 datasets (Table 7.7 and 7.8) span 2 order of magnitude of organic solubility: indeed, organic solubility of Benorilate (Figure 7.8c) and TPBi (Figure 7.8h) have an order of magnitude of $10^{-3}$; 2-chloro-4-amino-6,7-dimethoxyquinazoline (Figure 7.8e) has an order of magnitude of $10^{-2}$, while the remaining ones (Figure 7.8a-b,d,f-g,i; Figure 7.9) have an order of magnitude of $10^{-1}$. This suggests that the proposed modelling approach holds

also with higher organic solubility values than the ones explored with the tested drug of sections 7.3.1-7.3.2.

As regards the first 8 literature datasets (Table 7.7), it is not possible to directly compare performance of the PLS models calibrated in this section with the one of the models calibrated in the original publication. In fact, they employed semi-empirical thermodynamic models such as Apelblat equation (Zhou et al., 2021; Wang et al., 2021; Yu et al., 2021; Li et al., 2020; Wu et al., 2020; Zhang et al., 2019;Huang et al., 2015), $\lambda$h equation (Zhou et al., 2021; Li et al., 2020; Wu et al., 2020; Zhang et al., 2019; Huang et al., 2015), van't Hoff model (Sadeghi and Rasmuson, 2020), which need re-calibration for every solvent type. Instead, the model proposed in this work allows to identify solvent types based on their functional groups, therefore one PLS model is calibrated with the solubility data of one solute in multiple solvents. Moreover, the abovementioned models were calibrated with the entire experimental dataset and were used to fit data; instead, in this section a small fraction of the literature datasets are used to calibrate the model and the rest is used for validation: the 8 dataset used for calibration are reduced of 81%, 76%, 78%, 60%, 78%, 78%, 55% and 64% (considering the datasets a-h of Table 7.7, respectively).

As regards the ninth literature dataset (Table 7.8), Khajir et al. (2024) used it to calibrate different cosolvency models. Two of them, namely Combined Nearly Ideal Binary Solvent (CNIBS)/Redlich-Kister model and modified Wilson equation, cannot be directly compared with the results of this section: they model the effect of mixture composition on solubility, but not the effect of temperature, therefore they must be re-calibrated for the 5 different temperatures. Moreover, the Jouyban-Acree-van't Hoff model was employed and it is able to represent both composition and temperature effects (for the same binary mixture ethanol-acetinitrile), therefore it can be compared with the PLS model proposed in this work. Khajir et al. (2024) calibrated the Jouyban-Acree-van't Hoff model with the same calibration data indicated in Table 7.8 (as anticipated in section 7.2.1) and they obtained a mean relative deviation ($\%MRD = \frac{100}{N}(\sum_{n=1}^{N}(|y_n - \hat{y}_n|/y_n)$) equal to 4.3% for the overall dataset. The same index is calculated with the PLS model proposed in this work and a MRD of 3.6% is obtained, therefore the accuracy of solubility predictions is improved. This further confirms the advantages of the proposed PLS model, besides the fact that it can be used to predict a much wider range of organic mixtures types and compositions.

## 7.4 Conclusions

A novel machine learning model to predict solubility of drug and drug-like molecules in mixtures of organic solvents has been proposed. Little information is required to predict solubility, namely temperature, solvent types and mixture composition. Solvent types are identified through their molecular structure and the issue of selecting proper molecular descriptors is overcome by the employment of the consolidated UNIFAC theory as a reference. In fact, solvents molecules are broken down into their subgroups defined as in UNIFAC model and both single subgroups and interactions among subgroups are used to define the matrix of regressors. Temperature effect is explicitly included in the model, too. These regressors are correlated among each other, but their correlation is handled through PLS regression.

The PLS model is calibrated with experimental data of a real drug substance, namely the mesylate salt of a real drug substance, with 14 organic solvents commonly employed in the pharmaceutical industry to design and operate crystallisation units. Both single solvents and binary mixtures are measured to calibrate the model. The time and resources employed to calibrate the model have been reduced by combining the usage of high-throughput technology with a proper selection of organic mixtures to fill one 96-vials plate.

Experiments at single solvents, binary mixtures and ternary mixtures at 20, 40 and 50°C are used to validate the model, supporting the adequacy of model predictions for crystallisation design and/or optimisation. In fact, 94% and 93% of calibration and validation data, respectively, have a prediction error statistically equivalent to 0 due to a measured solubility value falling between the limits of predicted 95%CIs. The accuracy of model predictions are further confirmed by $R^2$ equal to 0.92 and 0.90 for calibration and validation, respectively.

To further verify the adequacy of the modelling approach proposed, 9 literature datasets of organic solubility have been employed. Such datasets are selected because they use pharma-related compounds and/or for the diverse chemical structures involved, as well as for the variety of conditions explored. Accurate results have been obtained, with $R^2$ between 0.82 and 0.998 and a percentage of 97-100% of model predictions having a prediction error equivalent to 0 due to corresponding measured values between the predicted 95%CIs. This prediction accuracy is achieved using a calibration dataset made of only 19-45% of the overall data available, suggesting that the experimental burden can be more than halved while keeping satisfactory performance.

Further work will focus on the application of the proposed modelling approach to new organic solvents. Moreover, the representation of phenomena that may occur in experimental screening of organic solubility, such as solid form changes in solvents (e.g., formation of hydrates or solvates), boiling out or gelling, is not currently taken into account in the modelling approach and it will be tackled in future work.

# Chapter 8
# Prediction of intestinal solubility: food effects and inter- and intra- subject variability[6]

In this work, a Gaussian Process model is developed in order to improve the prediction of biorelevant solubility data measured in vitro and to improve the overall PBPK simulations performed with the commercial software Simcyp® (Simcyp Simulator V20TM, Certara, UK). The model is calibrated and validated with recently published data on the solubility of a real drug substance. Prediction accuracy is greatly improved with respect to the conventional model implemented in Simcyp, thanks to a coefficient of determination $R^2$ equal to 0.97 and a root-mean-squared-error RMSE equal to 0.33mM, thus comparable to the standard deviation of measurement errors that can reach up to 0.10mM in the fasted state and 0.77mM in the fed state. These satisfactory results suggest that the proposed model can be used to substitute the conventional solubility model in Simcyp to improve the representation of inter- and intra-subject variability of intestinal solubility.

## 8.1 Introduction

Intestinal solubility is one of the key properties for an Oral Solid Dosage form. In fact, the efficacy and safety of an orally administered drug is related to its bioavailability (Abrahamsson et al., 2020), defined as the fraction of drug that enters the systemic circulation. To be available in the bloodstream, a solid form entering the Gastrointestinal (GI) tract must be absorbed, namely it must permeate across the gut wall. In turn, intestinal absorption is the result of an interplay of several phenomena, including: release of drug particles; dissolution; precipitation

[6] Cenci, F., Stamatopoulos, K. , Diab, S., Ferrini, P., Barolo, M., Bezzo, F. and Facco, P.. Machine-Learning approach to represent food effect and inter- and intra-subject variability of intestinal solubility [preparation].

(if solubility is exceeded); resolubilisation of precipitated particles; ionisation, based on drug acid-base properties; partition to micelles, based on drug lipophilicity, leading to an enhanced solubilisation (Stamatopoulos, 2022; see Appendix J for more details). As a consequence, solubility is a crucial property for intestinal absorption: if it is too low, it limits intestinal absorption; even when it is not the rate-limiting step, it can influence dissolution rate and precipitation (Augustijns et al., 2014).

Poor solubility is an issue in drug substance discovery, as well as in early and late phases of development, therefore solubility must be assessed from the early stages of compounds design and optimisation (Stegemann et al., 2007). This need is also in line with the QbD approach, which consists in focusing on product quality at the initial stage of the development. Solubility studies are especially important with recent molecules discovered through combinatorial chemistry and high throughput screening, which lead to increasing molecular weight and lipophilicity, thus decreasing aqueous solubility (Stegemann et al., 2007; Lipinski, 2000; Lipinski et al., 1997). Moreover, intestinal solubility cannot be determined uniquely through studies of aqueous solubility, because it has a high level of variability that depends on three main types of factors: physico-chemical properties of the drug molecule; properties of the formulation; the environment in the intestine (Abrahamsson et al., 2020). Relevant drug properties for intestinal solubility include chemical structure, lipophilicity and acidic, basic or neutral properties (Stamatopoulos, 2022; Ainousah et al., 2017). Moreover, formulation can enhance solubilisation through a proper design of excipients and/or solid-state form of the drug (Abrahamsson et al., 2020). Finally, different physiological and pathophysiological factors can alter intestine solubility and they may vary among different subjects, due to individual characteristics such as age, sex, race and diseases, and/or within the same subjects, e.g., location in the GI tract, fed or fasted state and/or meal composition (Abrahamsson et al., 2020; Salehi et al., 2021; Jamei et al., 2009). As a consequence, intestinal solubility is a range, not a single value (Abuhassan et al., 2022).

To assess intestinal solubility, both experimental and modelling approaches have been developed. The final aim is to characterise Human Intestinal Fluids (HIFs; Rosenberger et al., 2018; de la Cruz-Moreno et al., 2017), which change within the same individual and among individuals. As regards the single individual, HIFs can change in space (e.g., at different locations of the gastrointestinal tract) and in time (e.g., at the same point in different time instants), for instance due to changes of pH and of concentrations of the compounds deriving from endogenous excretions or food digestion (Pyper et al., 2020). As regards the differences

in HIFs samples coming from different individuals, they are mainly due to differences in physiological factors and in the different food, drink, and sampling protocols adopted. Although experimental data of HIFs are considered as the "gold standard", they cannot be used routinely because their extraction from human volunteers is difficult, invasive and expensive (Silva et al., 2022; Dahlgren et al., 2021). Therefore, Simulated Intestinal Fluids (SIFs) have been developed: they are biorelevant media used for in vitro experiments mimicking the behavior of HIFs. Several options for SIFs representing the fasted conditions have been proposed in the Literature, from the simplest ones made of a buffer, bile salt and lecithin, to more complex ones containing free fatty acid, monoglyceride and enzyme components. Similarly, different alternatives can be proposed for media representing fed conditions, for instance by changing buffers, composition and ratio of bile salts and lecithin and/or by adding monoglycerides or fatty acids. More details on fasted and fed media employed in the Literature can be found in Abuhassan et al. (2022) and Zhou et al., (2017), respectively. There is still no consensus on the most appropriate composition for biorelevant media to mimic HIFs in vitro. However, the main goal of SIFs should be representing the solubilising effects of HIFs on drugs, rather than their exact composition (Augustijns et al., 2014).

Since it is not feasible to experimentally test all the conditions that can be encountered in the human body of one or more individuals, mathematical models are useful to support drug development. A certain amount of experimental data is still needed to identify model equations and/or parameters (see Chapter 1), but the overall experimental effort can be reduced by means of a reliable model. For instance, PK profiles provide crucial information on drug bioavailability and clearance (Heller et al., 2018) and they are employed from drug discovery to drug development: in pre-clinical trials, to assess drug safety and dosing metrics (Heller et al., 2018); in drug development, to study a variety of aspects such as drug–drug interactions, targeted tissue exposure and disease effect (Yuan et al., 2022). Plasma-concentration profiles can be obtained through physiologically-based pharmacokinetic models, which represent different organs as compartments linked by the bloodstream. PBPK models allow to simulate the four main steps of body's interaction with drugs, namely absorption, distribution, metabolism and excretion, thanks to a combination of three key contributions: (*i*) data on drugs and formulations; (*ii*) parameters on species physiology; (*iii*) good understanding of the processes affecting drug properties (Zhuang and Lu, 2016). Complete PBPK simulations can be performed with commercial software like Simcyp. In particular, the Simcyp Population-based ADME Simulator allows to study not only the average individual, but populations with

inter-subject variability: the user can select healthy and diseased populations, as well as specific ethnic populations, and the software generates different virtual subjects in the population by varying demographic, genetic, anatomical and physiological factors within plausible ranges (Zhuang et al.,2016; Jamei et al., 2009). However, the current Simcyp model representing the solubility in biorelevant media considers only two input variables, namely pH and total bile salts concentration (BSs), while more variables should be included to represent the complex interactions within HIFs in fasted and fed conditions. In fact, some attempts have been made in the Literature to investigate the effects of multiple factors in vitro, mainly by designing the composition of biorelevant media through statistical design of experiments (DoE; Montgomery 2013). For instance, Zhou et al. (2017) collected 92 experiments by varying 8 factors based on a factorial DoE: pH, bile salt, lecithin, sodium oleate, monoglyceride, buffer, salt and pancreatin. Ainousah et al. (2017) selected 20 conditions among the ones designed by a factorial DoE with 8 factors: bile salt, lecithin, sodium oleate, monoglyceride, cholesterol, pH, and bile salt phospholipid molar ratio. Moreover, Zhou et al. (2017) considered a media containing a fixed total concentration of 4 amphiphiles and performed a 4 components Mixture Design to variate their ratio in the experiments. A similar approach was applied by Dunn et al. (2019), who performed 4 components Mixture Designs at 3 pH values and 3 total amphiphile concentrations, providing 351 experimental points. However, few attempts have been made to develop models that are able to accurately represent the effects of multiple components besides pH and BSs. This holds with respect to both mechanistic (or first-principles) models and data-driven (or black-box) models. As regards the former type of models, Henderson–Hasselbalch equations allow to predict solubility for monoprotic acids, monoprotic basis or ampholytes, but they express the dependence of solubility on pH only. Moreover, equations relating solubility to bile salts concentration are available, too, as described in Stamatopoulos (2022). However, such models do not distinguish among bile salts types and do not consider food digestion products. As regards data-driven approaches, several studies analyse data through statistical indices and/or tests without developing a proper model for solubility prediction. Statistical analysis often involves: tests to assess data normality or to compare the means of two groups; scalar indices, like median and percentiles, to represent viability of measured and/or simulated bioavailability; standardised effects values, to evaluate the statistical significance of every factor included in the study (Silva et al., 2023; Inês Silva et al., 2022; Pyper et al., 2020; Dunn et al., 2019; Perrier et al., 2018; Ainousah et al., 2017; Zhou et al., 2017; Jamei et al., 2009). Moreover, data-driven approaches have been applied by Augustijns et al., (2014), who

correlated solubility to one specific factor at a time, namely pH, total concentration of phospholipids or total concentration of bile acids, and by Rabbie et al. (2015), who calibrated a liner regression model to relate measured solubility to pH and buffer capacity. Fagerberg et al. (2015) developed a PLS model for solubility prediction using measured solubility values of 86 lipophilic drugs in fasted SIFs (FaSSIFs), HIFs and a phosphate buffer at pH 6.5. The original list of input variables (or regressors) was obtained considering molecular descriptors of the compounds chemical structure; then, only significant regressors were retained, namely only regressors having a significant variable importance in projection and impacting on model prediction accuracy. The proposed model provided a coefficient of determination $R^2$ between 0.69 and 0.86; the highest accuracy of solubility predictions in HIFs and FaSSIF was obtained by adding pH-dependent solubility as a descriptor, which is commonly available at early stages of drug development. However, this type of model has not been used within the PBPK modelling framework of Simcyp to predict PK profiles for all the virtual subjects of a simulated population.

In this work, a novel machine-learning model for the prediction of intestinal solubility is proposed. The aim is to improve the representation of the solubilising effects of human intestinal fluids in fasted and fed conditions, to allow the determination of inter- and intra-subject variability and to allow the validation of model predictions with human data commonly available at the early stages of drug development. To better represent the interactions encountered in vivo, solubility measurements of a real Active Pharmaceutical Ingredient, indicated as API A in this Chapter, in the biorelevant media proposed by Stamatopoulos et al. (2023) are considered. In this dataset, four different types of bile salts are considered to reflect inter-subject differences in the composition of human duodenal aspirates (De la Cruz Moreno et al., 2006; Riethorst et al., 2016). Moreover, oleic acid and cholesterol in fasted and fed levels are included to mimic food effects; finally, different values of pH are employed in the experiments. Besides the improved representation of food effects on solubility, the experimental data do not cover completely some conditions that may be encountered in vivo. Consequently, Gaussian Process regression is used since it inherently provides an estimation of prediction uncertainty. The GP model represents the dependence of solubility on bile salts, lecithin, oleic acid, cholesterol and pH and to have an estimation of prediction uncertainty. Besides the accuracy of the model in predicting in vitro solubility measurements, also the accuracy with respect to in vivo performance should be assessed. The GP model built with in vitro data cannot be validated directly with HIFs solubility data of the same API, since this type

of data is not available. However, plasma-concentration profiles of API A in healthy volunteers are available (Spinner et al., 2022; Johnson et al., 2022; Joshi et al., 2020; Riedmaier et al., 2020). Therefore, the GP model is defined in such a way as it can be integrated within the overall PBPK model implemented in the Simcyp simulator, thus allowing the prediction of plasma-concentration profiles. This integration gives also the opportunity to use the proposed GP model to improve the representation of virtual subjects in a population, therefore to improve the representation of inter- and intra-subject variability of the absorption of a drug. In this work, the GP model is built and analysed using the in vitro experiments in biorelevant media; ongoing work focuses on its final implementation and validation in Simcyp.

Section 8.2 illustrates the literature dataset considered and the solubility model developed. In Section 8.3, the performance of the proposed GP model is discussed and compared to the one of the conventional solubility model implemented in Simcyp. Finally, conclusions and future work are explained in Section 8.4.

## 8.2 Materials and methods

In this Section, the literature dataset employed to build and validate the solubility model is explained (Subsection 8.2.1). Moreover, the mathematical modelling approach developed to assess biorelevant solubility is explained (Subsection 8.2.2).

### *8.2.1 Intestinal solubility: in vitro experiments in biorelevant media*

API A is poorly soluble and poorly permeable . Moreover, it is a zwitterionic drug, therefore its intestinal solubility depends on its ionisation state in the pH range encountered in the GI tract and on the charge (or zeta potential) of the micelles (Takács-Novák et al., 2013; Stamatopoulos et al., 2023). Due to the poor solubility in water, volunteers involved in clinical trials ingested a moderate-fat meal before drug administration to increase its absorption; more details on the API and its clinical trials can be found in Joshi et al. (2020) and Spinner et al. (2022).

The in-vitro experimentation of Stamatopoulos et al. (2023) was performed using different solution types for both fasted and fed states, as shown in Table 8.1 and 8.2, respectively. The levels of bile salts (BS), lecithin (namelu, L-alpha-phosphatidylcholine from egg yolk, indicated as PC), oleic acid (OA) and cholesterol (CH) in the fasted state are equal to, respectively: 3 mM, 0.75 mM, 0.53 mg/mL, 0.027 mg/mL. In the fed state they are equal to, respectively: 15 mM, 3.75 mM, 6.50 mg/mL and 0.72 mg/mL. Notice that pairs of solution

types indicated by the same number and the letters “a” and ”b” (i.e., solutions 6-10 in the fasted state, Table 8.1, and solutions 3-4 in the fed state, Table 8.2) have equal amounts of total BS, PC, OA and CH, but they have different proportions of the four BS types: namely sodium taurocholate hydrate (TC), sodium taurochenodeoxycholate (TCDC), sodium glycocholate hydrate (GLC), sodium glycodeoxy-cholate (GDC).

**Table 8.1.** *Description of the dataset of API A solubility in biorelevant media representing the fasted state. All data points are measured at a nominal pH of 6.5 (Stamatopoulos et al., 2023).*

| Solution | Total BS (mM) | TC (%) | TCDC (%) | GLC (%) | GDC (%) | PC (mM) | OA (mg/mL) | CH (mg/mL) |
|---|---|---|---|---|---|---|---|---|
| 1 | 3 | 100.0 | 0.0 | 0.0 | 0.0 | 0.000 | 0.000 | 0.000 |
| 2 | 3 | 100.0 | 0.0 | 0.0 | 0.0 | 0.750 | 0.000 | 0.000 |
| 3 | 3 | 100.0 | 0.0 | 0.0 | 0.0 | 0.750 | 0.000 | 0.027 |
| 4 | 3 | 100.0 | 0.0 | 0.0 | 0.0 | 0.750 | 0.530 | 0.027 |
| 5 | 3 | 100.0 | 0.0 | 0.0 | 0.0 | 0.750 | 0.530 | 0.000 |
| 6a | 3 | 14.9 | 15.1 | 45.1 | 24.9 | 0.000 | 0.000 | 0.000 |
| 6b | 3 | 47.1 | 8.8 | 24.6 | 19.5 | 0.000 | 0.000 | 0.000 |
| 7a | 3 | 14.9 | 15.1 | 45.1 | 24.9 | 0.750 | 0.000 | 0.000 |
| 7b | 3 | 47.1 | 8.8 | 24.6 | 19.5 | 0.750 | 0.000 | 0.000 |
| 8a | 3 | 14.9 | 15.1 | 45.1 | 24.9 | 0.750 | 0.000 | 0.027 |
| 8b | 3 | 47.1 | 8.8 | 24.6 | 19.5 | 0.750 | 0.000 | 0.027 |
| 9a | 3 | 14.9 | 15.1 | 45.1 | 24.9 | 0.750 | 0.530 | 0.027 |
| 9b | 3 | 47.1 | 8.8 | 24.6 | 19.5 | 0.750 | 0.530 | 0.027 |
| 10a | 3 | 14.9 | 15.1 | 45.1 | 24.9 | 0.750 | 0.530 | 0.000 |
| 10b | 3 | 47.1 | 8.8 | 24.6 | 19.5 | 0.750 | 0.530 | 0.000 |

**Table 8.2.** *Description of the dataset of SPI A solubility in biorelevant media representing the fed state. All the solutions are measured at three nominal pH values: 5, 6.5 and 7. The asterisk (*) indicate the solutions that do not contain PC in the measurements at nominal pH equal to 5 (Stamatopoulos et al., 2023).*

| Solution | Total BS (mM) | TC (%) | TCDC (%) | GLC (%) | GDC (%) | PC (mM) | OA (mg/mL) | CH (mg/mL) |
|---|---|---|---|---|---|---|---|---|
| 1 | 15 | 100.0 | 0.0 | 0.0 | 0.0 | 3.75 | 0.00 | 0.00 |
| 2 | 15 | 100.0 | 0.0 | 0.0 | 0.0 | 3.75 | 6.50 | 0.72 |
| 3a* | 15 | 14.9 | 15.1 | 45.1 | 24.9 | 3.75 | 0.00 | 0.00 |
| 3b* | 15 | 47.1 | 8.8 | 24.6 | 19.5 | 3.75 | 0.00 | 0.00 |
| 4a | 15 | 14.9 | 15.1 | 45.1 | 24.9 | 3.75 | 6.50 | 0.72 |
| 4b | 15 | 47.1 | 8.8 | 24.6 | 19.5 | 3.75 | 6.50 | 0.72 |

Biorelevant media representing the fasted state (Table 8.1) are measured at a nominal pH equal to 6.5; three replicates are measured for every experimental condition. Therefore, 45 data points are available. Instead, every solution type of the fed state (Table 8.2) is measured at three nominal pH values: 5, 6.5 and 7. During the experimentation, in the time lapse between the addition of the liquid to the vial with the solid and the achievement of solid-liquid equilibrium, pH may be subjected to variations, thus the actual pH slightly oscillates around the abovementioned nominal values (the variance of the measured pH values around the same

nominal value are between 0 and $10^{-3}$). The pH values at equilibrium are measured for the fasted conditions and they are the ones used to build the GP model (Section 8.2.2). Also in the fed dataset, three replicates are performed for every condition; one experiment does not have a valid measurement. Therefore, 53 data points are available.

Both fasted and fed experiments are performed with a high-throughput technology that uses 24-vials plate, therefore it allows to measure 24 experimental conditions in parallel. During the experiments, the automated powder dispensing platform dispenses a controlled weight of solid powder in each vial; the liquid medium is added manually, but the rest of the experiment is carried out by an automated liquid handling platform: it mixes samples, it controls the system temperature and it performs the sampling at the scheduled time points. More details on the materials and experimental setup employed can be found in Stamatopoulos et al. (2023).

## *8.2.2 Mathematical modelling to predict intestinal solubility*

This section illustrates the mathematical models that are compared and discussed in Section 8.3. First, the state-of-the-art model for intestinal solubility implemented in Simcyp is explained, focusing on the types of input and output variables that can be simulated in Simcyp. Then, the features of the GP model proposed in this work are presented.

### 8.2.2.1 State-of-the-art solubility model

The model implemented in Simcyp to predict intestinal solubility considers the sum of different contributions:

$$S_T = S_{\mathrm{pH}} + S_{\mathrm{BS,u}} + S_{\mathrm{BS,i}}, \tag{8.1}$$

where $S_{\mathrm{pH}}$ is the drug solubility in the luminal fluids depending on pH, $S_{\mathrm{BS,u}}$ and $S_{\mathrm{BS,i}}$ represent the bile-salt mediated enhancement of solubility for unionised and ionised species, respectively, while $S_T$ is the total solubility (namely, intestinal solubility).

The solubility $S_{\mathrm{pH}}$ of the drug in the lumen is calculated as:

$$S_{\mathrm{pH}} = S_{\mathrm{pH,i}} + S_0, \tag{8.2}$$

where $S_0$ indicates the intrinsic solubility of the drug, while $S_{\mathrm{pH,i}}$ is the pH-dependent solubility of ionised species. The API considered is an ampholyte, namely it may act as an acid or as a base, therefore the expression of $S_{\mathrm{pH,i}}$ corresponds to:

$$S_{\mathrm{pH,i}} = S_0(10^{\mathrm{pH}-\mathrm{pK}_{a,1}} + 10^{\mathrm{pK}_{a,2}-\mathrm{pH}} + 10^{\mathrm{pK}_{a,2}-\mathrm{pK}_{a,1}}), \tag{8.3}$$

where $\mathrm{pK}_{a,1}$ and $\mathrm{pK}_{a,2}$ are the two dissociation constants (see Appendix J), while pH changes in different segments of the GI tract.

The bile-salt mediated enhancement of solubility $S_{\mathrm{BS,u}}$ and $S_{\mathrm{BS,i}}$ are calculated as:

$$S_{\mathrm{BS,u}} = [\mathrm{BS}] \frac{S_0}{[\mathrm{H_2O}]} K_{m:w,u} \,, \tag{8.4}$$

$$S_{\mathrm{BS,i}} = [\mathrm{BS}] \frac{S_0}{[\mathrm{H_2O}]} K_{m:w,i} \,, \tag{8.5}$$

where $[\cdot]$ indicates the concentration, thus [BS] is the bile salts concentration and $[\mathrm{H_2O}]$ is water concentration, namely 55560mM. Moreover, $K_{m:w,u}$ and $K_{m:w,i}$ are the water-to-micelle partition coefficients of unionised and ionised species, respectively (see Appendix J for more details). However, if [BS] is below the critical micelles concentration (CMC) of 1 mM, $S_{\mathrm{BS,u}}$ and $S_{\mathrm{BS,i}}$ (Eq.s 8.4, 8.5) are set equal to 0.

The values of $S_0$, $\mathrm{pK}_{a,1}$, $\mathrm{pK}_{a,2}$, $K_{m:w,u}$ and $K_{m:w,i}$ depend on the system under study; the most suitable values for API A (Sub-section 8.2.1), are retrieved from Stamatopoulos et al. (2023) and shown in Table 8.3.

**Table 8.3.** *Properties of API A used to predict total solubility in Simcyp (Stamatopoulos et al., 2023).*

| Drug properties | Value |
|---|---|
| $S_0$(mg/mL) | 0.00015 |
| $\mathrm{pK}_{a,1}$ | 4.63 |
| $\mathrm{pK}_{a,2}$ | 8.65 |
| $K_{m:w,u}$ | 0.019 |
| $K_{m:w,i}$ | 4.559 |

As introduced in Section 8.1, only two inputs are included in the model of Eq. (8.1): pH and [BS]. These are the variables that change across the GI tract or at the same location in different time instants during Simcyp simulations. Therefore, the GP model developed in this work should adapt to these possibilities, as explained in Sub-subsection 8.2.2.2.

### 8.2.2.2 Proposed GP model

The experimental dataset of API A contains 8 different input variables: pH and concentrations of TC, TCDC, GLC, GDC, PC, OA, CH. To model their effects and, at the same time, have a model suitable for PBPK simulations in Simcyp, an apparent bile salts concentration is considered:

$$[\mathrm{BS}]_{\mathrm{app}} = [\mathrm{TC}] + [\mathrm{TCDC}] + [\mathrm{GLC}] + [\mathrm{GDC}] + [\mathrm{PC}] + [\mathrm{OA}] + [\mathrm{CH}] \,, \tag{8.6}$$

where all concentrations are expressed as mM. Consequently, $[BS]_{app}$ ranges between 3 mM and 43.6 mM: experimental points representing fasted conditions are between 3 and 5.7 mM, while those of fed conditions are between 15 and 43.6 mM.

Therefore, pH and $[BS]_{app}$ (in mM) are the two inputs of the model, while the output of the model is the API solubility expressed as mM.

In turn, these input-output data are used to calibrate a GP model characterised by the mean $\bar{\mathbf{f}}_*$ and covariance $\text{cov}(\mathbf{f}_*)$ functions of Eq. (2.36, 2.37), considering the squared-exponential function (Eq. 2.34) as kernel function. The best parameters $\boldsymbol{\theta}^* = \{\sigma_{SE}^2, \ell, \sigma_y^2\}$ are obtained by using the GPy package (Sheffield machine learning group, 2020); the marginal likelihood (Eq. 2.40) is optimised by using the default algorithm of GPy, i.e., L-BFGS-B (Byrd et al., 1995).

If a prior mean equal to zero is assumed for the GP model (as common practice), the results may be unsatisfactory in interpolation, especially at experimental conditions relatively far from the ones explored in vitro. Figure 8.1 shows the mean solubility values [mM] predicted by the GP model having a prior mean equal to zero and built with a training dataset made of all data of fasted and fed states. Results are not as expected in the range of $[BS]_{app}$ between 0 and 50 mM: the solubility is expected to increase with $[BS]_{app}$, but the values predicted by the GP model between (approximately) 5 and 15 mM and between 20 and 40 mM (thus, between the levels explored experimentally) are close to 0 mM.

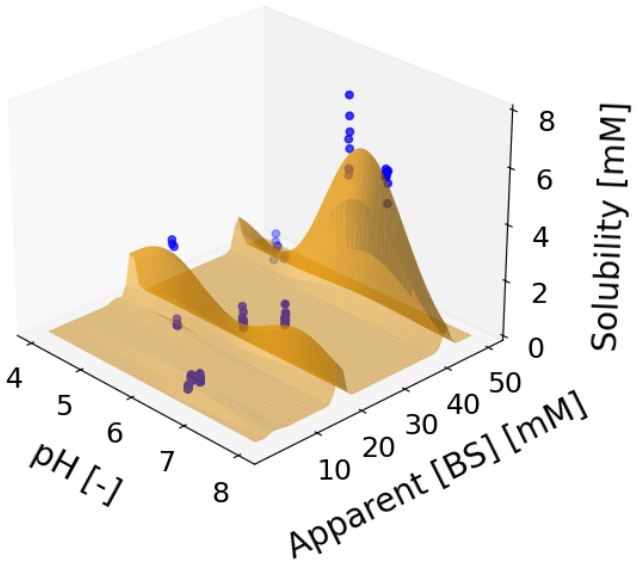



**Figure 8.1**. *Mean solubility values predicted by the GP model having prior mean equal to zero. Blue dots represent experimental data of API A in fasted and fed states.*

To improve the model performance, a prior mean function is specified. A simple linear model is sufficient to express the increasing or decreasing trend of solubility with respect to pH and

$[\mathrm{BS}]_{\mathrm{app}}$, while possible non-linearities are represented by the posterior mean of the GP model. The step-by-step procedure to calibrate the GP model with a specified prior mean function is:

- step 1: calibrate a multi-linear regression model to approximate the dependence of solubility on pH and $[\mathrm{BS}]_{\mathrm{app}}$; the approximated solubility values predicted by the linear regression model are indicated as $\hat{\mathbf{y}}_{\mathrm{LR}}$;
- step 2: rescale the measured solubility values by subtracting the values predicted by the linear model;
- step 3: use the rescaled solubility values, namely $\mathbf{y} - \hat{\mathbf{y}}_{\mathrm{LR}}$, to calibrate a Gaussian Process model. Use the assumption of prior mean function equal to zero and find the best hyperparameters $\boldsymbol{\theta} = \{\sigma_{\mathrm{SE}}^2, \ell, \sigma_y^2\}$ by maximising the maximum likelihood of Eq. (2.40). Then, the mean and covariance functions can be calculated (Eq.s 2.36, 2.37). The mean solubility value predicted by the GP model is indicated as $\hat{\mathbf{y}}_{\mathrm{GP}}$ from now on;
- step 4: make predictions of drug solubility $\hat{\mathbf{y}}$ summing up the two contributions (as in Eq. (2.41)), namely:

$$\hat{\mathbf{y}} = \hat{\mathbf{y}}_{\mathrm{LR}} + \hat{\mathbf{y}}_{\mathrm{GP}} \tag{8.7}$$

The multi-linear regression model considered for this purpose is:

$$\hat{y}_{\mathrm{LR}} = \beta_1 \mathrm{pH} + \beta_2 [\mathrm{BS}]_{\mathrm{app}} , \tag{8.8}$$

The linear model leading to the most reliable GP model is retained, as discussed in Subsection 8.3.1.

Finally, the performance of the GP model is evaluated in terms of prediction precision and accuracy. Prediction precision is estimated in terms of 95% confidence intervals (CIs) using the posterior covariance function $\mathrm{cov}(\mathbf{f}_*)$ (see Chapter 2). Moreover, model prediction accuracy is calculated through the coefficient of determination $R^2$ (Eq. 2.14) and the root mean squared error (RMSE, Eq. 2.12).

## 8.3 Results and discussion

In this section, the GP model is calibrated using biorelevant solubility experiments in fasted and fed conditions. Section 8.3.1 compares the performance of the conventional solubility model implemented in Simcyp with the one of the proposed GP model. Ongoing work focuses on the validation of the proposed GP model with new experiments; in this work, the modelling approach proposed is validated by calibrating the model with a subset of the in vitro

experiments available and evaluating the performance with the remaining ones (Section 8.3.2). The implementation in Simcyp is preliminarily discussed in Section 8.3.3. and will be concluded in future work.

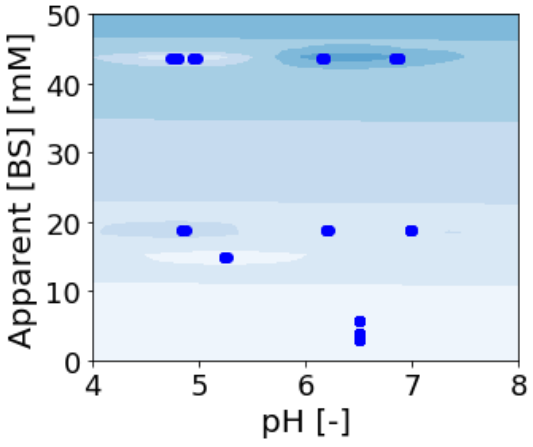
### *8.3.1 Model calibration*

The GP model described in Subsection 8.2.2 is built with in vitro solubility experiments of API A in fasted and fed states (Tables 8.1 and 8.2, Subsection 8.2.1). First, the most suitable linear approximation of the system must be chosen between Eq. (8.8) and (8.9). A basic requirement for the overall GP model is that the mean predicted solubility (namely, $\hat{\mathbf{y}} = \hat{\mathbf{y}}_{\mathrm{LR}} + \hat{\mathbf{y}}_{\mathrm{GP}}$) is non-negative. Table 8.4 shows the parameters obtained. Parameters of the linear models are obtained through Maximum Likelihood Estimation (MLE, Eq. (2.2)) and the optimisation is carried out in Python 3.9 by using the Nelder-Mead algorithm of the scipy.m package. Parameters of the GP regression are optimised with the default method of the GPy package.

**Table 8.4.** *Parameters estimated for the overall multi-linear regression and GP model.*

| **Model** | $\beta_1$ | $\beta_2$ | $\sigma_{SE}^2$ | $\ell$ | $\sigma_y^2$ |
|---|---|---|---|---|---|
| $\hat{\mathbf{y}} = \hat{\mathbf{y}}_{\mathrm{LR}} + \hat{\mathbf{y}}_{\mathrm{GP}}$ | 0.010 | 0.085 | 1.516 | 0.604 | 0.126 |

The GP model obtained (Table 8.4) is used to predict solubility at pH values between 4 and 8 and at $[\mathrm{BS}]_{\mathrm{app}}$ values between 0 and 50 mM. The contour plots of Figure 8.3 show the results in terms of mean predicted values obtained with Eq. (8.7), namely $\hat{\mathbf{y}} = \hat{\mathbf{y}}_{\mathrm{LR}} + \hat{\mathbf{y}}_{\mathrm{GP}}$.

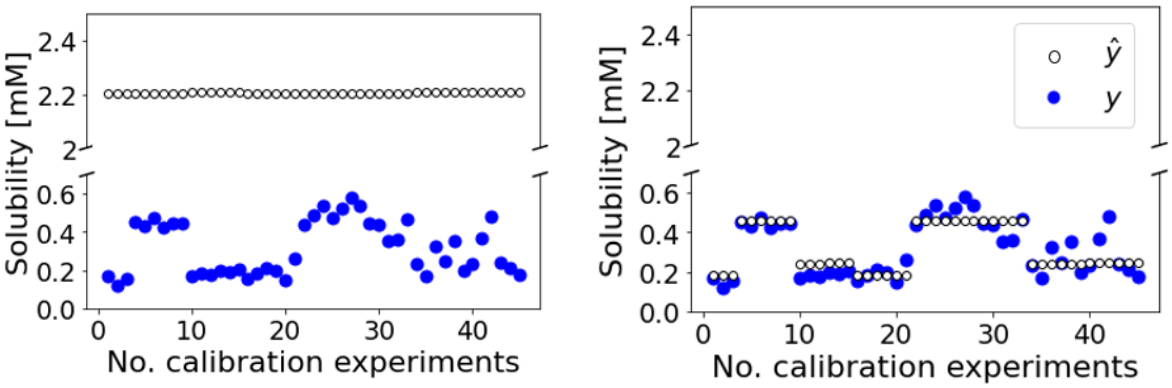



(a)

**Figure 8.3**. *Contour plots representing the predicted values of drug solubility [$\hat{\mathbf{y}}$, mM] with the overall GP model: $\hat{\mathbf{y}} = \hat{\mathbf{y}}_{LR} + \hat{\mathbf{y}}_{GP}$. Blue dots represent experimental conditions of API A in fasted and fed states.*

Figure 8.3 shows that the predicted solubility is always positive, therefore the basic requirement of the model is satisfied.

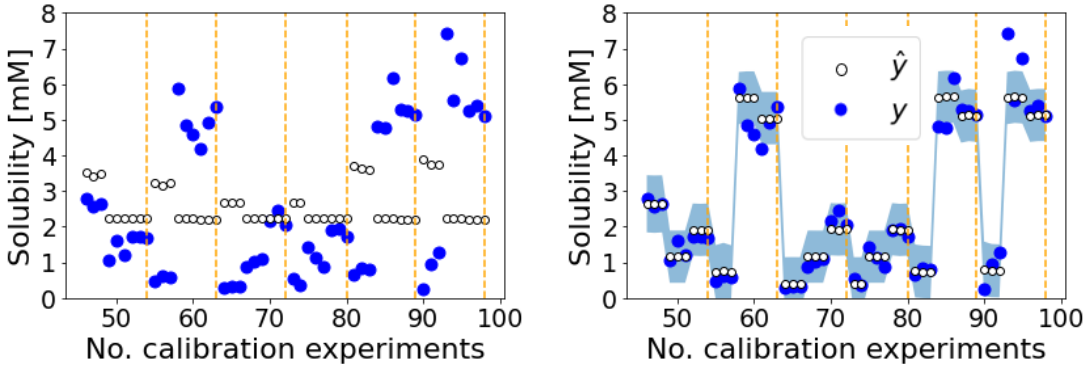
The performance of the conventional model (Eq. 8.1) and of the proposed GP model (Eq. 8.7) are compared in Figure 8.4 with data in the fasted state and in Figure 8.5 with data in the fed state. In the latter case, vertical dotted lines are used to separate the 6 solution types presented in Table 8.2 (section 8.2.1); for the same solution type, the measures obtained at the three nominal pH values (namely, 5, 6.5 and 7) are included.




(a) (b)

**Figure 8.4**. *Comparison between measured solubility values* **y** *and predicted solubility values* $\hat{\mathbf{y}}$ *obtained with: (a) conventional solubility model (Eq. 8.1); (b) proposed GP model (Eq.s 8.7). Only the fasted state is considered.*

As shown in Figure 8.4a, the prediction of biorelevant solubility in fasted conditions are, on average, 7 times higher than the predicted values if the conventional model is employed. Hence, the proposed GP model considerably improves the prediction accuracy, as proved by the predicted points very close to the measured ones in Figure 8.4b. This is further confirmed by the fact that the RMSE obtained with the GP model is equal to 0.065 mM in the fasted state, thus comparable to or even smaller than the standard deviation of the measurement errors which can reach up to 0.101 mM in the fasted state.




(a) (b)

**Figure 8.5**. *Comparison between measured solubility values* **y** *and predicted solubility values* $\hat{\mathbf{y}}$ *obtained with: (a) conventional solubility model (Eq. 8.1); (b) proposed GP model (Eq. 8.7). Only the fed state is considered. Vertical dotted lines separate different solution types (see Table 8.2).*

As shown in Figure 8.5a, the conventional solubility model does not have a satisfactory accuracy. Even the trend of solubility is not well represented. For instance, solution type 2 (namely, calibration experiments no. 55-63) is measured at three nominal pH values, each of which has three replicates. The first three points of this solution (i.e., calibration experiments no. 55-57), which are measured at the smallest pH, are characterised by a relatively small solubility value ($y$ <1 mM), while the remaining 6 points at higher pH have higher measured solubility ($y$ >4 mM). However, solubility predictions through the conventional model of Eq. (8.1) display an opposite behaviour: the highest predicted solubility ($S_T$ ≅3.2 mM) is found at the smallest pH and then it decreases ($S_T$ ≅2.2 mM) at higher pH values. Analogous results are found with the subsequent 4 solution types.

Instead, the GP model is able to represent the correct trend of solubility variation based on pH and $[\mathrm{BS}]_{\mathrm{app}}$ for all solution types (Figure 8.5b). Similarly to the results in the fasted state, the RMSE is comparable to or smaller than the standard deviation of the experimental replicates: in fact, the RMSE in the fed state is equal to 0.45mM, while the standard deviation of the measurement errors can reach up to 0.77mM.

Considering both fasted and fed data, GP model predictions have a coefficient of determination $R^2$ equal to 0.97 and RMSE equal to 0.33 mM, therefore prediction accuracy is satisfactory and the model is expected to improve the overall PBPK simulation in Simcyp. Besides the improved accuracy, the GP model has the advantage of providing an estimation of model prediction uncertainty through the calculation of the covariance function (Eq. 2.37). The predicted 95%CIs for the fasted and fed states are shown in Figure 8.6.

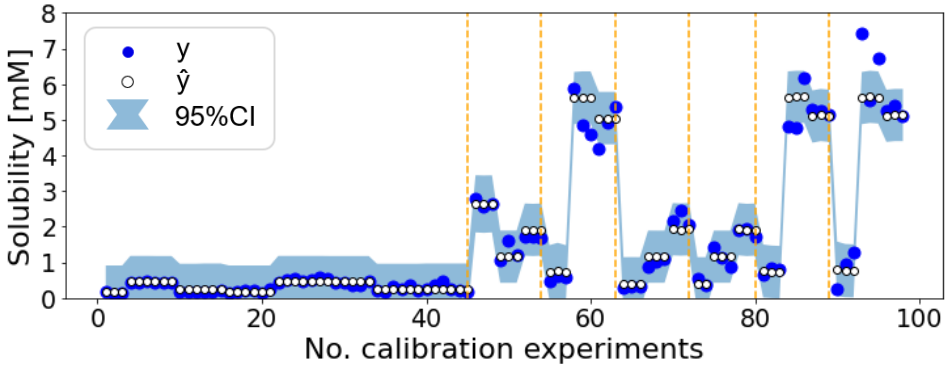



**Figure 8.6**. *Comparison between measured solubility values* **y** *and predicted solubility values* $\hat{\mathbf{y}}$ *obtained with the proposed GP model (Eq. 8.7) in both fasted and fed states. Blue shaded areas represent 95%CIs (Eq. 2.38), vertical dotted lines delimit solution types of fed state having three nominal pH values each.*

Results in Figure 8.6 show that the uncertainty of model predictions contains the majority of measured data (93% of the data), especially in the fasted state, meaning that prediction errors are statistically negligible. A few points in the fed state exceed the 95%CIs, but this happens only at experimental conditions having high variability within replicates.

Based on this result, the proposed GP model is deemed adequate for the integration into the Simcyp PBPK model. At first, the predicted mean solubility (Eq. 8.7) will be used in the simulations; in a second step, the predicted covariance function will be used to simulate random variations in solubility. For visualisation purposes, the mean function (Eq. 8.7) that will be implemented in Simcyp is displayed in Figure 8.7 as a 3-D surface.

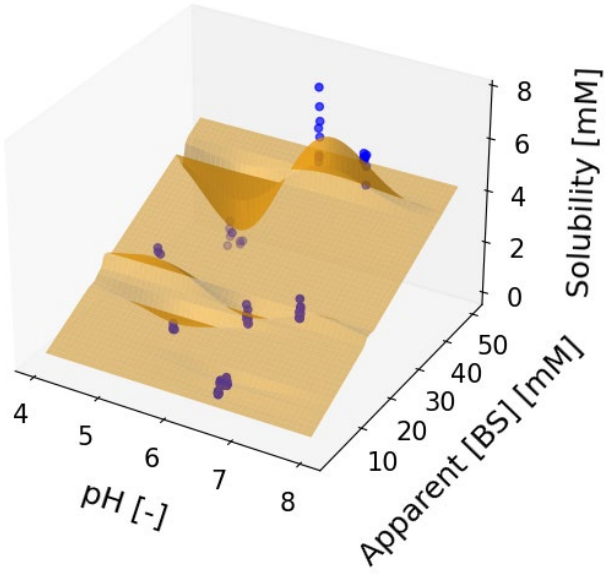



**Figure 8.7**. *Mean predicted solubility values obtained with the proposed GP model (Eq. 8.7) in the range [4,8] for pH and [0,50] mM for the apparent [BS]. Blue dots represent in vitro data in fasted and fed states.*

As shown in Figure 8.7, non-linearities are present, especially when the effect of pH on solubility is considered (as can be seen in the surface representing the model at pH between 4 and 8 in the Figure). This non-linearity cannot be predicted in regions quite far from the in vitro data, such as $[BS]_{app}$ between 20 and 40 mM, but the predicted values are higher than 0 and an increasing trend is predicted for solubility at increasing values of $[BS]_{app}$, as expected. Therefore, model predictions in these regions are greatly improved with respect to the ones obtained with the GP model without prior mean function (see Figure 8.1, section 8.2.2.2).

A validation of the proposed modelling approach based on GP regression is discussed in section 8.3.2.

### *8.3.2 Model validation*

To further assess the adequacy of the proposed GP model, future work will consist in the collection of new measurements in unexplored experimental conditions, such as in the region with $[BS]_{app}$ between 20 and 40mM.

Before performing new experiments, the modelling approach proposed in Figure 8.2 (Sub-subsection 8.2.2.2) is validated with the available literature data, by calibrating the model with a subset of data points. The fasted dataset is reduced by considering only the following solution types for calibration: a) BS+PC; b) BS+PC+OA+CH. The former is selected because bile salts and lecithin are often encountered in literature as biorelevant media; the latter is chosen because it includes all the novel elements of the chosen dataset, namely the species mimicking food effects. Therefore, the number of data points of the fasted state used in calibration is 18 over 45. The fed dataset, which contains a smaller number of solution types (see Table 8.2) with respect to the fasted one, is reduced considering all solution, types but excluding data at the intermediate value of pH. Therefore, 35 data points over the 53 representing fed conditions are used for calibration. Overall, 54% of the literature data are used to calibrate the GP model (Eq. 8.7) and the remaining 46% is used to validate it.

Table 8.5 shows the parameters optimised for the linear model and the GP regression, while predicted and measured data are compared in Figure 8.8 considering both calibration and validation data.

**Table 8.5.** *Parameters estimated for the overall GP model calibrated with the reduced dataset.*

| $\beta_1$ | $\beta_2$ | $\sigma_{SE}^2$ | $\ell$ | $\sigma_n^2$ |
|---|---|---|---|---|
| 0.010 | 0.085 | 1.895 | 0.616 | 0.051 |

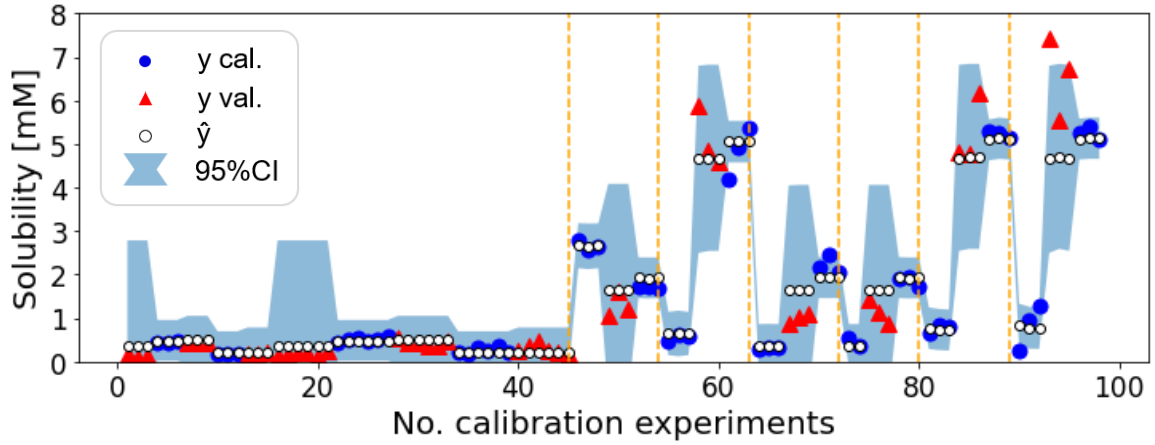



**Figure 8.8**. *Mean predicted solubility values obtained with the proposed GP model (Eq. 8.7) calibrated with the reduced dataset. Both calibration and validation data are shown. Different solutions in the fed state are delimited by vertical dotted lines.*

As shown in Figure 8.8, the accuracy in validation is still satisfactory besides the reduction of the calibration dataset. This is further confirmed by a $R^2$ close to 1 and RMSE comparable to or smaller than the standard deviation of the measurement errors (up to 0.101mM in the fasted state and up to 0.77mM in the fed state): $R^2$ and RMSE are respectively equal to 0.985 and 0.208 in calibration and they are respectively equal to 0.905 and 0.657 in validation.

Considering intermediate pH values in the fed state, the model is able to correctly represent the measured variation in solubility. For instance, the first solution type (calibration experiments no. 46-54; see Table 8.2) has higher solubility values in the first 3 measurements, then it decreases in the remaining 6 and this is well predicted by the model. Moreover, solution types 2, 4a, 4b (calibration experiments no. 55-63, 81-89, 90-98, respectively; see Table 8.2) have small solubility values in the first 3 experiments, then it considerably increases in the remaining 6 points and also in this case the model is able to represent this variation appropriately. The GP model tends to slightly overpredict the mean of the solubility values at intermediate pH values of solution types 3a and 3b (calibration experiments no. 64-72, 73-80, respectively; see Table 8.2), however the performance is still satisfactory: in fact, the measured values are within the predicted 95% confidence intervals; moreover, the predicted solubility at intermediate pH is higher than the one predicted at the smallest pH (first 3 experiments of the specific solution type) and smaller than one predicted at the highest pH value (last 3 experiments), thus coherent with the increasing trend of solubility encountered with in vitro data.

Overall, this validation is satisfactory and suggests that the modelling approach proposed allows to appropriately represent the dependence of solubility on pH, bile salts and food components also at experimental conditions not used for calibration.

### *8.3.3 Discussion of the implementation in Simcyp*

The GP model calibrated with the entire fasted and fed dataset is suitable for the implementation in Simcyp, because it predicts mean values having an accuracy comparable to the measurement errors and its 95% confidence intervals contain the majority of experimental measurements. Moreover, only two input variables are used, therefore the model can be integrated without requiring modifications of the conventional PBPK simulations with Simcyp. Therefore, ongoing work is focusing on the implementation of the proposed GP model in Lua (Ierusalimschy et al., 2005; PUC-Rio, Brazil), which is the programming software available in Simcyp to customise mathematical models.

Based on the results discussed in Subsections 8.3.1 and 8.3.2, the GP model is expected to provide more reliable solubility estimation with respect to the conventional model. On the other side, a possible limitation may be the fact that $[BS]_{app}$ used to calibrate the model ranges between 3 mM and 43.62 mM. In fact, that range is physiologically meaningful, since HIFs samples extracted from 20 volunteers displayed BS concentrations between 0 and 100 mM in the work of Riethorst et al. (2016); however, PBPK simulations usually consider [BS] values in the small intestine between 0 and 15 mM. The the higher values for the $[BS]_{app}$ used in calibration may have a negative impact on the final PBPK predictions. If this is the case (ongoing work is focusing on the verification of this aspect), a modification to the proposed GP model can be done: instead of considering $[BS]_{app}$ as defined in Eq. (8.6), the GP model can be built following the procedure of Figure 8.2 (Sub-subsection 8.2.2.2), but using [BS] only, namely the sum of the 4 bile salts types TC, TCDC, GLC and GDC. Figure 8.9 compares the pH and bile salts concentrations used as input of the GP models and simulated by Simcyp in the first segment of the small intestine, namely Duodenum.

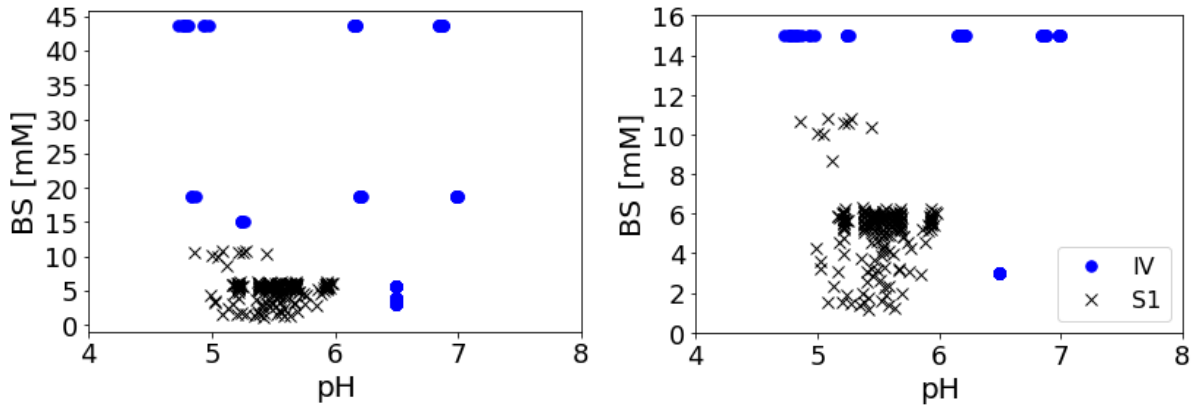



**Figure 8.9**. *Input values: pH and bile salts concentration. In the legend, 'IV' stands for 'in vitro' data, while 'S1' stands for 'segment 1' of the small intestine. In vitro data used to calibrate the GP models are displayed: a) apparent bile salts concentration, namely [TC]+[TCDC]+[GLC]+[GDC]+[PC]+[OA]+[CH]; b) total bile salts concentration, namely [TC]+[TCDC]+[GLC]+[GDC].*

As shown in Figure 8.9b, the concentration made by the sum of bile salts only is closer to the values typically encountered in PBPK simulations. However, if [BS] only is used to build the GP model, the second input variable has only two levels with the selected literature data, namely 3 mM for the fasted state (Table 8.1) and 15 mM for the fed state (Table 8.2), regardless of the fact other solubilising factors such as OA and CH are present. This leads to a considerable increase of solubility variability associated to a given level of [BS]: for instance, the variance $\sigma_y^2$ of the solubility measurements in the fed state (namely, at [BS]=15mM) at pH≅6.2 is equal

5.507 $mM^2$ (i.e., $\sigma_y$ =2.35 mM). Even though the mean change of solubility values due to the presence of PC, OA and CH cannot be predicted with this type of input variables, the GP model can still account for their effect on solubility through the estimation of the covariance function (Eq. 2.37). This can be seen in Figure 8.10, where solubility values are predicted at BS concentrations of 3 mM and 15 mM for a range of pH values equal to [4,8]. To better represent the variability of this dataset, the variance of the experimental data is fixed in the GP model (namely $\sigma_y^2$ is set equal to 5.507 $mM^2$ in Eq.s 2.37, 2.41), and only the hyperparameters $\sigma_{SE}^2$ and $\ell$ are optimised to obtain the GP model of Figure 8.10.

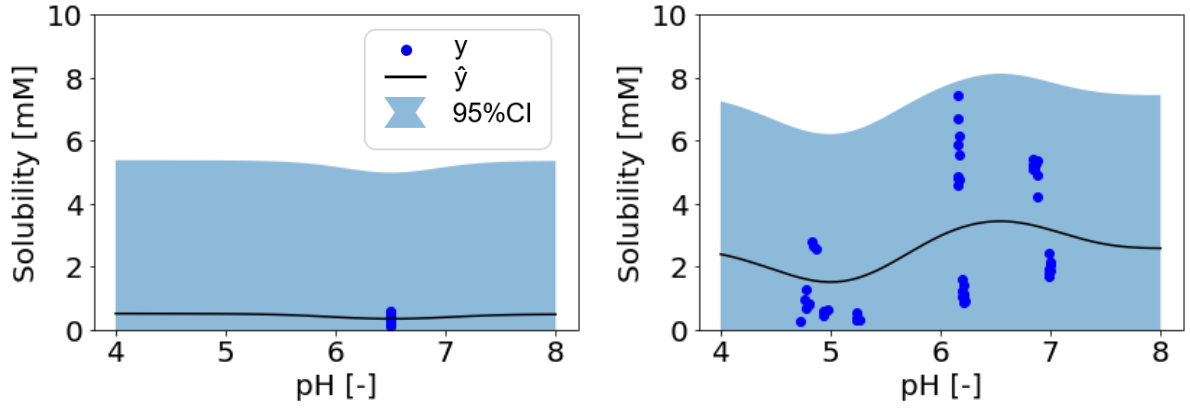



**Figure 8.10**. *Solubility predictions in the (a) fasted and (b) fed states obtained with the GP model calibrated with pH and [BS]= [TC]+[TCDC]+[GLC]+[GDC]. The black line represents the posterior mean function of the GP model; the light-blue shaded area represents the predicted 95%CI; the blue dots represent in vitro measurements in (a) fasted and (b) fed states.*

As shown in Figure 8.10, the GP model represents the variability of the fed state correctly at different values of pH (Figure 8.10b), but the uncertainty in the fasted state reaches values that are quite higher than the measured ones (Figure 8.10a). However, the mean predicted values (black line) are able to catch the correct order of magnitude of measured solubility at both fasted and fed states, therefore this GP model still improves the prediction with respect to the conventional solubility model implemented in Simcyp (see section 8.3.1). Consequently, the GP model proposed in Sub-subsection 8.2.2.2 and discussed in Subsection 8.3.1 remains the first choice for the implementation in Simcyp, but the variation proposed in this section can be taken into account if the results are not as expected due to apparent bile salts concentrations that are higher than the [BS] values simulated in the GI tract.

## 8.4 Conclusions and future work

Intestinal solubility is a crucial property to assess the bioavailability of oral solid dosage forms. Different biorelevant media have been proposed in literature to mimic the solubilizing effects of human intestinal fluids, but there is no consensus on the best composition. Moreover, commercial software like Simcyp are widely used in the pharmaceutical industry to simulate drug efficacy and safety besides inter- and intra-subject variations of demographic, genetic and/or physiological factors, but solubility models currently employed in such software are not suitable to represent the complex interactions occurring in HIFs, especially after food intake.

In this work, a novel machine-learning approach has been developed to accurately represent drug solubility in fasted and fed conditions. To do so, a recently published dataset on a real API is used; besides common factors such as pH, bile salts (BS) and lecithin (PC), this dataset allows to study food effects, thanks to the presence of oleic acid (OA) and cholesterol (CH), and the effects of different proportions of 4 bile salts species. This dataset is used to calibrate a Gaussian Process (GP) model with a prior mean function made of a linear model. The GP model is built in such a way as it can be integrated within Simcyp simulator: since only pH and total BS concentration are simulated in the PBPK model, an apparent BS concentration is considered as input of the GP model (besides pH) by summing up the concentrations of BS, PC, OA and CH. The integration into Simcyp is important for two main reasons: (1) validation with bioavailability data from human volunteers in clinical trials; (2) representation of inter- and intra-subject variability in intestinal solubility. As regards the former, comparing simulated and in vivo data is important during drug development, but samples of HIFs are often scarce, therefore intestinal solubility cannot be measured directly in an extensive way. On the other side, plasma-concentration profiles are often retrieved from human volunteers in clinical trials, therefore they can be used as a reference to assess the reliability of the plasma-concentration profiles predicted by Simcyp after simulating a variety of phenomena occurring in the human body, including the drug solubilisation in the luminal fluid. Moreover, plasma-concentration profiles simulated with Simcyp can vary in different virtual subjects or within the same subject in time due to different simulated conditions in the GI tract and the solubility model should be able to make reliable predictions despite this high level of variability.

The results obtained with in vitro data suggest that the proposed GP model can improve considerably the accuracy of the PBPK simulations performed in Simcyp. In fact, the GP model outperforms the conventional solubility model in terms of solubility prediction, giving a

coefficient of determination $R^2$ equal to 0.97 and a root-mean-squared-error RMSE equal to 0.33 mM, thus comparable to the standard deviation of measurement errors that can reach up to 0.10 mM in the fasted state and 0.77 mM in the fed state. Moreover, the GP model provides an estimation of model prediction uncertainty that is not provided by conventional solubility models; the estimated prediction uncertainty contains the majority of the experimental points, therefore the prediction error is statistically negligible. Moreover, the GP model is validated by using approximately half of the solubility dataset in calibration and the remaining half in validation and the satisfactory results are confirmed with $R^2$=0.985 and RMSE=0.208mM in calibration and $R^2$=0.905 and RMSE=0.657mM in validation.

Thanks to these satisfactory results, future work will consist in: (i) further validation of the proposed model with new data at experimental conditions not used in calibration; (ii) final implementation in Simcyp. The plasma-concentration profiles obtained with the PBPK model will be compared with profiles of API A administered to human volunteers in clinical trials for validation purposes. If the satisfactory results obtained in this work with in vitro data are confirmed by the PBPK simulations, the proposed modelling approach will aid the identification of poorly soluble oral dosage forms from the early stages of drug discovery and development, thus reducing waste of time and resources on drugs that will be unsuccessful later on in the development. Moreover, it will allow to improve the representation of inter- and intra-subject variability in different populations, thus ensuring safer and more efficient drugs to patients in a reduced timeline.

# Conclusions and future perspectives

This Dissertation proposed model-based methods that streamline pharmaceutical R&D, while at the same time improving product and process understanding and ensuring product quality and process robustness. Two main approaches were used: model-based design of experiments; data-driven modelling. The methods developed in this work aimed at:

- streamlining the design of tablets lubrication;
- minimising model prediction uncertainty in the whole design space, while ensuring parameters precision and minimum experimental burden;
- achieving autonomous decision-making in continuous-flow microreactor platforms for the minimisation of model prediction variance;
- accurately predicting drug solubility in mixtures of organic solvents for the design and/or optimisation of crystallisation processes;
- predicting the effects of food and population variability on drug solubility in human intestinal fluids, to improve the understanding of drug absorption in the intestine.

**Streamlining the design of tablets lubrication**

In the pharmaceutical industry, the extended Kushner and Moore model proposed by Nassar et al. (2021) is often used to design the best tablets lubrication. In fact, this model allows to predict tablets tensile strength (TS) based on tablets solid fraction (SF, related to compression pressure) and lubrication extent (K, related to blending time). However, it is an algebraic model with 5 correlated parameters in one equation, with one parameter having a small influence on the response variable. Therefore, a considerable number of experiments is required to precisely estimate model parameters, involving the preparation and compression of multiple powder blends, each one with a different lubrication extent. This leads to time-consuming experiments and, especially, to an excessive usage of API, which is the most expensive compound involved and may be scarcely available at early stages of drug development. Therefore, a novel MBDoE method to design highly informative experiments was developed. Being an MBDoE method, the model equation was used to calculate the sensitivities of the response variable (TS) with respect to every model parameter, which in turn was used to calculate the Fisher information matrix. However, the state-of-the-art MBDoE would select a different (SF,K) point at every optimal design, but it is not feasible with the equipment available for lubrication experiments.

In fact, to calibrate the lubrication model, a tablet press is used to compress the same blend at $N_{SF}$ different compression pressures, thus obtaining a set of $N_{SF}$ tablets with the same lubrication extent and different solid fractions (named "profile"). Therefore, the method proposed in this Dissertation consisted in modifying the mathematical structure of the FIM in order to design optimal profiles instead of optimal points: this was achieved by imposing that $N_{SF}$ subsequent rows of the sensitivity matrix used to calculate the FIM had the same lubrication extent, while only SF could change. Moreover, two MBDoE procedures were compared: 1) a sequential one, consisting in designing one optimal profile at a time and having the advantage of updating model parameters, thus updating the estimation of experiments information content, as soon as a new profile was measured; 2) a parallel one, consisting in the design of multiple optimal profiles at once and having practical advantages such as the possibility to prepare different blends in advance and to better organise the experimentation schedule. The proposed MBDoE method was tested with two placebo blends with different lubrication sensitivity and the results showed that 3-4 optimal profiles designed through the proposed MBDoE were sufficient to precisely estimate model parameters instead of the 7-9 profiles typically used in the pharmaceutical industry, thus reducing the experimental burden of 60-70%. The results were similar between sequential and parallel designs, therefore the parallel one can be applied with great practical benefits and without losing valuable information. This conclusion was further supported by the accuracy of model predictions: in fact, the model predicted with 3-4 optimal experiments allowed to predict tablets TS with an absolute error below 0.25 MPa, which is the maximum threshold tolerated in industrial applications.

**Minimising model prediction uncertainty in the whole design space, while ensuring parameters precision and minimum experimental burden**

A novel exploratory MBDoE method was proposed in order to minimise model prediction variance in the whole design space with minimum experimental burden, also ensuring statistically sound parameters estimates. This was done through a mapping of G-optimality: first, a map of information content, H-map, and a map of G-optimality, G-map, were built considering the whole design space; then, points having a G-optimality $J_G$ higher than a user-defined threshold $J_{G,\mathrm{thr}}$were retained as candidate design points; finally, the candidate design point having maximum information content was selected as the experiment to be performed.

This optimisation of the information content in a subset region of the design space determined based on G-optimality allowed to find a trade-off between information maximisation and space

exploration. In turn, this trade-off was handled by means of the threshold $J_{G,\mathrm{thr}}$. The proposed G-map eMBDoE method was tested in silico with two models of increasing complexity: 1) an algebraic model with two control variables and one output; 2) a differential equation model with two constant control variables and two dynamic outputs sampled at three sampling points. In both cases, G-map eMBDoE allowed to increase space exploration with respect to state-of-the-art MBDoE. Moreover, in both cases G-map eMBDoE with $J_{G,\mathrm{thr}}$=0.75 allowed to minimise model prediction variance in the whole design space, as shown by the fastest reduction of the mean and maximum $J_G$ and by the visualisation of G maps. Although space exploration was increased, experiments information content was still adequate for parameters estimation, since G-map eMBDoE with $J_{G,\mathrm{thr}}$=0.75 allowed to pass all $t$-tests with a number of experiments that was equal to or less than the one required by state-of-the-art MBDoE.

Since in silico results showed the advantages of the proposed G-map eMBDoE over conventional information-based methods, like MBDoE, and over exploration-based methods, like LH and factorial DoE, the novel eMBDoE method was validated with experimental data. The platform employed allowed to perform catalytic reactions of total methane oxidation, but the method proposed can be applied to any type of automated platforms, including those used in the (bio)pharmaceutic industry. Accordingly, the G-map eMBDoE method was adapted to the kinetic model representing Mars-van Krevelen mechanism. First, the G-optimality constraint was selected as suggested by preliminary simulations: candidate design points had a G-optimality value below a user-defined threshold $J_{G,\mathrm{thr}}$. Then, three different designs were compared: MBDoE; G-map eMBDoE with $J_{G,\mathrm{thr}}$=0.70; G-map eMBDoE with $J_{G,\mathrm{thr}}$=0.60. The experiments were carried out with the following procedure: (*i*) the G-map eMBDoE method was used to design experimental conditions off-line; (*ii*) the design was sent to the platform, which set the control variables to the designed values in an automated way and automatically measured and recorded the composition of the outlet stream from the reactor; (*iii*) the input-output data measured by the platform were used to design the new experiment and send it to the platform; and so on until the experimental budget was reached. Results showed that the exploration of the design space was enhanced by the G-map eMBDoE with $J_{G,\mathrm{thr}}$=0.60. However, this did not translate into a loss of information: in fact, MBDoE required 23 optimal experiments to have statistically precise model parameters, while G-map eMBDoE with $J_{G,\mathrm{thr}}$=0.60 required only 5 optimal experiments, thus reducing the experimental burden of 78%.

Moreover, G-map eMBDoE with $J_{G,\text{thr}}$=0.60 was able to reduce model prediction variance more efficiently among the three methods considered.

**Implementing autonomous decision-making in continuous-flow microreactors for the minimisation of model prediction variance**

As an improvement of the first versions of the G-map eMBDoE method, a novel method was proposed to automatically select the best G-optimality constraint during the experimentation without needing any prior knowledge on the system. This is especially useful for completely new systems for which plausible initial parameters values may not be available. Furthermore, the lack of human intervention in the determination of the G-optimality constraint allows to implement autonomous decision-making in Industry 4.0 technologies, such as automated chemical platform.

The novel method was named *adaptative G-map eMBDoE* because it automatically adapted the G-optimality constraintonly based on the overlap between H-maps and G-maps. In fact, if (a) the points located in the regions of the H-map with highest information content overlapped with the regions of the G-map with highest model prediction variance, then space exploration was favored by selecting candidate design points having lower G-optimality; instead, if (b) they overlapped with the regions of the G-map with lowest model prediction variance, then space exploration was favored by selecting candidate design points having higher G-optimality. To have a fair comparison, the adaptative G-map eMBDoE method was applied to the two simulated systems used with the first version of G-map eMBDoE, namely the algebraic model with two inputs and one output and the differential equation model with two inputs and two dynamic outputs. Also the same simulation settings were used: preliminary experiments; parameters initial values and upper/lower bounds; measurement errors; ranges for the control variables; E-optimal objective function. Then, the adaptative G-map eMBDoE was compared to the results previously obtained by means of MBDoE, G-map eMBDoE with $J_{G,\text{thr}}$={0.25, 0.75} (with $J_{G,\text{thr}}$=0.75 having a better performance than $J_{G,\text{thr}}$=0.75 in terms of space exploration, parameters precision and minimisation of model prediction variance); LH and factorial DoE. First of all, both inequality types (i.e., $\leq$ and $\geq$) for the G-optimality constraints were selected by the adaptative G-map eMBDoE during the simulated experimental campaigns, proving that there was no contradiction in the use of the two inequality types in the previous works. Moreover, the results showed that the adaptative G-map eMBDoE increased space exploration with respect to information-based methods, namely MBDoE, and it improved

parameters precision with respect to exploration-based methods, like LH and factorial DoE. As regards the comparison between the adaptative G-map eMBDoE and the original G-map eMBDoE, the former one achieved a level of space exploration that is intermediate between the one of eMBDoE with $J_{G,\mathrm{thr}}$=0.25 and the one of eMBDoE with $J_{G,\mathrm{thr}}$=0.75. Moreover, the reduction of model prediction variance of the algebraic model achieved by the adaptative G-map eMBDoE was lower than the one obtained with a fixed threshold of 0.75, but it was considerably more marked than the one achieved with $J_{G,\mathrm{thr}}$=0.25. With the differential equation model, G-map eMBDoE with $J_{G,\mathrm{thr}}$=0.75 and the adaptative method achieves the greatest reduction of model prediction variance among all the methods involved. Finally, with both models, the adaptative G-map eMBDoE required a number of experiments to precisely estimate parameters that was comparable to the one of MBDoE and G-map eMBDoE with $J_{G,\mathrm{thr}}$=0.75. Therefore, these results suggested that the proposed adaptative procedure was able to find a satisfactory trade-off between space exploration and information maximisation without requiring human intervention.

**Accurately predicting drug solubility in mixtures of organic solvents for the design and/or optimisation of crystallisation processes**

A novel data-driven model was proposed to predict drug solubility in a variety of single solvents, binary mixtures and ternary mixtures at different temperatures and compositions. To overcome the issue of selecting the most suitable molecular descriptors among the several possibilities available in the Literature, the UNIFAC theory was used as a reference: the entities identified within the organic mixtures corresponded to UNIFAC subgroups. Overall, a few input variables were needed to make predictions: temperature, mixture composition before API dissolution, UNIFAC subgroups. Correlation can be present among these inputs, but it was automatically handled through the use of a PLS model. The proposed modelling approach was validated experimentally using a real drug substance and 14 organic solvents typically employed in crystallisation units. The calibration dataset was made of the 14 single solvents and a few binary mixtures at two temperatures, 20 and 40°C, while the validation dataset included: (*i*) binary mixtures with the same pairs of solvents used in calibration but different composition and/or temperature; (*ii*) binary mixtures with pairs of solvents not mixed in calibration; (*iii*) ternary mixtures; (*iv*) systems (single solvents, binary mixtures and ternary mixtures) at 50°C, thus a temperature higher than the range used in calibration. The results

showed that the PLS model proposed was able to accurately predict API solubility in all the conditions explored, with 93% of the predictions of validation data having a prediction error statistically equivalent to zero, with a coefficient of determination 0.90. To further test the proposed modelling approach, 9 literature solubility datasets involving organic solvents were used (mainly single solvents and a few binary mixtures, due to the scarcity of published datasets of drug solubility in binary and ternary mixtures of organic solvents). The results confirmed the good prediction accuracy achieved with the PLS model built with the experimental data: the majority of validation data had a coefficient of determination between 0.95 and 0.99; moreover, a percentage of 97-100% of the measured solubility values had a prediction error statistically equivalent to zero.

**Predicting the effects of food and population variability on drug solubility in the human intestine**

A data-driven model was proposed to improve the prediction of intestinal solubility measured in vitro (with biorelevant media mimicking human intestinal fluids) and to facilitate the study of intra- and inter-subject variability. Based on the fact that solubility is highly variable in humans and that in vitro data cover only a restricted set of possible scenarios, a Gaussian Process model was selected, because it inherently provides an estimation of model prediction uncertainty. The Gaussian Process model should be able to represent the effects of all compounds in the biorelevant media, but among them only variations of pH and bile salts can be simulated with Simcyp. Nevertheless, the aim was to integrate the Gaussian Process model into Simcyp, in order to improve the description of drug absorption while still being able to study intra- and inter-subject variability through dynamic simulations of virtual populations. Therefore, two inputs were considered for the Gaussian Process model: (1) pH and (2) an apparent bile salts concentration, given by the sum of the concentrations of the 4 bile salts, lecithin, oleic acid and cholesterol. This model was able to improve considerably the predictions of all in-vitro data: the coefficient of determination was equal to 0.97 (thus, close to 1) and the root mean squared error RMSE equal to 0.33 mM, thus comparable to the standard deviation of measurement errors that could reach up to 0.10 mM in the fasted state and 0.77 mM in the fed state. Therefore, this Gaussian Process model is expected to improve the predictions of PBPK phenomena in Simcyp. Moreover, the improved representation of intestinal solubility is expected to give a valuable support to drug development: even though the intestinal absorption of a drug is the result of a complex interplay of several phenomena, solubility directly influences

dissolution rate and precipitation and, if too low, it can limit absorption itself. In turn, a proper intestinal absorption is required for the drug to reach the bloodstream and, consequently, to have the desired effect on the patients.

**Future perspcetives** to extend the usefulness of the methods proposed in this Dissertation are:

- Based on the results of MBDoE applied to the lubrication unit (Chapter 3), MBDoE can be applied to different units of pharmaceutical manufacturing processes and adapted based on the specific features of the equipment involved. This will improve robustness of the processes developed, while simultaneously reducing time, labour, and costs for process development.
- The satisfactory results obtained with exploratory MBDoE, both in silico and experimentally, support the conclusion that the adaptative G-map eMBDoE is ready to be implemented in a fully autonomous platform. Ongoing work is focusing on its application to the kinetic model of Chapter 5 for total methane oxidation and on its integration within LabView. This will allow to experimentally confirm the advantages of the adaptative procedure in terms of space exploration, parameters precision and minimisation of model prediction variance without the need for human intervention. Besides this specific application, the adaptative G-map eMBDoE procedure can be applied to any type of automated (bio)chemical platform employed in the pharmaceutical industry, thus exploiting the full potential of Industry 4.0 technologies;
- The PLS model to predict drug solubility in organic mixtures (Chapter 3) has shown its good performance both in calibration and validation with a real drug substance and with 9 literature datasets on drug and drug-like molecules. Therefore, future work may focus on the systematic application of this modelling approach to drugs of different physico-chemical properties, in order to identify possible deviations of specific systems from the expected behaviour and to identify possible causes of such deviations;
- The accurate predictions of biorelevant solubility achieved by the GP model (Chapter 8) suggest that this solubility model is ready for the integration into the available commercial software for PBPK simulations. This will allow to predict plasma-concentration profiles in all the subjects of a virtual population, which in turn can be compared to the profiles experimentally measured in human volunteers. The GP model improves the prediction of the in vitro solubility data, therefore it is expected to improve the overall PBPK simulation. A possible limitation may be the fact that the apparent BS concentration reaches values up

to 43 mM, which is not unfeasible since BS measured in human volunteers can reach up to 100 mM, but it is still higher than the average values typically simulated in the intestine, which range between 0 and 15 mM. In case this degrades the predictions of plasma-concentration profile, a different version of the GP model that uses only pH and BS as inputs can be considered. The GP model leading to the most accurate prediction of the plasma concentration profiles, besides patients' variability, will be implemented in the PBPK software for future development of drugs alongside clinical trials.

# Appendix A
# Lubrication model

The following strategy is proposed to solve $\mathbf{S}(\widehat{\boldsymbol{\theta}}, \boldsymbol{\varphi})$ ill-conditioning:

1. the sensitivity matrix ($\mathbf{S}(\widehat{\boldsymbol{\theta}}, \boldsymbol{\varphi})$) is evaluated with currently available experiments and its condition number ($\kappa_{\text{cond}}$) is calculated;
2. if $\kappa_{\text{cond}} \geq \kappa_{\text{cond,max}}$, $\mathbf{S}(\widehat{\boldsymbol{\theta}}, \boldsymbol{\varphi})$ is deemed ill-conditioned and a regularization technique must be applied, as suggested by Grah (2004), we set $\kappa_{\text{cond,max}} = 1000$;
3. data points of the form $[SF_{\text{reg}}, K_{\text{reg}}]_l^T$, $l = 1, ..., N_{\text{reg}}$, are selected in their domain and they are used to calculate additional rows $\mathbf{s}_{\text{reg},l}(\widehat{\boldsymbol{\theta}})$, $l = 1, \dots, N_{\text{reg}}$:

$$\mathbf{S}(\widehat{\boldsymbol{\theta}}, \boldsymbol{\varphi}) = \begin{pmatrix} \mathbf{s}_{11}(\widehat{\boldsymbol{\theta}}, \boldsymbol{\varphi}) \\ \vdots \\ \mathbf{s}_{N_{SF}N_K}(\widehat{\boldsymbol{\theta}}, \boldsymbol{\varphi}) \\ \hline \mathbf{s}_{\text{reg},1}(\widehat{\boldsymbol{\theta}}, \boldsymbol{\varphi}) \\ \vdots \\ \mathbf{s}_{\text{reg},N_{\text{reg}}}(\widehat{\boldsymbol{\theta}}, \boldsymbol{\varphi}) \end{pmatrix} \tag{A.1}$$

4. the condition number of the regularized $\mathbf{S}(\widehat{\boldsymbol{\theta}}, \boldsymbol{\varphi})$ is evaluated in order to verify that $\kappa_{\text{cond}} < \kappa_{\text{cond,max}}$; if this is not the case, repeat points 3., 4., otherwise stop.

The resulting well-conditioned sensitivity matrix is used to calculate the FIM and to solve the optimisation problem (Eq. 2.5, section 2). However, the $N_{\text{reg}}$ random experiments are never measured experimentally and they are removed from the sensitivity matrix as soon as the optimal experiment ($\boldsymbol{\varphi}_{\text{opt}}$) is designed; for this reason, they are named *ghost* data.

We empirically observed that this strategy is able to efficiently solve ill-conditioning with a limited number of ghost data.

# Appendix B
# G-map eMBDoE: selection of the threshold and results reproducibility

Appendix B shows additional results for case study 1 (algebraic model of Chapter 4) including the rationale for the selection of G-optimality threshold (Section B.1) and parameters accuracy (difference between estimated and true parameters values) and precision (confidence intervals) evaluated at each iteration of the sequential procedure (Section B.2); reproducibility of the LH results besides random variations in the LH designs when this method is simulated in different runs (Section B.3)

## B.1 Selection of G-optimality threshold

Graphical visualisation of the designed experiments shows that space exploration increases with a threshold $J_{G,\text{thr}}$ between 0.50 and 0.85 with respect to classical MBDoE (see main text, Section 4.3.1). Moreover, Figure B.1 shows that eMBDoE with $J_{G,\text{thr}}$=0.85 reduced the maximum G-optimality calculated across the entire design space more than all the other methods.

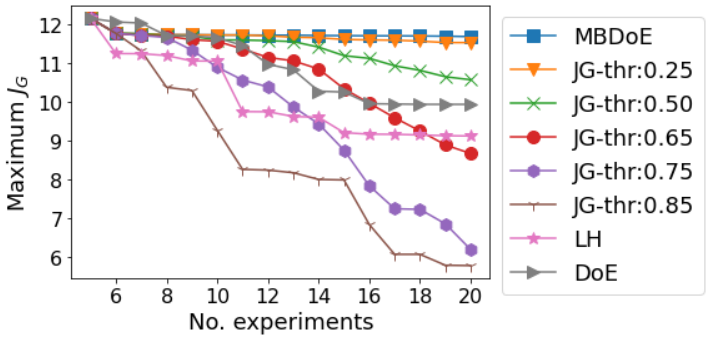



**Figure B.1** *Maximum G-optimality calculated across the whole design space. Different curves represent different methods: classical MBDoE (i.e., eMBDoE with $J_{G,thr}$=0); G-map eMBDoE with $J_{G,thr}$ ∈{0.25, 0.50, 0.65,0.75,0.85}; Latin Hypercube (LH) and factorial DoE.*

However, Figure B.2 shows that eMBDoE with $J_{G,\text{thr}}$=0.85 increases the number of calibration experiments to precisely estimate parameter $\hat{\theta}_5$ with respect to all the other eMBDoE scenarios.

The remaining parameters are omitted for sake of conciseness, since $\hat{\theta}_5$ is the one requiring more calibration data to be estimated.

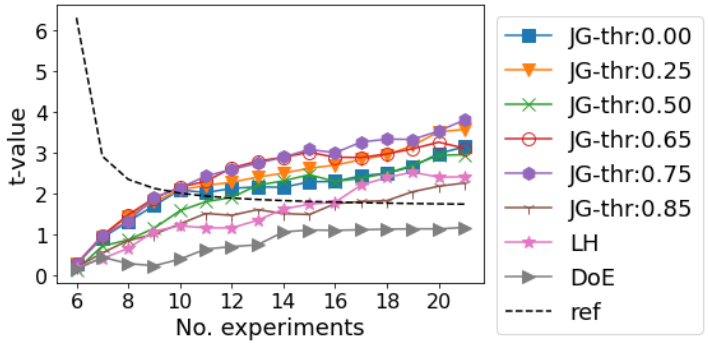



**Figure B.2** *Parameters precision of parameter 5 in terms of t-tests for all the scenarios considered: classical MBDoE (i.e., eMBDoE with $J_{G,thr}$=0); $J_{G,thr}$ ∈{0.25, 0.50, 0.65,0.75,0.85}; Latin Hypercube (LH) and factorial DoE.*

By analysing the results from Figure B.1 and B.2, a threshold $J_{G,\text{thr}}$=0.75 is selected: it reduces significantly model prediction variance across the whole design space (Figure B.1) without increasing the number of experiments to precisely estimate model parameters (Figure B.2).

## B.2 Parameters accuracy and precision

Parameters accuracy is evaluated by comparing the true parameter values, which are the values used to generate data in silico, to the estimated parameter values at every iteration.

Figures B.3-B.7 show that a good accuracy is achieved in all scenarios as the point estimate is close to the true parameter value, which in turn lies in the range delimited by the 95% C.I..

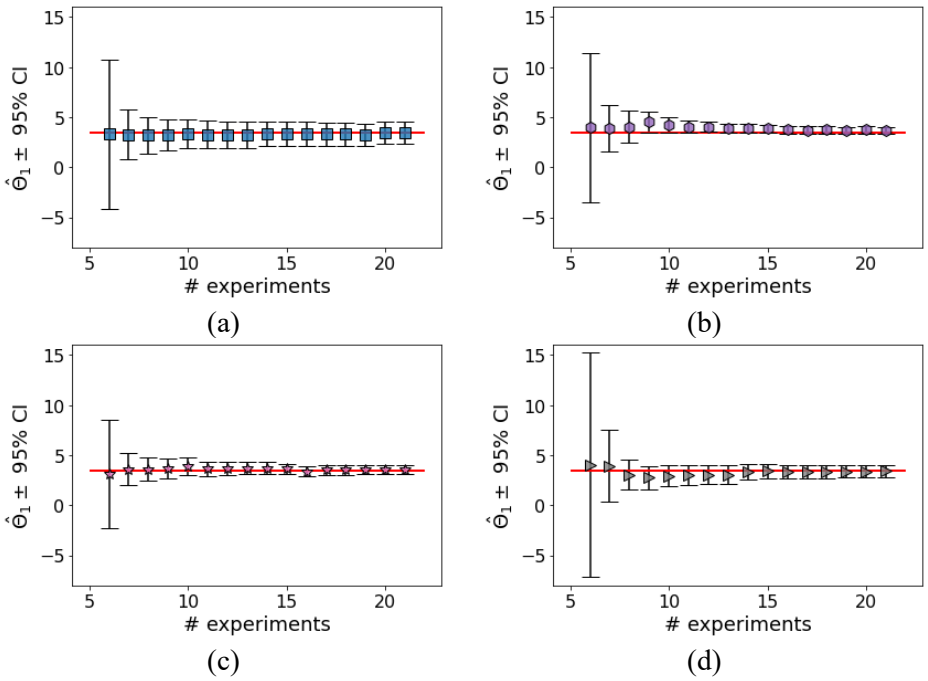



**Figure B.3** *Analysis of the accuracy of parameter 1: the true parameter value (red line) is compared against the point estimate together with their 95% CI (black vertical lines). Four methods are compared: (a) MBDoE (blue squares); (b) G-map eMBDoE with $J_{G,thr}$=0.75 (lilac hexagons); (c) LH (pink stars); (d) $4^2$-level full factorial DoE (grey triangles).*

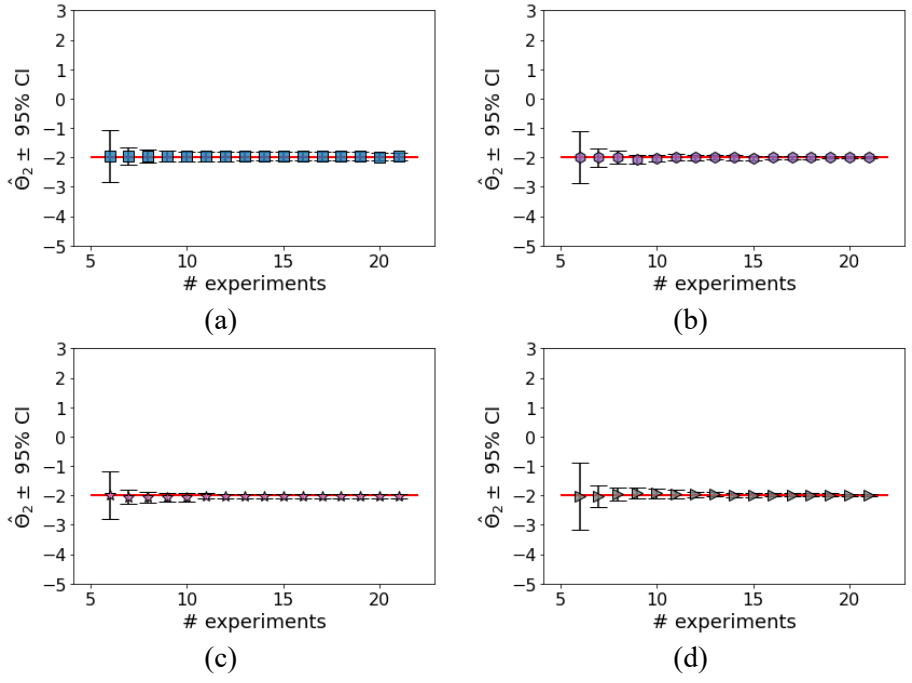



**Figure B.4** *Analysis of the accuracy of parameter 2: the true parameter value (red line) is compared against the point estimate together with their 95% confidence intervals (95% CI; black vertical lines). Four methods are compared: (a) MBDoE (blue squares); (b) G-map eMBDoE with $J_{G,thr}$=0.75 (lilac hexagons); (c) LH (pink stars); (d) $4^2$-level full factorial DoE (grey triangles).*

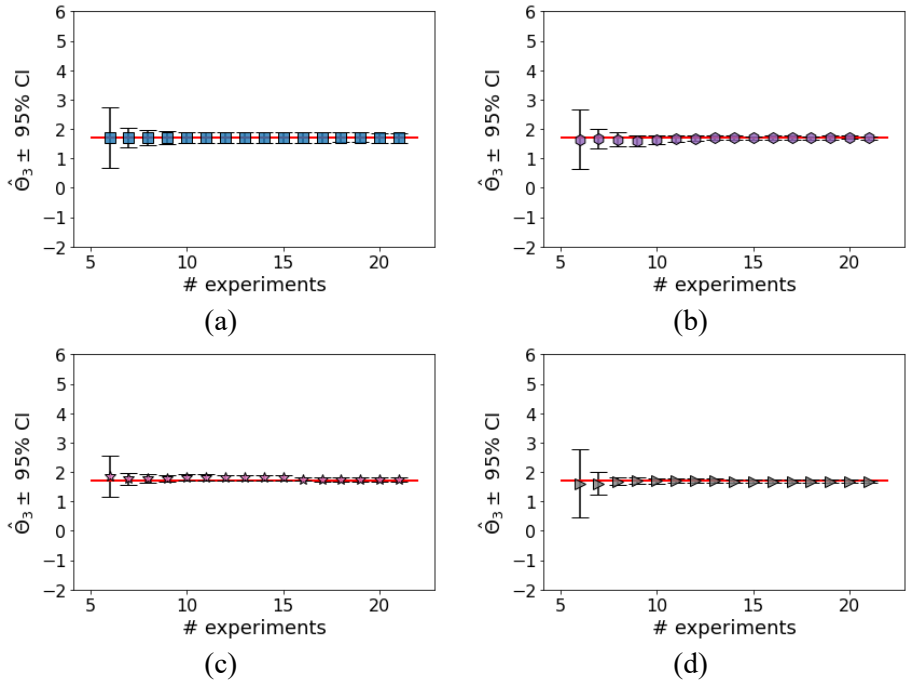



**Figure B.5** *Analysis of the accuracy of parameter 3: the true parameter value (red line) is compared against the point estimate together with their 95% confidence intervals (95% CI; black vertical lines). Four methods are compared: (a) MBDoE (blue squares); (b) G-map eMBDoE with $J_{G,thr}$=0.75 (lilac hexagons); (c) LH (pink stars); (d) $4^2$-level full factorial DoE (grey triangles).*

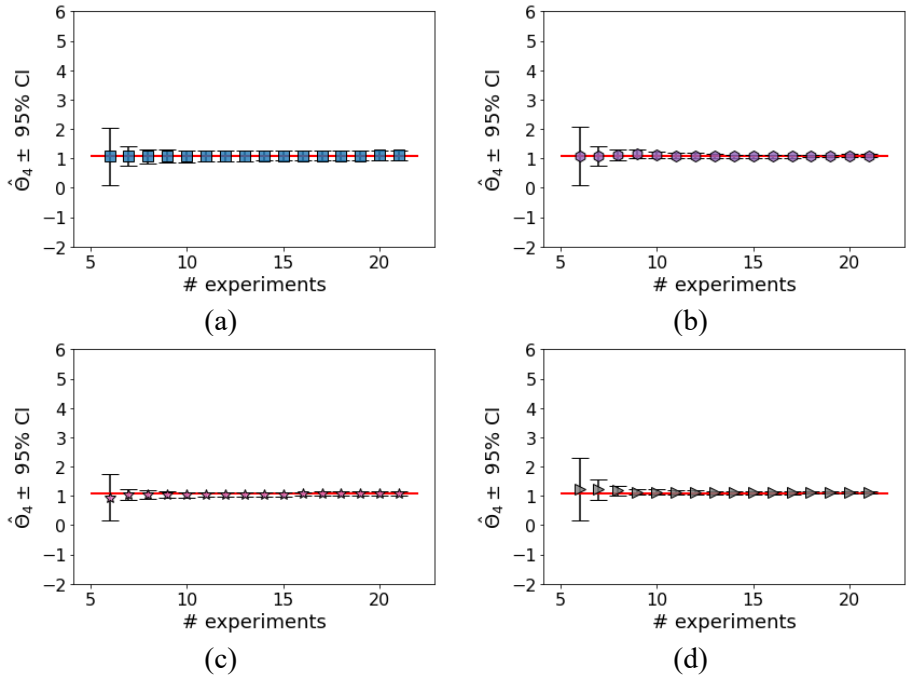



**Figure B.6** *Analysis of the accuracy of parameter 4: the true parameter value (red line) is compared against the point estimate together with their 95% confidence intervals (95% CI; black vertical lines). Four methods are compared: (a) MBDoE (blue squares); (b) G-map eMBDoE with $J_{G,thr}$=0.75 (lilac hexagons); (c) LH (pink stars); (d) $4^2$-level full factorial DoE (grey triangles).*

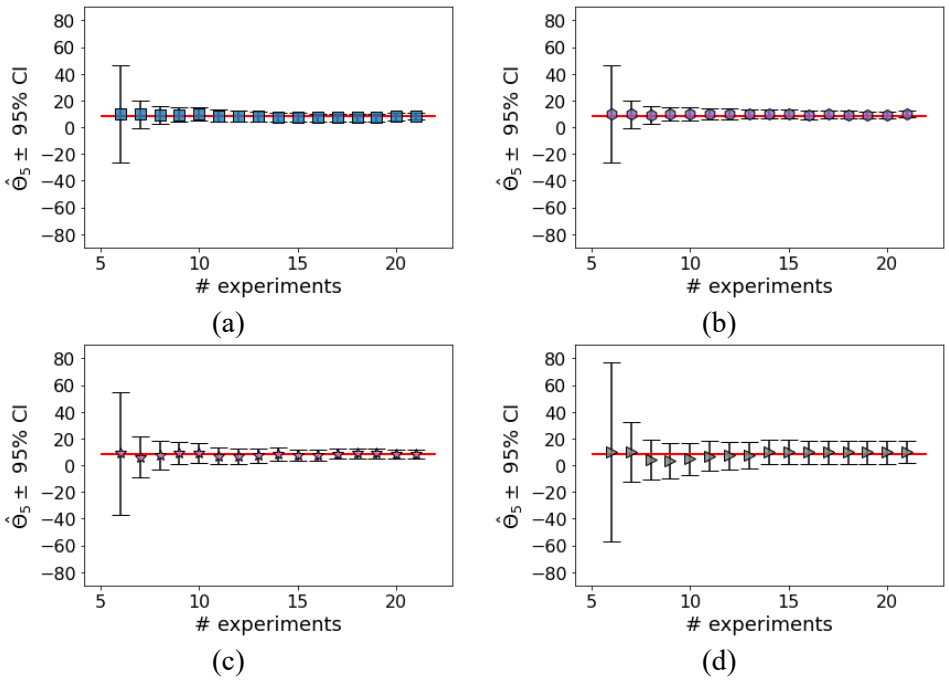



**Figure B.7** *Analysis of the accuracy of parameter 5: the true parameter value (red line) is compared against the point estimate together with their 95% confidence intervals (95% CI; black vertical lines). Four methods are compared: (a) MBDoE (blue squares); (b) G-map eMBDoE with $J_{G,thr}$=0.75 (lilac hexagons); (c) LH (pink stars); (d) $4^2$-level full factorial DoE (grey triangles).*

## B.3 Reproducibility of the LH results

For a fixed design space, different runs of the LH method can lead to randomly different designs. However, the conclusions drawn in the main text are general and not strictly related to a particular realization of the LH design. As a proof, Figure A.8 shows the results in terms of parameters precision obtained with 5 different LH runs.

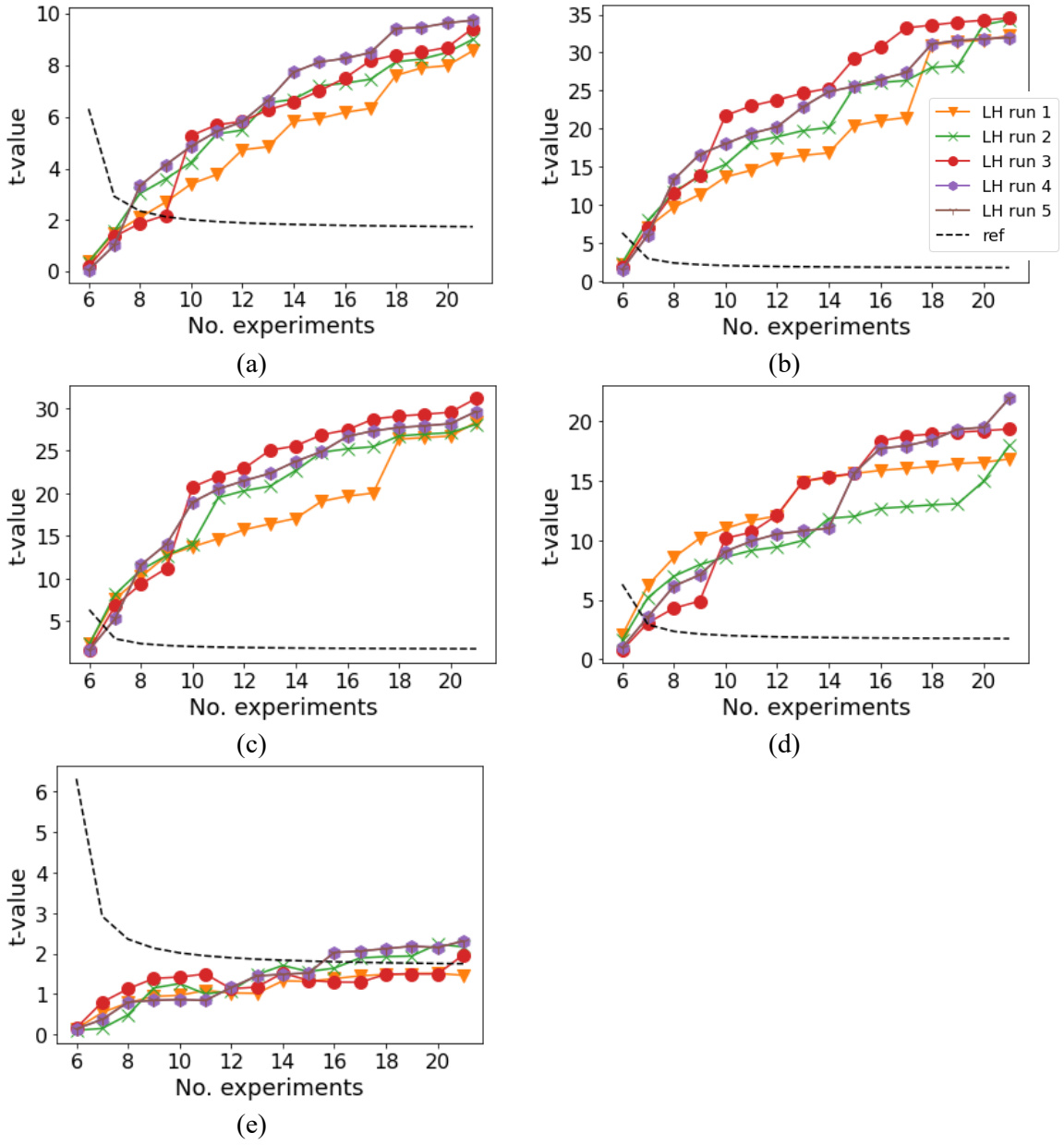



**Figure A.8** *Parameters precision tests performed with 5 different LH designs. Figures (a)-(e) show results of parameters 1-5, respectively.*

Figure A.8 shows that the LH performance in terms of number of experiments needed to pass the t-tests is not strictly dependent on the singular realisation of the LH design. Moreover, the number of experiments needed to estimate all parameters is always higher to the one required by G-map eMBDoE and MBDoE (which need 10 experiments, as shown in Section 4.3.1 of the main text), therefore the conclusions on the comparison among the explorative and/or optimal methods remains unchanged.

# Appendix C
# G-map eMBDoE: additional results with Model 2

Additional results for the Model 2 (baker's yeast fermentation model) of Chapter 4 are presented: Section C.1 illustrates the rationale for selection of the G-optimality threshold; Section C.2 shows the results in terms of parameter accuracy (i.e., difference between estimated and true values) and precision (through confidence intervals) at each iteration of the sequential procedure; Section C.3 shows the reproducibility of the LH results when the method is simulated several times; Section C.4 shows the contributions to $J_G$ given by both responses $x_1$ and $x_2$ at all time points $t_1$, $t_2$ and $t_3$; section C.5 shows G-maps and H-maps at the last iteration of experiments design.

## C.1 Selection of the G-optimality threshold

The maximum G-optimality calculated in the whole design space is shown in Figure C.1. The figure suggests that completely explorative methods such as DoE and LH show higher values of $J_{G,\max}$ in the whole experimental campaign, followed by MBDoE, i.e., eMBDoE with $J_{G,\text{thr}}$=0. Instead, $J_{G,\text{thr}}$=0.85 has the best performance in terms of model prediction variance reduction, followed by $J_{G,\text{thr}}$=0.75, which has the second-best performance.

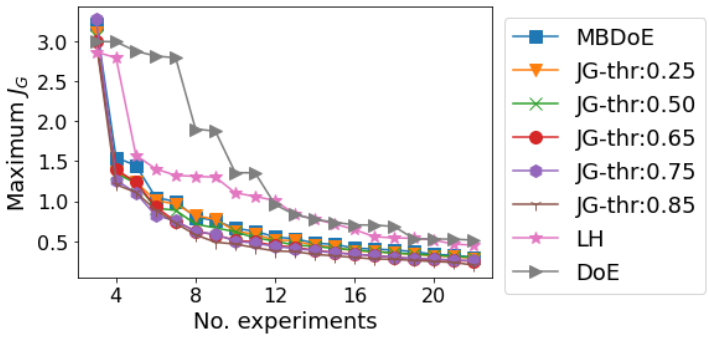



**Figure B.1** *Maximum G-optimality in the whole design space. Different methods: classical MBDoE (i.e., eMBDoE with $J_{G,thr}$=0); G-map eMBDoE with $J_{G,thr}$ ∈{0.25, 0.50, 0.65,0.75,0.85}; Latin Hypercube (LH) and factorial DoE.*

However, eMBDoE with $J_{G,\text{thr}}$=0.85 requires a higher number of calibration experiments to precisely estimate the model parameters. Figure C.2 shows that the two eMBDoE methods that

allow to estimate parameter $\hat{\theta}_4$ (the most critical model parameters in terms of experiments required for estimation) with the minimum number of experiments have $J_{G,\text{thr}} \in \{0.50, 0.75\}$.

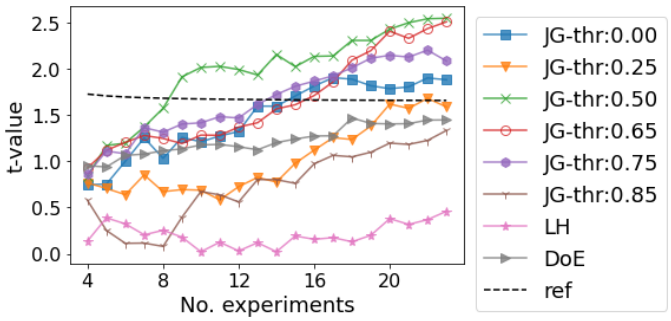



**Figure C.2** *Parameters precision in terms of t-tests for all the scenarios considered: classical MBDoE (i.e., eMBDoE with $J_{G,thr}$=0); $J_{G,thr}$ ∈{0.25, 0.50, 0.65,0.75,0.85}; Latin Hypercube (LH) and factorial DoE.*

Therefore, to find a trade-off between reduction of model prediction variance and maximisation of parameters precision, a threshold of 0.75 is selected for this case study.

## C.2 Parameters accuracy and precision

Besides the differences in terms of precision (as shown in the $t$-tests), a satisfactory accuracy is achieved in all scenarios and all the 4 parameters do not depart considerably from the assumed true values for the 4 methods (Figure C.3-6).

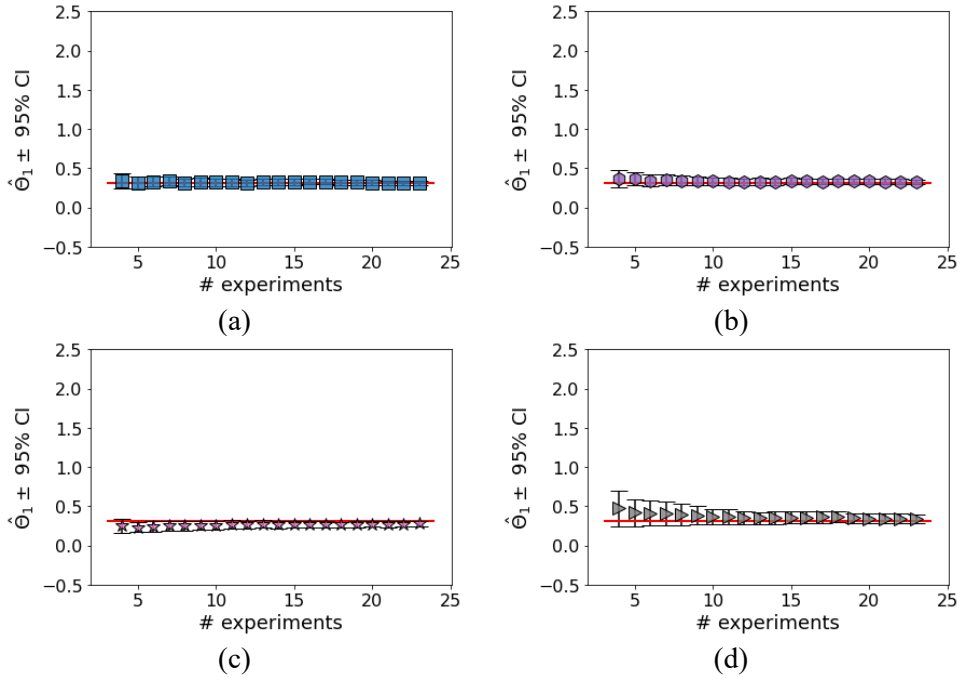



**Figure C.3** *Analysis of the accuracy of parameter 1: true value (red line) compared against the point estimate with 95% CI (black vertical lines). Methods are compared: (a) MBDoE (blue squares); (b) G-map eMBDoE, $J_{G,thr}$=0.75 (lilac hexagons); (c) LH (pink stars); (d) DoE (grey triangles).*

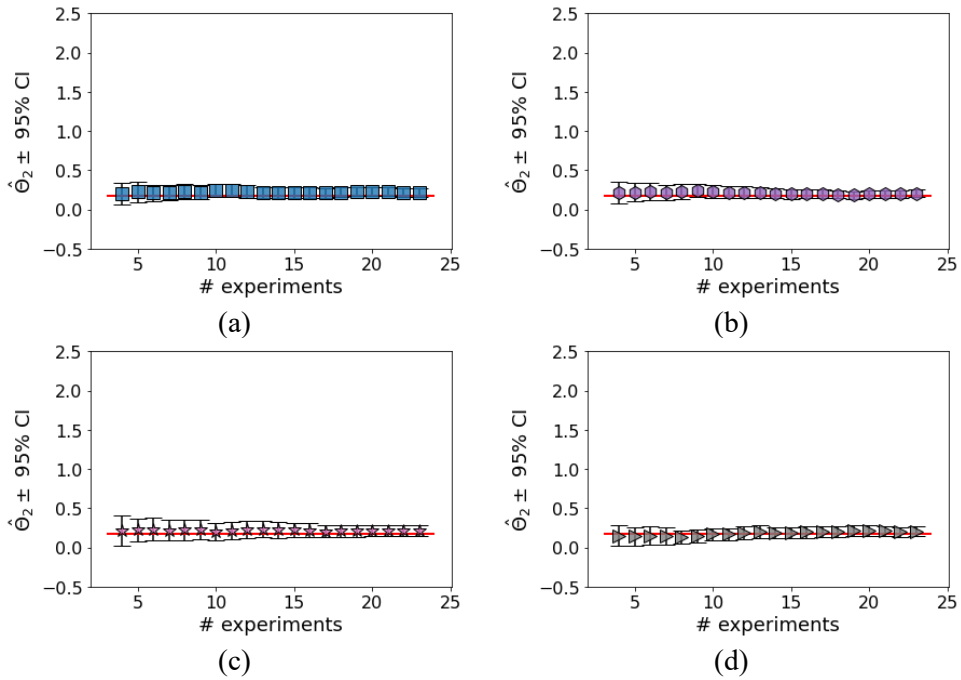



**Figure C.4** *Analysis of the accuracy of parameter 2: the true parameter value (red line) is compared against the point estimate (squares, hexagons, stars, triangles for MBDoE, eMBDoE with $J_{G,thr}$=0.75, LH and $4^2$-level full-factorial DoE, respectively) together with their 95% confidence intervals (95% CI; black vertical lines).*

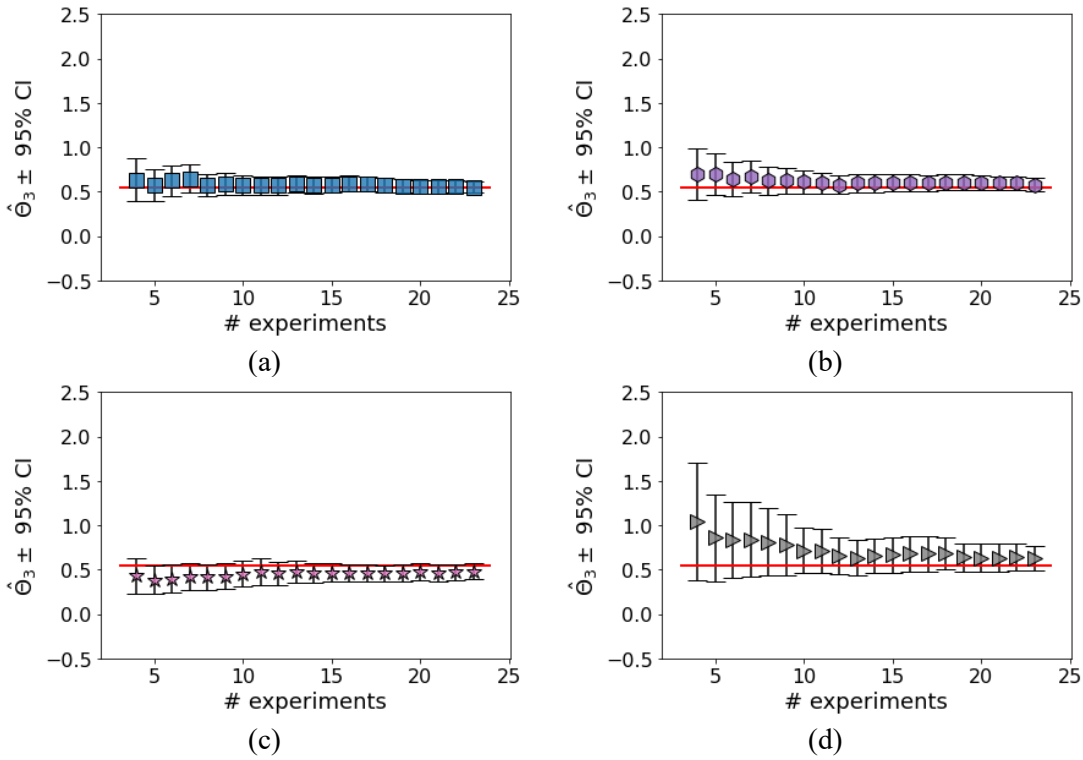



**Figure C.5** *Analysis of the accuracy of parameter 3: the true parameter value (red line) is compared against the point estimate together with their 95% confidence intervals (95% CI; black vertical lines). Four methods are compared: (a) MBDoE (blue squares); (b) G-map eMBDoE with $J_{G,thr}$=0.75 (lilac hexagons); (c) LH (pink stars); (d) $4^2$-level full factorial DoE (grey triangles).*

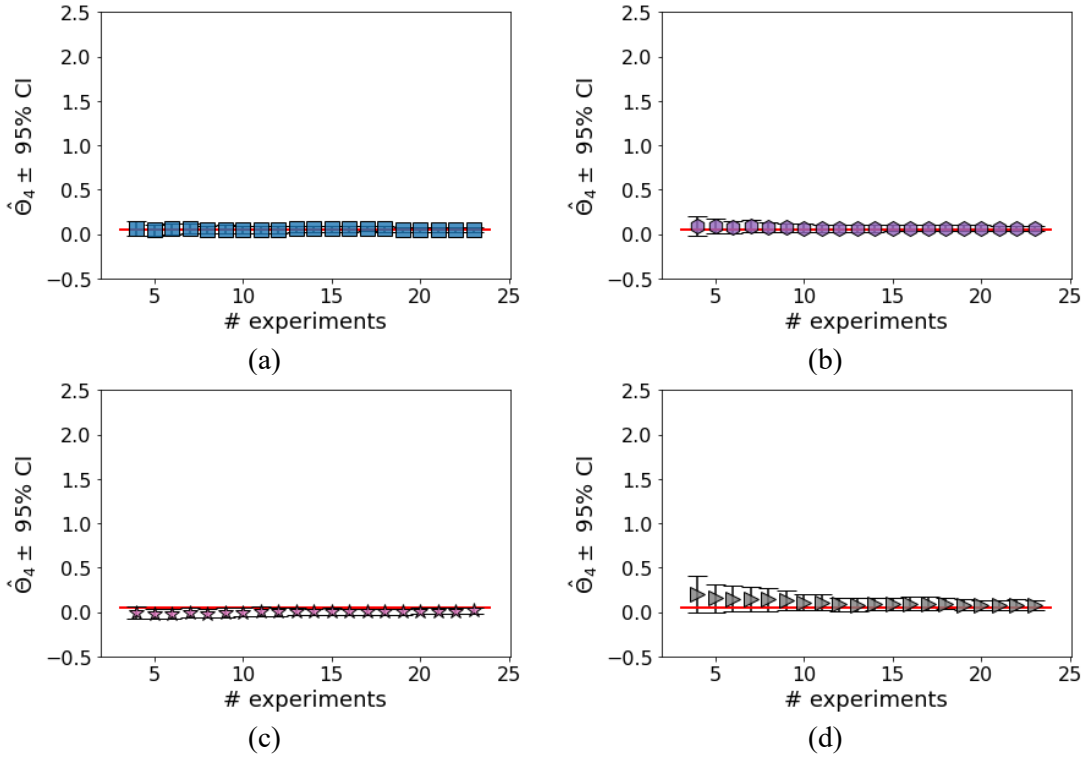



**Figure C.6** *Analysis of the accuracy of parameter 4: the true value (red line) is compared against the point estimate with itd 95% CI (black vertical lines). Methods compared: (a) MBDoE (blue squares); (b) G-map eMBDoE, $J_{G,thr}$=0.75 (lilac hexagons); (c) LH (pink stars); (d) $4^2$-level full factorial DoE (grey triangles).*

## C.3 Reproducibility of the LH results

Different runs of LH method lead to random variations in the corresponding LH designs. Therefore, 5 different runs are analysed in term of parameters precision tests (Figure C.7) in order to assess reproducibility of the results shown in the main text (Section 4.3.2).

Figure C.7 shows that the number of experiments required to estimate a given model parameter does not change considerably with random variations in the LH design: indeed, a maximum difference of two experiments is found with parameters 2 and 3, while parameter 4 is never statistically sound for all considered scenarios. These results are analogous to the ones shown in the main text, thus proving reproducibility of the LH results

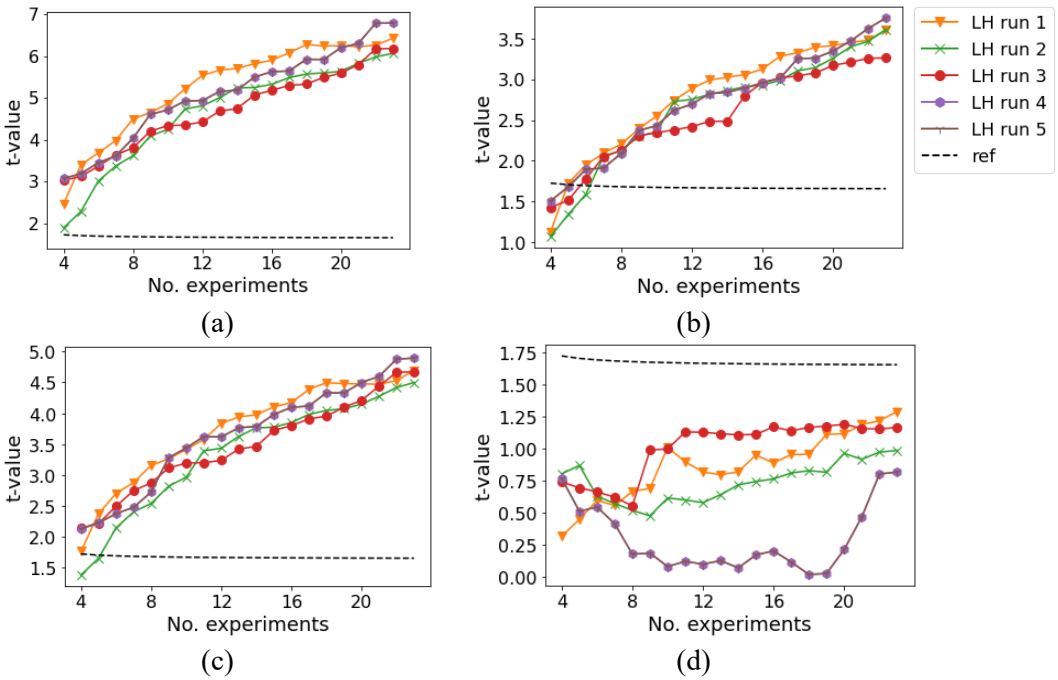



**Figure C.7** *Parameters precision tests performed with 5 different LH designs. Figures (a)-(d) show results of parameters 1-4, respectively.*

## C.4 Single contributions to the scalar measures of G-optimality

In the G-map eMBDoE, each point of the grid covering the design space is characterised in terms of both G-optimality and information content (i.e., a FIM-based metric). However, dynamic models such as the one of Model 2 presents a higher level of complexity: each point of the grid corresponds not only to a scalar value, but to an entire dynamic profile for each response variable. In this case, 2 response variables ($x_1$ and $x_2$) are measured at 3 different sampling points ($t_1$, $t_2$ and $t_3$). The model prediction variance $\mathbf{V}_y$ can be calculated as in Chapter 2 for $x_1$ at $t_1$, $t_2$ and $t_3$ and for $x_2$ at $t_1$, $t_2$ and $t_3$; this holds for every point in the grid. To summarise this information, mean and maximum values of the $\mathbf{V}_y$ estimated for every point in the design space are calculated for the 6 time contributions of model prediction variances; here, only maximum values are shown in Figure C.5 for sake of conciseness, but similar results hold for the mean values.

Considering both responses $x_1$ and $x_2$, the maximum model prediction variances increase from time point $t_1$ (Figure C.8a,b) to $t_2$ (Figure C.8c,d) to $t_3$ (Figure C.8e,f). Moreover, at the same time point, $x_2$ has a higher model prediction variance with respect to $x_1$ (Figure C.8b,d,f compared to Figure C.8a,c,e, respectively). The advantage of G-map eMBDoE in terms of model prediction variance reduction is most evident when the highest G-optimality values are found, namely with $x_2$ at $t_3$ and with $x_1$ at $t_3$.

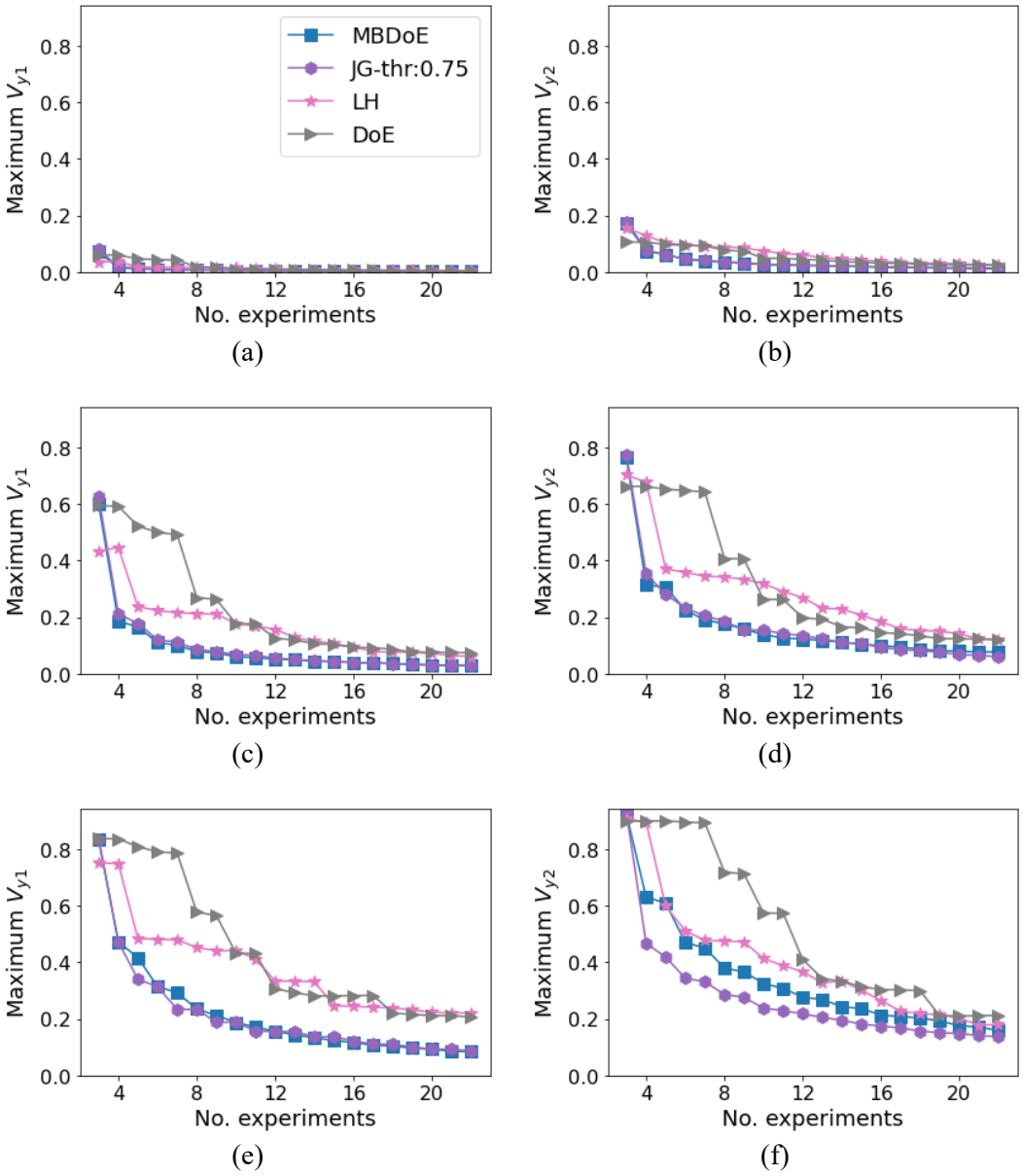



**Figure C.8** *Maximum values of the single contributions to $J_G$ at every iteration; comparison among MBDoE, G-map eMBDoE, LH and $4^2$-level full factorial DoE. All sampling points, are considered for both responses, namely $x_1$ and $x_2$: (a)-(b) $t_1$; (c)-(d) $t_2$; (e)-(f) $t_3$.*

## C.5 G-maps and H-maps at the last experiment design iteration

When 22 experiments are used in calibration and the 23rd experiment must be designed, the G-maps in Figure C.9 are obtained: completely explorative methods like LH and factorial DoE have still higher values of G-optimality at $u_2$ between 30 and 35 g/L; instead, MBDoE and eMBDoE have reduced considerably G-optimality values in the whole design space.

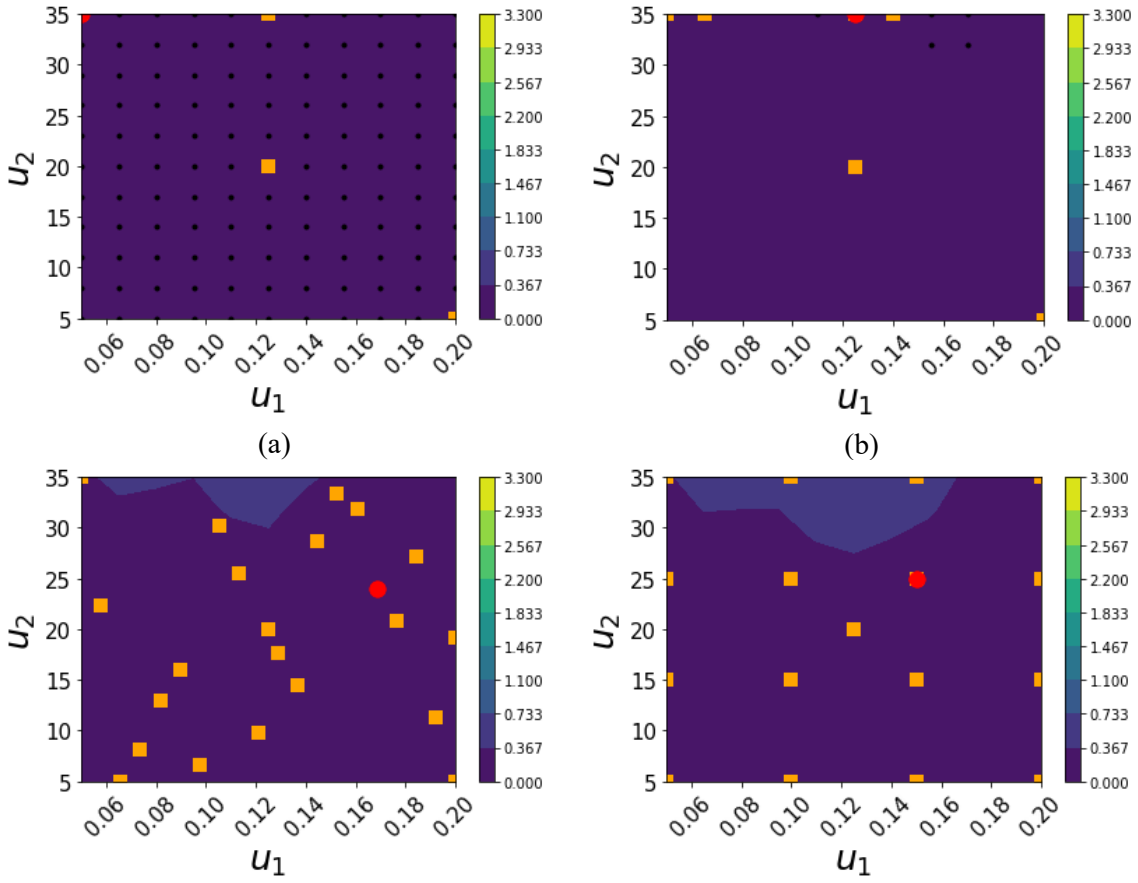



**Figure C.9** *G-maps generated after 22 calibration experiments. Four methods are compared: (a) MBDoE; (b) G-map eMBDoE with $J_{G,thr}$=0.75; (c) LH; (d) $4^2$-level full factorial DoE. Orange squares indicates already measured data (namely, data used to calibrate the model); black dots indicate candidate design points; the red point indicates the experiment designed at the current iteration.*

Finally, the corresponding maps of information content are shown in Figure C.10: as expected, completely explorative methods like LH and factorial DoE lead to the smallest information content, while standard MBDoE ensures the highest information content. Explorative MBDoE with $J_{G,thr}$=0.75 provides the second best performance in terms of information maximization.

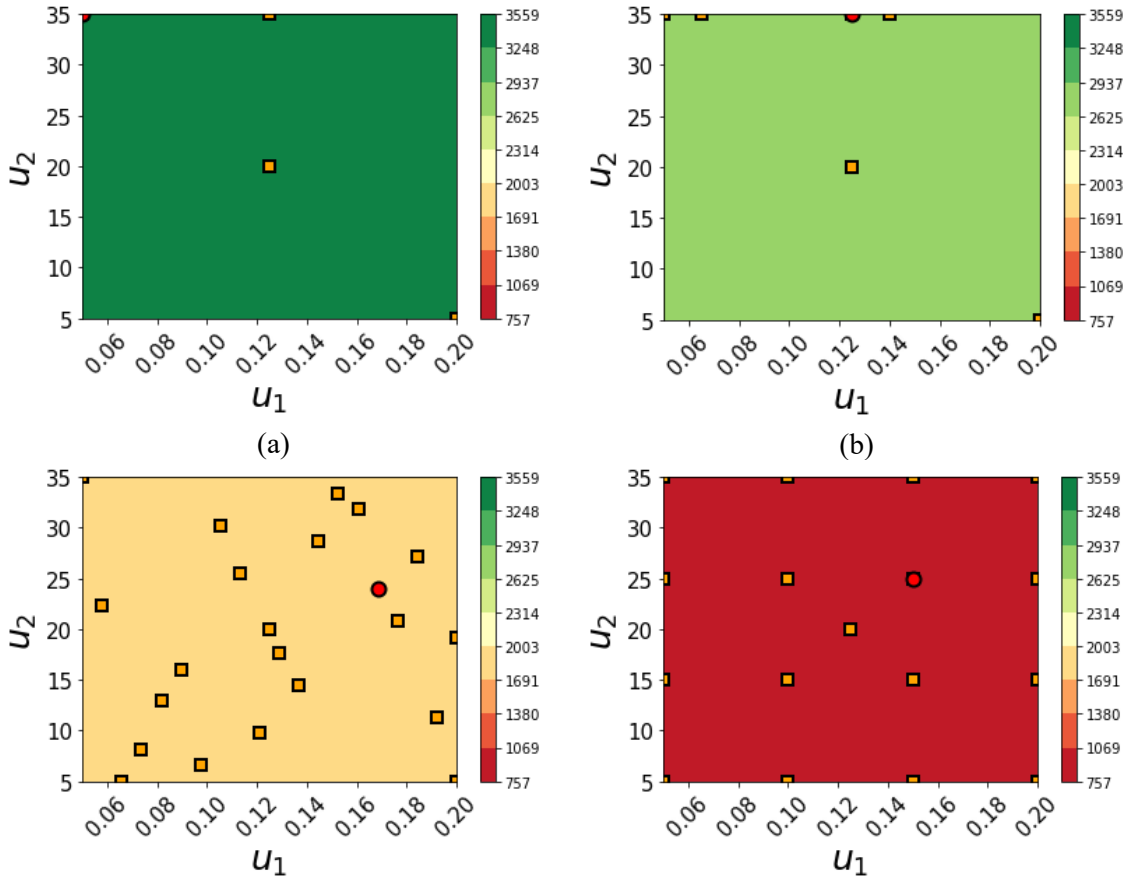



**Figure C.10** *H-maps generated after 22 calibration experiments. Four methods are compared: (a) MBDoE; (b) G-map eMBDoE with $J_{G,thr}$=0.75; (c) LH; (d) $4^2$-level full factorial DoE. Orange squares indicates already measured data (namely, data used to calibrate the model); the red point indicates the experiment designed at the current iteration.*

# Appendix D
# G-map eMBDoE: supplementary material

In this Appendix, two aspects regarding the robustness of the results are considered: in Section D.1, the effect of different sampling points for the dynamic system (Model 2 in Subsection 4.3.2 of the main text); in Section D.2, the effect of different realisations of measurement noise for both Model 1 and 2 (Subsections 4.3.1 and 4.3.2 of the main text, respectively).

## D.1 Effect of sampling points selection

In Model 2 of Chapter 4, 3 equally spaced sampling points $N_{\mathrm{sp}}$are selected within the experiments duration of 21h. In this Appendix, we show that the results obtained in the main text are general and different choices of sampling points do not change significantly the conclusions on the comparison among model-based design of experiments (MBDoE), explorative model-based design of experiments (eMBDoE), Latin Hypercube (LH), factorial design of experiments (DoE).

To show this, different number of sampling points, equally distributed in the fixed experiments duration of 21h, are considered:

- 4 sampling points in Section D.1.1;
- 20 sampling points in section D.1.2.

The remaining settings of the simulations are the same as in the main text of the paper. Two keys analyses are performed for each scenario:

- t-tests for parameters precision;
- profiles of scalar indices of G-optimality.

### *D.1.1. Small increase of sampling points: $N_{\mathrm{sp}}$=4*

Figure D.1 shows that purely explorative designs such as LH and DoE are not able to estimate parameter 4 within the experimental budget. Both MBDoE and eMBDoE are able to estimate all model parameters, but the latter reduces the number of experiments required (Figure D.1.d).

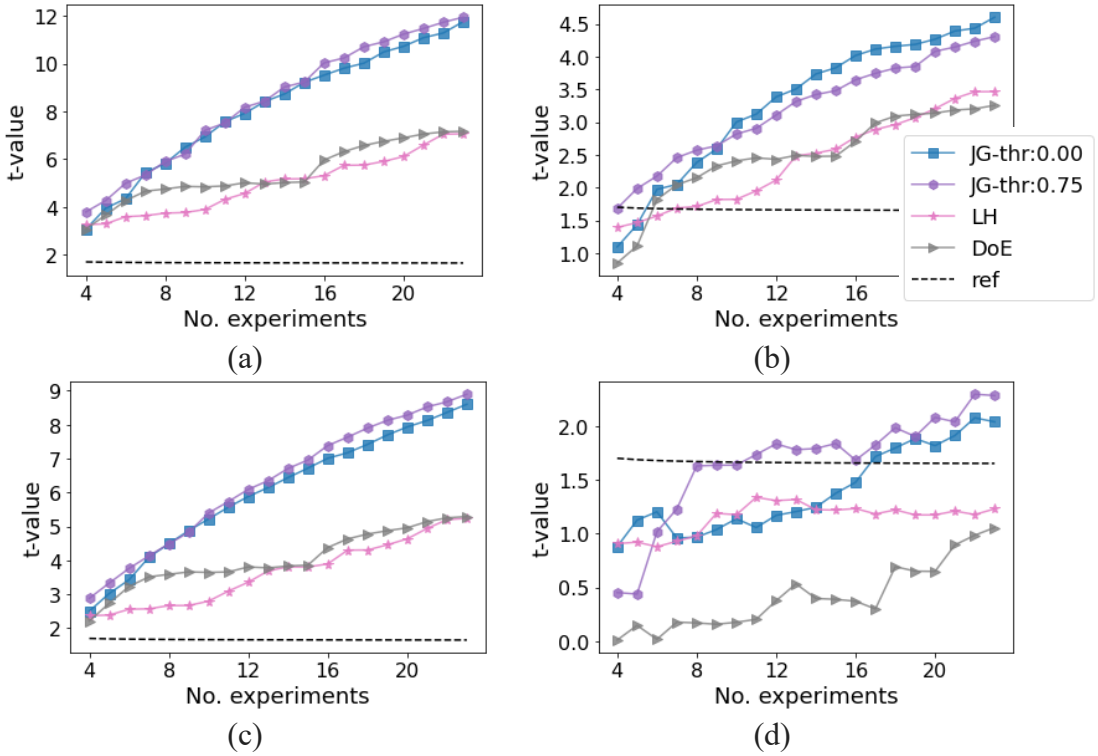



**Figure D.1** *Parameters precision through t-tests: parameters 1-4 in figures a-d, respectively. Results obtained with 4 sampling points of the response variable.*

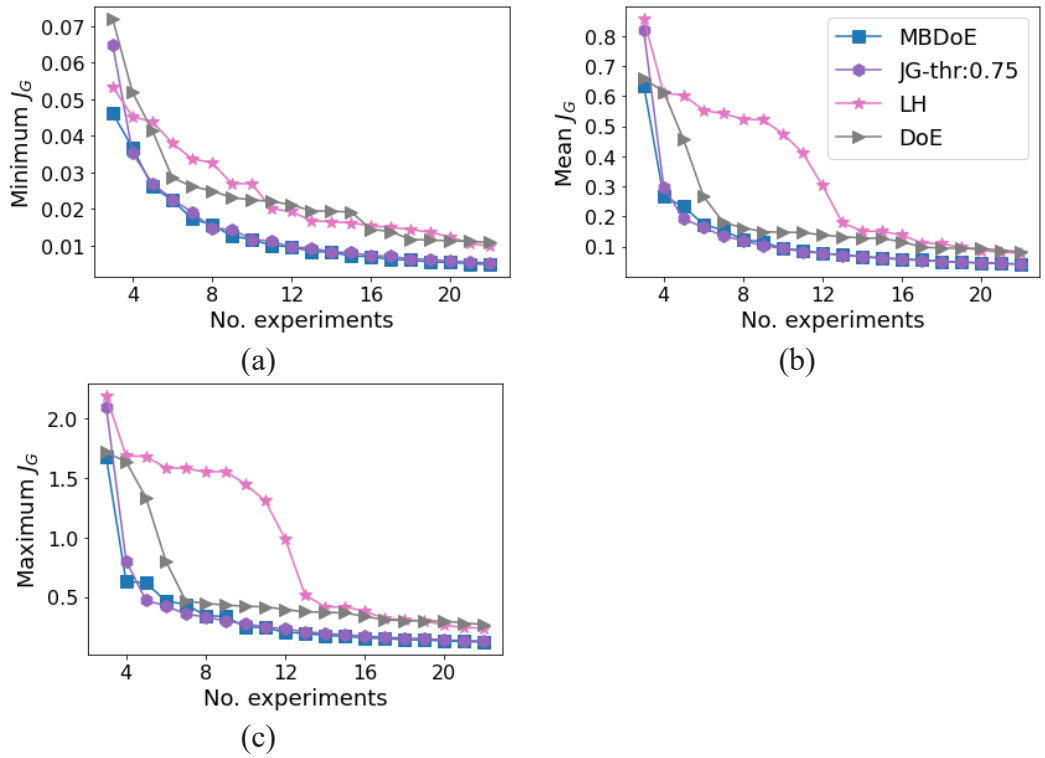



**Figure D.2** *Scalar indices of G-optimality: (a) minimum, (b) mean and (c) maximum G-optimality calculated in the whole design space at every iteration. 4 sampling points are used in the dynamic profiles.*

Figures D.2a-c show that optimal designs such as MBDoE and eMBDoE reduce G-optimality in the whole design space more efficiently than purely explorative ones like LH and DoE.

Moreover, eMBDoE has the smallest values of mean and maximum G-optimality from the fifth iteration onwards.

This is analogous of what is shown in the paper with 3 sampling points for the dynamic profiles, suggesting that the results are not impacted in a relevant way by the $N_{sp}$ considered.

### *D.1.2. Large increase of sampling points: $N_{sp}$=20*

Figure D.3d shows that there is not a clear difference in the number of experiments needed to estimate the most critical parameter ($\theta_4$); this is likely due to the fact that a very large number of sampling points is selected, therefore one experiment provides large information. However, all parameters have a smaller parameters precision (namely, smaller t-values in Figures D.3a-d) when experiments are designed with purely explorative methods such as LH and factorial DoE.

Figure D.4 shows that the G-optimality of purely explorative methods (LH and DoE) is always higher than the one of optimal methods. Moreover, eMBDoE has a smaller G-optimality than MBDoE; the fact that the between the two difference is less evident is likely due to the large information provided by the high number of sampling points in every dynamic profile.

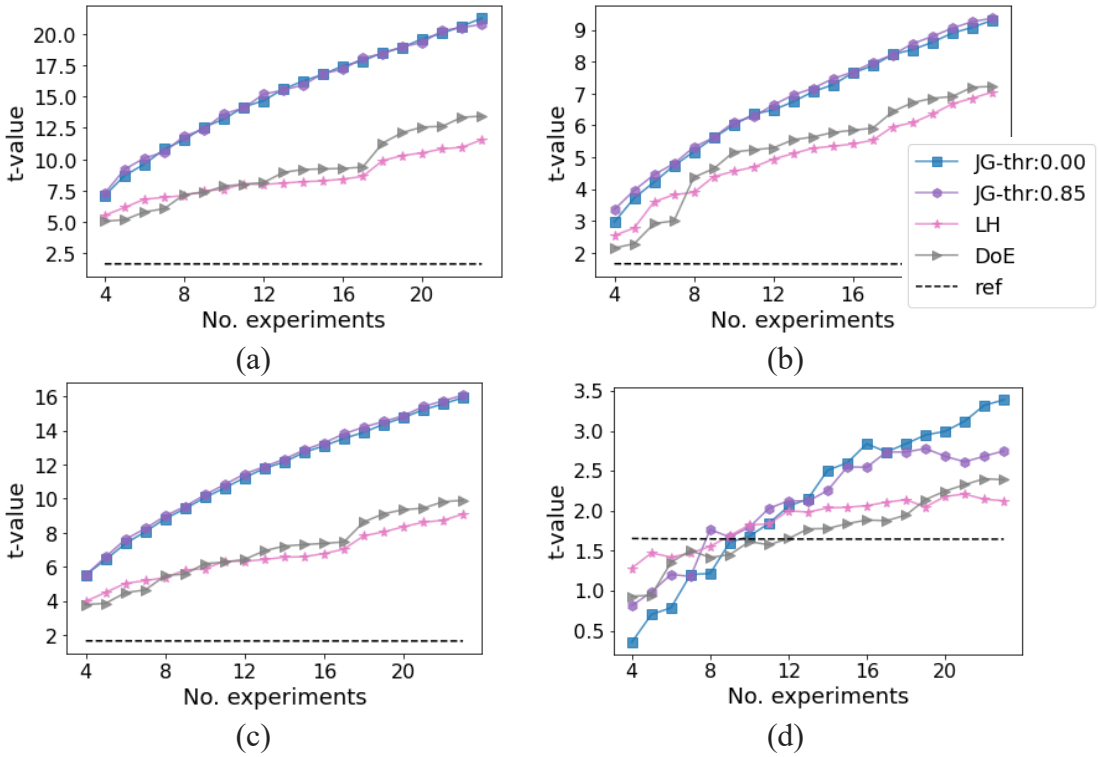



**Figure D.3** *Parameters precision through t-tests: parameters 1-4 in figures a-d, respectively. Results obtained with 20 sampling points of the response variables.*

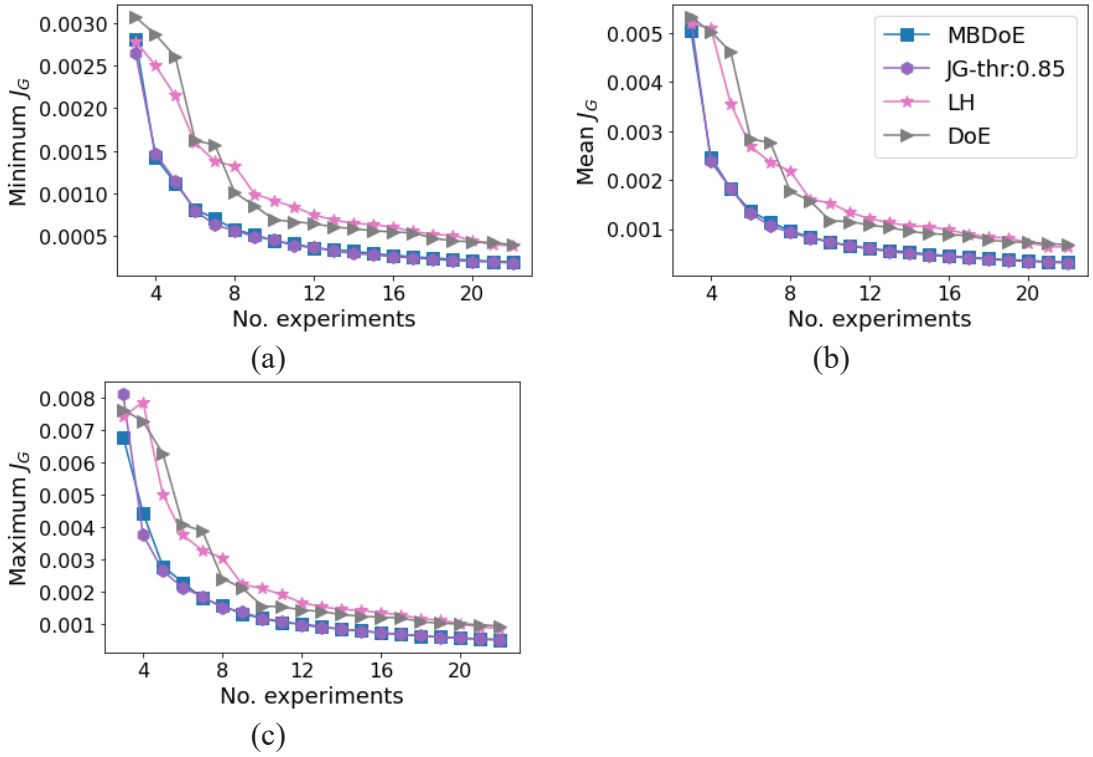


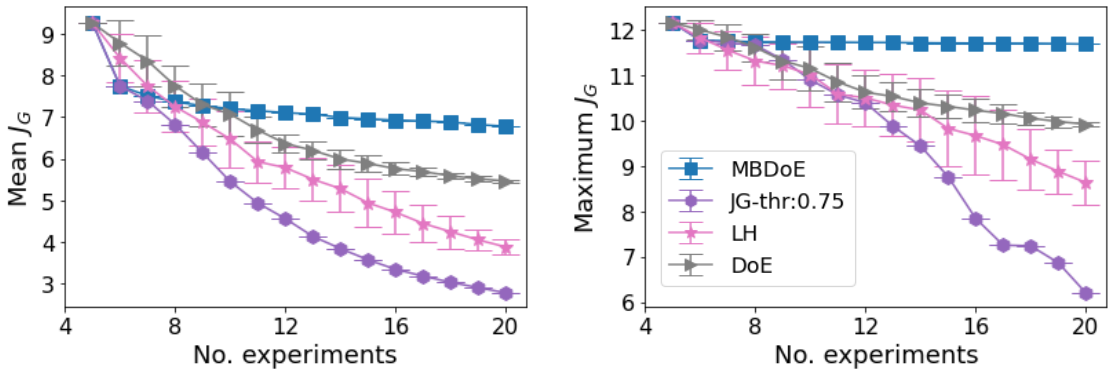

**Figure 4** *Scalar indices of G-optimality: (a) minimum, (b) mean and (c) maximum G-optimality calculated in the whole design space at every iteration. 20 sampling points are used in the dynamic profiles.*

To conclude, both scenarios explored (namely, $N_{sp}$=4 and $N_{sp}$=20) confirm that G-map eMBDoE finds the best trade-off between space exploration and information maximisation and the advantages with respect to conventional design methods is not affected by the choice of the sampling points.

## D.2. Effect of different noise realisations

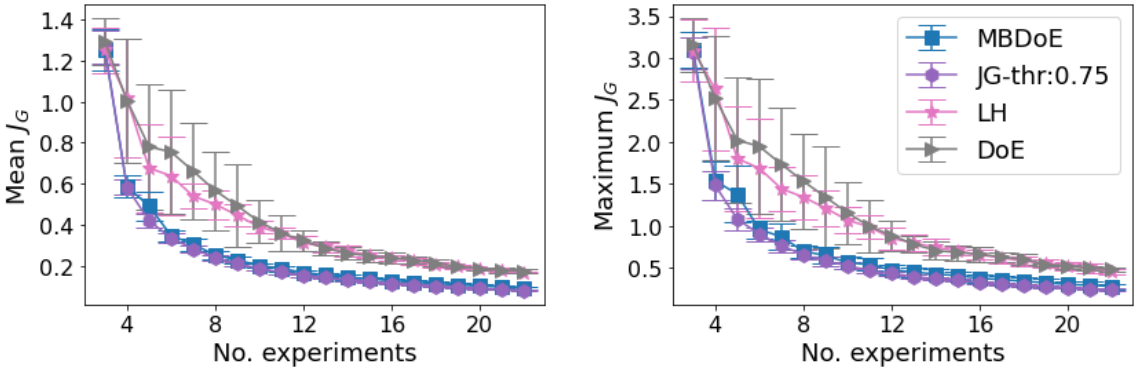
In this Section, a study on the effect of measurement noise on the performance evaluation of different experimental campaigns, namely eMBDoE, MBDoE, factorial Doe and LH, is presented. Therefore, eMBDoE, MBDoE, DoE and LH experiments for Model 1 and 2 are simulated 10 times by using the same settings as in Section 4.3.1 and 4.3.2, respectively, of the main text. More specifically, the settings set equal to those of the main text are: set of preliminary experiments; initial parameters values; ranges of control variables; ranges of parameters estimates; standard deviation of the randomly generated gaussian noise; optimality criterion; G-optimality threshold; experimental budget. In every simulation, the results in terms of mean and maximum G-optimality are stored and their mean and standard deviations are plotted in Figures D.5 and D.6. Notice that in case of DoE and LH, not only the noise of

response variables change for different simulations of the entire experimental campaign, but also the order in which single experiments are added may change in a random way. In fact, there is no systematic method to select the best order to execute DoE or LH experiments sequentially throughout the experimental campaign.




(a) (b)

**Figure D.5** *Scalar indices of G-optimality for case study 1: (a) mean and (b) maximum G-optimality calculated in the whole design space at every iteration. In the plot, symbols indicate the mean value calculated with 10 different simulations (namely, 10 different noise realisations), while vertical bars indicate the calculated standard deviation.*




(a) (b)

**Figure D.6** *Scalar indices of G-optimality for case study 2: (a) mean and (b) maximum G-optimality calculated in the whole design space at every iteration. In the plot, symbols indicate the mean value calculated with 10 different simulations (namely, 10 different noise realisations), while vertical bars indicate the calculated standard deviation.*

As shown in Figures D.5 and D.6, the highest variability is found with DoE and LH, but this is partially due to random noise and partially to the different order with which experiments are progressively added to the calibration dataset. Instead, the variability of the G-optimality obtained with MBDoE and G-map eMBDoE is very low with respect to the mean value.

Despite the variability of DoE and LH, the same conclusions can be made on the reduction of model prediction variance as in the main text: as regards Model 1 (Figure D.5), explorative designs such as LH and DoE reduce G-optimality with respect to MBDoE, but the best

performance is found with eMBDoE with a threshold of 0.75; as far as Model 2 is concerned, explorative designs have always higher G-optimality with respect to MBDoE and eMBDoE (Figure D.6) and the latter one has, overall, the best performance. This proves that the better performance of eMBDoE over conventional design methods is robust and not bound to specific noise realisations.

# Appendix E
# Kinetic model of total methane oxidation

The reaction of total methane oxidation over Pd/Al2O3 catalyst (Chapter 5) is assumed to be unaffected by mass transfer resistances and isothermal within the catalyst particle. Moreover, axial dispersion is negligible in the packed bed reactor. More details can be found in Bawa et al. (2022) and in Pankajakshan et al. (2023). The reactor, modelled as an isothermal plug flow reactor (PFR), can then be described by the equations:

$$\begin{aligned}
&\frac{\mathrm{d}x_1}{\mathrm{d}w} = \frac{Ru_1}{u_2 P_{\mathrm{avg}}}(-r)\,, \quad x_1(0) = u_3\,,\\
&\frac{\mathrm{d}x_2}{\mathrm{d}w} = \frac{Ru_1}{u_2 P_{\mathrm{avg}}}(-2r)\,, \quad x_2(0) = u_3 u_4\,,\\
&\frac{\mathrm{d}x_3}{\mathrm{d}w} = \frac{Ru_1}{u_2 P_{\mathrm{avg}}}(r)\,,\\
&\frac{\mathrm{d}x_4}{\mathrm{d}w} = \frac{Ru_1}{u_2 P_{\mathrm{avg}}}(2r)\,,\\
&y_i = x_i\,, \quad i = 1,2,3
\end{aligned} \tag{E.1}$$

where $x_1, x_2, x_3$ and $x_4$ represent the steady-state mole fractions [molmol$^{-1}$] of $CH_4$, $O_2$, $CO_2$ and $H_2O$, while $u_i$ with $i = 1, \dots, 4$ represent the control variables described in Chapter 5. Moreover, $R$ [J mol$^{-1}$ K$^{-1}$] is the universal gas constant, $w$ [g] is the catalyst mass along the reactor, $P_{\mathrm{avg}}$ [bar] is the average pressure along the reactor, calculated using an empirical pressure drop model based on the Ergun equation (Bawa et al., 2023). Finally, the reaction rate $r$ [mol g$^{-1}$ min$^{-1}$] is given by Mars-van Krevelen kinetic mechanism:

$$r_{CH_4} = \frac{k_1 k_2 P_{CH_4} P_{O_2}}{k_1 P_{O_2} + 2k_2 P_{CH_4} + (k_1 k_2 / k_3) P_{O_2} P_{CH_4}} \tag{E.2}$$

Moreover, reaction rate constants $k_i$ [mol bar$^{-1}$ g$^{-1}$ min$^{-1}$] and adsorption equilibrium constants $K_i$ [bar$^{-1}$] are expressed by Eq. E.3 and A.4, respectively:

$$k_i = k_{i,\mathrm{ref}} \exp\left(\frac{-E_{a,i}}{R}\left[\frac{1}{T} - \frac{1}{T_{\mathrm{ref}}}\right]\right), \qquad i = 1, \dots, N_r\,, \tag{E.3}$$

$$K_i = K_{i,\mathrm{ref}} \exp\left(\frac{-\Delta H_i}{R}\left[\frac{1}{T} - \frac{1}{T_{\mathrm{ref}}}\right]\right), \qquad i = 1, \dots, N_r\,, \tag{E.4}$$

where $k_{i,\mathrm{ref}}$ [mol bar$^{-1}$ g$^{-1}$ min$^{-1}$] is the rate constant at reference temperature, chosen as the average temperature used, $E_{a,i}$ [J mol$^{-1}$] is the activation energy, $T$ [K] is the reaction temperature; $T_{\mathrm{ref}}$ [K] is the reference temperature, chosen as the mean temperature, $K_{i,\mathrm{ref}}$ [bar$^{-1}$] is the adsorption constant at reference (namely, mean) temperature, $\Delta H_i$ [J mol$^{-1}$] is the heat of adsorption, $N_r$ is the number of reactions.

Both Eq. E.3 and E.4 are reparametrised to facilitate model estimation by reducing parameters correlation (Bawa et al., 2022), obtaining Eq. E.5-E.6, respectively:

$$k_i = \exp\left(\theta_{1,i} - \frac{\theta_{2,i}10^4}{R}\left[\frac{1}{T} - \frac{1}{T_{\mathrm{ref}}}\right]\right)$$
$$\theta_{1,i} = \ln(k_{i,\mathrm{ref}}), \qquad \theta_{2,i} = \frac{E_{a,i}}{10^4} \tag{E.5}$$

$$K_i = \exp\left(\theta_{3,i} - \frac{\theta_{4,i}10^4}{R}\left[\frac{1}{T} - \frac{1}{T_{\mathrm{ref}}}\right]\right)$$
$$\theta_{3,i} = \ln(K_{i,\mathrm{ref}}), \qquad \theta_{4,i} = \frac{\Delta H_i}{10^4} \tag{E.6}$$

The reparametrised forms are the ones used during parameters estimation.

# Appendix F
# G-map eMBDoE on total methane oxidation

In this Appendix, more details are found on parameters precision and distribution of information content in the whole design space.

## F.1 Tables of t-values

Tables of $t$-values calculated at every iteration of the G-map eMBDoE procedure are shown in Tables F.1, F.2 and F.3: the best performance is achieved by G-map eMBDoE with $J_{G,\mathrm{thr}}$=0.60, since 5 eMBDoE experiments are enough to estimate all parameters. Instead, conventional MBDoE is able to estimate all parameters at the 23rd experiment, while G-map eMBDoE with $J_{G,\mathrm{thr}}$=0.70 would need a higher number of experiments.

**Table F.1.** *Parameters precision tests for conventional MBDoE.*

| No. MBDoE | t-value $\hat{\theta}_1$ | t-value $\hat{\theta}_2$ | t-value $\hat{\theta}_3$ | t-value $\hat{\theta}_4$ | t-value $\hat{\theta}_5$ | t-value $\hat{\theta}_6$ | t-ref |
|---|---|---|---|---|---|---|---|
| **0** | 17.0274 | 2.4840 | 1.0130 | 1.2404 | **29.8692** | **5.2055** | 1.6973 |
| **1** | 9.8928 | 1.3272 | 5.8495 | 1.9975 | 9.7777 | 1.8016 | 1.6924 |
| **2** | 10.2600 | 1.8029 | 1.1586 | 1.0581 | 41.6467 | 4.8258 | 1.6883 |
| **3** | 12.1550 | 1.8100 | **1.6893** | 1.3893 | 100.3264 | 8.2836 | 1.6849 |
| **4** | 9.2607 | 1.3432 | 7.4699 | 1.6677 | 75.6153 | 6.0163 | 1.6820 |
| **5** | 10.9329 | 1.5101 | 10.6955 | **1.9421** | 71.0188 | 5.6149 | 1.6794 |
| **6** | 2.4819 | 0.6322 | 22.0141 | 3.5922 | 40.5772 | 4.1540 | 1.6772 |
| **7** | 1.1809 | 0.9197 | 23.0338 | 3.9051 | 34.5474 | 3.3905 | 1.6753 |
| **8** | 2.1422 | 0.6194 | 23.9950 | 4.1141 | 37.0895 | 4.1222 | 1.6736 |
| **9** | 2.2709 | 0.8274 | 24.4454 | 4.2008 | 35.6524 | 3.9355 | 1.6720 |
| **10** | 1.3865 | 0.7427 | 25.4787 | 4.3154 | 34.7234 | 3.6920 | 1.6706 |
| **11** | 0.9967 | 0.8506 | 25.4751 | 4.6044 | 34.9845 | 4.1152 | 1.6694 |
| **12** | 2.2084 | 1.1035 | 24.2316 | 4.0237 | 38.4493 | 3.9361 | 1.6683 |
| **13** | 1.9788 | 1.3179 | 25.1187 | 4.4170 | 37.7544 | 4.3245 | 1.6672 |
| **14** | 1.1914 | 0.5619 | 27.5534 | 4.7412 | 34.4522 | 3.4104 | 1.6663 |
| **15** | **3.0797** | 1.3280 | 26.6527 | 4.5932 | 34.1011 | 3.8504 | 1.6654 |
| **16** | 1.7564 | 1.1041 | 25.8311 | 4.2298 | 41.8158 | 3.9606 | 1.6646 |
| **17** | 3.2979 | 0.9710 | 26.6492 | 4.5013 | 38.0398 | 3.9767 | 1.6639 |
| **18** | 5.5333 | 1.3498 | 26.6730 | 4.1703 | 38.8337 | 3.6279 | 1.6632 |
| **19** | 6.0389 | 1.2272 | 26.0571 | 4.1996 | 39.7109 | 3.9137 | 1.6626 |
| **20** | 6.5573 | 1.4326 | 26.1222 | 4.3342 | 39.8329 | 3.5705 | 1.6620 |
| **21** | 6.0102 | 1.1389 | 27.2243 | 4.4628 | 37.9242 | 3.7648 | 1.6614 |
| **22** | 6.1984 | 1.1398 | 26.9772 | 4.4274 | 39.4925 | 3.7996 | 1.6609 |
| **23** | 8.2387 | **1.7800** | 26.1796 | 4.0127 | 42.0700 | 3.9031 | 1.6604 |

**Table F.2.** *Parameters precision tests for G-map eMBDoE with $J_{G,\mathrm{thr}}$=0.70*

| No. eMBDoE | t-value $\widehat{\boldsymbol{\theta}}_1$ | t-value $\widehat{\boldsymbol{\theta}}_2$ | t-value $\widehat{\boldsymbol{\theta}}_3$ | t-value $\widehat{\boldsymbol{\theta}}_4$ | t-value $\widehat{\boldsymbol{\theta}}_5$ | t-value $\widehat{\boldsymbol{\theta}}_6$ | t-ref |
|---|---|---|---|---|---|---|---|
| **0** | 17.0274 | 2.4840 | 1.0130 | 1.2404 | **29.8692** | **5.2055** | 1.6973 |
| **1** | 9.8928 | 1.3272 | 5.8495 | 1.9975 | 9.7777 | 1.8016 | 1.6924 |
| **2** | 10.2600 | 1.8029 | 1.1586 | 1.0581 | 41.6467 | 4.8258 | 1.6883 |
| **3** | 12.1550 | 1.8100 | **1.6893** | 1.3893 | 100.3264 | 8.2836 | 1.6849 |
| **4** | 8.6500 | 1.6603 | 6.3686 | 1.5475 | 92.7433 | 7.3826 | 1.6820 |
| **5** | 10.8325 | 1.6554 | 9.4718 | **2.0946** | 73.2537 | 5.6790 | 1.6794 |
| **6** | 0.4798 | 1.5191 | 22.5975 | 4.0740 | 43.0235 | 3.8311 | 1.6772 |
| **7** | 1.9526 | 0.5055 | 26.7579 | 4.5806 | 35.5143 | 3.7515 | 1.6753 |
| **8** | 0.9430 | 0.8003 | 27.5418 | 4.6295 | 35.6747 | 3.5501 | 1.6736 |
| **9** | 3.4051 | 1.0652 | 27.0508 | 4.4380 | 31.5154 | 3.2857 | 1.6720 |
| **10** | 2.1596 | 0.7598 | 26.2992 | 4.3142 | 42.3413 | 4.0461 | 1.6706 |
| **11** | 1.1153 | 0.8331 | 28.3421 | 4.6323 | 35.9922 | 3.1995 | 1.6694 |
| **12** | **2.8823** | 1.1303 | 27.1236 | 4.2035 | 41.5440 | 3.6986 | 1.6683 |
| **13** | 3.4449 | 1.2278 | 25.1008 | 3.8696 | 48.6858 | 4.1606 | 1.6672 |
| **14** | 3.9144 | 1.0686 | 24.8123 | 4.1482 | 53.4494 | 4.1493 | 1.6663 |
| **15** | 4.0856 | 1.0854 | 25.1671 | 4.1913 | 58.0273 | 5.8025 | 1.6654 |
| **16** | 2.9337 | 1.0418 | 26.3870 | 4.2927 | 53.3763 | 4.5749 | 1.6646 |
| **17** | 2.0566 | 1.0898 | 26.7288 | 4.5988 | 60.2529 | 4.7745 | 1.6639 |
| **18** | 2.2161 | 0.8676 | 27.7055 | 4.8070 | 54.3127 | 5.5972 | 1.6632 |
| **19** | 4.5330 | 1.6150 | 22.7255 | 3.8079 | 81.0074 | 6.3587 | 1.6626 |
| **20** | 2.8178 | 1.2192 | 28.3806 | 4.5119 | 63.1830 | 5.6179 | 1.6620 |
| **21** | 4.7328 | 1.2561 | 26.4000 | 4.1211 | 68.4597 | 5.9146 | 1.6614 |
| **22** | 6.6201 | 1.2250 | 25.8597 | 4.0436 | 68.6939 | 6.4150 | 1.6609 |
| **23** | 4.3769 | 1.3604 | 26.2679 | 4.0969 | 83.8053 | 6.5734 | 1.6604 |

**Table F.3.** *Parameters precision tests for G-map eMBDoE with $J_{G,\mathrm{thr}}$=0.60*

| No. eMBDoE | t-value $\widehat{\boldsymbol{\theta}}_1$ | t-value $\widehat{\boldsymbol{\theta}}_2$ | t-value $\widehat{\boldsymbol{\theta}}_3$ | t-value $\widehat{\boldsymbol{\theta}}_4$ | t-value $\widehat{\boldsymbol{\theta}}_5$ | t-value $\widehat{\boldsymbol{\theta}}_6$ | t-ref |
|---|---|---|---|---|---|---|---|
| **0** | 17.0274 | 2.4840 | 1.0130 | 1.2404 | **29.8692** | 5.2055 | 1.6973 |
| **1** | 15.2018 | 2.5630 | **5.2642** | 2.0054 | 13.0288 | 1.6326 | 1.6924 |
| **2** | 0.2557 | 1.0591 | 18.2675 | 3.4300 | 21.5058 | **2.4591** | 1.6883 |
| **3** | **23.5655** | **3.5975** | 2.8862 | 1.6873 | 42.0451 | 3.6741 | 1.6849 |
| **4** | 22.5605 | 3.5281 | 6.3270 | 1.4909 | 43.1566 | 3.4588 | 1.6820 |
| **5** | 21.9537 | 3.3424 | 6.9861 | **1.7224** | 56.0482 | 4.0992 | 1.6794 |
| **6** | 19.0756 | 2.7389 | 8.8858 | 2.0907 | 60.3679 | 3.9570 | 1.6772 |
| **7** | 19.8097 | 3.5308 | 14.0504 | 2.5956 | 59.9215 | 3.8347 | 1.6753 |
| **8** | 19.1960 | 3.1453 | 13.7473 | 2.7305 | 60.8739 | 4.0752 | 1.6736 |
| **9** | 22.2704 | 3.5130 | 13.6123 | 2.7935 | 55.0802 | 3.7608 | 1.6720 |
| **10** | 20.4348 | 3.0920 | 12.3087 | 2.4451 | 64.4571 | 5.3062 | 1.6706 |
| **11** | 28.2268 | 4.3460 | 14.2304 | 2.5749 | 58.9940 | 4.3630 | 1.6694 |
| **12** | 19.7636 | 2.9611 | 21.0222 | 3.4707 | 57.7133 | 4.7714 | 1.6683 |
| **13** | 22.8066 | 3.6937 | 22.2314 | 3.6977 | 51.2748 | 3.8319 | 1.6672 |
| **14** | 21.0258 | 3.4867 | 23.3802 | 3.6796 | 69.6797 | 4.8195 | 1.6663 |
| **15** | 16.3236 | 2.6964 | 22.9846 | 3.7232 | 96.9638 | 6.1995 | 1.6654 |
| **16** | 18.0052 | 2.7700 | 23.3649 | 3.8769 | 97.6486 | 5.9891 | 1.6646 |
| **17** | 18.7089 | 2.8850 | 25.3435 | 4.0612 | 93.7648 | 5.9024 | 1.6639 |
| **18** | 17.3838 | 2.6607 | 23.1809 | 3.8085 | 111.4196 | 7.1495 | 1.6632 |
| **19** | 17.0912 | 2.6597 | 22.9867 | 3.7590 | 119.3977 | 7.1632 | 1.6626 |
| **20** | 17.9229 | 2.7546 | 22.2085 | 3.7286 | 122.3977 | 7.2531 | 1.6620 |
| **21** | 20.6074 | 3.0560 | 22.6773 | 3.9009 | 116.0847 | 6.6831 | 1.6614 |
| **22** | 19.9154 | 2.9464 | 23.2681 | 4.0798 | 119.2387 | 6.9285 | 1.6609 |
| **23** | 20.2832 | 2.9576 | 24.3938 | 4.1423 | 113.9940 | 6.8429 | 1.6604 |

## F.2 H-maps at different iterations

G-map eMBoE uses H-maps (namely, maps of information content) to select the most informative experiment among the candidates that satisfy the G-optimality constraint, as described in Chapter 6 of the main text. Figure F.1 shows the H-maps obtained with 1 optimal experiment and it can be used to compare the distribution of information with the distribution of model prediction variance (Chapter 6) in the design space.

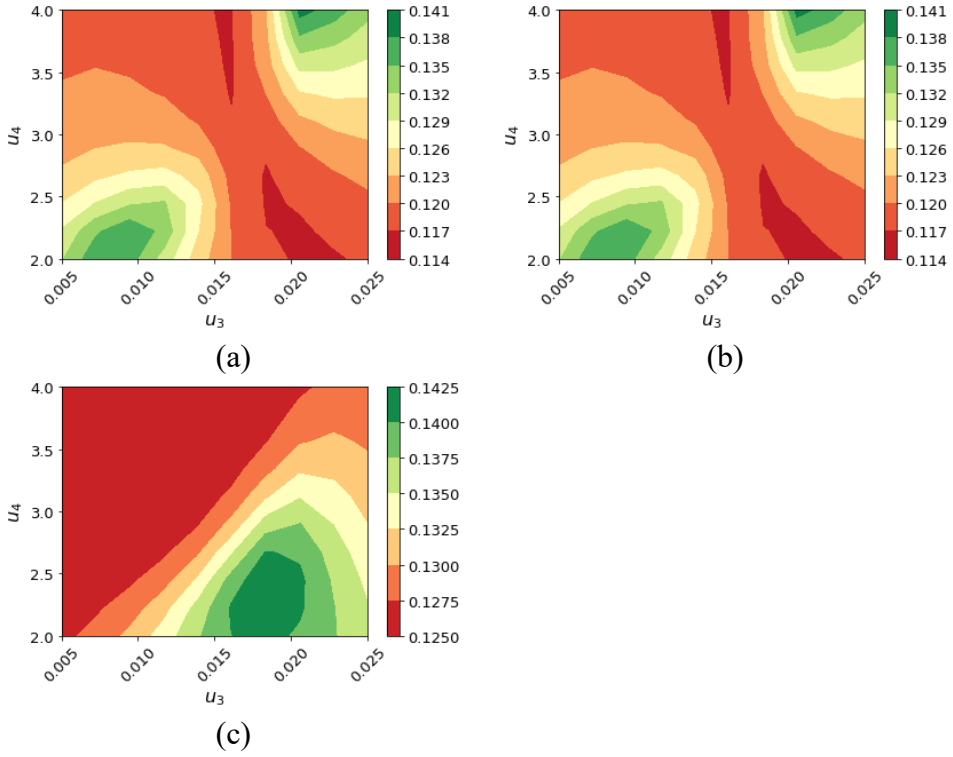
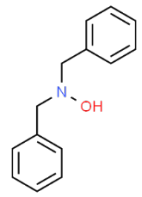



**Figure F.1**. *H-maps (maps of information content) obtained with 1 optimal experiment (besides the 12 preliminary ones) calculated by: (a) MBDoE; (b) G-map eMBDoE with* $J_{G,thr}$*=0.70; (c) G-map eMBDoE with* $J_{G,thr}$*=0.60.*

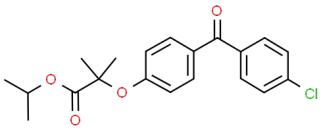
The comparison between H-maps in Figures F.1a-c and G-maps in Chapter 6 of the main text suggests that there is not necessarily a correspondence between regions of high information content and regions of small model prediction variance. For instance, considering the results of conventional MBDoE and G-map eMBDoE with $J_{G,\text{thr}}$=0.70 after one optimal experiment, the region with $u_3$ less than 0.015 molmol$^{-1}$ and $u_4$ less than 2.8 molmol$^{-1}$ is characterised by high information content (Figure B.1a-b) and low model prediction variance (Figure in Chapter 6), but the region with $u_3$ higher than 0.018 molmol$^{-1}$ and $u_4$ higher than 3 molmol$^{-1}$ has high information content (Figure B.1a-b) but also medium-high values of model prediction variance (Figure in Chapter 6). Similarly, G-map eMBDoE with $J_{G,thr}$=0.60 after one optimal experiment has a high information where $u_3$ is higher than 0.015 molmol$^{-1}$ (Figure B.1c), which is also where the highest model prediction variance values are found (Figure in Chapter 6).

# Appendix G
# Drug and drug-like molecules to develop the organic solubility model

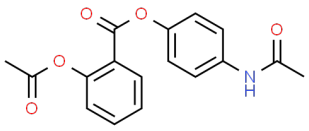
This appendix shows the chemical formula of the solutes employed to test the PLS model proposed in Chapter 7.



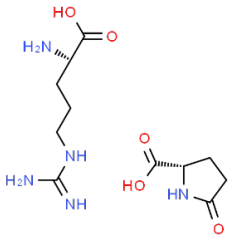
1) N,N-Dibenzylhydroxylamine (DBHA)
2) $C_{14}H_{15}NO$
3) 621-07-8



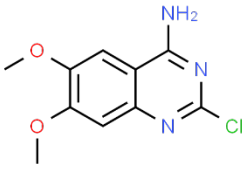
1) Fenofibrate
2) $C_{20}H_{21}ClO_4$
3) 49562-28-9



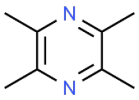
1) Benorilate
2) $C_{17}H_{15}NO_5$
3) 5003-48-5



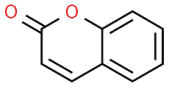
1) L-Arginine L-pyroglutamate
2) $C_{11}H_{21}N_5O_5$
3) 56265-06-6



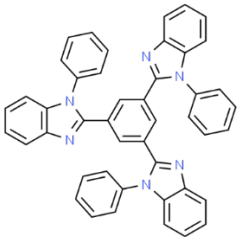
1) 2-chloro-4-amino-6,7-dimethoxyquinazoline
2) $C_{10}H_{10}ClN_3O_2$
3) 23680-84-4



1) Tetramethylpyrazine
2) $C_8H_{12}N_2$
3) 1124-11-4

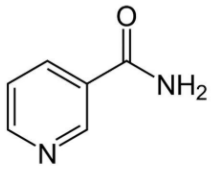


1) Coumarin
2) $C_9H_6O_2$
3) 91-64-5



1) 1,3,5-Tris(1-phenyl-1H-benzimidazol-2-yl)benzene (TPBi)
2) $C_{45}H_{30}N_6$
3) 192198-85-9

**Figure G.1**. *Solutes of the datasets retrieved from Krasnov et al., (2022): 1) name; 2) chemical formula; 3) CAS number.*




1) Nicotinamide
2) $C_6H_6N_2O$
3) 98-92-0

**Figure G.2**. *Chemical structure of the solute of the dataset retrieved from Khajir et al. (2024). The molecule is identified through: 1) name; 2) chemical formula; 3) CAS number.*

# Appendix H
# In silico screening of miscibility and evaporation issues

As explained in Section 7.2 of the main text, separated phases as well as excessive solvent evaporation should be avoided during solubility experiments. The purpose of this appendix is to explain how candidate binary mixtures are retained based on in silico predictions of their miscibility and evaporation issues. In order to screen those solvent pairs where immiscibility may arise under the conditions of interest, different binary solvent pairs are simulated at $T$ = {20, 50} °C, $P$ = 1 atm and molar fraction $x$ = 0–100 mol% from the candidate solvent list using UNIFAC (Fredenslund et al., 1975), NRTL (Renon et al., 1978), UNIFAC Modified (Gmehling et al., 1993), UNIFAC LLE (Magnussen et al., 1981) and UNIQUAC (Abrams et al., 1975) activity coefficient models via the Vapour-Liquid and Liquid-Liquid Equilibrium utility in DynoChem® by Scale-up Systems Ltd. (https://www.scale-up.com/). Those solvent pairs where immiscibility issues are flagged by any of the considered activity coefficient models are not proposed as experiments in the workflows. The same models are used to screen those solvent pairs where evaporation issues may arise under the conditions of interest at $P$ = 1 atm and $x$ = 0–100 mol%. If a given binary solvent mixture is not predicted to form an azeotrope, the maximum and minimum boiling points (and the corresponding mixture compositions) are those of each of the pure components in the mixture. If a given binary solvent mixture is predicted to form an azeotrope, the maximum or minimum boiling point (depending if a maximum or minimum azeotrope is formed, respectively) and the corresponding mixture composition is the azeotrope point. Those solvent pairs whose minimum boiling point is equal to or less than 55°C (i.e., within 5°C of the upper temperature of 50 °C at which solubility is measured) are not proposed as experiments in the workflows.

# Appendix I
# Effect of non-ideal mixing of organic solvents

Mixtures of two or three organic solvents are prepared in practice by handling specified volumes, as explained in section 7.2 of Chapter 7. Therefore, the molar fractions $x_i$ to build **U** (see Eq.s 7.8, 7.9 of the main text) must be calculated starting from the volumetric quantities employed experimentally.

Results of section 7.3.1-7.3.2 of the main text consider molar fractions of organic mixtures before API dissolution, under the assumption of no volumetric effects due to liquids mixing. Therefore, volumetric fractions are considered:

$$x_i^V = V_i/V_{\text{tot},0}, \qquad i=1,\ldots, N_L, \tag{I.1}$$

where $V_i$ is the volume of the i-th solvent added to the mixture, while $V_{\text{tot},0}$ is the total nominal volume $V_{\text{tot},0} = V_1 + \ldots + V_{N_L}$. In turn, volumetric fractions are used to calculate the number of moles $n_i$ of solvents in the vial:

$$n_i \cong \rho_i x_i^V V_{\text{vial}}/MW_i, \qquad i=1,\ldots, N_L, \tag{I.2}$$

where $N_L$ is the number of organic solvents in the mixture, $V_{\text{vial}}$ is the volume of liquid mixture in the vial, $\rho_i$ and $MW_i$ are density and molecular weight of the $i$-th solvent, respectively.

Therefore, molar fractions of organic mixture before API dissolution and assuming no volumetric effects become:

$$x_i \cong n_i/\sum_{j=1}^{N_o} n_j \qquad i=1,\ldots, N_L, \tag{I.3}$$

Molar fractions of Eq.I.3 are used to obtain the results in the main text.

In this appendix, the same modelling approach is applied to **U** matrices built with molar fractions obtained removing the assumption of absence of volumetric effects and considering the presence of API. The total volume after mixing $V_{\text{tot,meas}}$ is experimentally measured for every solution (as explained in section 7.2 of the main text) and used to calculate volumetric concentrations for every solvent in the mixture as $c_i^V = V_i/V_{\text{tot,meas}}$ . Therefore, molar fractions of solvents in the vial are calculated as:

$$n_i^{\mathrm{eq}} = \rho_i c_i^V V_{\mathrm{vial}}/MW_i, \qquad i=1,\ldots, N_L. \tag{I.4}$$

API solubility is measured as mass concentration $c_{\mathrm{API}}^{\mathrm{eq}}$ [mg/mL] in the liquid solution at equilibrium, therefore it is converted into number of moles in the vial $n_{\mathrm{API}}^{\mathrm{eq}}$:

$$n_{\mathrm{API}}^{\mathrm{eq}} \cong c_{\mathrm{API}}^{\mathrm{eq}}\, V_{\mathrm{vial}}/MW_{\mathrm{API}}, \qquad i=1,\ldots, N_L. \tag{I.5}$$

Eq. (I.5) assumes that $V_{\mathrm{vial}}$ is the same as in Eq. (I.4), namely it assumes that API dissolution does not change total volume considerably.

This results in the following molar fractions for the organic solvents:

$$x_i^{\mathrm{eq}} \cong n_i^{\mathrm{eq}}/(\textstyle\sum_{j=1}^{N_o} n_j^{\mathrm{eq}} + n_{\mathrm{API}}^{\mathrm{eq}}) \qquad i=1,\ldots, N_L, \tag{I.6}$$

Figure I.1a-b shows the results of model calibration and validation with the same datasets employed in section 3.1-3.2 of the main text, but using molar fractions of Eq. (I.6) to build **U**.

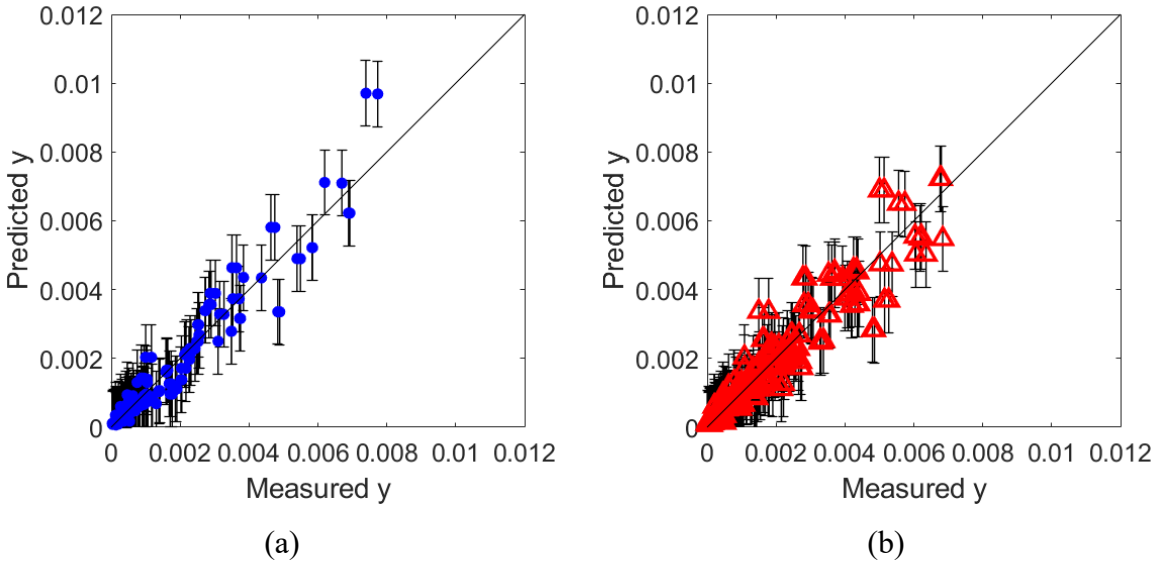



**Figure I.1**. *Results of the PLS model with: a) calibration experiments; b) all validation experiments. Molar fractions at equilibrium are used to build the matrix of regressors, removing the assumptions of ideal mixing of liquid solvents.*

The results obtained with molar fractions $x_i^{\mathrm{eq}}$ (Eq. I.6) calculated considering mixing effects and the dissolution of API are almost identical to the ones obtained with approximated molar fractions $x_i$ (Eq. I.3) in the main text. This is confirmed by the determination coefficient: $R^2$=0.92 and $R^2$=0.90 for calibration and validation data, respectively, exactly as it was in sections 7.3.1-7.3.2 of the main text.

# Appendix J
# Intestinal absorption in PBPK studies

In this Appendix, more details on the biological phenomena leading to intestinal absorption are provided, in order to contextualise the PBPK models discussed in Chapter 8. After providing a general overview of the main modelling framework, the attention focuses on the interplay of phenomena that can favor or limit the drug absorption at intestinal level. The role of drug physico-chemical properties and of food digestion products is explained, too.

## J.1 PBPK modelling

PBPK models are made of a series of differential equations that describe several phenomena involved in ADME (Figure J.1):

- Absorption, i.e. the movement of the drug from the administration site to the bloodstream;
- Distribution, i.e. the journey of the drug through the bloodstream to the different tissues;
- Metabolism, i.e. the process that breaks down the drug;
- Excretion, i.e. the removal of the drug from the body.

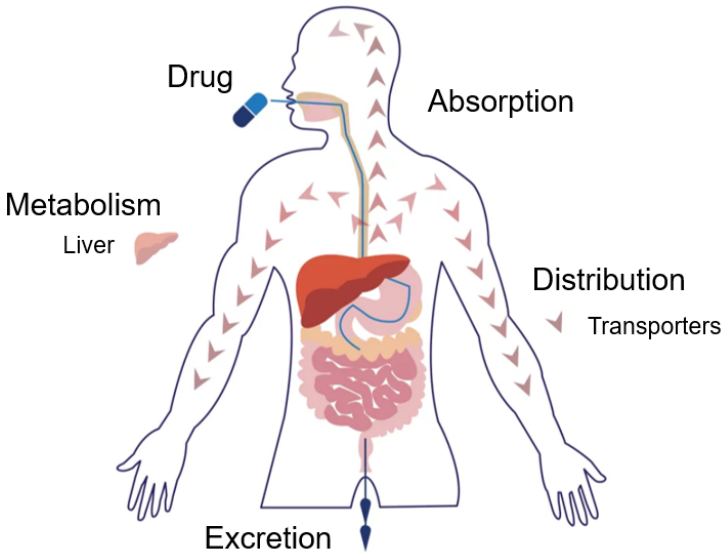



**Figure J.1**. *Illustration of ADMDE (adapted from www.toolbox.eupati.eu)*

In a PBPK model, the different organs or tissues of the human body are represented as different compartments linked by blood flows. In turn, a compartment can be divided into different segments: for instance, the gut is further compartmentalised into 9 segments in order to represent the variations in physiological conditions that characterise every segment (Figure J.2).

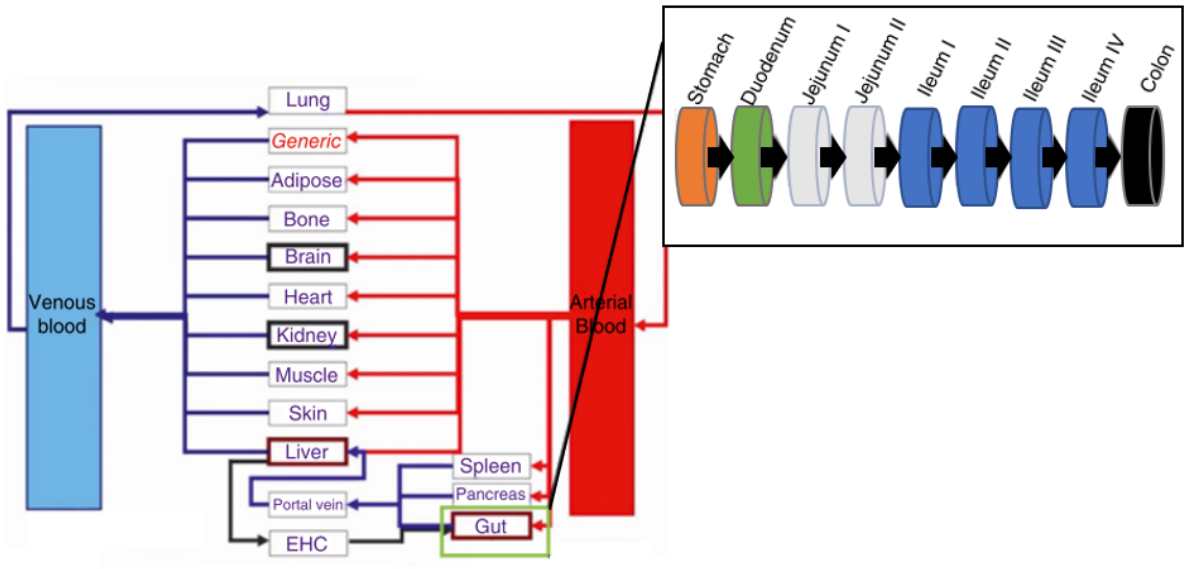


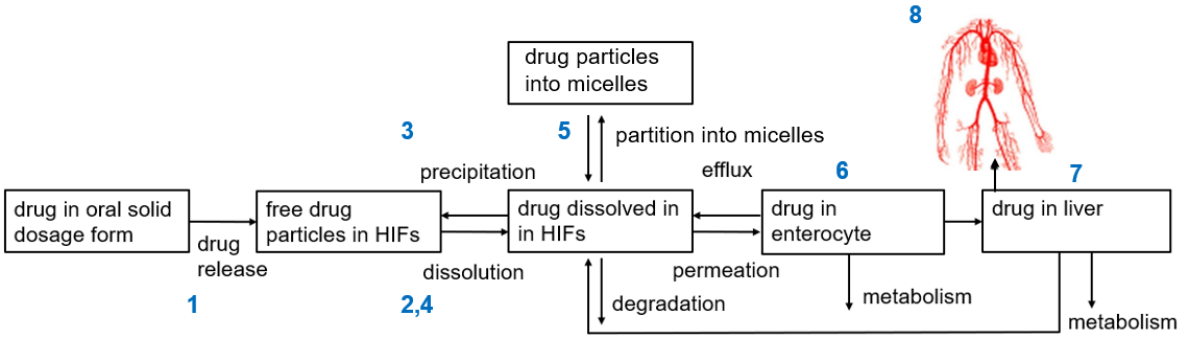

**Figure J.2**. *Schematic of the compartments of the human body represented in PBPK models. The 9 segments of the gut compartment are shown, too (adapted from Stamatopoulos, 2022).*

The differential equations of every compartment combine physiological conditions (e.g., pH, bile salts concentrations, transit time, etc.) with drug properties (e.g., solubility, diffusion coefficient, water to micelle partition coefficient, etc.). When intestinal absorption (which is of main interest for this Dissertation) is considered, solubility is one of the key drug properties.

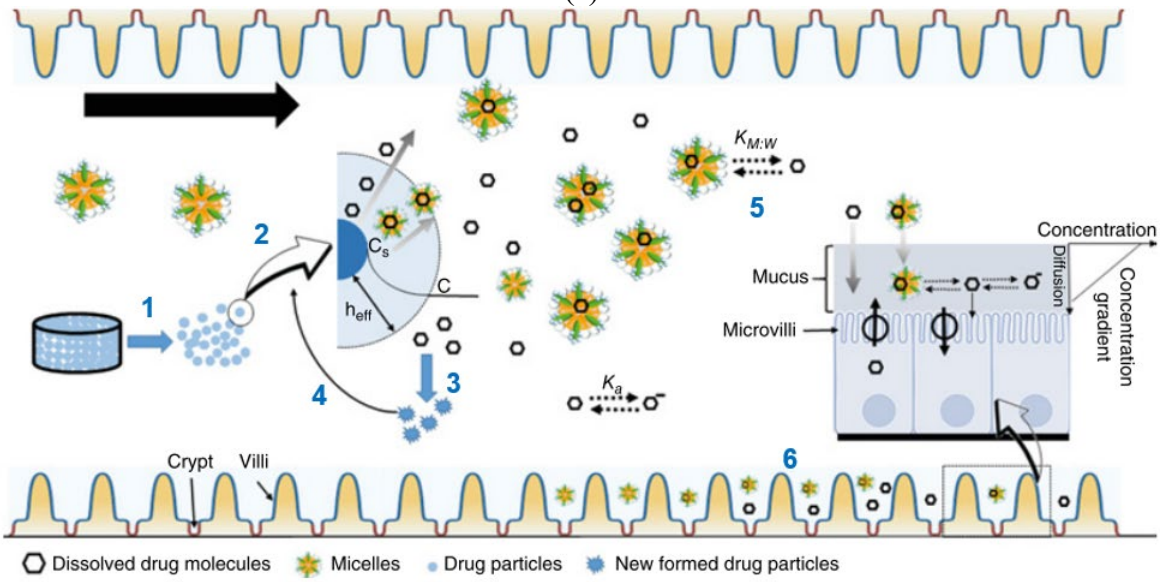
Figure J.3 shown the main steps involved in drug absorption from the administration site to the bloodstream (Figure J.3a), focusing the attention to the luminal environment (Figure J.3b; Stamatopoulos, 2022; Abrahamsson et al., 2020):

- once the oral solid dosage form is ingested by the patient, it reaches the gastrointestinal tract and it starts disintegrating, releasing its particles to the human intestinal fluid (step 1 in Figure J.3a-b);
- the released drug particles are dissolved into the human intestinal fluid (step 2 in Figure J.3a-b). Based on the dissociation constant $K_a$ (or its logarithm $\log(K_a)$) of the drug, ionisation of the dissolved molecules can take place; consider that the dissociation constant must be considered for acidic species ($K_{a,1}$), basic species ($K_{a,2}$) and ampholytes (both $K_{a,1}$ and $K_{a,2}$);

- if the concentration of the drug overcomes the maximum concentration that can be dissolved in that fluid, precipitation occurs (step 3 in Figure J.3a-b);
- the precipitated particles can be resolubilised (step 4 in Figure J.3a-b);
- based on the lipophilicity of the drug, partitioning into the micelles can take place (step 5 in Figure J.3a-b), thus enhancing drug solubilisation. The drug partitioning into micelles is characterised through the micelle-water partition coefficient $K_{m:w}$, which can be calculated for both unionised $K_{m:w,u}$ and ionised $K_{m:w,i}$ drug particles;
- drug molecules, both free and bound to micelles, diffuse through the mucus layer on the surface of the enterocytes, then free molecules (if coming from micelles, they must be released) permeate the membrane of the enterocytes (step 6 in Figure J.3a-b);
- then, excluding particles subjected to degradation, drug particles can reach liver (step 7 in Figure J.3a) and, ultimately, the bloodstream (step 8 in Figure J.3a).




(a)




(b)

**Figure J.3**. *Illustration of the main phenomena leading to drug absorption: (a) from administration to bloodstream; (b) in the luminal environment (adapted from Stamatopoulos, 2022; Abrahamsson et al., 2020).*

In the human intestinal fluids, bile salts form micelles that can enhance the solubilisation of the drug. This effect is further enhanced by the presence of lecithin. Moreover, food intake has usually a positive effect on drug solubility: *i*) bile salts and lecithin concentrations increase in the fed condition with respect to the fasted one; *ii*) food digestion products, such as lipid degradation products, form mixed micelles together with bile salts and lecithin, enhancing the overall solubilising effect (Clarysse et al., 2011).

Finally, the plasma-concentration profiles that are commonly measured during clinical trials concern the concentration of the drug that reaches the bloodstream (step 8) after all the abovementioned phenomena (from step 1 to 8). Predicted plasma-concentration profiles can be obtained by the simulator, for instance by Simcyp, by modelling every aspect of intestinal absorption. In other terms, all the phenomena are described by mathematical mdoels, including the properties of the drug and of the luminal environment. In this Dissertation, only the solubility model is analysed in order to be improved, while the remaining mathematical framework of Simcyp is retained. In fact, ongoing work is focusing on the implementation of the proposed solubility model within the set of equations implemented in Simcyp to describe all the other properties or phenomena, with the aim of improving the final prediction of plasma-concentration profiles.

# References


Abboud, L., Hensley, S., 2003. New prescription for drug makers: update the plants. Wall Str. J. 09/23/2003. Available at: https://www.wsj.com/articles/SB10625358 403931000. Last accessed on: 29/08/2023

Abt, V., Barz, T., Cruz -Bournazou, M. N., Herwig, C., Kroll, P., Möller, J., Pörtner, R., Schenkendorf, R., 2018. Model-based tools for optimal experiments in bioprocess engineering, Model-based tools for optimal experiments in bioprocess engineering, Current Opinion in Chemical Engineering, 22, 244–252

Abuhassan, Q., Khadra, I., Pyper, K., Augustijns, P., Brouwers, J., & Halbert, G. W. (2022). Fasted intestinal solubility limits and distributions applied to the biopharmaceutics and developability classification systems. *European Journal of Pharmaceutics and Biopharmaceutics*, *170*(September 2021), 160–169.

Ainousah, B. E., Perrier, J., Dunn, C., Khadra, I., Wilson, C. G., & Halbert, G. (2017). Dual Level Statistical Investigation of Equilibrium Solubility in Simulated Fasted and Fed Intestinal Fluid. *Molecular Pharmaceutics*, *14*(12), 4170–4180. https://doi.org/10.1021/acs.molpharmaceut.7b00869

Akkermans, S., Nimmegeers, P., Van Impe, J. F. (2018). Comparing design of experiments and optimal experimental design techniques for modelling the microbial growth rate under static environmental conditions. *Food Microbiology*, 76, 504–512.

Allison, K., Patel, D., & Kaur, R. (2022). Assessing Multiple Factors Affecting Minority Participation in Clinical Trials: Development of the Clinical Trials Participation Barriers Survey. *Cureus*, *14*(4), e24424.

Alpizar-Ramos, S., González-de la Parra, M. (2017). Application of Sequential Design of Experiments to Develop Ibuprofen (400 mg) Tablets by Direct Compression. *Asian Journal of Chemistry and Pharmaceutical Sciences*, 2(1), 10.

Apelblat, A.; Manzurola, E. Solubilities of o -acetylsalicylic, 4- aminosalicylic, 3,5-dinitrosalicylic, and p -toluic acid, and magnesium-, ja:math -aspartate in water from T = (278 to 348) K. J. Chem. Thermodyn. 1999, 31,85−91

Arshad, M. S., Saman Zafar, Bushra Yousef, Yasmine Alyassin, Radeyah Ali, Ali AlAsiri, Ming-Wei Chang, Zeeshan Ahmad, Amal Ali Elkordy, Ahmed Faheem, Kendal Pitt, A review of emerging technologies enabling improved solid oral dosage form manufacturing and processing, Advanced Drug Delivery Reviews, Volume 178, 2021, 113840, ISSN 0169-409X, https://doi.org/10.1016/j.addr.2021.113840.

Asprey, S. P., Macchietto, S. (2000). Statistical Tools for Optimal Dynamic Model Building. *Computers and Chemical Engineering*, 24, 1261-1267.

Asprey, S. P., Naka, Y. (1999). Mathematical problems in fitting kinetic models-some new perspectives. *Journal of Chemical Engineering of Japan*, 32(3), 328–337.

Asprion, N., Böttcher, R., Pack, R., Stavrou, M. E., Höller, J., Schwientek, J., & Bortz, M. (2019). Gray-Box Modeling for the Optimisation of Chemical Processes. *Chemie-Ingenieur-Technik*, *91*(3), 305–313.

Augustijns, P., Wuyts, B., Hens, B., Annaert, P., Butler, J., & Brouwers, J. (2014). A review of drug solubility in human intestinal fluids: Implications for the prediction of oral absorption. *European Journal of Pharmaceutical Sciences*, *57*(1), 322–332. https://doi.org/10.1016/j.ejps.2013.08.027

Bard, Y. (1974). Nonlinear Parameter Estimation. New York, NY: Academic Press.

Barz,T, Kager,J., Herwig,C., Neubauer, P., Bournazou,M. N. C., Galvanin,F., 2022. Chapter 11 - Characterization of reactions and growth in automated continuous flow and bioreactor platforms—From linear DoE to model-based approaches. Michael Bortz, Norbert Asprion, Ed., in Simulation and Optimisation in Process Engineering,Elsevier,Pages 273-319

Basak, S., Petit, S., Bect, J., Vazquez, E. (2022). Numerical Issues in Maximum Likelihood Parameter Estimation for Gaussian Process Interpolation. In: Nicosia, G., *et al.* Machine Learning, Optimization, and Data Science. LOD 2021. Lecture Notes in Computer Science(), vol 13164. Springer, Cham. https://doi.org/10.1007/978-3-030-95470-3_9

Beg, Sarwar. ed., 2021. Design of Experiments for Pharmaceutical Product Development: Volume II : Applications and Practical Case Studies. SpringerLink (Online service), 1st ed., http://lib.ugent.be/catalog/ebk01:4100000011728437

Bernaerts, K., Servaes, R. D., Kooyman, S., Van Impe, J. F. (2001). Iterative Optimal Experiment Design for Estimation of Microbial Growth Kinetics as Function of Temperature. *IFAC Proceedings Volumes*, 34(5), 19–24.

Bhattamisra, S. K., Banerjee, P., Gupta, P., Mayuren, J., Patra, S., & Candasamy, M. (2023). Artificial Intelligence in Pharmaceutical and Healthcare Research. *Big Data and Cognitive Computing*, *7*, 10.

Bogacka, B., Patan, M., Johnson, P. J., Youdim, K., Atkinson, A. C. (2011). Optimum design of experiments for enzyme inhibition kinetic models. *Journal of Biopharmaceutical Statistics*, 21(3), 555–572.

Bonate, P. L. (2011). Pharmacokinetics. *Wiley Interdisciplinary Reviews: Computational Statistics*, *3*(4), 332–342. https://doi.org/10.1002/wics.153

Bonvin, D., Georgakis, C., Pantelides, C. C., Barolo, M., Grover, M. A., Rodrigues, D., Schneider, R., & Dochain, D. (2016). Linking Models and Experiments. *Industrial and Engineering Chemistry Research*, *55*, 6891–6903.

Boobier, S., Hose, D. R. J., Blacker, A. J., Nguyen, B. N. (2020). Machine learning with physicochemical relationships: solubility prediction in organic solvents and water. *Nature Communications*, *11*, 5753. https://doi.org/10.1038/s41467-020-19594-z

Borhani, T. N., García-Muñoz, S., Vanesa Luciani, C., Galindo, A., & Adjiman, C. S. (2019). Hybrid QSPR models for the prediction of the free energy of solvation of organic solute/solvent pairs. *Physical Chemistry Chemical Physics*, *21*, 13706–13720. https://doi.org/10.1039/c8cp07562j

Borkar, S., Ghutke, P., Patil, W., Joshi, S., & Sorte, S. (2023). A Review of Pick and Place Robots for the Pharmaceutical Industry. *11th International Conference on Emerging*

*Trends in Engineering & Technology - Signal and Information Processing (ICETET - SIP) A*. https://doi.org/10.1109/icetet-sip58143.2023.10151652

Bouillot, B., Teychené, S., & Biscans, B. (2011). An evaluation of thermodynamic models for the prediction of drug and drug-like molecule solubility in organic solvents. *Fluid Phase Equilibria*, *309*(1), 36–52. https://doi.org/10.1016/j.fluid.2011.06.032

Box, G. E. P., Lucas, H. L. (1959). Design of experiments in non-linear situations. *Biometrika*, 46, 77-90.

Box, M. J. 1968. The Occurrence of Replications in Optimal Designs of Experiments to Estimate Parameters in Non-Linear Models, *Journal of the Royal Statistical Society: Series B (Methodological)*, 30 (2), 290–302. https://doi.org/10.1111/j.2517-6161.1968.tb00728.x

Box, G.E.P., Hill, W.J.,1967. Discrimination among mechanistic models. Technometrics, 9, 57–71.

Buchowski, H.; Ksiazczak, A.; Pietrzyk, S. Solvent activity along a saturation line and solubility of hydrogen-bonding solids. J. Phys. Chem. 1980, 84, 975−979

Buxbaum JD, Chernew ME, Fendrick AM, Cutler DM. Contributions Of Public Health, Pharmaceuticals, And Other Medical Care To US Life Expectancy Changes, 1990-2015. Health Aff (Millwood). 2020 Sep;39(9):1546-1556. doi: 10.1377/hlthaff.2020.00284. PMID: 32897792.

Buzzi-Ferraris, G., Forzatti, P., 1983. A new sequential experimental design procedure for discriminating among rival models, Chemical Engineering Science, 38 (2), 225-232. https://doi.org/10.1016/0009-2509(83)85004-0

Buzzi-Ferraris, G., Forzatti, P., Canu, P.,1990. An improved version of a sequential design criterion for discriminating among rival multiresponse models. ChemicalEngineeringScience45,477–481.

Buzzi-Ferraris, G.,Forzatti,P.,Emig,G.,Hofmann,H.,1984.Sequential experimental design for model discriminating in the case of multiresponse models. ChemicalEngineeringScience39,81–85.

Byrd, R., Lu, P., Nocedal, J., and Zhu, C. (1995). A limited memory algorithm for bound constrained optimization. SIAMJ. Sci. Comput., 16(5):1190–1208

Cherkasov, N., Bai, Y., Expósito, A. J., & Rebrov, E. V. (2018). OpenFlowChem-a platform for quick, robust and flexible automation and self-optimisation of flow chemistry. *Reaction Chemistry and Engineering*, *3*(5), 769–780. https://doi.org/10.1039/c8re00046h

Chinta, S., Rengaswamy, R., 2019. Machine Learning Derived Quantitative Structure Property Relationship (QSPR) to Predict Drug Solubility in Binary Solvent Systems. Industrial and Engineering Chemistry Research, 58, 3082−3092.

Chong, I.G., and C.H. Jun (2005). Performance of some variable selection methods when multicollinearity is present. *Chemom. Intell. Lab. Syst.*, 78, 103-112.

Chung, S. H., Ma, D. L., Braatz, R. D.., 2000. Optimal model-based experimental design in batch crystallization, Chemometrics and Intelligent Laboratory Systems 50, 83–90

Clarysse, S., Brouwers, J., Tack, J., Annaert, P., & Augustijns, P. (2011). Intestinal drug solubility estimation based on simulated intestinal fluids: Comparison with solubility

in human intestinal fluids. *European Journal of Pharmaceutical Sciences*, *43*(4), 260–269.

Collins, P. C. (2018). Chemical engineering and the culmination of quality by design in pharmaceuticals. *AIChE Journal*, *64*(5), 1502–1510. https://doi.org/10.1002/aic.16154

Constable, D.J., Jiménez-González, C., Henderson, R.K., 2007. Perspective on solvent use in the pharmaceutical industry. Organ. Process Res. Develop. 11 (1), 133– 137.

Cox Gad, S. 2008. Pharmaceutical Manufacturing Handbook: Production and Processes, John Wiley & Sons, Inc.

Cysewski, P., Jeliński, T., Przybyłek, M., Nowak, W., & Olczak, M. (2022). Solubility Characteristics of Acetaminophen and Phenacetin in Binary Mixtures of Aqueous Organic Solvents: Experimental and Deep Machine Learning Screening of Green Dissolution Media. *Pharmaceutics*, *14*, 2828. https://doi.org/10.3390/pharmaceutics14122828

Czitrom, V. 1999. One-factor-at-a-time versus designed experiments. *American Statistician*, 53(2), 126–131. https://doi.org/10.1080/00031305.1999.10474445

Dahlgren, D., Venczel, M., Ridoux, J. P., Skjöld, C., Müllertz, A., Holm, R., Augustijns, P., Hellström, P. M., & Lennernäs, H. (2021). Fasted and fed state human duodenal fluids: Characterization, drug solubility, and comparison to simulated fluids and with human bioavailability. *European Journal of Pharmaceutics and Biopharmaceutics*, *163*(February), 240–251. https://doi.org/10.1016/j.ejpb.2021.04.005

Dansereau, R. and G. Peck. (1987). The Effect of the Variability in the Physical and Chemical Properties of Magnesium Stearate on the Properties of Compressed Tablets. *Drug Development and Industrial Pharmacy*, 13, 975-999.

D'Argenio, 1981. Optimal Sampling Times for Pharmacokinetic Experiments, Journal of Pharmacokinetics and Biopharmaeeutics, 9, 6

Dasgupta, S. , Mukhopadhyay, S., Keith, J. 2021. G-optimal grid designs for kriging models. arXiv. https://doi.org/10.48550/arXiv.2111.06632

De la Cruz Moreno, M.P.; Oth, M.; Deferme, S.; Lammert, F.; Tack, J.; Dressman, J.; Augustijns, P. Characterization of Fasted- State Human Intestinal Fluids Collected from Duodenum and Jejunum. J. Pharm. Pharmacol. 2006, 58, 1079–1089.

De la Cruz-Moreno, Mariangeles Pérez; Consuelo Montejo, Antonio Aguilar-Ros, Walthère Dewe, Benoît Beck, Jef Stappaerts, Jan Tack, Patrick Augustijns, 2017. Exploring drug solubility in fasted human intestinal fluid aspirates: Impact of inter-individual variability, sampling site and dilution, International Journal of Pharmaceutics, Volume 528, Issues 1–2, Pages 471-484

de Prada, C., Pantelides, C.C, Pitarch, J.L., 2019. Special Issue on “Process Modelling and Simulation”. Processes, 7(8), 511, https://doi.org/10.3390/pr7080511

De-Luca, R., Bano, G., Tomba, E., Bezzo, F., Barolo, M., 2020. Accelerating the Development and Transfer of Freeze-Drying Operations for the Manufacturing of Biopharmaceuticals by Model- Based Design of Experiments, Industrial and Engineering Chemistry Research, 59, 20071−20085

Dette, H., Bretz, F.,Pepelyshev, A.,Pinheiro, J.., 2008. Optimal designs for dose-finding studies, Journal of the American Statistical Association, 103 (483), pp. 1225-1237

Diedrichs, A., Gmehling, J. (2011). Solubility calculation of active pharmaceutical ingredients in alkanes, alcohols, water and their mixtures using various activity coefficient models. *Industrial and Engineering Chemistry Research*, *50*, 1757–1769. https://doi.org/10.1021/ie101373k

DiStefano, J. J.. 1981. Optimized blood sampling protocols and sequential design of kinetic experiments, American Journal of Physiology, 9, 259-265.

DiStefano, J. J.., 1982. Algorithms, software and sequential optimal sampling schedule design for pharmacokinetic and physiologic experiments, Mathematics and Computers in Simulation, 24, 531-534

Djuris, J., and Djuric, Z. (2017). Modeling in the quality by design environment: Regulatory requirements and recommendations for design space and control strategy appointment. *International Journal of Pharmaceutics*, *533*(2), 346–356.

Dong, Y., Yang, T., Xing, Y., Du, J., & Meng, Q. (2023). Data-Driven Modeling Methods and Techniques for Pharmaceutical Processes. *Processes*, *11*, 2096.

Dragalin, V., Fedorov, V.., 2006. Adaptive designs for dose-finding based on efficacy–toxicity response, Journal of Statistical Planning and Inference, 136,1800–1823

Duarte, B., Saraiva, P., Pantelides, C. 2004. Combined Mechanistic and Empirical Modelling. International Journal of Chemical Reactor Engineering, 2(1). https://doi.org/10.2202/1542-6580.1128.

Duffield, S., Da Vià, L., Bellman, A. C., Chiti, F., 2021. Automated High-Throughput Partition Coefficient Determination with Image Analysis for Rapid Reaction Workup Process Development and Modeling. Organic Process Research & Development, 25(12), 2738-2746

Ebden, Mark, 2015. Gaussian Processes: A Quick Introduction, arXiv. https://api.semanticscholar.org/CorpusID:31606842

Echtermeyer, A., Amar, Y., Zakrzewski, J.,l Lapkin, A. 2017. Self-optimisation and model-based design of experiments for developing a C-H activation flow process. *Beilstein Journal of Organic Chemistry*, 13, 150-163. https://doi.org/10.3762/bjoc.13.18

EFPIA, 2020. The Pharmaceutical Industry in Figures. Available at: chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://www.efpia.eu/media/554521/efpia_pharmafigures_2020_web.pdf. Last access: 27/08/2023

Elliott, J. R. and Lira, C. T., 2012. Introductory Chemical Engineering Thermodynamics, 2nd Ed., Prentice Hall, New York

Espie, D., Macchietto, S., 1989. The optimal design of dynamic experiments. AIChE Journal. 35 (2), 223–229.

Faber, K. and Kowalski, B.R. (1997), Propagation of measurement errors for the validation of predictions obtained by principal component regression and partial least squares. J. Chemometrics, 11: 181-238. https://doi.org/10.1002/(SICI)1099-128X(199705)11:3<181::AID-CEM459>3.0.CO;2-7

Facco, P., Dal Pastro, F., Meneghetti, N., Bezzo, F., Barolo, M. 2015. Bracketing the Design Space within the Knowledge Space in Pharmaceutical Product Development. Industrial & Engineering Chemistry Research, 54, 18, 5128–5138

FDA (2004b). Pharmaceutical CGMPs for the 21st century – A risk based approach. Final report. *U.S. Department of Health and Human Services*. *U.S. Food and Drug Administration*.

FDA, 2018. https://www.fda.gov/patients/drug-development-process/step-2-preclinical-research. Last access: 27/08/2023

Fedorov, V. V., Leonov., S.L.., 2001. Optimal design of dose response experiments: a modeloriented approach, Drug lnformation Journal, Vol. 35, pp. 1373-1383

Fisher, J. A., and Kalbaugh, C. A. (2011). Challenging assumptions about minority participation in US clinical research. *American Journal of Public Health*, *101*(12), 2217–2222.

Fisher, R.A., 1950. Contributions to Mathematical Statistics. John Wiley and Sons, New York.

Foracchia, M., Hooker, A., Vicini, P.,Ruggeri, 2004. POPED, a software for optimal experiment design in population kinetics, Computer Methods and Programs in Biomedicine, 74, 29—46

Franceschini, G., Macchietto, S., 2008a. Model-based design of experiments for parameter precision: State of the art, Chemical Engineering Science,63 (19), 4846-4872

Franceschini, G. and Macchietto, S., 2008b, Novel anticorrelation criteria for model-based experiment design: Theory and formulations. AIChE J., 54: 1009-1024.

Frank, A. G., Dalenogare, L. S., & Ayala, N. F. (2019). Industry 4.0 technologies: Implementation patterns in manufacturing companies. *International Journal of Production Economics*, *210*, 15–26.

Fredenslund, A.; Jones, R.L.; Prausnitz J.M. Group-contribution estimation of activity coefficients in nonideal liquid mixtures. *AIChE J.* 1975, *21*, 1086

Fukuda, I. M., Pinto, C. F. F., Moreira, C. D. S., Saviano, A. M., & Lourenço, F. R. (2018). Design of experiments (DoE) applied to pharmaceutical and analytical quality by design (QbD). *Brazilian Journal of Pharmaceutical Sciences*, *54*, e01006.

Galvanin, F., Ballan, C. C., Barolo, M., Bezzo, F. 2013. A general model-based design of experiments approach to achieve practical identifiability of pharmacokinetic and pharmacodynamic models. *Journal of Pharmacokinetics and Pharmacodynamics*, 40(4), 451–467. https://doi.org/10.1007/s10928-013-9321-5

Galvanin, F., Barolo, M., Bezzo, F. (2009). Online model-based redesign of experiments for parameter estimation in dynamic systems. *Industrial and Engineering Chemistry Research*, *48*(9), 4415–4427.

Galvanin, F., Cao, E., Al-Rifai, N., Gavriilidis, A., Dua, V. 2016. A joint model-based experimental design approach for the identification of kinetic models in continuous flow laboratory reactors. *Computers and Chemical Engineering*, 95, 202–215. https://doi.org/10.1016/j.compchemeng.2016.05.009

Galvanin, F., Macchietto, S., Bezzo, F., 2007. Model-based design of parallel experiments. Industrial and Engineering Chemistry Research, 46, 871–882. https://doi.org/10.1021/ie0611406

Gao, Z., Rohani, S., Gong, J., Wang, J. (2017). Recent Developments in the Crystallization Process: Toward the Pharmaceutical Industry. *Engineering*, *3*(3), 343–353. https://doi.org/10.1016/J.ENG.2017.03.022

Garud, S.S, Karimi, I.A., Kraft, M., 2017. Design of computer experiments: A review. Computers and Chemical Engineering, 106, 71-95. https://doi.org/10.1016/j.compchemeng.2017.05.010

Geladi,P., and Kowalski,B. R., 1986. Partial least-squares regression: a tutorial, Analytica Chimica Acta, 185,1-17

Geremia, M., Bano, G., Tomba, E., Barolo, M., Bezzo, F.., 2022. Practical use of primary drying models in an industrial environment with limited availability of equipment sensors, International Journal of Pharmaceutics, 619, 121699

Geremia, M., Diab, S., Christodoulou, C., Bano, G., Barolo, M., Bezzo, F.., 2023. A general procedure for the evaluation of the prediction fidelity of pharmaceutical systems models, Chemical Engineering Science 280, 118972

Getreskilled, 2022. https://www.getreskilled.com/types-of-pharma-jobs/. Last accessed: 27/08/2023

Gharagheizi, F., Eslamimanesh, A., Mohammadi, A. H., Richon, D. (2011). Representation/Prediction of solubilities of pure compounds in water using artificial neural network-group contribution method. *Journal of Chemical and Engineering Data*, *56*, 720–726. https://doi.org/10.1021/je101061t

Gill, P.E., W. Murray, M.A. Saunders, and M.H. Wright (1984). Procedures for Optimization Problems with a Mixture of Bounds and General Linear Constraints. *CM Trans. Math. Software*, 10, 282-298.

Global Regulatory Authority Websites, 2023. Available at: https://www.pda.org/scientific-and-regulatory-affairs/regulatory-resources/global-regulatory-authority-websites#europe

Gmehling, J. G., Anderson, T. F., Prausnitz, J. M., 1978. Solid-Liquid Equilibria Using UNIFAC. Industrial & Engineering Chemistry Fundamentals, 17, 269-273. https://doi.org/10.1021/i160068a008

González Peña OI, López Zavala MÁ, Cabral Ruelas H. Pharmaceuticals Market, Consumption Trends and Disease Incidence Are Not Driving the Pharmaceutical Research on Water and Wastewater. Int J Environ Res Public Health. 2021 Mar 4;18(5):2532. doi: 10.3390/ijerph18052532. PMID: 33806343; PMCID: PMC7967517.

Grah, A. (2004). Entwicklung und anwendung modularer software zur Sim- ulation und Parameterschaetzung in gaskatalytischen Festbettreaktoren. *Ph.D. thesis*, Martin Luther University Halle-Wittenberg.

Granato, D., de Araújo Calado, V.M., 2013. The use and importance of design of experiments (DOE) in process modelling in food science and technology. In: D. Granato, D. and Ares, G. (Eds), Mathematical and Statistical Methods in Food Science and Technology, John Wiley & Sons, Ltd. pp 1-18. https://doi.org/10.1002/9781118434635.ch1

Grangeia, H. B., Silva, C., Simões, S. P., & Reis, M. S. (2020). Quality by design in pharmaceutical manufacturing: A systematic review of current status, challenges and future perspectives. *European Journal of Pharmaceutics and Biopharmaceutics*, *147*, 19–37.

Heller, A. A., Lockwood, S. Y., Janes, T. M., & Spence, D. M. (2018). Technologies for Measuring Pharmacokinetic Profiles. *Annual Review of Analytical Chemistry*, *11*, 79–100.

Hone, C. A., Holmes, N., Akien, G. R., Bourne, R. A., Muller, F. L. (2017). Rapid multistep kinetic model generation from transient flow data. *React. Chem. Eng.*, *2*, 103–108.

Hooker, A. C., Karlsson, M. O., 2009. Optimisation of the intravenous glucose tolerance test in T2DM patients using optimal experimental design, Journal of Pharmacokinetics and Pharmacodynamics. 36, 281–295

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *ournal of Educational Psychology*, 24, 417-441.

Hurtado, P., Ordóñez, S., Sastre, H., Dıez, F. V., Appl. Catal. B 2004, 51, 229–238

https://doepy.readthedocs.io/en/latest/

https://www.scientificlabs.ie/description/DOXAZOSIN%20RELATED%20COMPOUND%20C%20UNITED%20STA#overview

Hu, S., Han, J., Liu, H., Qiu, J., Zhao, Y., Guo, Y., Huang, H., He, H., Wang, P., 2021. Solubility Behavior and Polymorphism of l-Arginine l-Pyroglutamate in Nine Pure Solvents and a Binary Water + Ethanol System. Journal of Chemical & Engineering Data, 66,6, 2383-2390

Huang, X., Wang, J., Hao, H., Ouyang, J., Gao, Y., Bao, Y., Wang, Y., & Yin, Q. (2015). Determination and correlation of solubility and solution thermodynamics of coumarin in different pure solvents. Fluid Phase Equilibria, 394, 148–155. https://doi.org/10.1016/j.fluid.2015.03.022

Hunter,W.G.,Reiner,A.M.,1965.Designs for discriminating between two rival models. Technometrics, 7, 307–323.

ICH Points to Consider (R2), 2011. ICH-Endorsed Guide for ICH Q8/Q9/Q10 Implementation

Ierusalimschy, R, de Figueiredo, L.H., Celes, W.., 2005.The Implementation of Lua 5.0. Journal of Universal Computer Science, 11 (7), 1159-1176

Imarc, 2022. https://www.imarcgroup.com/pharmerging-market

Indeed, August 2023. https://uk.indeed.com/career-advice/finding-a-job/pharmaceutical-sector#:~:text=The%20pharmaceutical%20sector%20is%20an,biopharmaceuticals%20and%20sales%20and%20marketing. Last access: 26/08/2023

Inês Silva, M., Khadra, I., Pyper, K., & Halbert, G. W. (2022). Small scale in vitro method to determine a potential bioequivalent equilibrium solubility range for fed human intestinal fluid. *European Journal of Pharmaceutics and Biopharmaceutics*, *177*, 126–134. https://doi.org/10.1016/j.ejpb.2022.06.005

Insights10, 2022. https://www.insights10.com/report/us-pharmaceutical-market-analysis/. Last access: 27/08/2023

Jamei, M., Turner, D., Yang, J., Neuhoff, S., Polak, S., Rostami-Hodjegan, A., & Tucker, G. (2009). Population-based mechanistic prediction of oral drug absorption. *AAPS Journal*, *11*(2), 225–237. https://doi.org/10.1208/s12248-009-9099-y

Jamei,M., Steve Marciniak , Kairui Feng , Adrian Barnett , Geoffrey Tucker & Amin Rostami-Hodjegan, 2009. The Simcyp ® Population-based ADME Simulator. Expert Opinion on Drug Metabolism & Toxicology, 5 (2), 211 - 223

Joshi SR, Fernando D, Igwe S, McKenzie L, Krishnatry AS, Halliday F, Zhan J, Greene TJ, Xu J, Ferron-Brady G, Lataillade M, Min S. Phase I evaluation of the safety, tolerability, and pharmacokinetics of GSK3640254, a next-generation HIV-1 maturation inhibitor. Pharmacol Res Perspect. 2020 Dec;8(6):e00671. doi: 10.1002/prp2.671. PMID: 33200887; PMCID: PMC7670640.

Jung JY, Choi Y, Suh CH, Yoon D, Kim HA. Effect of fenofibrate on uric acid level in patients with gout. Scientific Reports. 2018;8(1):16767. doi: 10.1038/s41598-018-35175-z.

Juran, J.M., De Feo, J.A. Juran's Quality Handbook: The Complete Guide to Performance Excellence, sixth ed., McGraw Hill, New York, 2010.

Kalicka, R. and Bochen, D.., 2006. A new OSS design based on parameter sensitivity to changes in measurements, Control and Cybernetics, 35,No. 2

Kawakita, K.; Lüdde, K.-H. Some considerations on powder compression equations. Powder Technol. 1971, 2, 61–68.

Khajir, S., Shayanfar, A., Martinez, F., Rahimpour, E., Jouyban, A. (2024). Nicotinamide Solubility in Ethanol + Acetonitrile at Different Temperatures. *Physical Chemistry Research*, *12*(1), 33–45. https://doi.org/10.22036/pcr.2023.368643.2233

Kiefer, J. and Wolfowitz, J. 1960. The equivalence of two extremum problems. Canadian Journal of Mathematics, 12, 363–366.

Kiefer, J. and Wolfowitz, J., 1959. Optimum designs in regression problems. Annals of Mathematical Statistics, 30, 271–294. https://doi.org/10.1214/AOMS/1177706252

Kikuta, J. and Kitamori, N. (1994). Effect of mixing time on the lubricating properties of magnesium stearate and the final characteristics of the compressed tablets. *Drug Development and Industrial Pharmacy*, 20, 343-355.

Killeen, P. R. 2005. An Alternative to Null-Hypothesis Significance Tests. Psychological Science, 16(5), 345–353. https://doi.org/10.1111/j.0956-7976.2005.01538.x

Kiriiri, G. K., Njogu, P. M., & Mwangi, A. N. (2020). Exploring different approaches to improve the success of drug discovery and development projects: a review. *Future Journal of Pharmaceutical Sciences*, *6*(27).

Kroll, P., Hofer, A., Ulonska, S., Kager, J., Herwig, C. (2017). Model-Based Methods in the Biopharmaceutical Process Lifecycle. *Pharmaceutical Research*, *34*(12), 2596–2613.

Kushner, J., Moore, F. (2010). Scale-up model describing the impact of lubrication on tablet tensile strength. *International Journal Pharmaceutics*, 399, 19–30.

Lemmer, H. J.R. and Liebenberg, W. ,. 'Crystallization: Its Mechanisms and Pharmaceutical Applications', Crystal Growth and Chirality - Technologies and Applications. IntechOpen, May 17, 2023. doi: 10.5772/intechopen.105056.

Lev Krasnov, Simon Mikhaylov, Maxim V. Fedorov, Sergey Sosnin. (2022). BigSolDB: solubility dataset of compounds in organic solvents and water in a wide range of temperatures [Data set]. Zenodo. https://doi.org/10.5281/zenodo.6809669

Li, Y., Li, X., & Wu, K. (2020). Measurement and Modeling of the Solubility of N,N-Dibenzylhydroxylamine in 17 Solvents from T = 273.15 to 323.35 K and Thermodynamic Properties of Solution. Journal of Chemical and Engineering Data, 65(2), 828–840. https://doi.org/10.1021/acs.jced.9b01028

Lin, J.,Wang, Q., Zhou, S., Xu, S., Yao, K. 2022. Tetramethylpyrazine: A review on its mechanisms and functions, Biomedicine & Pharmacotherapy,150,113005

Lipinski, C.A., 2000. Drug-like properties and the causes of poor solubility and poor permeability. J. Pharmacol. Toxicol. Met. 44, 235–249

Lipinski, C.A., et al., 1997. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv. Drug Del. 23, 3–25.

Lipsky, M.S., Sharp, L.K., 2001. From idea to market: The drug approval process. J. Am. Board Fam. Pract. 14, 362–367.

López C., D. C., Barz, T., Körkel, S., Wozny, G. (2015). Nonlinear ill-posed problem analysis in model-based parameter estimation and experimental design. *Computers and Chemical Engineering*, 77, 24–42.

Lu, Y., Kim, S., & Park, K. (2011). In vitro-in vivo correlation: Perspectives on model development. *International Journal of Pharmaceutics*, *418*(1), 142–148.

Maghsoodi, M. (2015). Role of solvents in improvement of dissolution rate of drugs: Crystal habit and crystal agglomeration. *Advanced Pharmaceutical Bulletin*, *5*(1), 13–18. https://doi.org/10.5681/apb.2015.002

Mani S., Tabil L. G. and Sokhansanj S. (2004). Evaluation of compaction equations applied to four biomass species. Can. Biosyst. Eng., 46, 355-361.

Mardia, K.V., Kent, J.T. and Bibby, J.M. (1979). *Multivariate analysis*. Academic Press Limited, London (U.K.).

Marino M, Jamal Z, Zito PM. Pharmacodynamics. 2023 Jan 29. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; PMID: 29939568.

Martinez, E., Cristaldi, M., Grau, R., 2009. Design of Dynamic Experiments in Modeling for Optimisation of Batch Processes, Industrial & Engineering Chemistry Research. 48, 3453–3465

McKinsey & Company, 2021. Smart quality: Reimagining the way quality works. Available at: https://www.mckinsey.com/industries/life-sciences/our-insights/sm art-quality-reimagining-the-way-quality-works.

McLean, K. A. P., & McAuley, K. B. (2012). Mathematical modelling of chemical processes-obtaining the best model predictions and parameter estimates using identifiability and estimability procedures. *Canadian Journal of Chemical Engineering*, *90*(2), 351–366.

McMullen J. P. and Jensen, K. F.. 2011. Org. Process Res. Dev. , 15, 398–407.

Mennen, S. M., Alhambra, C., Allen, C. L., Barberis, M., Berritt, S., Brandt, T. A., Campbell, A. D., Castañón, J., Cherney, A. H., Christensen, M., Damon, D. B., Eugenio De Diego, J., García-Cerrada, S., García-Losada, P., Haro, R., Janey, J., Leitch, D. C., Li, L., Liu, F., … Zajac, M. A. (2019). The Evolution of High-Throughput Experimentation in Pharmaceutical Development and Perspectives on the Future. *Organic Process Research and Development*, *23*, 1213–1242.

Mihaluta, M., Martin, P., Dantan, J. Y. 2008. Manufacturing Process Modeling and Simulation. ICME, Italy. 7p. ffhal-00999479f. https://hal.archives-ouvertes.fr/hal-00999479

Miller, J. , Rodríguez-Hornedo, N., Blackburn, A., Macikenas, D., Collman, B.. (2007). Solvent Systems for Crystallization and Polymorph Selection.In book: Solvent Systems and Their Selection in Pharmaceutics and Biopharmaceutics 10.1007/978-0-387-69154-1_3.

Montgomery, D., 2013. Design and Analysis of Experiments. John Wiley & Sons, Hoboken, NJ.

Mori, F.; DiStefano, J. 1979. Optimal Nonunifom Sampling Interval and Test-Input Design for Identification of Physiological Systems from Very Limited Data, IEEE Transactions on Automatic Control, 24(6), 893–900.

Mortier, S. T. F. C., De Beer, T., Gernaey, K. V., Remon, J. P., Vervaet, C., & Nopens, I. (2011). Mechanistic modelling of fluidized bed drying processes of wet porous granules: A review. European Journal of Pharmaceutics and Biopharmaceutics, 79(2), 205–225. https://doi.org/10.1016/j.ejpb.2011.05.013

Muhieddine, M. H., Viswanath, S. K., Armstrong, A., Galindo, A., Adjiman, C. S. (2022). Model-based solvent selection for the synthesis and crystallisation of pharmaceutical compounds. *Chemical Engineering Science*, *264*, 118125. https://doi.org/10.1016/j.ces.2022.118125

Musa MA, Cooperwood JS, Khan MO. A review of coumarin derivatives in pharmacotherapy of breast cancer. Current Medicinal Chemistry , 2008;15(26):2664-79. doi: 10.2174/092986708786242877.

Nassar, J., Williams, B., Davies, C., Lief, K., Elkes, R. (2021). Lubrication empirical model to predict tensile strength of directly compressed powder blends. *International Journal Pharmaceutics*, 592, 119980.

Nyberg, J., Karlsson, M. O., Hooker, A. C.., 2009. Simultaneous optimal experimental design on dose and sample times, Journal of Pharmacokinetics and Pharmacodynamics,36, 125–145

Ogungbenro, K., Dokoumetzidis, A., Aarons, L.., 2009. Application of optimal design methodologies in clinical pharmacology experiments, Pharmaceutical Statistics, 8, 239–252

Orehek, J., Teslić, D., & Likozar, B. (2021). Continuous Crystallization Processes in Pharmaceutical Manufacturing: A Review. *Organic Process Research and Development*, *25*, 16–42.

Pankajakshan, A., Bawa, S. G., Gavriilidis, A., & Galvanin, F. (2023). Autonomous kinetic model identification using optimal experimental design and retrospective data analysis: methane complete oxidation as a case study. *Reaction Chemistry & Engineering*. DOI: 10.1039/d3re00156c

Panteli D and Edwards S. Ensuring access to medicines: How to stimulate innovation to meet patients' needs? [Internet]. Richardson E, Palm W, Mossialos E, editors. Copenhagen (Denmark): European Observatory on Health Systems and Policies; 2018. PMID: 30272894.

Papadakis, E., Tula, A. K., Gani, R. (2016). Solvent selection methodology for pharmaceutical processes: Solvent swap. *Chemical Engineering Research and Design*, *115*, 443–461. https://doi.org/10.1016/j.cherd.2016.09.004

Papadakis, E., Woodley, J.M., Gani, R., 2018. Perspective on PSE in pharmaceutical process development and innovation. In: Computer Aided Chemical Engineering, volume 41, pages 597–656. Elsevier

Perrier, J., Zhou, Z., Dunn, C., Khadra, I., Wilson, C. G., & Halbert, G. (2018). Statistical investigation of the full concentration range of fasted and fed simulated intestinal fluid on the equilibrium solubility of oral drugs. *European Journal of Pharmaceutical Sciences*, *111*(June 2017), 247–256. https://doi.org/10.1016/j.ejps.2017.10.007

Petersen, B., Gernaey, K., & Vanrolleghem, P. A. (2001). Practical identifiability of model parameters by combined respirometric-titrimetric measurements. *Water Science and Technology*, *43*(7), 347–355. https://doi.org/10.2166/wst.2001.0444

Peters, R., 2019. "Bio/Pharma Needs Ideas and Incentives to Advance Manufacturing, "P*harmaceutical Technology* 43 (12) 2019

Petsagkourakis, P., & Galvanin, F. (2021). Safe model-based design of experiments using Gaussian processes. *Computers and Chemical Engineering*, *151*, 107339.

Pitt, K. G., Newton, J. M., Stanley, P. (1988). Tensile fracture of doubly-convex cylindrical discs under diametral loading. *Journal of Materials Science*, 23(8), 2723–2728.

Podczeck, F., Miah, Y., (1994). The influence of particle size and shape on the angle of internal friction and the flow factor of unlubricated and lubricated powders. *International Journal Pharmaceutics*, 144, 187- 194

Prescott, L. F.. Paracetamol: Past, Present, and Future. American Journal of Therapeutics 7(2):p 143-148, March 2000.

Prus, M. 2019. Various optimality criteria for the prediction of individual response curves. Statistics and Probability Letters, 146, 36-41. https://doi.org/10.1016/j.spl.2018.10.022

Pukelsheim, F. (1993). Optimal Design of Experiments. J. Wiley & Sons, New York (U.S.A.).

Pyper, K., Brouwers, J., Augustijns, P., Khadra, I., Dunn, C., Wilson, C. G., & Halbert, G. W. (2020). Multidimensional analysis of human intestinal fluid composition. *European Journal of Pharmaceutics and Biopharmaceutics*, *153*(January), 226–240. https://doi.org/10.1016/j.ejpb.2020.06.011

Quaglio, M., Fraga, E. S., Cao, E., Gavriilidis, A., Galvanin, F. 2018. A model-based data mining approach for determining the domain of validity of approximated models, Chemometrics and Intelligent Laboratory Systems, 172, 58-67. https://doi.org/10.1016/j.chemolab.2017.11.010

Rabbie, S. C., Flanagan, T., Martin, P. D., & Basit, A. W. (2015). Inter-subject variability in intestinal drug solubility. *International Journal of Pharmaceutics*, *485*(1–2), 229–234. https://doi.org/10.1016/j.ijpharm.2015.03.006

Rasmussen, C. , Williams, C. , 2006. Gaussian processes for machine learning. MIT Press, Cambridge, MA, USA

Reichert, I., Olney, P., Lahmer, T. 2021. Combined approach for optimal sensor placement and experimental verification in the context of tower-like structures. *Journal of Civil Structural Health Monitoring*, 11(1), 223–234. https://doi.org/10.1007/s13349-020-00448-7

Riethorst, D.; Mols, R.; Duchateau, G.; Tack, J.; Brouwers, J.; Augustijns, P. Characterization of Human Duodenal Fluids in Fasted and Fed State Conditions. J. Pharm. Sci. 2016, 105, 673–681.

Reizman, B. J., Jensen, K. F. (2016). Feedback in Flow for Accelerated Reaction Development. *Accounts of Chemical Research*, *49*(9), 1786–1796.

Reizman, B. J., Wang, Y. M., Buchwald, S. L., & Jensen, K. F. (2016). Suzuki-Miyaura cross-coupling optimization enabled by automated feedback. Reaction Chemistry and Engineering, 1(6), 658–666.

Rosenberger, Julian., James Butler, Jennifer Dressman,2018. A Refined Developability Classification System, Journal of Pharmaceutical Sciences,Volume 107, Issue 8,Pages 2020-2032

Ruether F, Sadowski G. Modeling the solubility of pharmaceuticals in pure solvents and solvent mixtures for drug process design. J Pharm Sci. 2009 Nov;98(11):4205-15. doi: 10.1002/jps.21725. PMID: 19283772.

Sabir, A., Evans, B., Jain, S. (2001). Formulation and process optimization to eliminate picking from market image tablets. *International Journal of Pharmaceutics,* 215, 123-135.

Salehi, N., Kuminek, G., Al-Gousous, J., Sperry, D. C., Greenwood, D. E., Waltz, N. M., Amidon, G. L., Ziff, R. M., & Amidon, G. E. (2021). Improving Dissolution Behavior and Oral Absorption of Drugs with pH-Dependent Solubility Using pH Modifiers: A Physiologically Realistic Mass Transport Analysis. *Molecular Pharmaceutics*, *18*(9), 3326–3341. https://doi.org/10.1021/acs.molpharmaceut.1c00262

Sansana, J., Joswiak, M. N., Castillo, I., Wang, Z., Rendall, R., Chiang, L. H., & Reis, M. S. (2021). Recent trends on hybrid modeling for Industry 4.0. *Computers and Chemical Engineering*, *151*, 107365.

Šantl, M., Ilić, I., Vrečer, F., Baumgartner, S. (2011). A compressibility and compactibility study of real tableting mixtures: The impact of wet and dry granulation versus a direct tableting mixture. *International Journal of Pharmaceutics*, 414(1–2), 131–139.

Sato, A., Miyao, T., Jasial, S. *et al.* Comparing predictive ability of QSAR/QSPR models using 2D and 3D molecular representations. *J Comput Aided Mol Des* 35, 179–193 (2021).

Schittkowski, K., 2008. Parameter identification and model verification in systems of partial differential equations applied to transdermal drug delivery, Mathematics and Computers in Simulation, 79, 521–538

Shahmohammadi, A., McAuley, K. B. (2019). Sequential model-based A- and V-optimal design of experiments for building fundamental models of pharmaceutical production processes. *Computers and Chemical Engineering*, 129, 106504.

Shahmohammadi, A., McAuley, K. B. (2020). Using prior parameter knowledge in model-based design of experiments for pharmaceutical production. *AIChE Journal*, 66(11), 1–20.

Sheskey, P. J., Robb, R. T., Moore, R. D., and Boyce, B. M. (1995). Effects of lubricant level, method of mixing, and duration of mixing on a controlled-release matrix tablet containing hydroxypropyl methylcellulose. *Drug development and industrial pharmacy*, 21(19), 2151– 2165.

Shivva, V., Korell, J., Tucker, I. G., & Duffull, S. B. (2013). An approach for identifiability of population pharmacokinetic- pharmacodynamic models. CPT: Pharmacometrics and Systems Pharmacology, 2(6).

Silber, H. E., Nyberg, J.

Silva, M. I., Khadra, I., Pyper, K., & Halbert, G. W. (2023). Fed intestinal solubility limits and distributions applied to the Developability classification system. *European Journal of Pharmaceutics and Biopharmaceutics*, *186*(January), 74–84. https://doi.org/10.1016/j.ejpb.2023.03.005

Singh, B., Kumar, R., & Ahuja, N. (2005). Optimizing drug delivery systems using systematic "design of experiments." Part I: Fundamental aspects. *Critical Reviews in Therapeutic Drug Carrier Systems*, *22*(1), 27–105.

Singh, B., Dahiya, M., Saharan, V., & Ahuja, N. (2005). Optimizing drug delivery systems using systematic "design of experiments." Part II: Retrospect and prospects. *Critical Reviews in Therapeutic Drug Carrier Systems*, *22*(3), 215–293.

Sivaraman, A. and Banga, A. (2015). Quality by design approaches for topical dermatological dosage forms. *Research and Reports in Transdermal Drug Delivery*, *4*, 9–21.

Smith, K., 1918. On the Standard Deviations of Adjusted and Interpolated Values of an Observed Polynomial Function and its Constants and the Guidance they give Towards a Proper Choice of the Distribution of Observations. *Biometrika*, 12 (1/2) 1-85.

Specchia, S., Conti, F., Specchia, V. Ind. Eng. Chem. Res. 2010, 49, 11101–11111.

Spinner, C.D.; Felizarta, F.; Rizzardini, G.; Philibert, P.; Mitha, E.; Domingo, P.; Stephan, C.J.; DeGrosky, M.; Bainbridge, V.; Zhan, J.; et al. Phase Iia Proof-of-Concept Evaluation of the Antiviral Efficacy, Safety, Tolerability, and Pharmacokinetics of the Next-Generation Maturation Inhibitor Gsk3640254. Clin. Infect. Dis. 2022, 75, 786–794.

Stamatopoulos, K. (2022). Integrating Biopharmaceutics to Predict Oral Absorption Using PBPK Modelling. In Biopharmaceutics, H. Batchelor (Ed.). https://doi.org/10.1002/9781119678366.ch12

Stasevych, M., and Zvarych, V. (2023). Innovative Robotic Technologies and Artificial Intelligence in Pharmacy and Medicine : Paving the Way for the Future of Health Care — A Review patient. *Big Data and Cognitive Computing*, *7*, 147.

Statista, 2022. https://www.statista.com/topics/5056/pharmaceutical-industry-in-the-uk/#topicOverview. Last access: 27/08/2023

Statista, 2023. https://www.statista.com/statistics/263102/pharmaceutical-market-worldwide-revenue-since-2001/#:~:text=The%20global%20pharmaceutical%20market%20has,at%201.42%20trillion%20U.S.%20dollars.

Stigler, S. M., 1971. Optimal Experimental Design for Polynomial Regression. *Journal of the American Statistical Association*, 66 (334), 311-318. https://doi.org/10.2307/2283928

Svejstrup TD, Ruffoni A, Juliá F, Aubert VM, Leonori D. Synthesis of Arylamines via Aminium Radicals. Angewandte Chemie International Edition, 2017;56(47):14948-14952. doi: 10.1002/anie.201708693.

Sverdlov, O., Ryeznik, Y., Wong, W. K.., 2020. On Optimal Designs for Clinical Trials: An Updated Review, Journal of Statistical Theory and Practice, 14, 10

Takács-Novák, K., Vera Szőke, Gergely Völgyi, Péter Horváth, Rita Ambrus, Piroska Szabó-Révész, 2013. Biorelevant solubility of poorly soluble drugs: Rivaroxaban, furosemide, papaverine and niflumic acid, Journal of Pharmaceutical and Biomedical Analysis, 83,279-285

Tamargo, J., Kaski, J. C., Kimura, T., Barton, J. C., Yamamoto, K., Komiyama, M., Drexel, H., Lewis, B. S., Agewall, S., & Hasegawa, K. (2022). Racial and ethnic differences in pharmacotherapy to prevent coronary artery disease and thrombotic events. *European Heart Journal - Cardiovascular Pharmacotherapy*, *8*, 738–751.

Taylor, C. J., Booth, M., Manson, J. A., Willis, M. J., Clemens, G., Taylor, B. A., Chamberlain, T. W., & Bourne, R. A. (2021). Rapid, automated determination of reaction models and kinetic parameters. *Chemical Engineering Journal*, *413*, 127017.

Thompson, D. E., McAuley, K. B., McLellan, P. J. (2009). Parameter estimation in a simplified MWD model for HDPE produced by a ziegler-natta catalyst. *Macromolecular Reaction Engineering*, 3(4), 160–177.

Tsamandouras, N., Rostami-Hodjegan, A., & Aarons, L. (2015). Combining the "bottom up" and "top down" approaches in pharmacokinetic modelling: Fitting PBPK models to observed clinical data. *British Journal of Clinical Pharmacology*, *79*(1), 48–55.

Uchimoto, T., Iwao, Y., Yamamoto, T., Sawaguchi, K., Moriuchi, T., Noguchi, S., Itai, S. (2013). Newly developed surface modification punches treated with alloying techniques reduce sticking during the manufacture of ibuprofen tablets. *International Journal Pharmaceutics*, 441(1-2), 128-34.

Valle, S., Li, W., Qin, S. J., 1999. Selection of the Number of Principal Components: The Variance of the Reconstruction Error Criterion with a Comparison to Other Methods.Industrial & Engineering Chemistry Research, 38 (11), 4389-4401

Violet, L., Loubière, K., Rabion, A., Samuel, R., Hattou, S., Cabassud, M., Prat, L. (2016). Stoichio-kinetic model discrimination and parameter identification in continuous microreactors. *Chemical Engineering Research and Design*, 114, 39–51.

Vo, A. D. D., Shahmohammadi, A., McAuley, K. B. 2021. Model-based design of experiments for polyether production from bio-based 1,3-propanediol. *AIChE Journal*, 67, e17394. https://doi.org/10.1002/aic.17394

Waldron, C., Pankajakshan, A., Quaglio, M., Cao, E., Galvanin, F., & Gavriilidis, A. (2020). Model-based design of transient flow experiments for the identification of kinetic parameters. *Reaction Chemistry and Engineering*, *5*(1), 112–123. https://doi.org/10.1039/c9re00342h

Waldron, C., Pankajakshan, A., Quaglio, M., Cao, E., Galvanin, F., Gavriilidis, A. 2019. Closed-Loop Model-Based Design of Experiments for Kinetic Model Discrimination

and Parameter Estimation: Benzoic Acid Esterification on a Heterogeneous Catalyst. *Industrial and Engineering Chemistry Research*, 22165–22177. https://doi.org/10.1021/acs.iecr.9b04089

Wang, J., Wen, H., Desai, D. (2010). Lubrication in tablet formulations. *European Journal of Pharmaceutics and Biopharmaceutics*, 75(1), 1–15.

Wang, M., Risuleo, R.S., Jacobsen, E.W., Chotteau, V., Hjalmarsson, H., 2020. Identification of nonlinear kinetics of macroscopic bio-reactions using multilinear gaussian processes. Computers & Chemical Engineering 133, 106671.

Wang, R., Zou, Y., Guo, J., Pu, Y., & Wang, D. (2021). Solubility and Solubility Modeling of 1,3,5-Tris(1-phenyl-1 H-benzimidazol-2-yl)benzene toward Nanodispersions in Organic Solvents. Journal of Chemical and Engineering Data, 66(6), 2568–2575. https://doi.org/10.1021/acs.jced.1c00163

Wang, Y., Li, B., Jiang, C., Fang, Y., Bai, P., Wang, Y., 2021. Study on Electron Transport Characterization in TPBi Thin Films and OLED Application., The Journal of Physical Chemistry C, 125, 16753 - 16758

Weissman, S. A., & Anderson, N. G. (2015). Design of Experiments (DoE) and Process Optimisation. A Review of Recent Publications. *Organic Process Research and Development*, *19*(11), 1605–1633.

Wold, S., Martens, H. and H. Wold (1983). The multivariate calibration problem in chemistry solved by the PLS method. *Lecture Notes in Math.*, 973, 286-293.

Wold, Svante. 1978. Cross-Validatory Estimation of the Number of Components in Factor and Principal Components Models. *Technometrics*, 20 (4), , 397–405. *JSTOR*,

Wong, W. K., 1995. On the equivalence of D and G-optimal designs in heteroscedastic models. Statistics and Probability Letters, 25, 317-321.

Woodcock, J. (2004). The concept of pharmaceutical quality. *American Pharmaceutical Review*, *7*, 10–15.

Wright V. A review of benorylate - a new antirheumatic drug. Scandinavian Journal of Rheumatology, 1975;13:5-8.

Wu, Y., Ren, M., & Zhang, X. (2020). Solubility Determination and Model Correlation of Benorilate between T = 278.18 and 318.15 K. Journal of Chemical and Engineering Data, 65(7), 3690–3695. https://doi.org/10.1021/acs.jced.0c00301

www.efpia.eu/more-than-medicine. Las access: 27/08/2023

Yamamura, T., Ohta T, Taira T, Ogawa Y, Sakai Y, Moribe K, Yamamoto K. (2009). Effects of automated external lubrication on tablet properties and the stability of eprazinone hydrochloride. *International Journal Pharmaceutics*, 370(1-2), 1-7.

Ye, Z., Ouyang, D. (2021). Prediction of small-molecule compound solubility in organic solvents by machine learning algorithms. *Journal of Cheminformatics*, *13*, 98. https://doi.org/10.1186/s13321-021-00575-3

Yu, L. X. (2008). Pharmaceutical quality by design: Product and process development, understanding, and control. *Pharmaceutical Research*, *25*(4), 781–791.

Yu, L.X., Kopcha, M., 2017. The future of pharmaceutical quality and the path to get there. Int. J. Pharm. 528, 354–359

Yuan, Y., He, Q., Zhang, S., Li, M., Tang, Z., Zhu, X., Jiao, Z., Cai, W., & Xiang, X. (2022). Application of Physiologically Based Pharmacokinetic Modeling in Preclinical Studies: A Feasible Strategy to Practice the Principles of 3Rs. *Frontiers in Pharmacology*, *13*(May), 1–18. https://doi.org/10.3389/fphar.2022.895556

Zhang, N., Li, S., Yang, H., Li, M., Yang, Y., & Tang, W. (2019). Measurement and Correlation of the Solubility of Tetramethylpyrazine in Nine Monosolvents and Two Binary Solvent Systems. Journal of Chemical and Engineering Data, 64(3), 995–1006. https://doi.org/10.1021/acs.jced.8b00888

Zhou, G., Chen, K., Yang, Z., Shao, D., & Fan, H. (2021). Research on the 2-Chloro-4-amino-6,7-dimethoxyquinazoline Solubility in 12 Monosolvents at Various Temperatures: Experimental Measurement and Thermodynamic Correlation. Journal of Chemical and Engineering Data, 66(1), 170–177. https://doi.org/10.1021/acs.jced.0c00506

Zhou, Z., Dunn, C., Khadra, I., Wilson, C. G., & Halbert, G. W. (2017). Statistical investigation of simulated fed intestinal media composition on the equilibrium solubility of oral drugs. *European Journal of Pharmaceutical Sciences*, *99*, 95–104. https://doi.org/10.1016/j.ejps.2016.12.008

Zhuang, X., & Lu, C. (2016). PBPK modeling and simulation in drug research and development. *Acta Pharmaceutica Sinica B*, *6*(5), 430–440. https://doi.org/10.1016/j.apsb.2016.04.004

Zullo, L. 1991. Computer aided design of experiments. An engineering approach. Ph.D. Thesis, The University of London, U.K.

# Acknowledgments

Coming to the end of this journey, there are so many people I would like to thank. First of all, I thank my Supervisor, Prof. Pierantonio Facco, for his careful guidance and for all the technical advices that helped me improve throughout the entire PhD. Thank you for believing in my potential and for giving me this amazing opportunity that will impact on my professional and personal growth in upcoming years. Thank you also for your constant commitment to making the workplace a positive and cohesive environment for my colleagues and me. Moreover, I would like to thank Prof. Massimiliano Barolo e Fabrizio Bezzo for all the suggestions and contributions to the scientific work carried out and for all the meetings where I could learn a lot from their experience and expertise.
I would like to thank Dr. Simeone Zomer for giving me the opportunity to collaborate with amazing GSK teams on stimulating and impactful projects and for allowing me to remain part of the team. Many thanks to Dr Gabriele Bano, Dr Charalampos Christodoulou, Dr. Yuliya Vueva, Dr. Samir Diab, Dr. Paola Ferrini, Dr. Konstantinos Stamatopoulos and Ms Katy Harabajiu for their precious contribution to the scientific work carried out. Looking forward to starting new challenging projects with you.
Furthermore, I would like to express my gratitude to Prof. Federico Galvanin, who made my experience at UCL memorable. Both his remarkable competence and his positive and encouraging attitude allowed me to persevere besides difficulties, to solve technical doubts, to be inspired and to enjoy my work even more. I would like to thank also Prof. Asterios Gavriilidis, Dr. Arun Pankajakshan and Mr. Solomon Gajere Bawa for helping me improve the quality of my research.
I would like to thank all the amazing professionals and friends I've had the pleasure of meeting at CAPE-Lab, GSK and UCL during this journey.

Vorrei ringraziare i miei genitori, Sergio e Marta, per il supporto, l'affetto e la comprensione che hanno sempre dimostrato e per essere un esempio di vita che mi ispira ogni giorno. Un ringraziamento speciale anche a mio fratello, Giovanni, per essere sempre presente, incoraggiante e affettuoso anche nei momenti più difficili. Un ringraziamento ai miei nonni, Antonio, Assunta, Francesco e Maria, per avermi insegnato il senso del sacrificio e per avermi supportata in ogni mia scelta. Un ringraziamento ai miei zii, Jessica, Mauro e Renata, e cugini, Antea, Elisa, Gianmarco, Irene e Mirco, per tutti i momenti speciali che abbiamo condiviso e per tutti quelli che verranno. Inoltre, vorrei ringraziare Giovanni, Sara, Beatrice e tutta la famiglia Guiotto-Bettiati per avermi accolta sin dal primo giorno con caloroso affetto e per essere diventati parte integrante della mia famiglia.

Un ringraziamento a tutti amici che mi hanno accompagnata in questo percorso. In particolare, ringrazio Daniel per avermi regalato molti sorrisi con la sua simpatia e gentilezza; ringrazio Elia per i confronti interessanti, per la disponibilità e l'aiuto disinteressato; ringrazio Alberto per essere un buon amico e per avermi presentato la persona che mi ha cambiato la vita; ringrazio Samir per essere non solo un fantastico *tutor aziendale*, sempre preciso, attento e disponibile, ma anche un amico sincero; I thank Theresa and Monica for the beautiful friendship that made my experience in

London even more special. Grazie a Palak per i bei momenti trascorsi insieme, per i dialoghi sinceri, per l'amicizia e l'affetto che mi dimostra sempre. Vorrei ringraziare Anna, Andrea, Salvatore e Monica, per il loro supporto, per l'affetto che hanno sempre dimostrato e che porto sempre nei miei pensieri. Ringrazio le mie amiche di sempre, Caterina, Carolina, Ilaria, Laura e Eniana, per i bei momenti trascorsi e per il supporto dimostrato. Inoltre, ringrazio Alessandra per essere stata una buona amica e preziosa confidente. Grazie anche alle mie amiche del Forcellini per aver lasciato dei ricordi indelebili, in particolare a Maria Giulia per aver reso più piacevoli i miei anni universitari.

Dulcis in fundo, vorrei ringraziare Carlo. Ti ringrazio per i bei momenti passati insieme, per la felicità e serenità che mi regali ogni giorno, per il supporto e la pazienza nei momenti difficili, per tutti i dialoghi profondi che potrei fare solo con te perché *Ci vuole gente intelligente/Anche per parlare, per stare in compagnia./ Ci vuole gente intelligente/ Anche per non essere, non essere banali, volgari./ Ma cosa vuoi di più?*