THE UNIVERSITY *of* EDINBURGH

Metabolomics and machine learning to assist biotechnology culture optimisation

Ricardo Gabriel Valencia Albornoz

Thesis submitted for the Degree of

Doctor of Philosophy

School of Biological Sciences, The University of Edinburgh, United Kingdom

2024

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

Ricardo Valencia Albornoz

Abstract

Optimisation of product titre or yield in a bioprocess is crucial for the economic and technical success of its operation. This optimisation problem is usually a challenge, as it involves several factors or variables. For example, in a bioprocess, medium components are important factors in the final titre, and the concentrations of each component need to be manipulated to achieve optimal conditions. Here, we present an active learning/Bayesian optimisation framework to enhance surfactin titres by adjusting medium component concentrations. Surfactin, produced by bacteria of the genus *Bacillus*, is a promising biosurfactant due to its physical and chemical properties. However, reported laboratory titres are typically low because of its complex molecular assembly pathway. We used active learning to refine the culture medium composition through iterative experimentation, enhancing Surfactin C levels in *Bacillus subtilis* DSM 3256. Growth curves and other central metabolites were measured as part of the experimental loop. The final medium mixture resulted in approximately a 1.6-fold increase after three rounds, compared to the M9 medium standard. Reanalysis of the optimisation data reveals trade-offs when comparing the production of lipopeptides, such as Surfactin D and Iturin A, with the maximum OD in the growth curve data. Organic acids in the supernatant positively correlate with Surfactin C levels, suggesting an impact on central carbon metabolism. For some metabolites, including certain amino acids and sugars, the change in their abundance around the optimal surfactin C mix is not uniform, indicating an "anisotropy" in how metabolism reacts to shifts in carbon and nitrogen levels. Thus, our framework addresses the challenges of data handling and analysis, offering several visual tools, data analysis techniques, and analytical methods (using mass spectrometry), which promised to be a contribution to Design, Build, Test & Learn cycles.

After addressing the challenge of modifying the concentrations of two components in the culture medium, we scaled up our approach to optimise surfactin production by modifying all seven components of the M9 medium, transforming it into a multidimensional optimisation protocol. However, performing the mixing and medium preparation became technically challenging.

Thus, this experiment was made possible through a high degree of automation, both computationally and experimentally. Two pipelines were built: the first one addresses the initial sampling and first robotic experiment in a Bayesian optimisation loop, while the second one execute data analysis following data acquisition from the mass spectrometer and can couple with the concentration mixing protocol in the Opentrons OT-2 robot for subsequent iterations. The Opentrons scripts were able to calculate and transfer the correct volumes of each component based on the stock concentration and desired concentration in the wells, generating robot-ready instructions to perform the mixing. Similar protocols were employed for quenching and sample preparation, enabling a full experimental cycle in 2-3 days.

In the experimental design part, we opted for an off-line approach, whereby sufficient samples—specifically 42 combinations with four replicates each, plus quality control and biological controls—were obtained from a single space-filling design to cover the full seven-factor space. The results indicate that combinations close to the M9 reference composition are the highest producers of surfactin C, confirming the optimal carbon and nitrogen conditions from the previous 2D iterative experiment. We then generated a high-quality surrogate model of production outputs, including lipopeptide production and biomass, measured as OD. This model serves as a realistic benchmark for testing single-objective and multi-objective lipopeptide production optimisation using Bayesian optimisation. From the single objective optimisation, results showed that the optimal number of initial samples and batch size can be adjusted to achieve the maximum Surfactin C yield in fewer iterations. However, the greater number of factors and the observed variance in the measurements mean the iterations cannot be reduced further, with approximately 10 iterations required using our current experimental setup of seven initial samples and seven combinations per batch (with 4 replicates). In the case of multi-objective optimisation, A Bayesian optimisation framework was able to identify the Pareto Front between lipopeptide production and biomass in the 7-dimensional factor space, with batch sizes and number of iterations comparable to those obtained from microplate experiments. This thesis tested that Bayesian optimisation is a feasible option for optimisation of secondary metabolites such as lipopeptides and that this approach can integrate with automation for high-throughput microbial metabolism studies.

Lay Summary

Bacteria is a source of relevant chemicals for modern society, as well as a promised system for sustainable (bio)chemical production of complex molecules, that often have large carbon footprints. A bacterial genus commonly employed in bioprocess is *Bacillus*. This genus has demonstrated being a source of many chemicals of relevance, as well as a chassis for bioproduction of custom molecules.

In this project, we focus on increasing the production of a natural product called Surfactin, made by *Bacillus*. Surfactin is a biosurfactant, a type of molecule that has various uses in cleaning, oil recovery, and other industries due to its special properties. However, getting bacteria to produce high amounts of Surfactin is difficult because the synthesis process inside the bacterium is quite complex.

To solve this, we used machine learning and experimental design to adjust the ingredients in the bacterium's growth medium. By carefully changing the amounts of certain components, we were able to increase Surfactin production by about 1.6 times compared to standard methods. In addition to boosting Surfactin levels, we also learned more about how the bacterium's metabolism changes when conditions like carbon and nitrogen levels are adjusted. We found that certain substances, such as organic acids, were linked to higher Surfactin production. Understanding these relationships helps us fine-tune the production process.

Then, with the goal of changing every component in the growth medium, we used robots to prepare the growth medium and computer programs to analyse the results, helping us decide how to improve the process further.

Overall, this project shows how combining automation, machine learning, and fast experimentation can make bioprocesses more efficient. This is especially useful for producing valuable and renewable products like Surfactin and could lead to more sustainable (bio-)chemical production in the future.

Publications

The contents of this thesis have been reproduced, fully or in part, in the following scientific publications:

Valencia Albornoz, R., Oyarzún, D., & Burgess, K. (2024). Optimisation of surfactin yield in *Bacillus* using data-efficient active learning and high-throughput mass spectrometry. *Computational and Structural Biotechnology Journal*, *23*, 1226–1233. https://doi.org/10.1016/j.csbj.2024.02.012

Ricardo Valencia Albornoz

Acknowledgments

I would like to thank my family for accompanying me in this journey. Firstly, my parents Angelina and Ricardo, whose continuous love and support across continents were essential to complete this dream. They are my inspiration, and I am really proud of them every day. I would like to thank my Tata Gabriel, who is always around, and to who I have great affection, even in the distance. What he taught me is now helping me to develop my career. Also, to my family in Valparaiso (Veronica, Gabriel, Francisca, Sebastian), Santiago (Gabriela, Victor, Laurana, Emiliano) and Williamstown (Nardy, Claudio, Benjamin, Gabriel), who always have been checking that I'm doing well, and have follow my path over all these years, undergraduate and before. With a special feeling, this thesis is also dedicated to my Lela Angelica, who passed away in 2019. I think she would be proud of where I am stepping right now and how life is going.

This thesis could not have been possible without the support of my supervisors, Karl and Diego, who not only introduced me to the worlds of metabolomics and machine learning for biology, but also gave me the tools and the freedom to pursuit new scientific horizons. They have been of invaluable help from the moment I arrived at the UK to the conclusion of this project and are truly models of how to do great science in a human way. With them, there is an amazing group of people, not only because they make science possible at Edinburgh, but also because of their quality as colleagues. I would like to thank Luke, Dasha, Jessica, Aya, Georgie, Jess, Lisa, Susana, Fongling; thanks for accompanying in the office and make my day by day better. Shoutout to the amazing team at EdinOmics, including Tessa, Fraser and Andrew. They were simply instrumental for the development of this thesis. From Diego's lab, I would like to thank Yuxin, Arin, Lucas, Zuzanna, Charlotte, Evangelos, Michael, Achille, and Michal, thanks for the fruitful discussions and meals.

I would like to specially thank Vanessa, who welcomed me for Christmas in my first year, and whose friendship and support has been profound to me, at many levels: the science, the dance, and the personal level too. My experience in Edinburgh has been richer thanks to you. Also, to Joan and Rebekah, who made my PhD time a great time with several food and tennis meetings. Of course, with Joan and Alán, we had

good chess games together. To Yu and Justin, great friends along these four years, with an unlimited hospitality and it is nice to see that we are growing professionally together. To Hazel, for introducing me to the tennis club and for a long-term friendship, inside and outside the court. And to Manon, for the most random stories and gigs that have happened in the city.

The biology community at Edinburgh is welcoming and I found there a great place to develop science and to enjoy simple things. Many thanks to the third-floor people, that creates a comfortable space for working and for sharing our PhD conundrums. Also, to the second-floor people, who I have shared from small snack moments to tennis cheerleading squad, and then struggling preparing posters and everything you can imagine for a PhD life. Thanks to Sofia, Mehak, Wei, Niki, Chu, you are wonderful people.

I would like to thank friends, in the UK and Chile. Special mention to Pat and Francisco that for many years bear with me and accompany me in every celebration and every struggle, every time with nice Chinese food. To Elisa and Natsuki, their friendship at Richmond Place was very important for me, and I am glad we were able to celebrate our achievements. Many people from there formed part of my daily life and I am very happy that I met them. To the people at Craigmillar Park Tennis Club, Aly, Sandy, Mike, John, Paul, Alex, Gerald and other players, who adopt me as a new player, but also as a friend. From Chile to my friends at the Universidad Tecnica Federico Santa Maria, that at the distance have cheered me up.

I would like to acknowledge the tremendous support, which was the spark to enrol in a PhD programme, given by the Actino Lab at the Universidad Tecnica Federico Santa Maria in Chile. Special thanks to Beatriz and Agustina, whose energy and help made possible for me to start a scientific career in the UK.

Finally, I would like to thank my fiancée Aink Acrie, who has been a fundamental pillar of my life over the last two years. Thank you for the love and happiness you give me every day and for taking care of me, which was crucial for the writing of this document. Thanks for showing me that things always get better, that dreams can come true when we work and eat together, and that the future looks bright. I cannot wait to start our new life.

Table of Contents

Declaration ii

Abstract iii

Lay Summary v

Publications vi

Acknowledgments vii

Table of Contents ix

Figure Index xii

Table Index xv

1. **Introduction** 1

1.1. Brief introduction to bioprocesses 1

1.2. Surfactin production by bacteria 3

1.2.1. Surfactants and biosurfactants 3

1.2.2. Surfactin characterisation and properties 5

1.2.3. Biosynthesis and regulation 6

1.3. Design of experiments in bioprocesses 8

1.3.1. Common design of experiments approaches in bioprocesses 9

1.4. Modelling beyond polynomial models 10

1.4.1. Gaussian processes 13

1.4.2. Other machine learning algorithms and cross-validation 17

1.5. Bayesian optimisation: an adaptive experimental design for optimisation 26

1.5.1. Initial sampling 27

1.5.2. Probabilistic surrogate modelling 29

1.5.3. Acquisition functions 29

1.5.4. Stopping of a Bayesian optimisation loop 34

1.5.5. Batch Bayesian optimisation 34

1.5.6. Uses of Bayesian optimisation routines in biology 35

1.6. Metabolomics 36

1.6.1. Technological foundations 37
1.6.2. Application of metabolomics 39
1.6.3. Integration with other omics 41
1.6.4. Historical context 42
1.7. Mass spectrometry basics 42
1.7.1. Main concept 43
1.7.2. The instrument 43
1.7.3. Historical notes 47
1.8. Functioning of a triple quadrupole mass spectrometer 47
1.9. Mass spectrometry-based metabolomics sample preparation 49
1.10. Mass spectrometry-based metabolomics data analysis 50
1.10.1. Filtering 50
1.10.2. Peak/feature detection 51
1.10.3. Alignment 51
1.10.4. Area under the curve, normalisation and batch correction 51
1.10.5. Metabolite annotation 52
1.10.6. Common statistical analysis in metabolomics 53
1.10.7. Technical concepts to consider when acquiring metabolomics data 54
1.11. Probing bacterial biochemistry through mass spectrometry-based metabolomics 56
1.12. Metabolomics profiling for strain optimisation 57
1.13. Hypothesis and aims 57
1.13.1. Hypothesis 58
1.13.2. Aims 58
2. **Optimisation of surfactin yield in Bacillus using active learning and high-throughput mass spectrometry** 59
2.1. Introduction 60
2.2. Material and Methods 64
2.3. Results 68
2.4. Discussion 84

2.5. Conclusion 86
2.6. Acknowledgements 87
2.7. Data availability 87
3. **A robotic platform for semi-automated iterative experimentation and additional protocols** 88
3.1. Introduction 89
3.2. Material and Methods 92
3.3. Results 97
3.4. Discussion 120
3.5. Conclusion 124
3.6. Acknowledgements 124
3.7. Data availability 124
4. **Single and multi-objective optimisation of titres from a *Bacillus* space-filling experiment** 125
4.1. Introduction 126
4.2. Material and Methods 129
4.3. Results 135
4.4. Discussion 153
4.5. Conclusion 155
4.6. Acknowledgements 156
4.7. Data availability 156
5. **Discussion and Conclusion** 157
References 162
Appendix A. 222
A.1. Thin layer chromatography 222
A.2. Development of a 3D-printed electroporation tip for Opentrons 225
A.3. Mathematical explanation for the sigmoidal shape in ordered samples 227
A.4. List of materials for building the robotic arm 229
Appendix B. Supplementary Figures and Tables 231

Figure Index

Figure 1.1. Development in upstream processing from the nanolitre to thousands of litres..... 2

Figure 1.2. Chemical structure of Surfactin C and other lipopeptides found in Bacillus..... 7

Figure 1.3. Central composite design in 2D and 3D design spaces 10

Figure 1.4. An example of a Gaussian process defined in unidimensional domain ... 13

Figure 1.5. Gaussian process regression on a 1D dataset with 4 points 14

Figure 1.6. Gaussian process draws with a Matern kernel as covariance function 15

Figure 1.7. A visualisation of k-fold cross-validation 20

Figure 1.8. PCA rotates a dataset when performed in 2 dimensions 23

Figure 1.9. PCA biplot on the Iris dataset 26

Figure 1.10. A diagram of a typical Bayesian optimisation loop 27

Figure 1.11. A comparison between random sampling, Sobol sampling and Latin hypercube sampling in a 2D design space 28

Figure 1.12. Branin function and Ackley function in a 2d space as contour plots 34

Figure 1.13. A diagram of a generic mass spectrometer that employs a magnetic sector analyser 43

Figure 1.14. Diagram of an electrospray 44

Figure 1.15. Components of a triple quadrupole mass spectrometer 48

Figure 1.16. A typical sample order for a set of samples in a metabolomics experiment 52

Figure 1.17. Example of an extracted ion chromatogram 55

Figure 1.18. Example of a total ion chromatogram 55

Figure 2.1. Metabolic pathways converging into Surfactin synthesis 61

Figure 2.2. Growth curves for Bacillus subtilis DSM 3256 69

Figure 2.3. Benchmarking parameters of a Bayesian optimisation loop in a test function 72

Figure 2.4. Active learning for optimisation of Surfactin titres..... 74

Figure 2.5. Pareto front obtained by simulating the lipopeptide production and biomass 76

Figure 2.6. Multivariate analysis of additional metabolites in the experiment 78

Figure 2.7. Detailed pathways and key intermediates for the synthesis of amino acids and fatty acids in Bacillus subtilis 79

Figure 2.8. Directional analysis reveals which metabolites are sensible to slight and simultaneous changes in carbon/nitrogen composition 81

Figure 2.9. Predicted vs actual mean values plot for the fitted Surfactin C Gaussian process regression (GPR) model after each iteration 83

Figure 2.10. Predicted vs actual uncertainty values plot for the fitted Surfactin C Gaussian process regression (GPR) model after each iteration 84

Figure 3.1. Comparison of different solvent mixes as mobile phase for a flow injection run of *B. subtilis* DSM 3256 supernatant 98

Figure 3.2. Test of pipetting accuracy on the Opentrons OT-2 robot at normal aspiration and dispense speed 100

Figure 3.3. Test of pipetting accuracy on the Opentrons OT-2 robot at normal aspiration and dispense rate of 0.5 101

Figure 3.4. Picture of the Opentrons OT-2 robot in a step during a typical run 102

Figure 3.5. Initial sampling and initial Opentrons run pipeline 103

Figure 3.6. Snippet of code from the 1_initial_samples.py script 104

Figure 3.7. Brief description of the scripts of the initial pipeline 105

Figure 3.8. Pipeline for subsequent iterations 106

Figure 3.9. Snippet of code from the a3_batch_correction_norm_QC.R script 107

Figure 3.10. Description of files for the iteration pipeline 108

Figure 3.11. Example of Apache Airflow code to execute and monitor the initial sampling pipeline 109

Figure 3.12. A view from the console in Apache Airflow 110

Figure 3.13. UV lamp located in a corner of the Opentrons robot 111

Figure 3.14. The small servo is installed on the grip case and the cable extension is evident 112

Figure 3.15. Forearm section with integrated grip part 113

Figure 3.16. Main arm assembly displaying the shoulder 20 kg servo shaft 114

Figure 3.17. The shoulder joint is attached to the base 114

Figure 3.18. A view to the circuitry of the robotic arm 115

Figure 3.19. Diagram for the connections in the robotic arm 116

Figure 3.20. The fully assembled robotic arm 117

Figure 3.21. A 3d printed cover for 96-well plates with holes 118

Figure 3.22. 3D printed racks used in Opentrons runs 119

Figure 3.23. A 3D printed cuvette and microtube holder. 119

Figure 4.1. The concept of a Pareto front 127

Figure 4.2. Extracted ion chromatograms for surfactin C across the space-filling samples 137

Figure 4.3. Intra and inter batch correction of area under the curve quantification of extracted peaks using the pmp package in R 138

Figure 4.4. Surfactant Concentration across different batches and combinations from the space-filling design 139

Figure 4.5. Contour plot of Surfactin C titres against pairs of culture medium components 141

Figure 4.6. Optical Density (OD) measurements after 24 hours across different batches and combinations 142

Figure 4.7. Contour plot of 24-hour Optical Density (OD) measurements against pairs of culture medium components 143

Figure 4.8. A “naïve” training of a Heteroskedastic Gaussian process to the processed Surfactin C titre data from the space-filling design 145

Figure 4.9. Diagram of a probabilistic ensemble learning algorithm using Bayesian averaging 146

Figure 4.10. Predictions of the Bayesian averaging ensemble models for a test set in the Optical Density (OD) data 148

Figure 4.11. Feature importance as retrieved from the Random Forest base learner in the Bayesian averaging ensemble model 149

Figure 4.12. Performance of our single-objective optimisation Bayesian optimisation pipeline in the Surfactin C surrogate model 151

Figure 4.13. Performance of our single-objective optimisation Bayesian optimisation pipeline in the Optical Density (OD) surrogate model 151

Figure 4.14. Performance of a multi-objective optimisation of Surfactin C titres and Optical Density, trying to maximise both simultaneously 153

Figure A1. A TLC run of a Bacillus subtilis supernatant sample 224

Figure A2. Imaging of plate after staining with Rhodamine G..... 224

Figure A3. Insufficient drying and quick heating can lead to bubbling 225

Figure A4. An Opentrons tip to perform electroporation on 48-well plates 226
Figure A5. The built glass slide smearer 227
Figure A6. The probit function or quantile function of the normal distribution..... 227
Figure A7. Plots of values of a linear function over 5 variables, in increasing order... 228
Figure A8. Plot of values of a non-linear function over 5 variables 229
Figure A9. Ordered GFP synthesis rate vs individual strains, extracted from Zhang .229
Figure B.1. Surfaces for the 4 lipopeptides 231
Figure B.2. Uncertainty surfaces for the 4 lipopeptides 232
Figure B.3. Surfaces for 18 carbon/TCA-related compounds 233
Figure B.4. Uncertainty surfaces for 18 carbon/TCA-related compounds 234
Figure B.5. Surfaces for 2 biomass measurements 235
Figure B.6. Uncertainty surfaces for 2 biomass measurements 235
Figure B.7. Diagram showing what data serves as input to the model 235
Figure B.8. Growth curves for the microplates 236
Figure B.9. OD curves for the space-filling design experiment 236
Figure B.10. Contour plot of Surfactin B titres against pair of components concentrations 237
Figure B.11. Results of the Bayesian averaging ensemble prediction on test set for Surfactin B 237
Figure B.12. Single objective optimisation of Surfactin B using the surrogate data 238

Table index

Table 3.1. Summary of accuracy and precision values 101
Table BT1. Parameters used on the triple quadrupole mass spectrometer 239
Table BT2. Precursor and product masses used for MRM 239
Table BT3. Estimation of growth rates for OFAT experiments 240
Table BT4. Pipetting results for normal speed test 241
Table BT5. Pipetting results for slow speed test 244

1. Introduction

1.1. Brief introduction to bioprocesses

Bioprocess engineering is a discipline that integrates biology, chemistry and engineering to optimise the use of living cells or their components to produce valuable substances (O'Brien & Hu, 2020). This approach is pivotal across various industries, from pharmaceuticals to food and biofuels, by turning raw biological materials into useful products under controlled and optimized conditions (Koutinas et al., 2012).

In the pharmaceutical sector, bioprocesses enable the production of critical biopharmaceuticals, including vaccines and monoclonal antibodies, through techniques involving the cultivation of cells, extraction, and purification processes (Jozala et al., 2016; Szkodny & Lee, 2022). For example, during World War II, bioprocess engineering facilitated the scaled-up production of penicillin, transforming it from a laboratory-scale promise to a mass-produced antibiotic (Rokem et al., 2007; Wang et al., 2014; Bandyopadhyay et al., 2017).

The food and beverage industry also benefits significantly from bioprocesses. Yeast and bacteria have been used from ancient times in the fermentation of products like beer, wine, and cheese (Maicas, 2020; Siddiqui et al., 2023), where modern bioprocess engineering have helped to optimise production and final quality in terms of flavour, texture and other organoleptic properties (Bibra et al., 2021). Two interesting examples are the production of plant-based proteins and alternative meat products. These innovations leverage bioprocessing techniques, such as precision fermentation and cell culture, to create sustainable and environmentally friendly alternatives to traditional animal-based products (Post et al., 2020; Rubio et al., 2020; Chen et al., 2022).

Furthermore, in the energy sector, bioprocess engineering contributes to sustainable practices by facilitating the production of biofuels such as bioethanol from agricultural feedstocks (Melendez et al., 2022; Periyasamy et al., 2023). This application fits into the global concept of biomanufacturing, where renewable sources are harnessed to produce bulk chemicals such as bioplastics (Rosenboom et al., 2022;

de Souza & Gupta, 2024) and bio solvents (Li et al., 2016), or fine chemicals such as taxol (DeJong et al., 2006; Xie et al., 2024) or artemisinin (Paddon et al., 2013), using biosynthesis or biological transformations.

Bioprocess engineering involves two main development phases: upstream and downstream processing. Upstream processes focus on the preparation and cultivation of microbial or cell cultures to optimize their growth and productivity (Matanguihan & Wu, 2022). Downstream processing, meanwhile, involves the recovery and purification of bioproducts from the culture medium, which is essential for achieving high purity and quality in the final product (Baumann & Hubbuch, 2017; Cramer & Holstein, 2011).

On upstream development, advancements in genetic engineering, process analytics, and the integration of new bioreactor technologies, have made culturing of microbes a reliable and efficient process (Boodhoo et al., 2022). Moreover, several efforts are focused on testing growth conditions and process parameters in smaller system in comparison to larger cultures, which are more expensive to test (Lattermann & Büchs, 2015). These scale-down approaches benefit of lower cost and higher throughput for parameter optimisation, given valuable physical and biochemical information for scaling-up development (Delvigne et al., 2017). Thus, several platforms going from nanolitres, such as microfluidic devices (Funke et al., 2010; Brás & Fernandes, 2024), to plate experiments in microliters (Auld et al., 2004), offers a comprehensive solution for screening and optimizing microbial cultures (Figure 1.1).

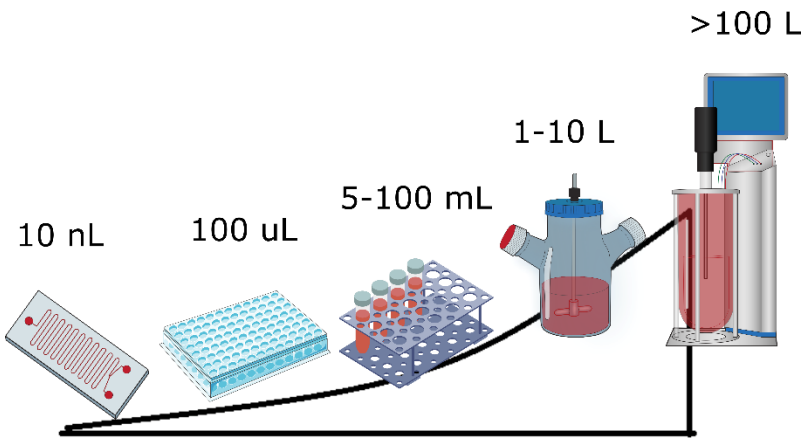


Figure 1.1. Development in upstream processing from the nanolitre to thousands of litres. Different platforms are used to grow microorganisms on scale. Each of them offers advantages and disadvantages, where the cost per sample goes down with the size, but the physical environment might be less representative for bulk process development.

Typical measure of performance in bioprocess development are titres, rates and yield (TRY). Titres correspond to the concentration of desired compound in the final solution (Konzock & Nielsen, 2024), which can be reported as relative abundances, concentration on molar units or g/L. Rates refer to the speed at which the bioprocess produces the desired product, often measured as volumetric productivity (g/L/h), while yield indicates the efficiency of the conversion process, representing the amount of desired product generated per unit of substrate consumed (Konzock & Nielsen, 2024). Optimisation techniques are key to increasing these numbers, by overcoming trade-off in bioprocess conditions, and therefore expanding profit margins at operation (Banerjee & Mukhopadhyay, 2023). These are focused on manipulating medium components or abiotic components, such as temperature and agitation, via a proper experimental design, so that the organism can grow optimally and/or synthesise more of the desired compound (Mandenius & Brundin, 2008; Mondal et al., 2023; Zhou et al., 2023). Optimisation can also be performed at the molecular level, by engineering pathways for enhanced production and coupling pathways to growth for laboratory evolution of process strains (Jones & Koffas, 2016; Cheng et al., 2023). Finding optimal conditions can later inform intensification of the bioprocess, where production is taken to bigger scales and with more resources, such that cultures cycles are faster and robust, i.e., consistent production over time and cell generations (Boodhoo et al., 2022; Olsson et al., 2022)

The project uses surfactin production in *Bacillus* as the bioprocess to analyse and optimise titres by changing culture medium composition.

1.2. Surfactin production by bacteria

1.2.1. Surfactants and biosurfactants.

Surfactants are molecules capable of altering the surface tension of interfaces (Cullum, 1994). A common system is the water-air interface, where surfactants added

to the water decrease its surface tension (Cullum, 1994). Due to this effect, they are a fundamental resource in modern industry, including wetting agents, emulsifiers and adjuvants, as well as in the home as detergents and cleaners (Shaban et al., 2020).

These molecules are generally amphiphilic, exhibiting a hydrophilic "head" and a hydrophobic "tail" (Myers, 2020). The resulting polarity induces the surfactant molecules to organize in a way that the hydrophilic heads are oriented towards the water, while the hydrophobic tails are directed away from it, leading to the formation of layered structures at interfaces, and altering water's cohesive forces (Myers, 2020). If the concentration of surfactant molecules in solution exceeds a certain threshold known as the critical micelle concentration (CMC), micelle structures can form (Perinelli et al., 2020). In micelles, the surfactant molecules arrange themselves in spherical structures with the hydrophobic tails pointed inward, away from the water, and the hydrophilic heads pointed outward. This arrangement minimizes the unfavourable interactions between the hydrophobic tails and the aqueous environment, thereby stabilizing the structure of the micelles in solution (Perumal et al., 2022). Micelles are important in various applications, including in detergents for breaking down oils and in drug delivery systems for encapsulating hydrophobic drug molecules (Perumal et al., 2022).

Most of the daily-used surfactants are derived from the petroleum or vegetable oil industry (Bhadani et al., 2020). The chemical synthesis process of these molecules involves a pipeline where the hydrophilic head is added to a long carbon chain with a terminal reactive group, as in the case of sodium lauryl ether sulphate (Zumpano et al., 2024). On the renewable side, agricultural feedstocks can be transformed via a chemical process into fatty acids and other biomolecules, that can be further used as reagents for surfactant manufacturing (Bhadani et al., 2020).

Oil-derived surfactants are a constant environmental problem and their corresponding long-term impact on the environment is still being investigated (Johnson et al., 2021). Alternatives have been proposed to replace oil-based surfactants, where these options need to fulfil several criteria to be used in high-duty applications expected for industrial use. Thus, an effort has been made to produce biosurfactants, which are environmentally friendly, safe for human health, and

biodegradable (Nikolova & Gutierrez, 2021; Markande et al., 2021; Eras-Muñoz et al., 2022). Biosurfactants, or bio-based surfactants, are defined as surface-active compounds produced by living organisms, such as bacteria, fungi, and plants (Henkel et al., 2017). When microorganisms are only considered as producers, i.e., microbial surfactants, these molecules alone encompass a wide range of chemical classes, including lipopeptides, glycolipids, phospholipids, polymeric surfactants, among others (Jahan et al., 2020; Twigg et al., 2021; Vieira et al., 2021) and are synthesized by strains from different phyla (Soberón-Chávez & Maier, 2011; Kumari et al., 2023).

Industrial production and demand for surfactants are increasing year after year, with a valuation of USD ~43.5 billion in 2022 (MarketsAndMarkets, 2024), with an expected biosurfactant market growth from approximately USD 1.2 billion to USD 2.3 billion by 2028 (MarketsAndMarkets, 2024). Thus, to keep the pace with demand, there is increasing pressure to optimize titres for biosurfactant biorefineries, finding stable microbial surfactants for high-duty applications and overcoming the natural trade-offs between higher production and cell density due to biosurfactant-induced foaming and membrane disruption (Kanwal et al., 2023).

1.2.2. Surfactin characterisation and properties.

Surfactin is a promising biosurfactant for various applications, and its biosynthesis has been thoroughly studied, making it a well-known representative of microbial biosurfactants (Arima et al., 1968; Théatre et al., 2021). Originally isolated from a *Bacillus* sp. culture (Arima et al., 1968), researchers later realized it can be produced by several other species of the genus Bacillus (Steinke et al., 2021). It currently has a wide range of applications in industry, agriculture, and medicine, including uses as emulsifiers, dispersants, and biocontrol agents (Ongena and Jacques, 2008; Jacques et al., 2011; Zhen et al., 2023).

In nature, surfactin secretion provides bacteria with additional capabilities that aid in survival or feeding exploration. It is known that in situ surfactin production facilitates cell motility and colonization (Raaijmakers et al., 2010), with cells swimming along on waves of surfactin searching for better feeding conditions (Angelini et al., 2009). When interacting with other bacteria, surfactin production is induced as a spatially distributed antibiotic, disrupting the membranes of nearby cells

(Stein, 2005; Hoefler et al., 2012; Luzzatto-Knaan et al., 2019). Interestingly, it has also been proposed that membrane depolarization mediated by surfactin presence provides a mechanism of cell survival under conditions of oxygen depletion (Arjes et al., 2020).

1.2.3. Biosynthesis and regulation

On the chemical side, surfactin is classified as a lipopeptide, possessing a hydrophilic amino acid ring "head" and a hydrophobic fatty acid "tail". Surfactin biosynthesis is carried out by a non-ribosomal peptide synthetase (NRPS) system (Koglin et al., 2008; Süssmuth & Mainz, 2017; Théatre et al., 2021). NRPSs are gene clusters that encode mega enzymes which produce molecules in tandem, like an assembly line. These enzymes possess multiple domains, catalysing different biosynthetic reactions in coordinated action (Weissman, 2015). Each module comprises a condensation, adenylation, and a thiolation domain, catalysing the integration of an amino acid, with rules for amino acid insertion determined based on enzyme motifs (Süssmuth & Mainz, 2017). This module structure is not fixed, as additional reaction domains can be present, such as epimerization domains (Miller & Gulick, 2016), which is observed in the final surfactin molecule by the presence of the two stereoisomers of the amino acid Leucine. In the surfactin gene cluster, 3 biosynthetic genes encode for megaenzymes containing seven modules in total, each of them facilitating the incorporation of a single amino acid, and thus synthesising the final heptameric amino acid ring (Théatre et al., 2021). Surfactin C, comprising an n-carbon fatty acid and a ring composition of L-Glu1-L-Leu2-D-Leu3-L-Val4-L-Asp5-D-Leu6-L-Leu7 (Figure 1.2), is one of the most well-known variants and therefore a representative of the family (Théatre et al., 2021).

Due to its combinatorial assembly mechanism, there is not a single, defined biosynthetic pathway leading to surfactin production. Instead, multiple pathways supply the required molecular precursors in the cytoplasm for assembly. These precursors include the branched-chain fatty acids and amino acids that form the final molecule (Hu et al., 2019; Xia and Wen, 2022). This flexibility in assembly implies that variants of surfactin can be produced, especially if a precursor undergoes slight modifications or if there is an error in the amino acid assembly. This results in a family

of related molecules, known plurally as the surfactin family or surfactin-like molecules (Ongena and Jacques, 2008).

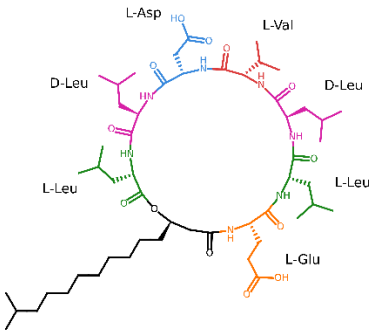


Figure 1.2. Chemical structure of Surfactin C and other lipopeptides found in *Bacillus*. The amino acid residues are coloured for easy visualisation. L-Glutamate is in orange, L-Leucine in green, D-Leucine in purple, L-Valine in dark red and L-Aspartate in blue. The ester bond links the fatty acid to the ring.

From a regulatory perspective, several genetic systems influence the regulation of the surfactin biosynthetic gene cluster. A notably significant one is quorum sensing, primarily via the ComQXPA system (Qiao et al., 2024), the competence system in *Bacillus*. Within this system, the NRPS cluster promoter gene, P_{srf}, is stringently regulated within a specific subpopulation (López and Kolter, 2010; Kalamara et al., 2018; Rahman et al., 2021). Specifically, there is a cascade of responses based on environmental conditions, such as nutrient availability. The competence system, including the genes ComA-P, ComS and ComK, senses the presence of ComX in the extracellular medium due to cell density (Comella & Grossman, 2005). ComA has been demonstrated to positively regulate the surfactin cluster, triggering the production as response to increasing neighbour population (Nakano et al., 1991). One of the mentioned genes, ComK, is also induced by AbrB, which is in turn repressed by Spo0A. The gene Spo0A controls several genetic programs in *Bacillus* towards sporulation (López et al. 2009; López and Kolter, 2010). Indeed, it has been recognised that sporulation and competence are alternative responses when cells are in starvation or general environment perturbation (Schultz et al., 2009).

In addition to sporulation-competence processes, other genes either activates or represses the surfactin cluster expression (SubtiWiki) (Elfmann, 2024), including

PerR, which is an iron uptake repressor (activation) (Hayashi et al., 2005), PhoP, a two-component systems involved in adaption to manganese concentration (activation) (Salzberg et al., 2015), Abh, that is related to biofilm formation and secondary metabolite production (repression) (Chumsakul et al., 2011), CodY, which senses and regulates metabolic states and it is one of the genes responsible for catabolic control (repression) (Serror & Sonenshein, 1996), and Spx, involved in stress response (repression) (Nakano, 2003).

Additionally, surfactin itself is considered a quorum molecule, causing important metabolic shifts to other *Bacillus* strains (Wen et al., 2021)

1.3. Design of experiments in bioprocesses.

Bioprocess optimisation is performed by tweaking parameters or inputs and quantify the influence of them in the output, which is commonly a titre and yield. An experimental scheme which is commonly used to test these inputs is one-factor-at-a-time (OFAT) experiments. As the name indicates, the experiments are recording by keeping all factor fixed, except for one that it is varied. This procedure is repeated until every factor has been tested (Czitrom, 1999). Although this method is still popular in the literature, it has been described as inadequate to optimise bioprocesses (Toms et al., 2017). It possesses clear disadvantages, such that no interactions between factors/inputs can be resolved (Czitrom, 1999) and the chosen level for the fixed factors may be not an appropriate one for a realistic bioprocess operation (Toms et al., 2017). OFAT may have a use in determining the range of factors affecting culture growth, since above a value of specific component, prominent cell death might be observed and that determines the maximum value of the range. Examples of One-factor-at-a-time (OFAT) experiments in bioprocesses, often in comparison with other experimental designs, include enzymes and valuable amino acids (Abdel-Rahman et al., 2020; Bhaturiwala et al., 2022; Nor et al., 2017).

Therefore, better experimental designs are needed that can be sample-efficient, while obtaining detailed information about factors' influence. Thus, classical design of experiments (DoE) has been extensively used for optimisation of bioprocesses, and specifically in biosurfactant production (Mandenius & Brundin, 2008; Bertrand et al.,

2018; Kasemiire et al., 2021). Although experimental designs are abundant in the literature for different process objectives, a few of them are often selected by their generality and modelling complexity. These are factorial designs, response surface methodology using central composite design (CCD) and screening designs such as Plackett-Burman (Balakrishnan et al., 2022).

1.3.1. Common design of experiments approaches in bioprocesses

Factorial design involves testing 2 or more levels for each factor to manipulate and measure the output for each combination of these levels (Lawson, 2015). In the simpler case, a minimum value and a maximum value per factor is taking into consideration (2 levels), so the number of combinations and required experimental samples is 2^k, where k is the number of factors (Lawson, 2015). Outputs of full factorial designs and their variation, using 2^k experiments, can be modelled using a linear model over the factors, now variables in the model, and their interactions, represented as product of the variables (Kaltenbach, 2021a). When a lower level of detail is required, it is possible to halve a full factorial design into a fractional factorial design, which requires less experiments, but the interactions are confused with other terms in the model, changing the interpretation of the coefficient in the linear model related to variable importance (Kaltenbach, 2021b). Factorial designs have been successfully employed to understand bioprocess parameters, both at medium composition and process operation (Mandenius & Brundin, 2008. There are reported uses in biosurfactant production, including analysing the effect on medium components in surfactin yield (Fonseca et al., 2007; Zouari et al., 2014).

Response surface methodology (RSM) aims to model surfaces of outputs as approximations for a true output (Breig & Luti, 2021). The most popular design to perform RSM is central composite design (CCD). CCD can obtain extra information about the measured output, since more points are added in the area between the min-max range of the factors, in comparison to a factorial design. These points allow for higher flexibility modelling, with quadratic models being commonly employed to predict the output over the design space (Lawson, 2015). A CCD can be split into cube points, which reassembles a factorial design and star points, which correspond to a radially shape set points inside the cube domain (Figure 1.3). CCD satisfy a series of

desirable statistical properties, such as being rotatable (all points are at the same distance of the centre) when proper samples are selected (Lawson, 2015). CCD have a profound influence in engineering science, and it is a popular choice in bioprocess modelling, *e.g.*, in the optimisation of bioethanol production (Pereira et al., 2021).

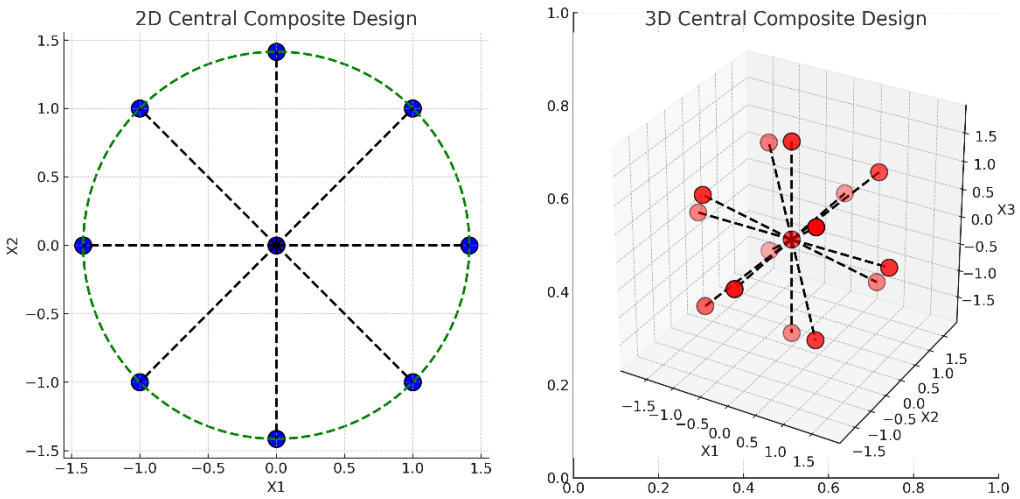


Figure 1.3. Central composite design in 2D and 3D design spaces. Central composite design can be decomposed into cube points that are 2-level factorial designs over the design space, and star points that add the flexibility to fit higher-order models to the data.

Finally screening designs are alternative design that tries to obtain information about which factors have influence in the output using a minimum number of designs. Like fractional factorial designs, interactions are lost in the modelling and only contribution of single factors can be quantified. Such designs, such, as Plackett-Burman have been for determination of important components in defined and rich culture medium (Srinivas et al., 1994; Nanthakumar et al., 2013) and for selection of process parameters (Agarabi et al., 2015). List of design of experiment approaches and literature examples in bioprocesses can be consulted in (Kasemiire et al., 2021; Mandenius & Brundin, 2008) and for biosurfactant in (Bertrand et al., 2018).

1.4. Modelling beyond polynomial models

Polynomial models are a powerful family of models that can be used to understand several quantitative biological processes. However, they have some limitations. While polynomial models can capture nonlinearity, they assume a specific

form of nonlinearity, which may not align with the true nature of the biological output to model (Mead, 1971). Thus, machine learning models have been employed as flexible models that can approximate outputs coming from different biological data structures.

Machine learning algorithms deal with the problem of analysing data, which is the input of the algorithm, and predict a property or number when new inputs are presented (Murphy, 2012). A dataset is normally written as:

$$\mathcal{D} = \{(\boldsymbol{x}_i, y_i) | i = 1, \ldots, n\}$$

with $\mathbf{x}_i$ an input point (with arbitrary number of dimensions), corresponding to a dependent variable and y_i the values or labels to predict. Thus, the goal of machine learning is to find a good predictor for y_i, i.e, a model f(**x**) that can "learn" to accurately estimate the values of y_i,, given the input data $\mathbf{x}_i$ and can predict unseen data with similar properties to the input (Ben-David & Shalev-Shwartz, 2014).

A machine learning model can be parametric, containing parameters similar to the coefficients in polynomial regression. Tweaking these parameters by following a statistical criterion generates the predictor and the relationship between the parameters is the base of the model structure (James et al., 2023). On the other hand, there are non-parametric models, which use the data to infer the appropriate structure for model. In both cases, training is the process by which the parameters or the structure of a model are determined as optimal given a train dataset and constitutes the learning part (Ben-David & Shalev-Shwartz, 2014).

Although these algorithms tackle a diversity of prediction tasks, two main classes are important for their discussion: classification and regression. Classification deals with assigning input to discrete outputs or category, known as labels. Thus, the goal of a classification algorithm is to predict which category or class an input belongs to, based on the features of that input (James et al., 2023. For example, in a medical diagnosis system, a classification algorithm might be used to determine whether a patient has a particular disease (e.g., "disease" or "no disease") based on symptoms and medical test results (Ahsan et al., 2022).

In contrast, regression is concerned with predicting continuous numerical values. Instead of assigning inputs to discrete categories, regression algorithms predict a value that can take any number within a range (Gelman et al., 2020). For instance, in a house pricing model, regression could be used to predict the price of a house based on features like square footage, number of bedrooms, and location.

Both classification and regression are supervised learning tasks, meaning they require a labelled dataset where the correct output (label or value) is known for each input (Hastie et al., 2009a). The algorithm learns from this data to make predictions on new, unseen data. While classification and regression are distinct in their goals, they often share similar underlying techniques and models, such as decision trees, neural networks, and support vector machines, which can be adapted to either type of task depending on the nature of the problem at hand (Hastie et al., 2009a). In contrast, unsupervised algorithms try to uncover patterns in the input data without predicting a label (Hastie et al., 2009b). These techniques often embed a dataset into a lower-dimensional representation, expecting to find distinctive patters and/or use distance-based techniques to analyse similarity between points (Izenman, 2008a, 2008b).

When talking about the best predictor for a given dataset, accuracy is important, but it is not everything to consider. Certain models can adjust to any dataset, but they lack predictive power for unseen data, and therefore there are not generalisable. This situation is called overfitting (James et al., 2023). Similarly, a model could be too “stiff” and do not represent the variation observed in the dataset, resulting in underfitting (Hastie et al., 2009a). Therefore, there is a trade-off, often called in the literature bias–variance trade-off, that guides the selection of the model, improving accuracy but choosing an appropriate complexity or number of parameters in the model (Briscoe & Feldman, 2011; Belkin et al., 2019; Doroudi, 2020; Fong & Holmes, 2020).

For this work, we will focus solely on regression, since we are looking to predict quantitative data obtained by metabolomics or optical density measurements. We will dive first into Gaussian processes, which are core to the modelling part of the thesis and then consider other machine learning approaches for titre modelling.

1.4.1. Gaussian processes regression

A Gaussian process is a stochastic process such that any finite collection of random variables, indexed by a continuous variable, should follow a multivariate Gaussian distribution (Rasmussen & Williams, 2006). In a more intuitive sense, these random variables can be seen as values of functions or infinite length vector over a continuous domain. Thus, Gaussian processes (GPs) are defined as a probability distribution over functions. This means that when we "draw" a sample from a Gaussian process, the outcome is not a single number or vector, but rather an entire function defined over the specified domain (Rasmussen & Williams, 2006) (Figure 1.4).

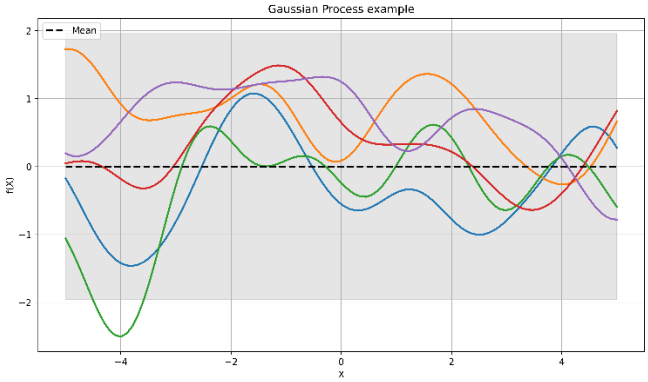


Figure 1.4. An example of a Gaussian process defined in a unidimensional domain (x). The Gaussian process was defined with a mean of 0 and the covariance function is the exponential kernel, also called the radial basis function. The colour curves are draws from this process.

This might initially suggest that working with such a model would be computationally impossible, given the infinite number of values involved. However, the Gaussian nature of the process is particularly powerful because it allows us to compute the distributions of interest exactly, using finite means (Rasmussen & Williams, 2006). Specifically, when we are concerned with the function values at a finite number of points, the joint distribution of these values is a multivariate Gaussian, characterized by a mean vector m(**x**) and a covariance function or kernel k(**x**,**x'**) (Rasmussen & Williams, 2006) and it is normally written as (Equation 1):

$$f(\boldsymbol{x}) \sim \mathcal{GP}(m(\boldsymbol{x}), k(\boldsymbol{x}, \boldsymbol{x'})) \qquad (1)$$

For Gaussian process regression (GPR), a GP is conditioned on observed data points, so we update our prior beliefs about the underlying function to form a posterior distribution that reflects both the observed data, and our assumptions encoded in the kernel function (Rasmussen & Williams, 2006). This posterior distribution is still a Gaussian process, but it is now centred around the observed data points, with uncertainty quantification that accounts for how far other points are from the observed data and how much the kernel function suggests they should be correlated (Rasmussen & Williams, 2006; Murphy, 2012) (Figure 1.5).

Mathematically, given a set of observed input-output pairs *D,*

$$\mathcal{D} = \{(\boldsymbol{x}_i, y_i) | i = 1, \ldots, n\}$$

and a Gaussian process prior with mean 0, it is possible to analytically obtain the mean prediction and variance predictions (Equation 2 & 3) when modelling the data using a GPR:

$$\mu(\boldsymbol{x}^*) = \boldsymbol{k}(\boldsymbol{x}^*)^\top \left[\boldsymbol{K} + \sigma^2 \mathbf{I}\right]^{-1} \boldsymbol{y} \tag{2}$$

$$\sigma^2(\boldsymbol{x}^*) = k(\boldsymbol{x}^*, \boldsymbol{x}^*) - \boldsymbol{k}(\boldsymbol{x}^*)^\top \left[\boldsymbol{K} + \sigma^2 \mathbf{I}\right]^{-1} \boldsymbol{k}(\boldsymbol{x}^*), \tag{3}$$

where $\boldsymbol{K}$ is the covariance matrix, such that $\left(\boldsymbol{K}\right)_{ij} = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$, $\boldsymbol{k}(\boldsymbol{x}^*) = [k(\boldsymbol{x}_1, \boldsymbol{x}^*), \ldots, k(\boldsymbol{x}_n, \boldsymbol{x}^*)]^\top$, and $\boldsymbol{x}^*$ is the point where the GP is evaluated.

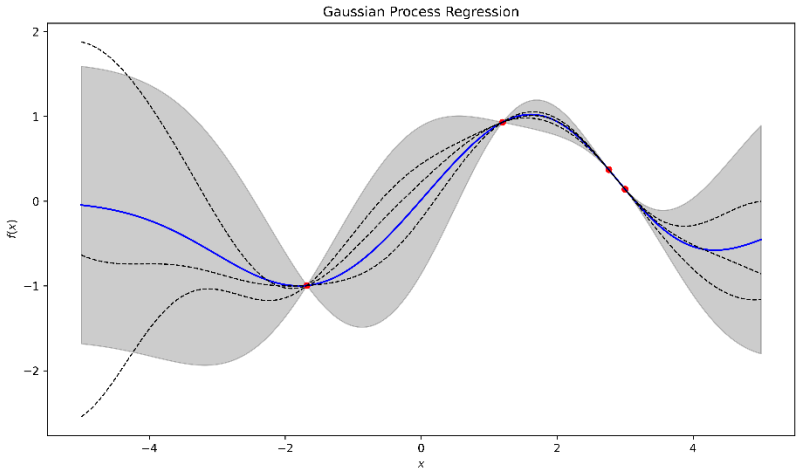


Figure 1.5. Gaussian process regression on a 1D dataset with 4 points. The dataset points are shown in red. The blue line indicates the mean prediction, the grey area represents a 95% confidence interval based on uncertainty prediction. The dashed lines show three samples of the conditioned process.

The covariance function, or kernel, indicated in the GP formulas as k(**x**,**x'**) (**x** and **x'** are two different points), plays a crucial role in Gaussian process regression. It captures the degree of similarity or correlation between those points (Duvenaud, 2014). The choice of this kernel function determines the characteristics of the functions that the Gaussian process models—such as smoothness, periodicity, and amplitude. Common kernels include the squared exponential or the Radial Basis Function (RBF) kernel, and the Matérn kernel, which allows to control model smoothness.

The squared exponential or the Radial Basis Function (RBF) kernel (Equation 4) is inspired by the normal distribution and can fit general datasets as a smooth approximation (Duvenaud, 2014). This kernel was used for the simulation in Figure 1.5.

$$k_{\mathrm{SE}}(\boldsymbol{x}, \boldsymbol{x}') = \sigma_f^2 \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{x}'\|^2}{2\ell^2}\right) \tag{4}$$

The more complex Matérn kernel (Equation 5) can deal with rougher approximations, with the presence of a shape or smoothness parameter *v* a modulator of the "roughness" of the output to estimate (Stein, 2012; Matern, 2013) (Figure 1.6).

$$k_{\mathrm{Matérn}}(\boldsymbol{x}, \boldsymbol{x}') = \sigma_f^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu}\|\boldsymbol{x} - \boldsymbol{x}'\|}{\ell}\right)^{\nu} K_\nu\left(\frac{\sqrt{2\nu}\|\boldsymbol{x} - \boldsymbol{x}'\|}{\ell}\right) \tag{5}$$

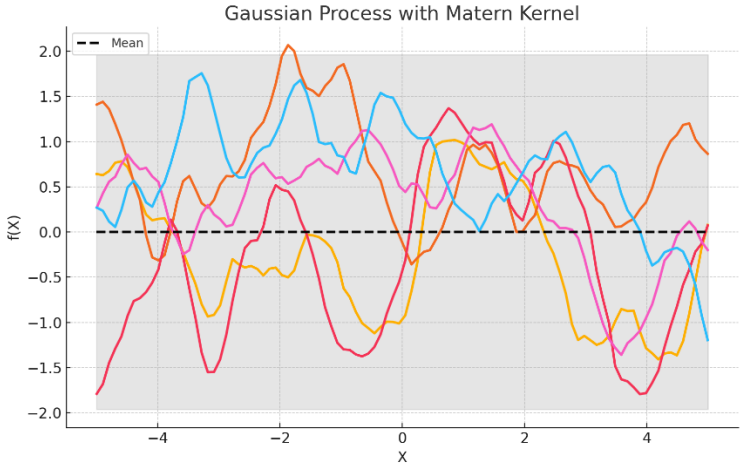


Figure 1.6. Gaussian process draws with a Matern kernel as covariance function. The Gaussian process was defined with a mean of 0 and the covariance function is the Matern kernel, with shape parameter v = 0.5. The colour curves are draws from this process.

An important parameter of kernels used in GPR is the lengthscale, which adjust the "granularity" of the model, i.e., smaller lengthscales means that the model is very fine-grained with many features in small zones compared to the domain, while larger lengthscales means that the model gets smoother with patterns comparable with the domain size. Very small lenghtscales can induce overfitting, while big ones can underfit the data. To choose reasonable lenghtscales, one can use a marginal likelihood approach to "condense" model complexity or by using a cross-validation method (Rasmussen & Williams, 2006)

Other kernels such as the rational quadratic kernel and the linear kernel can be seen as versions to perform Bayesian polynomial regression. Periodic function can be modelled better using a periodic kernel, involving the exponential sine function. Indeed, it is indicated in Rasmussen & Williams, 2006, that sophisticated kernels can "absorb" the details of diverse classes of data and developing new kernels for GPR is key for the expressivity of these models.

This ability to model uncertainty is one of the strengths of Gaussian process regression. In particular, the posterior variance allows us to identify regions of the input space where the model is uncertain and could benefit from additional data, a property that is exploited in strategies such as active learning and Bayesian optimization (Shahriari et al., 2016; Snoek et al., 2012). Another important property is that is non-parametric in the sense of the model (except for the kernel parameter), because the structure of model is inferred from the observed points (following the kernel information) (Murphy, 2012).

In practice, the computational demands of Gaussian processes can be significant, particularly as the size of the dataset grows. The computation of the covariance matrix, which involves an operation on the order of $O(n^3)$ for n data points, can become a bottleneck (Rasmussen & Williams, 2006). However, data sets with dimensions not exceeding the thousands, which are typically seen in physical

experiments, GPR models can be computed in reasonable time in home computers (Rasmussen & Williams, 2006). Additionally, various global (such as sparse approaches) and local approximation methods have been developed to make Gaussian processes more scalable (Liu et al., 2020). These methods allow the use of Gaussian processes in large-scale applications while retaining much of the model's inherent flexibility and uncertainty quantification capabilities.

1.4.1.1 Heteroskedastic Gaussian process regression

Gaussian process regression can be adapted to heteroskedastic noise in experimental outputs (Goldberg et al., 1997). Heteroskedasticity means that observed variance varies depending on the inputs, and it is a common feature observed in biological data (Cleasby & Nakagawa, 2011). To achieve the goal of modelling heteroskedastic data using GPs, two Gaussian processes can be used, one for the mean and for the variance, generating a hierarchical model. The prior of both GPs can be handled jointly and a posterior for the variance or noise can be sampled using Monte Carlo techniques (Goldberg et al., 1997; Balandat et al., 2020).

1.4.2. Other machine learning algorithms and cross-validation

1.4.2.1. Supervised algorithms

Supervised learning algorithms form the backbone of machine learning, wherein models are trained on labelled data to make predictions or classifications. Linear regression, one of the simplest and most well-known supervised learning algorithms, aims to model the relationship between a dependent variable (vector) **y** and one or more independent variables **X** as a matrix, by fitting a linear equation to the observed data (Christensen, 2020). The fundamental equation for simple linear regression (Equation 6), where there is only one independent variable, can be expressed as:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{6}$$

Here, **y** represents the dependent variable, **β** is the coefficient vector for the independent variable (matrix) **X**, and epsilon is the error term, which accounts for the deviation of the observed data from the predicted line.

While linear regression provides an easy-to-interpret model for understanding linear relationships, it is not always adequate for capturing complex, non-linear patterns in the data. To address this limitation, more sophisticated supervised learning algorithms are employed. We will examine a specific model that will be used in Chapter 4: random forest. A random forest is a model that builds multiple decision trees and aggregates their predictions to improve the accuracy and robustness of the model (Breiman, 2001; Hastie et al., 2009). The prediction for a new instance in a random forest for regression (Equation 7) is given by:

$$\hat{y} = \frac{1}{T}\sum_{t=1}^{T} h_t(\mathbf{x}) \tag{7}$$

where is the prediction of the t-th decision tree, and T is the total number of trees in the forest (Hastie et al., 2009). The prediction given by each decision tree, h(**x**) (Equation 8) can be derived from the tree structure (Hastie et al., 2009):

$$h_t(x) = \sum_{l=1}^{L_t} c_{tl} \cdot I\left(x \in R_{tl}\right)$$

The performance of a random forest is influenced by various hyperparameters, which are parameters not learned from the data but set before the learning process begins. Key hyperparameters in random forests include the number of trees T, the maximum depth of each tree, the minimum number of samples required to split a node, and the number of features considered for splitting at each node. The process of selecting the optimal values for these hyperparameters, known as hyperparameter tuning, is paramount for enhancing the model's performance and ensuring it generalizes well to unseen data (Bernard et al., 2009; Probst et al., 2019).

Hyperparameter tuning is often performed using techniques like grid search, where a predefined set of hyperparameters is exhaustively searched, or random search, where hyperparameters are randomly sampled from a distribution. More advanced methods, such as Bayesian optimisation, can also be employed to explore the hyperparameter space more efficiently by modelling the relationship between hyperparameters and model performance (Yang & Shami, 2020).

Linear regression and random forests are just two examples of the variety of machine learning, that also includes kernel methods such as support vector machines, gradient boosting, logistic regression, among others (Ray, 2019). Each of them has a specific internal structure that accommodates to certain characteristics of the dataset to model. Small neural networks provide a link between classical machine learning algorithms, which usually consist of a small or medium of parameters, and large parametric models involving neural networks with multiples layers and the field of deep learning (Goodfellow et al., 2016).

1.4.2.2. Cross-validation

Cross-validation is a statistical method used to estimate the generalizability of a predictive model (James et al., 2023). It serves to mitigate the risk of overfitting, ensuring that the model performs well not just on the data it was trained on, but also on new data. This technique is particularly important in scenarios where the goal is to make the most out of limited data samples.

The crux of cross-validation involves partitioning the available data into subsets, performing the analysis on one subset (known as the training set), and validating the analysis on the other subset (referred to as the testing set) (James et al., 2023). To reduce variability, multiple rounds of cross-validation are performed using different partitions, and the validation results are averaged over the rounds (James et al., 2023).

One of the most common methods of cross-validation is k-fold cross-validation (Figure 1.7). In k-fold cross-validation, the original sample is randomly partitioned into k equal sized subsamples. Of the k subsamples, a single subsample is retained as the validation data for testing the model, and the remaining k-1 subsamples are used as training data. The cross-validation process is then repeated k times, with each of the k subsamples used exactly once as the validation data. The k results from the folds can then be averaged to produce a single estimation (Refaeilzadeh et al., 2009).

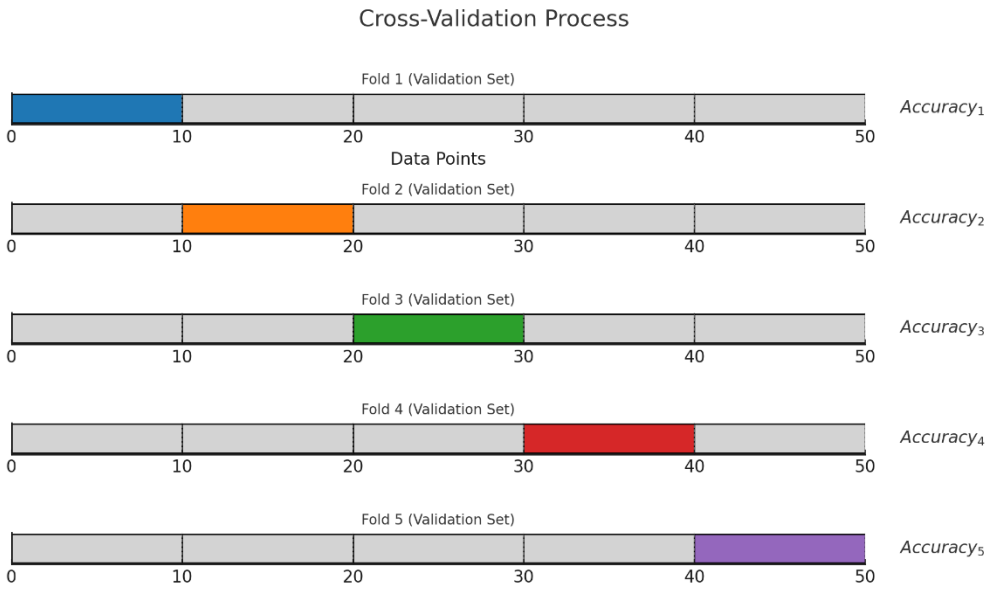


Figure 1.7. A visualisation of k-fold cross-validation. Suppose that you have 50 samples in your dataset and 5-fold cross-validation is required. The dataset is partitioned into 5 subsets and in the first instance, the model is trained using 4 out of these 5 subsets and tested on the remaining subset. This protocol is repeated, until all subsets have been used as the test set once, and the accuracies can be summarised in terms of mean accuracy and corresponding variance.

The advantage of k-fold cross-validation is that all observations are used for both training and validation, and each observation is used for validation exactly once. Variants of cross-validation, such as stratified and leave-one-out cross-validation, are used to address specific problems of model sensitivity and data variability. Stratified cross-validation is often used to ensure that each fold of the dataset contains roughly the same proportion of samples of each target class as the complete set (Refaeilzadeh et al., 2009). In contrast, leave-one-out cross-validation (LOOCV) is a special case of cross-validation where the number of folds equals the number of instances in the dataset, which means that each fold contains only one instance. This method, while expensive in terms of computation, is very useful when the dataset is limited in size (Refaeilzadeh et al., 2009).

1.4.2.3. Unsupervised algorithms

Unsupervised algorithms are designed to analyse input data that lacks predefined labels or target outcomes. These algorithms explore the underlying

structure of the data without any prior knowledge of the categories or values they might contain (Hastie et al., 2009). A key subset of unsupervised learning techniques is dimensionality reduction, which seeks to reduce the number of variables under consideration by creating a lower-dimensional representation of the data (Hastie et al., 2009).

Dimensionality reduction techniques, such as Principal Component Analysis (PCA) and t-Distributed Stochastic Neighbour Embedding (t-SNE), serve multiple purposes. They can simplify data for easier visualization, enhance computational efficiency, and help eliminate noise by focusing on the most informative aspects of the data (Hastie et al., 2009). These techniques act like compression algorithms, condensing the data into a more compact form that retains the essential patterns and relationships.

1.4.2.3.1. Principal component analysis (PCA)

Principal Component Analysis (PCA) is a powerful statistical technique used in data analysis and machine learning for dimensionality reduction. By reducing the number of variables while retaining as much linear information as possible, PCA simplifies the complexity of the data, making it easier to analyse and interpret without significant loss of information (Hotelling, 1933; Jolliffe, 2002; Izenman, 2008).

The core idea behind PCA is to transform the original data into a new coordinate system. These new coordinates, called principal components, are orthogonal, meaning they are uncorrelated with each other. Essentially, PCA identifies the directions or components in which the variation in the data is maximized (Deisenroth et al., 2020).

PCA begins by standardizing the data, a crucial step since PCA is sensitive to the scale of the variables (Härdle & Simar, 2019). After standardization, the next step is to compute the covariance matrix, which captures the relationships (covariances) between pairs of variables in the data. For a dataset matrix **X** with *n* observations and *p* variables, the covariance matrix **C** is calculated as (Deisenroth et al., 2020) (Equation 9):

$$\mathbf{C} = \frac{1}{n-1}\mathbf{X}^\top \mathbf{X} \tag{9}$$

The covariance matrix **C** is a *p* x *p* symmetric matrix, where each element C_{ij} represents the covariance between the i-th and j-th variables.

The key to PCA is solving the eigenvalue problem for the covariance matrix. This involves finding the eigenvalues λ and the corresponding eigenvectors ***v*** that satisfy the equation C **v** = λ **v** (Ben-David & Shalev-Shwartz, 2014; Deisenroth et al., 2020). The eigenvectors represent the directions of the principal components, while the eigenvalues indicate the amount of variance explained by each principal component (Johnson & Wichern, 2007). The eigenvectors and their corresponding eigenvalues are then sorted in descending order of the eigenvalues. The eigenvector associated with the largest eigenvalue becomes the first principal component, the one associated with the second largest eigenvalue becomes the second principal component, and so forth (Ben-David & Shalev-Shwartz, 2014; Wang, 2012).

Once the principal components are identified, the original standardized data is projected onto the new coordinate system defined by these components (Figure 1.8). The resulting data matrix, often referred to as the scores or principal components matrix, is obtained by multiplying the standardized data matrix by the matrix of eigenvectors **V** (Deisenroth et al., 2020): **Z=X V**, where **Z** represents the data in the reduced dimensional space, and **V** is the matrix of eigenvectors.

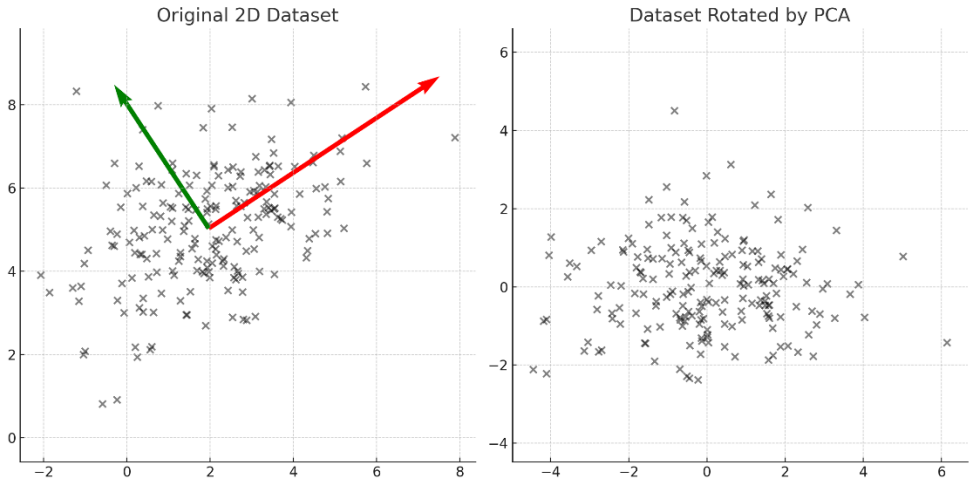


Figure 1.8. PCA rotates a dataset when performed in 2 dimensions. The red and green vectors indicate the (orthogonal) directions of biggest variance in the dataset. These vectors are obtained from the eigenvectors of the covariance matrix. Then, the points are projected into these vectors to effectively rotate the data such that new axes are the principal components.

An important aspect of PCA is determining how much of the total variance in the data is captured by each principal component. The variance explained by each principal component is calculated as the ratio of its corresponding eigenvalue to the sum of all eigenvalues. (Johnson & Wichern, 2007). Thus, the cumulative variance explained by the first few principal components is often used to decide how many components to retain.

PCA is widely used in various applications. In data visualization, PCA can reduce the dimensionality of data to two or three principal components, allowing for easy visualization of complex datasets and revealing underlying patterns or clusters, by generating uncorrelated features (Deisenroth et al., 2020). In noise reduction, PCA can help by retaining only the principal components that explain most of the variance, effectively filtering out noise (though some experimentally relevant information might be lost in the process) (Ben-David & Shalev-Shwartz, 2014). It is also used in feature extraction, where the most important features from the data are extracted and used in subsequent analysis or machine learning models (Katsaggelos et al., 2016). However, PCA is not without limitations. It assumes that the expected relationships between variables are linear, which may not hold for the dataset in mind, and that

would mean that no significant patterns would manifest in the reduction (Lever et al., 2017). Additionally, the principal components are linear combinations of the original variables, which can make them difficult to interpret. This problem may be alleviated by employing a biplot visualisation.

1.4.2.3.2. PCA and biplots

Related to the problem of interpreting the components, a biplot is a graphical representation that simultaneously displays the information from both observations and variables of a multivariate dataset in a single plot by combining scatter plot with vector representations, thereby providing a comprehensive view of the structure within the data (Gower et al., 2011) (Figure 1.9).

The biplot was first introduced by Gabriel in 1971, and its development was primarily associated with principal component analysis (PCA) (Gabriel, 1971). However, it can also be used in the context of other multivariate techniques, such as multidimensional scaling, correspondence analysis or canonical variate analysis.

To construct a biplot, the multivariate data is first subjected to dimensionality reduction, usually through PCA. As described before, PCA transforms the original data into a set of orthogonal components (principal components) that capture the maximum variance within the data and are linear combinations of the original variable. The first two or three components are usually selected as the axes for the biplot, as they typically account for most of the variance, and because of the natural limitations of visualisation in higher dimensions (Deisenroth et al., 2020).

In a biplot, the observations are represented as points, and the variables are depicted as vectors (arrows). The position of a point along the axes reflects its values with respect to the selected principal components, indicating the similarities or dissimilarities between observations. Points that are closer to each other represent observations with similar patterns across the variables, while those further apart are dissimilar (Gower et al., 2011).

The vectors, also known as loadings, representing the variables are essential for interpreting the biplot. The direction of a loading indicates the orientation of a variable in the principal component space, and the length of the loading reflects the

magnitude of the variable's contribution to the principal components (Gower et al., 2011). Variables that have loadings pointing in the same direction are positively correlated, while those pointing in opposite directions are negatively correlated. The angle between two vectors gives an indication of the correlation between the variables they represent; smaller angles suggest stronger correlations (Greenacre, 2010; Gower et al., 2011).

The projections of the points onto the loadings provide insight into the relationship between observations and variables. An observation that projects strongly onto a particular vector suggests that the corresponding variable has a significant influence on that observation (Greenacre, 2010; Gower et al., 2011). Similarly, the proximity of an observation to the origin in the biplot indicates that it has average or typical values for the variables under consideration.

Interpretation of a biplot requires caution, particularly when the explained variance by the chosen principal components is not high. In such cases, the projection might distort the true relationships, and important information might be lost (Greenacre, 2010). However, when the first few principal components capture a substantial amount of the variance, the biplot serves as a powerful tool for visualizing the relationship between components and original variables in a more comprehensible form.

Figure 1.9. PCA biplot on the Iris dataset. After taking the first two components in the PCA of the iris dataset, these components can be interpreted as linear combinations of the original variables (sepal length, sepal width, petal width and petal length). The last three of these variables has loadings pointing in almost the same direction, meaning that they are correlated and are important contribution to the first principal component. On the other side, sepal width is the major (linear) contribution to the second component and could be attributed as the differentiating variable that separates the *setosa* species form the other iris species.

1.5. Bayesian optimisation: an adaptive experimental design for optimisation

Bayesian optimisation (BO) is a method for searching for the global maximum (or minimum) value of an unknown, black-box function, when no gradient information is known a priori (Močkus, 1975; Shahriari et al., 2015; Garnett, 2023). It is a sequential experimental design that begins with an initial set of points in a design space, and then iteratively selects the next point in this space that is most likely to maximise the function based on a surrogate model (Snoek et al., 2012; Shahriari et al., 2015; Frazier, 2018; Wang et al., 2022) (Figure 1.10). The surrogate model is updated at each iteration using the data collected from the previous points and mimics the true function based on the available data (Shahriari et al., 2015). Bayesian optimisation has been utilised in a variety of applications including robotics and is currently a popular method for optimising machine learning algorithms (Wang et al., 2022; Bai et al., 2023). There are several software packages that implement Bayesian optimisation,

including BayesOpt (Martinez-Cantin, 2014), Bayesian-Optimisation (Nogueira, 2014), GPyOpt (GPyOpt, 2016), and Botorch (Balandat et al., 2019), among others.

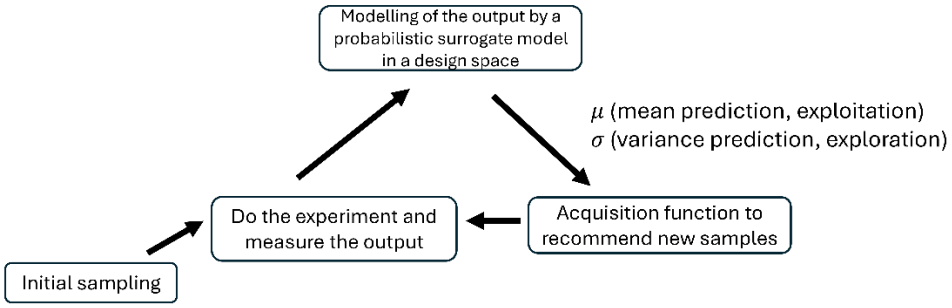


Figure 1.10. A diagram of a typical Bayesian optimisation loop. There are 4 main steps in Bayesian optimisation: generate an initial sampling/combinations of factors, measuring these samples, model the output data using a probabilistic surrogate, and use the mean and uncertainty prediction from this model to generate new samples/factor combinations to test in the next iteration. The cycle stops until no further improvement is achieved (criterion can be specified).

1.5.1. Initial sampling

Examining the different parts of the Bayesian optimisation loop, the method starts with an initial sampling of the factors/inputs of the experiments. A natural choice would be performing a random sampling in the design space. However, random sampling has a particular drawback that the samples are not “equidistributed” in multidimensional space, but rather it exhibits clumps and zones with low number of samples (Jäckel, 2002). Therefore, space-filling designs are considered, as they can ameliorate these both problems. Inside the category of space-filling designs, Sobol sequences are particularly popular. Sobol sequences are a type of quasi-random sequence used in the field of numerical integration and experimental design (Sobol’, 1967). Unlike purely random sampling methods, Sobol sequences ensure that the sample points are more evenly distributed across the entire experimental space (Jäckel, 2002). It achieves this property by ensuring low discrepancy, where the discrepancy is a measure that basically quantifies if the number of points in an arbitrary set, which can be a volume defined in the design space, differ to the measure of a set. When this number is small, one can say the number of points is approximately

proportional to the volume measure and equidistribution is ensured (Keng & Yuan, 1981; Jäckel, 2002).

Apart from these sequences, there are others space-filling design such as Latin Hypercube design (LHD) that can provide homogeneous sampling on the design (McKay et al., 1979). The LHD has an analogy to the problem of positioning rooks in a chess board without any of them threatening another. If you have *n* rooks in a chess board, a configuration of the rooks where none of the attacking paths intersect each other effectively distributes the rooks in an even way in the board (Santner et al., 2018). This idea can be generalised by partitioning the design space in a grid with size depending on the number of samples to draw and finding this optimal configuration, which correspond to find a Latin square. This can be done by well-known algorithms (Santner et al., 2018). Since the design space is continuous, the samples can be defined in the centre of the selected "cubes" in the (multi-dimensional) grid or in a random position in the "cubes". Both Sobol and LHD sampling have been successfully used in initial sampling for Bayesian optimisation protocols (Bossek et al., 2020) (Figure 1.11).

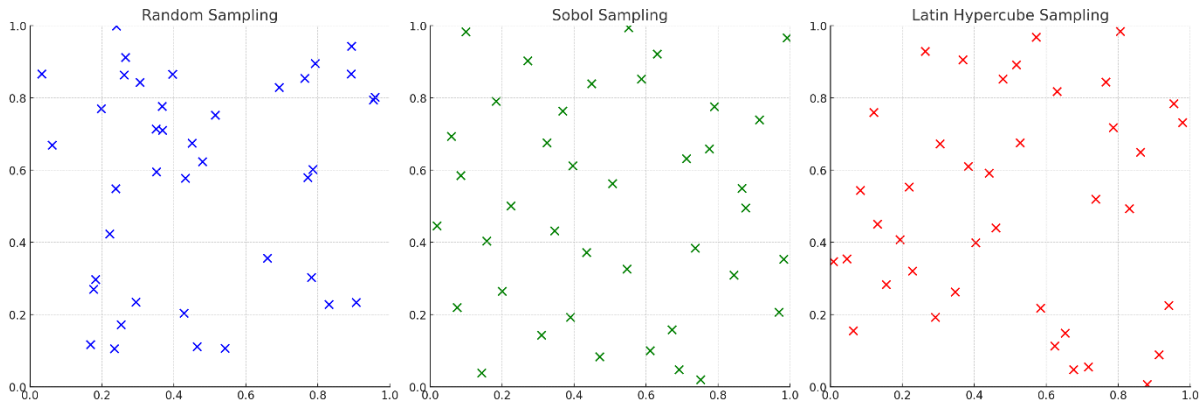


Figure 1.11. A comparison between random sampling, Sobol sampling and Latin hypercube sampling in a 2D design space. 40 samples were draws from these sampling algorithms. Random sampling tends to generate clumps and zones of the design space with low number of samples, while Sobol sampling and Latin hypercube sampling cover the space more homogenously.

1.5.2. Probabilistic surrogate modelling

Then after measuring the outputs for this initial set of points, a surrogate model is employed to generate prediction over the whole design space based on the measured points. For this surrogate, any probabilistic model can be used, being Gaussian process regression one of the standard choices (and it also the overall choice in the project) (Brochu et al., 2010; Snoek et al., 2012; Shahriari et al., 2016; Frazier, 2018), but Bayesian ensemble models and Bayesian neural networks have been also extensively in the literature. These models can fit the data and provide a mean prediction for the values of the outputs in observed. Additionally, it considers a model of noise, which can be defined explicitly in the model or can be inferred from replicate data (Balandat et al., 2019). Both the mean prediction and uncertainty prediction coming from the noise model are important to understand how much information is obtained from the design space, and to determine which points are promising to test further.

1.5.3. Acquisition functions

To formalise this idea, an acquisition function can be calculated using these prediction terms. The acquisition function is a mathematical formula that balances exploration, represented by the mean prediction of the model in the domain space, and exploitation, given by the uncertainty prediction of the model (Snoek et al., 2012; Frazier, 2018). The mean prediction provides promising points to test later, since it would be expected that points with high function value (in case of maximisation) might be closer to previous high value points. However, it might happen that maximum of the function is not located nearby of any of the previous tested points. This information is given by the variance prediction, since it will be higher in under sampled zones of the design space (Frazier, 2018).

1.5.3.1 Upper confidence bound.

The simplest acquisition is Upper Confidence Bound (Equation 10), which is simple a linear combination of the mean prediction and the variance prediction (Snoek et al., 2012; Garnett, 2023):

$$\mathrm{UCB}(x) = \mu(x) + \beta\sigma(x) \tag{10}$$

where:

- μ(x) is the predicted mean at point **x**,
- σ(x) is the predicted standard deviation at point **x**,
- β is a parameter that controls the balance between exploration and exploitation.

In this formula, β determines how much importance is placed on exploration. A higher β favours exploration by placing more weight on the uncertainty σ(x), while a lower β favours exploitation by focusing more on the predicted mean μ(x).

1.5.3.2. Probability of Improvement (PI)

The Probability of Improvement (PI) acquisition function (Equation 11) focuses on the probability that a given point **x** will improve upon the best observed value y_{best} = f(***x***$^+$) so far. This function is particularly useful when the primary objective is to simply increase the likelihood of finding better points rather than balancing the trade-off between exploration and exploitation explicitly (Kushner, 1964; Shahriari et al., 2016; Garnett, 2023).

The PI is defined as:

$$\mathrm{PI}(x) = \Phi\left(\frac{\mu(x) - f(x^+) - \xi}{\sigma(x)}\right) \tag{11}$$

Where μ(x) is the predicted mean at point, σ(x) is the predicted standard deviation at point **x**, f(***x***$^+$) is the best observed value so far, Φ (·) is the cumulative distribution function (CDF) of the standard normal distribution, and ξ is a small positive constant that adds exploration by making the function less greedy.

The PI acquisition function tends to favour exploitation, especially when ξ is small, as it emphasizes points with a high probability of exceeding f(***x***$^+$). However, by tuning ξ, the balance can be adjusted to introduce more exploration.

1.5.3.3. Expected Improvement (EI)

A modified version of the PI acquisition function is the **Expected Improvement (EI)** acquisition function (Equation 12 & 13), one of the most used in practice. It not only considers the probability of improvement but also the expected magnitude of that improvement.

The EI is defined (in the case of a Gaussian process regression surrogate) as (Snoek et al., 2012; Garnett, 2023):

$$\mathrm{EI}(x) = \mathbb{E}\Big[(f(x) - f(x^+))_+\Big] = \big[\mu(x) - f(x^+) - \xi\big]\,\Phi(Z) + \sigma(x)\,\phi(Z) \qquad (12)$$

$$Z = \frac{\mu(x) - f(x^+) - \xi}{\sigma(x)} \qquad (13)$$

using the same definitions as in the PI acquisition function. $\phi(\cdot)$, which did not appear in the PI function, is the probability density function (PDF) of the standard normal distribution.

The first term of the EI formula represents the improvement over the best observed $f(\boldsymbol{x}^+)$ scaled by the probability that the improvement occurs, while the second term accounts for the variability around the predicted mean. Similar to PI, this acquisition function is flexible and can be tuned to encourage more exploration or exploitation by adjusting the ξ parameter.

1.5.3.4. Thompson Sampling

Thompson Sampling is a different approach compared to the previous acquisition functions, since it does not consider a mathematical formula dependent of the mean and variance prediction. Rather, it involves drawing a sample from the posterior distribution over the objective function and selecting the next point that maximizes this sampled function (Thompson, 1933; Russo et al., 2018).

Thompson Sampling naturally balances exploration and exploitation by selecting different functions from the posterior distribution in each iteration. The inherent randomness ensures that the algorithm explores different areas of the design space, and it has proven particularly effective when performing Bayesian optimisation for a large number of factors/dimensions (Pleiss et al., 2020).

1.5.3.5. Comparison and Application

Each of these acquisition functions has its own strengths and is suitable for different scenarios:

- **UCB** is simple and intuitive, providing a clear trade-off between exploration and exploitation. It's particularly useful when there's a need for a transparent and interpretable method.
- **PI** is straightforward and easy to compute, focusing on maximizing the probability of improvement, which can be useful in scenarios where exploitation is highly desirable.
- **EI** provides a more nuanced approach by considering the magnitude of potential improvements, making it effective in balancing exploration and exploitation.
- **Thompson Sampling** offers a probabilistic method that can explore the design space efficiently, particularly in complex or high-dimensional problems.

Choosing the right acquisition function often depends on the specific problem and how explicitly the balance between exploration and exploitation must be controlled. Experimentation and cross-validation can help determine which acquisition function works best for a given scenario (Snoek et al., 2012). Nevertheless, the reviewed acquisition functions have theoretical guarantees, in terms that the global optimum can be approximated within a margin of error (also known as “regret”) in a finite number of iterations (Auer et al., 2002; Bull, 2011; Agrawal & Goyal, 2012; Russo & Roy, 2016). This list of functions is by no means exhaustive, with more option in the literature, such as Knowledge Gradient (Frazier et al., 2008), the recent log Expected Improvement (Daulton et al., 2024), and neural acquisition function (Volpp et al., 2019), being an active line of research in Bayesian optimisation (Wang et al., 2022).

1.5.3.6. Optimisation of the continuous acquisition function

After defining an acquisition function, the next sample or recommended sample to test in the next iteration can be found by maximising this function using

standard continuous methods, since it is continually defined over the design space. A typical approach is to choose quasi-Newton methods such L-BFGS-B (Nocedal & Wright, 2006; Balandat et al., 2020) or evolutionary algorithms such as NSGA-II (Cowen-Rivers et al., 2022), but other algorithms have been used specially in high-dimensional, rugged cases (Wilson et al., 2018; Daulton et al., 2024; Song et al., 2024). The recommendation for the next sample in the loop is therefore the point where the acquisition function is maximised.

Bayesian optimisation routines can be tested *in silico* before implementing them in an experimental setting, to guarantee a minimum performance based on the chosen hyperparameters in the loop. For this purpose, several functions are used as benchmarks for testing new BO approaches. These functions differ in the number of input dimensions and the complexity of their landscape, i.e., some of them present a clear maximum (or minimum point) with a small number of local maxima (or minima), while others exhibit a very rugged landscape with many critical points, making them a challenge for optimisation. The Branin function (number of dimensions, d=2) and Hartmann function (number of dimensions, d=6) are common synthetic functions employed for optimisation benchmarks, since they possess several global maxima (minima) in different zones of the space (Dixon & Szegö, 1975). As an example of a rugged counterpart, the Ackley function is useful to test the robustness of the algorithm in terms of avoiding getting stuck in a local minimum for several iterations (Ackley, 1987; Bäck, 1996) (Figure 1.12).

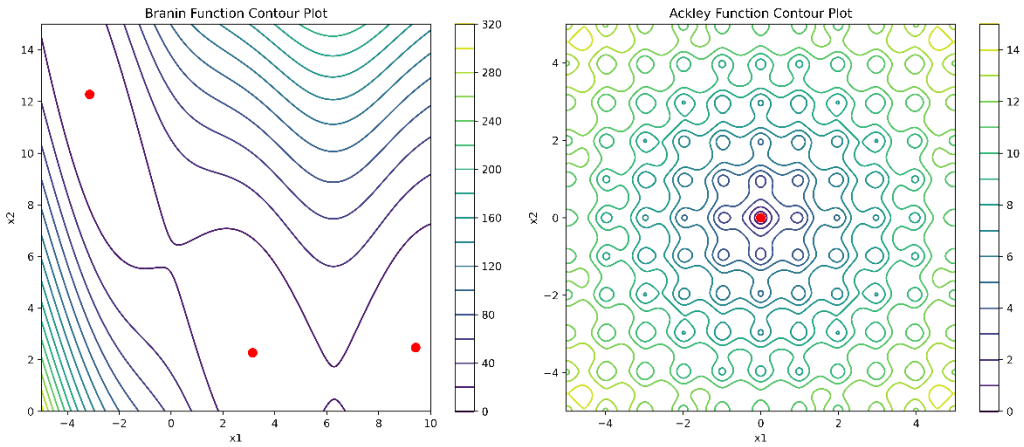


Figure 1.12. Branin function and Ackley function in a 2d space as contour plots. The values of the function can be read from the colour of the level lines, and the global minima is depicted with a red point.

1.5.4. Stopping of a Bayesian optimisation loop.

Stopping a Bayesian optimisation loop relies on a criterion, that will depend on how much effort the experimenter is willing to put to improve the optimal point on subsequent iterations. If the optimum is known beforehand, such as the case of the previously shown benchmark function, one can calculate the regret, which is the best value observed at an iteration subtracted by the optimal value (Wilson, 2024). Then, a threshold is defined to stop the loop if the regret gets below a certain value.

If the optimal value is not known, it is possible to define a similar threshold, but in this case, it must be calculated over the difference between best values at contiguous iterations. Indeed, it could be any measure that indicates the amount of effort needed for running more interactions versus the observed recent gains (Wilson, 2024). Eventually, one can restrict the number of iterations based if the design space has been adequately sampled.

1.5.5 Batch Bayesian optimisation.

This method of maximising the acquisition function works if only one sample is tested per iteration. However, in many applications, we would like to recommend several new points, so we can get more information of the black-box function on an iteration and thus, obtain better recommendations in the next one. Considering this, we would like this batches of recommendations to be as informative as possible. In

this regard, recommended batch should try the mimic the sequential (one-by-one) recommendation policy in terms of gained information (Gonzalez et al., 2016). One of the first approaches to achieve this was to find the maximum and then penalise the acquisition function in that area, such that maximising again would yield a different point, and the process of penalising is repeated until the batch is formed (Gonzalez et al., 2016). Subsequent approaches employed different ways to attack the problem of obtaining optimal batches from optimizing joint acquisition functions that consider all points in the batch simultaneously (Balandat et al., 2020) to using methods that diversify the selected points based on uncertainty or mutual information criteria (Nava et al., 2022). These approaches generate intractable integrals that must be approximated using, e.g., Monte Carlo approaches (Balandat et al., 2020).

1.5.6 Uses of Bayesian optimisation routines in biology.

In recent years, there has been a development of active learning/Bayesian optimisation protocols for chemistry and materials science, leading the way in integrating advanced artificial intelligence techniques, such as large language models, into automated pipelines (Griffiths et al., 2020; Eyke et al., 2020; Eyke et al., 2021; Wang et al., 2021; Griffiths et al., 2022; Hickman et al., 2022; Jorayev et al., 2022; Wang et al., 2022; Kumar et al., 2023; Basha et al., 2023; Ranković et al., 2024). However, the utilization of active learning for experimental design in quantitative biology experiments is still emerging. An early instance involves the suggestion to employ BO for synthetic gene design (González et al., 2015). Subsequently, BO has been applied to automate large-scale optimization of tryptophan production via metabolic engineering (Radivojević et al., 2020; Zhang et al., 2020) and for the high throughput in silico design of synthetic circuits (Merzebacher et al., 2023). In cell-free systems, BO has been utilized to enhance green fluorescent protein (GFP) production by simultaneously optimizing chemical components in the system (Borkowski et al., 2020). Recently, a code-friendly framework has been introduced for the general optimization of genetic and metabolic networks, demonstrating improvement in various cell-free systems' titre and increased productivity and efficiency of a synthetic carbon fixation cycle (Pandi et al., 2022). Particularly, for media optimisation, enhancing production using active learning in mammalian cells has been reported in

the last couple of years (Cosenza et al., 2022; Cosenza et al., 2023; Hashizume et al., 2023).

1.6. Metabolomics

For several biotechnological solutions, it is necessary to know the range of molecules present in an organism or, more generally, in a relevant biological sample. Thus, it is natural to ask if there is a general method to capture this catalogue of molecules comprehensively. The field addressing this problem is called metabolomics and is defined as the large-scale study of small molecules, with a mass range of 50–1500 Da (mass range might vary in the literature) (Hollywood et al., 2006; Kuehnbaum & Britz-McKibbin, 2013; Wishart, 2013), commonly known as metabolites, within organisms or other samples of biological origin (Johnson et al., 2016; Liu & Locasale, 2017). The outcome of detecting and measuring these molecules, known as the metabolome (Wishart, 2019), is an instantaneous picture of the metabolism present in a bio sample. It has been used successfully as both a milestone for basic physiological studies (Wishart, 2019) and as a decision-making tool for biotechnology and bioengineering, including agricultural, ecological and medical (German et al., 2005; Danzi et al., 2023) applications, among others. It is said that together with proteomics, metabolomics applied to organisms provides the closest biological information to the actual phenotype (Guijas et al., 2018), and therefore provides a window to probe and analyse biochemical processes at a molecular scale.

The impact of the field can be equated with its technological development, where the overcoming of experimental, computational and statistical hurdles has been a central part over the last decades (Miggiels et al., 2019; Ebbels et al., 2023). This improvement allowed to identify and quantify a vast range of molecules in complex samples coming from different bio-sources and material substrates. However, metabolomics still faces the challenge of a “dark matter” of molecules that cannot detected or identified due to the explicit selection of a specific analytical chemistry approach to measure the metabolome (da Silva et al., 2015). Thus, a single metabolomics protocol can only reveal a small portion of the total molecules present. This fundamental limitation has been a driver of metabolomics advances, and it is

expected that the field will keep evolving towards higher coverage, higher resolutions, higher sensitivities and consistent analysis pipelines (Ghafari & Sleno, 2024).

Metabolomics is often subdivided into two approaches. The first one is called targeted metabolomics. It is based in the principle that the sample is analysed by extracting information about a specific list of compounds (Roberts et al., 2012; Costanzo et al., 2022). Thus, a list of masses or chemical shifts is defined to build a database of metabolites to search. In contrast, the second approach, non-targeted or untargeted metabolomics, does not presuppose any list of metabolites to search and aims to span the largest number of them (Gertsman & Barshop, 2018). However, that also means, that most detected molecules will not match any entry on available compound libraries, hindering further annotation (Gertsman & Barshop, 2018). Thus, in comparison, a targeted approached is preferred when there exists a specific hypothesis about biochemical processes and metabolites involved and the goal is to quantify these metabolites of interest (Zhou & Yin, 2016). On the other side, an untargeted approach has a discovery component, where the focus is in obtaining information of possible unreported molecules and mechanisms, as in the case with biomarkers (Ribbenstedt et al., 2018), or to provide a global perspective of the metabolome in the experimental context (Sévin et al., 2015).

1.6.1. Technological foundations

Regarding analytic techniques, the investigation of the metabolome can be performed by various techniques, being mass spectrometry (MS) and nuclear magnetic resonance (NMR) the main ones in terms of use in academia and industry (Azad & Shulaev, 2019). Each of these techniques works by a specific physical principle and as technology has specific advantages and drawbacks moulding the current metabolomics research.

1.6.1.1. Mass spectrometry

Mass spectrometry (MS) is based in the drift and/or control of ion trajectories by electric or magnetic forces, which it is dependent of the ion's mass-to-charge ratio (m/z) (Glish & Vachet, 2003). The resulting manipulated ion stream is measured by a detector, and the data is used to determine the molecular composition of the sample,

providing information on the mass of the molecules present and further properties if this mass can be matched with a MS compound library (Milman & Zhurkovich, 2016).

Mass spectrometry has some advantages including its ability to detect a wide range of molecules with high sensitivity and specificity (Silberring & Smoluch, 2019). Additionally, the technique allows for the analysis of small sample quantities, in the micro to nanoliter order (Ho et al., 2003), and can process multiple samples in shorter times, enabling high-throughput analysis (Dueñas et al., 2023; Williams et al., 2024). This makes mass spectrometry an invaluable tool not only in metabolomics, but also in fields such as proteomics, and drug discovery, where rapid and accurate analysis of complex mixtures is essential (Sinha & Mann, 2020; Babele & Yadav, 2023; Williams et al., 2024). In addition, mass spectrometry equipment can be combined in tandem (MS/MS) providing fragmentation fingerprints of metabolites that can be used later for enhanced identification (Heiles, 2021; Neagu et al., 2022; Thomas et al., 2022).

1.6.1.2. Chromatography

Independent of the technology used for molecule detection, samples can be split into its components before entering the machine inlet, a method called chromatography. Chromatography is a fundamental technique in analytical chemistry for separating and analysing mixtures of chemical substances (Miller, 2009). It operates on the principle that different compounds will exhibit varied affinities for a stationary phase when carried by a mobile phase across it (Miller, 2009). The properties affecting this affinity could include molecular weight, polarity, or solubility (Pitt, 2009). Components that have a lower affinity for the stationary phase move more quickly, thus separating from those that interact more strongly and travel slowly (Miller, 2009). The stationary phase can be a solid, or a liquid supported on a solid, and the mobile phase might be a liquid or a gas (Poole, 2003).

This method has several applications, including the purification of chemicals and the analysis of complex biological samples (Ovbude et al., 2024). It is also essential in industries such as pharmaceuticals for the quality control of products (D'Atri et al., 2019). Depending on the mobile phase employed, chromatography can be subclassified into gas chromatography (GC), liquid chromatography (LC), or thin-layer chromatography (TLC), each suited to specific types of samples and analytical

needs (Poole, 2003). A relevant method for metabolomics is modern high-performance liquid chromatography (HPLC) systems can separate complex samples in shorter times by working with high pressures (Ovbude et al., 2024).

1.6.1.3 Flow injection mass spectrometry

Flow injection mass spectrometry (also known as direct injection or direct infusion), as the name suggests, is a method where the sample is directly injected to the mass spectrometer without a chromatographic separation (Kirwan et al., 2014). The rationale behind this approach is that sample processing can be done very fast, and therefore it can allow measurement of hundreds of samples in a few hours. The drawback is that without separation, molecules can interact and therefore matrix effect are observed, affecting the acquisition signal-to-noise ratio. One of the most common matrix effects is ion suppression. Ion suppression is a phenomenon where ionization of a molecule is not achieved completely due to competence with other chemical species or analytes (Annesley, 2003; Volmer & Jessome, 2006) and can lead to reduced detection.

Flow injection can work even if the case of supernatant is directly injected without prior extraction, where masses can be more easily detected if they are in range far from typical small central metabolites. (Ciasca et al., 2020; Fuhrer et al., 2011; Kaiser et al., 2018; Nanita & Kaldon, 2016; Sarvin et al., 2020; Vaidyanathan et al., 2002).

1.6.2. Application of metabolomics

In human health, metabolomics has become integral to disease diagnosis and treatment. By analysing metabolic profiles, researchers can identify biomarkers (Zhang et al., 2015; Qiu et al., 2023;) for diseases like cancer (Danzi et al., 2023; Wang et al., 2023), diabetes (Jin & Ma, 2021), and neurodegenerative disorders (Shao & Le, 2019; Gonzalez-Riano et al., 2021), enabling early detection and personalized medicine (Jacob et al., 2019; Castelli et al., 2022). For instance, lipidomics, a branch of metabolomics dealing with lipid detection, has revealed specific signatures associated with psoriasis, offering potential targets for new therapies (Zeng et al., 2017; Liu et al., 2023). On the treatment side, metabolomics is a powerful tool to

explore new drugs in terms of molecular diversity and mechanisms of degradation by patient's metabolism (pharmacokinetics) (Pang & Hu, 2023).

In agriculture, metabolomics is applied to improve crop resilience, quality, and yield (Razzaq et al., 2019; Benkeblia, 2022). By studying plant metabolic responses to environmental stresses, such as drought or pathogen attacks, researchers can breed or engineer crops with enhanced resistance (Villate et al., 2021). Furthermore, metabolomics is used to monitor later effects of genetic modifications in plants, ensuring that these changes do not inadvertently produce harmful metabolites (Rischer & Oksman-Caldentey, 2006; Ricroch et al., 2011).

The food industry utilizes metabolomics to ensure food quality, safety, and authenticity. By analysing the metabolite profiles of food products, companies and laboratories can detect adulteration, monitor freshness, and assess nutritional content (García-Pérez et al., 2024). Additionally, nutritional metabolomics explores the relationship between diet and health, identifying biomarkers that reflect dietary intake and nutritional status (O'Gorman & Brennan, 2017; Kortesniemi et al., 2023). This information may guide dietary recommendations and interventions aimed at preventing diet-related diseases.

In synthetic biology, metabolomics is essential for optimising and/or identifying targets in metabolic pathway engineering (Dromms & Styczynski, 2012; Costello & Martin, 2018; Khanijou et al., 2022; Cortada-Garcia et al., 2023). This application is crucial for producing biofuels, pharmaceuticals, and other valuable chemicals sustainably (Babele & Young, 2020; Ellis & Goodacre, 2012; Hill et al., 2015; Hollywood et al., 2018; Muhamadali et al., 2023). By understanding the metabolic networks within these organisms, researchers can enhance production efficiency and minimize unwanted by-products, making the production processes more sustainable and cost-effective, as in the case of natural products coming from secondary metabolites (Nguyen et al., 2012).

Environmental sciences also benefit from metabolomics, which is used to monitor the impact of pollutants on ecosystems (Miller, 2007; Deng et al., 2019; Kovacevic & Simpson, 2020; Dumas et al., 2022). By analysing the metabolic responses of organisms to environmental stressors, scientists can assess ecosystem

health and develop strategies for environmental conservation (Waller et al., 2023; Brown et al., 2024; Song et al., 2024). This application is particularly important in the context of climate change, where metabolomics can help understand how organisms adapt to changing environmental conditions, informing conservation efforts (Romero et al., 2021).

Additional applications can be found in other areas such as aging research (Panyard et al., 2022), and the range of uses expand every year, making metabolomics a versatile tool for investigating general biochemical questions.

1.6.3 Integration with other omics

Integrating metabolomics with other omics techniques such as genomics and transcriptomics can offer additional insights into biological systems, enhancing our understanding of metabolic regulation at a molecular level (Chen et al., 2023). Recent advances in high-throughput technologies have facilitated this integration by allowing for comprehensive profiling of multiple biological layers—each capturing a unique aspect of cellular function and thus, providing a holistic view of an organism's biological status (Bersanelli et al., 2016).

For instance, when metabolomics is combined with genomics, researchers can observe how genetic variations influence metabolic pathways directly. Such studies often employ methods like QTL-based integration where quantitative trait loci identified in genome-wide association studies are used to correlate genetic variations with observed metabolic profiles (Adamski, 2012). Another example could be linking secondary metabolites with its respective genetic units, biosynthetic gene clusters (Leão et al., 2022; Schorn et al., 2021). This provides an expanded view of how the involved biosynthetic genes are relevant for the chemistry observed and can inform biosynthesis strategies and engineering in the future (Hooft et al., 2020).

Moreover, incorporating transcriptomics allows scientists to connect gene expression levels with metabolic changes, offering insights into how genes are regulated and expressed in different conditions (Patt et al., 2019). This is crucial for identifying key regulatory genes that impact metabolic functions in tissues, potentially highlighting new therapeutic targets or biomarkers for diseases (Maan et al., 2023). In

metabolic engineering, metabolic and transcriptomics are a useful combination to constrain genome-scale metabolic models and obtain a broader picture of distribution of fluxes at the organism or community level (Sen & Orešič, 2023; Zampieri et al., 2023).

The integration typically involves sophisticated data analysis techniques to handle the massive and heterogeneous data sets, such as network inference or high dimensional statistical methods for data fusion (Subramanian et al., 2020), which help in identifying correlation structures and causal relationships between different biological layers.

1.6.4. Historical context

Advances in analytical chemistry equipment were responsible for the development of metabolomics as field. In that sense, around the mid-twentieth century, scientists started thinking about the utility of high-throughput molecule detection. Williams, in the late 1940s, discussed that it would be possible to obtain a "metabolic profile" from biological fluids and proposed how to use this information for biomedical purposes (Gates & Sweeley, 1978). Decades later, improvement in mass spectrometry and nuclear magnetic resonance made the possible detection of molecule *en masse* with enough sensitivity and accuracy to yield significate biological information. Thus, reports as earlier as 1970s, include metabolic profile measurement using gas chromatography mass spectrometry in clinically relevant samples such as human urine and tissue extracts (Horning & Horning, 1971; van der Greef & Smilde, 2005; Griffiths & Wang, 2009) and NMR for analysis of unmodified biological samples (Hoult et al., 1974). Continuous analytical improvement followed, with major breakthroughs in the 90s thanks to better processing algorithms, increase in computational power and availability of bigger compound libraries (Kell & Oliver, 2016).

1.7. Mass spectrometry basics

A mass spectrometer is an analytical instrument used to determine masses of molecules present in a sample (Gross, 2017). It is widely employed in various scientific fields, including chemistry, physics, biology, and environmental science (Dettmer et

al., 2007). Understanding the fundamental components and operation of a mass spectrometer is crucial for comprehending its applications and interpreting the data it generates.

1.7.1 Main concept

The primary concept of mass spectrometry involves ionising chemical compounds to generate charged molecules or molecule fragments and measuring their mass-to-charge ratios (m/z). This discrimination is achieved by subjecting these ions to electric and/or magnetic fields, which direct them through a flight path or trajectory in the spectrometer (Silberring & Smoluch, 2019).

1.7.2. The instrument

At its core, a mass spectrometer consists of four main components: an ion source, a mass analyser, a detector, and a data analysis system (Gross, 2017) (Figure 1.13). A brief description of each of these components is presented:

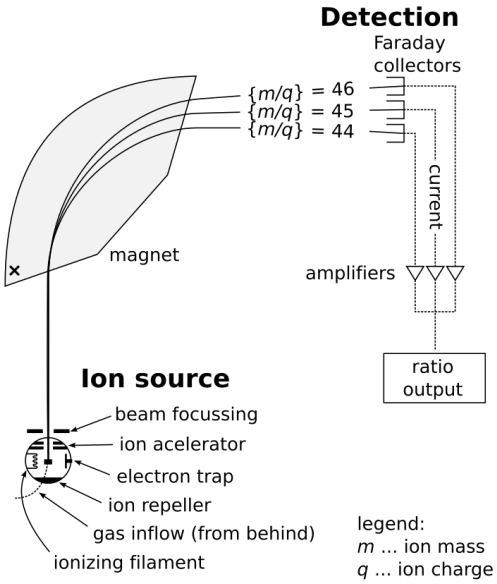


Figure 1.13. A diagram of a generic mass spectrometer that employs a magnetic sector analyser. The molecules are ionised at the ion source, then the ions are focused into a magnet that separate the trajectories of different ions and create an ion traversal profile dependent on the mass-to-charge ratio. A detector waits at the end to collect the ions and transform the signal into a current that can be analysed by the electronic system in the machine. Open-source image from Wikimedia Commons.

1.7.2.1 Ion Source

The ion source is responsible for converting the sample molecules into ions, i.e., electrically charged chemical species. Ionization is necessary to enable the subsequent separation, detection, and analysis of these ions in the mass spectrometer. Commonly used ionization techniques include electrospray ionization (ESI), electron impact ionization (EI), matrix-assisted laser desorption/ionization (MALDI), and atmospheric pressure chemical ionization (APCI) (Bhardwaj & Hanley, 2014). The choice of the ion source plays a critical role in determining the ionisation efficiency and selectivity of the instrument (Gross, 2017).

In our laboratory, electrospray ionisation is the main method used for the available triple quadrupole mass spectrometer. ESI is based on applying high voltage between a dispenser or nozzle and a machine inlet (Konermann et al., 2013). Then, the sample undergoes a complex ionization dynamic in a vacuum, where the sample volume forms in first instance an electrostatic cone (Taylor cone), and a droplet is released (Wilm & Mann, 1994). Then, this droplet splits into sub droplets due to superficial electric charges and continue splitting on the way to the inlet, until the small size of these nanodroplets cause the rapid evaporation of the surrounding solvent and the entering of the sole ions in the inlet (Banerjee & Mazumdar, 2012; Konermann et al., 2013) (Figure 1.14).

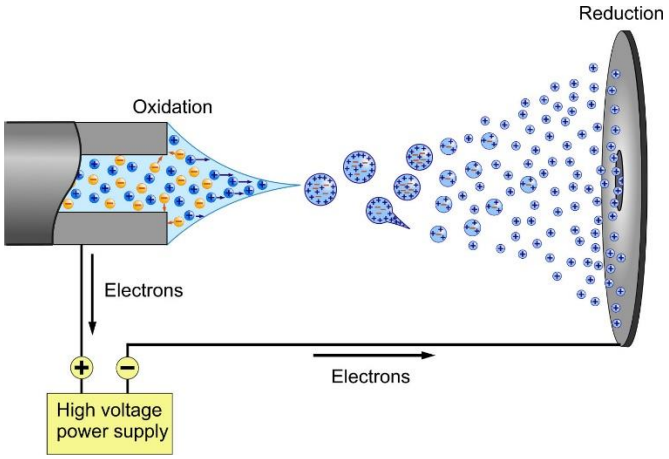


Figure 1.14. Diagram of an electrospray. The voltage applied between the nozzle and the inlet induce several physical phenomena. First, a cone of charged solvent, called the Taylor cone, is formed by an equilibrium between the solvent superficial tension and the involved electrostatic force. The equilibrium breaks at the tip where charged droplets are generated.

These charged droplets keep splitting on the way to the inlet, ending in nanodroplets where ionisation of sample components is approximately complete. Open-source image from Wikimedia Commons.

Electron ionisation, as the name indicates, use an energetic electron beam to expel electrons from the sample's molecules by collision and generating the ions (Gross, 2017). However, this method can be too energetic, and fragmentation can occur before entering the mass spectrometer (Siuzdak, 2004). MALDI uses a laser to generate an ion cloud from a sample's material (El-Aneed et al., 2009) and has been successfully employed for spatial metabolomics, i.e, metabolite identification over a spatial section, such as a tissue preparation (Aichler & Walch, 2015; Li et al., 2022). APCI employs a different method since it uses chemical reactions in a gas phase to obtain the ions (Rebane et al., 2016), and it is used as a soft ionisation method for sensitive molecules, without requiring vacuum conditions such as the EI case (Rebane et al., 2016).

1.7.2.2 Mass Analyser

The mass analyser separates ions based on their mass-to-charge ratio (m/z). Various types of mass analysers are used in mass spectrometers, each employing different principles to achieve ion discrimination. Some commonly used include quadrupole (Paul, 1990), time-of-flight (TOF) (Price, 1993; Boesl, 2017), magnetic sector (Chen et al., 2015), and ion trap (Stafford, 2002) analysers. Each of them has its advantages and limitations in terms of mass resolution, mass accuracy, sensitivity, and scan speed. Quadrupole systems utilise a set of rods to generate electric fields and filter ions based on radiofrequency (Leary & Schmidt, 1996). Time-of-flight analysers push the ions into a tube via a pulser and measure the time they take to reach the detector, which is proportional to the m/z ratio (Boesl, 2017). Magnetic sectors were primarily used in the first spectrometers and employed a magnetic field to deflect ions into trajectories. Then, the longitudinal position on the arrival to the detector would determine the m/z ratio (Nier, 1991). Finally, ion trap analysers keep the ions in confined space by carefully manipulating magnetic or electrostatic fields in a geometric fashion (Stafford, 2002). By manipulating this field, ions can be pulsed out to keep a group of ions with specific mass. Several high-resolution analysers,

going to the scale of 1 ppm, have been derived from ion traps, including magnetic traps such as Fourier Transform ion cyclotron resonance (FT-ICR) MS (Comisarow & Marshall, 1974) and electrostatic traps such as the Orbitrap (Hu et al., 2005; Makarov et al., 2006; Zubarev & Makarov, 2013). Development of analysers is a continuous process, powered by sophisticated ion optics and materials simulations, where manufacturing companies patent new designs every year, in a race to increase resolution and sensitivity (Li et al., 2021; Kuster et al., 2024).

1.7.2.3 Detector

The detector in a mass spectrometer measures the abundance of ions passing through the mass analyser. Different types of detectors can be used, such as electron multipliers, photomultiplier tubes, or solid-state detectors (Koppenaal et al., 2005; John Roboz, 2016). The choice of detector depends on the specific requirements of the experiment, such as sensitivity, dynamic range, and response time (Koppenaal et al., 2005). A detector basically transforms a signal triggered by an ion-related event, such as resonance frequency in quadrupole or arrival in time-of-flight into an analogous electrical signal, which can be further processed by the machine electronics (Li et al., 2021).

1.7.2.4 Data Analysis System

The data analysis system is responsible for processing and interpreting the signals generated by the detector. Modern mass spectrometers are equipped with sophisticated (and often proprietary) data acquisition software that allows for real-time data analysis, peak integration, deconvolution, and identification of mass spectra. Raw data can be also pre-processed by using open-source software options (Röst et al., 2016). Data analysis may also involve comparing experimental spectra with reference databases to identify unknown compounds, as it will be described later.

1.7.2.5. How a sample is processed

In summary, the path of a sample in a mass spectrometer involves a series of steps. First, the sample is introduced into the instrument, either directly or via a separation technique such as liquid chromatography or gas chromatography. The sample molecules are then ionized in the ion source, generating ions with different

masses and charges. These ions are accelerated and directed into the mass analyser, where they are separated based on their mass-to-charge ratios. The separated ions are then detected, and their abundance is recorded by the detector. Finally, the data analysis system processes the recorded signals and generates mass spectra, which provide information about the composition and structure of the sample.

1.7.3. Historical notes

The history of the mass spectrometer has its beginning in JJ Thomson's work, who demonstrated how particles can be deflected by magnetic fields and was able to determine the existence of electrons and isotopes, work that granted him a Nobel Prize (Gross, 2017). The first-called mass spectrometer was developed around the start of the 20th century, thanks to the work of Aston and Dempster in more sophisticated magnetic sector analysers for isotope measurements (Sutton, 2022). The current mass spectrometer market involves a handful of analytical chemistry companies, together with small developers at startups or academic lab, where new ion optics design and analysers can be tested.

1.8. Functioning of a triple quadrupole mass spectrometer

A quadrupole, as the name suggests, is composed of 4 parallel rods. These rods are connected to an alternate current circuit and a direct current circuit and can generate strong electric fields (March, 2009). This field is controlled by the (radio-) frequency (RF) potential and the polarities of the rods (March, 2009). When the molecules pass through the rods, and a specific RF frequency is set up, only molecules with a specific mass will be retained in oscillatory motion inside the volume of the quadrupole, while other molecules will be expelled out (Miller & Denton, 1986).

A triple quadrupole mass spectrometer (QqQ-MS) (Yost & Enke, 1978; Yost, 2022) operates by utilizing a sequence of three quadrupoles aligned in series (Figure 1.15). The first quadrupole acts as a mass filter, allowing only ions of a specific mass-to-charge ratio to pass through based on the user's selection (Dass, 2007). Subsequently, these selected ions enter the second quadrupole, which as a collision chamber, where they undergo collision-induced dissociation. In this chamber, neutral gas molecules collide with the ions, causing them to fragment into smaller ions (Dass,

2007). The third quadrupole then functions as another mass filter, selecting one or more of the resulting fragment ions for final detection (Dass, 2007). Therefore, it can be considered as a tandem mass spectrometer.

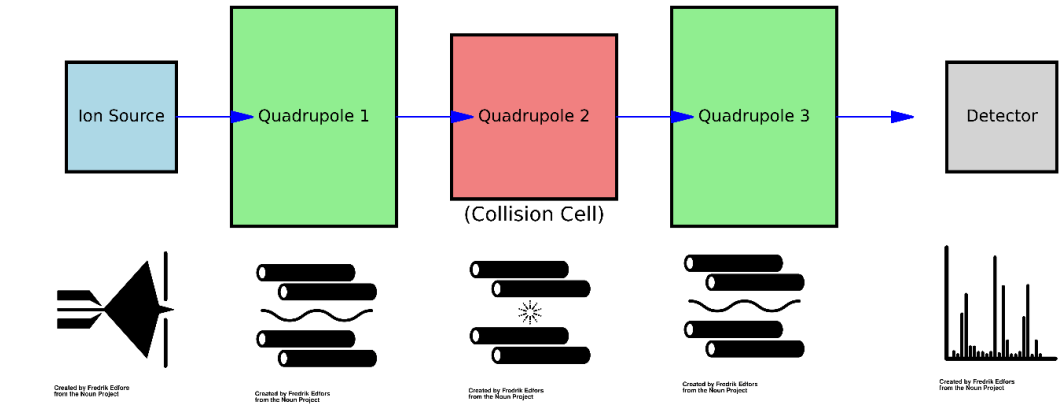


Figure 1.15. Components of a triple quadrupole mass spectrometer. The metabolite in the sample is ionised and focused into the first quadrupole, where a filtering is performed. The second quadrupole is configured as a collision cell, where the molecules are fragmented due to impact with inert gas molecules. The third quadrupole is similar in operation to the first one and provides another filter step. Finally, the ions arrived at the detector so that the masses' signals can be processed. Icons provided by Fredrik Edfors/Noun Icons Project.

Triple quadrupole machines can use their quadrupole sections independently to perform versatile mass analysis, and indeed they might have different resolutions. The most common mode to perform quantification is called selected reaction monitoring (SRM). In SRM, a specific mass is selected by the user and filtered in the first quadrupole. This mass is called the precursor mass. Then this mass undergoes fragmentation, and another mass is selected for filtering in the third quadrupole. These are fragment or product masses, normally calls as transitions. Together, the precursor and fragment mass can determine a chemical species that is unique to a specific structure, allowing distinction even between chemical isomers in some cases. The fragment mass depends on the collision energy, measured in eV, and can be retrieved experimentally or by databases, reports or simulations. If several metabolites are being tracked, with its given precursor and fragment masses, it can be called multiple reaction monitoring (MRM), although SRM and MRM are often interchangeable. In this mode, the dwell time variable is important, because it will

define the number of points/ion events acquired over time for each precursor/fragment mass (Thomas et al., 2022). These can also be defined as the number of defined transitions over the total time window for acquisition, called the cycle time (Thomas et al., 2022).

Triple quadrupole mass spectrometers can perform other kinds of chemical analysis, apart from SRM/MRM, due to flexible configuration of the quadrupoles. In the above paragraph, it was discussed how the fragment mass can be achieved experimentally. This can be done by only filtering a precursor mass in the first scan and perform a full scan over a m/z window in the third quadrupole (product ion scan). Thus, the masses of all fragments can be detected for that specific ions and collision energy optimisation can be employed to get useful fragment information and quantification for SRM.

Similarly, it can be inverted to find possible precursors for a fixed fragment (precursor ion scan). Full scans can be also independently performed by each quadrupole filter. This allows tracking of neutral losses in the fragmentation part, by defining a specific difference mass in the last filter (neutral loss scan). Finally, any quadrupole can be used a single analyser, which can be used for mass measurement of pure compounds or as overview of the masses that are present in the sample.

1.9. Mass spectrometry-based metabolomics sample preparation

Sample preparation for mass spectrometry-based metabolomics involves a series of time-sensitive procedures aimed at ensuring accurate detection and quantification of metabolites. The process begins with the collection of biological samples, such as blood, urine, tissue extracts, or cell cultures, which dictates the subsequent preparation methods required.

Following collection, rapid cooling or treatment with cold solvents, a process known as quenching, is implemented to halt metabolic activity and preserve the metabolic state at the time of collection (de Koning & van Dam, 1992; Faijes et al., 2007; Teng et al., 2009). Metabolites are then extracted using appropriate solvents, the choice of which, like water, methanol, or chloroform, depends on the polarity of the target metabolites (William Allwood et al., 2013; Ser et al., 2015). This extraction might

involve phase separation to partition metabolites into aqueous and organic layers (William Allwood et al., 2013). In gas chromatography MS, derivatization of metabolites is often necessary to enhance their stability, volatility, or detectability (Moros et al., 2017).

When needed, a cleanup stage can be useful for removing proteins, and other contaminants that could interfere with the analysis. Techniques such as solid-phase extraction or filtration are commonly employed (Vuckovic, 2020). Finally, the extracts may be kept in cold storage, or dried and then reconstituted in a suitable solvent that is compatible with the mass spectrometry system (William Allwood et al., 2013; Vuckovic, 2020). For the cell cultures, there may be cases where the intracellular and extracellular metabolome wants to be studied separately and cells are separated from the spent medium by centrifugation right after the quenching (William Allwood et al., 2013; Mohd Kamal et al., 2022). Each step in the preparation process must be carefully optimised to minimise the loss of metabolites, either by the rapid turnover of biomolecules in physiologically adverse conditions (León et al., 2013; Lu et al., 2017; Mohd Kamal et al., 2022), or by disrupting cellular membranes and “leaking” these intracellular metabolites to the medium (Siegel et al., 2014; Lu et al., 2017).

1.10. Mass spectrometry-based metabolomics data analysis

Managing and interpreting metabolomics data, especially when consider a huge number of samples. is a challenging task that requires a systematic approach. The process of metabolomics data analysis involves several key steps, each of which is critical for deriving a feature matrix, which is a representation of the abundances of putative metabolites detected in the samples that can be used for downstream statistical analysis (Cambiaghi et al., 2017). The steps described the path for a typical untargeted metabolomics, but some of these tasks are also performed in targeted metabolomics.

1.10.1. Filtering

The first step in metabolomics data analysis is filtering, which involves cleaning the dataset to remove noise and irrelevant data points. This is crucial because raw metabolomics data typically contain many signals, including background noise,

contaminants, and artifacts, which do not correspond to true metabolites (Schiffman et al., 2019). Common filtering techniques include setting thresholds based on signal intensity or frequency of detection across samples (Schiffman et al., 2019). In the case of multiple reaction monitoring, filtering is only necessary if the peak is deemed noisy for a specific precursor/fragment.

1.10.2. Peak/Feature Detection

The next step, which is important in both untargeted and targeted metabolomics, is peak or feature detection. In this part, the software or the user identifies significant peaks in the chromatograms or spectra, corresponding to potential metabolites. Advanced algorithms are often employed to distinguish peaks from noise and to handle overlapping peaks. This step also includes deconvolution (for untargeted metabolomics), which separates co-eluting compounds that might appear as a single peak (Lu et al., 2008) and baseline correction, where an offset signal is subtracted from the retrieved peaks (Sun & Xia, 2024). The outcome is a list of peaks with their corresponding retention times, mass-to-charge ratios (m/z), and intensities. In MRM, the number of acquired points forming the peak is given by the dwell time (Thomas et al., 2022).

1.10.3. Alignment

Alignment is the process of matching features across different samples to account for variations in retention times or m/z values that may arise due to technical variations in the analytical platform. Without proper alignment, the same metabolite could be recorded with slightly different retention times or m/z values across different samples, leading to erroneous conclusions. Advanced computational algorithm, such as warping methods, are used to correct for these discrepancies and ensure that the same metabolite is consistently identified across all samples (Bloemberg et al., 2013). In MRM, alignment is not of vital importance, since a proper peak assignation can be achieved using the precursor/fragment mass definition.

1.10.4. Area under the curve, normalisation and batch correction

Abundance for each peak can be obtained by integrating the area under the curve using numerical methods, such as the trapezoidal rule (Sun & Xia, 2024). Then,

normalization is a critical step that adjusts the data to account for differences in sample concentration, instrumental variation, or batch effects (De Livera et al., 2012; Wu & Li, 2016). Various normalization techniques can be applied depending on the specific dataset and experimental design. These include methods such as total ion current (TIC) normalization, probabilistic quotient normalization (PQN) performed using quality control (QC) samples (Dieterle et al., 2006), and internal standard normalization (De Livera et al., 2012; Wu & Li, 2016). Additionally, if batch effects are not corrected completely, there are specific batch correction procedures that can systematically correct intra- and/or inter-batch drifts using QC sample data (Han & Li, 2022). Intra-batch drifts are result of decreasing sensitivity across the sample order due to accumulation of dirt g, i.e., solids or non-ionised molecules that stick to the inlet or inside the mass spectrometer, while inter-batch drifts arise when samples are not measured concurrently, but rather in different times (Han & Li, 2022). Randomising the sample order is a simple but important consideration to alleviate intra-batch effects (Souza & Patti, 2021) (Figure 1.16).

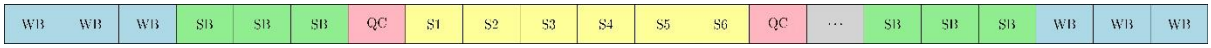


Figure 1.16. A typical sample order for a set of samples in a metabolomics experiment. The sequence begins with HPLC-grade water blanks, shown in blue. Next, solvent blanks are used, consisting of the solvent employed for extraction. A quality control sample is then measured, followed by a set of experimental samples. This process is repeated, with a quality control sample interspersed after every number of samples, six in this case. The run concludes with solvent blanks and HPLC-water blanks. Randomisation of experimental samples is essential to guarantee robustness of results, and decoupling any effect related to the treatments.

1.10.5. Metabolite annotation

The final challenging steps in untargeted metabolomics data analysis is metabolite annotation. This step involves comparing the spectral data (retention time, m/z, and fragmentation patterns) to reference databases or libraries. However, due to the vast diversity of metabolites and the limitations of current databases, not all features can be confidently assigned to a chemical species in untargeted metabolomics (Gertsman & Barshop, 2018). Techniques such as tandem mass spectrometry (MS/MS), isotope labelling, and high-resolution mass spectrometry

(HRMS) are often employed to improve the accuracy of identification (Qiu et al., 2023). Proper annotation is crucial for linking the observed features to biological pathways and interpreting the biological significance of the results (de Jonge et al., 2022). It is an intensively researched topic, where many directions have been explored for improvements, and has a strong connection with the development of new statistical and deep learning methods (Zhou et al., 2022).

1.10.6. Common statistical analyses in metabolomics

After retrieving the feature matrix, there are several options to analyse this multivariate data. The most common approaches involve dimensionality reduction algorithms (PCA, PLS-DA), pathway enrichment analysis, clustering, and machine learning approaches.

Dimensionality reduction techniques like PCA are essential for simplifying complex datasets. This algorithm allows for the visualization of patterns in metabolomics sample data and can help with potential identification of outliers that may not be apparent in the full-dimensional space (Worley & Powers, 2013). PCA is particularly useful in metabolomics exploratory data analysis, where the goal is to uncover the underlying metabolic relationships of the samples without any prior assumptions (Worley & Powers, 2013).

PLS-DA, on the other hand, is a supervised method that combines features of both PCA and linear regression (Ruiz-Perez et al., 2020). It is particularly valuable when the objective is not only to reduce dimensionality but also to maximize the separation between predefined groups or classes within the data (Ruiz-Perez et al., 2020). By incorporating information about the class labels, PLS-DA ensures that the resulting components are not only informative about the data's variability but also about the differences between groups (Gromski et al., 2015). This makes PLS-DA especially useful in classification tasks, such as distinguishing between healthy and diseased states in metabolic data (Gromski et al., 2015; Kalivodová et al., 2015).

Pathway enrichment analysis is another option that it is used to interpret the features in a biological context. This method involves mapping the features to known biological pathways and identifying which pathways are overrepresented in the

dataset (Lu et al., 2023). Pathway enrichment analysis is a powerful tool for generating hypotheses about the biological processes that may be driving the observed metabolic patterns, but consideration should be taken to avoid biases in algorithm parameter selection and experimental procedures (Wieder et al., 2021). It also allows researchers to move beyond statistical correlations and begin to explore potential mechanisms of action, which is particularly important in combination with other omics such as genomics and proteomics (Paczkowska et al., 2020).

Clustering techniques such as hierarchical clustering or k-means can be employed to group similar observations based on their feature profiles (Heinemann, 2019). Clustering is useful for identifying subgroups or patterns within the data that may not correspond to predefined classes (Heinemann, 2019). For example, it can reveal subpopulations with similar metabolic profiles correlating with specific dietary intake patterns (O'Sullivan et al., 2011).

Finally, classical machine learning algorithms like support vector machines (SVM), random forests, or even deep neural networks can be applied for predictive modelling and classification. These models can be trained on the reduced and enriched data to develop predictive tools for applications such as disease diagnosis, prognosis, or treatment response prediction (Pomyen et al., 2020; Sen et al., 2021; Galal et al., 2022; Zhang et al., 2023).

1.10.7. Technical concepts to consider when acquiring metabolomics data.

The acquisition of data can have different approaches, depending on if the focus resides in a single analyte or a set of them. Thus, mass chromatograms can be divided into Extracted Ion Chromatogram (XIC) and Total Ion Chromatogram (TIC). Each of them serves distinct analytical purposes that must be considered in downstream analysis.

1. Extracted Ion Chromatogram (XIC) (Figure 1.17): This method is utilized to isolate and monitor specific ions, or a group of ions associated with a particular metabolite or a class of metabolites. XIC facilitates the detailed examination of targeted compounds within a complex matrix, enabling precise quantification and temporal monitoring across the chromatographic process. XIC are

normally labelled based on precursor and fragment masses (Smoluch & Piechura, 2019).

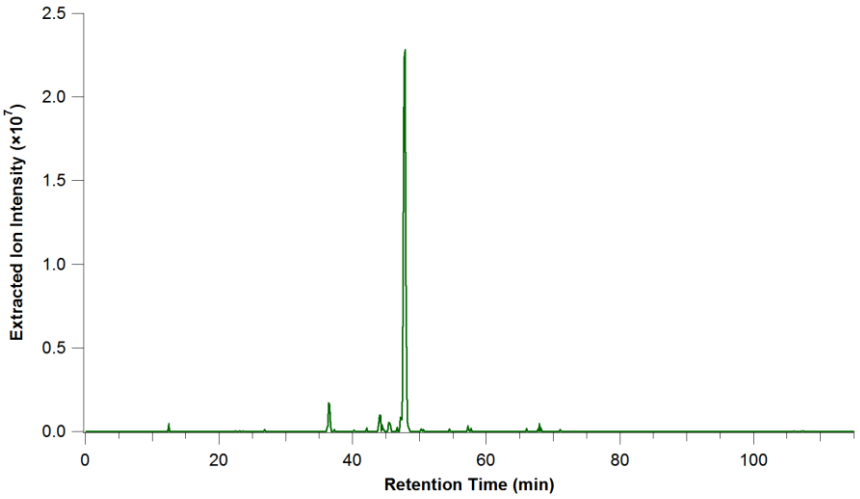


Figure 1.17. Example of an extracted ion chromatogram. The peak is narrow, since only one ion (precursor/product mass) is being tracked over the acquisition time. Credits CWenger at English Wikipedia.

2. Total Ion Chromatogram (TIC) (Figure 1.17): This technique aggregates the intensities of all ions detected by the mass spectrometer throughout the chromatographic separation. The TIC produces a composite profile that reflects the overall chemical composition of the sample at successive time intervals, providing a comprehensive snapshot of the metabolic diversity present (Smoluch & Piechura, 2019).

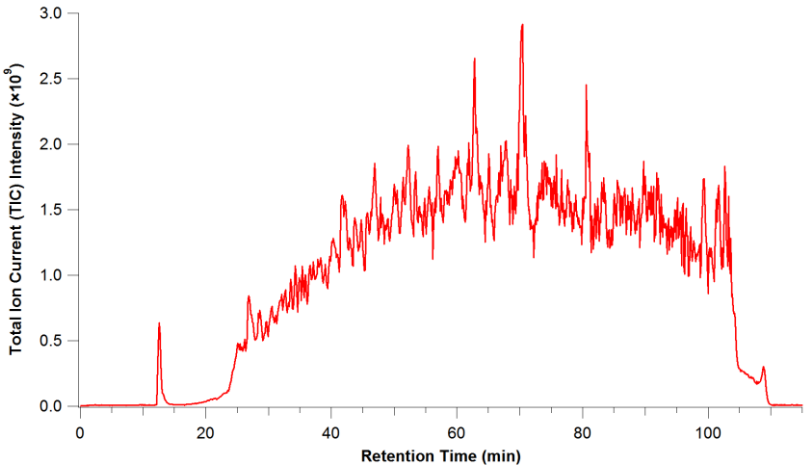


Figure 1.18. Example of a total ion chromatogram. The observed intensities are a sum of intensities coming from multiple ion signals. Credits CWenger at English Wikipedia

1.11. Probing bacterial biochemistry through mass spectrometry-based metabolomics

Bacterial metabolomics, defined as the comprehensive study of metabolites within bacterial systems, has become an import tool in physiological, ecological and biotechnological research. This subfield focuses on the detailed analysis of metabolic processes within bacteria, both at the intracellular and extracellular level, providing insights into their physiological states and interactions with their environment (Tang, 2011; van der Velden & Jansen, 2023).

Bacterial metabolomics has been instrumental in identifying which specific compounds or families of compounds are produced within complex microbial communities (Kellogg & Kang, 2020; Thukral et al., 2023; Bhattacharjya et al., 2024). This is particularly relevant in the study of natural environments where diverse microbial species coexist and interact. The chemical interactions between bacteria and organisms from other kingdoms—such as plants, fungi, and animals—are of great interest, as these interactions often involve the exchange of signalling molecules and metabolites that can influence community structure and function (Kellogg & Kang, 2020). Understanding these interactions through metabolomics can reveal novel biochemical pathways and potential biotechnological applications.

Another of the most significant applications of microbial metabolomics is in the discovery of novel bioactive compounds, particularly secondary metabolites. These compounds, often produced by bacteria in response to environmental stress or interspecies competition, have considerable potential as pharmaceuticals, agrochemicals, and industrial enzymes (Demarque et al., 2020; Sieniawska & Georgiev, 2022). However, the re-discovery of known compounds has been a persistent challenge in natural product research (Deutsch et al., 2022). Metabolomics offers a solution by enabling the rapid identification and characterisation of novel metabolites, thus streamlining the discovery process and reducing redundancy in compound identification (Deutsch et al., 2022).

Another specific research challenge addressed in this context is the optimisation of microbial metabolite production within a model organism. This involves systematic

testing of various culture conditions to determine their effects on the production of specific metabolites (Tanaka et al., 2024). Existing literature underscores the value of integrating multiple metabolic measurements to dynamically capture the network of biochemical reactions within microbial cells. This information is vital for the fine-tune of conditions for industrial metabolite production (Vavricka et al., 2020; Tanaka et al., 2024).

Finally, metabolomics can be applied to answer basic biochemical questions in bacteria. For example, *E. coli* serves as an ideal model for studying microbial biochemistry, due to its well-characterised physiology and gene function information. Metabolomics in *E. coli* has derived into insights related to metabolism stability, response to internal and external signals, and determine molecular response to stress (Carvalho et al., 2019; Lempp et al., 2019; Radoš et al., 2022). Physiological discoveries can be also found in metabolomics approaches for the Gram-positive model *Bacillus subtilis,* including metabolome change for different subpopulations and sporulation (Huang et al., 2024; Meyer et al., 2013) and conditions for using *Bacillus* for biological control (Horak et al., 2019).

1.12. Metabolomics profiling for strain optimisation

Since metabolomics allows for comprehensive analysis of cellular metabolites, it provides insights into metabolic states and bottlenecks that can guide strain engineering efforts. The typical workflow involves pathway analysis to determine affected metabolic pathways in an experimental design and identification of targets based on rate-limiting steps or bottlenecks. Thus, high-throughput metabolomics methods can accelerate strain selection and improvement, as well as provide targets for future metabolic engineering (Hou et al., 2012; Iman et al., 2022; Khanijou et al., 2022; Cortada-Garcia et al., 2023)

1.13. Hypothesis and aims.

This thesis focuses in understanding how iterative experimentation frameworks, specifically Bayesian optimisation, can be coupled with targeted metabolomics for sample-efficient optimisation of culture medium composition in a complex bioprocess, as it is surfactin production. To reach a working platform for Bayesian

optimisation of surfactin production in *Bacillus*, several technical developments were proposed, from reducing time in mass spectrometry data acquisition using flow injection to the integration of robotic platforms for sample manipulation.

1.13.1. Hypothesis

The main hypothesis of the thesis is:

Bayesian optimisation is an effective method to optimise titre of surfactin in *Bacillus subtilis* by manipulating culture medium composition.

1.13.2. Aims

There are 4 aims that are considered in order to assess the rejection of the proposed hypothesis:

- Develop and validate a Bayesian optimisation framework for optimisation of surfactin titres In Bacillus by manipulating culture medium composition.

- Develop a flow injection-mass spectrometry protocol for rapid relative quantification of surfactin variants in complex biological samples.

- Automate and streamline experimental procedures using robotic platforms for large sample handling in iterative experiments.

- Demonstrate the applicability of the developed platform for bioprocess optimisation via optimisation of several components in culture medium.

2. Optimisation of surfactin yield in *Bacillus* using active learning and high-throughput mass spectrometry

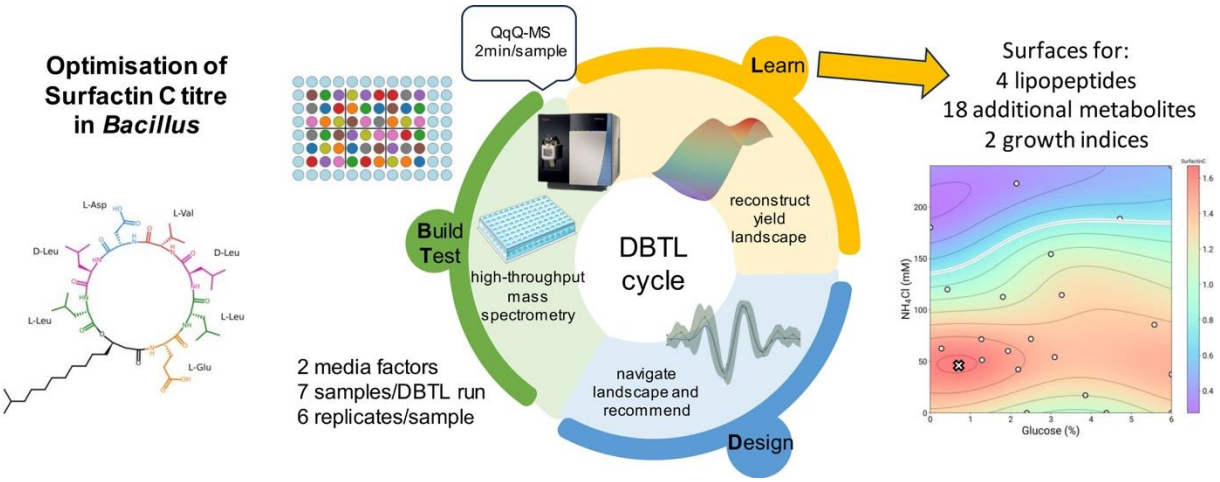
Work conducted by Ricardo Valencia Albornoz[1], Diego Oyarzún[1,2] & Karl Burgess[1*]

[1] Institute of Quantitative Biology, Biochemistry & Biotechnology, School of Biological Sciences, University of Edinburgh, King's Buildings, Edinburgh, United Kingdom

[2] School of Informatics, University of Edinburgh, Edinburgh, EH8 9AB, United Kingdom

Ricardo Valencia Albornoz (thesis author) performed the conceptualisation and testing of the experiments, and the development and testing of the algorithm/pipeline for the optimisation of culture medium. DO and KB provided guidance on methodology development.

Graphical Summary



Outline

Optimisation of a product titre or yield in a bioprocess is crucial for the economic and technical success of its operation. This problem is usually complex, since several factors or variables are involved in the outcome, where some of them can be modified by the experimenter, while others are not controllable. In a bioprocess, medium components are important factors in the final yield and each of the component concentrations can be adjusted to get optimal conditions. Here we present an active learning/Bayesian optimisation framework to enhance surfactin titres by changing medium component concentrations. Surfactin, produced by

Bacillus, is a promising biosurfactant, because of its chemical properties. Reported laboratory titres are typically low, due to its complex molecule assembly pathway. Thus, we used active learning to refine the culture medium composition through iterative experimentation, which enhanced surfactin C levels in *Bacillus subtilis* DSM 3256. Growth curves and other central metabolites were measured as part of the experimental loop.

The final medium mixture led to ~1.6-fold increase after three rounds, compared to the M9 medium standard. Reanalysis of the optimisation data reveals trade-offs when comparing the production of other lipopeptides, Surfactin D and Iturin A, with the maximum OD in the growth curve data. Organic acids in the supernatant positively correlate with surfactin C levels, suggesting an impact on central carbon metabolism. For some metabolites, including certain amino acids and sugars, the change in their abundance around the optimal surfactin C mix is not uniform, indicating an "anisotropy" in how metabolism reacts to shifts in carbon and nitrogen levels. Our framework addresses the challenges of data handling and analysis, offering several visual tools, data analysis techniques, and analytical methods (using mass spectrometry), promising to be valuable components of the Design, Build, Test & Learn cycle of synthetic biology.

2.1. Introduction

Surfactin is a family of lipopeptides produced by strains of the genus *Bacillus*. They are promising candidates to replace oil-based surfactants in intensive industrial activities, due to its temperature resistance. These molecules show prominent tensioactive properties and are considered high value chemicals.

2.1.1. Surfactin metabolism in detail.

Surfactin variants are synthetised by non-ribosomal peptide synthetases system (NRPS). NRPS are responsible for the production of very important compounds for modern society, such as antibiotics, and the surfactin cluster may be consider a prototype of this kind of biosynthetic gene clusters (BGC). Specifically for this molecule, the corresponding NRPS draws molecules from different pathways, assembling them in sequence (Figure 2.1). First, the branched fatty acid in the

surfactin molecule originates from malonyl-CoA and pyruvate in central metabolism, after several steps of elongation (Xia and Wen, 2022). Indeed, leucine and valine are also considered as intermediates in the pyruvate pathway, enabling the synthesis of several CoA variants (Coutte et al., 2015; Xia and Wen, 2022). The origin of glutamate for the lipopeptide is intrinsically related to nitrogen assimilation (Gunka & Commichau, 2012) and aspartate comes from the central carbon metabolite via the oxaloacetate pathway (Park & Lee, 2010). These substrates are available in the cytoplasm, where the NRPS can take the necessary building for the final surfactin variant (Théatre et al., 2021). Thus, optimisation of NRPS pathways is not trivial, as several metabolic fluxes need to be modified to increase the final titre.

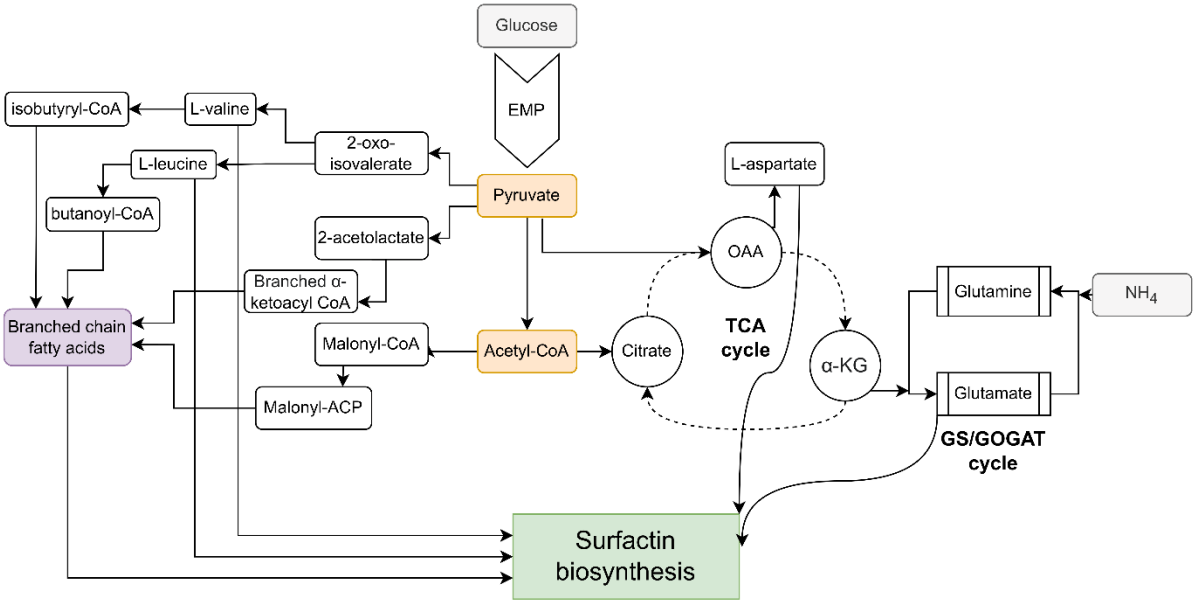


Figure 2.1. Metabolic pathways converging into Surfactin synthesis. Adapted from Hu et al., 2019; Xia & Wen, 2022.

2.1.2. Attempts to increase and quantify surfactin production.

Production of surfactin in liquid medium demonstrates high variability and low titres (Hu et al., 2019; Xia and Wen, 2022). Scale also needs to be considered as working volume can affect drastically surfactin output (Alonso & Martin, 2016). Due to these challenges, several strategies to increase surfactin production has been proposed, including genetic engineering and optimization of fermentation conditions, among others (Hu et al., 2019; Xia and Wen, 2022).

An example of the genetic engineering approach is the overexpression of the exporter gene *swr*C (synonym: *yer*P), which has been reported to result in a 1.2–15.6-fold increase in surfactin production (Hu et al., 2019). Additionally, focus has been given to replacing the promoter P_{srf} by a highly expressible promoter, resulting in titres between 0.04-1.5 g/L (Hu et al., 2019). In terms of metabolic engineering, a comprehensive rewiring of *Bacillus* metabolism was carried out to generate a surfactin hyperproducer by extensively knocking out pathways and overexpressed genes unrelated to the main amino acid and fatty acid pathways (Wu et al., 2019).

Regarding culture medium optimisation, several studies have employed statistical design approaches to maximize surfactin titres by altering the concentrations of the medium components (Bertrand et al., 2018; Czinkóczky et al., 2023). Components include pure compounds such as glucose and more complex substrates from industrial waste (Fonseca et al., 2007; Mohanty et al., 2021). Additional work has been made to study the influence of specific carbon sources and metal ions in surfactin production (Bartal et al., 2018). Here, it was determined that fructose and xylose were the carbon sources with most impact in the distribution of surfactin variants produced. Metals ions as well can stimulate the synthesis of additional variants, such as methyl sterified surfactin forms (Bartal et al., 2018). Before this publication, Wei et al. 2004 also analysed the effects of iron ions in production, reaching 3 g/L by supplementing the medium with 4 mM of Fe^{+2}. Thus, iron and manganese ions has been identified as the main metal stimulators of surfactin production. In an industrial setting, a patent was filled regarding long fermentation of *B. subtilis* for surfactin production (Yoneda et al., 2006). This patent shows that concentration between 8 and 50 g/L can be achieved using soybean flour as feedstock.

On the metabolomics side, a few reports of metabolic studies on surfactin production in *B. subtilis* are present in the literature (Valdés-Velasco et al., 2022), and it has also been investigated in *B. velezensis* (Wang et al., 2018) and in the case of other lipopeptides (Sánchez-Lozano et al., 2023; Yang et al., 2023).

In terms of quantification of surfactin, a few colorimetric methods have been developed for surfactin and general (anionic) lipopeptide quantification (Zhu et al.,

2014; Yang et al., 2015; Ong and Wu, 2018; Heuson et al., 2019; Kubicki et al., 2020), with diverse accuracy across them. Most of the methodologies use a system composed of a reporter and a complementary cationic detergent. The reporter is added first, then the cationic detergent is added to form a complex with the reporter, but surfactin competes with this interaction due to its negative charge, therefore releasing the reporter in the medium in a proportional concentration to the surfactin concentration.

2.1.3. Active learning methods for optimisation of titres

There are a few reports in the literature utilising active learning or Bayesian optimisation techniques specifically for bioprocess optimisation. Active learning has been employed in the optimisation of culture medium for mammalian cell growth, specifically HeLa-S3 cells (Hashizume et al., 2023). Interestingly, this work utilises a decision tree as a surrogate model and considers two different experimental setups: a normal mode where the cells are grown for the usual cultivation period of 168 hours and a time-saving mode where the culture time is shortened to 96 hours, allowing faster iterations. Concentrations of 29 compounds in the composition were varied, and the biomass was measured as absorbance at 450nm. They compared the performance of the active learning on the proposed modes, establishing that in the normal mode, the loop can find a better medium (~1.5 fold) than the reference (EMEM medium) after four rounds (Hashizume et al., 2023). More advanced Bayesian optimisation pipelines, such as multi-objective and multi-fidelity optimisation, have been used for the optimisation of desired properties in cell-based meat production. Meat production by cell growth in reactors, also known as cellular agriculture, depends strongly on the medium combination and the corresponding passage number. In addition, other objectives need to be optimised simultaneously apart from cell density, such as economic cost. (Cosenza et al., 2023) found a low-cost, high-growth medium by using an active learning approach with a Gaussian process regression surrogate and an acquisition function related to expected hypervolume proposals. In the case of a multi-fidelity approach, the cell counts information obtained in Passage 1 can be seen as cheaper but less reliable data to model the effect of medium concentrations

before testing recommended iterations in high-fidelity Passage 2 cells (Cosenza et al., 2022).

2.1.4. Aims and objectives

The chapter consider the following aims:

- Validate a Bayesian optimisation framework for optimisation of surfactin titres in *Bacillus* by manipulating culture medium composition.

- Develop a flow injection-mass spectrometry protocol for rapid relative quantification of surfactin variants in complex biological samples derived from cultures.

The specific objectives outlined for this chapter are:

- Determine the range of the components' concentrations to test

- Test the Bayesian optimisation *in silico* using the Branin function.

- Determine the optimal carbon and nitrogen concentration in medium for maximum Surfactin titre.

- Analyse trade-offs and links between surfactin production, biomass and other metabolites

2.2. Materials and Methods

2.2.1. Strains and media employed.

Bacillus subtilis DSM 3256, a *Bacillus* strain known for producing surfactin, was obtained from the Deutsche Sammlung von Mikroorganismen und Zellkulturen (DSMZ) repository. According to DSMZ, this strain has been classified as a surfactin producer (DSMZ, 2024). Further investigation into related strains confirmed that the genome of *Bacillus subtilis* ATCC 21332, a synonym strain, harbours the surfactin biosynthetic gene cluster (ATCC, 2024). Before employing the initial strain stock in experiments, a Gram stain was performed to confirm the Gram-positive nature of the bacterium. Later, each time a new stock is generated for maintenance at – 80 °C, the species identity was assessed by 16S rRNA sequencing.

M9 medium was prepared using the following formula: Glucose: 0.4% (as the carbon source), NH_4Cl: 18.7 mM (as the nitrogen source), NaCl: 8.5 mM, $MgSO_4$: 2 mM, $CaCl_2$: 0.1 mM, Na_2HPO_4: 42.2 mM, KH_2PO_4: 22 mM. Care was taken during sterilization of the stock components. Glucose tends to show visible browning after autoclaving at both 20% and 50% (g/L), so filtration is recommended (Leitzen et al., 2021). The remaining components were autoclaved. Additionally, stock solutions were prepared based on the reported solubility of each compound.

For optimization experiments, glucose and ammonium chloride were omitted from the formula, resulting in a basic M9 salts/buffer solution to which carbon and nitrogen sources could be added later. All components of the M9 medium were procured from Sigma Aldrich. Microplate experiments were conducted using ultra-low attachment surface flat bottom 96-well microplates from Corning.

2.2.2. Microplate cultures

Bacillus subtilis DSM 3256 frozen stock was revived by culturing on an LB agar plate at 37°C overnight. From this plate, individual colonies were inoculated into six precultures consisting of 5ml of M9 medium. The cells were then grown in a shaking incubator at 37°C and 180 rpm for 12 hrs.

Conditions corresponding to the carbon-nitrogen concentrations in the medium, as well as controls, were block-randomised across a 96-well microplate, considering six replicates (blocks). These blocks are physically distinct groups of 9 wells, comprising 7 conditions, 1 M9 medium well with added bacteria and 1 M9 medium well with no bacteria, acting as control. These were manually pipetted into place. The layout for block randomisation was obtained from an R script utilising the agricolae package (Mendiburu and Yaseen, 2020). In each well, 100µl of 2X M9 salts (M9 medium without a carbon or nitrogen source), 80µl of the glucose/ammonium mix, and 20µl of the preculture were added. To help with the pipetting process, the corresponding concentrations of glucose/ammonium chloride for each well were achieved first in microtubes by diluting the stock and then transferring a constant 80µl volume to the well, as described. For the inoculum, the starter cultures were assigned to each block and adjusted to achieve an initial optical density at 600nm (OD_{600}) of 0.1 in each well. This initial OD is easy to measure in the plate reader and can be obtained

consistently across the samples. The OD was measured in a Biochrom Ultrospec 10 Cell Density Meter.

The microplate culture was conducted using a Tecan Infinite M200 PRO plate reader, recording an OD_{600} measurement every 10 minutes for 36 hours at 37°C. This temperature was selected according to Sen and Swaminathan, 1997, though it is recognised that other reports mention a temperature of 30°C for production (Xia et al., 2024). The fermentation time aligns with publications reporting that a peak on the yield of surfactin is reached between 24-48 hrs (Guo et al., 2024), though longer times might decrease surfactin concentration, since it can be used as alternative carbon source by the remaining cells (Shaligram & Singhal, 2010). The agitation was set to a maximum amplitude of 6mm (shaking frequency is a function of the amplitude, but it is not given my manufacturer). Growth data for each well were collected in an Excel file for post-processing. Growth rates and maximum OD were extracted using the amiga package v. 3.0.2 (Midani et al., 2021).

After culturing, the microplate was centrifuged for 20 minutes at 3220 x g (4000 rpm in machine) and 4°C to separate the cell pellet from the supernatant. The employed centrifuge was an Eppendorf Centrifuge 5810 R, because this model allows for plate holders.

80µl of supernatant was taken from each well in the cold room at 4C and transferred into Axygen 96-well PCR plates. These plates possess a deep bottom. Several difficulties were found in this transfer, because of the foamy and slightly viscous nature of the supernatant/spent media, and the depth accuracy of the pipetting, to avoid perturbing the pellet. After failed attempts, a 3D printed cover was built such that the tip only penetrates the well to a depth where the tip end is 3mm above the pellet. Details of the 3D model are discussed in the results and the stl file is available in the Data Availability section. At the end, the samples were sealed with a cap and stored in a -80°C freezer until mass spectrometry quantitation.

2.2.3. Surfactin and additional metabolites quantification

We employed flow injection of spent media with a ThermoFisher Dionex Ultimate 3000 autosampler. The sample volume was set at 1 µl. The mobile phase

comprised a 1:1 ratio of acetonitrile to water with 0.1% formic acid, and the flow rate was 200 µl/min. These conditions were optimised experimentally and are discussed in Chapter 3. The sample acquisition time spanned 1 minute, and to quantify specified molecules, selected reaction monitoring (SRM) on a ThermoFisher TSQ Quantiva triple quadrupole (QqQ) mass spectrometer (MS) was implemented. The MS parameters, as well as the precursor/product masses, collision energy and dwell time table, can be found in Supplementary Tables BT1 and BT2 in Appendix B, respectively. The precursor/product masses for surfactin variants were obtained and confirmed using PubChem open mass spectrometry data (PubChem, 2024).

Peak extraction was accomplished using the rawrr package in R (Kockmann and Panse, 2021). In the script, each of the scans in a raw file corresponds to a specific precursor/product mass for each retention. If there are more than one product associated with a given precursor, the intensity is summed across the products. The number of acquired points along the 1 min run is controlled by the dwell time. Higher values of dwell time (in ms). The custom script is available at the link from the Data Availability section. The script outputs a file called a "hypertable". This hypertable consists of three columns with retention time (s), scan information, intensity, and the raw file name from which the measurements come.

Subsequently, metabolites peaks were baseline corrected using the asymmetric least squares algorithm from the Python pybaselines package (Erb, 2022) set to default parameters. The corrected peaks were integrated over the 1-minute run using the trapezoid rule function from the NumPy package via a custom Python script. The integrated intensity data were then compiled into a table.

Outliers were identified and removed based on the interquartile range (IQR), keeping only values within the range of the median - 1.5IQR to median + 1.5IQR. These values were then normalized by dividing them by the average M9 titres observed for each batch. A table comparing the relative surfactin C titre to the observed M9 titre for every condition or combination was constructed for the subsequent active learning prediction step. Similar tables were generated for additional metabolites.

2.2.4. Active learning loop

From the minimum and maximum concentrations of glucose and ammonium chloride that were selected for testing, a 2D design space was defined. Seven initial conditions were obtained from a Latin hypercube design (LHD) (McKay et al., 1979). The centred LHD was implemented in Python using the pyDOE2 package (Sjögren and Svensson, et al., 2018). After getting the relative surfactin C titre to the observed M9 titre from the MS quantification, these values were fitted using a heteroskedastic Gaussian process regression (GPR) model (Rasmussen & Williams, 2006; Balandat et al. 2019), where the corresponding glucose/ammonium concentrations are inputted as variables/features and the titre is the observed output. From the GPR predictions, the q-Noisy expected improvement (q-NEI) acquisition function (Letham et al., 2017, Balandat et al. 2019) is calculated for the design space, and optimising this function retrieve seven combinations to be tested on the next iteration of the loop.

The model and the acquisition function were implemented using the Ax and Botorch library in Python (Balandat et al. 2019), and default parameters were used. Several additional scripts used in intermediate steps for formatting data tables and are described in the supplementary material. Surface plots, principal component analysis (PCA) and radar charts to explore and analyse the data were implemented using the matplotlib, seaborn and plotly packages in Python. PCA biplots were generated using the pca package in Python (Taskesen, 2020). The surfactin molecule diagram was generated using the Pikachu package (Terlouw et al., 2022).

The Branin function (https://www.sfu.ca/~ssurjano/branin.html) is defined between 0 and 15 in the x axis, and –5 and 10 in the y axis. It possesses three global minima with same function value 0.397887, at coordinates (-π, 12.275), (π, 2.275), and (9.42478, 2.475). Mathematically, it is defined as (Equation 14):

$$\mathrm{Branin}(x_1, x_2) = \left(x_2 - \frac{5.1}{4\pi^2}x_1^2 + \frac{5}{\pi}x_1 - 6\right)^2 + 10\left(1 - \frac{1}{8\pi}\right)\cos(x_1) + 10 \quad (14)$$

2.3. Results

2.3.1. Determination of glucose/ammonium chloride concentration range using a one-at-time approach

In active learning, many of the algorithm parameters have to be tested or recommended values are considered. However, the ranges of the values of the factors involved in the cycles should be given by the scientist, based on previous knowledge or an educated guess on the limits of the system under study. In our case, to determine the glucose and ammonium chloride concentration range for the active learning experiment, we conduct a growth experiment using as medium the composition of M9 medium, but either carbon or nitrogen concentration is fixed (Figure 2.2). This approach is usually known in the literature as one-factor-at-a-time (OFAT). As mentioned in the Introduction, for statistical analysis this design is not reliable, but it may offer a quick solution for determining concentrations' maximum range.

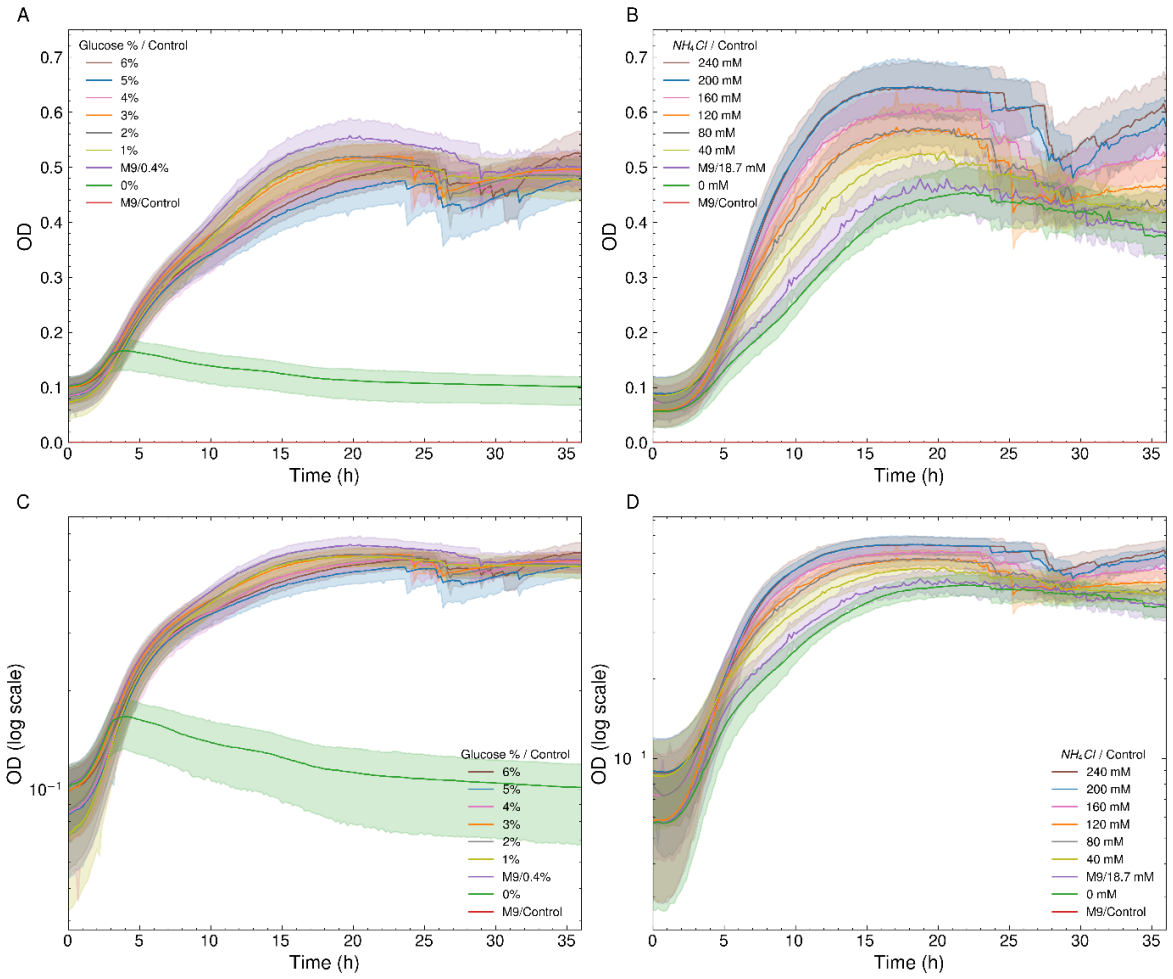


Figure 2.2. Growth curves for *Bacillus subtilis* DSM 3256, when changing carbon and nitrogen separately via the one-factor-at-a-time (OFAT) approach. Curve colouring is associated with a specific carbon and nitrogen concentration according to the legend. Bacteria were grown for 36 hrs. **A)** Growth curves for variable glucose concentration, from added 0 to 6%, and fixed ammonium concentration of 18.7mM **B)** Growth curves for variable ammonium chloride

concentration, from added 0 to 240mM, and fixed glucose concentration of 0.4%. No cell wash was performed in the transfer from the pre-cultures to the microplate wells. Therefore, some carryover can be observed, including the non-zero growth for the added 0 mM of nitrogen condition in **B.** Cells reach stationary phase around 24 hrs. **C)** Semi-log plot of OD vs time for fixed ammonium and variable glucose. Biphasic growth is observed where the rate changes at the time point of 6 hr, due to nitrogen limitation. **D)** Semi-log plot of OD vs time for fixed glucose and variable ammonium. Biphasic growth is only observed in low nitrogen conditions. For higher nitrogen concentrations, exponential growth is observed, and the rate is proportional to the amount of nitrogen added.

This process not only helped determining the design space for the active learning loop but also provided valuable insights into the effect of main nutrients in the microbial physiology of *Bacillus subtilis* DSM 3256.

When glucose concentration is varied and ammonium chloride concentration is fixed, we observed a slight detrimental effect of glucose concentration in maximum OD, especially in higher concentrations. After 5% glucose, the bacteria grow, but the maximum OD has been reduced by 21% compared to lower concentrations (Figure 2.2, A). As it is evidenced in the semi-log plot of OD vs time, every combination displays biphasic growth with a change on growth rate (slope in the semi-log plot) at 6 hrs, due to nitrogen limitation (Figure 2.2, C). Due to practical considerations, we decided for 6% as upper limit for glucose. This may avoid any scenario of carbon limitation when testing over the full range of glucose concentrations. In addition, making glucose stocks with higher percentage is more complicated (if we make a 10x stock, it is a 60% glucose solution stock, which already takes long time to dissolve and prepare).

For the experiment where the glucose concentration is fixed, and the ammonium concentration is varied, higher concentration of NH_4Cl in the medium are correlated with higher maximum OD (Figure 2.2, B). In addition, a higher nitrogen concentration, above added 120 mM, results in a regrowth after 24 hrs. This regrowth is described in the literature, and it is caused by the recycling of available mass from dead cells (Zhu et al., 2023). The maximum OD is proportional to the nitrogen concentration (Pearson's r=0.967, p=8.887e-05) (Supplementary Table BT3). Biphasic growth is only observed in the low ammonium chloride combinations, where nitrogen would be limited (Figure 2.2, D). After 240 mM of ammonium chloride concentration,

no further increase in the growth is observed, and this will be considered as the upper limit for ammonium chloride concentration). In the lower ammonium limit, growth is observed even when no nitrogen is added to the culture medium. This is explained by carry-over, i.e, a certain concentration of nitrogen is transferred from the spent M9 medium in the preculture accompanying the inoculum.

The number of generations reached at 24 hrs is 2.16-2.7 (2 complete doublings). This indicates that choosing an initial OD of 0.1 may not the best choice for the experiment, since means that the culture did not reach “maturity”, and this may affect surfactin production dynamics (Horvath, 1970). However, this value is amended in the culturing protocol for Chapter 4 (initial OD of 0.01). The iterative experiment presented in this Chapter 2 keeps the value of 0.1, for consideration. Additionally, onwards, maximum OD and last measured OD will be the measurements considered for the optimisation analysis as representatives of biomass.

2.3.2. *In silico* benchmark of iterative strategies using a synthetic function

Before employing the active learning pipeline to a biological experiment, its performance can be evaluated by substituting the objective function of interest in the experiment by a synthetic function, i.e., a function that can be described mathematically in an exact way. Not every function is suitable for the benchmarking of global optimisation algorithms, because it should exhibit a landscape with multiple maxima and minima to test. Therefore, we consider the Branin function for this evaluation (Figure 2.3). This function has been thoroughly employed for testing optimisation algorithms because it possesses three separated global minima with exactly the same value. Therefore, a good black-box optimisation algorithm should be able to identify every minimum and cannot get stuck in one of them, as this behaviour would affect the performance of the algorithm on more complicated, realistic functions.

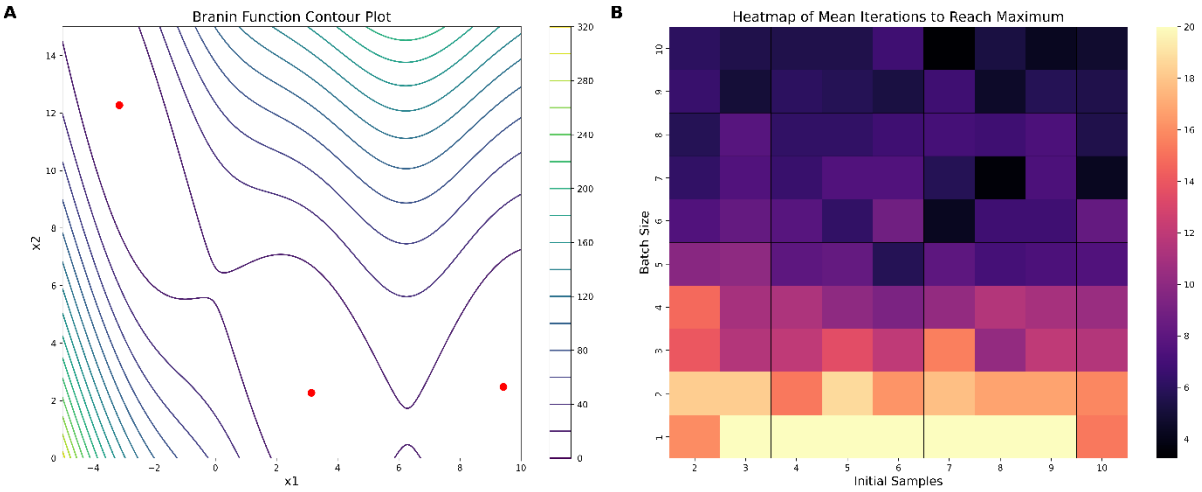


Figure 2.3. Benchmarking parameters of a Bayesian optimisation loop in a test function. A) Contour plot of the negative Branin function (ground truth), highlighting the three maxima on the 2D space with a red dot. **B)** A comparison on the number of iterations needed to converge to an optimal value, when changing the initial samples and number of samples per batch in the active learning framework. Darker colours indicate a smaller number of iterations. The iterations are stopped when the current best value is closer to the optimal function value by less of 0.5%.

We test our Bayesian optimisation pipeline by modifying the number of initial samples, which is related to how much initial information we have about the function landscape and the number of samples per batch. The iterations are stopped when the current best value for the observed iteration is closer to the optimal function value by less of 0.5%. As expected, higher numbers of initial and batch samples imply less iterations, although medium initial and batch sizes also perform well for this complex function.

2.3.3. Production surfaces for surfactin C and observed biomass.

The culture conditions were optimised to enhance the titre of Surfactin C in the spent media (Figure 2.4) from *Bacillus subtilis* DSM 3256 using an iterative active learning loop. The medium composition variables, specifically the carbon concentration sourced from glucose and nitrogen concentration derived from ammonium chloride, were adjusted in a base M9 medium. While temperature and agitation are pivotal to surfactin production (Bertrand et al., 2018), they remained constant to simplify the experimental design.

The active learning cycle tested seven different carbon/nitrogen concentration combinations for Surfactin C production in each iteration. Each iteration can be seen as stages in the widely recognised Design, Build, Test, and Learn concept in synthetic biology (Figure 2.4A).

The Build and Test stages involve cultivating the bacteria in microplates with varying combinations and then measuring the metabolites using mass spectrometry. Surfactin C (Figure 2.4B) and three other lipopeptides (Surfactin B, Surfactin D, and Iturin A) were identified using a flow injection-mass spectrometry method between iterations (Figure B.1 and B.2). This method was developed specifically for the experiment, given its ability to measure a single sample within 2 minutes and a full 96-well microplate in approximately 2 hours, thus streamlining the process. The microplate experiments were appropriately randomized using a custom script that can be used in general experiments (see Data Availability), optical density (OD) was measured (Figure B.8) and Surfactin C titre results for iteration 0 are depicted in Figure 2.4C. We decided to reduce the number of combinations per plate and increase the number of replicates (n=6), since quality over quantity in the data should be beneficial for the loop and the subsequent data analysis, thorough a precise estimation of the biological noise. The effectiveness of a small number of samples has been reported previously (Pandi et al., 2022).

The Learn phase pertains to using a Gaussian process regression model as a surrogate model, predicting the Surfactin C titre landscape alongside its associated prediction uncertainty. Lastly, the Design stage identifies with suggesting new combinations, facilitated by the acquisition function. This function is a mathematical equation that balances between exploitation (selecting combinations close to previously tested high-titre ones) and exploration (areas of the design space that remain uncertain).

Initial combinations were derived from a Latin hypercube design, and will be denoted as Iteration 0, with subsequent combinations recommended by maximising the acquisition function. After three iterations, the active learning loop does not show any further improvement, obtaining a ~1.6-fold titre improvement relative to the surfactin titre in M9 medium (Figure 2.4E). The optimum is reached at 0.8% glucose

and 50 mM NH4Cl. Previously reported values for optimal surfactin production, including modifications to the Cooper and Landy media regarding carbon and nitrogen concentrations, correspond to 0.8% glucose and 100 mM NH4Cl (Willenbacher et al. 2015), in agreement with the obtained maximum for carbon concentration. When updating the models after each iteration (Figure B.7), the average uncertainty in the predictions decreases from 0.45 to 0.3 (Figure 2.4D), indicating that model is gaining more information about the system across iterations.

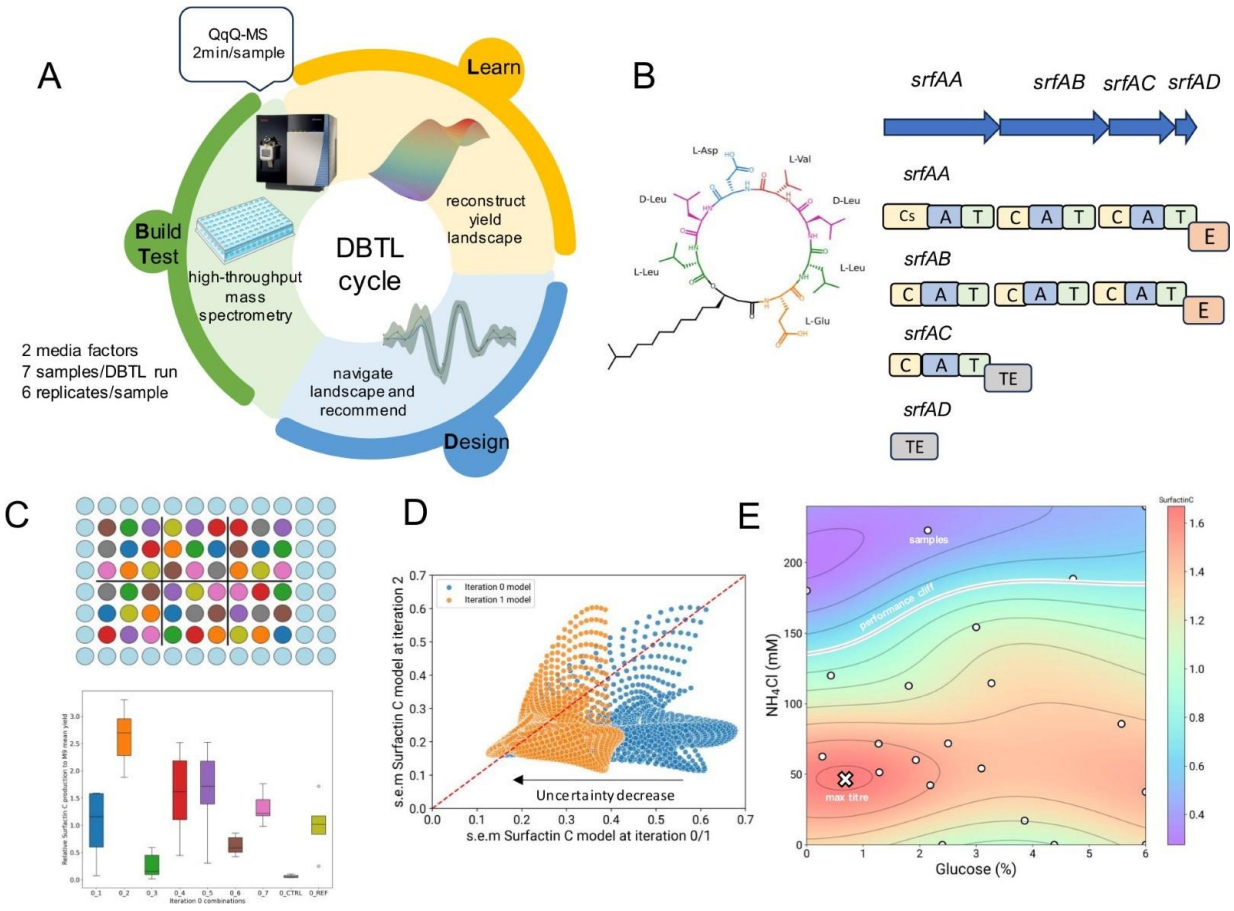


Figure 2.4. Active learning for optimisation of Surfactin titres. A) This diagram illustrates the active learning loop employed and its association with the standard stages of a Design, Build, Test, Learn (DBTL) cycle. The Build and Test stages encompass the cultivation of bacteria in microplates using machine learning-suggested combinations, followed by metabolite measurements via mass spectrometry. In the Learn stage, a Gaussian process regression model serves as a surrogate, forecasting the Surfactin C titre landscape together with its prediction uncertainty. The Design phase, meanwhile, focuses on proposing new combinations, aided by the acquisition function. **B)** The structure of Surfactin C is displayed, emphasising the ring amino acids. The sequence for these amino acids is L-Glu1-L-Leu2-D-

Leu3-L-Val4-L-Asp5-D-Leu6-L-Leu7. A simplified surfactin biosynthetic cluster diagram presents the four biosynthetic genes within the cluster. Each gene codifies for a megaenzyme and the domains in each of these enzymes follows the nomenclature: C for condensation; A for adenylation; T for thiolation or PCP; E for epimerase; and TE for thioesterase. The initial condensation domain Cs in *srfAA* facilitates the integration of the fatty acid at the synthesis commencement. **C)** This section shows the layout for the microplate experiment during the initial iteration of active learning. Combination treatments, alongside control and M9 reference treatments, underwent block randomisation, resulting in six blocks. These are associated with biological replicates. A boxplot displays the Surfactin C titres achieved in the initial iteration. **D)** A dense grid of carbon/nitrogen combinations was employed to estimate uncertainty levels, expressed as the standard error of the mean, for models updated after iterations 0, 1 and 2. Models are compared pairwise, showcasing the uncertainty for identical grid points when comparing the model from iteration 2 against those from iterations 0 and 1.**E)** The final Surfactin C landscape following all three iterations is shown. The combination predicted to yield the maximum titre is marked with a cross. A zone exhibiting reduced titre (less than 0.5 in relation to the M9 titre, known as the performance cliff) is delineated with a white line.

In the case of surfactin variants, such as Surfactin B and Surfactin D, as well as the lipopeptide Iturin A from a different biosynthetic cluster, their production profiles are similar to that of Surfactin C (Figure B.1). This is consistent with previous reports on co-production of fengycin and surfactin (Yaseen et al., 2017), phenomena that also is observed in other categories of biosynthetic gene clusters (Qi et al., 2021).

2.3.4. Identifying trade-offs between lipopeptide production and biomass.

Interestingly, these production profiles can be used to investigate trade-offs between lipopeptide production and maximal biomass *in silico* (Figure 2.5). By inputting simulated combinations into the final model, it is possible to identify Pareto fronts between lipopeptide titre and the maximum optical density (OD). The results show no evident trade-off for Surfactin C and Surfactin B in relation to biomass (Figure 2.5A). In other words, there's minimal sacrifice in biomass to achieve a high titre. Conversely, a slight trade-off is observed for Surfactin D and Iturin A, where a reduction in biomass of about 10% facilitates a 5% increase in the observed titre (Figure 2.5B), suggesting that the metabolic resources are diverted away from growth and towards the production of the lipopeptides/variants. Conversely, as growth rates increase, the

production of Iturin B decreases, highlighting the metabolic burden associated with its synthesis.

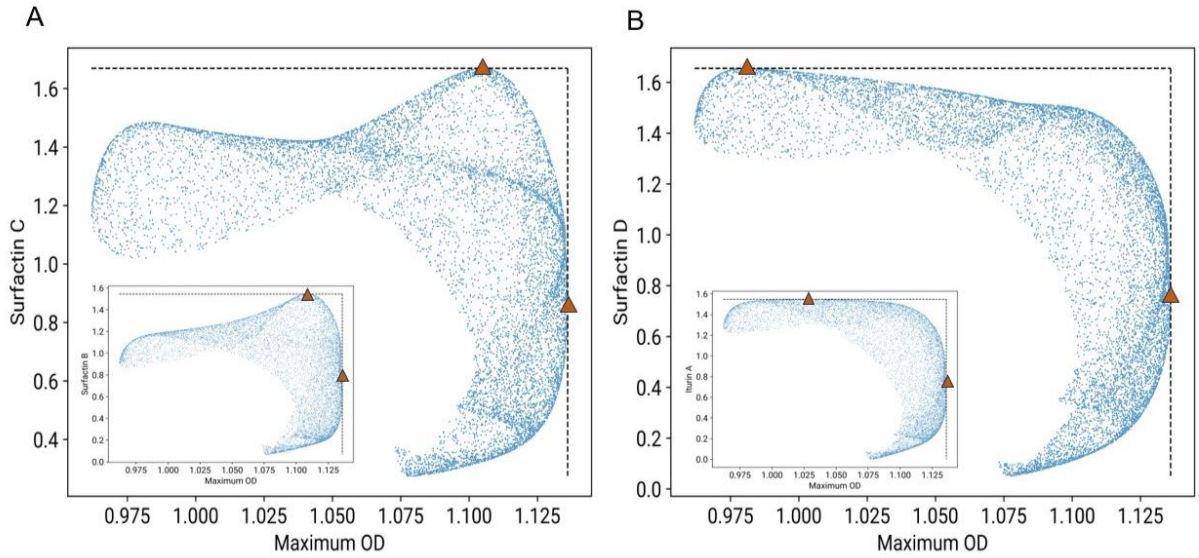


Figure 2.5. Pareto front obtained by simulating the lipopeptide production and biomass with the final (GPR) model using a random and dense set of combinations. The limits of the Pareto fronts are shown using red triangles. **A)** For Surfactin C and D, growth can be minimally sacrificed to obtain a maximum of production, therefore, no trade-off is observed. The ratio of biomass vs lipopeptide, that has to be sacrificed to obtain the maximum production, is approx. 0.025/0.8 = 0.03 titre/OD for Surfactin C and 0.025/0.7 =0.035 titre/OD for Surfactin B. **B)** For Iturin A and Surfactin D, the Pareto front indicates an evident trade-off between lipopeptide production and bacterial growth. The maximum production of Iturin A and Surfactin D occurs at lower growth rates. The ratio of biomass vs lipopeptide, that has to be sacrificed to obtain the maximum production, is approx. 0.15/0.8 = 0.188 titre/OD for Surfactin D and 0.1/0.7 = 0.142 titre/OD for Iturin A.

2.3.5. Carbon-related metabolites and correlation with lipopeptide production.

In addition to the lipopeptide list, various metabolites associated with carbon metabolism and the tricarboxylic acid cycle (TCA) can be measured using the flow injection method (Figure B.3 and B.4). As these are not included to the medium, they are found in the spent medium due to two reasons: export of molecules from the cell and the release of cytoplasmic content from membrane disruption, primarily caused by surfactin itself. These measurements provide valuable insights into the bioprocess.

Data on lipopeptide production, carbon/TCA-related metabolites in the spent medium, and growth curves can be organised for further analysis. We divide the analysis into two subtasks: correlation/PCA/pathway analysis and a novel directional analysis. The correlation/PCA/pathway approach is commonly used in metabolomics. It involves creating a table with metabolites as features and employing visualisation techniques like correlation heatmaps and dimension reduction methods. The results can be embedded into a metabolic pathway diagram and with it, one can infer useful biochemical connections to the synthesis of the desire compound (Bartel, 2013). On the other hand, the proposed directional analysis reveals how metabolism reacts to simultaneous changes in medium composition, focusing on the production surfaces and conceivable asymmetries in the metabolic response.

2.3.5.1. Correlation and PCA analysis.

Using solely the data from the loop samples, we performed a correlation analysis amongst the available 24 features: 4 lipopeptides, 18 other metabolites, and 2 derived from growth data. A hierarchical clustering dendrogram highlights two primary clusters of metabolites, further divided into four closely correlated subgroups (Figure 2.6A). The first (6) and second groups (7) of the primary cluster contain a combination of amino acids and additional metabolites from the glycolysis/glucogenesis pathway. In contrast, the third (7) and fourth groups (5) consist of lipopeptides (4), and organic acids related to carbon metabolism (6) respectively. Both Canonical Correlation Analysis (CCA) and a PERMANOVA test validate the distinct nature of these groups (CCA: p-value approximates 0 across all dimensions; PERMANOVA: p-value 1e-4). The organic acids in the fourth group showing a positive correlation with lipopeptide production, are easily recognisable as components or affiliates of the tricarboxylic acid cycle, suggesting heightened activity in this pathway.

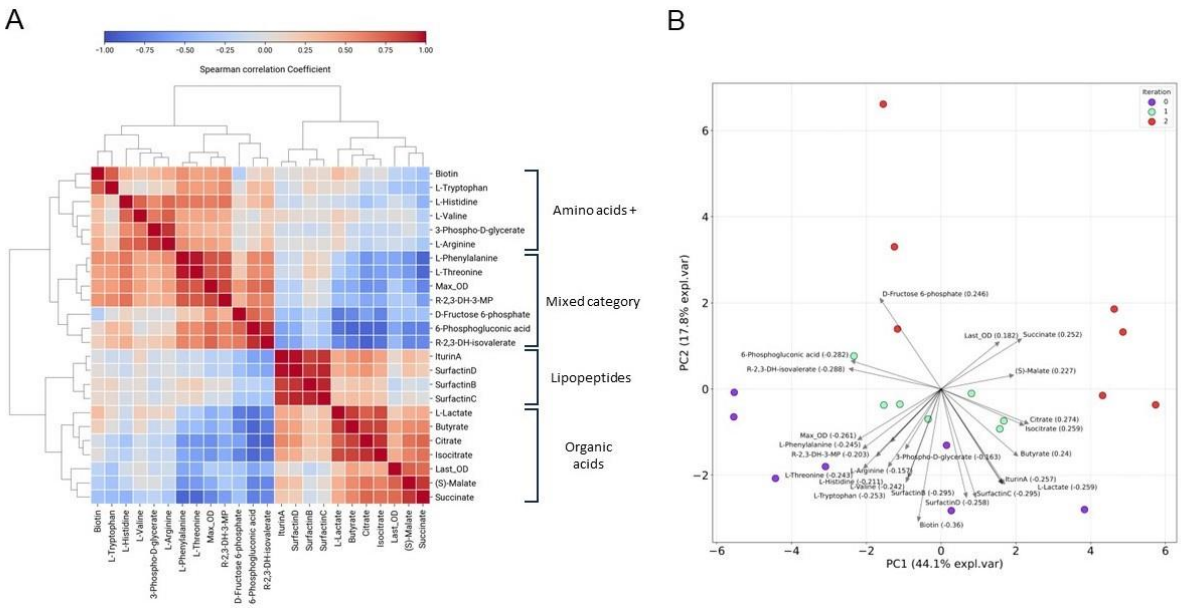


Figure 2.6. Multivariate analysis of additional metabolites in the experiment. A) Spearman correlation between the measured metabolites/biomass across loop samples. The rows and columns are ordered by hierarchical clustering, shown as a dendrogram. Red colours indicate high positive correlation, while bluer colours indicate high negative correlation. Two primary clusters were identified from the dendrogram, which can be subdivided into 4 groups. Each of the groups is labelled within general biochemical categories. **B)** PCA biplot of the active learning loop samples. Iteration samples are depicted as points with distinct colours; purple for iteration 0, green for iteration 1, red for iteration 2. The arrows in the biplot represent the loading scores of each metabolite, indicating their influence on the principal components. This information is also presented in the scores accompanying the metabolite labels. To facilitate reading in the biplot, the metabolite names are abbreviated when necessary.

We can complement these insights by looking at a principal component analysis (PCA) biplot (Figure 2.6B) on the experimental samples. The first and second components account for 44.1% and 17.8% of the explained variance, respectively. Notably, the lipopeptides and TCA organic acids account for the variance in the positive values of the first PCA component. In contrast, the metabolites from the primary cluster show the opposite trend. The associated loading scores are similar, indicating that no individual metabolite dominates the variance contribution after the reduction.

Although PCA and other dimension reduction methods are used in metabolomics as a useful visualization to gain insights about treatments' sample

similarities and cluster detection (Bartel et al., 2013), here we will also employ it to get a grasp on how “metabolically diverse” the recommended samples in each iteration are. Remarkably, the samples are indeed metabolically diverse, as shown by the sparsity of the points in the PCA, and the distribution/score of the loadings. Therefore, exploring the design space also explore the metabolic space in a comprehensive way, which is not trivial.

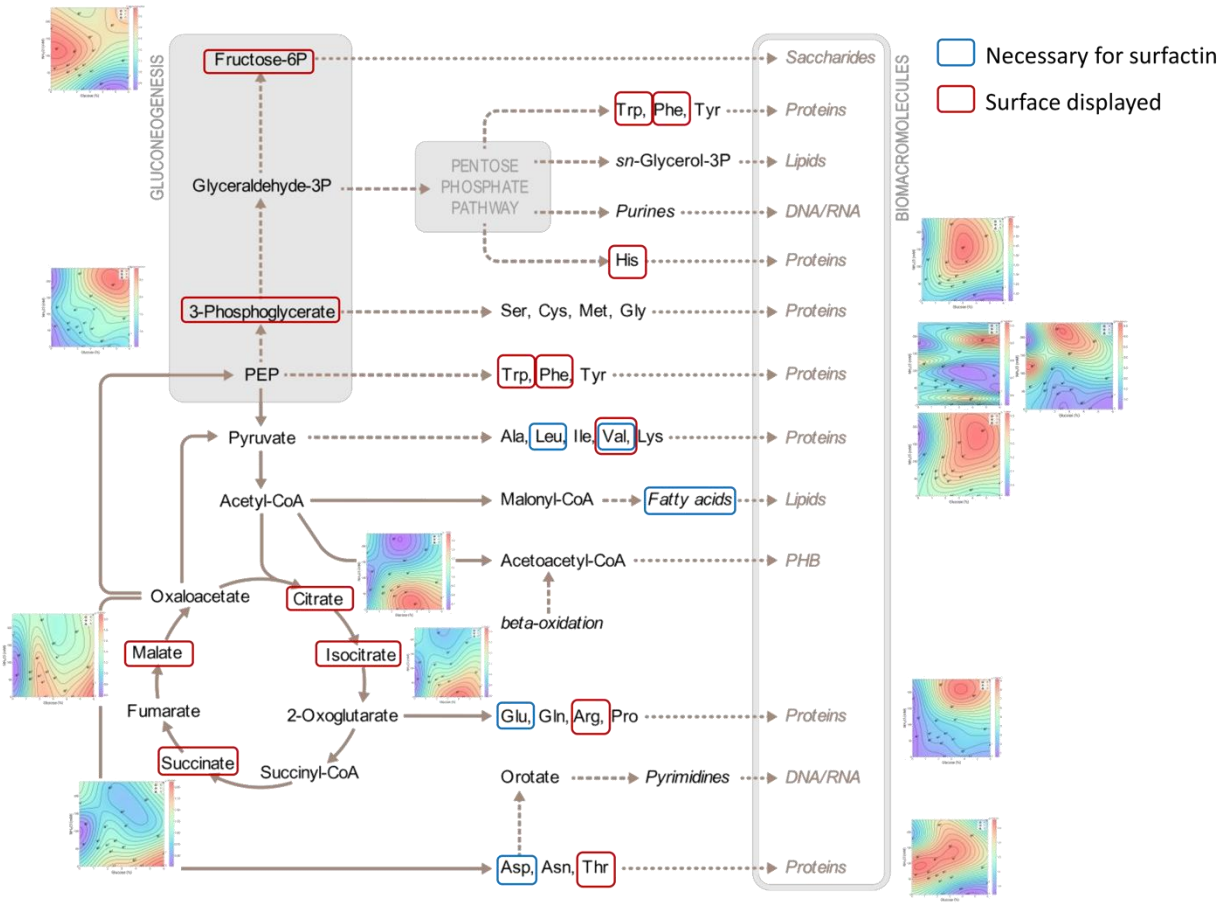


Figure 2.7. Detailed pathways and key intermediates for the synthesis of amino acids and fatty acids in *Bacillus subtilis*. The central carbon metabolism, including glycolysis/gluconeogenesis, the pentose phosphate pathway and the citric acid cycle, is depicted on the left side. From the metabolites on these pathways, several branches diverge into several amino acid synthesis pathways. Fatty acids have their origins in acetyl-CoA/malonyl-coA metabolism. 2-Oxoglutarate is linked with the glutamine-glutamate cycle, which is crucial for nitrogen assimilation. Metabolites with an available titre surface predicted by the model are enclosed in a red box and the surface is shown next to the label, while metabolites necessary for synthesis are enclosed in a blue box. From MetaboMaps, (Koblitz et al., 2020).

Finally, the surfaces can be embedded into a simplified pathway diagram for anabolism in *Bacillus subtilis* (Figure 2.7) (Koblitz et al., 2020), mirroring the conclusions from the previous steps. However, some extra information can be obtained. L-arginine, which is linked with 2-oxoglutarate and is in the same pathway with glutamate synthesis, an amino acid in surfactin, possesses high abundancy with higher carbon-nitrogen concentration, suggesting higher activity in the glutamine synthetase-glutamate synthase (GS/GOGAT) cycle and overflowing of this compound to the medium (Gunka et al., 2012, He et al., 2023).

2.3.5.2. Sensitivity of metabolite production to media component changes: directional analysis.

It is natural to hypothesise that the microbial metabolism responds and adapt in diverse ways respect to slight changes in carbon/nitrogen concentration. In the same way, the direction (or mathematically a vector) to which carbon and nitrogen changes might have specific effect in the metabolic enzymes and reactions. This concept is encapsulated in microbiology as part of the global topic of catabolic control. Catabolic control corresponds to the mechanisms involved in the preference in the utilisation of a nutrient source over another, for example a simple sugar over a more complex sugar for carbon intake, depending on how easily can be assimilate into the metabolism (Görke & Stülke, 2008). These mechanisms not only regulate the utilisation of unique nutrients but also control the metabolic intersections between nutrients' catabolic pathways, which happens in the intersection of carbon and nitrogen utilisation. For *Bacillus*, the 2-oxoglutarate-glutamine-glutamine (GS/GOGAT) cycle determines downstream flow of nitrogen for the synthesis of several important molecules for the cell, for example amino acids. This cycle is affected by global regulators, such as CcpA (activation), CodY (activation) and TnrA (repression). Additionally, energy utilisation in this cycle, together with energy requirement by glycolysis enzymes shapes the flow distribution (Sonenshein, 2007)

Considering this context, nevertheless, it would be good to consider a name for the quantitative method itself, in the context of titre modelling and metabolite control, and we will use the term "metabolic anisotropy" when referring to it, inspired by the meaning of anisotropy in physics and geostatistics, among other science areas, i.e,

non-uniformity in different directions or specifically in experimental terms, in simultaneous changes in medium composition. Using the production surface data, we can explore this idea in a preliminary fashion.

The approach consists of stepping on the observed maximum of surfactin titre and trace orthogonal trajectories to the level curves on different angle directions, until they reach a specified radius. This allows us to calculate the gradient of metabolite/biomass/lipopeptide levels on angles from 0° to 360° respect to the Surfactin C maximum (Figure 2.8). For example, 0° corresponds to only increasing glucose, while 90° corresponds to increasing NH4Cl concentration. Taking a radius of 0.6% glucose and 24 mM of nitrogen around the maximum carbon/nitrogen models, we used the trained models to step on the same optimum combination in each surface and calculate the gradient along the depicted blue circle (Figure 2.8A).

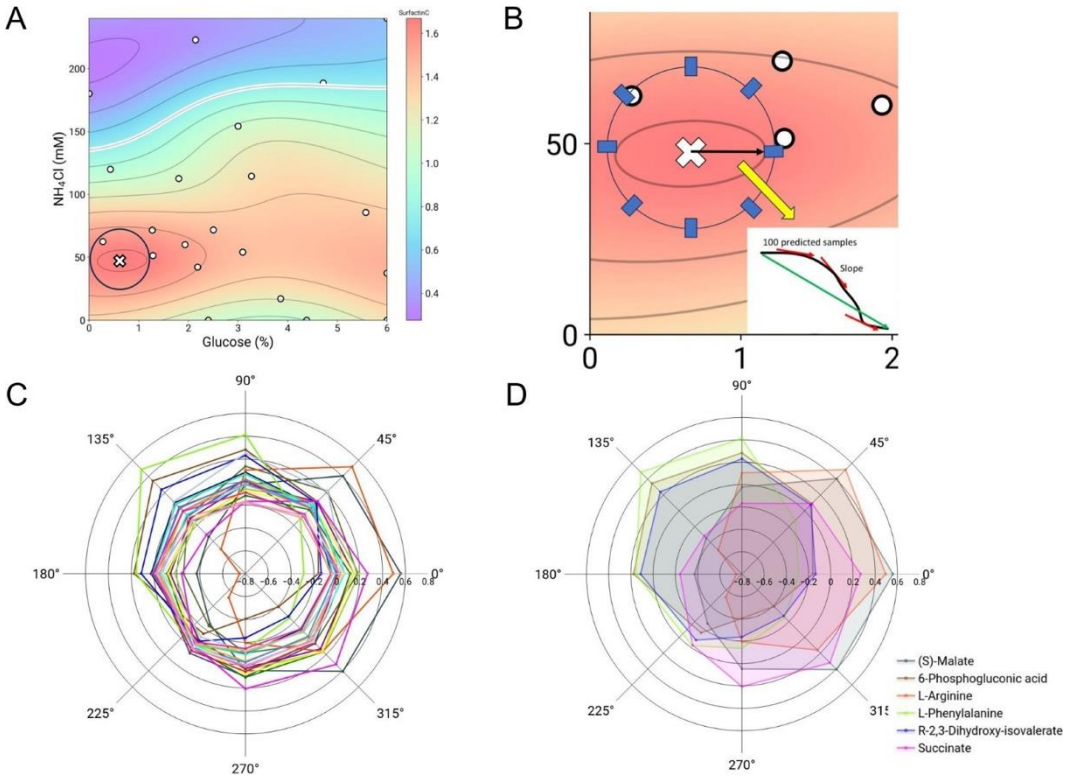


Figure 2.8. Directional analysis reveals which metabolites are sensible to slight and simultaneous changes in carbon/nitrogen composition around the Surfactin C maximum. **A)** Radius around the maximum titre in the Surfactin C surface to calculate gradients in each metabolite production surface **B)** The gradients are calculated by simulating 100 samples in

a path between the maximum titre point and a specific direction in the defined radius. Then the average gradient is obtained by averaging the (approximated) slopes for these samples. **C)** Profile of average gradient for different directions depicted as a radar chart. Each colour corresponds to a specific metabolite. The production surface for each metabolite was used to make the calculations. **D)** For specific metabolites, the profile is highly asymmetric, suggesting flux redistribution or enzyme kinetic changes when changing media composition simultaneously in a specific direction. The color legend is associated to each metabolite: (S)-Malate, 6-Phosphogluconic acid, L-Arginine, L-Phenylalanine, R-2,3-Dihidroxy-isovalerate and Succinate. This phenomenon might be associated with catabolic control. For example, L-arginine levels react rapidly to change in carbon and nitrogen concentrations, since its synthesis directly depends on the GS/GOGAT cycle activity (Sonenshein, 2007).

A radar chart shows that several metabolites show near symmetric profiles, i.e., the rate of change of its abundance in every direction is similar, while others possess unusual long gradients for certain angles (Figure 2.8B). After filtering the gradient profiles that possess overall symmetry across every angle, we found 6 metabolites which present an asymmetric gradient profile, i.e., when stepping down of the optimum carbon/nitrogen concentration and moving towards another combination in a specific direction, the decrease/increase on that metabolite production is significantly different to when choosing another direction. Specifically, L-Phenylalanine, 6-Phosphogluconic acid and R-2,3-Dihidroxy-isovalerate exhibits higher gradient (abundance change) at the 135° direction (0.35 titre change vs -0.2 titre change on the opposite direction), corresponding to decreasing glucose concentration and increasing nitrogen concentration (Figure 2.8D). On the other hand, Succinate, (S)-Malate, and L-Arginine show higher gradients on the right half of the angle plane (0.5 titre change vs -0.4 titre change on the opposite direction), moving towards increasing glucose concentration (Figure 2.8D). For L-arginine, as we have shown before, has a production profile strongly associated with increasing carbon/nitrogen, pointing about fast responses in the nitrogen metabolism, since its synthesis is downstream respect to glutamate and therefore, it depends on the GS/GOGAT cycle activity (Sonenshein, 2007).

. The observed anisotropy on abundance changes for certain spent media metabolites could be related to multiple factors, including rapid enzymatic action,

transient metabolic fluxes, overexpression of exporting systems, among others, and has not been thoroughly studied before on this experimental context, as far it is known.

2.3.6 Evaluation of the titre models using cross-validation and accuracy metrics

To evaluate the predictive capacity of the fitted Gaussian process regression model for Surfactin C, and how information is gained across iterations, we compared the actual vs predicted values for the mean and standard error of the mean on each sample (Figure 2.9). Ideally, both values should coincide, giving the maximum R-squared (R^2) of 1. Additionally, the mean square error from the cross-validation procedure should be as low as possible.

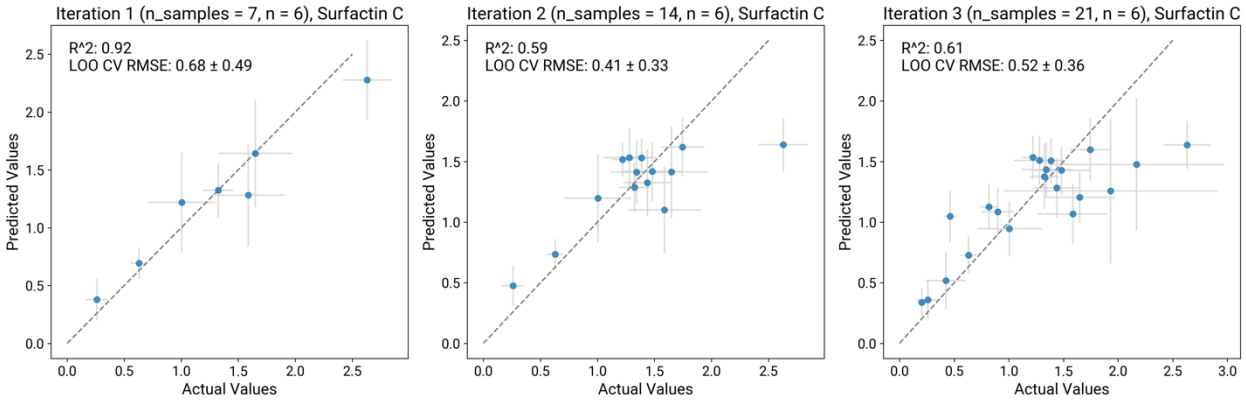


Figure 2.9. Predicted vs actual mean values plot for the fitted Surfactin C Gaussian process regression (GPR) model after each iteration. The average value for each sample considering replicates is indicated in the x-axis, while the mean prediction from the model on each of them is depicted in the y-axis. The standard error of the mean (s.e.m) from the actual samples and the predicted s.e.m from the model is shown using the errors bars.

Thus, the analysis presents a multi-iteration diagnostic of the predictive model for Surfactin C, our compound of interest, with each plot corresponding to a subsequent iteration. In the first iteration, the model is trained on a small dataset of 7 samples and yields a high R-squared (R^2) value of 0.92. However, this iteration also comes with a Leave-One-Out Cross-Validation Root Mean Square Error (LOO CV RMSE) of 0.68 ± 0.49, indicating high variability in the model's prediction accuracy across different subsets of the data, which is not desirable. On the second iteration, although the R^2 is reduced, the error in the cross-validation analysis is also slightly

reduced, indicating better generalisation capacity of the model. This pattern remains in the third iteration model, with some points adjusting to the predicted values but other possessing more error in the prediction, suggesting that the smoothing properties of the Gaussian regression might be playing a role on the latter points.

It is also convenient to invert the analysis and plot the predicted uncertainty (as standard error of the mean) vs the actual observed uncertainty (Figure 2.10). Here we appreciate that for the first model the calculated R^2 for this prediction is negative. This means that there is a shift or bias in the predicted value, rendering the model as an overestimation. This behaviour is corrected on the next iterations, showing higher R^2 (0.65 and 0.69 respectively) and low LOO CV RSME, indicating that the second and third iteration models generalise much better in terms of uncertainty prediction.

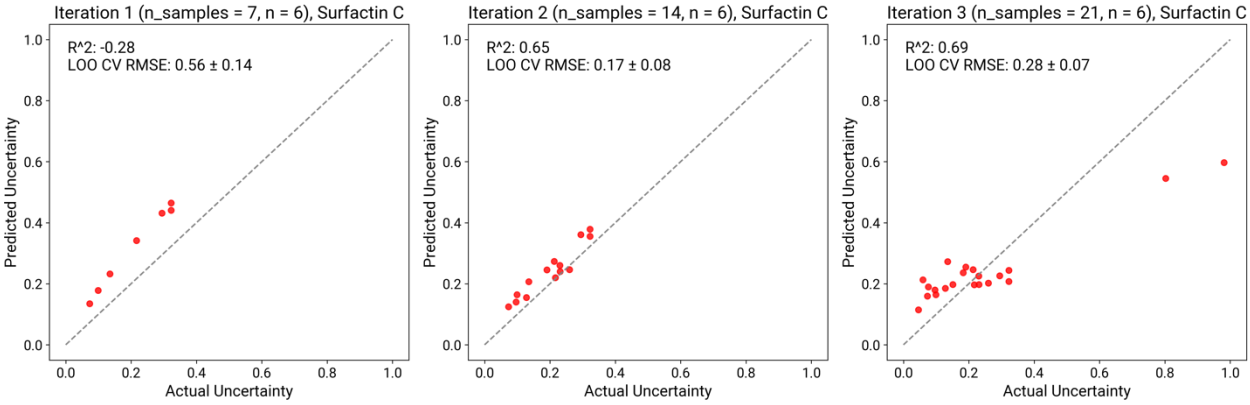


Figure 2.10. Predicted vs actual uncertainty values plot for the fitted Surfactin C Gaussian process regression (GPR) model after each iteration. The average standard error of the mean (s.e.m) for each sample considering replicates is indicated in the actual value (x-axis), while the mean s.e.m from the model on each of them is depicted in the y-axis. The mean from the actual samples and the predicted mean from the model is shown using the errors bars. This plot is a dual from the information given on Figure 2.9.

2.4. Discussion.

Surfactin production in a liquid culture involving several medium components is still a bioprocess challenge (Dobler et al., 2022). Therefore, we believe there is a need for high-throughput experimentation to optimize surfactin production and overcome the trade-off between population growth and biosynthesis. Our method achieves a 1.6-fold increase in surfactin titre, which in comparison to reported

increases using genetic engineering approaches (Xia & Wen, 2002), corresponds to a modest but significant improvement, given that only two factors were fine-tuned. Genetic approaches have reach between 1.4-fold improvement to 15.6-fold improvement in the extreme case, corresponding to the replacement of the native promoter P_{srf} with the strong chimeric promoter P_{g3} (Xia & Wen, 2002). In the case of solely performing medium optimisation, it has been reported titre improvements from 2 to 10-fold (excepting hydrocarbon addition experiments) and these are mostly related to tweaking mineral traces concentrations, such as Fe^{+2}, Mn^{+2}, Mg^{+2} or K^{+} in the medium. Interestingly, production can be very sensitive to these ions, resulting in decreasing titres for some conditions (Xia & Wen, 2002). Mg^{+2} and K^{+} concentrations are tested as part of the space-filling experiment in Chapter 4.

Upon consideration of the surface visualisation, it is observed that there are many combinations of carbon and nitrogen concentration around the optimal combination that exhibit a similar Surfactin C titre. Given the uncertainty associated with the model, the differences between the mean values of Surfactin C within the radius defined for the directional analysis are not statistically significant (p-value >> 0.05, ANOVA, 10 points taken randomly within the radius). Therefore, in a statistical sense, selecting any combination within that radius may result in a culture with a high Surfactin C titre.

Titre variation is a fundamental factor to consider when performing a Bayesian optimisation loop. Sources of noise include the stochasticity of the underlying biological process, errors introduced by liquid handling, instrumental variability, among others. Fundamentally, one can establish a trade-off where the total number of samples is reduced and the number of replicates per batch is increased, to gain confidence in the model's predictions and thus achieve sample-efficient optimisation and reliable production data that can be used in downstream analysis, as was shown by performing simulations for the optimisation of the Branin function. We took this path by using 6 replicates in our plate experiments, following a similar strategy as shown in Pandi et al., 2023 and Hashizume et al., 2023. Interestingly, in Hachizume et al., 2023, the total number of initial samples is quite high (N=236, with replicates included) in comparison to our framework (N=42, with replicates included), but the

number of factors is much higher than our case, so it is reasonable that more information is required for the active learning loop to converge in a small number of iterations.

The selection of an appropriate model and acquisition function to account for this noise is paramount. Thus, it is recommendable to characterize the statistical nature of the variation using small, preliminary experiments to choose the best options for the BO loop. We demonstrate that after the preliminary experiments to determine carbon and nitrogen ranges, deciding for using heteroskedastic models as surrogate was an appropriate choice since this kind of models can capture the observed variability in a flexible way.

If a complex surrogate model is chosen, such as random forest, ensembles or neural networks, feature importance or Shapley values may be important to rank the most important medium components contributing to the optimisation, as has been previously reported (Hashizume et al., 2023). Additionally, high-throughput multi-omics measurements have been intensively developed over the last years and they are the perfect complement to setup a seamless bioprocess optimisation that is at the same time metabolically informative. These are already in practice for synthetic biology and metabolic engineering (Roy et al. 2021), but any bioprocess areas could benefit of this approach. With time-resolved multi-omics data in hand, metabolic anisotropy studies might uncover interesting *in vivo* enzyme kinetics, novel regulatory systems, and unusual pathway flow redistribution, expanding the literature on bacterial catabolic control and inform metabolic engineering.

2.5. Conclusion

This study provides a step further in different directions. First, it realises the potential of active learning and more generally, iterative improvements methods to optimise production of biosurfactants, complex molecules with special production regimes. Second, it shows the importance of replication in active learning/Bayesian optimisation experiments when facing with highly variable measurements. Lastly, it is an example that optimisation pipelines can deliver rich data, that after further re-analysis, can reveal important aspects of the metabolism of *Bacillus* in conditions of lipopeptide production.

The approach taken in the experiment consider 2 factors to change simultaneously. The method can be adapted to consider more factors, giving a platform to solve complex optimisations. However, adding more factors make the medium preparation protocol tedious in terms of time and number of pipetting operations. This complexity also means that the volume transfer is prone to manual error after several hours of pipetting. To overcome this problem, a robotic platform should be developed for the liquid handling (Chapter 3). This platform will unlock optimisation of culture medium for Surfactin production in several factors, a result shown in Chapter 4.

2.6. Acknowledgements.

We would like to thank EdinOmics for providing access to the mass spectrometry equipment.

2.7. Data availability.

The raw data from the QqQ-MS runs, code scripts, microplate experiment layouts, and intermediate data tables are deposited in Zenodo (doi: 10.5281/zenodo.10203481). Additionally, the scripts are available at the Github repository www.github.com/rvalenciaaz/surfactin_bayesian_optimisation.

Chapter 3

A robotic platform for semi-automated iterative experimentation and additional protocols

Work conducted by Ricardo Valencia Albornoz[1]

[1] Institute of Quantitative Biology, Biochemistry & Biotechnology, School of Biological Sciences, University of Edinburgh, King's Buildings, Edinburgh, United Kingdom

Ricardo Valencia Albornoz (thesis author) performed the conceptualisation and testing of the platform, including liquid handling tests and electronics building. Any other reported protocols were also performed and conceived by RVA.

Outline

This chapter presents the technical advances in sampling, robotics and software to perform high-dimensional (multi-factorial) bioprocess experiments, enabling the main results in the subsequent Chapter 4. Additionally, we mention work related to optimising a mass spectrometry protocol to quantify surfactin using flow injection.

On the microbial culturing aspect of the project, we integrated advanced robotics, custom Python scripting, and state-of-the-art analytical techniques to create a system capable of conducting multi-factorial experiments in a semi-automated manner. The core of the experimental automation was achieved using the Opentrons OT-2 robot, which handled liquid transfers and was programmed to perform tasks with precision across different labware via Python scripts. These scripts, together with data analysis scripts, are the base of an Apache Airflow pipeline to streamline Bayesian optimisation iterations. To accommodate the experimental requirements in terms of volumes and available labware, the work was also supported by a series of 3D-printed custom labware, downloaded from the web or made in-house.

Finally, the system's capabilities were further enhanced with the integration of a robotic arm, enabling versatile manipulations within the Opentrons deck, avoiding the need to open the door and thus preventing potential contamination during the

preparation of culture medium mixes. The resulting pipeline complements existing pipelines on automated systems for experimentation, reducing human error, and increasing reproducibility. We also identified critical points in common laboratory equipment, both software and hardware, which can hinder fully automated, closed-loop experiments and we discussed future proposals on how to fill this gap.

3.1. Introduction

3.1.1. Optimisation of conditions for mass spectrometry

Optimising of mass spectrometry conditions is a crucial step in ensuring high-quality data acquisition (Hecht et al., 2016), which is essential for accurate metabolic analysis. Optimal parameter settings can improve signal stability, reduce noise, and enhance the detection of low-abundance species (Williams et al., 2024). Optimising a run in a triple quadrupole mass spectrometry, the main equipment used in this thesis, typically involves changing a mix of numerical and categorical factors that most likely affect peak quality, such as mobile phase flow rate, collision energy and chromatographic parameters (Jiao et al., 2002; Soler et al., 2006). In the literature, this procedure is generally carried by performing one-factor-at-a-time approach, because many of the parameters are commonly taken from previous runs in the laboratory or facility or are extracted from literature, providing a useful initial guess. However, depending on the nature of the protocol or the target compounds, optimisation may be performed from scratch, employing a design of experiment approach (Jiao et al., 2002; Riter et al., 2005).

Automated tools, such as the proprietary ThermoFisher software optimisation utility and the recent OptiMS (Williams et al., 2024), can facilitate process by systematically testing the parameters using large grids over a batch of samples, and it is especially effective when optimising conditions for molecule detection in a complex sample matrix.

3.1.2. Pipeline development for automation in biology.

Automation in biological research has seen significant advancements with the development of automated pipelines that streamline complex experimental workflows. These pipelines are designed to handle large-scale data processing,

reduce manual intervention, and improve reproducibility by breaking the process into smaller subroutines. Thus, by integrating automation with experimental design, researchers can achieve higher throughput and greater accuracy. This has proven particularly beneficial in experimental workflows for several “omics” including genomics (Socea et al., 2023), transcriptomics (Berglund et al., 2020; Santacruz et al., 2022), proteomics (Fu et al., 2023), and high-throughput screening (Michael et al., 2008), where the volume and complexity of data require sophisticated automation solutions.

Typical biocomputational pipelines uses packages that provides functions for assembling steps in a path fashion and to assess if any of these steps might be giving incorrect output or is not being executed. The latter could be because either a condition is not met in a subroutine, or in the physical case, the interaction with the laboratory equipment (including the machine operation itself) is not working reliably. The steps in the pipeline can be composed of many elements or subroutines, including scripts in several languages, calls to servers or cloud services and interacting with packages and modules for actioning machines and other laboratory equipment (Reiter et al., 2021).

Snakemake and Nextflow are popular frameworks for generating pipelines in bioinformatics and experimentation in general (Jackson et al., 2021). These frameworks use a specific syntax (Python/GNU Make-inspired and Groovy, a java-based language respectively) to monitor and execute the different subroutines/scripts and have native integration with the most common programming languages, such as Python and R. Snakemake has easy integration features with Conda (a python package manager) and containerisation frameworks such as Docker (Köster & Rahmann, 2012). Nextflow has a focus in adaptability and big data by integrating outputs on different section of the pipeline and it also provides templates for typical bioinformatics pipelines, providing ease of use, but also flexibility by user editing (Di Tommaso et al., 2017).

Apart from these packages, other general pipeline automation packages, or more generally, workflow management software, can be used, such as Apache Airflow (*Apache/Airflow*, 2015/2024), Luigi (*Spotify/Luigi*, 2012/2024) and Toil

(*DataBiosphere/Toil*, 2015/2024). These are partially or purely Python-based, lowering the learning barrier and has become popular in diverse industries. Additionally, they have the capability to interface with a graphical window, tracing the performance of the pipeline across the different steps.

Examples of pipelines combining *in silico* and i*n vivo* approaches in synthetic biology include DBTL cycles for enhanced microbial biomanufacturing of fine chemicals (Carbonell et al., 2018), the integration of computational metabolic modelling with experimental validation to optimize biosynthetic pathways (Choi et al., 2019), and the design of synthetic gene circuits using bioinformatics tools followed by in vivo implementation (Jones et al., 2022).

3.1.3. Opentrons platforms for biological experimentation

Opentrons, a US company building liquid-handling robots, focuses on accessible automation solutions for biotech companies and laboratories. These platforms are designed to be user-friendly, allowing researchers to automate a wide range of biological protocols using a graphical interface or via Python scripts. The modular nature of Opentrons systems enables the integration of various proprietary tools (such as shaker or temperature modules), open-source tools (such as 3D-printed racks), and reagents, making it possible to tailor experimental setups to specific research needs (Councill et al., 2021).

Opentrons robots have been described in protocols for synthetic biology, such as plasmid assembly and transformation (Storch et al., 2020; Hérisson et al., 2022; Bryant et al., 2023). In metabolomics, Opentrons platforms have been used for fast sample preparation (Nemkov et al., 2022). There are reports where an Opentrons robot has been modified by adding a camera, providing additional information on the pipetting process and/or for biological imaging purposes (Ouyang et al., 2022), as well as custom connections between the pipette head and mini-bioreactor lines for straightforward sampling and manipulation of cell cultures (Bertaux et al., 2022).

3.1.4. Custom platforms for biological experimentation an integration

In addition to commercial solutions like Opentrons, custom robotic platforms for biological experimentation have been developed to meet needs of specific

research projects in the technical side, but also providing a cost-effective solution (Alisch et al., 2018; Chory et al., 2021; Hamm et al., 2024). These platforms are often designed to integrate with existing laboratory equipment and data analysis tools, providing a cohesive and efficient workflow (Haby et al., 2019). The ability to customise these platforms allows biological researchers to control unusual environmental conditions, complement real-time data acquisition, or integration with novel assay technologies (Kurtser et al., 2021). Thus, custom hardware has been successfully employed in engineering biology protocols (Oliveira & Densmore, 2022), such as bioreactor engineering (Haby et al., 2019), physical sensors for measuring biological properties (Katunin et al., 2021), adaptable liquid handling on diverse scales, such as open-source microfluidics (Kong et al., 2017; Shin & Choi, 2021), among other applications. An interesting initiative on this matter are Frugal Self-driving Labs, which integrates several low-cost platforms to provide an accessible option for high-throughput experimentation via open-source software drivers (Lo et al., 2024).

3.1.5. Aims and objectives

This chapter considers the following aim:

- Automate and streamline experimental procedures using robotic platforms for large sample handling in iterative experiments.

To fulfil this aim, the following objectives are considered:

- Optimise the flow injection protocol regarding flow rate and mobile phase composition

- Benchmark pipetting volume performance of the Opentrons OT-2 robot

- Develop a pipeline for experimental Bayesian optimisation assisted by the OT-2 robot

- Build a robotic arm to assist with operations inside the OT-2 working deck

3.2. Materials and Methods

3.2.1. Optimisation of flow injection mass spectrometry protocol

To optimize the flow injection mass spectrometry protocol, the flow rate of the mobile phase was adjusted from 100 µL/min to 500 µL/min. The optimum was

determined at a rate that reduces peak broadening and provided a sufficient signal without excessive dilution. For the mobile phase composition, different mixtures of organic solvents (acetonitrile and methanol), with water were tested, including pure acetonitrile, acetonitrile/water 50:50 (or 1:1) and methanol/water 50:50 (or 1:1). The final mobile phase composition was selected based on the best performance for the target analyte, Surfactin C, in terms of signal-to-noise ratio in the extracted peaks. To prevent carryover and ensure system consistency, blanks and quality control samples were injected at regular intervals throughout the analysis.

The injection volume was set to 1µL for all runs. For simplification, 0.1% formic acid (acting as an additive) is added to every mobile phase mixture tested. For the mass spectrometer settings, the samples were ran using the parameters in Supplementary Table BT1 (Appendix B) and the scan for Surfactin C is given in Supplementary Table BT2. The peaks were extracted using the ThermoFisher QuanBrowser software into tables and analysed using a Python script. Plot were generated using a Matplotlib function (Hunter, 2007). Signal-to-noise ratio (S/N) (Equation 15) is calculated using the formula:

$$S/N = \frac{H}{N} \tag{15}$$

where:

- H = Height of the peak (signal),
- N = Noise level (measured as baseline noise, often the standard deviation or peak-to-peak noise).

3.2.2. Characterisation of the Opentrons robot

For the mixing of culture medium components in multi-factorial experiments, we employed an Opentrons OT-2 liquid handling robot (Opentrons Inc, USA). Testing of accuracy and precision of OT-2 pipetting was performed by transferring defined volumes of water from a 15 ml Falcon tube reservoir into a 48 well plate, considering 16 replicates and three volumes tested per plate. The volumes chosen for P300 were 30, 150 and 300 ul (tested in one plate), and for P1000 we tested 100, 500 and 1000 ul (tested in one plate), following the same reference values in the official Opentrons white paper (Opentrons, 2019). Then, we employed recently calibrated manual

pipettes (P200 and P1000, Gilson) to aspire the water in the wells, modifying the volume regulator in the pipette until all the water from the well is in the tip and no air bubble is present. The data was collected in a spreadsheet and plotted using Matplotlib package in Python (Hunter, 2007). Calculations were performed using numpy functions (Harris et al., 2020) in Python. Accuracy, also known as systematic error, was calculated using the formula (Equation 16):

$$\text{Accuracy } (\%) = \frac{(\bar{x} - V_{test})}{V_{test}} \times 100 \tag{16}$$

where:

- $\bar{x}$ = Mean of the measured volumes,
- V_{test} = True or expected volume.

while the precision was obtained by the following formula (Equation 17):

$$\text{Precision (CV } \%) = \frac{\sigma}{\bar{x}} \times 100 \tag{17}$$

where:

- σ = Standard deviation of the measured volumes,
- $\bar{x}$ = Mean of the measured volumes.

An additional set of tests, on the same volumes and for both pipettes, was performed, setting the aspiration and dispense speed to 0.5, to check for accuracy and precision in slower protocols (0.5 x standard aspiration/dispense rate is 43.46µL/s =21.73 µL/s).

3.2.3. Development of scripts for Opentrons OT-2 protocols

For robot manipulation, a series of Python scripts were created. These employ the Opentrons Python package to define robot operations as functions. The main functions used in the script were: *aspirate*, which draws liquid into the tip; *dispense*, which releases the liquid from the tip; *pick_up_tip*, to take a tip from a tip box; *drop_tip*, to dispose of a tip into the waste bin; *load_labware*, to specify the position and layout of labware on the deck so it can be easily referenced by other functions; and *load_instrument*, which loads the pipette to be used.

Every operation is associated with a specific pipette, loaded using the *load_instrument* function. *aspirate* and *dispense* require a labware reference, along with a well label and the volume to be moved. The speed at which this operation is performed can be optionally set. The *pick_up_tip* and *drop_tip* functions do not require arguments, as the tip box is linked to the active pipette in the *load_instrument* declaration. Respectively, a tip is drawn from the tip box or dispensed into the waste bin, and these functions must be consistently called in the correct order to perform the pipetting operation (e.g., a tip should be attached before performing the aspiration). The code checking system in the Opentrons library can assess these conflicts before setting up the run. The *load_labware* function generates the labware references to be called in the operation functions and takes as arguments a label from the Opentrons Labware Library and the deck position. If necessary, custom labware can be defined with a new label and added to the local labware library. For the *load_instrument* position, it requires a label for the attached pipette to use, in which arm is located, either “left” or “right”, and the tip rack that should be linked to draw the correct tips for the pipetting.

For general volume testing, the Opentrons scripts were written manually. In contrast, for the developed Bayesian optimisation pipeline(s), the robot scripts for the first microtube stock, filling the plate with the mix of components, and the transfer to the quenching plate were generated by a custom Python script, using string templates and text file manipulations. The runs are performed by uploading the robot instruction script into the Opentrons GUI software v. 7.0.2, which enables the connection of the computer with the Opentrons OT-2 robot through a USB cable. For the connection to be successful, both the robot and the GUI software must have the same firmware version, which can also be updated via the software. The GUI allows the visualisation of the running time, and a live listing of the current robot task and the remaining tasks.

3.2.4. Assembly and characterization of robotic arm

A robotic arm was assembled to perform operations inside the Opentrons deck without the need to open the main door, thereby reducing the chances of contamination in plates. The robotics requires 3D-printed parts in PLA, an Arduino UNO development board, a servo driver, one 20kg servo, three RC servos, one micro

servo, cables, gears, and several small components such as screws, switches, and elastic bands, as detailed in Appendix A.4. It is powered by a variable voltage source, adjusted for common operation at approximately 6V. The robot was designed by Kalton Serra, and he open-sourced the STL files and circuit plan, available at the following link: https://www.thingiverse.com/thing:6313449. The detailed building process of the robotic arm is presented in the Results section.

3.2.5. Development of a pipeline for synchronised operation of robotic platforms

The pipeline consists in several Python and R coordinated using an Apache Airflow workflow, which is composed of a Python script where the tasks are defined, and the workflow is declared as a directed acyclic graph. The full Bayesian optimisation execution and analysis protocol can be split into two pipelines: one for the initial sampling, and a second one for arbitrary iterations. The initial sampling pipeline considers 4 main scripts written in Python using the BoTorch library and other packages for data manipulation and calculation, such as Pandas and numpy. The second pipeline employs as mix of Python and R scripts, oriented to extract, analyse and reformat mass spectrometry abundance data for modelling and sample recommendations in the next iteration. Several csv files are used as information input to perform the tasks. More details on the implementation are presented in the Results.

3.2.6. Design and manipulation of 3D-printed models.

The design and manipulation of 3D-printed models were carried out using two primary software tools: AutoCAD Inventor and OpenSCAD. AutoCAD Inventor is a graphical software for 3D modelling. It was used for the initial design stages, specifically the plate cover and the cuvette holder. OpenSCAD, which is an open alternative that defines the shapes via code, was used for some advanced models, such as the one shown in Appendix A.2.

Both tools facilitated seamless integration with 3D printing workflows, with the final models being exported as STL files, which are compatible with most 3D printers. The models were uploaded to Ultimaker Cura v.5.5.0 for placement on the printer deck, running time estimation, and slicing, which generates a file with instructions for the printer nozzle. The models were then printed using a high-resolution 3D printer, the

Ultimaker 3, available at the University of Edinburgh Library (UCreate Studio). The material used was polylactic acid (PLA) filament, with a width of 0.4 mm, and no polyvinyl alcohol (PVA) support was employed.

3.3 Results

3.3.1. Optimisation of mass spectrometry for surfactin detection

Compared to other protocols, flow injection mass spectrometry is simpler in terms of the number of factors to optimise, since the sample is directly injected into the mass spectrometer without prior separation. Therefore, we focused on flow rates, which is the speed the mobile phase is pumped into the injection tube and to the mass spectrometer, and mobile phase compositions, considering mixtures of organic solvents with water.

For the first factor, it is observed that the flow rate does not affect the overall quality of the peaks, but rather affects the arrival time of the sample to the mass spectrometry inlet and the spread of the peak over the acquisition time. Thus, higher flow rates mean narrow peaks arriving earlier. After testing between 100 µl/min and 500 ul/min, we conclude that 200 µl/min keeps the totality of the peak inside the 1 min acquisition window, and it is also a good option in terms of solvent use (200 µl/min * (1 min acquisition + 1 min between samples) *60 samples = 72 ml per plate).

Regarding the second factor, we chose three solvent systems to test: pure acetonitrile, acetonitrile/water 50:50 (or 1:1), and a slightly more polar methanol/water 50:50 (or 1:1) (Figure 3.1). The main reason to choose these solvent systems is that preliminary testing with a Surfactin C standard (5mM) show that clear peaks can be retrieved from these mobile phases.

Considering this, we need to verify whether these solvent choices are effective when analysing real samples, where Surfactin detection might be affected by other compounds in the supernatant matrix, i.e., any other metabolites secreted by the bacteria or compounds that have been added as part of the culture medium. The raw sample analysed in the test is the spent medium of *Bacillus subtilis* DSM 3256 grown in LB. We also used two additional samples, which are simple dilutions to half and quarter concentrations, using water.

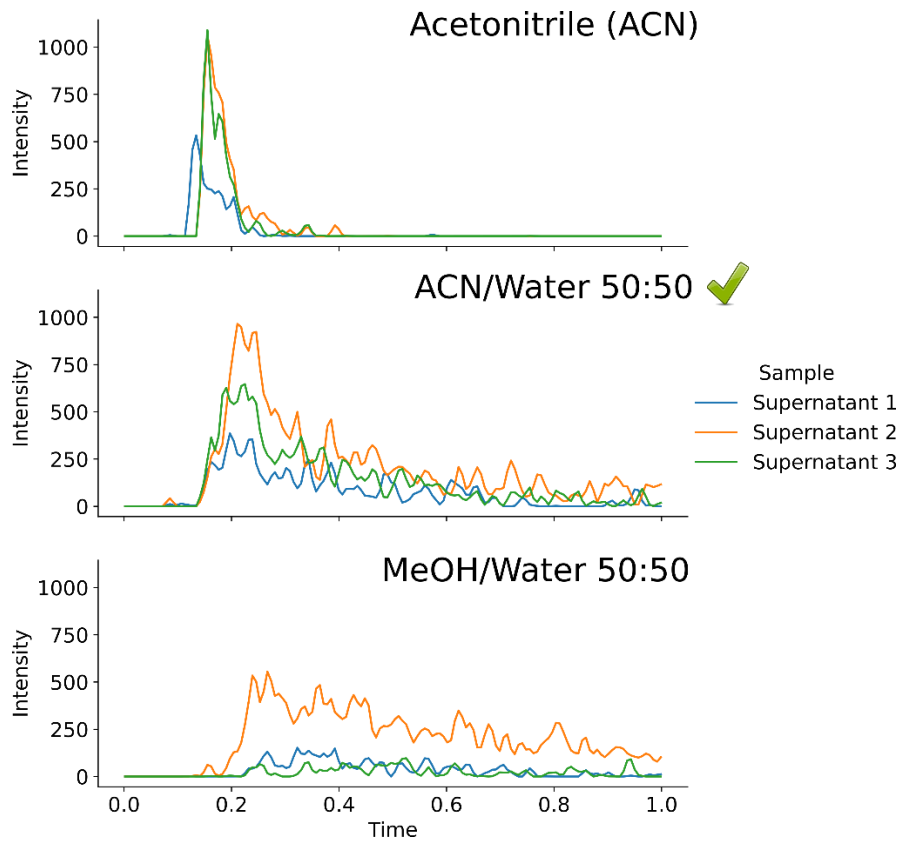


Figure 3.1. Comparison of different solvent mixes as mobile phase for a flow injection run of *B. subtilis* DSM 3256 supernatant from overnight M9 medium culture**.** Surfactin C was detected via single reaction monitoring, following the parameters (precursor/fragments mass) and collision values in Table BT1 and BT2. Supernatant 2 corresponds to the raw sample of *Bacillus subtilis* DSM 3256 supernatant grown in LB medium, while supernatant 1 correspond to a dilution 1:3 of this sample and supernatant 3 to a dilution 1:1 of this sample with water.

The signal-to-noise ratio for the Methanol/Water system was generally very low, at ~1.2:1, making it unreliable for quantification. In addition, the dilution effect is not consistent with the area under the curve (AUC). This issue is also observed in the Acetonitrile-only system, where, although the peaks are sharper and the signal-to-noise ratio is higher (~10:1), the dilution effect is not clearly reflected in the AUC. Therefore, the Acetonitrile/Water 50:50 mixture exhibits both desirable properties in terms of signal-to-noise (~6:1) and a (linear) dilution effect and has been selected for further mass spectrometry quantification.

3.3.2. Testing of liquid transfer and Python scripting for Opentrons OT-2 robot control.

For the mixing of culture medium components in a multifactorial experiment, we employed an Opentrons OT-2 robot. The Opentrons OT-2 is a two-arm liquid handling robot that facilitates pipetting and liquid transfer across a working deck, where the necessary labware is arranged. An automatic pipette can be attached to either of these arms, capable of aspirating volumes comparable to manual pipettes. Specifically, three single-channel pipettes, covering ranges of 1-20 µl (P20), 20-300 µl (P300), and 100-1000 µl (P1000), as well as a multi-channel pipette in the range of 10-50 µl (P50), are available for use with the robot. For the protocol described in Chapter 4, only the single-channel Generation 1 (GEN1) P300 and P1000 were utilised, as they can handle the volumes required by the protocol plan.

The first part of the Opentrons protocol involves mixing medium component stocks with water to create 2 ml pre-stocks. These pre-stocks contain the desired concentration (often 10x relative to the well concentration, though this can be adjusted), allowing for a fixed volume to be transferred from the microtube stocks to the plate wells (e.g., 80 µl for each component in a 7-component experiment). Given the pipettes being used, the smallest volume that can be manipulated is 30 µl. The rationale for this decision is that most volume transfers required to create the microtube stocks fall within the 30-1000 µl range. This means we can avoid constantly switching one of the pipettes for the P30 pipette during the protocol, which could introduce errors and problems, such as contamination, into the pipetting process.

Now, to be certain of the magnitude of these errors and about the reported accuracy, we created a manual Opentrons script to test the reported values by transferring water between a reservoir and plates well and measuring the water in the wells with a calibrated manual pipette. In every case, 18 replicates (or wells) were used (Table BT4 and BT5, Appendix B). For P300, the following volumes were tested in one plate: 30, 150 and 300 ul with reported Opentrons white paper accuracy of 3%, 1% and 0.6%; and precision, as coefficient of variation, of 1.5%, 0.4%, and 0.3%. For P1000 we tested 100, 500 and 1000 ul in one plate, with reported accuracy of 2%, 1%, 0.7%; and CVs provided by the company of 1%, 0.2 %, and 0.15%.

In terms of accuracy, the pipetting consistently produces slightly smaller volumes than the expected values. Specifically, the estimated accuracy for the P300 testing was -3.55%, -3.16%, and -2.56% for 30, 150, and 300 µl, respectively. In the P1000 testing, the results show an accuracy of -2.22%, -2.82%, and -2.78% for 100, 500, and 1000 µl, respectively (Figure 3.2). In terms of random error, or precision, the pipetting is relatively consistent, with the coefficient of variation in the P300 pipette test being 1.07%, 0.76%, and 1.02% for 30, 150, and 300 µl, and 1.46%, 1.33%, and 1.15% for 100, 500, and 1000 µl using the P1000 pipette.

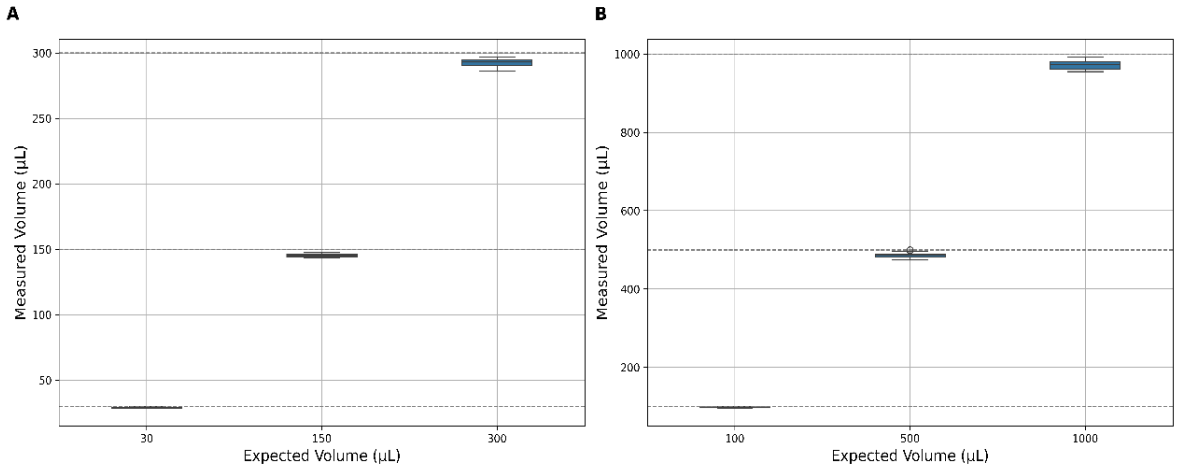


Figure 3.2. Test of pipetting accuracy on the Opentrons OT-2 robot at normal aspiration and dispense speed (43.46µL/s). **A)** The results of test with GEN1 P300 single pipette are shown as boxplots B) The measurements for GEN1 P1000 single pipette are shown.

A slight improvement, in line with reported values for both accuracy and precision, is achieved when reducing the pipetting aspiration and dispense speed to 0.5 of the normal speed. In this case, the accuracy for the P300 is in the range of -2.35%, -1.82%, and -1.39%, while for the P1000, we obtained -1.88%, -1.80%, and -1.13% (Figure 3.3). Precision, as CV, is in the range of 0.71%, 1.06%, and 0.77% for the P300, and 1.70%, 0.79%, and 0.60% for the P1000. The consistently lower-than-expected mean volume suggests that the pipette aspiration system requires some maintenance, which has been carried out since 2020. However, Opentrons does not have a technical service office in the UK, which involves additional time and costs. The values using the slower aspiration and dispense rate are similar to those reported by the company and are considered acceptable for running protocols.

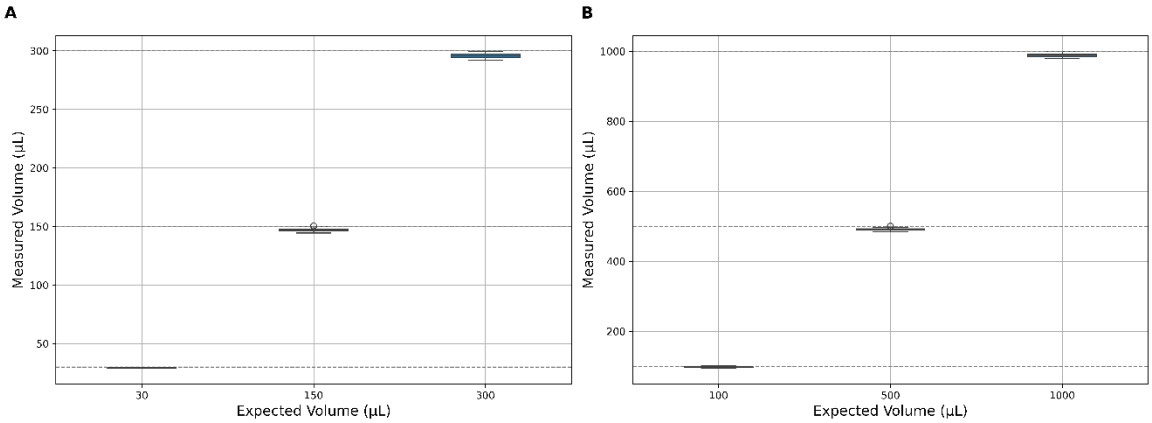


Figure 3.3. Test of pipetting accuracy on the Opentrons OT-2 robot at aspiration and dispense rate of 0.5 (21.73 µL/s). **A)** Results of test with GEN1 P300 single pipette. **B)** Measurements for GEN1 P1000 single pipette.

Table 3.1. Summary of accuracy and precision values for the testing of GEN P300 and P1000 pipette, in normal speed and slow speed (0.5 rate)

Pipette	Volume (ul)	Accuracy % (normal speed)	Precision % (normal speed)	Accuracy % (slow speed)	Precision % (slow speed)	Accuracy % (company)	Precision % (company)
P300	30	-3.55	1.07	- 2.35	0.71	+/- 3	1.5
P300	150	-3.16	0.76	- 1.82	1.06	+/- 1	0.4
P300	300	-2.56	1.02	- 1.39	0.77	+/- 0.6	0.3
P1000	100	-2.22	1.46	- 1.88	1.70	+/- 2	1
P1000	500	-2.82	1.33	- 1.80	0.79	+/- 1	0.2
P1000	1000	-2.78	1.15	- 1.13	0.60	+/- 0.7	0.15

A typical protocol in the Opentrons involves several pieces of labware that are placed in designated spaces on the deck (Figure 3.4). Before starting a run, the robot requires two steps of calibration. The first one, which is done monthly, corresponds to the main calibration. The second type of calibration is a labware position check. This check takes less time than the general calibration and is performed before every run to ensure fine alignment of the pipette head with the tips and wells. Notably, the adjustment is three-dimensional, meaning that not only the position of the tip relative to the deck can be modified, but also the height of the tip relative to the labware. In

this regard, a non-trivial adjustment was made during the Labware check step in the Opentrons application. When the tip enters a microtube or 96-deep well plate, if the tube or well is nearly full, the transferred volume plus the volume of the tip would cause the liquid to overflow. To avoid this, the vertical placement of the tip should be adjusted 10 mm above the reference line, which corresponds to the upper edge of the tube or plate.

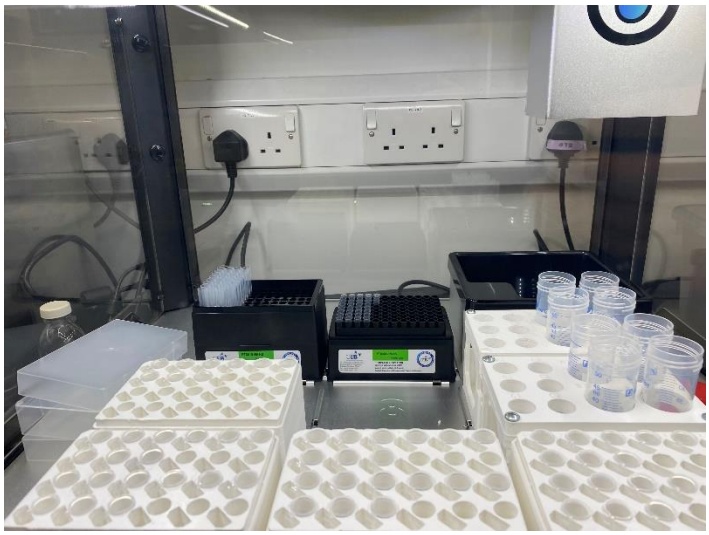


Figure 3.4. Picture of the Opentrons OT-2 robot in a step during a typical run.

3.3.3. Consolidated pipeline for data analysis and main robot operations

In an effort towards full automation, the main data analysis scripts and the Opentrons scripts were combined into a pipeline that can be run sequentially. However, in practice, the protocol must be split in two (sub-)pipelines: one only for the initial sampling and first run with the Opentrons robot, and another for mass spectrometry data analysis, sample recommendation for the next iteration using the Bayesian optimization routine, and the subsequent Opentrons OT-2 run. Here, I described how the subroutines in both pipelines work. Code is deposited in a Github repository described in the Data Availability section. Snippets of the code will be shown occasionally to highlight aspects of the algorithms.

3.3.3.1. Pipeline for initial sampling and first Opentrons run.

The pipeline for initial sampling and initial run is depicted in Figure 3.5. It starts with the script 1_initial_samples.py, that is used for initial sampling from a space filling

design, either Latin Hypercube sampling (LHS) or Sobol sampling (Figure 3.6). For LHS, the script uses a pyDOE2 function, while the Sobol sampling is generated using the SobolEngine function from BoTorch. Scipy has also recently implemented quasi-random sampling functions (Scipy, 2024), that can be easily adapt to this script. A version using Scipy is available in the ALTERNATIVES folder in the Github repository. The input for this script is the number of dimensions/factors *d* and the number of initial samples, together with a file containing labels of the components. These labels make the columns in the sample table. By default, both space-filing methods propose samples in the cube $[0,1]^d$.

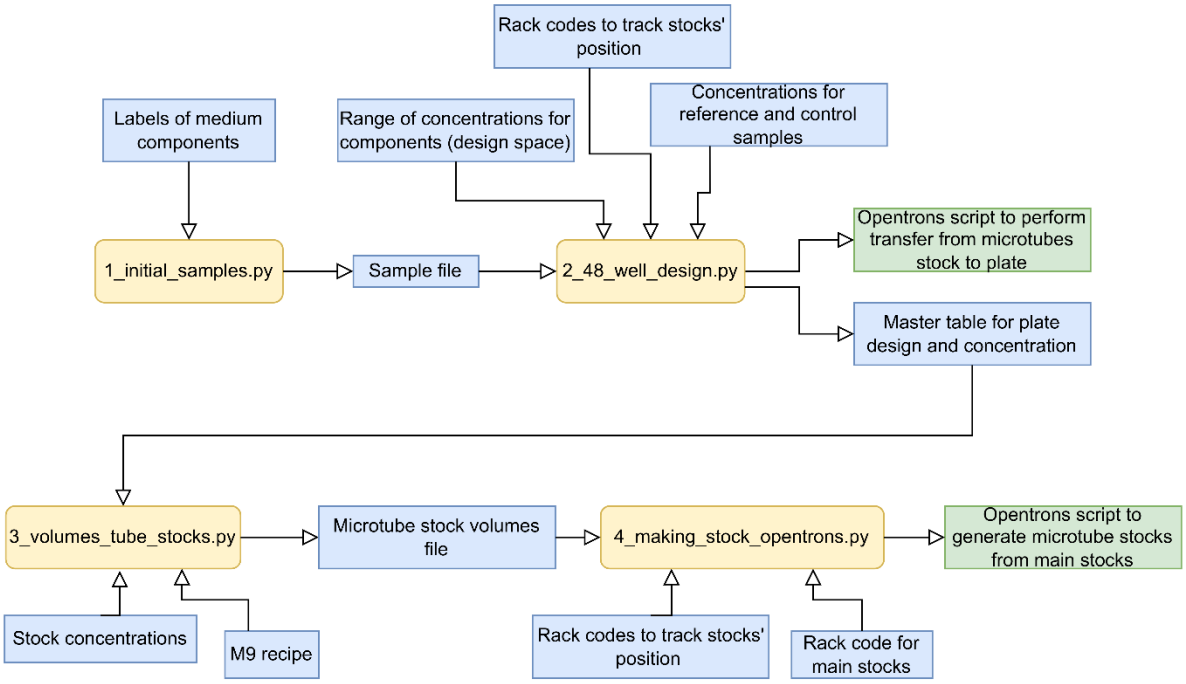


Figure 3.5. Initial sampling and initial Opentrons run pipeline. The scripts are coloured in yellow, while csv files employed in the process are blue. The generated Opentrons OT-2 scripts in the steps are depicted in green. If a big space-filling design is required instead of iterative experimentation, the –big argument in 1_initial_samples.py can be activated to split a space-filling design in small subsets for batch testing.

If, instead of an iterative experiment, you want to obtain a large space-filling design and then split these samples into batches, the script 1_initial_samples.py includes options –big and --batches that can be specified in the input arguments to create a space-filling set and divide it into the desired number of batches. Regardless of the type of experiment, it generates a sample file with index 0 (representing the first

iteration or first batch) corresponding to coordinates in the design space $[0,1]^d$ of the initial samples.

```python
# defining bounds
bounds = torch.stack([torch.zeros(dim), torch.ones(dim)])
if args.method == "sobol":
    samples = draw_sobol_samples(bounds=bounds,
    ↪ n=init_samples, q=1).reshape(init_samples, dim)
else:  # LHS method
    samples = torch.tensor(pyDOE2.lhs(n=dim,
    ↪ samples=init_samples))

#extract labels of components
labels_df = pd.read_csv(col_labels)
col_labels_list = labels_df["Component"].to_list()

#save initial sample table with components as columns
init_table = pd.DataFrame({"sample": [f"{iteration}_{i+1}"
↪ for i in range(init_samples)]})
init_table[col_labels_list] = samples
init_table.to_csv(f"samples/{iteration}_samples.csv",
↪ index=False)
```

Figure 3.6. Snippet of code from the 1_initial_samples.py script. In the snippet, a sample is generated depending on the chosen method. Then, the labels are extracted and a table with these initial samples coordinates is saved, with columns as components/labels.

The pipeline proceeds to the 2_48_well_design.py script, which performs three tasks: designing the layout of a 48-well plate, generating the script for Opentrons OT-2 liquid transfer from stock microtubes to the 48-well plate, and producing a CSV file with the sample order for the mass spectrometry run, which can be uploaded to theSequence Setup in the Thermo Scientific Xcalibur software. Additionally, it generates a master table detailing the distribution of the components across the wells, facilitating information for the next stages.

To determine the appropriate volumes for creating stock solutions in microtubes, the 3_volumes_tube_stocks.py script calculates the required microtube or 96-well stock volumes based on predefined stock concentrations, which are sourced from 15ml Falcon tubes in a deck rack during the Opentrons run. The stock in the Falcon tubes comes from big volumes stocks kept in Duran bottles, which are made at the start of the whole experiment and are not made again until the complete Bayesian optimisation loop or space-filling experiment finishes.

Finally, the 4_making_stock_opentrons.py script generates a script for the Opentrons OT-2 liquid handling system to automate the preparation of microtube stocks from the Falcon tube rack to the microtubes or deep 96-well plates. By automating this process, the script reduces the potential for human error and increases the efficiency of stock preparation (from approximately 6 hours manually to 2 hours of robot time).

A summary of the main characteristics of the pipeline scripts can be checked in Figure 3.7.

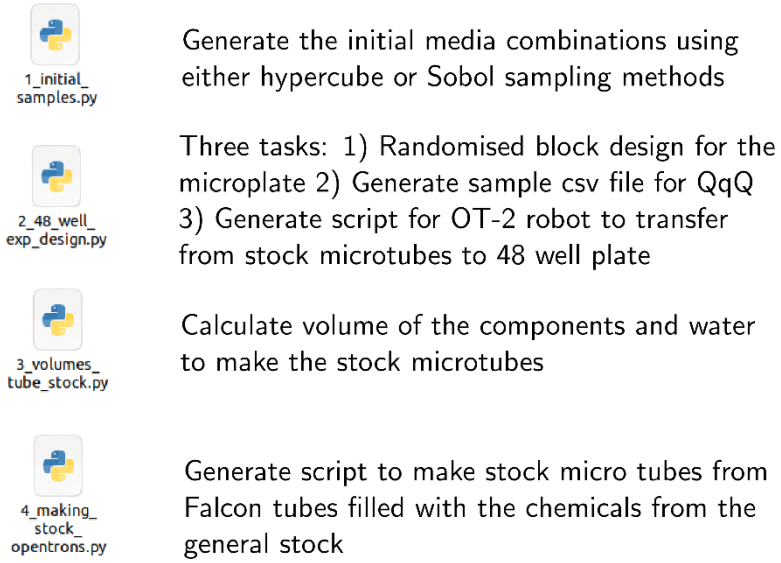


Figure 3.7. Brief description of the scripts of the initial pipeline. 4 out of 4 are Python scripts, using functions from diverse libraries including Numpy, Pandas, PyDOE2 for classical design of experiments and BoTorch for Sobol sampling.

3.3.3.2. Pipeline for an arbitrary iteration in Bayesian optimisation experiment

In the case of the pipeline for an arbitrary iteration in the Bayesian optimisation loop, the process consists of five main scripts (Figure 3.8). It begins with the script a1_processing_files_rawrr.R, which manages the conversion of Thermo RAW files into a usable format using the rawrr package in R (Kockmann & Panse, 2021). This script extracts multiple reaction monitoring (MRM) scans along with their corresponding intensities and retention times. The resulting table, referred to as the "hypertable" in the file naming, provides a "melted" data table containing sample identifiers, scans, and intensity values for subsequent analysis steps.

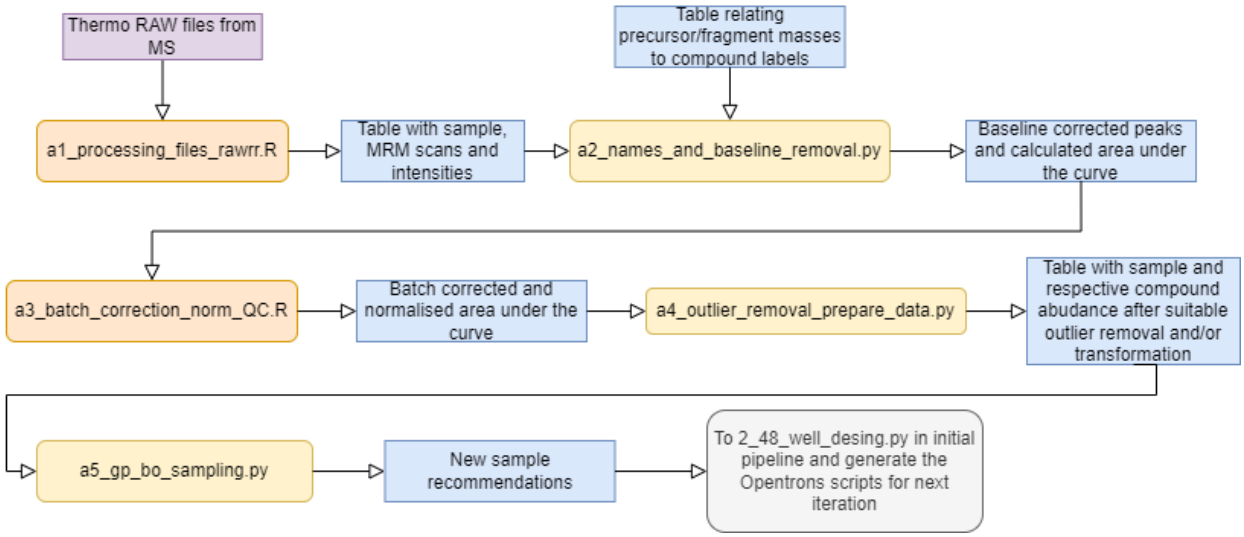


Figure 3.8. Pipeline for subsequent iterations. It starts with the manipulation of the raw files from the mass spectrometer, then performs several operations, including correction and normalisation, and ends with a new Opentrons robot action script for microtube stock and plate generation, linking to the next iteration.

Following the extraction of raw data, the script a2_names_and_baseline_removal.py is executed to associate the precursor and product mass information from the scan labels with known compound labels. In this step, the script also performs baseline correction of the peaks using the asymmetric least squares algorithm in the pybaselines package (Erb, 2024). After this correction, the area under the curve (AUC) for each peak is calculated, representing compound abundance.

Once baseline correction and compound naming are completed, the script a3_batch_correction_norm_QC.R is employed to address common batch effects in experiments with several iterations and a large number of samples per batch, as in our case. By applying batch correction using the QC-RSC algorithm provided by the pmp package (Kirwan et al., 2013), the script adjusts the AUC values, ensuring that intensities across different experimental runs are directly comparable. After the correction, a probabilistic quantile normalisation is applied (Dieterle et al., 2006; Wu & Li, 2016; Sun & Xia, 2024), improving data consistency and reproducibility, particularly when dealing with multiple batches of samples (Figure 3.9).

```r

# Create SummarizedExperiment object
new_df <- SummarizedExperiment(assays = list(counts =
    target_df)) # target_df is the dataframe of AUC data
    after baseline correction, for experimental samples
    and analysed compounds

# Perform batch correction
corrected_data <- QCRSC(df = new_df, order =
    sample_order, batch = batch, classes = class, spar =
    0, minQC = 4)

# Plot correction results
plots <- sbc_plot(df = new_df, corrected_df =
    corrected_data, classes = class, batch = batch,
    output = "correction_plot.pdf", indexes = seq(1, 58))
    #indexes indicates which features to plot

# Perform PQN normalization
data_normalised <- pqn_normalisation(df = corrected_data,
    classes = class, qc_label = "QC")

# Save the batch-corrected data
write.csv(data_normalised, paste("batch_correction/",
    iteration, "_batch_corrected.csv", sep = ""))
```

Figure 3.9. Snippet of code from the a3_batch_correction_norm_QC.R script.

To further refine the dataset, the pipeline incorporates a4_outlier_removal_prepare_data.py, a script designed to detect and remove outliers from the normalised data using the interquartile range (IQR) method (± 0.95 * IQR). Outliers can result from experimental errors, or other technical factors that could skew the analysis. The same script then links the normalised abundances with the initial coordinates provided by the sample file, either from the initial sampling pipeline or from a previous iteration.

The final script in the pipeline, a5_gp_bo_sampling.py, models the abundance data given the sample coordinates using a Gaussian Process regression model. The acquisition function is then calculated and maximised to suggest recommendations for combinations to test in the next iteration. This step connects to the second script in the initial samples' pipeline, thereby generating the mass spectrometry and Opentrons OT-2 files, and linking with the next iteration in the Bayesian optimisation loop.

As with the first pipeline, brief descriptions are provided in Figure 3.10 for quick reference.

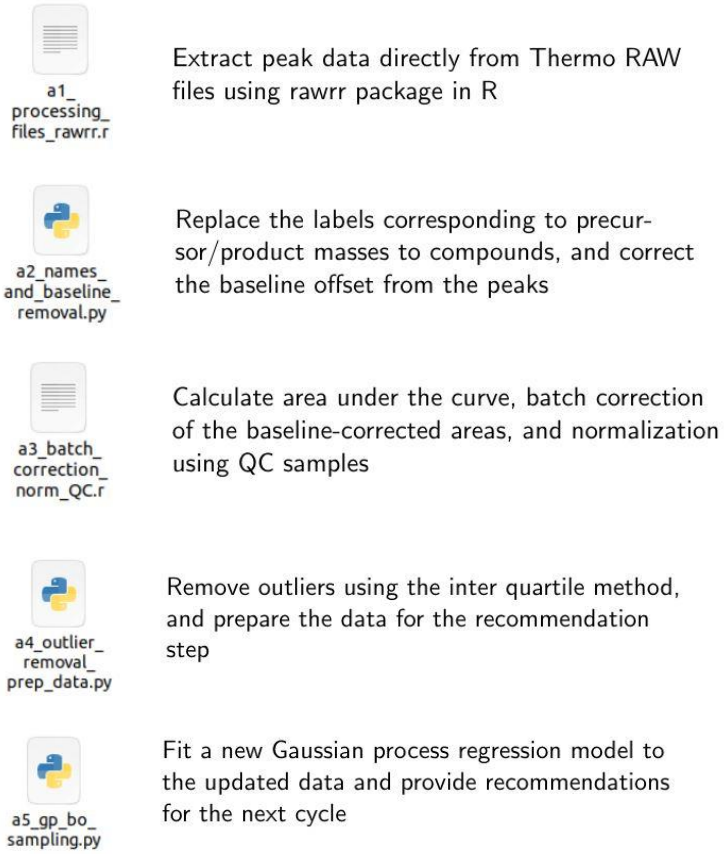


Figure 3.10. Description of files for the iteration pipeline. 3 scripts are written in Python, while 2 are in R, due to specific libraries being used, such as rawrr for Thermo RAW files data extraction and pmp for batch correction using the QSRSC algorithm.

There is one Opentrons operation script, quenching_protocol.py, which is currently not integrated to any of these pipelines. This script is used for the quenching protocol, where the supernatant from microtubes or the microplate is transferred to a cold methanol 96-deep well plate.

3.3.3.3. Consolidating pipelines using Apache Airflow

As seen for the diagrams, the pipelines are complex, requiring several interconnected parts. To reliable execute these pipelines, we propose the use of Apache Airflow workflow to consolidate the necessary steps. The pipelines are encoded as directed acyclic graphs (DAG), and this is represented in the code by

simply declaring the variables related to the execution of each of the scripts and linking them using the >> symbol at the end of the scripts (Figure 3.11). The >> means an edge in the pipeline graph, and gives the directionality, i.e, which scripts should finish before starting the next one.

```

from airflow import DAG
from airflow.operators.bash_operator import BashOperator
from datetime import datetime, timedelta
import os
import logging

#... Code defining DAG and function for python script execution in a mamba environment

# Task 1: Generate Initial Samples
generate_initial_samples = BashOperator(
    task_id='generate_initial_samples',
    bash_command=run_python_script_in_conda_env(
        script_name='1_initial_samples.py',
        conda_env=conda_env,
        script_args='--dim 7 --labels M9_7F_labels.csv --samplesper 7 --big --batches 6
        ↪  --method sobol',
        working_dir=working_dir,
    ),
    dag=dag,
)

# Task 2: Run Experiment Design
plate_design_ot2_ms = BashOperator(
    task_id='plate_design_ot2_ms',
    bash_command=run_python_script_in_conda_env(
        script_name='2_48_well_exp_design.py',
        conda_env=conda_env,
        script_args='component_data_M9_space.csv M9_rack_codes.csv
        ↪  M9_reference_samples.csv --iteration 0 --dim 7 --maxvol 80 --maxwell 800',
        working_dir=working_dir,
    ),
    dag=dag,
)

# Task 3: Calculate Volumes for Tube Stock
volumes_tube_stock = BashOperator(
    task_id='volumes_tube_stock',
    bash_command=run_python_script_in_conda_env(
        script_name='3_volumes_tube_stock.py',
        conda_env=conda_env,
        script_args='--dim 7 --iteration 0 --voltube 2000 --maxwell 800 --maxvol 80
        ↪  --stock stock_table_M9.csv --medium M9_preparation.csv --space
        ↪  component_data_M9_space.csv',
        working_dir=working_dir,
    ),
    dag=dag,
)

# Task 4: Prepare Stock for Opentrons
making_stock_opentrons = BashOperator(
    task_id='making_stock_opentrons',
    bash_command=run_python_script_in_conda_env(
        script_name='4_making_stock_opentrons.py',
        conda_env=conda_env,
        script_args='--iteration 0 --rackdata M9_rack_codes_all.csv --bigdata
        ↪  M9_big_codes.csv',
        working_dir=working_dir,
    ),
    dag=dag,
)

# Sequence for task execution
generate_initial_samples >> plate_design_ot2_ms >> volumes_tube_stock >>
↪  making_stock_opentrons

```

Figure 3.11. Example of Apache Airflow code to execute and monitor the initial sampling pipeline.

After running the Airflow client, the pipeline is available for control in a graphical interface that can be open in a browser by calling a local address (Figure 3.12). The interface starts with a login window, where the username and password specified in the client (can be modified by the programmer) is entered. It will open a dashboard with the pipeline as a list. A pipeline is determined by its name and on the right part, a series of indicators are shown. The first set of indicators compiles the historical statistics of the pipeline, like how many time the pipeline succeed, how many times it fails, and other error cases. Meanwhile, the larger second set of indicators, together with the time stamp, shows the status of the pipeline in real time, where each kind of error (execution errors, connection errors, etc.) can be individualise for easier visualisation. The pipeline can be started by clicking the play button in this dashboard.

If we click in the name of the pipeline, we are taken to a detailed view of the pipeline's execution. This view provides a graphical representation of the DAG, showing each task as a node and the dependencies between tasks as arrows. Each node is color-coded to represent its status: green for success, red for failure, yellow for running, and grey for tasks that are yet to execute. This visualization makes it easy to track the progress and debug any issues.

By clicking on any node, we can access detailed logs and metadata about that task's execution, such as the start time, duration, and any errors that occurred. Additionally, Airflow allows users to retry failed tasks or manually trigger specific tasks without re-running the entire pipeline.

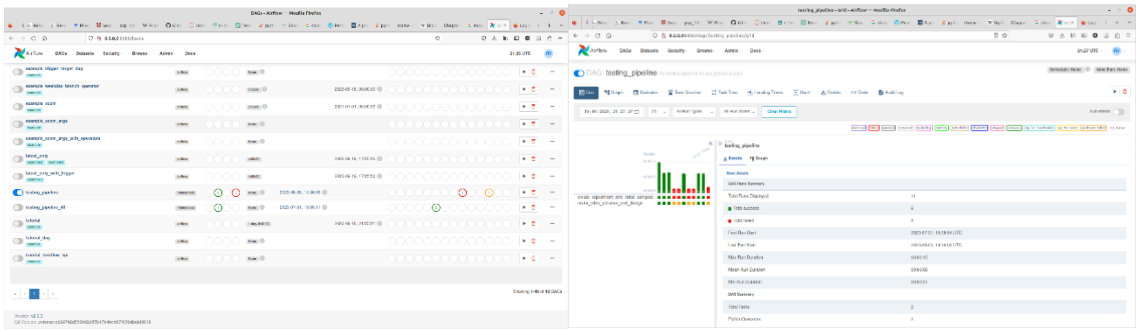


Figure 3.12. A view from the console in Apache Airflow. On the left side, the list of pipelines or DAGs is available for inspection. There, the pipeline execution can be tracked live, and a history of runs is displayed using a semaphore system. The pipeline can be executed using the play button. On the right side, the internal pipeline configuration is shown. The bar plot

indicates execution times for each of the different steps and the colour determines if the step was successful o an error was found. The steps can be organised in a graph plot and the associated Python or bash commands can be inspected.

3.3.4. A robotic arm for operations inside the Opentrons robot deck

Opentrons protocols involves some kind of stock tube movement. However, this might happen at the same time as the mixing procedure. Normally, the solution involves quickly opening the front door, perform the tube move and close the door. It is important to remember that the UV lamp was turn on before the mixing procedure for around 30 min to keep the inner space as sterile as possible (Figure 3.13). Therefore, even if the steps are done quickly (1 min), a considerable volume of air goes into the deck and may cause contamination in the plate well. To avoid this situation, we built a robotic arm that can perform simple operations inside the Opentrons deck, such as moving tubes, without opening the front door.

Figure 3.13. UV lamp located in a corner of the Opentrons robot. It can be operated via remote control. When the ON button is clicked, it takes 5 mins to start. On that time, alarm beeps to warn surrounding people about accidental exposition. The robot cover should be always used when operating the UV lamp.

The robotic arm was designed by Kalton Serra, who kindly shared the necessary files for 3D printing (in STL format) and a description of the circuitry online. STL files

are commonly used to save 3D models and can be opened by most proprietary and open-source 3D printing software, allowing dimensions and shapes to be easily extracted. Additionally, a tutorial for the general assembly is available on YouTube, with the link provided in the Data Availability section and the list of materials is available in the Appendix A.4. The criteria for choosing this design over others included size, simplicity of parts, simplicity of the circuit, flexibility, the choice of development board (with a preference for Arduino/Raspberry Pi options), and online reviews of performance.

The assembly process is shown in the following pictures. It begins with 3D printing the necessary parts. We used the university-wide service provided by UCreate Studio, located in the University of Edinburgh's main library, for the printing. The parts were printed from PLA material using an Ultimaker 3 printer. The filament width was set to 0.4 mm.

Then after collecting the printed parts, the servo cables should be extended enough to be able to reach the base of the design. The length will depend on the corresponding joint with the gripper servo being farther away from the base and therefore it has a longer cable. Then, the gripper servo is fit into the grip support and the cables move outside (Figure 3.14).

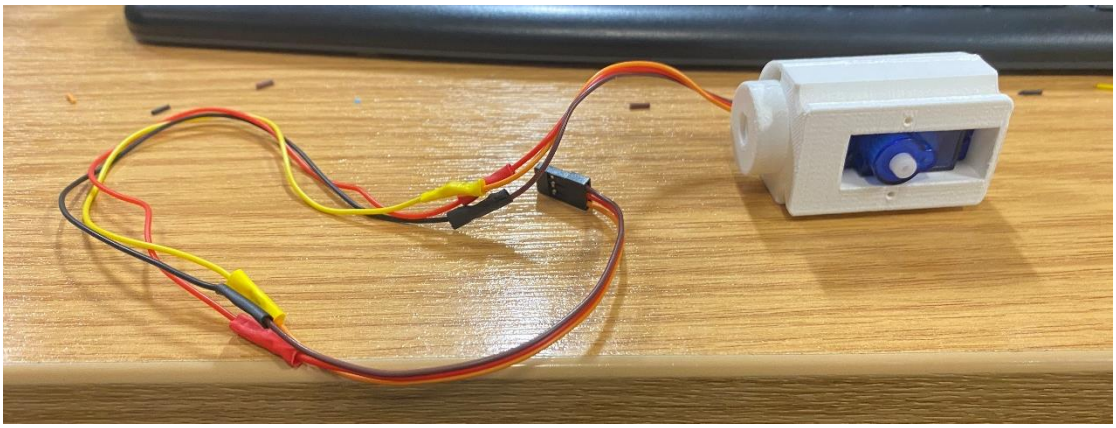


Figure 3.14. The small servo is installed on the grip case and the cable extension is visible.

Then, two of the RC servos are fitted into the forearm part. One of them is connected through gears with the gripper side, enabling a wrist-like motion of the gripper. The other one will provide the elbow joint, connecting with the rest of the arm. Additional gripper parts can be added at this stage, and a elastic band is used to keep the grip close, just for sake of manipulation easiness. Similar to the gripper case, all cable from the 3 servos should be displaced and taken out at the base hole in the forearm part (Figure 3.15).

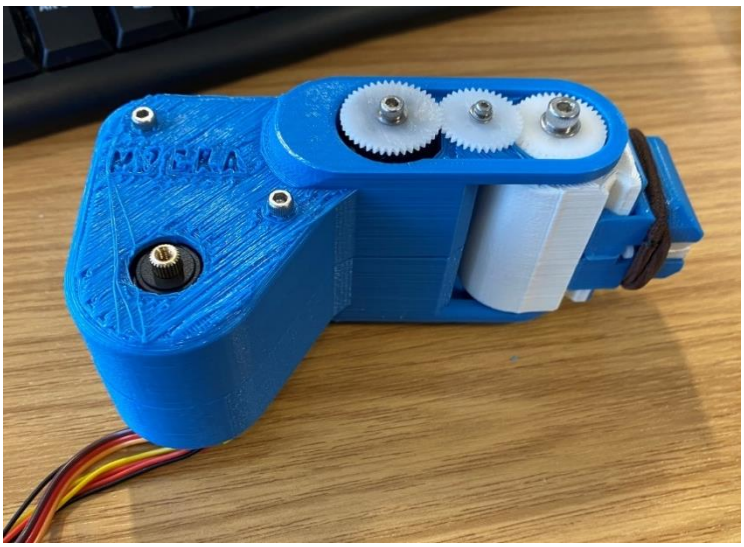


Figure 3.15. Forearm section with integrated grip part.

After working in the forearm, the focus turns in fitting the section between the elbow and the shoulder. There the 20kg servo is fitted as the should joint. The connection between forearm (blue) and the white section is due to a plastic support calls horns (Figure 3.16). Horns has a orifice, where the servo shaft is inserted. Some caution is needed, since the shaft, with its dented shape, moulds the inner part of the horn's orifice. Engaging and disengaging the servo shaft in the horn orifice may cause the orifice to get rounded and no longer provide a friction surface. Therefore, the servo shaft motion does not connect properly.

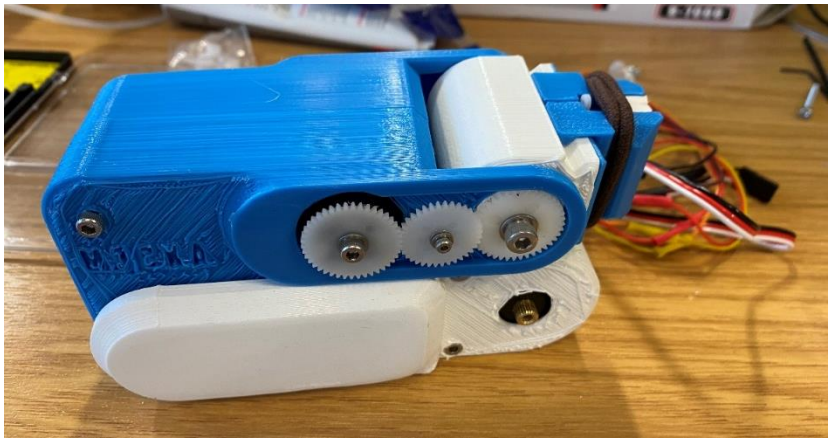


Figure 3.16. Main arm assembly displaying the shoulder 20 kg servo shaft.

This proto-arm is then assembled to the base using the green and blue shoulder joint supports. The last RC servo is attached to the base and the shaft is connected to a cross-shaped horn. The horn link with the base joint and provides the “almost” 360 degrees move of the base joint (it is constrained by the cables torsion) (Figure 3.17).

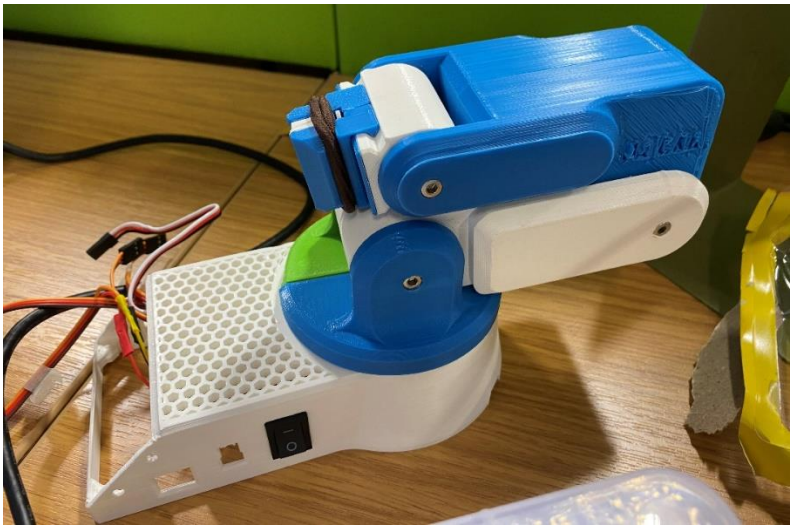


Figure 3.17. The shoulder joint is attached to the base.

After lining the cables towards the bottom of the base, the connection of the servos with the servo driver and the Arduino board is performed (Figure 3.18). The circuitry follows the diagram provided (Figure 3.19) and can be summarised as follows. The servos are connected to the servo driver in the corresponding pins. On the right side, the potentiometers from the mimic controller are connected to the Arduino board using the analogue input pins, except for the grip button that is connected with a digital pin in the board. Then, there is a group of jumpers that power

and link the servo driver board using the power pins output of the Arduino board. Thus, when the Arduino board is connected to the computer via USB cable, it can also power the servo driver and can execute code in the microcontroller present in the code. However, this power is not enough to move the servos. Thus, to provide enough power, a plug connector (together with a switch) is linked to the servo driver. This plug connector then links to a variable voltage source, adjusted for common operation at ~6V. In the case, the connection to the potentiometers is broken at the rear back, so the mini-arm controller can be connected to these pins.

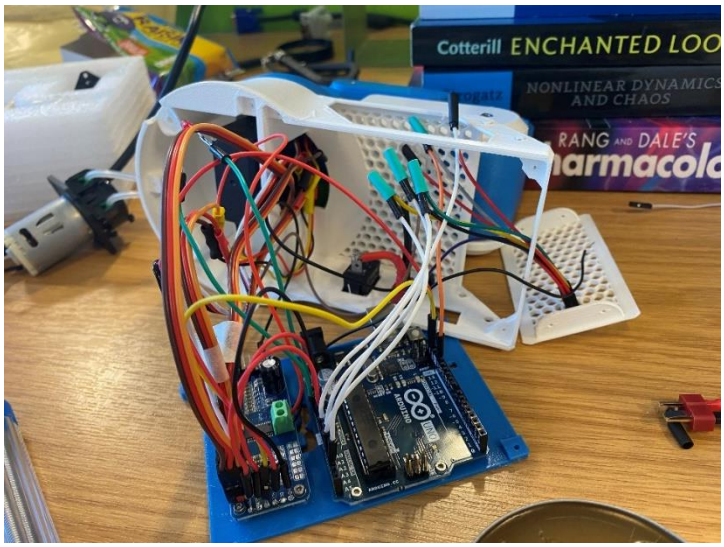


Figure 3.18. A view to the circuitry of the robotic arm. The Arduino board and the servo driver are screwed to the 3D-printed bottom. Then, the circuitry can be fitted inside the robot's body.

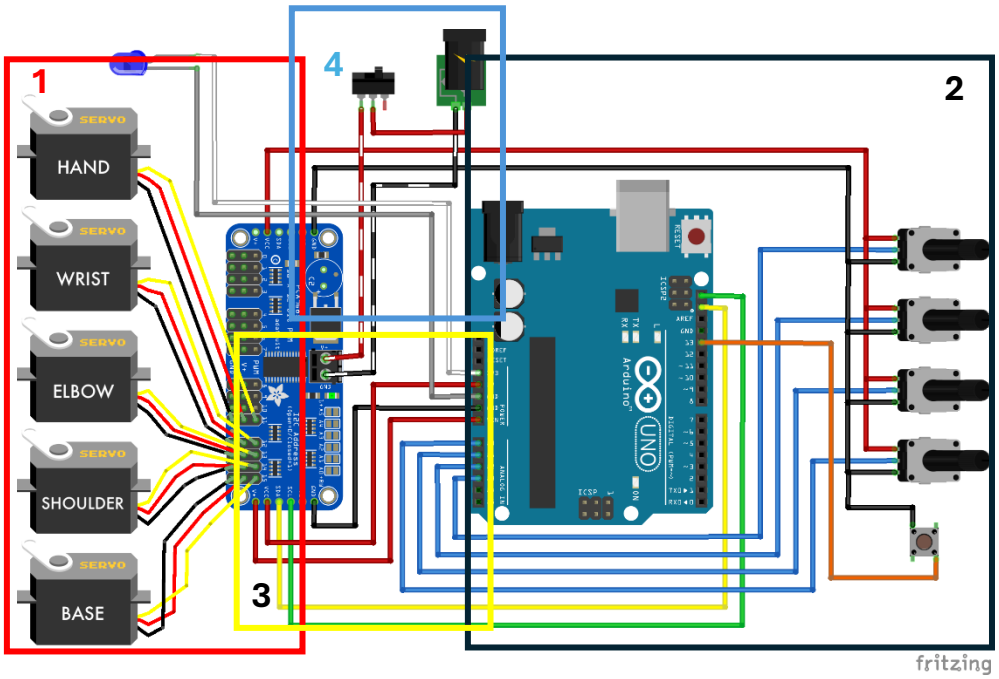


Figure 3.19. Diagram for the connections in the robotic arm. The diagram shows the wiring of the different components. To aid in the analysis of the circuit, the wiring can be broken down into four blocks: 1) The connections of the servos to the servo driver, 2) The connection of the potentiometers to the analogue inputs on the Arduino board, 3) Linking the Arduino board to the driver for code execution and power, and 4) External power supply to keep the servo driver running and execute the code without drawing power from the Arduino board. Provided by Kalton Serra.

After checking that the connections are fixed in the pins and well-soldered, the PLA base with the electronic cards can be screwed to rest of the scaffold and the arm is ready to work (Figure 3.20).

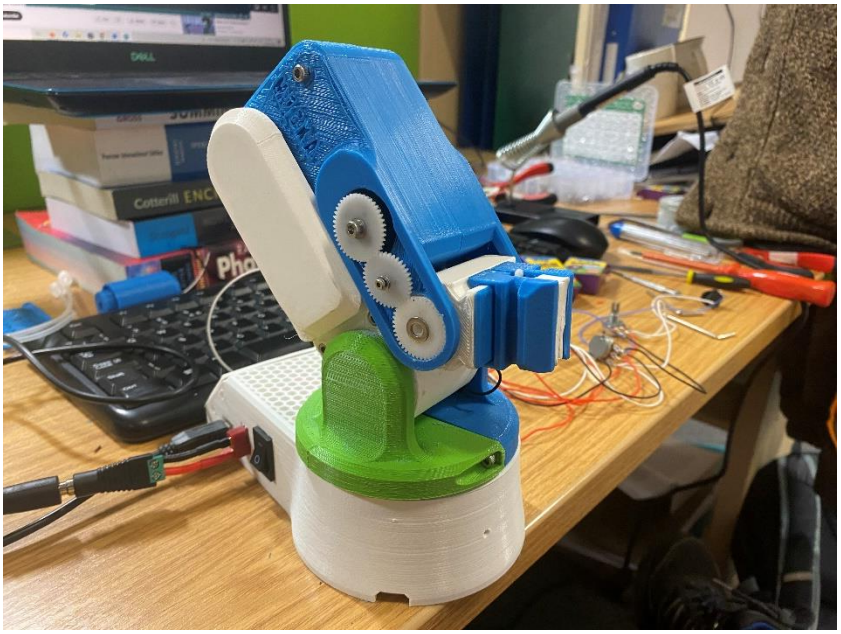


Figure 3.20. The fully assembled robotic arm. The external power plug and the switch is shown on the left side of the robot.

The robotic arm is controlled by the controller, which is a mini model of the arm with the input potentiometers. Since the potentiometers are connected to the Arduino board via the analogue input pins, we can use those readings to align the servo angles with the potentiometer's angles using the Arduino script. After that the servos' shafts mimic the movement of potentiometers' shafts, and the controller movement is translated into the arm.

The robot arm has been tested in bench, as also inside the Opentrons robot, where the controller is carefully extended outside the deck and able to be manually operated from a position closed to the Opentrons computer keyboard. The testing of the robot can be watched in the following Youtube link.

3.3.5. 3D-printed models to assist experiments.

Several 3D prints were employed over the PhD project, to facilitate manual pipetting steps, to help to set up the labware necessary for the robot protocols and for other miscellaneous applications. For most purposes, an Ultimaker 3 3D printer (the same used for the robot arm construction) was used to print the models.

The first one is a 96-well plate cover (Figure 3.21), whose holes are of diameter such that when the tip penetrates the cover, it goes down to a height just before the

pellet. This is when the centrifugation is performed to the full plate. The diameter calculation was obtained by trial and error, due to the difficulty to measure tip section with the available tools in the laboratory.

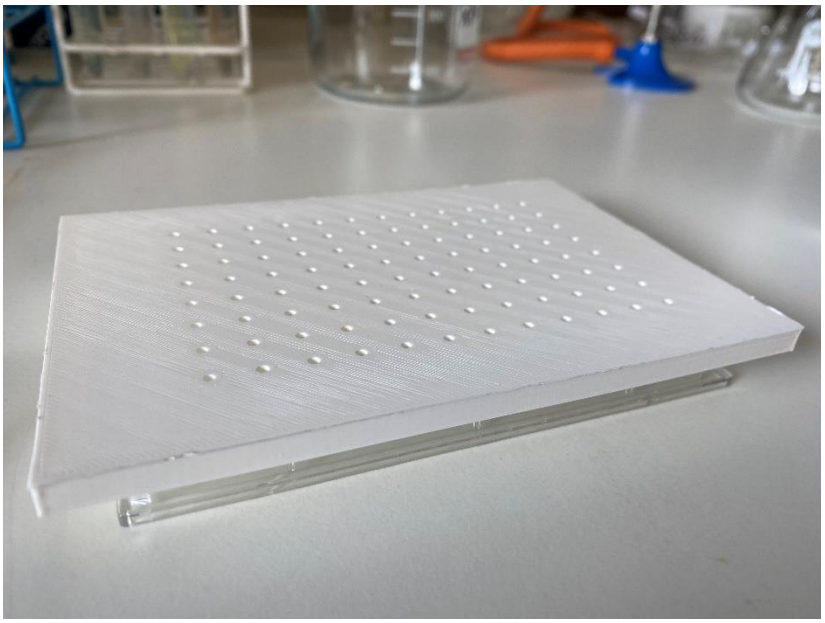


Figure 3.21. A 3d printed cover for 96-well plates with holes. The diameter of the holes was designed such that a 20 ul tip can penetrate the well to a depth just before the bottom (2mm before bottom), where the pellet is located after centrifugation. This design helps with supernatant extraction when performing manual pipetting.

The second set of prints are racks that can fit in the Opentrons deck. We printed microtube racks and racks for 5ml and 15 ml Falcon tubes. The latter are important because the tubes act as reservoir for solutions in the culture medium mix protocol (Figure 3.22). The microtube rack model is provided by Opentrons, while the Falcon tub rack is provided by Trevor Ho. The GitHub repository to the Falcon tube model is present in the Data Availability section.

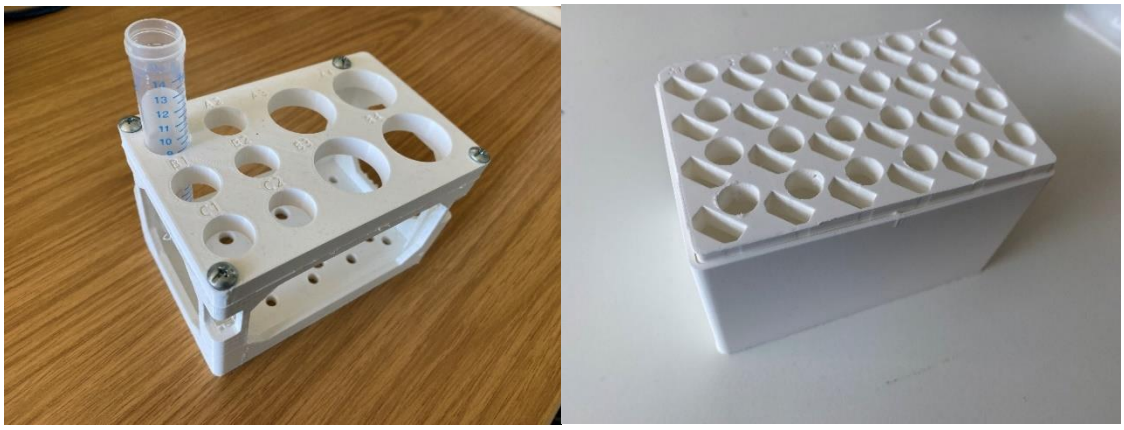


Figure 3.22. 3D printed racks used in Opentrons runs. The rack for 24 microtubes is provided by Opentrons, while the design on the left for 5ml and 15 ml Falcon tubes is an open-source design provided by Trevor Ho (github/tyhho).

Finally, the constant use of cuvettes for optical density measurement created the necessity to have cuvettes racks that can be small and easy to handle (Figure 3.23). Additional developed and built 3D-prints, which are not directly used in our protocols, are shown in Appendix A.2.

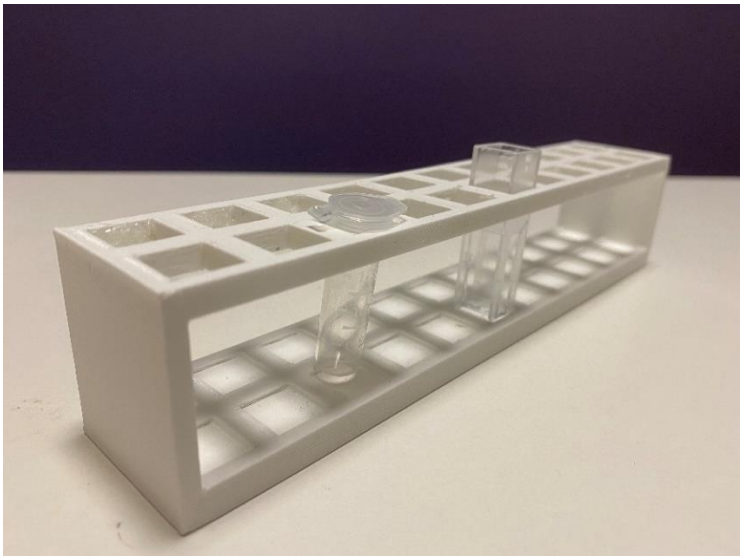


Figure 3.23. A 3D printed cuvette and microtube holder. The model was created in AutoCAD Inventor and can hold 20 cuvettes and/or microtubes.

3.4 Discussion

3.4.1. Liquid handling landscape in academia and industry

Liquid handling systems are essential tools in both academic research and industrial settings, playing a key role in automating routine and high-throughput tasks, including sample preparation, serial dilutions, compound dispensing, and plate replication. The adoption of liquid handling systems, such as pipetting robots and acoustic handling, has advanced significantly in recent years (Kong et al., 2012).

Open-source platforms often provide accessibility and customisation opportunities, but frequently lack the level of integration and throughput seen in more industrial-grade systems. Liquid handling robots can be found in specialist journals, such as HardwareX, as well as in general journals. But often, these models are not found in the academic literature but rather as standalone projects on the web or as part of other initiatives. Examples of open projects for liquid handling include OTTO (Florian et al., 2020), with reported pipetting accuracy of 2.5%, PHIL (Dettinger et al., 2022), a personal robotic pipettor with accuracy around 1-2%, and even transforming a 3D printer into a liquid handling robot by attaching hand pipettes (Kopyl et al., 2024), among others.

In contrast, the industrial sector tends to rely on high-end, proprietary systems that offer greater precision, reliability, and throughput. Liquid handling in the biopharmaceutical and biotechnology industries typically demands integration with other instruments, such as plate readers, centrifuges, and mass spectrometers. As a result, industrial liquid handling platforms, such as those from Beckman Coulter or Tecan, offer high levels of automation and integration capabilities. While these systems are robust and suitable for large-scale applications, they come with a high price tag, both in terms of initial investment and spare parts, and are often less customisable than open-source alternatives. Interestingly, acoustic dispensers, which allow for liquid manipulation in volumes of less than 1 µL in microplates, are becoming more common in industrial laboratories and foundries. These are specially designed to interface with other machines and can even interface with mass spectrometers, with proper modifications (Van Puyvelde et al., 2024).

However, there are efforts to bridge the gap between open and closed platforms. One example is the PyLabRobot package for Python (Wierenga et al., 2023), which enables control of multiple liquid-handling robots, including Opentrons, Tecan, and Hamilton models, using a single set of base functions (though specific configurations differ depending on the brand), as well as other laboratory equipment, such as ClarioStar plate readers, scales, heaters, and more.

3.4.2. Gaps found in our current pipeline and proposed solutions.

Main gaps exist in the interaction between the Opentrons OT-2 and the plate reader, the centrifuge and the mass spectrometer. The first issue, integration with the plate reader, has been described in the literature. One of the lateral windows in the OT-2 can be dismounted and a medium-sized plate reader can be fit, such that the plate tray aligns with one of the labware spaces in the deck. In case that the plate reader does not align by design, because for example the tray is pointing 90° degree respect the deck main orientation, a custom labware protocol can be implemented and added to library, such that the robot can correctly position the tip over the wells in this rotated plate (Martin et al., 2022). For the experiment, we did not have a dedicated plate reader, therefore this option was discarded in principle, but in the future arrangements could be to acquire a plate reader suitable for integration purposes.

For the second issue, centrifuge integration, two practical solutions have been discussed. Specially in biotech industries, there are specialised centrifuges with lateral plate entry that can be streamlined with other robotic operations (Hettich, 2024). But it requires a purposed investment in automation, and it might reduce flexibility in combined automated-manual protocols. Therefore, many laboratories, even with enough funding, has centrifuges that only can be manipulated manually. Thus, the second option is to create a specific robotic solution to operate the centrifuge in an analogous way a human would do it, but this would require a more sophisticated robotic state estimation system (Medra, 2024).

3.4.3. Use of databases in workflows

Apart from the division of the Bayesian optimisation protocol into two pipelines that can be understood as directed acyclic graphs (in contrast to the loop structure of

the original protocol), the flow resembles other pipelines found in bioinformatics and automated laboratories (Casas et al., 2024). At many stages, a series of files are queried as input, mainly to determine labels, such as component labels or position-related labels for Opentrons operations. These files also contain some redundant information. For instance, the M9 recipe file already includes the component labels, so a separate file containing just the labels as input for the first script in the initial sampling pipeline is unnecessary and could be considered cluttered from a data manipulation perspective. Therefore, an upgrade of the pipeline should incorporate the use of databases, such as SQL, to streamline the extraction of these labels and create a cleaner workflow from the user's point of view (Reiter et al., 2021).

The database could include tables related to experimental information, such as the number of samples per batch, as well as tables for stock consolidation, chemical information on the components, and culture medium recipes (with interaction with the DSMZ MediaDive database (Koblitz et al., 2022). This would make the pipeline more flexible for experimenting with other base culture media in future trials with our *Bacillus* and would also facilitate the transition to bioprocess optimisation with other organisms.

Database integration has proven effective in biological experimentation (Reiter et al., 2021), handling both tabular and unstructured or graph data, such as those commonly found in systems biology. One important thing to consider when creating database for experimentation are standards. In my current pipeline, labels are being used, but these often do not adhere to a specific standard, as the development was considered local. However, by adopting standards and ontologies, databases ensure that biological data can be shared and reused efficiently across different research projects (Lapatas et al., 2015), so proper standardisation will be considered for the future.

3.4.4. Open labware

Open labware refers to laboratory tools, equipment, and devices that are designed using open-source principles, allowing users to access, modify, and share designs and protocols freely (Baden et al., 2015). This concept is relevant to different types of labware found in biology, including microscopy, centrifugation, containers,

and span different levels of complexity or intended use, covering also the repurposing of old labware and open wet-lab resources, such as enzymes for biological engineering (Wenzel, 2023).

The advantages of open labware include lower costs compared to commercial solutions, flexibility, and the ability to integrate custom solutions with existing laboratory workflows. Normally, the designs are available in the web, and have to be built locally, which also increase the number of options available in terms of material and manufacturing accuracy.

Robotic arms are abundant in the open-source community, where designs often originate from the ingenuity of individuals or teams. They come in various sizes and can be generalist (able to perform different tasks) or specialist (designed for a specific task, e.g. pipetting). It is difficult to benchmark open-source robots for manipulation because the designs can be quite dissimilar, using different torques for servos. This is why several criteria were used to select the appropriate robotic arm model to replicate. Cultivating these types of experiences in the laboratory may help increase biologists' interest in automation (Holland & Davies, 2020).

However, open labware also presents challenges. While the flexibility and cost savings are significant advantages, there is often a lack of standardisation, which can lead to compatibility issues when integrating open labware with other systems or instruments, such as plate readers, centrifuges, or mass spectrometers (Wenzel, 2023). Additionally, the responsibility of troubleshooting and optimising these open-source systems falls on the user, which may not be ideal for all laboratories, particularly those in industry where reliability and throughput are critical. Developing engineering/electrical knowledge in a biology laboratory can take an arbitrary time and sometimes, an engineer has to be hired only for that purpose.

This democratisation of biological technology is levelling the playing field, allowing smaller labs or those in less developed regions to engage in cutting-edge research (Oellermann et al., 2022). In the UK, there are several examples of democratisation of biological tools, such as the OpenPlant initiative, which are being crucial to increase penetration of these tools in the community.

3.5 Conclusion

Automation in high-dimensional biological experiments is essential to perform multi-factorial experiments in reproducible and time-effective way. This can also contribute to reduce the experimenter load and organise the data to make accessible and findable. We expect that this pipeline development can help other biologists and engineers interested in developing automated workflows and show the power of custom approaches for reliable biological measurements.

3.6 Acknowledgments

Access to the Opentrons OT-2 liquid handling robot was kindly provided by the Edward Wallace Lab at the School of Biological Sciences, University of Edinburgh. I thank Laura for introducing me to the basics of the Opentrons OT-2 operation.

3.7. Data Availability

The scripts are contained in several repositories. However, the main scripts for the developed pipeline are compiled in the GitHub repository www.github.com/rvalenciaaz/pipeline_scripts. Interactive links for information related to the robotic arm can be also found in https://rvalenciaaz.github.io/portfolio/. The STL models of 3D printed tools is available in the repository https://github.com/rvalenciaaz/3D_models_bio_automation. Youtube link to the robot demo is here: https://www.youtube.com/watch?v=U1p1gtquFiM&ab_channel=RicardoValencia. The 3D model generated by Trevor Ho is available at https://github.com/tyhho/OT-2_3D_Designs/

Chapter 4

Single and multi-objective optimisation of titres from a *Bacillus* space-filling experiment

Work by Ricardo Valencia Albornoz[1], Jessica O'Loughlin[1], Diego Oyarzún[1,2] & Karl Burgess[1]

[1] Institute of Quantitative Biology, Biochemistry & Biotechnology, School of Biological Sciences, University of Edinburgh, King's Buildings, Edinburgh, United Kingdom

[2] School of Informatics, University of Edinburgh, Edinburgh, EH8 9AB, United Kingdom

Ricardo Valencia Albornoz (thesis author) performed the conceptualisation and execution of the experiment, the analytical chemistry measurements and the analysis and interpretation of the results. The experimental protocol was supported by JO (sample handling) and DO and KB provided guidance for the main experiment.

Outline

In biomanufacturing, controlling several factors is crucial to optimising production outputs. However, this also means an increased number of samples and subsequent costs to be able to map this multidimensional space to final titres. Therefore, we need to look for alternative designs of experiments that can efficiently sample the design space and adapt to laboratory constraints, such as maintenance periods. Thus, building from the optimisation framework presented in Chapter 2 and the robotic platform developed in Chapter 3, we designed a media optimisation experiment targeting seven key components of a typical M9 medium and employed a space-filling approach to ensure comprehensive coverage of the factors' domain. We used our mass spectrometry protocol to measure metabolites in the supernatant after the samples were collected and generate surrogate models of production outputs, including lipopeptide production and biomass as OD. The resulting model serves as a reproducible benchmark for subsequent single-objective and multi-objective lipopeptide production optimisation using Bayesian optimisation. From the single-objective optimisation, results showed that the optimal number of initial samples and

batch size can be slightly modified to achieve the optimum of Surfactin C in a smaller number of iterations, but the complexity of the objective and the observed variance of the measurements means that the iterations cannot be reduced further, reaching 90% of maximum titre in 10 iterations for the case of 7 initial samples and 7 combinations per batch. In the case of multi-objective optimisation, Bayesian optimisation frameworks were able to identify the Pareto Front between lipopeptide production and biomass in the 7-dimensional factor space, considering batch sizes similar to those obtained from single-objective loops. Thus, this approach, using iterative cycles of high-dimensional optimisation of complex natural production, promises to improve the scalability and reproducibility of process conditions and analyse data and information constraints arising from the biology and analytical chemistry perspectives that can hinder efficient adaptive optimisation.

4.1. Introduction

4.1.1 Space-filling designs in biological or metabolomics experiments

As mentioned in the section 1.5.1 in the introduction about initial sampling, space-filling designs are experimental designs that aims to comprehensively cover the domain or range of the factors involved. Space-filling designs can be used on bioprocessing experiments, particularly when exploring a big and heterogenous numbers of factors, such as culture medium composition, temperature and agitation. However, reports on the use of these designs related to biology or metabolomics experiments in the academic literature is scarce. A cause for this may be that high-dimensional bioproduction screening required for this design often needs automation and robotics to make it feasible, and most of the laboratories in academic environments do not have access to this infrastructure or the automation pipelines are used for other purposes (Kampers et al., 2022). Chilakwad in 2022 describes this situation, while working in Synthace, a company based in London that supply external options for large-scale biological experimentation, and Gilman in 2021 describes the use of space-fillings designs as a tool for “future” synthetic biology large-scale experimental design, mentioning only one example (Govindarajan et al., 2015).

4.1.2 Trade-offs between variables, Pareto fronts and bacterial metabolism

Multi-objective optimisation aims to find inputs to maximise or minimising several conflicting objectives (Collette & Siarry, 2004). For example, as shown in Chapter 1, changing the medium composition to increase the titre of a metabolic product (Surfactin D) might reduce the growth rate of the bacteria. Pareto fronts are used in these cases to identify the set of optimal solutions where no objective can be further improved without worsening other objective, i.e, there is a trade-off (Collette & Siarry, 2004). In practice, it can be visualised better in a scatter plot, like the one shown in Figure 4.1. Thus, when minimising two objectives, solely trying to minimise the second objective means adding value to the first objective, which is not desirable. The solutions that are not Pareto-efficient, i.e, there are still resources or inputs that can be distributed to decrease one of the objectives are said to be dominated by the Pareto-efficient points. This optimal set of points, defining the Pareto front, are called non-dominated solutions (Collette & Siarry, 2004).

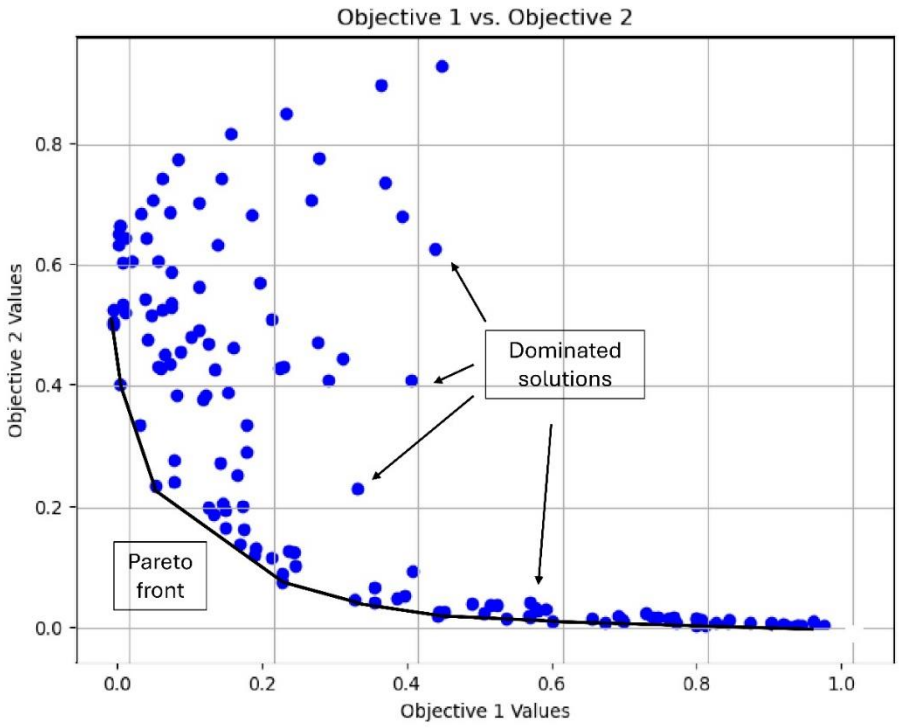


Figure 4.1. The concept of a Pareto front. A system with inputs and 2 optimisation objectives that are conflicting and needs to be minimised, because they use the same resources for example, has a series of compatible solutions where both objectives can reach a specified value. However, there are set of points that efficiently use every

resource available for the objectives and moving along these solutions are equally preferred if no constraints are imposed (Pareto efficiency). These solutions demarcate the Pareto front and are said to dominate the other compatible solutions, and therefore called dominated solutions.

Important for multi-objective optimisation purposes is the concept of hypervolume. The hypervolume can be broadly defined as the size of the dominated space and in practice it is calculated as a set of n-dimensional volumes delimited by the non-dominated solutions and a reference point (Guerreiro et al., 2021). In Figure 4.1, one can define a reference for calculation in the point (1,1), and calculate the area between the Pareto front, the reference and the lines x=1 and y=1. When performing multi-objective optimisation, the hypervolume serves an indicator of the quality of the non-dominated set, in the sense that this set is covering mostly all the optimal front (Guerreiro et al., 2021).

The concept of Pareto front and trade-offs has been proven important in cell physiology and metabolism. A clear example occurs in the allocation of ribosomal resources to either growth or division-related systems (Serbanescu et al., 2020). Another classic occurrence is in metabolic engineering, where trade-offs between production rates and cell burden must be carefully managed, and flows needs to be constrained to allow normal growth (Wortel et al., 2018; Han & Zhang, 2020; Banerjee & Mukhopadhyay, 2023). Trade-offs also happen in ecological interactions of microbes (Litchman et al., 2015; Manhart & Shakhnovich, 2018; Thingstad, 2022) and strategies for long-term survival in bacteria (Abram et al., 2021; Bruggeman et al., 2023; Behringer et al., 2024). The book on Economic Principles in Cell Biology (The Economic Cell Collective, 2024) covers a good portion of these examples.

4.1.3. Aims and objectives

The aim of this chapter is to:

- Demonstrate the applicability of the developed platform for bioprocess optimisation via optimisation of several components in culture medium.

The aim encompasses the following objectives:

- Apply the developed pipeline in Chapter 3 to prepare mixes of testing media with 7 components using the OT-2 robot

- Quantify surfactin, other lipopeptides, and biomass coming from a space-filling experiment over the medium components

- Use these measurements to fit a probabilistic model the titre of surfactin, as well as the titre of other lipopeptudes, and OD as biomass indicator

- Employ the mentioned models to test *in silico* the Bayesian optimisation pipeline for single-objective optimisation of surfactin titre and multi-objective optimisation of surfactin titre and maximum OD

4.2. Materials and Methods

4.2.1. Strains and media employed.

Bacillus subtilis DSM 3256, a surfactin-producing *Bacillus* strain, was employed for the experiment. Similar as in chapter 2, the strain genus was verified using 16S rRNA sequencing, to avoid any contamination in the stock that might affect the results.

M9 medium was prepared using the following recipe: Glucose: 0.4% (carbon source), NH_4Cl: 18.7 mM (nitrogen source), NaCl: 8.5 mM, $MgSO_4$: 2 mM, $CaCl_2$: 0.1mM, Na_2HPO_4: 42.2 mM, KH_2PO_4: 22 mM. Precaution was taken when sterilising the glucose stock, while remaining components were autoclaved at standard 121° C for 15 min.

For the optimization experiments, every component concentration was modified from this recipe, in contrast to Chapter 2. Components of the M9 medium were purchased from Sigma Aldrich. The microplate experiments were performed using 48-well flat bottom microplates (Corning), which allows to perform experiments with bigger volume. The working volume for every experiment was defined in 800 ul, leaving headspace for correct aeration of the cultures.

4.2.2. Creation of tick box list for the experimental and computational pipeline

To streamline the experimental and computational aspects of our study, we developed a detailed tick box list. This list serves as a comprehensive checklist that

ensures all necessary steps and components are accounted for before proceeding with each phase of the pipeline. The tick box list covers initial setup, data collection protocols, computational analysis steps, and final validation processes. By methodically checking off each item, we maintain consistency and accuracy throughout the experimental workflow, reducing the risk of oversight and enhancing the reproducibility of our results. The link to the tick box list document is described in the Data Availability section. The experimental iteration (for Bayesian optimisation cycles), together with basic experimental information, such as day and time, can be filled in the top of the document.

4.2.3. Generation of component mixes using Opentrons OT-2 robot

To obtain the desired concentration of each M9 component in the wells, we proceed in four steps: First, a Sobol sampling design containing 6 batches multiplied by 7 samples per batch/microplate is created using a custom Python script and the SobolEngine function from the BoTorch library. Second, a custom Python script translates the coordinates of the sampled points, which are constrained to the (0,1) 7-dimensional box into concentrations that need to be achieved in each well. For this, we need the ranges of concentrations to test. All components can have a minimum value of 0% or 0 mM in the combination, while the maximum is given by the following list (the M9 default concentrations are shown again in parenthesis for comparison): Glucose: 4% (0.4%), NH_4Cl: 240 mM (18.7 mM), $MgSO_4$: 2 mM (2mM), KH_2HPO_4: 50 mM (22.04 mM), Na_2HPO_4: 50 mM (42.26 mM), $CaCl_2$: 0.2mM (0.1 mM) and NaCl: 50 mM (8.55 mM).

Thus, combinations corresponding to modifications of the M9 medium recipe, as well as controls, were block-randomised across a 48-well microplate, considering 4 replicates (blocks). These blocks are physically distinct groups of 9 wells, comprising 7 conditions, 1 M9 medium well with added bacteria and 1 M9 medium well with no bacteria, acting as control. The layout for block randomisation was obtained from a custom Python script. Third, 2ml pre-stocks of the different components' stocks were prepared in microtubes by mixing the component with water in appropriated volumes, such that the concentration in the pre-stock is 10 times the required in the wells and labelled depending on which well will be used. Finally, for

mixing the component in the wells, a fixed volume of 80 ul was added to each well from each corresponding pre-stock for each M9 component and water was added to reach the final volume of 800 ul. Every liquid transfer was performed using an Opentrons OT-2 robotic system. The robot was programmed to prepare themes according to the formulations derived from the microplate layout.

4.2.4. Microplate cultures and sample preparation

Bacillus subtilis DSM 3256 frozen stock was revived by culturing on an LB agar plate at 37°C overnight. From this plate, individual colonies were inoculated into four precultures consisting of 5 ml of M9 medium. The cells were then grown in a shaking incubator at 37°C and 180 rpm for 12 hrs. For the added culture, the starter cultures were adjusted such to achieve an initial optical density at 600nm (OD600) of 0.01 in each well. The OD was measured directly in the Tecan Infinite M200 PRO plate reader, to keep consistency with later measurements.

The proper microplate culture was conducted using a Tecan Infinite M200 PRO plate reader, recording an OD600 measurement every 10 minutes for 24 hours at 37°C. The agitation was set to a maximum of 6 mm. Growth data for each well were collected in a excel file for blank subtraction and post-processing.

After 24 hours, the cultures are stopped and 500 ul from each sample on the 48-well microplate were transferred into microtubes and quenched with 500ul of cold methanol, available in a 96-well deep bottom plate (Greiner) and kept cold in dry ice (~−78.5 °C). After the quenching, the samples were quickly transferred into a centrifuge and spun to 3220 x g (10000 rpm) for 30 min to separate the cell pellet from the supernatant. 600 µl of supernatant was carefully transferred from each microtube to ice-cooled microtubes and later aliquoted into Axygen 96-well PCR plates for measuring on the mass spectrometer. The plates were sealed with a cap and stored in a -80°C freezer until the quantitation. In addition, 20 ul from each supernatant sample were mixed to generate a quality control (QC) pool for intra- and inter-batch abundance correction.

4.2.5. Surfactin and additional metabolites quantification

We utilised flow injection of spent media with a ThermoFisher Dionex Ultimate 3000 autosampler. The sample volume was set at 1 µl. The mobile phase comprised a 1:1 ratio of acetonitrile to water with 0.1% formic acid, and the flow rate was 200 µl/min. These conditions were optimised experimentally and are discussed in Chapter 3. The sample acquisition time spanned 1 minute, and to quantify specified molecules, selected reaction monitoring (SRM) on a ThermoFisher TSQ Quantiva triple quadrupole (QqQ) mass spectrometer (MS) was implemented. The MS parameters, as well as the precursor/product masses, collision energy and dwell time table, can be found in Supplementary Tables BT1 and BT2, respectively. The precursor/product masses for surfactin variants were obtained and confirmed using Pubchem open mass spectrometry data (Pubchem, 2024).

Peak extraction was accomplished using the rawrr package in R (Kockmann and Panse, 2021). In the script, each of the scans in a raw file corresponding to a specific precursor/product mass for each retention. If there are more than one product associated to a given precursor, the intensity is sum across the products. The number of acquired points along the 1 min run is controlled by the dwell time. Higher values of dwell time (in ms) mean a greater number of acquired points over the time window. The custom script is available on the link from the Data Availability section. The script outputs a file calls an “hypertable”. This hypertable consists of three columns with retention time (s), scan information, intensity and raw file name from which this measurement comes from.

Subsequently, metabolites underwent baseline correction using the asymmetric least squares algorithm from the Python pybaselines package (Erb, 2022) set to default parameters. The corrected peaks were integrated over the 1-minute run using the trapezoid rule function from the Numpy package via a custom Python script. The integrated intensity data was then compiled into a table.

Intra and inter batch correction was performed using the Quality Control Robust Spline Correction (QCRSC) algorithm implemented in the pmp package in R (Jankevics et al., 2024). This algorithm has been used before for batch correction in flow injection mass spectrometry runs (Kirwan et al, 2013) and employ cubic splines to fit the trend

in the QC samples associated with carryover (intra-batch) and heterogeneous count levels (inter-batch). After the correction, samples were probabilistic quantile normalised (Dieterle et al., 2006) using the QC samples' abundance over the batches and the pmp package. These normalised values are further referred as normalised Surfactin C titres or abundances.

4.2.6. Evaluation of the surrogate model fitted on the space-filling design.

To perform Bayesian optimisation on the output dataset from the space-filling design, we use surrogate models that act as a substitute of the "true" titres of metabolites and biomass as OD across the factors' domain. In general, a surrogate model is a mathematical model that approximates the outcome of a function of interest. First, we train a Heteroskedastic Gaussian Process regression (GPR) model with default parameters and priors for the lengthscales (Goldberg et al., 1997). The lenghtscales are associated with the granularity of the approximation of the Gaussian Process regression depending on which input it is being considered. Lower lenghtscales means more granularity in the prediction associated with an input and therefore, these variables (Rasmussen & Williams, 2006). The model was fit using the functions provided in the BoTorch package in Python. For the cross-validation of the model, the 42 combinations with its respected measured mean abundance and its variance as square error of the mean where split into 7 folds, leaving 6 samples in a test set, while the rest is used for training the model. After the training on each of these splits and quantifying the fit in the test split, a mean R^2 score and a mean square root of the mean metric (RMSE) can be obtained across the splits (Gelman et al., 2020).

In addition to Gaussian process regression, we also test an ensemble model with Bayesian model averaging (Hoeting et al., 1999). For this model several base learners are first trained on the dataset. Then, the predictions of the ensemble model are generated by merging the predictions from each learning using a probabilistic linear form. The outputs of each model are accompanied by a weight, which as distribution where the mean and variance are determined by the errors and variance found in the original model where the prediction comes from. We tested this approach using the following base learners from the scikit-learn package: Linear Regression, Decision Tree Regressor, Random Forest Regressor, Gradient Boosting Regressor, K-

Neighbours Regressor, Gaussian Process Regressor (non-heteroskedastic), Support Vector Regression, Kernel Ridge Regression and Multi-Layer Perceptron Regressor. Details on the models can be found in reference books, such as Murphy, 2012. The cross-validation was performed in the same way than in the testing of the Gaussian process regression model, using seven splits.

4.2.7. Simulation of active learning loops

From the minimum and maximum concentrations that were selected for the microplate layout concentrations, a 7D design space was defined by scaling into the box $[0,1]^7$. Initial conditions for both single and multi-objective optimisation of the lipopeptide production surrogate model and the biomass model were obtained from a Sobol sampling design implemented in Botorch. Several initial samples sizes and batch sizes were tested. The models and the acquisition function for the Bayesian optimisation loop were implemented using Botorch library in Python (Balandat et al. 2019), and default internal parameters were used. For the single-objective optimisation, a Heteroskedastic Gaussian Process regression (GPR) model was trained in the initial samples, then the q-Noisy Expected Improvement acquisition function was calculated and maximised to obtain candidates for the next cycle. In the case of multi-objective optimisation, the chosen model for the iterations is a multi-output Heteroskedastic Gaussian Process regression (GPR) model and the acquisition function correspond to q-Expected Hyper Volume Improvement. This function provides sample recommendations that balance between exploration and exploitation that aims to increase the volume delimited by non-dominated solutions, i.e the hypervolume, such that the Pareto front can be identified by expanding this set of solutions as much as possible. Several additional scripts used in intermediate steps for formatting data tables and are described in Chapter 3. Surface plots, principal component analysis (PCA) and radar charts to explore and analyse the data were implemented using the matplotlib, seaborn and plotly packages in Python. PCA biplot was generated using the pca package in Python (Taskesen, 2020).

4.3. Results

4.3.1. Space-filling design allows off-line lipopeptide quantification on a multi-factor medium composition experiment.

After optimising surfactin production by changing two factors, it is important to scale the approach to include multiple components (more than two) in the medium. This situation is commonly observed in bioindustry, where any improvement in final titres is paramount for the success of company operations.

For technical reasons, such as equipment maintenance periods, it is essential to employ an efficient sampling technique that minimises the number of experimental runs while maximising the information gained. In addition, this technique should also allow samples to be measured offline, in contrast to the iterative method presented in the previous chapter. This dataset will be important for generating a surrogate model of production outputs, including lipopeptide production and biomass, which will be used to test Bayesian optimisation pipelines in multi-factor culture medium components experiments. Estimating how many iterations will be needed to reach an optimal value of Surfactin C or biomass in the 7-factor space and therefore the experimental effort (costs and time) involved will be important for future bioprocess testing experiments for this strain or similar strains.

Thus, we design an experiment to modify all M9 culture medium components concentrations (glucose, ammonium chloride, magnesium sulphate, potassium phosphate, sodium phosphate, calcium chloride, and sodium chloride) for the estimation of lipopeptides titres and biomass in the 7-factor space. In this case, we will use Sobol sequences as a space-filling design to draw informative samples. Sobol samples, and quasi-random sequences in general, can cover the design space homogenously, maximising information about the lipopeptide production landscape for our *Bacillus subtilis* system. The limits of the design space are given in Methods. These ranges were chosen based on the ranges from the previous 2D experiments combined with literature on maximum concentrations allowing normal growth.

As observed in Figure 1.11, the distance between points in Sobol sampling is similar across the design space, avoiding gaps where there is little information for

further modelling and the uncertainty prediction of the outputs in these zones would be high. These gaps are common in random sampling from a normal distribution, where points agglomerations leave some combination subsets untested.

4.3.2. Extraction and correction of the mass spectrometry data from the space-filling experiment.

Several protocol optimisations were implemented in the space-filling experiment compared to the two-factor active learning loop. We moved from a 96-well plate format to a 48-well plate format, allowing for a higher culture volume and therefore easier liquid manipulation, including minimal perturbation of the pellet at the supernatant transfer step. However, the reduction in number of wells meant that some replicates were sacrificed, and therefore four biological replicates are used per combination.

The experiment consists of 6 microplates batches of 7 combinations each. Instead of creating the batches separately, a single space-filling design with Sobol sampling was created with 42 samples in the design space. Then, the 42 samples were split in 6 sets randomly, giving what combinations should go in each well. As with the iterative 2D experiment, a M9 reference and control with no bacteria were added to the list. For the layout of the plate, we performed block randomisation, this time with 4 blocks/replicates and the 7 combinations plus 2 reference and control samples were allocated in each block. Thus, in total we retrieved 206 experimental samples.

After the microplate planning, the mixing step was performed with a pipetting robot, the Opentrons OT-2, which allows automatic mixing of the components in the required concentrations for each well. A mixing run with the 7 components on different concentrations can be done in approximately 2 hours in comparison to manual pipetting that can take 6 hours or more and can be prone to manipulation error. After the quenching with cold methanol and supernatant transfer, 20 ul of the samples were mixed to generate a QC pools, and together with the rest of the samples were kept in -80 °C until mass spectrometry quantification. The raw peaks for Surfactin C extracted from the mass spectrometry files are depicted in Figure 4.2.

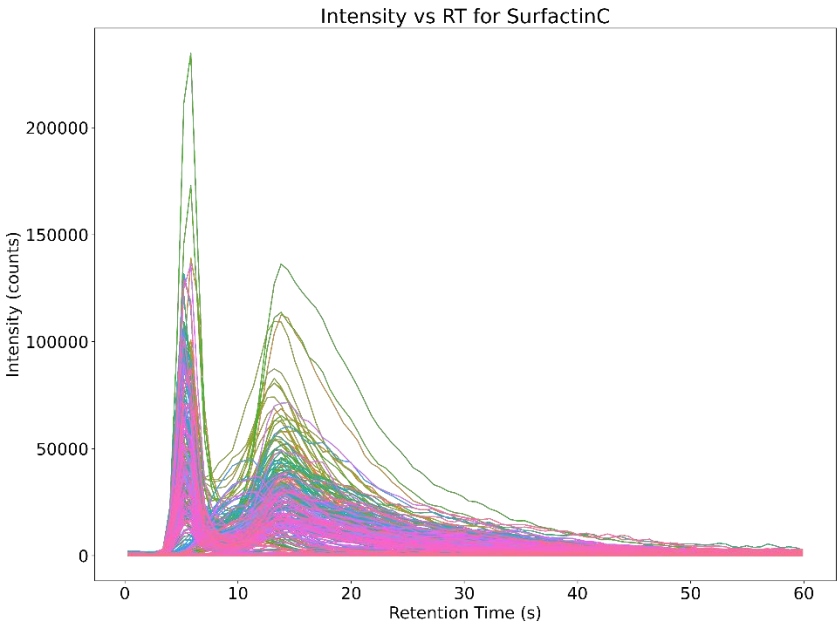


Figure 4.2. Extracted ion chromatograms for surfactin C across the space-filling samples, including 42 samples from different medium culture combinations, QC samples, and Reference and Control samples**.** Curves are shown in colours from the one with least area under the curve (in pink) vs those with highest area under the curve (green).

The extracted ion chromatograms over the 1 min mass spectrometry (MS) run for each of the samples show a characteristic double peak curve, consistent with ion suppression effects (Annesley, 2003; Volmer & Jessome, 2006; Furey et al., 2013). The area under the curve progresses softly from the smallest peaks to the biggest ones, as expected from a dense sampling of an experimental design space.

Later, the areas under the curve for each peak were corrected using quality control (QC) pool samples in the run. There are 3 QC samples at the start of the MS run, then 1 every 6 experimental samples, and three at the end corresponding to 11 QC samples per batch. These adds to a grand total of 206 + 11 x 6 = 272 samples that were quantified using the triple quadrapole mass spectrometer. Since the QC samples are extracted from the same pool, they should have the same abundance in each batch. This QC abundance is used to correct the samples, since carryover and shifts can affect the measured abundance as function of the sample order and the batch order (Figure 4.3).

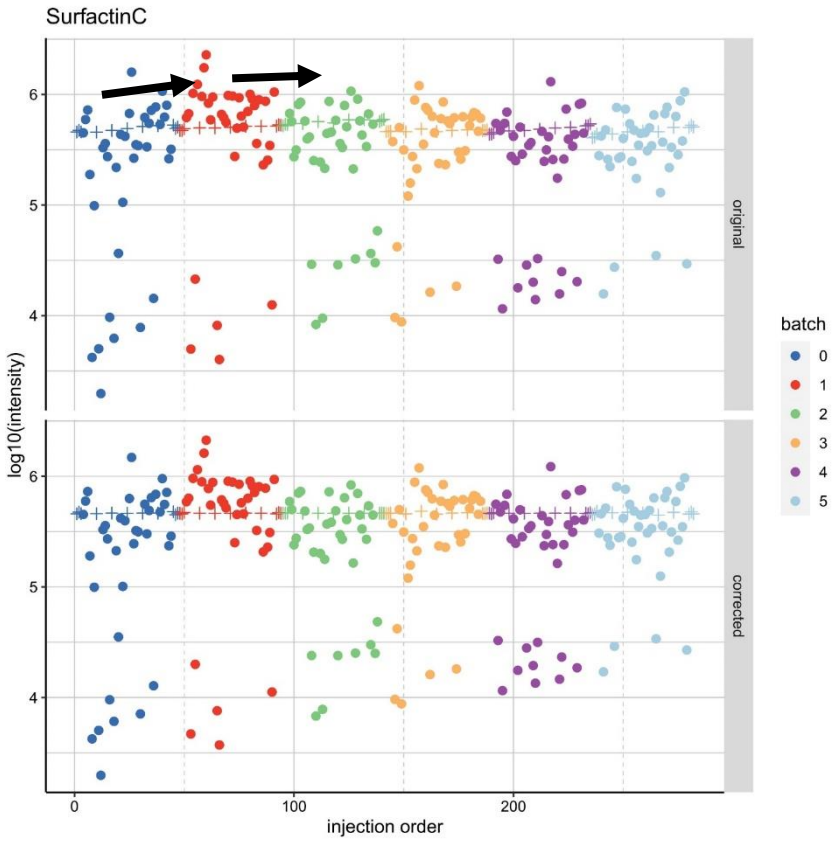


Figure 4.3. Intra and inter batch correction of area under the curve quantification of extracted peaks using the pmp package in R. On the top, the data is shown before correction. Black arrows depict the overall trend of quality control (QC) over the first and second batch. In contrast, at the bottom, the data has been corrected using the Quality Control Robust Spline correction (QCRSC), which basically fits a cubic spline curve over the QC samples at the intra- and inter-batch level and calculate factors to adjust the experimental samples.

In the top of Figure 4.3, it can be observed that carryover means that QC samples progressively increase in abundance over the batch, to rate of ~0.1 normalised titre/sample in the first batch and similar rates on the other batches. There are some shifts in the starting QC samples between the batches, of the order of also ~0.1 normalised titre between the first and second batch and similar to other contiguous pairs of batches. For the correction, we employed the QCRSC algorithm implemented in the R pmp package (see Methods), that fits a cubic spline to the QC samples and use this curve to calculate adjusting ratios for the other samples based on sample order and batch order. Thus, after the correction and normalisation, the samples are ready to be analysed in detail.

In the case of Surfactin C abundances, several patterns are striking. First, the M9 reference samples are on the top of the list (Figure 4.4), indicating that the specific M9 combination of components is indeed a good surfactin producer and is correlates with the optimal results found in the 2D experiment. Then other combinations can be highlighted as high producers, including 1_1 and 1_4. These are given by the following composition respectively, for 1_1: Glucose: 0.7%, NH_4Cl: 103 mM, $MgSO_4$: 1.52 mM, KH_2HPO_4: 46.42 mM, Na_2HPO_4: 33.98 mM, $CaCl_2$: 0.18 mM, NaCl: 46.57 mM and for 1_4: Glucose: 0.27%, NH_4Cl: 4.88 mM, $MgSO_4$: 0.63 mM, KH_2HPO_4: 27.5 mM, Na_2HPO_4: 22.7 mM, $CaCl_2$: 0.11 mM and NaCl: 26.9 mM. The former is a low glucose, medium nitrogen combination with respect to the concentration ranges for these compounds, while the latter is low glucose and nitrogen combination with similar composition to the M9 medium.

The controls are consistently close to 0, as expected from since no bacteria was added to the medium and therefore no surfactin should be present, except in the last batch (5_CTRL) where one of the measurements is a bit higher, due to a contaminated control well on that microplate. This can be confirmed by examining the growth curves on the Supplementary Figure B.9.

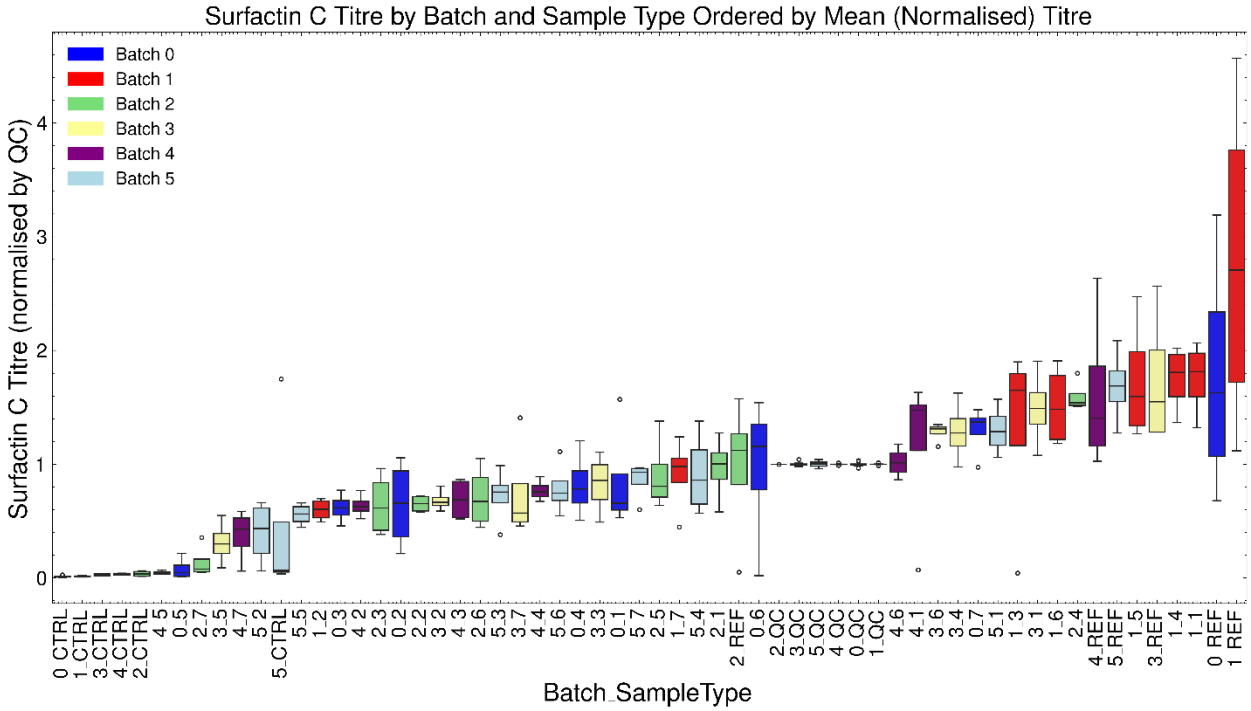


Figure 4.4. Surfactant Concentration across different batches and combinations from the space-filling design. The box plot displays the distribution of (normalised) surfactant

concentrations measured on the six different batches. Each box represents the interquartile range (IQR) of the concentration values, with the median indicated by a line within the box. Whiskers extend to the furthest point within 1.5 times the IQR from the upper and lower quartiles. Outliers are represented as individual points beyond the whiskers. Data show a general increase in the concentration values towards combinations close to the M9 reference, with Batch 1 exhibiting samples with high surfactin production. The analysis includes data from multiple measurement points, 4 replicates per combination and 9 sample types (7 combinations plus 2 control and reference) across each batch. QC pools correspond to normalised titre 1 in the plot.

Since visualising high dimensional datasets and functinos is a complex art, a simple way to look at dependencies on Surfactin C titres related to component concentratinos is to get contour plots using pair of variables/inputs/factors in the x- and y-axis. Since we do not want to establish a model at this point, but rather evaluate non-linearities in the data and other features, the countour plots are generated by simply fitting a spline curve to the observed values. In Figure 4.5, it is possible to pinpoint the Glucose vs NH_4Cl surface, with striking resemblance to the surface obtained in the 2D iterative experiment. For other component pairs, the relationship is unclear since it is possible to find several promising combination in the space. Since we are looking to slices, these might be compositions in which another variable is important. A clear example is the blue spot in the ammonium chloride vs sodium chloride, where that specific combination is a very low carbon combination. However, it is not clear from the slices in which scenario this low spot happens and if sodium chloride is relevant or not for the yield.

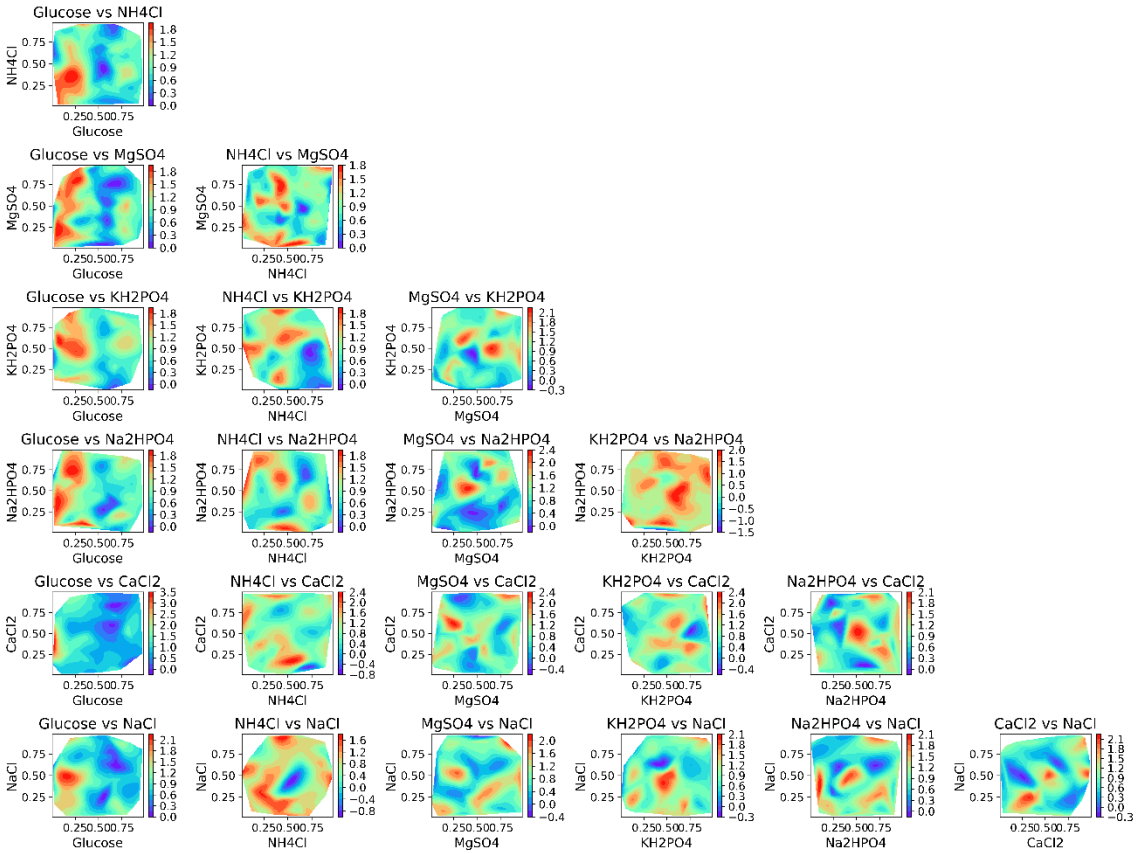


Figure 4.5. Contour plot of Surfactin C titres against pairs of culture medium components. These plots serve a quick way to analyse dependencies of Surfactin C titres with the underlying factors in the experiments. Red colours indicate higher normalised titres while blue colours mean lower normalised titres. The surfaces are not provided by Gaussian process models, but instead by cubic spline interpolation of the empirical data. This provides a visualisation of the data before model choosing.

Turning the attention into the growth data, the combinations that appears on the list comes from different batches (Figure 4.6). The combinations 5_3, 4_6 and 2_5 are in top 3. These correspond to low carbon and high nitrogen concentrations, over the 120 mM. Deviation in the OD measurements for each combination is similar, with a mean value of the coefficient of variation across the samples of 14%. However, there is a higher spread in the case of 1_REF. 1_REF is the second batch reference culture, using the original M9 recipe and with added bacteria. There are no indications of irregular growth at the visual level or any indications in the microplate reader run file that would be consider a cause of this spread.

Analogous to the Surfactin C case, in Figure 4.7, the carbon vs nitrogen contour plot is similar to the one obtained in the previous chapter, with optimum biomass attained in low carbon ~0.5% glucose and high nitrogen concentrations. Similar analysis was performed for the other lipopeptides and is shown in the Appendix B, Supplementary Figure B.10.

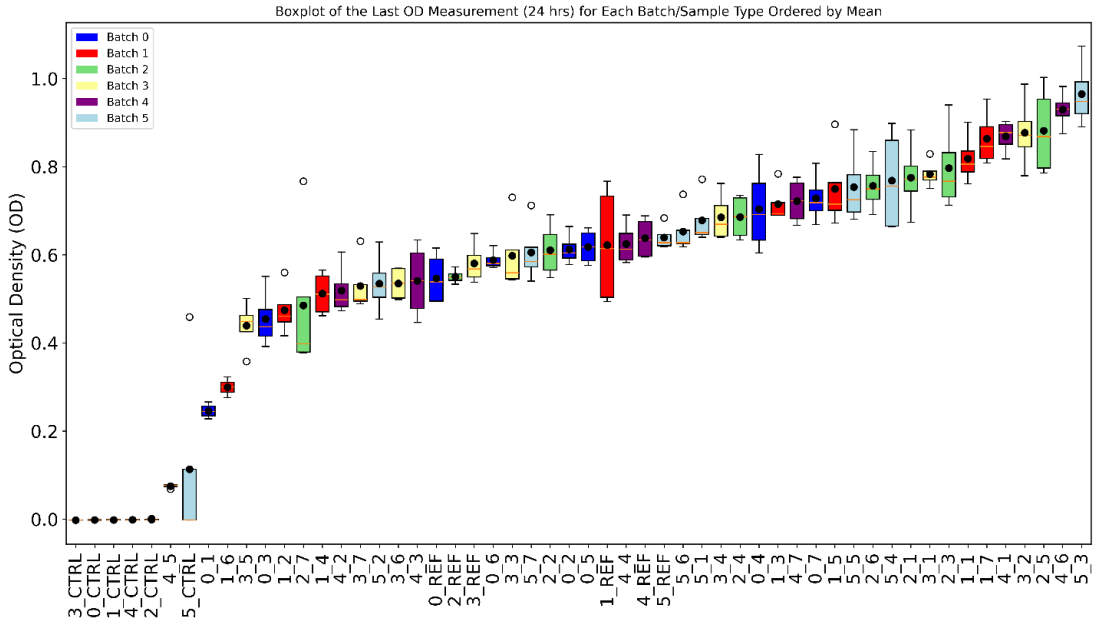


Figure 4.6. Optical Density (OD) measurements after 24 hours across different batches and combinations. This box plot displays the optical density measurements for the six batches over the multiple combinations and the control and reference samples. Each box indicates the interquartile range of OD measurements, with the median value represented by the line within each box. Whiskers extend to 1.5 times the interquartile range from the box, identifying the range of typical values. Outliers are shown as individual points. The data suggests several samples from different batch exhibiting high biomass, and the overall shape of the curve following the mean is a sigmoid shape, indicating a medium level of non-linearity of OD as function of the component concentrations.

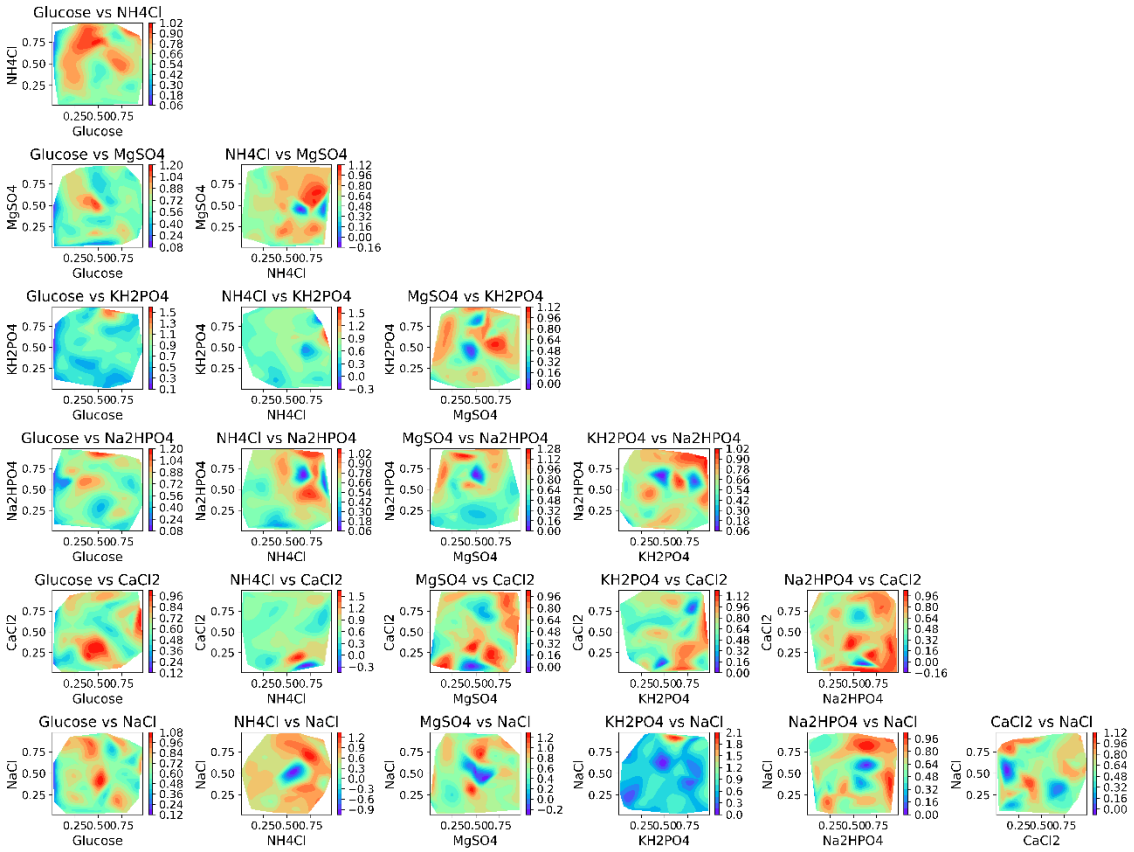


Figure 4.7. Contour plot of 24-hour Optical Density (OD) measurements against pairs of culture medium components. These plots serve a quick way to analyse dependencies of OD with the underlying factors in the experiments. Red colours indicate higher OD values while blue colours mean lower OD values. The surfaces are not provided by Gaussian process models, but instead by cubic spline interpolation of the empirical data.

When ordered the samples in the Surfactin C and growth data, the boxes follow a curve reminiscent of the logistic function. It is easy to show that ordering evaluation of random samples in a linear function gives a sigmoidal curve, known as probit function (Postelnicu, 2011). This is also known as the quantile function of the normal distribution, and mathematical details can be check in the Appendix A.3. Since our outputs possess are non-linear over the inputs and we used a sampling method different from drawing samples from a normal distribution, and that can be considered more uniform, it is expected that the shape of the ordered value as curve passing through the mean would be distorted, as seen in the left side of the samples from the Surfactin C quantification, suggesting non-linearity over the factors.

4.3.3. Selection of probabilistic surrogate model for the lipopeptide titres and growth measurements from space-filling design

For the benchmarking of Bayesian optimisation pipelines for single and multi-objective optimisation of lipopeptide production and corresponding growth measurements, the selection of an appropriate probabilistic surrogate model for the space-filling design data is critical. The chosen model must effectively handle the inherent variability and uncertainty associated with the lipopeptide titres shown in Figure 4.4, while also being robust enough to provide reliable predictions across the whole experimental design space.

The selection of a suitable probabilistic surrogate model involves several key criteria. The model must demonstrate high accuracy in predicting the lipopeptide titres and biomass measurements, to ensure that the model can reliably guide the optimisation process. This accuracy will be measured in terms of R^2 score and square root of the mean. Since the analysis of inputs and outputs in the previous section revealed nonlinearities of lipopeptide titres against medium composition such as Glucose, the surrogate model must be capable of capturing these relationships. Additionally, the model should be computationally efficient to allow for rapid predictions, which is particularly important when dealing with large datasets or when the model is used in real-time optimisation scenarios. Since we are working with small datasets (42 medium combination samples and 7 factors), classic probabilistic machine learning algorithms can run relatively faster, in contrast to scenario with hundreds or thousands of factors, where the model training can be become computationally intractable. Finally, the model should exhibit a generalization property, i.e, the capacity to predict unseen data. This property will be tested via k-fold cross-validation on our space-filling dataset.

Several probabilistic surrogate models are available, and I will focus on two of them: Gaussian Process Regression (GPR) and an Ensemble model with Bayesian model averaging, which aggregates several small machine learning algorithms predictions to provide probabilistic outputs. Other algorithm such as Bayesian linear regression and Bayesian neural network are no considered for different reasons; lack

of flexibility outside of the linear domain for the former and the latter usually involves huge datasets and is computationally expensive.

A Heteroskedastic Gaussian Process regression was fitted to the lipopeptide data after batch correction and normalisation. The fit to the data is almost perfect, reaching a R^2 score of 0.99. However, this a sign of overfitting, a typical situation with Gaussian process-based models in small datasets due to its flexibility (Figure 4.8, left). Indeed, cross-validation shows a low capacity of the model to extrapolate to unseen data (Figure 4.8, right).

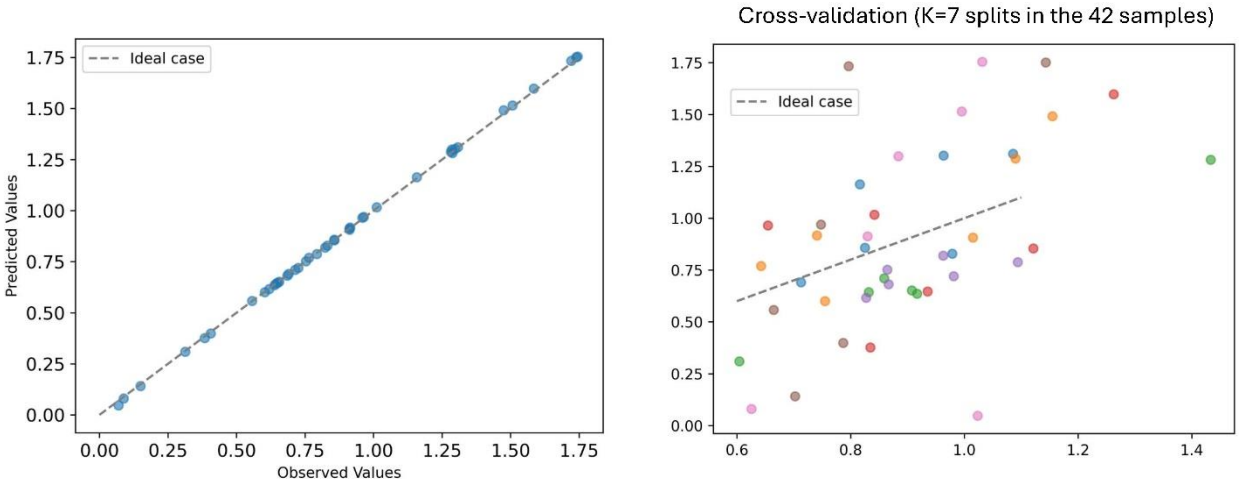


Figure 4.8. A "naïve" training of a Heteroskedastic Gaussian process to the processed Surfactin C titre data from the space-filling design. On the left, predicted values by the model are compared with the measured values in the samples. The fit is almost perfect, reaching a R^2 score of 0.99. However, on the right, sets of test points used for cross-validation (K=7 splits, N=42 samples from combinations) are shown with their corresponding predicted and observed. In this case, for each of the 7 sets of 6 points the fit is poor, and the mean R^2 score across the splits is 0.14.

There is an indication that the length scale associated with the kernel in this model is too small for a specific input, making the model too fine-grained and therefore exhibiting a good fit with the sampled points but poor generalisation. Effectively, after fitting the model, the inferred lengthscale for the Glucose variable (0.1663) is considerably lower than the other lengthscales (ammonium chloride: 0.4071, magnesium sulphate: 0.6388, potassium phosphate: 0.5140, sodium phosphate: 0.5334, calcium chloride: 0.5117, sodium chloride: 0.6020). Higher

lengthscales associated with a specific variable (closer to 1) indicate that the covariance function is almost independent of that input. That means that Glucose, as a factor, is important for the prediction, but at the same time it can bias the model in even smaller datasets, such as the ones obtained by splitting the data for cross-validation and therefore, the models lack explaining power for other combinations in the design space.

An alternative to fix this overfitting problem is restricting the lengthscale to higher values, such that the overall fit with the complete data will be poorer, but the cross-validation metric, and therefore the generalization property of the model, would increase. I tried a systematic constraining of the lenghtscale for the Glucose input from 0.2 to 1 and assess the fit and cross-validation results using R^2 scores and root square of the mean (RSME) as metrics. However, the results indicate that there is no further improvement in cross-validation metrics with the mean R^2 scores across split ranging between 0.1 to 0.2.

A second option is to use ensemble learning, where several smaller machine learning models (usually called base learners or weak learners) and fit into the data and their predictions are merged in a way that this new supra-prediction is better that of any individual base learner. In our case the base learners will correspond to machine learning models found in the scikit-learn library, including, and the ensemble framework is shown in Figure 4.9.

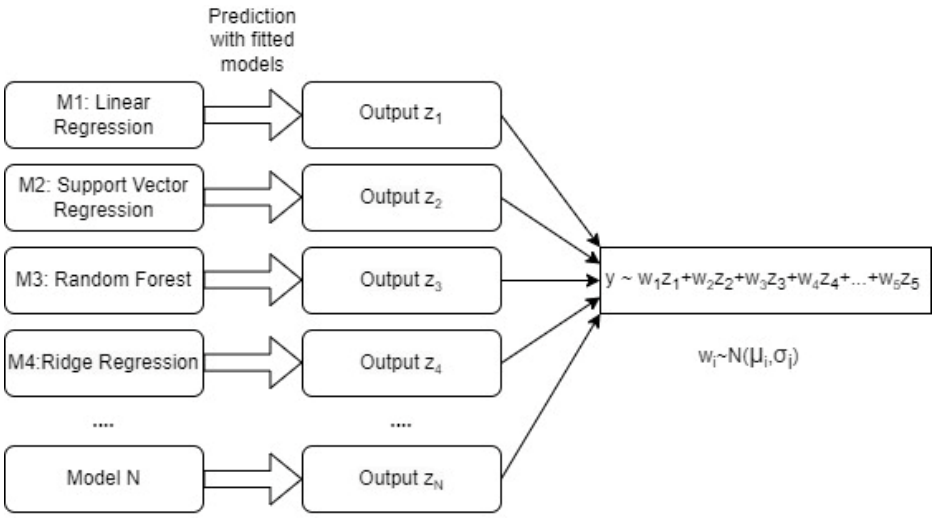


Figure 4.9. Diagram of a probabilistic ensemble learning algorithm using Bayesian averaging. Several machine learning models, known as base learners, weak learner, or

predictors, are trained in the dataset means. Then the outputs or predictions for the dataset inputs are aggregated into a probabilistic linear combination, where the weights are distributions depending on the prediction error for each base learner. Adapted from Radivojević et al., 2020.

In this case, the fit with the full data is reduced, reaching a R^2 score of 0.435, but obtaining similar mean R^2 score when performing cross-validation. That means that the model is least precise, but it has better generalisation properties. Similar scores have been reported in modelling of high-synthetic biology and metabolomics experiments, and therefore it is the chosen model for the Bayesian optimisation benchmarking.

We tested both approaches, now with the biomass data. In this case, the modelling results are slightly different. The biomass data seems be to be a smoother function of the medium components. To quantify this, we start by fitting as Heteroskedastic Gaussian process regression model as before, obtaining similar results in terms of prediction as in the case with Surfactin C (R^2 score=0.99). However, the cross-validation metrics shows a slightly improvement, with a mean R^2 score of 0.31 across splits. The importance of Glucose is also highlighted in this model, where the associated lenghtscale is 0.131, in comparison to the other variables in the order of 0.4-0.6. Then, turning into the Bayesian averaging ensemble model, the improvement is even higher in comparison to the ensemble Surfactin C model. Now the fit with the full data gives an R^2 score of 0.73, with similar scoring at the cross-validation level. The quality of the model is evident when testing the predictions in a small test set (Figure 4.10). Thus, this ensemble model is also selected for the benchmarking section.

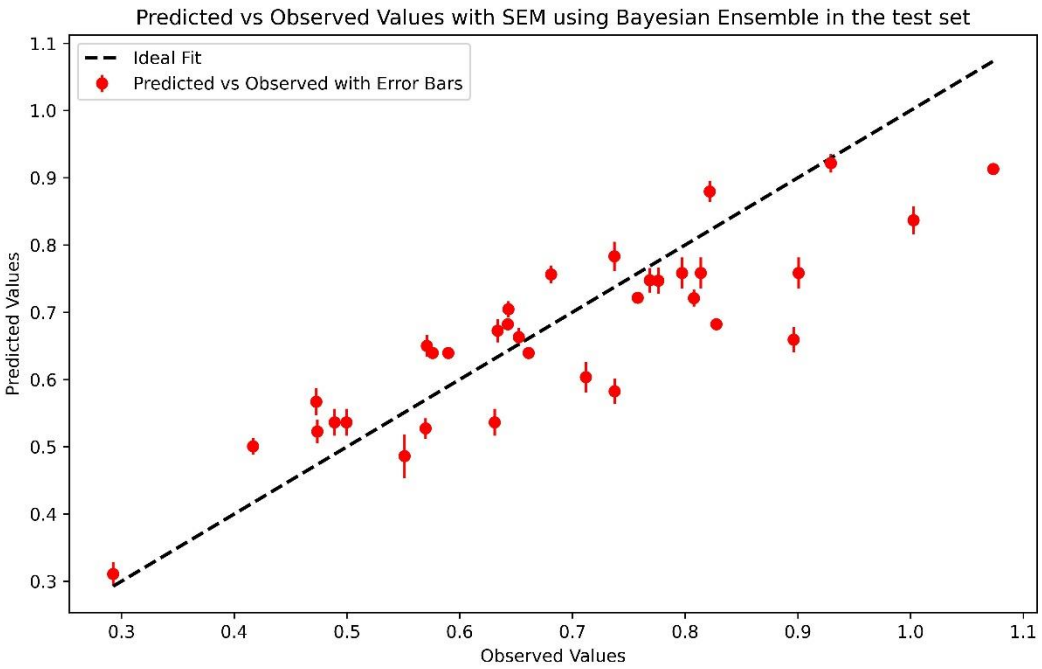


Figure 4.10. Predictions of the Bayesian averaging ensemble models for a test set in the Optical Density (OD) data. The predicted mean values can be read in y-axis, while the observed mean values can be traced in the x-axis. The bar represents the uncertainty on the prediction given by standard error of the mean.

The feature importance plots derived from the Random Forest base learner models in the two surrogate model scenarios—Surfactin C and OD—provide a clear comparative analysis of variable/input significance. In the Surfactin C and OD surrogate model, the feature importance plot illustrates a dominance of Glucose values as inputs with an importance value of 0.47 and 0.49 in each model respectively. This observation is in concordance with the inferred lenghtscales after fitting obtained from the Gaussian process regression model and suggests Glucose is the most critical input in the model (Figure 4.11).

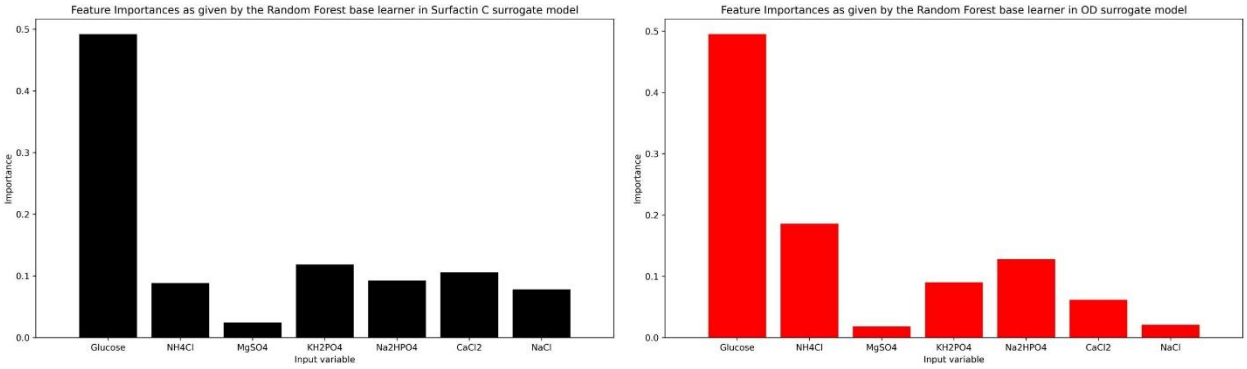


Figure 4.11. Feature importance as retrieved from the Random Forest base learner in the Bayesian averaging ensemble model. Random forest provides feature importance as part of the output. These results are obtained after training the model with the dataset from the space-filling experiment. On the left, the feature importance when fitting the Surfactin C titre is shown, while on the right the feature importance when fitting OD measurements is shown.

In the case of the other input variables, such as NH_4Cl, $MgSO_4$, K_2HPO_4, Na_2HPO_4, $CaCl_2$, and NaCl, they display significantly lower importance values, each below 0.2, indicating a lesser influence on the model's predictions. Interestingly, the Surfactin C and OD surrogate model presents a different pattern of feature importance for these inputs. Ammonium chloride shows a marked increase in importance in the OD model, reaching around 0.18, making it a substantial factor in this scenario. Then, it follows by the importance of the phosphates in the OD model, in contrast with the situation in the Surfactin C model where all other inputs except Glucose and Magnesium Sulphate are almost equally important.

4.3.4. Simulation of single- and multi-objective optimisation pipelines using the surrogate model as ground truth

We present the results of simulations performed on both single- and multi-objective Bayesian optimisation pipelines, using the surrogate models from the space-filling experiments as the ground truth. These simulations were designed to assess the efficacy of our proposed methods under *in silico* conditions in a high-dimensional space (7 factors), where the surrogate model is accurate enough to represent the metabolite tire or biomass being optimised. We tested several initial sampling sizes and different number of samples per batch, which is of utmost importance to estimate experimental effort in future experiments.

When considering benchmarking Bayesian optimisation pipelines for future experimental in *Bacillus* or other lipopeptide producers, there are practical decisions that need to be taken. Since bioprocess screening is an expensive and time-consuming process, we cannot go *ad infinitum* in terms of number of iterations. Therefore, we will establish a general 95% rule for the value to reach in the optimisation, taking as reference for the maximum "ideal" value the one obtained by dense simulation of points and evaluation in the Surfactin C surrogate model, as we considered the model as the "ground truth" for the *in silico* optimisation. This val. In this case, any combination of initial sampling size and batch size that does not reach 1.75 (0.95 x 1.85) in normalised Surfactin C titre before the 20 iterations will not consider as viable. This criterion establishes a clear optimisation goal, which can be achieved in reasonable time, since considering 20 iterations and 3 days of bench work per iteration involves 60 days (or 2 months) of theoretical continuous work. This timeframe is in some cases delayed, since experimental schedule can be affected by idle times, maintenance periods, accidents and errors in data manipulation. Also has a pragmatic consequence that it limits the number of iterations to test in our script, reducing execution time in the obtention of the results.

The results demonstrate that the single-objective optimization framework develop in Chapter 1 effectively identified optimal solutions in the 7D space, consistently converging towards the global optimum as defined by the surrogate model. For the case of 7 initial samples and 7 combinations per batch (Figure 4.12), the required value of 1.75 is obtained after 8 iterations (N=5 loop instances), which would correspond to ~ 1 month of continuous experimental work. When changing the number of initial samples and batch samples, a pattern emerges where it needs more or equal to 4 combinations per batch to reach the required values under the 20 iterations (Figure 4.12). Then, increasing the number of initial samples has a slight effect in reducing the number of iterations, and for the tested number of samples (between 2 and 10 initial samples and between 1 and 10 batches), the overall number of iterations necessary to reach the goal is around ~10 iterations, ranging from 20 iterations for 2 initial samples and 4 batch samples to 4 iterations in the extreme 10/10 case.

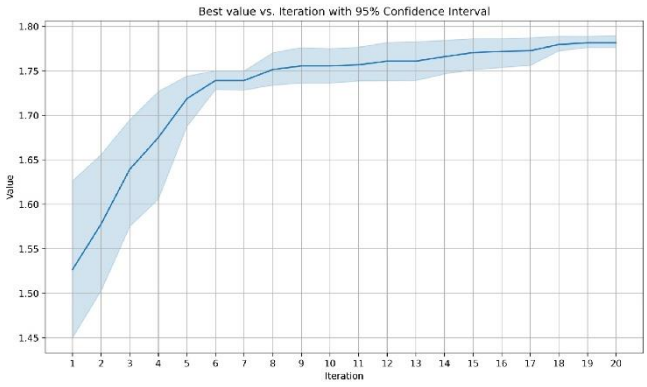


Figure 4.12. Performance of our single-objective optimisation Bayesian optimisation pipeline in the Surfactin C surrogate model. On the left, the best value for each iteration is plotted against the iteration number in the case of 7 initial samples and 7 combinations per batch, the same as our experimental protocol. 5 instances of full optimisation cycles were performed, so an estimation of the variance of the best value can be made since there some random components in the Bayesian optimisation loop, such as the initial Sobol sampling.

For the OD model, the maximum observed by dense simulation is 1.05, so the value to reach in the optimisation is 0.99 according to the 95% rule. Similar to the Surfactin C, the Bayesian optimisation loop reaches the value in 8 iterations for the case of 7 initial samples and 7 combinations per batch, which is the one currently established for experimental setup (Figure 4.13, A). When looking at the heatmap of the effect of sample sizes in the number of iterations to reach the goal, it differs slightly from the Surfactin C model, since 3 batch samples are the minimum necessary to get the required value in less than 20 iterations. Overall, for the different number of initial samples and batch sizes, it requires 9 iterations on average to reach 0.99.

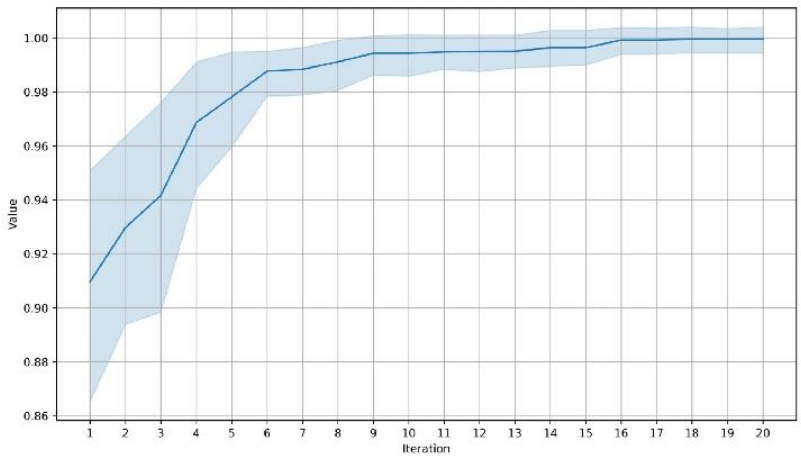


Figure 4.13. Performance of our single-objective optimisation Bayesian optimisation pipeline in the Optical Density (OD) surrogate model. On the left, the best value for each iteration is plotted against the iteration number in the case of 7 initial samples and 7 combinations per batch, the same as our experimental protocol. 5 instances of full optimisation cycles were performed, so an estimation of the variance of the best value can be made since there some random components in the Bayesian optimisation loop, such as the initial Sobol sampling.

After performing single objective optimisation, we proceed to test the pipeline for Pareto front discovery using multi-objective Bayesain optimisation, employing the q-Noisy Expected Improvement acquisition function. The multi-objective optimization showed robust performance in balancing the conflicting objectives of maximising Surfactin C titres and OD, achieving a set of Pareto-optimal solutions after 20 iterations with 7 initial/7 batch sample size configuration. Recalling from the introduction, the hypervolume for 2 objectives simply corresponds to the area delimited by the current Pareto front and a specific reference point, and for this optimisation we defined the reference in the origin (0,0). Thus, the hypervolume is an indicator of the progression in finding the Pareto plot across the iterations.

First, the maximum hypervolume for the inferred Pareto front after the evaluation of a dense simulation of points in the surrogate models reads 1.63 (Figure 4.14, A). According to our 95% rule, the goal of hypervolume to reach is 1.55. The Pareto front obtained after 20 iterations of the Bayesian optimisation loop, with 7 initial samples/7 batch samples is depicted in Figure 4.14 B. Tracing the hypervolume improvement across iterations, we verified that the goal is attained at 10 iterations for this setup (Figure 4.14 C), and this corresponds to the average number of iterations to reach the goal when checking different number of initial and batch samples.

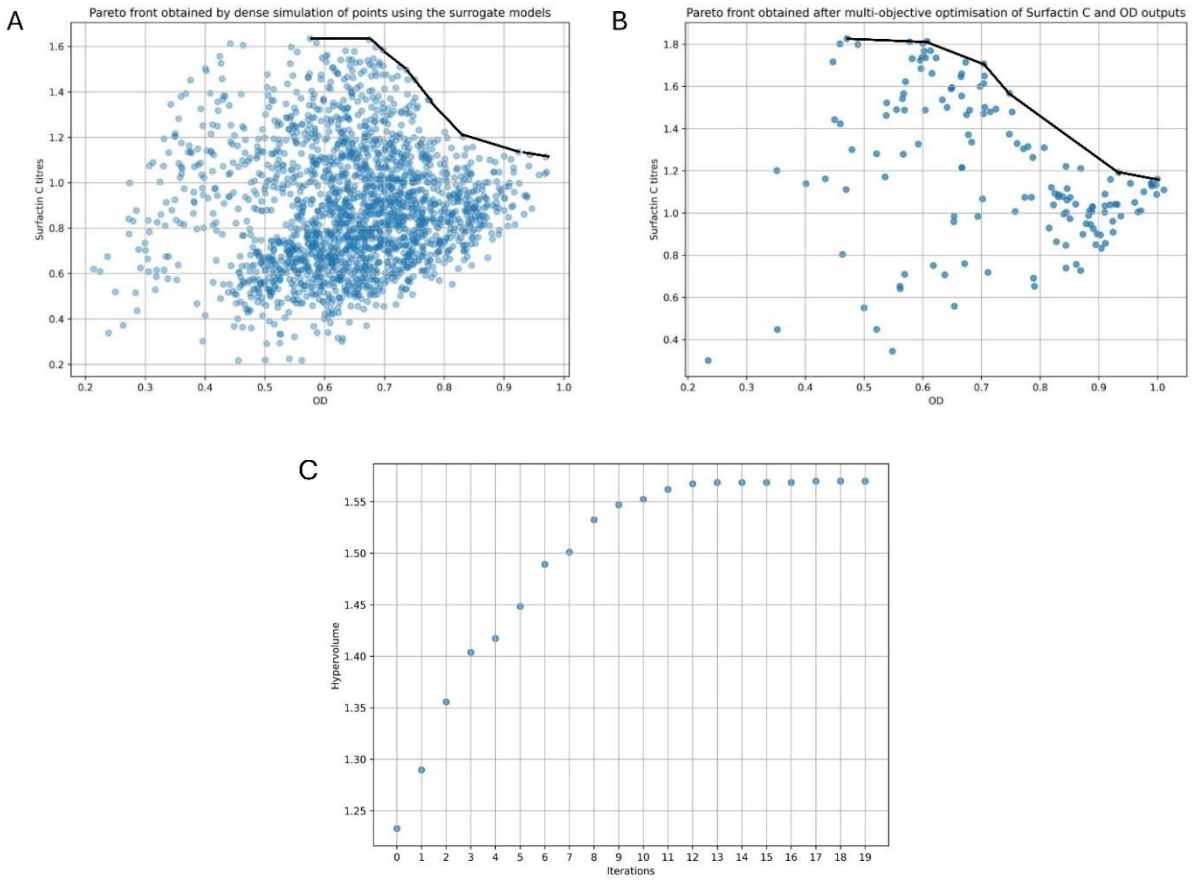


Figure 4.14. Performance of a multi-objective optimisation of Surfactin C titres and Optical Density, trying to maximise both simultaneously. A) The simulated pareto front by Sobol sampling 2000 points in the design space and evaluate them in the surrogate B) Pareto front found by our multi-objective Bayesian optimisation pipeline, using 7 initial samples and 7 combinations per batch over 20 iteration C) Hypervolume value for each iteration across the optimisation process

4.4. Discussion

4.4.1 Space-Filling Designs

Space-filling designs are invaluable in our study, facilitating thorough exploration of the design space with a minimal number of experimental runs. The uniform distribution of sampling points, as visualized in Figure 1.11, mitigates the risks associated with random sampling methods, such as information gaps and clustered points. This uniformity is pivotal for constructing accurate surrogate models, which serve as the foundation for subsequent optimization processes. Sobol sequences enabled a balanced and exhaustive sampling strategy, important when dealing with

high-dimensional spaces, but there is no limitation to using other quasi-random sequences in theory. Another important topic to consider, in contrast with literature, is the presence of replicates. Chiladwak (2022), as well as the other considered studies using space-filling designs, do not consider replicates on their final designs and the tools used for modelling are not probabilistic in nature. This scenario is the common in simulation, where a dense sampling of a ground truth function, with no noise, is easy to obtain and therefore Gaussian process regression may function as a simple interpolation on this data (an “emulator”) (Bastos & O’Hagan, 2009). However, this experiment and the trade-off between speed and accuracy discussed in Chapter 3, Indicates that efficient and reliable for noise measurements coming from biological systems needs replicates, and the saving in number of iterations can compensate the extra cost of samples and the bigger experimental effort per iteration.

4.4.2 Challenges and limitations in the modelling

Considering the noisy observations we have in our dataset; several challenges were encountered. One significant issue was the overfitting observed in the initial Gaussian Process Regression (GPR) models. As shown in Figure 4.8, heteroskedastic GPR models displayed high accuracy in fitting the training data but performed poorly in cross-validation. Efforts to address this discrepancy by constraining the lengthscale parameters did not yield substantial improvements, indicating the complexity of the modelled function. Thus, when turning to the ensemble learning approach, it resulted in reduced precision, but better generalization, enabling informative surrogate for the Bayesian optimisation pipeline testing. This trade-off between precision and generalization is a common challenge in high-dimensional optimization problems, underscoring the need for continuous refinement of modelling techniques. It is important to mention that ensembles are ubiquitous for the modelling of noisy biological data, spanning high-throughput transcript data (Cheng et al., 2024), microscopy image processing (Sorrentino et al., 2023) and prediction from multi-omics data (Tembhare et al., 2024), among others. The total number of parameters in these models, together with compute efficiency, allows for more flexibility in the fit and it can deal with the enormous sets produced today, making them a swiss-knife for biological big data analysis.

4.4.3 Optimising Metabolite Production and Biomass

Applying Bayesian optimization frameworks demonstrated their potential to identify optimal conditions for metabolite production and bacterial growth efficiently. The single-objective optimization consistently converged to high-producing solutions in around ~10 iteration, while the multi-objective optimization effectively delineated Pareto fronts in the same number of iterations, balancing the conflicting objectives of surfactant production and biomass (self-killing of the cells).

The simulation results validate the robustness of our optimisation pipelines. However, real-world application of these pipelines necessitates careful consideration of practical constraints, such as experimental costs and time limitations. Achieving experimentally a normalised Surfactin C titre of 1.7 within 20 iterations may provide a realistic benchmark for future experimental setups, ensuring the optimization process remains feasible within typical laboratory schedules. In this context, transitioning from a 96-well to a 48-well plate format, coupled with the use of robotic liquid handling systems, helped to streamline the cycles and reduce the burden on the human experimenter. Several examples exist on how this switch in the experimental paradigm has accelerate synthetic biology development in the academic environment (Holland & Davies, 2020) and more companies are now providing automated biology experiments services, with the expectation of this market becoming mainstream in the following years.

4.5. Conclusion

We systematically explored a seven-dimensional factor space, ensuring comprehensive coverage and robust modelling of the bioprocess experimental landscape. This data provides evidence of a trade-off between surfactin production and biomass inherent to biosurfactant synthesis, and that remains even when increasing the design space. Relevantly, glucose and nitrogen are still main factors influencing surfactin titre and OD in this multidimensional component space, but also other components acting as buffers might have an enough effect in titre to be considered later in further testing.

Future work will focus on further refining surrogate models to enhance their predictive accuracy and generalization capabilities. Exploring alternative modelling techniques, such as Bayesian neural networks or other advanced ensemble methods, could provide new insights into the complex dynamics of metabolite production, if they have better cross-validation behaviour. Expanding the experimental framework to include more factors, specially more complicated to manipulate reagents such as manganese and iron solutions (quick oxidation issues), and conditions will also enrich the experiment, enabling more comprehensive optimisation studies, as the ones found in the literature (Bertrand et al., 2018). We demonstrated that integrating automation in the media preparation protocol reduces the experimental burden on the scientist. Thus, the use of robotics, together with decrease of automated equipment prices, significantly reduce the time, money and effort required for experimental iterations, making the optimisation of bioprocesses more efficient and scalable.

4.6. Acknowledgments

We would like to thank EdinOmics for providing access to the mass spectrometry equipment. As in the previous chapter, we thank the Edward Wallace Lab for providing access to the Opentrons OT-2 platform.

4.7 Data availability.

Links to the raw data from the QqQ-MS runs, code scripts, microplate experiment layouts, and intermediate data tables are deposited at the Github repository https://github.com/rvalenciaaz/surfactin_bayesian_optimisation/tree/main/TEST/SOBOL_BIG/7F_48_WELL_COMPLETE_FILLING_M9_LIPOQQQ. Additionally, the scripts are available at the Github repository https://github.com/rvalenciaaz/surfactin_bayesian_optimisation/tree/main/TEST/SOBOL_BIG/7F_48_WELL_COMPLETE_FILLING_M9_LIPOQQQ.

Chapter 5

Discussion and Conclusion

5.1. Improving natural products' titres

Several secondary metabolites have intricated biosynthetic pathways. While some of these pathways have been extensively studied, and their regulation is well understood (O'Connor, 2015), many others remain elusive, with the metabolic origins of their precursors and the interactions between genetic components and corresponding biochemical reactions still unclear. Numerous efforts, including this thesis, are dedicated to enhancing the production of secondary metabolites, ranging from molecular strategies to optimising processes at the bioreactor scale. If we take *Bacillus* as an example, known for producing numerous commercially valuable molecules such as surfactin and iturin, as well as fengycins and bacteriocins (Abriouel et al., 2011, Harwood et al., 2018), there are several efforts to keep increasing titres and yields using, for example, medium optimisation (Mohanrasu et al., 2020) as in our project.

In many secondary metabolite bioprocesses, there remains considerable space for improving efficiency and safety. The integration of real-time monitoring and control systems using advanced sensors and machine learning algorithms can ensure that bioprocesses remain within optimal parameters, but often it is not enough to guarantee production goals. Thus, genetic strategies need to be combined with innovative fermentation technologies and medium optimisation on the long term, to improve the scalability and sustainability of secondary metabolite production. Secondary production is often not couple to growth, and long terms culture can select for fitness, reducing production. Media optimisation can be a way to modulate trade-offs inherent to secondary metabolite production, and to stimulate yield after several generations, as exemplified in the literature via a laboratory evolution approach (Jouhten et al., 2022)

5.2. Extracting the most from complex experimental data

Statistical analysis and data visualisation techniques are typically chosen based on the nature of the data and its inherent 'geometry'. Contour plots can be effective for exploring production surfaces in two-dimensional factor/component spaces. However, visualising multidimensional data is more challenging. Reducing multidimensional results into two-dimensional representations is the common method, such as PCA. It may introduce distortions, which must be carefully considered to ensure accurate conclusions are drawn.

Typically slicing over a set of dimensions, i.e., obtaining level sets may be an option, as in the case of Chapter 4 contour plots with pairs of components or using pairwise scatterplots, but without a model, it is difficult to assign importance to the variables and reflect this information in the visualisation process. Other approach in the literature is parallel coordinates (Inselberg, 2009) which aids to visualise correlations, but as mentioned, it is applicable mostly to multivariate data, instead of multivariable data, complementing the metabolite heatmap or the PCA. Thus, there is a need of more methods for visualising multivariable data, so it can be easier for humans to extract patterns in metabolomic data, apart from the traditional methods.

Utilising directional (or gradient) methods to analyse production surface data, presents a novel means of addressing new metabolic interactions. It can be seen as a local method of determining concentration trends, so their conclusions are tied to point in the design space chosen for the calculation. However, a good feature is that it can be applied to a function with any number of variables, just by increasing the number of representative directions. Even if a polar plot is not available, asymmetry can be estimated numerically. It can be considered a variant of the level set method, since gradients are perpendicular to level curves, and are naturally amenable to quantification. Future work will develop a new set of visualisations deriving from this gradient method.

5.3. The speed vs accuracy trade-off in protocol design

Biological experiments often require flexibility in their design. Optimisation workflows, therefore, must account for a wide range of variables and conditions and

requires consistency in quantification. Once the equipment is prepared for this, the number of factors can be increased to an unspecified limit, but then the speed vs accuracy trade-off becomes important.

In this regard, a method such as the flow injection mass spectrometry protocol developed in the project is of great importance to reduce optimisation cycles and experimenter effort, in comparison to HPLC-MS, which is more accurate due to separation, but much slower. Here, the presence of replicates was essential to get robust results in terms of noise for the correct performance of the Bayesian optimisation algorithm.

Employing the robot is also an accuracy vs speed compromise, that indeed in the long term, it is seen as more effective than human pipetting, simply because of degradation of pipetting performance over long experiments. As the optimisation process continues, leveraging these automated systems not only increases throughput but also maintains a higher degree of reproducibility.

On a small side note, speed vs accuracy trade-offs are also observed in biological systems. Decentralised decision systems such as ant colonies exhibit speed accuracy trade-offs in tasks such as house-hunting (Marshall et al., 2005). Indeed, decentralised and/or distributed laboratory systems might help to increase performance of optimisation by distributing biological testing over several platforms, with examples in the literature for chemical optimisation (Bai et al., 2024). But this kind of systems requires a higher degree of automation, which is the last topic in consideration.

5.4. The automation of the metabolic sciences

Optimisation is a versatile approach, applicable not only to achieving optimal titres but also to refining experimental protocols and pipelines. These decisions often require expert input, but there are ways to perform this process automatically, by exploiting the latest advances in artificial intelligence.

Indeed, many elements of the scientific process can be framed as optimisation problems: for example, which experimental protocol best tests a given hypothesis? To answer this question, there are additional trade-offs to consider, such as choosing

between a simpler, reagent-intensive protocol versus a more complex but sustainable one, and therefore a multi-objective optimisation approach would be feasible if proper quantification of these characteristics is achieved. It is also observed in power analysis, where determining the most efficient experimental design with sufficient power is critical for robust and significant results, but the number of samples also depends on funding factors and laboratory conditions.

We demonstrated that a pipeline can be generated for quasi-automatic optimisation of titres, and in the biological and analytical chemistry field, automation is starting to cover every aspect of the experiment, including expert decisions (Gao et al., 2024). In metabolomics, deep learning is already part of the processing pipeline, but it has not penetrated yet to hypothesis design/construction, as far as I know. Some work has been done regarding automated explainability of models for metabolomics (Bifarin & Fernández, 2024), and it would be interesting to see this approach implemented in large scale decision processes.

Safety is also important when considering fully automated metabolic experiments to ensure that all potential hazards are effectively managed. Automated systems can streamline complex processes, reduce human error, and increase reproducibility, but they also introduce new risks such as equipment malfunctions, software glitches, or unforeseen interactions between automated components and biological materials. Implementing rigorous safety protocols and workflow checkpoints, including regular maintenance, real-time monitoring, and fail-safe mechanisms, is crucial to prevent accidents and ensure reliable operation.

Can a whole metabolic experiment be fully automated, starting from *tabula rasa* or from a simple hypothesis given a scientist? With enough resources, it is probable. What kind of insights would be able to get from those experiments? It is difficult to know. It will depend on where the system is embedded (an academic lab or company) and the reliability of the technical decisions by the automated system, which is improving by the year. These insights may transform our vision on microbial metabolism in the next decades.

5.5. Conclusion

This project has explored optimisation of titres in the case of the natural product Surfactin in *Bacillus*. However, the methods and pipelines developed in this thesis are general, and we hope that it could be serve as a blueprint for more complex biotechnology and synthetic biology protocols. Thus, we demonstrate the value and effectiveness of adaptive, flexible Bayesian optimisation workflows that allow for refinement of culture medium composition, in line with the main hypothesis, benchmarking the efficiency and reliability of bioprocessing testing at the microplate scale.

References

Abdel-Rahman, M. A., Hassan, S. E. D., El-Din, M. N., Azab, M. S., El-Belely, E. F., Alrefaey, H. M. A., & Elsakhawy, T. (2020). One-factor-at-a-time and response surface statistical designs for improved lactic acid production from beet molasses by Enterococcus hirae ds10. SN Applied Sciences, 2(4), 573. https://doi.org/10.1007/s42452-020-2351-x

Abram, F., Arcari, T., Guerreiro, D., & O'Byrne, C. P. (2021). Chapter Four - Evolutionary trade-offs between growth and survival: The delicate balance between reproductive success and longevity in bacteria. In R. K. Poole & D. J. Kelly (Eds.), Advances in Microbial Physiology (Vol. 79, pp. 133–162). Academic Press. https://doi.org/10.1016/bs.ampbs.2021.07.002

Abriouel, H., Franz, C. M. A. P., Omar, N. B., & Gálvez, A. (2011). Diversity and applications of Bacillus bacteriocins. FEMS Microbiology Reviews, 35(1), 201–232. https://doi.org/10.1111/j.1574-6976.2010.00244.x

Ackley, D. (2012). A Connectionist Machine for Genetic Hillclimbing. Springer Science & Business Media.

Adamski, J. (2012). Genome-wide association studies with metabolomics. Genome Medicine, 4(4), 34. https://doi.org/10.1186/gm333

Agrawal, S., & Goyal, N. (2012). Analysis of Thompson Sampling for the Multi-armed Bandit Problem. Proceedings of the 25th Annual Conference on Learning Theory, 39.1-39.26. https://proceedings.mlr.press/v23/agrawal12.html

Ahsan, M. M., Luna, S. A., & Siddique, Z. (2022). Machine-Learning-Based Disease Diagnosis: A Comprehensive Review. Healthcare, 10(3), Article 3. https://doi.org/10.3390/healthcare10030541

Aichler, M., & Walch, A. (2015). MALDI Imaging mass spectrometry: Current frontiers and perspectives in pathology research and practice. Laboratory Investigation, 95(4), 422–431. https://doi.org/10.1038/labinvest.2014.156

Alisch, T., Crall, J. D., Kao, A. B., Zucker, D., & de Bivort, B. L. (2018). MAPLE (modular automated platform for large-scale experiments), a robot for integrated organism-handling and phenotyping. eLife, 7, e37166. https://doi.org/10.7554/eLife.37166

Alonso, S., & Martin, P. J. (2016). Impact of foaming on surfactin production by Bacillus subtilis: Implications on the development of integrated in situ foam fractionation removal systems. Biochemical Engineering Journal, 110, 125–133. https://doi.org/10.1016/j.bej.2016.02.006

Annesley, T. M. (2003). Ion Suppression in Mass Spectrometry. Clinical Chemistry, 49(7), 1041–1044. https://doi.org/10.1373/49.7.1041

Apache/airflow. (2024). [Python]. The Apache Software Foundation. https://github.com/apache/airflow (Original work published 2015)

ATCC. (n.d.). Bacillus subtilis (Ehrenberg) Cohn—21332 | ATCC. Retrieved 29 September 2024, from https://www.atcc.org/products/21332

Auer, P., Cesa-Bianchi, N., & Fischer, P. (2002). Finite-time Analysis of the Multiarmed Bandit Problem. Machine Learning, 47(2), 235–256. https://doi.org/10.1023/A:1013689704352

Azad, R. K., & Shulaev, V. (2019). Metabolomics technology and bioinformatics for precision medicine. Briefings in Bioinformatics, 20(6), 1957–1971. https://doi.org/10.1093/bib/bbx170

Babele, P. K., & Young, J. D. (2020). Applications of stable isotope-based metabolomics and fluxomics toward synthetic biology of cyanobacteria. WIREs Systems Biology and Medicine, 12(3), e1472. https://doi.org/10.1002/wsbm.1472

Babele, P., & Yadav, A. K. (2023). Back2Basics: Mass-to-charge ratio (m/z) in proteomics. Journal of Proteins and Proteomics, 14(4), 223–226. https://doi.org/10.1007/s42485-023-00115-7

Bäck, T. (1996). Artificial Landscapes. In T. Bäck (Ed.), Evolutionary Algorithms in Theory and Practice: Evolution Strategies, Evolutionary Programming, Genetic Algorithms (p. 0). Oxford University Press. https://doi.org/10.1093/oso/9780195099713.003.0008

Baden, T., Chagas, A. M., Gage, G., Marzullo, T., Prieto-Godino, L. L., & Euler, T. (2015). Open Labware: 3-D Printing Your Own Lab Equipment. PLOS Biology, 13(3), e1002086. https://doi.org/10.1371/journal.pbio.1002086

Bai, J., Mosbach, S., Taylor, C. J., Karan, D., Lee, K. F., Rihm, S. D., Akroyd, J., Lapkin, A. A., & Kraft, M. (2024). A dynamic knowledge graph approach to distributed self-driving laboratories. Nature Communications, 15(1), 462. https://doi.org/10.1038/s41467-023-44599-9

Balakrishnan, R., Mohan, N., & Sivaprakasam, S. (2022). Chapter 11 - Application of design of experiments in bioprocessing: Process analysis, optimization, and reliability. In R. Sirohi, A. Pandey, M. J. Taherzadeh, & C. Larroche (Eds.), Current Developments in Biotechnology and Bioengineering (pp. 289–319). Elsevier. https://doi.org/10.1016/B978-0-323-91167-2.00013-7

Balandat, M., Karrer, B., Jiang, D., Daulton, S., Letham, B., Wilson, A. G., & Bakshy, E. (2020). BoTorch: A Framework for Efficient Monte-Carlo Bayesian Optimization. Advances in Neural Information Processing Systems, 33, 21524–21538. https://proceedings.neurips.cc/paper/2020/hash/f5b1b89d98b7286673128a5fb112cb9a-Abstract.html

Banerjee, D., Eng, T., Lau, A. K., Sasaki, Y., Wang, B., Chen, Y., Prahl, J.-P., Singan, V. R., Herbert, R. A., Liu, Y., Tanjore, D., Petzold, C. J., Keasling, J. D., & Mukhopadhyay, A. (2020). Genome-scale metabolic rewiring improves titers rates and yields of the non-native product indigoidine at scale. Nature Communications, 11(1), Article 1. https://doi.org/10.1038/s41467-020-19171-4

Banerjee, D., & Mukhopadhyay, A. (2023). Perspectives in growth production trade-off in microbial bioproduction. RSC Sustainability, 1(2), 224–233. https://doi.org/10.1039/D2SU00066K

Banerjee, S., & Mazumdar, S. (2012). Electrospray Ionization Mass Spectrometry: A Technique to Access the Information beyond the Molecular Weight of the Analyte. International Journal of Analytical Chemistry, 2012(1), 282574. https://doi.org/10.1155/2012/282574

Barale, S. S., Ghane, S. G., & Sonawane, K. D. (2022). Purification and characterization of antibacterial surfactin isoforms produced by Bacillus velezensis SK. AMB Express, 12(1), 7. https://doi.org/10.1186/s13568-022-01348-3

Barrett, G. (2024). Development of high throughput metabolomics to aid the synthetic biology 'design-build-test-learn' cycle. https://doi.org/10.7488/era/4507

Bartal, A., Vigneshwari, A., Bóka, B., Vörös, M., Takács, I., Kredics, L., Manczinger, L., Varga, M., Vágvölgyi, C., & Szekeres, A. (2018). Effects of Different Cultivation Parameters on the Production of Surfactin Variants by a Bacillus subtilis Strain. Molecules, 23(10), Article 10. https://doi.org/10.3390/molecules23102675

Barton, R. H., Nicholson, J. K., Elliott, P., & Holmes, E. (2008). High-throughput 1H NMR-based metabolic analysis of human serum and urine for large-scale epidemiological studies: Validation study. International Journal of Epidemiology, 37(suppl_1), i31–i40. https://doi.org/10.1093/ije/dym284

Basha, N., Savage, T., McDonough, J., del Rio Chanona, E. A., & Matar, O. K. (2023). Discovery of mixing characteristics for enhancing coiled reactor performance through a Bayesian optimisation-CFD approach. Chemical Engineering Journal, 473, 145217.

Bates, S., Hastie, T., & Tibshirani, R. (2024). Cross-Validation: What Does It Estimate and How Well Does It Do It? Journal of the American Statistical Association, 119(546), 1434–1445. https://doi.org/10.1080/01621459.2023.2197686

Baumann, P., & Hubbuch, J. (2017). Downstream process development strategies for effective bioprocesses: Trends, progress, and combinatorial approaches. Engineering in Life Sciences, 17(11), 1142–1158. https://doi.org/10.1002/elsc.201600033

Behringer, M. G., Ho, W.-C., Miller, S. F., Worthan, S. B., Cen, Z., Stikeleather, R., & Lynch, M. (2024). Trade-offs, trade-ups, and high mutational parallelism underlie microbial adaptation during extreme cycles of feast and famine. Current Biology, 34(7), 1403-1413.e5. https://doi.org/10.1016/j.cub.2024.02.040

Belkin, M., Hsu, D., Ma, S., & Mandal, S. (2019). Reconciling modern machine-learning practice and the classical bias–variance trade-off. Proceedings of the National

Academy of Sciences, 116(32), 15849–15854. https://doi.org/10.1073/pnas.1903070116

Ben-David, S., & Shalev-Shwartz, S. (Eds.). (2014a). A Formal Learning Model. In Understanding Machine Learning: From Theory to Algorithms (pp. 22–30). Cambridge University Press. https://doi.org/10.1017/CBO9781107298019.004

Ben-David, S., & Shalev-Shwartz, S. (Eds.). (2014b). Dimensionality Reduction. In Understanding Machine Learning: From Theory to Algorithms (pp. 278–294). Cambridge University Press. https://doi.org/10.1017/CBO9781107298019.024

Ben-David, S., & Shalev-Shwartz, S. (Eds.). (2014c). Introduction. In Understanding Machine Learning: From Theory to Algorithms (pp. 1–10). Cambridge University Press. https://doi.org/10.1017/CBO9781107298019.002

Benkeblia, N. (2022). Metabolomics and sustainable agriculture: Concepts, applications, and perspectives. In Bioinformatics in Agriculture (pp. 123–138). Academic Press. https://doi.org/10.1016/B978-0-323-89778-5.00038-6

Berglund, E., Saarenpää, S., Jemt, A., Gruselius, J., Larsson, L., Bergenstråhle, L., Lundeberg, J., & Giacomello, S. (2020). Automation of Spatial Transcriptomics library preparation to enable rapid and robust insights into spatial organization of tissues. BMC Genomics, 21(1), 298. https://doi.org/10.1186/s12864-020-6631-z

Bernard, S., Heutte, L., & Adam, S. (2009). Influence of Hyperparameters on Random Forest Accuracy. In J. A. Benediktsson, J. Kittler, & F. Roli (Eds.), Multiple Classifier Systems (pp. 171–180). Springer. https://doi.org/10.1007/978-3-642-02326-2_18

Bersanelli, M., Mosca, E., Remondini, D., Giampieri, E., Sala, C., Castellani, G., & Milanesi, L. (2016). Methods for the integration of multi-omics data: Mathematical aspects. BMC Bioinformatics, 17(2), S15. https://doi.org/10.1186/s12859-015-0857-9

Bertaux, F., Sosa-Carrillo, S., Gross, V., Fraisse, A., Aditya, C., Furstenheim, M., & Batt, G. (2022). Enhancing bioreactor arrays for automated measurements and reactive control with ReacSight. Nature Communications, 13(1), 3363. https://doi.org/10.1038/s41467-022-31033-9

Bertrand, B., Martínez-Morales, F., Rosas-Galván, N. S., Morales-Guzmán, D., & Trejo-Hernández, M. R. (2018). Statistical Design, a Powerful Tool for Optimizing Biosurfactant Production: A Review. Colloids and Interfaces, 2(3), Article 3. https://doi.org/10.3390/colloids2030036

Bhadani, A., Kafle, A., Ogura, T., Akamatsu, M., Sakai, K., Sakai, H., & Abe, M. (2020). Current perspective of sustainable surfactants based on renewable building blocks. Current Opinion in Colloid & Interface Science, 45, 124–135. https://doi.org/10.1016/j.cocis.2020.01.002

Bhardwaj, C., & Hanley, L. (2014). Ion sources for mass spectrometric identification and imaging of molecular species. Natural Product Reports, 31(6), 756–767. https://doi.org/10.1039/C3NP70094A

Bhattacharjya, S., Ghosh, A., Sahu, A., Agnihotri, R., Pal, N., Sharma, P., Manna, M. C., Sharma, M. P., & Singh, A. B. (2024). Utilizing soil metabolomics to investigate the untapped metabolic potential of soil microbial communities and their role in driving soil ecosystem processes: A review. Applied Soil Ecology, 195, 105238. https://doi.org/10.1016/j.apsoil.2023.105238

Bhaturiwala, R., Bagban, M., Mansuri, A., & Modi, H. (2022). Successive approach of medium optimization using one-factor-at-a-time and response surface methodology for improved β-mannanase production from Streptomyces sp. Bioresource Technology Reports, 18, 101087. https://doi.org/10.1016/j.biteb.2022.101087

Bifarin, O. O., & Fernández, F. M. (2024). Automated Machine Learning and Explainable AI (AutoML-XAI) for Metabolomics: Improving Cancer Diagnostics. Journal of the American Society for Mass Spectrometry, 35(6), 1089–1100. https://doi.org/10.1021/jasms.3c00403

Biosurfactants Market, Industry Size Forecast, [Latest]. (n.d.). MarketsandMarkets. Retrieved 29 August 2024, from https://www.marketsandmarkets.com/Market-Reports/biosurfactant-market-163644922.html

Bloemberg, T. G., Gerretzen, J., Lunshof, A., Wehrens, R., & Buydens, L. M. C. (2013). Warping methods for spectroscopic and chromatographic signal alignment: A tutorial. Analytica Chimica Acta, 781, 14–32. https://doi.org/10.1016/j.aca.2013.03.048

Boesl, U. (2017). Time-of-flight mass spectrometry: Introduction to the basics. Mass Spectrometry Reviews, 36(1), 86–109. https://doi.org/10.1002/mas.21520

Boodhoo, K. V. K., Flickinger, M. C., Woodley, J. M., & Emanuelsson, E. A. C. (2022). Bioprocess intensification: A route to efficient and sustainable biocatalytic transformations for the future. Chemical Engineering and Processing - Process Intensification, 172, 108793. https://doi.org/10.1016/j.cep.2022.108793

Borkowski, O., Koch, M., Zettor, A., Pandi, A., Batista, A. C., Soudier, P., & Faulon, J.-L. (2020). Large scale active-learning-guided exploration for in vitro protein production optimization. Nature Communications, 11(1), Article 1. https://doi.org/10.1038/s41467-020-15798-5

Bossek, J., Doerr, C., & Kerschke, P. (2020). Initial design strategies and their effects on sequential model-based optimization: An exploratory case study based on BBOB. Proceedings of the 2020 Genetic and Evolutionary Computation Conference, 778–786. https://doi.org/10.1145/3377930.3390155

Brás, E. J. S., & Fernandes, P. C. de B. (2024). Miniaturization and microfluidic devices: An overview of basic concepts, fabrication techniques, and applications. Physical Sciences Reviews, 9(5), 2009–2036. https://doi.org/10.1515/psr-2022-0102

Breig, S. J. M., & Luti, K. J. K. (2021). Response surface methodology: A review on its applications and challenges in microbial cultures. Materials Today: Proceedings, 42, 2277–2284. https://doi.org/10.1016/j.matpr.2020.12.316

Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

Briscoe, E., & Feldman, J. (2011). Conceptual complexity and the bias/variance tradeoff. Cognition, 118(1), 2–16. https://doi.org/10.1016/j.cognition.2010.10.004

Brochu, E., Cora, V. M., & de Freitas, N. (2010). A Tutorial on Bayesian Optimization of Expensive Cost Functions, with Application to Active User Modeling and Hierarchical Reinforcement Learning. arXiv:1012.2599 [Cs]. http://arxiv.org/abs/1012.2599

Brown, R. W., Reay, M. K., Centler, F., Chadwick, D. R., Bull, I. D., McDonald, J. E., Evershed, R. P., & Jones, D. L. (2024). Soil metabolomics—Current challenges and

future perspectives. Soil Biology and Biochemistry, 193, 109382. https://doi.org/10.1016/j.soilbio.2024.109382

Bruggeman, F. J., Teusink, B., & Steuer, R. (2023). Trade-offs between the instantaneous growth rate and long-term fitness: Consequences for microbial physiology and predictive computational models. BioEssays, 45(10), 2300015. https://doi.org/10.1002/bies.202300015

Bryant, J. A., Jr., Kellinger, M., Longmire, C., Miller, R., & Wright, R. C. (2023). AssemblyTron: Flexible automation of DNA assembly with Opentrons OT-2 lab robots. Synthetic Biology, 8(1), ysac032. https://doi.org/10.1093/synbio/ysac032

Bull, A. D. (2011). Convergence Rates of Efficient Global Optimization Algorithms. Journal of Machine Learning Research, 12(88), 2879–2904.

Cambiaghi, A., Ferrario, M., & Masseroli, M. (2017). Analysis of metabolomic data: Tools, current strategies and future challenges for omics data integration. Briefings in Bioinformatics, 18(3), 498–510. https://doi.org/10.1093/bib/bbw031

Carbonell, P., Jervis, A. J., Robinson, C. J., Yan, C., Dunstan, M., Swainston, N., Vinaixa, M., Hollywood, K. A., Currin, A., Rattray, N. J. W., Taylor, S., Spiess, R., Sung, R., Williams, A. R., Fellows, D., Stanford, N. J., Mulherin, P., Le Feuvre, R., Barran, P., … Scrutton, N. S. (2018). An automated Design-Build-Test-Learn pipeline for enhanced microbial production of fine chemicals. Communications Biology, 1(1), 1–10. https://doi.org/10.1038/s42003-018-0076-9

Carvalho, S. M., Marques, J., Romão, C. C., & Saraiva, L. M. (2019). Metabolomics of Escherichia coli Treated with the Antimicrobial Carbon Monoxide-Releasing Molecule CORM-3 Reveals Tricarboxylic Acid Cycle as Major Target. Antimicrobial Agents and Chemotherapy, 63(10), 10.1128/aac.00643-19. https://doi.org/10.1128/aac.00643-19

Casas, A., Bultelle, M., & Kitney, R. (2024). An engineering biology approach to automated workflow and biodesign. Synthetic Biology, 9(1), ysae009. https://doi.org/10.1093/synbio/ysae009

Castelli, F. A., Rosati, G., Moguet, C., Fuentes, C., Marrugo-Ramírez, J., Lefebvre, T., Volland, H., Merkoçi, A., Simon, S., Fenaille, F., & Junot, C. (2022). Metabolomics for personalized medicine: The input of analytical chemistry from biomarker discovery to point-of-care tests. Analytical and Bioanalytical Chemistry, 414(2), 759–789. https://doi.org/10.1007/s00216-021-03586-z

Comella, N., & Grossman, A. D. (2005). Conservation of genes and processes controlled by the quorum response in bacteria: Characterization of genes controlled by the quorum-sensing transcription factor ComA in Bacillus subtilis. Molecular Microbiology, 57(4), 1159–1174. https://doi.org/10.1111/j.1365-2958.2005.04749.x

Chen, C., Wang, J., Pan, D., Wang, X., Xu, Y., Yan, J., Wang, L., Yang, X., Yang, M., & Liu, G.-P. (2023). Applications of multi-omics analysis in human diseases. MedComm, 4(4), e315. https://doi.org/10.1002/mco2.315

Chen, E. X., Russell, Z. E., Amsden, J. J., Wolter, S. D., Danell, R. M., Parker, C. B., Stoner, B. R., Gehm, M. E., Glass, J. T., & Brady, D. J. (2015). Order of Magnitude Signal Gain in Magnetic Sector Mass Spectrometry Via Aperture Coding. Journal of the American Society for Mass Spectrometry, 26(9), 1633–1640. https://doi.org/10.1007/s13361-015-1178-y

Chen, L., Guttieres, D., Koenigsberg, A., Barone, P. W., Sinskey, A. J., & Springs, S. L. (2022). Large-scale cultured meat production: Trends, challenges and promising biomanufacturing technologies. Biomaterials, 280, 121274. https://doi.org/10.1016/j.biomaterials.2021.121274

Cheng, Y., Bi, X., Xu, Y., Liu, Y., Li, J., Du, G., Lv, X., & Liu, L. (2023). Machine learning for metabolic pathway optimization: A review. Computational and Structural Biotechnology Journal, 21, 2381–2393. https://doi.org/10.1016/j.csbj.2023.03.045

Cheng, Y., Xu, S.-M., Santucci, K., Lindner, G., & Janitz, M. (2024). Machine learning and related approaches in transcriptomics. Biochemical and Biophysical Research Communications, 724, 150225. https://doi.org/10.1016/j.bbrc.2024.150225

Chilakwad, S. (2022). Automation and analysis of high-dimensionality experiments in biocatalytic reaction screening [Doctoral, UCL (University College London)]. In

Doctoral thesis, UCL (University College London). https://discovery.ucl.ac.uk/id/eprint/10156237/

Choi, K. R., Jang, W. D., Yang, D., Cho, J. S., Park, D., & Lee, S. Y. (2019). Systems Metabolic Engineering Strategies: Integrating Systems and Synthetic Biology with Metabolic Engineering. Trends in Biotechnology, 37(8), 817–837. https://doi.org/10.1016/j.tibtech.2019.01.003

Chory, E. J., Gretton, D. W., DeBenedictis, E. A., & Esvelt, K. M. (2021). Enabling high-throughput biology with flexible open-source automation. Molecular Systems Biology, 17(3), e9942. https://doi.org/10.15252/msb.20209942

Christensen, R. (2020). Introduction. In R. Christensen (Ed.), Plane Answers to Complex Questions: The Theory of Linear Models (pp. 1–20). Springer International Publishing. https://doi.org/10.1007/978-3-030-32097-3_1

Chumsakul, O., Takahashi, H., Oshima, T., Hishimoto, T., Kanaya, S., Ogasawara, N., & Ishikawa, S. (2011). Genome-wide binding profiles of the Bacillus subtilis transition state regulator AbrB and its homolog Abh reveals their interactive role in transcriptional regulation. Nucleic Acids Research, 39(2), 414–428. https://doi.org/10.1093/nar/gkq780

Ciasca, B., Pecorelli, I., Lepore, L., Paoloni, A., Catucci, L., Pascale, M., & Lattanzio, V. M. T. (2020). Rapid and reliable detection of glyphosate in pome fruits, berries, pulses and cereals by flow injection – Mass spectrometry. Food Chemistry, 310, 125813. https://doi.org/10.1016/j.foodchem.2019.125813

Cleasby, I. R., & Nakagawa, S. (2011). Neglected biological patterns in the residuals: A behavioural ecologist's guide to co-operating with heteroscedasticity. Behavioral Ecology and Sociobiology, 65(12), 2361–2372.

Clish, C. B. (2015). Metabolomics: An emerging but powerful tool for precision medicine. Molecular Case Studies, 1(1), a000588. https://doi.org/10.1101/mcs.a000588

Clomburg, J. M., Crumbley, A. M., & Gonzalez, R. (2017). Industrial biomanufacturing: The future of chemical production. Science, 355(6320), aag0804. https://doi.org/10.1126/science.aag0804

Collette, Y., & Siarry, P. (2004). Introduction: Multiobjective optimization and domination. In Y. Collette & P. Siarry (Eds.), Multiobjective Optimization: Principles and Case Studies (pp. 15–43). Springer. https://doi.org/10.1007/978-3-662-08883-8_1

Collins, S. L., Koo, I., Peters, J. M., Smith, P. B., & Patterson, A. D. (2021). Current Challenges and Recent Developments in Mass Spectrometry–Based Metabolomics. Annual Review of Analytical Chemistry, 14(Volume 14, 2021), 467–487. https://doi.org/10.1146/annurev-anchem-091620-015205

Comisarow, M. B., & Marshall, A. G. (1974). Fourier transform ion cyclotron resonance spectroscopy. Chemical Physics Letters, 25(2), 282–283. https://doi.org/10.1016/0009-2614(74)89137-2

Connor, M. C., Glass, B. H., Finkenstaedt-Quinn, S. A., & Shultz, G. V. (2021). Developing Expertise in 1H NMR Spectral Interpretation. The Journal of Organic Chemistry, 86(2), 1385–1395. https://doi.org/10.1021/acs.joc.0c01398

Cortada-Garcia, J., Daly, R., Arnold, S. A., & Burgess, K. (2023a). Streamlined identification of strain engineering targets for bioprocess improvement using metabolic pathway enrichment analysis. Scientific Reports, 13(1), 12990. https://doi.org/10.1038/s41598-023-39661-x

Cortada-Garcia, J., Daly, R., Arnold, S. A., & Burgess, K. (2023b). Streamlined identification of strain engineering targets for bioprocess improvement using metabolic pathway enrichment analysis. Scientific Reports, 13(1), 12990. https://doi.org/10.1038/s41598-023-39661-x

Cosenza, Z., Astudillo, R., Frazier, P. I., Baar, K., & Block, D. E. (2022). Multi-information source Bayesian optimization of culture media for cellular agriculture. Biotechnology and Bioengineering, 119(9), 2447–2458. https://doi.org/10.1002/bit.28132

Cosenza, Z., Block, D. E., Baar, K., & Chen, X. (2023). Multi-objective Bayesian algorithm automatically discovers low-cost high-growth serum-free media for cellular

agriculture application. Engineering in Life Sciences, 23(8), e2300005. https://doi.org/10.1002/elsc.202300005

Costanzo, M., Caterino, M., & Ruoppolo, M. (2022). Chapter 6—Targeted metabolomics. In J. Troisi (Ed.), Metabolomics Perspectives (pp. 219–236). Academic Press. https://doi.org/10.1016/B978-0-323-85062-9.00006-4

Costello, Z., & Martin, H. G. (2018). A machine learning approach to predict metabolic pathway dynamics from time-series multiomics data. Npj Systems Biology and Applications, 4(1), 1–14. https://doi.org/10.1038/s41540-018-0054-3

Couacault, P., Avella, D., Londoño-Osorio, S., Lorenzo, A. S., Gradillas, A., Kärkkäinen, O., Want, E., & Witting, M. (2024). Targeted and untargeted metabolomics and lipidomics in dried blood microsampling: Recent applications and perspectives. Analytical Science Advances, 5(5–6), e2400002. https://doi.org/10.1002/ansa.202400002

Councill, E. E. A. W., Axtell, N. B., Truong, T., Liang, Y., Aposhian, A. L., Webber, K. G. I., Zhu, Y., Cong, Y., Carson, R. H., & Kelly, R. T. (2021). Adapting a Low-Cost and Open-Source Commercial Pipetting Robot for Nanoliter Liquid Handling. SLAS Technology, 26(3), 311–319. https://doi.org/10.1177/2472630320973591

Cowen-Rivers, A. I., Lyu, W., Tutunov, R., Wang, Z., Grosnit, A., Griffiths, R. R., Maraval, A. M., Jianye, H., Wang, J., Peters, J., & Bou-Ammar, H. (2022). HEBO: Pushing The Limits of Sample-Efficient Hyper-parameter Optimisation. J. Artif. Int. Res., 74. https://doi.org/10.1613/jair.1.13643

Cramer, S. M., & Holstein, M. A. (2011). Downstream bioprocessing: Recent advances and future promise. Current Opinion in Chemical Engineering, 1(1), 27–37. https://doi.org/10.1016/j.coche.2011.08.008

Crutchfield, C. A., Thomas, S. N., Sokoll, L. J., & Chan, D. W. (2016). Advances in mass spectrometry-based clinical biomarker discovery. Clinical Proteomics, 13(1), 1. https://doi.org/10.1186/s12014-015-9102-9

Cullum, D. C. (1994). Surfactant types; classification, identification, separation. In D. C. Cullum (Ed.), Introduction to Surfactant Analysis (pp. 17–41). Springer Netherlands. https://doi.org/10.1007/978-94-011-1316-8_2

Czitrom, V. (1999). One-Factor-at-a-Time versus Designed Experiments. The American Statistician, 53(2), 126–131. https://doi.org/10.2307/2685731

da Silva, R. R., Dorrestein, P. C., & Quinn, R. A. (2015). Illuminating the dark matter in metabolomics. Proceedings of the National Academy of Sciences, 112(41), 12549–12550. https://doi.org/10.1073/pnas.1516878112

Danzi, F., Pacchiana, R., Mafficini, A., Scupoli, M. T., Scarpa, A., Donadelli, M., & Fiore, A. (2023a). To metabolomics and beyond: A technological portfolio to investigate cancer metabolism. Signal Transduction and Targeted Therapy, 8(1), 1–22. https://doi.org/10.1038/s41392-023-01380-0

Danzi, F., Pacchiana, R., Mafficini, A., Scupoli, M. T., Scarpa, A., Donadelli, M., & Fiore, A. (2023b). To metabolomics and beyond: A technological portfolio to investigate cancer metabolism. Signal Transduction and Targeted Therapy, 8(1), 1–22. https://doi.org/10.1038/s41392-023-01380-0

Dass, C. (2007). Tandem Mass Spectrometry. In Fundamentals of Contemporary Mass Spectrometry (pp. 119–150). John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470118498.ch4

DataBiosphere/toil. (2024). [Python]. Data Biosphere. https://github.com/DataBiosphere/toil (Original work published 2015)

D'Atri, V., Fekete, S., Clarke, A., Veuthey, J.-L., & Guillarme, D. (2019). Recent Advances in Chromatography for Pharmaceutical Analysis. Analytical Chemistry, 91(1), 210–239. https://doi.org/10.1021/acs.analchem.8b05026

Daulton, S., Ament, S., Eriksson, D., Balandat, M., & Bakshy, E. (2024). Unexpected improvements to expected improvement for Bayesian optimization. Proceedings of the 37th International Conference on Neural Information Processing Systems, 20577–20612.

Daulton, S., Wan, X., Eriksson, D., Balandat, M., Osborne, M. A., & Bakshy, E. (2024). Bayesian optimization over discrete and mixed spaces via probabilistic reparameterization. Proceedings of the 36th International Conference on Neural Information Processing Systems, 12760–12774.

de Jonge, N. F., Mildau, K., Meijer, D., Louwen, J. J. R., Bueschl, C., Huber, F., & van der Hooft, J. J. J. (2022). Good practices and recommendations for using and benchmarking computational metabolomics metabolite annotation tools. Metabolomics, 18(12), 103. https://doi.org/10.1007/s11306-022-01963-y

de Koning, W., & van Dam, K. (1992). A method for the determination of changes of glycolytic metabolites in yeast on a subsecond time scale using extraction at neutral pH. Analytical Biochemistry, 204(1), 118–123. https://doi.org/10.1016/0003-2697(92)90149-2

De Livera, A. M., Dias, D. A., De Souza, D., Rupasinghe, T., Pyke, J., Tull, D., Roessner, U., McConville, M., & Speed, T. P. (2012). Normalizing and Integrating Metabolomics Data. Analytical Chemistry, 84(24), 10768–10776. https://doi.org/10.1021/ac302748b

de Souza, F. M., & Gupta, R. K. (2024). Bacteria for Bioplastics: Progress, Applications, and Challenges. ACS Omega, 9(8), 8666–8686. https://doi.org/10.1021/acsomega.3c07372

Deisenroth, M. P., Faisal, A. A., & Ong, C. S. (2020, April 23). Mathematics for Machine Learning. Higher Education from Cambridge University Press; Cambridge University Press. https://doi.org/10.1017/9781108679930

DeJong, J. M., Liu, Y., Bollon, A. P., Long, R. M., Jennewein, S., Williams, D., & Croteau, R. B. (2006). Genetic engineering of taxol biosynthetic genes in Saccharomyces cerevisiae. Biotechnology and Bioengineering, 93(2), 212–224. https://doi.org/10.1002/bit.20694

Delvigne, F., Takors, R., Mudde, R., van Gulik, W., & Noorman, H. (2017). Bioprocess scale-up/down as integrative enabling technology: From fluid mechanics to systems biology and beyond. Microbial Biotechnology, 10(5), 1267–1274. https://doi.org/10.1111/1751-7915.12803

Demarque, D. P., Dusi, R. G., de Sousa, F. D. M., Grossi, S. M., Silvério, M. R. S., Lopes, N. P., & Espindola, L. S. (2020). Mass spectrometry-based metabolomics approach in the isolation of bioactive natural products. Scientific Reports, 10(1), 1051. https://doi.org/10.1038/s41598-020-58046-y

Deng, P., Li, X., Petriello, M. C., Wang, C., Morris, A. J., & Hennig, B. (2019). Application of metabolomics to characterize environmental pollutant toxicity and disease risks. Reviews on Environmental Health, 34(3), 251–259. https://doi.org/10.1515/reveh-2019-0030

Dettinger, P., Kull, T., Arekatla, G., Ahmed, N., Zhang, Y., Schneiter, F., Wehling, A., Schirmacher, D., Kawamura, S., Loeffler, D., & Schroeder, T. (2022). Open-source personal pipetting robots with live-cell incubation and microscopy compatibility. Nature Communications, 13(1), 2999. https://doi.org/10.1038/s41467-022-30643-7

Dettmer, K., Aronov, P. A., & Hammock, B. D. (2007). Mass spectrometry-based metabolomics. Mass Spectrometry Reviews, 26(1), 51–78. https://doi.org/10.1002/mas.20108

Deutsch, J. M., Mandelare-Ruiz, P., Yang, Y., Foster, G., Routhu, A., Houk, J., De La Flor, Y. T., Ushijima, B., Meyer, J. L., Paul, V. J., & Garg, N. (2022). Metabolomics Approaches to Dereplicate Natural Products from Coral-Derived Bioactive Bacteria. Journal of Natural Products, 85(3), 462–478. https://doi.org/10.1021/acs.jnatprod.1c01110

Di Tommaso, P., Chatzou, M., Floden, E. W., Barja, P. P., Palumbo, E., & Notredame, C. (2017). Nextflow enables reproducible computational workflows. Nature Biotechnology, 35(4), 316–319. https://doi.org/10.1038/nbt.3820

Dieterle, F., Ross, A., Schlotterbeck, G., & Senn, H. (2006). Probabilistic Quotient Normalization as Robust Method to Account for Dilution of Complex Biological Mixtures. Application in 1H NMR Metabonomics. Analytical Chemistry, 78(13), 4281–4290. https://doi.org/10.1021/ac051632c

Dixon, R. A., Gang, D. R., Charlton, A. J., Fiehn, O., Kuiper, H. A., Reynolds, T. L., Tjeerdema, R. S., Jeffery, E. H., German, J. B., Ridley, W. P., & Seiber, J. N. (2006). Applications of Metabolomics in Agriculture. Journal of Agricultural and Food Chemistry, 54(24), 8984–8994. https://doi.org/10.1021/jf061218t

Dlamini, B., Rangarajan, V., & Clarke, K. G. (2020). A simple thin layer chromatography based method for the quantitative analysis of biosurfactant surfactin vis-a-vis the presence of lipid and protein impurities in the processing liquid. Biocatalysis and Agricultural Biotechnology, 25, 101587. https://doi.org/10.1016/j.bcab.2020.101587

Doroudi, S. (2020). The Bias-Variance Tradeoff: How Data Science Can Inform Educational Debates. AERA Open, 6(4), 2332858420977208. https://doi.org/10.1177/2332858420977208

Dromms, R. A., & Styczynski, M. P. (2012). Systematic Applications of Metabolomics in Metabolic Engineering. Metabolites, 2(4), Article 4. https://doi.org/10.3390/metabo2041090

DSMZ. (n.d.). Leibniz Institute DSMZ: Details. Retrieved 29 September 2024, from https://www.dsmz.de/collection/catalogue/details/culture/DSM-3256

Dueñas, M. E., Peltier-Heap, R. E., Leveridge, M., Annan, R. S., Büttner, F. H., & Trost, M. (2023). Advances in high-throughput mass spectrometry in drug discovery. EMBO Molecular Medicine, 15(1), e14850. https://doi.org/10.15252/emmm.202114850

Dumas, T., Courant, F., Fenet, H., & Gomez, E. (2022). Environmental Metabolomics Promises and Achievements in the Field of Aquatic Ecotoxicology: Viewed through the Pharmaceutical Lens. Metabolites, 12(2), 186. https://doi.org/10.3390/metabo12020186

Ebbels, T. M. D., van der Hooft, J. J. J., Chatelaine, H., Broeckling, C., Zamboni, N., Hassoun, S., & Mathé, E. A. (2023). Recent advances in mass spectrometry-based computational metabolomics. Current Opinion in Chemical Biology, 74, 102288. https://doi.org/10.1016/j.cbpa.2023.102288

El-Aneed, A., Cohen, A., & Banoub, J. (2009). Mass Spectrometry, Review of the Basics: Electrospray, MALDI, and Commonly Used Mass Analyzers. Applied Spectroscopy Reviews, 44(3), 210–230. https://doi.org/10.1080/05704920902717872

Elfmann, C., Dumann, V., van den Berg, T., & Stülke, J. (2025). A new framework for SubtiWiki, the database for the model organism Bacillus subtilis. Nucleic Acids Research, 53(D1), D864–D870. https://doi.org/10.1093/nar/gkae95

Ellis, D. I., & Goodacre, R. (2012). Metabolomics-assisted synthetic biology. Current Opinion in Biotechnology, 23(1), 22–28. https://doi.org/10.1016/j.copbio.2011.10.014

Elyashberg, M. (2015). Identification and structure elucidation by NMR spectroscopy. TrAC Trends in Analytical Chemistry, 69, 88–97. https://doi.org/10.1016/j.trac.2015.02.014

Erb, D. (2024). pybaselines: A Python library of algorithms for the baseline correction of experimental data (Version v1.1.0) [Computer software]. Zenodo. https://doi.org/10.5281/zenodo.10676584

Eyke, N. S., Koscher, B. A., & Jensen, K. F. (2021). Toward Machine Learning-Enhanced High-Throughput Experimentation. Trends in Chemistry, 3(2), 120–132. https://doi.org/10.1016/j.trechm.2020.12.001

Faijes, M., Mars, A. E., & Smid, E. J. (2007). Comparison of quenching and extraction methodologies for metabolome analysis of Lactobacillus plantarum. Microbial Cell Factories, 6(1), 27. https://doi.org/10.1186/1475-2859-6-27

Florian, D. C., Odziomek, M., Ock, C. L., Chen, H., & Guelcher, S. A. (2020). Principles of computer-controlled linear motion applied to an open-source affordable liquid handler for automated micropipetting. Scientific Reports, 10(1), 13663. https://doi.org/10.1038/s41598-020-70465-5

Fong, E., & Holmes, C. C. (2020). On the marginal likelihood and cross-validation. Biometrika, 107(2), 489–496. https://doi.org/10.1093/biomet/asz077

Fonseca, R. R., Silva, A. J. R., França, F. P. D., Cardoso, V. L., & Sérvulo, E. F. C. (2007). Optimizing carbon/nitrogen ratio for biosurfactant production by a Bacillus subtilis strain. Applied Biochemistry and Biotechnology, 137(1), 471–486. https://doi.org/10.1007/s12010-007-9073-z

Fontes, G. C., Amaral, P. F. F., Nele, M., & Coelho, M. A. Z. (2010). Factorial design to optimize biosurfactant production by Yarrowia lipolytica. Journal of Biomedicine & Biotechnology, 2010, 821306. https://doi.org/10.1155/2010/821306

Frazier, P. I. (2018). A Tutorial on Bayesian Optimization. arXiv:1807.02811 [Cs, Math, Stat]. http://arxiv.org/abs/1807.02811

Frazier, P. I., Powell, W. B., & Dayanik, S. (2008). A Knowledge-Gradient Policy for Sequential Information Collection. SIAM Journal on Control and Optimization, 47(5), 2410–2439. https://doi.org/10.1137/070693424

Fu, Q., Murray, C. I., Karpov, O. A., & Van Eyk, J. E. (2023). Automated proteomic sample preparation: The key component for high throughput and quantitative mass spectrometry analysis. Mass Spectrometry Reviews, 42(2), e21750. https://doi.org/10.1002/mas.21750

Fuhrer, T., Heer, D., Begemann, B., & Zamboni, N. (2011). High-Throughput, Accurate Mass Metabolome Profiling of Cellular Extracts by Flow Injection–Time-of-Flight Mass Spectrometry. Analytical Chemistry, 83(18), 7074–7080. https://doi.org/10.1021/ac201267k

Furey, A., Moriarty, M., Bane, V., Kinsella, B., & Lehane, M. (2013). Ion suppression; A critical review on causes, evaluation, prevention and applications. Talanta, 115, 104–122. https://doi.org/10.1016/j.talanta.2013.03.048

Gabriel, K. R. (1971). The Biplot Graphic Display of Matrices with Application to Principal Component Analysis. Biometrika, 58(3), 453–467. https://doi.org/10.2307/2334381

Galal, A., Talal, M., & Moustafa, A. (2022). Applications of machine learning in metabolomics: Disease modeling and classification. Frontiers in Genetics, 13. https://doi.org/10.3389/fgene.2022.1017340

Gao, S., Fang, A., Huang, Y., Giunchiglia, V., Noori, A., Schwarz, J. R., Ektefaie, Y., Kondic, J., & Zitnik, M. (2024). Empowering Biomedical Discovery with AI Agents (arXiv:2404.02831). arXiv. https://doi.org/10.48550/arXiv.2404.02831

García-Pérez, P., Becchi, P. P., Zhang, L., Rocchetti, G., & Lucini, L. (2024). Metabolomics and chemometrics: The next-generation analytical toolkit for the evaluation of food quality and authenticity. Trends in Food Science & Technology, 147, 104481. https://doi.org/10.1016/j.tifs.2024.104481

Gates, S. C., & Sweeley, C. C. (1978). Quantitative metabolic profiling based on gas chromatography. Clinical Chemistry, 24(10), 1663–1673. https://doi.org/10.1093/clinchem/24.10.1663

Geissler, M., Oellig, C., Moss, K., Schwack, W., Henkel, M., & Hausmann, R. (2017). High-performance thin-layer chromatography (HPTLC) for the simultaneous quantification of the cyclic lipopeptides Surfactin, Iturin A and Fengycin in culture samples of Bacillus species. Journal of Chromatography B, 1044–1045, 214–224. https://doi.org/10.1016/j.jchromb.2016.11.013

Gelman, A., Hill, J., & Vehtari, A. (2020). Assumptions, diagnostics, and model evaluation. In Regression and Other Stories (pp. 153–182). Cambridge University Press.

German, J. B., Hammock, B. D., & Watkins, S. M. (2005a). Metabolomics: Building on a century of biochemistry to guide human health. Metabolomics, 1(1), 3–9. https://doi.org/10.1007/s11306-005-1102-8

German, J. B., Hammock, B. D., & Watkins, S. M. (2005b). Metabolomics: Building on a century of biochemistry to guide human health. Metabolomics, 1(1), 3–9. https://doi.org/10.1007/s11306-005-1102-8

Gertsman, I., & Barshop, B. A. (2018). Promises and pitfalls of untargeted metabolomics. Journal of Inherited Metabolic Disease, 41(3), 355–366. https://doi.org/10.1007/s10545-017-0130-7

Ghafari, N., & Sleno, L. (2024). Challenges and recent advances in quantitative mass spectrometry-based metabolomics. Analytical Science Advances, 5(5–6), e2400007. https://doi.org/10.1002/ansa.202400007

Gilman, J., Walls, L., Bandiera, L., & Menolascina, F. (2021). Statistical Design of Experiments for Synthetic Biology. ACS Synthetic Biology, 10(1), 1–18. https://doi.org/10.1021/acssynbio.0c00385

Glish, G. L., & Vachet, R. W. (2003). The basics of mass spectrometry in the twenty-first century. Nature Reviews Drug Discovery, 2(2), 140–150. https://doi.org/10.1038/nrd1011

Goldberg, P., Williams, C., & Bishop, C. (1997). Regression with Input-dependent Noise: A Gaussian Process Treatment. Advances in Neural Information Processing Systems, 10. https://papers.nips.cc/paper_files/paper/1997/hash/afe434653a898da20044041262b3ac74-Abstract.html

Gonzalez, J., Dai, Z., Hennig, P., & Lawrence, N. (2016). Batch Bayesian Optimization via Local Penalization. Proceedings of the 19th International Conference on Artificial Intelligence and Statistics, 648–657. https://proceedings.mlr.press/v51/gonzalez16a.html

González, J., Longworth, J., James, D. C., & Lawrence, N. D. (2015). Bayesian Optimization for Synthetic Gene Design (arXiv:1505.01627). arXiv. https://doi.org/10.48550/arXiv.1505.01627

Gonzalez-Riano, C., Saiz, J., Barbas, C., Bergareche, A., Huerta, J. M., Ardanaz, E., Konjevod, M., Mondragon, E., Erro, M. E., Chirlaque, M. D., Abilleira, E., Goñi-Irigoyen, F., & Amiano, P. (2021). Prognostic biomarkers of Parkinson's disease in the Spanish EPIC cohort: A multiplatform metabolomics approach. Npj Parkinson's Disease, 7(1), 1–12. https://doi.org/10.1038/s41531-021-00216-4

Görke, B., & Stülke, J. (2008). Carbon catabolite repression in bacteria: Many ways to make the most out of nutrients. Nature Reviews Microbiology, 6(8), 613–624. https://doi.org/10.1038/nrmicro1932

Govindarajan, S., Mannervik, B., Silverman, J. A., Wright, K., Regitsky, D., Hegazy, U., Purcell, T. J., Welch, M., Minshull, J., & Gustafsson, C. (2015). Mapping of Amino Acid Substitutions Conferring Herbicide Resistance in Wheat Glutathione Transferase. ACS Synthetic Biology, 4(3), 221–227. https://doi.org/10.1021/sb500242x

Gower, J., Lubbe, S., & le Roux, N. (2011a). Biplot Basics. In Understanding Biplots (pp. 11–66). John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470973196.ch2

Gower, J., Lubbe, S., & le Roux, N. (2011b). Principal Component Analysis Biplots. In Understanding Biplots (pp. 67–144). John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470973196.ch3

Greenacre, M. J. (2010). Biplots in practice. Fundacion BBVA. https://books.google.com/books?hl=en&lr=&id=dv4LrFP7U_EC&oi=fnd&pg=PA9&dq=info:fjwhW45BXssJ:scholar.google.com&ots=yFLGlNvNIN&sig=y2pPxy6Dch2gvXbP7VuxPf9fJg0

Greenhill, S., Rana, S., Gupta, S., Vellanki, P., & Venkatesh, S. (2020). Bayesian Optimization for Adaptive Experimental Design: A Review. IEEE Access, 8, 13937–13948. IEEE Access. https://doi.org/10.1109/ACCESS.2020.2966228

Griffiths, R.-R., Greenfield, J. L., Thawani, A. R., Jamasb, A. R., Moss, H. B., Bourached, A., Jones, P., McCorkindale, W., Aldrick, A. A., Fuchter, M. J., & Lee, A. A. (2022). Data-driven discovery of molecular photoswitches with multioutput Gaussian processes. Chemical Science, 13(45), 13541–13551. https://doi.org/10.1039/D2SC04306H

Griffiths, R.-R., & Hernández-Lobato, J. M. (2020). Constrained Bayesian optimization for automatic chemical design using variational autoencoders. Chemical Science, 11(2), 577–586.

Griffiths, W. J., & Wang, Y. (2009). Mass spectrometry: From proteomics to metabolomics and lipidomics. Chemical Society Reviews, 38(7), 1882–1896. https://doi.org/10.1039/B618553N

Gromski, P. S., Muhamadali, H., Ellis, D. I., Xu, Y., Correa, E., Turner, M. L., & Goodacre, R. (2015). A tutorial review: Metabolomics and partial least squares-discriminant analysis--a marriage of convenience or a shotgun wedding. Analytica Chimica Acta, 879, 10–23. https://doi.org/10.1016/j.aca.2015.02.012

Gross, J. H. (2017a). Instrumentation. In J. H. Gross (Ed.), Mass Spectrometry: A Textbook (pp. 151–292). Springer International Publishing. https://doi.org/10.1007/978-3-319-54398-7_4

Gross, J. H. (2017b). Introduction. In J. H. Gross (Ed.), Mass Spectrometry: A Textbook (pp. 1–28). Springer International Publishing. https://doi.org/10.1007/978-3-319-54398-7_1

Gross, J. H. (2017c). Practical Aspects of Electron Ionization. In J. H. Gross (Ed.), Mass Spectrometry: A Textbook (pp. 293–324). Springer International Publishing. https://doi.org/10.1007/978-3-319-54398-7_5

Gross, J. H. (2017d). Principles of Ionization and Ion Dissociation. In J. H. Gross (Ed.), Mass Spectrometry: A Textbook (pp. 29–84). Springer International Publishing. https://doi.org/10.1007/978-3-319-54398-7_2

Guerreiro, A. P., Fonseca, C. M., & Paquete, L. (2021). The Hypervolume Indicator: Computational Problems and Algorithms. ACM Comput. Surv., 54(6), 119:1-119:42. https://doi.org/10.1145/3453474

Guijas, C., Montenegro-Burke, J. R., Warth, B., Spilker, M. E., & Siuzdak, G. (2018). Metabolomics activity screening for identifying metabolites that modulate phenotype. Nature Biotechnology, 36(4), 316–320. https://doi.org/10.1038/nbt.4101

Gunka, K., & Commichau, F. M. (2012). Control of glutamate homeostasis in Bacillus subtilis: A complex interplay between ammonium assimilation, glutamate biosynthesis and degradation. Molecular Microbiology, 85(2), 213–224. https://doi.org/10.1111/j.1365-2958.2012.08105.x

Guo, Z., Sun, J., Ma, Q., Li, M., Dou, Y., Yang, S., & Gao, X. (2024). Improving Surfactin Production in Bacillus subtilis 168 by Metabolic Engineering. Microorganisms, 12(5), Article 5. https://doi.org/10.3390/microorganisms1205099

Haby, B., Hans, S., Anane, E., Sawatzki, A., Krausch, N., Neubauer, P., & Cruz Bournazou, M. N. (2019). Integrated Robotic Mini Bioreactor Platform for Automated, Parallel Microbial Cultivation With Online Data Handling and Process Control. SLAS Technology, 24(6), 569–582. https://doi.org/10.1177/2472630319860775

Hamm, J., Lim, S., Park, J., Kang, J., Lee, I., Lee, Y., Kang, J., Jo, Y., Lee, J., Lee, S., Ratri, M. C., Brilian, A. I., Lee, S., Jeong, S., & Shin, K. (2024). A Modular Robotic Platform for Biological Research: Cell Culture Automation and Remote Experimentation. Advanced Intelligent Systems, 6(5), 2300566. https://doi.org/10.1002/aisy.202300566

Han, W., & Li, L. (2022). Evaluating and minimizing batch effects in metabolomics. Mass Spectrometry Reviews, 41(3), 421–442. https://doi.org/10.1002/mas.21672

Han, Y., & Zhang, F. (2020). Control strategies to manage trade-offs during microbial production. Current Opinion in Biotechnology, 66, 158–164. https://doi.org/10.1016/j.copbio.2020.07.004

Härdle, W. K., & Simar, L. (2019). Principal Components Analysis. In W. K. Härdle & L. Simar (Eds.), Applied Multivariate Statistical Analysis (pp. 299–336). Springer International Publishing. https://doi.org/10.1007/978-3-030-26006-4_11

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., … Oliphant, T. E. (2020). Array programming with NumPy. Nature, 585(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

Harwood, C. R., Mouillon, J.-M., Pohl, S., & Arnau, J. (2018). Secondary metabolite production and the safety of industrially important members of the Bacillus subtilis group. FEMS Microbiology Reviews, 42(6), 721–738. https://doi.org/10.1093/femsre/fuy028

Hashizume, T., Ozawa, Y., & Ying, B.-W. (2023). Employing active learning in the optimization of culture medium for mammalian cells. Npj Systems Biology and Applications, 9(1), Article 1. https://doi.org/10.1038/s41540-023-00284-7

Hashizume, T., & Ying, B.-W. (2024). Challenges in developing cell culture media using machine learning. Biotechnology Advances, 70, 108293. https://doi.org/10.1016/j.biotechadv.2023.108293

Hastie, T., Tibshirani, R., & Friedman, J. (2009a). Overview of Supervised Learning. In T. Hastie, R. Tibshirani, & J. Friedman (Eds.), The Elements of Statistical Learning: Data Mining, Inference, and Prediction (pp. 9–41). Springer. https://doi.org/10.1007/978-0-387-84858-7_2

Hastie, T., Tibshirani, R., & Friedman, J. (2009b). Random Forests. In T. Hastie, R. Tibshirani, & J. Friedman (Eds.), The Elements of Statistical Learning: Data Mining, Inference, and Prediction (pp. 587–604). Springer. https://doi.org/10.1007/978-0-387-84858-7_15

Hastie, T., Tibshirani, R., & Friedman, J. (2009c). Unsupervised Learning. In T. Hastie, R. Tibshirani, & J. Friedman (Eds.), The Elements of Statistical Learning: Data Mining, Inference, and Prediction (pp. 485–585). Springer. https://doi.org/10.1007/978-0-387-84858-7_14

Hastie, T., Tibshirani, R., & Friedman, J. (2009d). Unsupervised Learning. In T. Hastie, R. Tibshirani, & J. Friedman (Eds.), The Elements of Statistical Learning: Data Mining, Inference, and Prediction (pp. 485–585). Springer. https://doi.org/10.1007/978-0-387-84858-7_14

Hayashi, K., Ohsawa, T., Kobayashi, K., Ogasawara, N., & Ogura, M. (2005). The H2O2 Stress-Responsive Regulator PerR Positively Regulates srfA Expression in Bacillus subtilis. Journal of Bacteriology, 187(19), 6659–6667. https://doi.org/10.1128/jb.187.19.6659-6667.2005

Hecht, E. S., Oberg, A. L., & Muddiman, D. C. (2016). Optimizing Mass Spectrometry Analyses: A Tailored Review on the Utility of Design of Experiments. Journal of The American Society for Mass Spectrometry, 27(5), 767–785. https://doi.org/10.1007/s13361-016-1344-x

Heiles, S. (2021). Advanced tandem mass spectrometry in metabolomics and lipidomics—Methods and applications. Analytical and Bioanalytical Chemistry, 413(24), 5927–5948. https://doi.org/10.1007/s00216-021-03425-1

Heinemann, J. (2019). Cluster Analysis of Untargeted Metabolomic Experiments. In E. E. K. Baidoo (Ed.), Microbial Metabolomics: Methods and Protocols (pp. 275–285). Springer. https://doi.org/10.1007/978-1-4939-8757-3_16

Henkel, M., Geissler, M., Weggenmann, F., & Hausmann, R. (2017). Production of microbial biosurfactants: Status quo of rhamnolipid and surfactin towards large-scale production. Biotechnology Journal, 12(7), 1600561. https://doi.org/10.1002/biot.201600561

Hérisson, J., Duigou, T., du Lac, M., Bazi-Kabbaj, K., Sabeti Azad, M., Buldum, G., Telle, O., El Moubayed, Y., Carbonell, P., Swainston, N., Zulkower, V., Kushwaha, M., Baldwin, G. S., & Faulon, J.-L. (2022). The automated Galaxy-SynBioCAD pipeline for synthetic

biology design and engineering. Nature Communications, 13(1), 5082. https://doi.org/10.1038/s41467-022-32661-x

Hettich. (2024). Automated Centrifuges. Retrieved 29 September 2024, from https://www.hettichlab.com/products/centrifuges/automated-centrifuges/

Hickman, R. J., Aldeghi, M., Häse, F., & Aspuru-Guzik, A. (2022). Bayesian optimization with known experimental and design constraints for chemistry applications. Digital Discovery, 1(5), 732–744.

Hill, C. B., Czauderna, T., Klapperstück, M., Roessner, U., & Schreiber, F. (2015). Metabolomics, Standards, and Metabolic Modeling for Synthetic Biology in Plants. Frontiers in Bioengineering and Biotechnology, 3. https://doi.org/10.3389/fbioe.2015.00167

Ho, C., Lam, C., Chan, M., Cheung, R., Law, L., Lit, L., Ng, K., Suen, M., & Tai, H. (2003). Electrospray Ionisation Mass Spectrometry: Principles and Clinical Applications. The Clinical Biochemist Reviews, 24(1), 3–12.

Hoeting, J. A., Madigan, D., Raftery, A. E., & Volinsky, C. T. (1999). Bayesian model averaging: A tutorial (with comments by M. Clyde, David Draper and E. I. George, and a rejoinder by the authors. Statistical Science, 14(4), 382–417. https://doi.org/10.1214/ss/1009212519

Holland, I., & Davies, J. A. (2020). Automation in the Life Science Research Laboratory. Frontiers in Bioengineering and Biotechnology, 8. https://doi.org/10.3389/fbioe.2020.571777

Hollywood, K. A., Schmidt, K., Takano, E., & Breitling, R. (2018). Metabolomics tools for the synthetic biology of natural products. Current Opinion in Biotechnology, 54, 114–120. https://doi.org/10.1016/j.copbio.2018.02.015

Hollywood, K., Brison, D. R., & Goodacre, R. (2006). Metabolomics: Current technologies and future trends. PROTEOMICS, 6(17), 4716–4723. https://doi.org/10.1002/pmic.200600106

Home. (n.d.). Apache Airflow. Retrieved 14 September 2024, from https://airflow.apache.org/

Hooft, J. J. J. van der, Mohimani, H., Bauermeister, A., Dorrestein, P. C., Duncan, K. R., & Medema, M. H. (2020). Linking genomics and metabolomics to chart specialized metabolic diversity. Chemical Society Reviews, 49(11), 3297–3314. https://doi.org/10.1039/D0CS00162G

Horak, I., Engelbrecht, G., van Rensburg, P. J. J., & Claassens, S. (2019). Microbial metabolomics: Essential definitions and the importance of cultivation conditions for utilizing Bacillus species as bionematicides. Journal of Applied Microbiology, 127(2), 326–343. https://doi.org/10.1111/jam.14218

Horning, E. C., & Horning, M. G. (1971). Metabolic Profiles: Gas-Phase Methods for Analysis of Metabolites. Clinical Chemistry, 17(8), 802–809. https://doi.org/10.1093/clinchem/17.8.802

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. Journal of Educational Psychology, 24(6), 417–441. https://doi.org/10.1037/h0071325

Hou, Y., Braun, D. R., Michel, C. R., Klassen, J. L., Adnani, N., Wyche, T. P., & Bugni, T. S. (2012). Microbial Strain Prioritization Using Metabolomics Tools for the Discovery of Natural Products. Analytical Chemistry, 84(10), 4277–4283. https://doi.org/10.1021/ac202623g

Hoult, D. I., Busby, S. J. W., Gadian, D. G., Radda, G. K., Richards, R. E., & Seeley, P. J. (1974). Observation of tissue metabolites using 31P nuclear magnetic resonance. Nature, 252(5481), 285–287. https://doi.org/10.1038/252285a0

Hu, F., Liu, Y., & Li, S. (2019). Rational strain improvement for surfactin production: Enhancing the yield and generating novel structures. Microbial Cell Factories, 18(1), 42. https://doi.org/10.1186/s12934-019-1089-x

Hu, Q., Noll, R. J., Li, H., Makarov, A., Hardman, M., & Graham Cooks, R. (2005). The Orbitrap: A new mass spectrometer. Journal of Mass Spectrometry, 40(4), 430–443. https://doi.org/10.1002/jms.856

Huang, Y., Swarge, B. N., Roseboom, W., Bleeker, J. D., Brul, S., Setlow, P., & Kramer, G. (2024). Integrative Metabolomics and Proteomics Allow the Global Intracellular

Characterization of Bacillus subtilis Cells and Spores. Journal of Proteome Research, 23(2), 596–608. https://doi.org/10.1021/acs.jproteome.3c00386

Hunt, M. M., Meng, G., Rancourt, D. E., Gates, I. D., & Kallos, M. S. (2014). Factorial Experimental Design for the Culture of Human Embryonic Stem Cells as Aggregates in Stirred Suspension Bioreactors Reveals the Potential for Interaction Effects Between Bioprocess Parameters. Tissue Engineering Part C: Methods, 20(1), 76–89. https://doi.org/10.1089/ten.tec.2013.0040

Iman, M. N., Herawati, E., Fukusaki, E., & Putri, S. P. (2022). Metabolomics-driven strain improvement: A mini review. Frontiers in Molecular Biosciences, 9. https://doi.org/10.3389/fmolb.2022.1057709

Inselberg, A. (2009). Parallel Coordinates: Visual Multidimensional Geometry and Its Applications. Springer. https://doi.org/10.1007/978-0-387-68628-8

Izenman, A. J. (2008a). Cluster Analysis. In A. J. Izenman (Ed.), Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning (pp. 407–462). Springer. https://doi.org/10.1007/978-0-387-78189-1_12

Izenman, A. J. (2008b). Linear Dimensionality Reduction. In A. J. Izenman (Ed.), Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning (pp. 195–236). Springer. https://doi.org/10.1007/978-0-387-78189-1_7

Jackson, M., Kavoussanakis, K., & Wallace, E. W. J. (2021). Using prototyping to choose a bioinformatics workflow management system. PLOS Computational Biology, 17(2), e1008622. https://doi.org/10.1371/journal.pcbi.1008622

Jacob, M., Lopata, A. L., Dasouki, M., & Abdel Rahman, A. M. (2019). Metabolomics toward personalized medicine. Mass Spectrometry Reviews, 38(3), 221–238. https://doi.org/10.1002/mas.21548

Jahan, R., Bodratti, A. M., Tsianou, M., & Alexandridis, P. (2020). Biosurfactants, natural alternatives to synthetic surfactants: Physicochemical properties and applications. Advances in Colloid and Interface Science, 275, 102061. https://doi.org/10.1016/j.cis.2019.102061

James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). Statistical Learning. In G. James, D. Witten, T. Hastie, R. Tibshirani, & J. Taylor (Eds.), An Introduction to Statistical Learning: With Applications in Python (pp. 15–67). Springer International Publishing. https://doi.org/10.1007/978-3-031-38747-0_2

Jankevics, A., Lloyd, G., & Weber, R. (2024). Pmp. Bioconductor. http://bioconductor.org/packages/pmp/

Jiao, J., Carella, A. J., Steeno, G. S., & Darrington, R. T. (2002). Optimization of triple quadrupole mass spectrometer for quantitation of trace degradants of pharmaceutical compounds. International Journal of Mass Spectrometry, 216(2), 209–218. https://doi.org/10.1016/S1387-3806(02)00593-6

Jin, Q., & Ma, R. C. W. (2021). Metabolomics in Diabetes and Diabetic Complications: Insights from Epidemiological Studies. Cells, 10(11), Article 11. https://doi.org/10.3390/cells10112832

John Roboz. (2016). A History of Ion Current Detectors for Mass Spectrometry. In M. L. Gross & R. M. Caprioli (Eds.), The Encyclopedia of Mass Spectrometry (pp. 183–188). Elsevier. https://doi.org/10.1016/B978-0-08-043848-1.00023-7

Johnson, C. H., Ivanisevic, J., & Siuzdak, G. (2016). Metabolomics: Beyond biomarkers and towards mechanisms. Nature Reviews Molecular Cell Biology, 17(7), 451–459. https://doi.org/10.1038/nrm.2016.25

Jolliffe, I. T. (Ed.). (2002). Introduction. In Principal Component Analysis (pp. 1–9). Springer. https://doi.org/10.1007/0-387-22440-8_1

Jones, J. A., & Koffas, M. A. G. (2016). Chapter Eight—Optimizing Metabolic Pathways for the Improved Production of Natural Products. In S. E. O'Connor (Ed.), Methods in Enzymology (Vol. 575, pp. 179–193). Academic Press. https://doi.org/10.1016/bs.mie.2016.02.010

Jones, T. S., Oliveira, S. M. D., Myers, C. J., Voigt, C. A., & Densmore, D. (2022). Genetic circuit design automation with Cello 2.0. Nature Protocols, 17(4), 1097–1113. https://doi.org/10.1038/s41596-021-00675-2

Jorayev, P., Russo, D., Tibbetts, J. D., Schweidtmann, A. M., Deutsch, P., Bull, S. D., & Lapkin, A. A. (2022). Multi-objective Bayesian optimisation of a two-step synthesis of p-cymene from crude sulphate turpentine. Chemical Engineering Science, 247, 116938.

Jouhten, P., Konstantinidis, D., Pereira, F., Andrejev, S., Grkovska, K., Castillo, S., Ghiachi, P., Beltran, G., Almaas, E., Mas, A., Warringer, J., Gonzalez, R., Morales, P., & Patil, K. R. (2022). Predictive evolution of metabolic phenotypes using model-designed environments. Molecular Systems Biology, 18(10), e10980. https://doi.org/10.15252/msb.202210980

Jozala, A. F., Geraldes, D. C., Tundisi, L. L., Feitosa, V. de A., Breyer, C. A., Cardoso, S. L., Mazzola, P. G., Oliveira-Nascimento, L. de, Rangel-Yagui, C. de O., Magalhães, P. de O., Oliveira, M. A. de, & Pessoa, A. (2016). Biopharmaceuticals from microorganisms: From production to purification. Brazilian Journal of Microbiology, 47, 51–63. https://doi.org/10.1016/j.bjm.2016.10.007

Kaiser, S., Dias, J. C., Ardila, J. A., Soares, F. L. F., Marcelo, M. C. A., Porte, L. M. F., Gonçalves, C., Canova, L. dos S., Pontes, O. F. S., & Sabin, G. P. (2018). High-throughput simultaneous quantitation of multi-analytes in tobacco by flow injection coupled to high-resolution mass spectrometry. Talanta, 190, 363–374. https://doi.org/10.1016/j.talanta.2018.08.007

Kalil, S. J., Maugeri, F., & Rodrigues, M. I. (2000). Response surface analysis and simulation as a tool for bioprocess design and optimization. Process Biochemistry, 35(6), 539–550. https://doi.org/10.1016/S0032-9592(99)00101-6

Kalivodová, A., Hron, K., Filzmoser, P., Najdekr, L., Janečková, H., & Adam, T. (2015). PLS-DA for compositional data with application to metabolomics. Journal of Chemometrics, 29(1), 21–28. https://doi.org/10.1002/cem.2657

Kaltenbach, H.-M. (2021a). Many Treatment Factors: Fractional Factorial Designs. In H.-M. Kaltenbach (Ed.), Statistical Design and Analysis of Biological Experiments (pp. 213–240). Springer International Publishing. https://doi.org/10.1007/978-3-030-69641-2_9

Kaltenbach, H.-M. (2021b). Multiple Treatment Factors: Factorial Designs. In H.-M. Kaltenbach (Ed.), Statistical Design and Analysis of Biological Experiments (pp. 121–156). Springer International Publishing. https://doi.org/10.1007/978-3-030-69641-2_6

Kampers, L. F. C., Asin-Garcia, E., Schaap, P. J., Wagemakers, A., & Martins dos Santos, V. A. P. (2022). Navigating the Valley of Death: Perceptions of Industry and Academia on Production Platforms and Opportunities in Biotechnology. EFB Bioeconomy Journal, 2, 100033. https://doi.org/10.1016/j.bioeco.2022.100033

Kanwal, M., Wattoo, A. G., Khushnood, R. A., Liaqat, A., Iqbal, R., & Song, Z. (2023). Chapter 12—Advancements and challenges in production of biosurfactants. In Inamuddin & C. O. Adetunji (Eds.), Applications of Next Generation Biosurfactants in the Food Sector (pp. 239–259). Academic Press. https://doi.org/10.1016/B978-0-12-824283-4.00019-8

Kasemiire, A., Avohou, H. T., De Bleye, C., Sacre, P.-Y., Dumont, E., Hubert, P., & Ziemons, E. (2021). Design of experiments and design space approaches in the pharmaceutical bioprocess optimization. European Journal of Pharmaceutics and Biopharmaceutics, 166, 144–154. https://doi.org/10.1016/j.ejpb.2021.06.004

Katsaggelos, A. K., Watt, J., & Borhani, R. (Eds.). (2016). Dimension reduction techniques. In Machine Learning Refined: Foundations, Algorithms, and Applications (pp. 245–262). Cambridge University Press. https://doi.org/10.1017/CBO9781316402276.013

Katunin, P., Zhou, J., Shehata, O. M., Peden, A. A., Cadby, A., & Nikolaev, A. (2021). An Open-Source Framework for Automated High-Throughput Cell Biology Experiments. Frontiers in Cell and Developmental Biology, 9. https://doi.org/10.3389/fcell.2021.697584

Kell, D. B., & Oliver, S. G. (2016). The metabolome 18 years on: A concept comes of age. Metabolomics, 12(9), 148. https://doi.org/10.1007/s11306-016-1108-4

Kellogg, J., & Kang, S. (2020). Metabolomics, an Essential Tool in Exploring and Harnessing Microbial Chemical Ecology. Phytobiomes Journal, 4(3), 195–210. https://doi.org/10.1094/PBIOMES-04-20-0032-RVW

Keng, H. L., & Yuan, W. (1981). Estimation of Discrepancy. In H. L. Keng & W. Yuan (Eds.), Applications of Number Theory to Numerical Analysis (pp. 70–98). Springer. https://doi.org/10.1007/978-3-642-67829-5_4

Khanijou, J. K., Kulyk, H., Bergès, C., Khoo, L. W., Ng, P., Yeo, H. C., Helmy, M., Bellvert, F., Chew, W., & Selvarajoo, K. (2022a). Metabolomics and modelling approaches for systems metabolic engineering. Metabolic Engineering Communications, 15, e00209. https://doi.org/10.1016/j.mec.2022.e00209

Khanijou, J. K., Kulyk, H., Bergès, C., Khoo, L. W., Ng, P., Yeo, H. C., Helmy, M., Bellvert, F., Chew, W., & Selvarajoo, K. (2022b). Metabolomics and modelling approaches for systems metabolic engineering. Metabolic Engineering Communications, 15, e00209. https://doi.org/10.1016/j.mec.2022.e00209

Kirwan, J. A., Broadhurst, D. I., Davidson, R. L., & Viant, M. R. (2013). Characterising and correcting batch variation in an automated direct infusion mass spectrometry (DIMS) metabolomics workflow. Analytical and Bioanalytical Chemistry, 405(15), 5147–5157. https://doi.org/10.1007/s00216-013-6856-7

Kirwan, J. A., Weber, R. J. M., Broadhurst, D. I., & Viant, M. R. (2014). Direct infusion mass spectrometry metabolomics dataset: A benchmark for data processing and quality control. Scientific Data, 1(1), Article 1. https://doi.org/10.1038/sdata.2014.12

Koblitz, J., Halama, P., Spring, S., Thiel, V., Baschien, C., Hahnke, R. L., Pester, M., Overmann, J., & Reimer, L. C. (2022). MediaDive: The expert-curated cultivation media database. Nucleic Acids Research, 51(D1), D1531–D1538. https://doi.org/10.1093/nar/gkac803

Kockmann, T., & Panse, C. (2021). The rawrr R Package: Direct Access to Orbitrap Data and Beyond. Journal of Proteome Research, 20(4), 2028–2034. https://doi.org/10.1021/acs.jproteome.0c00866

Konermann, L., Ahadi, E., Rodriguez, A. D., & Vahidi, S. (2013). Unraveling the Mechanism of Electrospray Ionization. Analytical Chemistry, 85(1), 2–9. https://doi.org/10.1021/ac302789c

Kong, D. S., Thorsen, T. A., Babb, J., Wick, S. T., Gam, J. J., Weiss, R., & Carr, P. A. (2017). Open-source, community-driven microfluidics with Metafluidics. Nature Biotechnology, 35(6), 523–529. https://doi.org/10.1038/nbt.3873

Kong, F., Yuan, L., Zheng, Y. F., & Chen, W. (2012). Automatic Liquid Handling for Life Science: A Critical Review of the Current State of the Art. SLAS Technology, 17(3), 169–185. https://doi.org/10.1177/2211068211435302

Konzock, O., & Nielsen, J. (2024). TRYing to evaluate production costs in microbial biotechnology. Trends in Biotechnology, 0(0). https://doi.org/10.1016/j.tibtech.2024.04.007

Koppenaal, D. W., Barinaga, C. J., Denton, M. B., Sperline, R. P., Hieftje, G. M., Schilling, G. D., Andrade, F. J., & Barnes, J. H. (2005). MS detectors. Analytical Chemistry, 77(21), 418A-427A. https://doi.org/10.1021/ac053495p

Kopyl, A., Yew, Y., Ong, J. W., Hiscox, T., Young, C., Muradoglu, M., & Ng, T. W. (2024). Automated Liquid Handler from a 3D Printer. Journal of Chemical Education, 101(2), 640–646. https://doi.org/10.1021/acs.jchemed.3c00855

Kortesniemi, M., Noerman, S., Kårlund, A., Raita, J., Meuronen, T., Koistinen, V., Landberg, R., & Hanhineva, K. (2023). Nutritional metabolomics: Recent developments and future needs. Current Opinion in Chemical Biology, 77, 102400. https://doi.org/10.1016/j.cbpa.2023.102400

Köster, J., & Rahmann, S. (2012). Snakemake—A scalable bioinformatics workflow engine. Bioinformatics, 28(19), 2520–2522. https://doi.org/10.1093/bioinformatics/bts480

Koutinas, M., Kiparissides, A., Pistikopoulos, E. N., & Mantalaris, A. (2012). BIOPROCESS SYSTEMS ENGINEERING: TRANSFERRING TRADITIONAL PROCESS ENGINEERING PRINCIPLES TO INDUSTRIAL BIOTECHNOLOGY. Computational and Structural Biotechnology Journal, 3(4). https://doi.org/10.5936/csbj.201210022

Kovacevic, V., & Simpson, M. (2020). Fundamentals of environmental metabolomics. In Environmental Metabolomics (pp. 1–33). Elsevier. https://doi.org/10.1016/B978-0-12-818196-6.00001-7

Kuehnbaum, N. L., & Britz-McKibbin, P. (2013). New Advances in Separation Science for Metabolomics: Resolving Chemical Diversity in a Post-Genomic Era. Chemical Reviews, 113(4), 2437–2468. https://doi.org/10.1021/cr300484s

Kumar, P. V., & Jin, Y. (2023). Bayesian Optimisation for Efficient Material Discovery: A Mini Review. Nanoscale.

Kumar, V., Bhalla, A., & Rathore, A. S. (2014). Design of experiments applications in bioprocessing: Concepts and approach. Biotechnology Progress, 30(1), 86–99. https://doi.org/10.1002/btpr.1821

Kumari, R., Singha, L. P., & Shukla, P. (2023). Biotechnological potential of microbial bio-surfactants, their significance, and diverse applications. FEMS Microbes, 4, xtad015. https://doi.org/10.1093/femsmc/xtad015

Kurtser, P., Castro-Alves, V., Arunachalam, A., Sjöberg, V., Hanell, U., Hyötyläinen, T., & Andreasson, H. (2021). Development of novel robotic platforms for mechanical stress induction, and their effects on plant morphology, elements, and metabolism. Scientific Reports, 11(1), 23876. https://doi.org/10.1038/s41598-021-02581-9

Kushner, H. J. (1964). A New Method of Locating the Maximum Point of an Arbitrary Multipeak Curve in the Presence of Noise. Journal of Basic Engineering, 86(1), 97–106. https://doi.org/10.1115/1.3653121

Kuster, B., Tüshaus, J., & Bayer, F. P. (2024). A new mass analyzer shakes up the proteomics field. Nature Biotechnology, 1–2. https://doi.org/10.1038/s41587-024-02129-y

Kwan, E. E., & Huang, S. G. (2008). Structural Elucidation with NMR Spectroscopy: Practical Strategies for Organic Chemists. European Journal of Organic Chemistry, 2008(16), 2671–2688. https://doi.org/10.1002/ejoc.200700966

Lapatas, V., Stefanidakis, M., Jimenez, R. C., Via, A., & Schneider, M. V. (2015). Data integration in biological research: An overview. Journal of Biological Research-Thessaloniki, 22(1), 9. https://doi.org/10.1186/s40709-015-0032-5

Lattermann, C., & Büchs, J. (2015). Microscale and miniscale fermentation and screening. Current Opinion in Biotechnology, 35, 1–6. https://doi.org/10.1016/j.copbio.2014.12.005

Lawson, J. (2015a). Factorial Designs. In Design and Analysis of Experiments with R. Chapman and Hall/CRC.

Lawson, J. (2015b). Response Surface Designs. In Design and Analysis of Experiments with R. Chapman and Hall/CRC.

Leão, T. F., Wang, M., da Silva, R., Gurevich, A., Bauermeister, A., Gomes, P. W. P., Brejnrod, A., Glukhov, E., Aron, A. T., Louwen, J. J. R., Kim, H. W., Reher, R., Fiore, M. F., van der Hooft, J. J. J., Gerwick, L., Gerwick, W. H., Bandeira, N., & Dorrestein, P. C. (2022). NPOmix: A machine learning classifier to connect mass spectrometry fragmentation data to biosynthetic gene clusters. PNAS Nexus, 1(5), pgac257. https://doi.org/10.1093/pnasnexus/pgac257

Leary, J. J., & Schmidt, R. L. (1996). Quadrupole Mass Spectrometers: An Intuitive Look at the Math. Journal of Chemical Education, 73(12), 1142. https://doi.org/10.1021/ed073p1142

Leitzen, S., Vogel, M., Steffens, M., Zapf, T., Müller, C. E., & Brandl, M. (2021). Quantification of Degradation Products Formed during Heat Sterilization of Glucose Solutions by LC-MS/MS: Impact of Autoclaving Temperature and Duration on Degradation. Pharmaceuticals, 14(11), 1121. https://doi.org/10.3390/ph14111121

Lempp, M., Farke, N., Kuntz, M., Freibert, S. A., Lill, R., & Link, H. (2019). Systematic identification of metabolites controlling gene expression in E. coli. Nature Communications, 10(1), 4463. https://doi.org/10.1038/s41467-019-12474-1

León, Z., García-Cañaveras, J. C., Donato, M. T., & Lahoz, A. (2013). Mammalian cell metabolomics: Experimental design and sample preparation. Electrophoresis, 34(19), 2762–2775. https://doi.org/10.1002/elps.201200605

Lever, J., Krzywinski, M., & Altman, N. (2017). Principal component analysis. Nature Methods, 14(7), 641–642. https://doi.org/10.1038/nmeth.4346

Li, C., Chu, S., Tan, S., Yin, X., Jiang, Y., Dai, X., Gong, X., Fang, X., & Tian, D. (2021). Towards Higher Sensitivity of Mass Spectrometry: A Perspective From the Mass Analyzers. Frontiers in Chemistry, 9. https://doi.org/10.3389/fchem.2021.813359

Li, D., Yi, J., Han, G., & Qiao, L. (2022). MALDI-TOF Mass Spectrometry in Clinical Analysis and Research. ACS Measurement Science Au, 2(5), 385–404. https://doi.org/10.1021/acsmeasuresciau.2c00019

Li, Z., Smith, K. H., & Stevens, G. W. (2016). The use of environmentally sustainable bio-derived solvents in solvent extraction applications—A review. Chinese Journal of Chemical Engineering, 24(2), 215–220. https://doi.org/10.1016/j.cjche.2015.07.021

Litchman, E., Edwards, K. F., & Klausmeier, C. A. (2015). Microbial resource utilization traits and trade-offs: Implications for community structure, functioning, and biogeochemical impacts at present and in the future. Frontiers in Microbiology, 6. https://doi.org/10.3389/fmicb.2015.00254

Liu, H., Ong, Y.-S., Shen, X., & Cai, J. (2020). When Gaussian Process Meets Big Data: A Review of Scalable GPs. IEEE Transactions on Neural Networks and Learning Systems, 31(11), 4405–4423. IEEE Transactions on Neural Networks and Learning Systems. https://doi.org/10.1109/TNNLS.2019.2957109

Liu, P., Hou, G., Kuang, Y., Li, L., Chen, C., Yan, B., Zhu, W., Li, J., Chen, M., Su, J., Lin, L., Chen, X., & Peng, C. (2023). Lipidomic profiling reveals metabolic signatures in psoriatic skin lesions. Clinical Immunology (Orlando, Fla.), 246, 109212. https://doi.org/10.1016/j.clim.2022.109212

Liu, X., & Locasale, J. W. (2017). Metabolomics: A Primer. Trends in Biochemical Sciences, 42(4), 274–284. https://doi.org/10.1016/j.tibs.2017.01.004

Lo, S., Baird, S. G., Schrier, J., Blaiszik, B., Carson, N., Foster, I., Aguilar-Granda, A., Kalinin, S. V., Maruyama, B., Politi, M., Tran, H., Sparks, T. D., & Aspuru-Guzik, A. (2024). Review of low-cost self-driving laboratories in chemistry and materials science: The “frugal twin” concept. Digital Discovery, 3(5), 842–868. https://doi.org/10.1039/D3DD00223C

Lu, H., Liang, Y., Dunn, W. B., Shen, H., & Kell, D. B. (2008). Comparative evaluation of software for deconvolution of metabolomics data based on GC-TOF-MS. TrAC Trends in Analytical Chemistry, 27(3), 215–227. https://doi.org/10.1016/j.trac.2007.11.004

Lu, W., Su, X., Klein, M. S., Lewis, I. A., Fiehn, O., & Rabinowitz, J. D. (2017). Metabolite Measurement: Pitfalls to Avoid and Practices to Follow. Annual Review of Biochemistry, 86(Volume 86, 2017), 277–304. https://doi.org/10.1146/annurev-biochem-061516-044952

Lu, Y., Pang, Z., & Xia, J. (2023). Comprehensive investigation of pathway enrichment methods for functional interpretation of LC–MS global metabolomics data. Briefings in Bioinformatics, 24(1), bbac553. https://doi.org/10.1093/bib/bbac553

Maan, K., Baghel, R., Dhariwal, S., Sharma, A., Bakhshi, R., & Rana, P. (2023). Metabolomics and transcriptomics based multi-omics integration reveals radiation-induced altered pathway networking and underlying mechanism. Npj Systems Biology and Applications, 9(1), 1–13. https://doi.org/10.1038/s41540-023-00305-5

Makarov, A., Denisov, E., Kholomeev, A., Balschun, W., Lange, O., Strupat, K., & Horning, S. (2006). Performance Evaluation of a Hybrid Linear Ion Trap/Orbitrap Mass Spectrometer. Analytical Chemistry, 78(7), 2113–2120. https://doi.org/10.1021/ac0518811

Mandenius, C.-F., & Brundin, A. (2008). Bioprocess optimization using design-of-experiments methodology. Biotechnology Progress, 24(6), 1191–1203. https://doi.org/10.1002/btpr.67

Manhart, M., & Shakhnovich, E. I. (2018). Growth tradeoffs produce complex microbial communities on a single limiting resource. Nature Communications, 9(1), 3214. https://doi.org/10.1038/s41467-018-05703-6

March, R. E. (2009). Quadrupole ion traps. Mass Spectrometry Reviews, 28(6), 961–989. https://doi.org/10.1002/mas.20250

Marshall, J. A. R., Dornhaus, A., Franks, N. R., & Kovacs, T. (2005). Noise, cost and speed-accuracy trade-offs: Decision-making in a decentralized system. Journal of The Royal Society Interface, 3(7), 243–254. https://doi.org/10.1098/rsif.2005.0075

Martin, K. N., Rubsamen, M. S., Kaplan, N. P., & Hendricks, M. P. (2022). Method for Interfacing a Plate Reader Spectrometer Directly with an OT-2 Liquid Handling Robot. ChemRxiv. https://doi.org/10.26434/chemrxiv-2022-6z4q1

Matanguihan, C., & Wu, P. (2022). Upstream continuous processing: Recent advances in production of biopharmaceuticals and challenges in manufacturing. Current Opinion in Biotechnology, 78, 102828. https://doi.org/10.1016/j.copbio.2022.102828

Matern, B. (2013). Spatial Variation. Springer Science & Business Media.

McKay, M. D., Beckman, R. J., & Conover, W. J. (1979). A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code. Technometrics, 21(2), 239–245. https://doi.org/10.2307/1268522

Mead, R. (1971). A Note on the Use and Misuse of Regression Models in Ecology. Journal of Ecology, 59(1), 215–219. https://doi.org/10.2307/2258463

Medra. (2024). Medra. Retrieved 29 September 2024, from https://www.medra.ai/

Melendez, J. R., Mátyás, B., Hena, S., Lowy, D. A., & El Salous, A. (2022). Perspectives in the production of bioethanol: A review of sustainable methods, technologies, and bioprocesses. Renewable and Sustainable Energy Reviews, 160, 112260. https://doi.org/10.1016/j.rser.2022.112260

Merzbacher, C., Mac Aodha, O., & Oyarzún, D. A. (2023). Bayesian Optimization for Design of Multiscale Biological Circuits. ACS Synthetic Biology, 12(7), 2073–2082. https://doi.org/10.1021/acssynbio.3c00120

Meyer, H., Weidmann, H., & Lalk, M. (2013). Methodological approaches to help unravel the intracellular metabolome of Bacillus subtilis. Microbial Cell Factories, 12(1), 69. https://doi.org/10.1186/1475-2859-12-69

Michael, S., Auld, D., Klumpp, C., Jadhav, A., Zheng, W., Thorne, N., Austin, C. P., Inglese, J., & Simeonov, A. (2008). A Robotic Platform for Quantitative High-Throughput Screening. ASSAY and Drug Development Technologies, 6(5), 637–657. https://doi.org/10.1089/adt.2008.150

Midani, F. S., Collins, J., & Britton, R. A. (2021). AMiGA: Software for Automated Analysis of Microbial Growth Assays. mSystems, 6(4), 10.1128/msystems.00508-21. https://doi.org/10.1128/msystems.00508-21

Miggiels, P., Wouters, B., van Westen, G. J. P., Dubbelman, A.-C., & Hankemeier, T. (2019). Novel technologies for metabolomics: More for less. TrAC Trends in Analytical Chemistry, 120, 115323. https://doi.org/10.1016/j.trac.2018.11.021

Miller, B. R., & Gulick, A. M. (2016). Structural Biology of Nonribosomal Peptide Synthetases. In B. S. Evans (Ed.), Nonribosomal Peptide and Polyketide Biosynthesis: Methods and Protocols (pp. 3–29). Springer. https://doi.org/10.1007/978-1-4939-3375-4_1

Miller, J. (2009). Introduction to Chromatography. In Chromatography (pp. 35–66). John Wiley & Sons, Ltd. https://doi.org/10.1002/9780471980582.ch2

Miller, M. G. (2007). Environmental Metabolomics: A SWOT Analysis (Strengths, Weaknesses, Opportunities, and Threats). Journal of Proteome Research, 6(2), 540–545. https://doi.org/10.1021/pr060623x

Miller, P. E., & Denton, M. B. (1986). The quadrupole mass filter: Basic operating concepts. Journal of Chemical Education, 63(7), 617. https://doi.org/10.1021/ed063p617

Milman, B. L., & Zhurkovich, I. K. (2016). Mass spectral libraries: A statistical review of the visible use. TrAC Trends in Analytical Chemistry, 80, 636–640. https://doi.org/10.1016/j.trac.2016.04.024

Močkus, J. (1975). On bayesian methods for seeking the extremum. In G. I. Marchuk (Ed.), Optimization Techniques IFIP Technical Conference Novosibirsk, July 1–7, 1974 (pp. 400–404). Springer. https://doi.org/10.1007/3-540-07165-2_55

Mohanrasu, K., Rao, R. G. R., Dinesh, G. H., Zhang, K., Prakash, G. S., Song, D.-P., Muniyasamy, S., Pugazhendhi, A., Jeyakanthan, J., & Arun, A. (2020). Optimization of media components and culture conditions for polyhydroxyalkanoates production by Bacillus megaterium. Fuel, 271, 117522. https://doi.org/10.1016/j.fuel.2020.117522

Mohanty, S. S., Koul, Y., Varjani, S., Pandey, A., Ngo, H. H., Chang, J.-S., Wong, J. W. C., & Bui, X.-T. (2021). A critical review on various feedstocks as sustainable substrates for biosurfactants production: A way towards cleaner production. Microbial Cell Factories, 20(1), 120. https://doi.org/10.1186/s12934-021-01613-3

Mohd Kamal, K., Mahamad Maifiah, M. H., Abdul Rahim, N., Hashim, Y. Z. H.-Y., Abdullah Sani, M. S., & Azizan, K. A. (2022). Bacterial Metabolomics: Sample Preparation Methods. Biochemistry Research International, 2022(1), 9186536. https://doi.org/10.1155/2022/9186536

Mondal, P. P., Galodha, A., Verma, V. K., Singh, V., Show, P. L., Awasthi, M. K., Lall, B., Anees, S., Pollmann, K., & Jain, R. (2023). Review on machine learning-based bioprocess optimization, monitoring, and control systems. Bioresource Technology, 370, 128523. https://doi.org/10.1016/j.biortech.2022.128523

Moros, G., Chatziioannou, A. C., Gika, H. G., Raikos, N., & Theodoridis, G. (2017). Investigation of The Derivatization Conditions for GC–MS Metabolomics of Biological Samples. Bioanalysis, 9(1), 53–65. https://doi.org/10.4155/bio-2016-0224

Muhamadali, H., Winder, C. L., Dunn, W. B., & Goodacre, R. (2023). Unlocking the secrets of the microbiome: Exploring the dynamic microbial interplay with humans through metabolomics and their manipulation for synthetic biology applications. Biochemical Journal, 480(12), 891–908. https://doi.org/10.1042/BCJ20210534

Myers, D. (2020). The Classification of Surfactants. In Surfactant Science and Technology (pp. 17–59). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781119465829.ch2

Nakano, M. M., Xia, L. A., & Zuber, P. (1991). Transcription initiation region of the srfA operon, which is controlled by the comP-comA signal transduction system in Bacillus subtilis. Journal of Bacteriology, 173(17), 5487–5493. https://doi.org/10.1128/jb.173.17.5487-5493.1991

Nakano, S., Nakano, M. M., Zhang, Y., Leelakriangsak, M., & Zuber, P. (2003). A regulatory protein that interferes with activator-stimulated transcription in bacteria. Proceedings of the National Academy of Sciences, 100(7), 4233–4238. https://doi.org/10.1073/pnas.0637648100

Nanita, S. C., & Kaldon, L. G. (2016). Emerging flow injection mass spectrometry methods for high-throughput quantitative analysis. Analytical and Bioanalytical Chemistry, 408(1), 23–33. https://doi.org/10.1007/s00216-015-9193-1

Nava, E., Mutný, M., & Krause, A. (2022). Diversified Sampling for Batched Bayesian Optimization with Determinantal Point Processes (arXiv:2110.11665). arXiv. https://doi.org/10.48550/arXiv.2110.11665

Neagu, A.-N., Jayathirtha, M., Baxter, E., Donnelly, M., Petre, B. A., & Darie, C. C. (2022). Applications of Tandem Mass Spectrometry (MS/MS) in Protein Analysis for Biomedical Research. Molecules, 27(8), Article 8. https://doi.org/10.3390/molecules27082411

Neal, R. M. (1996). Introduction. In R. M. Neal (Ed.), Bayesian Learning for Neural Networks (pp. 1–28). Springer. https://doi.org/10.1007/978-1-4612-0745-0_1

Nemkov, T., Yoshida, T., Nikulina, M., & D'Alessandro, A. (2022). High-Throughput Metabolomics Platform for the Rapid Data-Driven Development of Novel Additive Solutions for Blood Storage. Frontiers in Physiology, 13, 833242. https://doi.org/10.3389/fphys.2022.833242

Nguyen, Q.-T., Merlo, M. E., Medema, M. H., Jankevics, A., Breitling, R., & Takano, E. (2012). Metabolomics methods for the synthetic biology of secondary metabolism. FEBS Letters, 586(15), 2177–2183. https://doi.org/10.1016/j.febslet.2012.02.008

Nier, A. O. (1991). The development of a high resolution mass spectrometer: A reminiscence. Journal of the American Society for Mass Spectrometry, 2(6), 447–452. https://doi.org/10.1016/1044-0305(91)80029-7

Nikolova, C., & Gutierrez, T. (2021). Biosurfactants and Their Applications in the Oil and Gas Industry: Current State of Knowledge and Future Perspectives. Frontiers in Bioengineering and Biotechnology, 9. https://www.frontiersin.org/articles/10.3389/fbioe.2021.626639

Nocedal, J., & Wright, S. J. (Eds.). (2006). Quasi-Newton Methods. In Numerical Optimization (pp. 135–163). Springer. https://doi.org/10.1007/978-0-387-40065-5_6

Nor, N. M., Mohamed, M. S., Loh, T. C., Foo, H. L., Rahim, R. A., Tan, J. S., & Mohamad, R. (2017). Comparative analyses on medium optimization using one-factor-at-a-time, response surface methodology, and artificial neural network for lysine–methionine biosynthesis by Pediococcus pentosaceus RF-1. Biotechnology & Biotechnological Equipment, 31(5), 935–947. https://doi.org/10.1080/13102818.2017.1335177

Nowak, J., Kothari, A., Li, H., Pannu, J., Algazi, D., & Prakash, M. (2023, April 20). Inkwell: Design and Validation of a Low-Cost Open Electricity-Free 3D Printed Device for Automated Thin Smearing of Whole Blood. arXiv.Org. https://arxiv.org/abs/2304.10200v1

O'Connor, S. E. (2015). Engineering of Secondary Metabolism. Annual Review of Genetics, 49(Volume 49, 2015), 71–94. https://doi.org/10.1146/annurev-genet-120213-092053

Oellermann, M., Jolles, J. W., Ortiz, D., Seabra, R., Wenzel, T., Wilson, H., & Tanner, R. L. (2022). Open Hardware in Science: The Benefits of Open Electronics. Integrative and Comparative Biology, 62(4), 1061–1075. https://doi.org/10.1093/icb/icac043

O'Gorman, A., & Brennan, L. (2017). The role of metabolomics in determination of new dietary biomarkers. Proceedings of the Nutrition Society, 76(3), 295–302. https://doi.org/10.1017/S0029665116002974

Oliveira, S. M. D., & Densmore, D. (2022). Hardware, Software, and Wetware Codesign Environment for Synthetic Biology. BioDesign Research, 2022, 9794510. https://doi.org/10.34133/2022/9794510

Olsson, L., Rugbjerg, P., Torello Pianale, L., & Trivellin, C. (2022). Robustness: Linking strain design to viable bioprocesses. Trends in Biotechnology, 40(8), 918–931. https://doi.org/10.1016/j.tibtech.2022.01.004

O'Sullivan, A., Gibney, M. J., & Brennan, L. (2011). Dietary intake patterns are reflected in metabolomic profiles: Potential role in dietary assessment studies. The American Journal of Clinical Nutrition, 93(2), 314–321. https://doi.org/10.3945/ajcn.110.000950

Ouyang, W., Bowman, R. W., Wang, H., Bumke, K. E., Collins, J. T., Spjuth, O., Carreras-Puigvert, J., & Diederich, B. (2022). An Open-Source Modular Framework for Automated Pipetting and Imaging Applications. Advanced Biology, 6(4), 2101063. https://doi.org/10.1002/adbi.202101063

Ovbude, S. T., Sharmeen, S., Kyei, I., Olupathage, H., Jones, J., Bell, R. J., Powers, R., & Hage, D. S. (2024). Applications of chromatographic methods in metabolomics: A review. Journal of Chromatography B, 1239, 124124. https://doi.org/10.1016/j.jchromb.2024.124124

Paczkowska, M., Barenboim, J., Sintupisut, N., Fox, N. S., Zhu, H., Abd-Rabbo, D., Mee, M. W., Boutros, P. C., & Reimand, J. (2020). Integrative pathway enrichment analysis of multivariate omics data. Nature Communications, 11(1), 735. https://doi.org/10.1038/s41467-019-13983-9

Paddon, C. J., Westfall, P. J., Pitera, D. J., Benjamin, K., Fisher, K., McPhee, D., Leavell, M. D., Tai, A., Main, A., Eng, D., Polichuk, D. R., Teoh, K. H., Reed, D. W., Treynor, T., Lenihan, J., Jiang, H., Fleck, M., Bajad, S., Dang, G., … Newman, J. D. (2013). High-level semi-synthetic production of the potent antimalarial artemisinin. Nature, 496(7446), 528–532. https://doi.org/10.1038/nature12051

Pandi, A., Diehl, C., Yazdizadeh Kharrazi, A., Scholz, S. A., Bobkova, E., Faure, L., Nattermann, M., Adam, D., Chapin, N., Foroughijabbari, Y., Moritz, C., Paczia, N., Cortina, N. S., Faulon, J.-L., & Erb, T. J. (2022). A versatile active learning workflow for optimization of genetic and metabolic networks. Nature Communications, 13(1), Article 1. https://doi.org/10.1038/s41467-022-31245-z

Pang, H., & Hu, Z. (2023). Metabolomics in drug research and development: The recent advances in technologies and applications. Acta Pharmaceutica Sinica B, 13(8), 3238–3251. https://doi.org/10.1016/j.apsb.2023.05.021

Panyard, D. J., Yu, B., & Snyder, M. P. (2022). The metabolomics of human aging: Advances, challenges, and opportunities. Science Advances, 8(42), eadd6155. https://doi.org/10.1126/sciadv.add6155

Park, J. H., & Lee, S. Y. (2010). Metabolic pathways and fermentative production of L-aspartate family amino acids. Biotechnology Journal, 5(6), 560–577. https://doi.org/10.1002/biot.201000032

Patt, A., Siddiqui, J., Zhang, B., & Mathé, E. (2019). Integration of Metabolomics and Transcriptomics to Identify Gene-Metabolite Relationships Specific to Phenotype. In M. Haznadar (Ed.), Cancer Metabolism: Methods and Protocols (pp. 441–468). Springer. https://doi.org/10.1007/978-1-4939-9027-6_23

Paul, W. (1990). Electromagnetic Traps for Charged and Neutral Particles (Nobel Lecture). Angewandte Chemie International Edition in English, 29(7), 739–748. https://doi.org/10.1002/anie.199007391

Perinelli, D. R., Cespi, M., Lorusso, N., Palmieri, G. F., Bonacucina, G., & Blasi, P. (2020). Surfactant Self-Assembling and Critical Micelle Concentration: One Approach Fits All? Langmuir, 36(21), 5745–5753. https://doi.org/10.1021/acs.langmuir.0c00420

Periyasamy, S., Beula Isabel, J., Kavitha, S., Karthik, V., Mohamed, B. A., Gizaw, D. G., Sivashanmugam, P., & Aminabhavi, T. M. (2023). Recent advances in consolidated bioprocessing for conversion of lignocellulosic biomass into bioethanol – A review. Chemical Engineering Journal, 453, 139783. https://doi.org/10.1016/j.cej.2022.139783

Perumal, S., Atchudan, R., & Lee, W. (2022). A Review of Polymeric Micelles and Their Applications. Polymers, 14(12), Article 12. https://doi.org/10.3390/polym14122510

Petrov, V. V. (1975). Sums of Independent Random Variables. Springer. https://doi.org/10.1007/978-3-642-65809-9

Pitt, J. J. (2009). Principles and Applications of Liquid Chromatography-Mass Spectrometry in Clinical Biochemistry. The Clinical Biochemist Reviews, 30(1), 19–34.

Pleiss, G., Jankowiak, M., Eriksson, D., Damle, A., & Gardner, J. (2020). Fast Matrix Square Roots with Applications to Gaussian Processes and Bayesian Optimization. Advances in Neural Information Processing Systems, 33, 22268–22281. https://proceedings.neurips.cc/paper_files/paper/2020/hash/fcf55a303b71b84d326fb1d06e332a26-Abstract.html

Pomyen, Y., Wanichthanarak, K., Poungsombat, P., Fahrmann, J., Grapov, D., & Khoomrung, S. (2020). Deep metabolome: Applications of deep learning in metabolomics. Computational and Structural Biotechnology Journal, 18, 2818–2825. https://doi.org/10.1016/j.csbj.2020.09.033

Poole, C. F. (2003). Chapter 1—General Concepts in Column Chromatography. In C. F. Poole (Ed.), The Essence of Chromatography (pp. 1–78). Elsevier Science. https://doi.org/10.1016/B978-044450198-1/50014-8

Post, M. J., Levenberg, S., Kaplan, D. L., Genovese, N., Fu, J., Bryant, C. J., Negowetti, N., Verzijden, K., & Moutsatsou, P. (2020). Scientific, sustainability and regulatory challenges of cultured meat. Nature Food, 1(7), 403–415. https://doi.org/10.1038/s43016-020-0112-z

Postelnicu, T. (2011). Probit Analysis. In M. Lovric (Ed.), International Encyclopedia of Statistical Science (pp. 1128–1131). Springer. https://doi.org/10.1007/978-3-642-04898-2_461

Prabhu, G. R. D., Williams, E. R., Wilm, M., & Urban, P. L. (2023). Mass spectrometry using electrospray ionization. Nature Reviews Methods Primers, 3(1), Article 1. https://doi.org/10.1038/s43586-023-00203-4

Price, D. (1993). Time-of-Flight Mass Spectrometry. In Time-of-Flight Mass Spectrometry (Vol. 549, pp. 1–15). American Chemical Society. https://doi.org/10.1021/bk-1994-0549.ch001

Probst, P., Wright, M., & Boulesteix, A.-L. (2019). Hyperparameters and Tuning Strategies for Random Forest. WIREs Data Mining and Knowledge Discovery, 9(3), e1301. https://doi.org/10.1002/widm.1301

Qi, Y., Nepal, K. K., & Blodgett, J. A. V. (2021). A comparative metabologenomic approach reveals mechanistic insights into Streptomyces antibiotic crypticity. Proceedings of the National Academy of Sciences, 118(31), e2103515118. https://doi.org/10.1073/pnas.2103515118

Qiu, S., Cai, Y., Yao, H., Lin, C., Xie, Y., Tang, S., & Zhang, A. (2023). Small molecule metabolites: Discovery of biomarkers and therapeutic targets. Signal Transduction and Targeted Therapy, 8(1), 1–37. https://doi.org/10.1038/s41392-023-01399-3

Qiu, S., Guo, S., Yang, Q., Xie, Y., Tang, S., & Zhang, A. (2023). Innovation in identifying metabolites from complex metabolome—Highlights of recent analytical platforms and protocols. Frontiers in Chemistry, 11, 1129717. https://doi.org/10.3389/fchem.2023.1129717

Radivojević, T., Costello, Z., Workman, K., & Garcia Martin, H. (2020). A machine learning Automated Recommendation Tool for synthetic biology. Nature Communications, 11(1), Article 1. https://doi.org/10.1038/s41467-020-18008-4

Radoš, D., Donati, S., Lempp, M., Rapp, J., & Link, H. (2022). Homeostasis of the biosynthetic E. coli metabolome. iScience, 25(7). https://doi.org/10.1016/j.isci.2022.104503

Ranković, B., Griffiths, R.-R., B. Moss, H., & Schwaller, P. (2024). Bayesian optimisation for additive screening and yield improvements – beyond one-hot encoding. Digital Discovery. https://doi.org/10.1039/D3DD00096F

Razzaq, A., Sadia, B., Raza, A., Khalid Hameed, M., & Saleem, F. (2019). Metabolomics: A Way Forward for Crop Improvement. Metabolites, 9(12), Article 12. https://doi.org/10.3390/metabo9120303

Rebane, R., Kruve, A., Liigand, P., Liigand, J., Herodes, K., & Leito, I. (2016). Establishing Atmospheric Pressure Chemical Ionization Efficiency Scale. Analytical Chemistry, 88(7), 3435–3439. https://doi.org/10.1021/acs.analchem.5b04852

Reel, P. S., Reel, S., Pearson, E., Trucco, E., & Jefferson, E. (2021). Using machine learning approaches for multi-omics data analysis: A review. Biotechnology Advances, 49, 107739. https://doi.org/10.1016/j.biotechadv.2021.107739

Refaeilzadeh, P., Tang, L., & Liu, H. (2009). Cross-Validation. In L. LIU & M. T. ÖZSU (Eds.), Encyclopedia of Database Systems (pp. 532–538). Springer US. https://doi.org/10.1007/978-0-387-39940-9_565

Reiter, T., Brooks†, P. T., Irber†, L., Joslin†, S. E. K., Reid†, C. M., Scott†, C., Brown, C. T., & Pierce-Ward, N. T. (2021). Streamlining data-intensive biology with workflow systems. GigaScience, 10(1), giaa140. https://doi.org/10.1093/gigascience/giaa140

Ribbenstedt, A., Ziarrusta, H., & Benskin, J. P. (2018). Development, characterization and comparisons of targeted and non-targeted metabolomics methods. PLOS ONE, 13(11), e0207082. https://doi.org/10.1371/journal.pone.0207082

Ricroch, A. E., Bergé, J. B., & Kuntz, M. (2011). Evaluation of Genetically Engineered Crops Using Transcriptomic, Proteomic, and Metabolomic Profiling Techniques. Plant Physiology, 155(4), 1752–1761. https://doi.org/10.1104/pp.111.173609

Rischer, H., & Oksman-Caldentey, K.-M. (2006). Unintended effects in genetically modified crops: Revealed by metabolomics? Trends in Biotechnology, 24(3), 102–104. https://doi.org/10.1016/j.tibtech.2006.01.009

Riter, L. S., Vitek, O., Gooding, K. M., Hodge, B. D., & Julian, R. K. (2005). Statistical design of experiments as a tool in mass spectrometry. Journal of Mass Spectrometry: JMS, 40(5), 565–579. https://doi.org/10.1002/jms.871

Roberts, L. D., Souza, A. L., Gerszten, R. E., & Clish, C. B. (2012). Targeted Metabolomics. Current Protocols in Molecular Biology, 98(1), 30.2.1-30.2.24. https://doi.org/10.1002/0471142727.mb3002s98

Romero, H., Pott, D. M., Vallarino, J. G., & Osorio, S. (2021). Metabolomics-Based Evaluation of Crop Quality Changes as a Consequence of Climate Change. Metabolites, 11(7), Article 7. https://doi.org/10.3390/metabo11070461

Rosenboom, J.-G., Langer, R., & Traverso, G. (2022). Bioplastics for a circular economy. Nature Reviews Materials, 7(2), 117–137. https://doi.org/10.1038/s41578-021-00407-8

Röst, H. L., Sachsenberg, T., Aiche, S., Bielow, C., Weisser, H., Aicheler, F., Andreotti, S., Ehrlich, H.-C., Gutenbrunner, P., Kenar, E., Liang, X., Nahnsen, S., Nilse, L., Pfeuffer, J., Rosenberger, G., Rurik, M., Schmitt, U., Veit, J., Walzer, M., … Kohlbacher, O. (2016). OpenMS: A flexible open-source software platform for mass spectrometry data analysis. Nature Methods, 13(9), 741–748. https://doi.org/10.1038/nmeth.3959

Rubio, N. R., Xiang, N., & Kaplan, D. L. (2020). Plant-based and cell-based approaches to meat production. Nature Communications, 11(1), 6276. https://doi.org/10.1038/s41467-020-20061-y

Rugbjerg, P., & Sommer, M. O. A. (2019). Overcoming genetic heterogeneity in industrial fermentations. Nature Biotechnology, 37(8), 869–876. https://doi.org/10.1038/s41587-019-0171-6

Ruiz-Perez, D., Guan, H., Madhivanan, P., Mathee, K., & Narasimhan, G. (2020). So you think you can PLS-DA? BMC Bioinformatics, 21(1), 2. https://doi.org/10.1186/s12859-019-3310-7

Russo, D. J., Van Roy, B., Kazerouni, A., Osband, I., & Wen, Z. (2018). A Tutorial on Thompson Sampling. Found. Trends Mach. Learn., 11(1), 1–96. https://doi.org/10.1561/2200000070

Russo, D., & Roy, B. V. (2016). An Information-Theoretic Analysis of Thompson Sampling. Journal of Machine Learning Research, 17(68), 1–30.

Santacruz, D., Enane, F. O., Fundel-Clemens, K., Giner, M., Wolf, G., Onstein, S., Klimek, C., Smith, Z., Wijayawardena, B., & Viollet, C. (2022). Automation of high-throughput mRNA-seq library preparation: A robust, hands-free and time efficient methodology. SLAS Discovery, 27(2), 140–147. https://doi.org/10.1016/j.slasd.2022.01.002

Santner, T. J., Williams, B. J., & Notz, W. I. (2018). Space-Filling Designs for Computer Experiments. In T. J. Santner, B. J. Williams, & W. I. Notz (Eds.), The Design and Analysis of Computer Experiments (pp. 145–200). Springer. https://doi.org/10.1007/978-1-4939-8847-1_5

Sarvin, B., Lagziel, S., Sarvin, N., Mukha, D., Kumar, P., Aizenshtein, E., & Shlomi, T. (2020). Fast and sensitive flow-injection mass spectrometry metabolomics by analyzing sample-specific ion distributions. Nature Communications, 11(1), Article 1. https://doi.org/10.1038/s41467-020-17026-6

Schiffman, C., Petrick, L., Perttula, K., Yano, Y., Carlsson, H., Whitehead, T., Metayer, C., Hayes, J., Rappaport, S., & Dudoit, S. (2019). Filtering procedures for untargeted LC-

MS metabolomics data. BMC Bioinformatics, 20(1), 334. https://doi.org/10.1186/s12859-019-2871-9

Schorn, M. A., Verhoeven, S., Ridder, L., Huber, F., Acharya, D. D., Aksenov, A. A., Aleti, G., Moghaddam, J. A., Aron, A. T., Aziz, S., Bauermeister, A., Bauman, K. D., Baunach, M., Beemelmanns, C., Beman, J. M., Berlanga-Clavero, M. V., Blacutt, A. A., Bode, H. B., Boullie, A., … van der Hooft, J. J. J. (2021). A community resource for paired genomic and metabolomic data mining. Nature Chemical Biology, 17(4), 363–368. https://doi.org/10.1038/s41589-020-00724-z

Schultz, D., Wolynes, P. G., Jacob, E. B., & Onuchic, J. N. (2009). Deciding fate in adverse times: Sporulation and competence in Bacillus subtilis. Proceedings of the National Academy of Sciences, 106(50), 21027–21034. https://doi.org/10.1073/pnas.0912185106

Scipy. (n.d.). Quasi-Monte Carlo submodule (scipy.stats.qmc)—SciPy v1.14.1 Manual. Retrieved 23 September 2024, from https://docs.scipy.org/doc/scipy/reference/stats.qmc.html

Salzberg, L. I., Botella, E., Hokamp, K., Antelmann, H., Maaß, S., Becher, D., Noone, D., & Devine, K. M. (2015). Genome-Wide Analysis of Phosphorylated PhoP Binding to Chromosomal DNA Reveals Several Novel Features of the PhoPR-Mediated Phosphate Limitation Response in Bacillus subtilis. Journal of Bacteriology, 197(8), 1492–1506. https://doi.org/10.1128/jb.02570-14

Sen, P., Lamichhane, S., Mathema, V. B., McGlinchey, A., Dickens, A. M., Khoomrung, S., & Orešič, M. (2021). Deep learning meets metabolomics: A methodological perspective. Briefings in Bioinformatics, 22(2), 1531–1542. https://doi.org/10.1093/bib/bbaa204

Sen, P., & Orešič, M. (2023). Integrating Omics Data in Genome-Scale Metabolic Modeling: A Methodological Perspective for Precision Medicine. Metabolites, 13(7), Article 7. https://doi.org/10.3390/metabo13070855

Ser, Z., Liu, X., Tang, N. N., & Locasale, J. W. (2015). Extraction parameters for metabolomics from cultured cells. Analytical Biochemistry, 475, 22–28. https://doi.org/10.1016/j.ab.2015.01.003

Serbanescu, D., Ojkic, N., & Banerjee, S. (2020). Nutrient-Dependent Trade-Offs between Ribosomes and Division Protein Synthesis Control Bacterial Cell Size and Growth. Cell Reports, 32(12). https://doi.org/10.1016/j.celrep.2020.108183

Serror, P., & Sonenshein, A. L. (1996). CodY is required for nutritional repression of Bacillus subtilis genetic competence. Journal of Bacteriology, 178(20), 5910–5915. https://doi.org/10.1128/jb.178.20.5910-5915.1996

Sévin, D. C., Kuehne, A., Zamboni, N., & Sauer, U. (2015). Biological insights through nontargeted metabolomics. Current Opinion in Biotechnology, 34, 1–8. https://doi.org/10.1016/j.copbio.2014.10.001

Shaban, S. M., Kang, J., & Kim, D.-H. (2020). Surfactants: Recent advances and their applications. Composites Communications, 22, 100537. https://doi.org/10.1016/j.coco.2020.100537

Shahriari, B., Swersky, K., Wang, Z., Adams, R. P., & Freitas, N. de. (2016). Taking the Human Out of the Loop: A Review of Bayesian Optimization. Proceedings of the IEEE, 104(1), 148–175. Proceedings of the IEEE. https://doi.org/10.1109/JPROC.2015.2494218

Shaligram, N. S., & Singhal, R. S. (2010). Surfactin – A Review on Biosynthesis, Fermentation, Purification and Applications. Food Technology and Biotechnology, 48(2), 119–134.

Shao, Y., & Le, W. (2019). Recent advances and perspectives of metabolomics-based investigations in Parkinson's disease. Molecular Neurodegeneration, 14(1), 3. https://doi.org/10.1186/s13024-018-0304-2

Shin, J. H., & Choi, S. (2021). Open-source and do-it-yourself microfluidics. Sensors and Actuators B: Chemical, 347, 130624. https://doi.org/10.1016/j.snb.2021.130624

Siegel, D., Permentier, H., Reijngoud, D.-J., & Bischoff, R. (2014). Chemical and technical challenges in the analysis of central carbon metabolites by liquid-chromatography mass spectrometry. Journal of Chromatography B, 966, 21–33. https://doi.org/10.1016/j.jchromb.2013.11.022

Sieniawska, E., & Georgiev, M. I. (2022). Metabolomics: Towards acceleration of antibacterial plant-based leads discovery. Phytochemistry Reviews, 21(3), 765–781. https://doi.org/10.1007/s11101-021-09762-4

Sinha, A., & Mann, M. (2020). A beginner's guide to mass spectrometry–based proteomics. The Biochemist, 42(5), 64–69. https://doi.org/10.1042/BIO20200057

Siuzdak, G. (2004). An Introduction to Mass Spectrometry Ionization: An Excerpt from The Expanding Role of Mass Spectrometry in Biotechnology, 2nd ed.; MCC Press: San Diego, 2005. JALA: Journal of the Association for Laboratory Automation, 9(2), 50–63. https://doi.org/10.1016/j.jala.2004.01.004

Smoluch, M., & Piechura, K. (2019). Basic Definitions. In Mass Spectrometry (pp. 9–12). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781119377368.ch3

Snoek, J., Larochelle, H., & Adams, R. P. (2012). Practical Bayesian Optimization of Machine Learning Algorithms. arXiv:1206.2944 [Cs, Stat]. http://arxiv.org/abs/1206.2944

Soberón-Chávez, G., & Maier, R. M. (2011). Biosurfactants: A General Overview. In G. Soberón-Chávez (Ed.), Biosurfactants: From Genes to Applications (pp. 1–11). Springer. https://doi.org/10.1007/978-3-642-14490-5_1

Sobol', I. M. (1967). On the distribution of points in a cube and the approximate evaluation of integrals. USSR Computational Mathematics and Mathematical Physics, 7(4), 86–112. https://doi.org/10.1016/0041-5553(67)90144-9

Socea, J. N., Stone, V. N., Qian, X., Gibbs, P. L., & Levinson, K. J. (2023). Implementing laboratory automation for next-generation sequencing: Benefits and challenges for library preparation. Frontiers in Public Health, 11. https://doi.org/10.3389/fpubh.2023.1195581

Soler, C., Hamilton, B., Furey, A., James, K. J., Mañes, J., & Picó, Y. (2006). Optimization of LC–MS/MS using triple quadrupole mass analyzer for the simultaneous analysis of carbosulfan and its main metabolites in oranges. Analytica Chimica Acta, 571(1), 1–11. https://doi.org/10.1016/j.aca.2006.04.033

Sonenshein, A. L. (2007). Control of key metabolic intersections in Bacillus subtilis. Nature Reviews Microbiology, 5(12), 917–927. https://doi.org/10.1038/nrmicro1772

Song, X., Zhang, Q., Lee, C., Fertig, E., Huang, T.-K., Belenki, L., Kochanski, G., Ariafar, S., Vasudevan, S., Perel, S., & Golovin, D. (2024). The Vizier Gaussian Process Bandit Algorithm (arXiv:2408.11527). arXiv. https://doi.org/10.48550/arXiv.2408.11527

Song, Y., Yao, S., Li, X., Wang, T., Jiang, X., Bolan, N., Warren, C. R., Northen, T. R., & Chang, S. X. (2024). Soil metabolomics: Deciphering underground metabolic webs in terrestrial ecosystems. Eco-Environment & Health, 3(2), 227–237. https://doi.org/10.1016/j.eehl.2024.03.001

Sorrentino, S., Manetti, F., Bresci, A., Vernuccio, F., Ceconello, C., Ghislanzoni, S., Bongarzone, I., Vanna, R., Cerullo, G., & Polli, D. (2023). Deep ensemble learning and transfer learning methods for classification of senescent cells from nonlinear optical microscopy images. Frontiers in Chemistry, 11. https://doi.org/10.3389/fchem.2023.1213981

Souza, A. L., & Patti, G. J. (2021). A Protocol for Untargeted Metabolomic Analysis: From Sample Preparation to Data Processing. In V. Weissig & M. Edeas (Eds.), Mitochondrial Medicine: Volume 2: Assessing Mitochondria (pp. 357–382). Springer US. https://doi.org/10.1007/978-1-0716-1266-8_27

Spotify/luigi. (2024). [Python]. Spotify. https://github.com/spotify/luigi (Original work published 2012)

Stafford, G. (2002). Ion trap mass spectrometry: A personal perspective. Journal of the American Society for Mass Spectrometry, 13(6), 589–596. https://doi.org/10.1016/S1044-0305(02)00385-9

Stavarache, C., Nicolescu, A., Duduianu, C., Ailiesei, G. L., Balan-Porcăraşu, M., Cristea, M., Macsim, A.-M., Popa, O., Stavarache, C., Hîrtopeanu, A., Barbeş, L., Stan, R., Iovu, H., & Deleanu, C. (2022). A Real-Life Reproducibility Assessment for NMR Metabolomics. Diagnostics, 12(3), 559. https://doi.org/10.3390/diagnostics12030559

Stein, M. L. (2012). Interpolation of Spatial Data: Some Theory for Kriging. Springer Science & Business Media.

Storch, M., Haines, M. C., & Baldwin, G. S. (2020). DNA-BOT: A low-cost, automated DNA assembly platform for synthetic biology. Synthetic Biology, 5(1), ysaa010. https://doi.org/10.1093/synbio/ysaa010

Subramanian, I., Verma, S., Kumar, S., Jere, A., & Anamika, K. (2020). Multi-omics Data Integration, Interpretation, and Its Application. Bioinformatics and Biology Insights, 14, 1177932219899051. https://doi.org/10.1177/1177932219899051

Sun, J., & Xia, Y. (2024). Pretreating and normalizing metabolomics data for statistical analysis. Genes & Diseases, 11(3), 100979. https://doi.org/10.1016/j.gendis.2023.04.018

Surfactants Market Size, Industry Share Growth Forecast, Global Trends Report, [Latest]. (n.d.). MarketsandMarkets. Retrieved 29 August 2024, from https://www.marketsandmarkets.com/Market-Reports/biosurfactants-market-493.html

Sutton, M. (2022). The discovery of mass spectrometry. Chemistry World. https://www.chemistryworld.com/features/the-discovery-of-mass-spectrometry/4016197.article

Szegö, G. P. (1975). Towards Global Optimisation: Proceedings of a Workshop at the University of Cagliari, Italy, October 1974. North-Holland Publishing Company.

Szkodny, A. C., & Lee, K. H. (2022). Biopharmaceutical Manufacturing: Historical Perspectives and Future Directions. Annual Review of Chemical and Biomolecular Engineering, 13(Volume 13, 2022), 141–165. https://doi.org/10.1146/annurev-chembioeng-092220-125832

Tanaka, K., Bamba, T., Kondo, A., & Hasunuma, T. (2024a). Metabolomics-based development of bioproduction processes toward industrial-scale production. Current Opinion in Biotechnology, 85, 103057. https://doi.org/10.1016/j.copbio.2023.103057

Tanaka, K., Bamba, T., Kondo, A., & Hasunuma, T. (2024b). Metabolomics-based development of bioproduction processes toward industrial-scale production. Current Opinion in Biotechnology, 85, 103057. https://doi.org/10.1016/j.copbio.2023.103057

Tang, J. (2011). Microbial Metabolomics. Current Genomics, 12(6), 391–403. https://doi.org/10.2174/138920211797248619

Tegally, H., San, J. E., Giandhari, J., & de Oliveira, T. (2020). Unlocking the efficiency of genomics laboratories with robotic liquid-handling. BMC Genomics, 21(1), 729. https://doi.org/10.1186/s12864-020-07137-1

Tembhare, K., Sharma, T., Kasibhatla, S. M., Achalere, A., & Joshi, R. (2024). Multi-ensemble machine learning framework for omics data integration: A case study using breast cancer samples. Informatics in Medicine Unlocked, 47, 101507. https://doi.org/10.1016/j.imu.2024.101507

Teng, Q., Huang, W., Collette, T. W., Ekman, D. R., & Tan, C. (2009). A direct cell quenching method for cell-culture based metabolomics. Metabolomics, 5(2), 199–208. https://doi.org/10.1007/s11306-008-0137-z

The Economic Cell Collective. (2024). Economic Principles in Cell Biology. No commercial publisher | Online open access book. https://doi.org/10.5281/ZENODO.12592398

Théatre, A., Cano-Prieto, C., Bartolini, M., Laurin, Y., Deleu, M., Niehren, J., Fida, T., Gerbinet, S., Alanjary, M., Medema, M. H., Léonard, A., Lins, L., Arabolaza, A., Gramajo, H., Gross, H., & Jacques, P. (2021). The Surfactin-Like Lipopeptides From Bacillus spp.: Natural Biodiversity and Synthetic Biology for a Broader Application Range. Frontiers in Bioengineering and Biotechnology, 9. https://doi.org/10.3389/fbioe.2021.623701

Thingstad, T. F. (2022). Competition–defense trade-offs in the microbial world. Proceedings of the National Academy of Sciences, 119(37), e2213092119. https://doi.org/10.1073/pnas.2213092119

Thomas, S. N., French, D., Jannetto, P. J., Rappold, B. A., & Clarke, W. A. (2022). Liquid chromatography–tandem mass spectrometry for clinical diagnostics. Nature Reviews Methods Primers, 2(1), 1–14. https://doi.org/10.1038/s43586-022-00175-x

Thompson, W. R. (1933). On the Likelihood that One Unknown Probability Exceeds Another in View of the Evidence of Two Samples. Biometrika, 25(3/4), 285–294. https://doi.org/10.2307/2332286

Thukral, M., Allen, A. E., & Petras, D. (2023). Progress and challenges in exploring aquatic microbial communities using non-targeted metabolomics. The ISME Journal, 17(12), 2147–2159. https://doi.org/10.1038/s41396-023-01532-8

Toms, D., Deardon, R., & Ungrin, M. (2017). Climbing the mountain: Experimental design for the efficient optimization of stem cell bioprocessing. Journal of Biological Engineering, 11(1), 35. https://doi.org/10.1186/s13036-017-0078-z

Twigg, M. S., Baccile, N., Banat, I. M., Déziel, E., Marchant, R., Roelants, S., & Van Bogaert, I. N. A. (2021). Microbial biosurfactant research: Time to improve the rigour in the reporting of synthesis, functional characterization and process development. Microbial Biotechnology, 14(1), 147–170. https://doi.org/10.1111/1751-7915.13704

Unal, D. N., Sadak, S., Erkmen, C., Selcuk, Ö., & Uslu, B. (n.d.). Review of Surfactants, Structural Properties and Their Role in Electrochemistry. Retrieved 22 August 2024, from https://books.rsc.org/books/edited-volume/2121/chapter/7715733/Review-of-Surfactants-Structural-Properties-and

Vaidyanathan, S., Kell, D. B., & Goodacre, R. (2002). Flow-injection electrospray ionization mass spectrometry of crude cell extracts for high-throughput bacterial identification. Journal of the American Society for Mass Spectrometry, 13(2), 118–128. https://doi.org/10.1016/S1044-0305(01)00339-7

van der Greef, J., & Smilde, A. K. (2005). Symbiosis of chemometrics and metabolomics: Past, present, and future. Journal of Chemometrics, 19(5–7), 376–386. https://doi.org/10.1002/cem.941

van der Velden, P. M. M., & Jansen, R. S. (2023). Microbial Metabolomics: An Overview of Applications. In V. Soni & T. E. Hartman (Eds.), Metabolomics: Recent Advances and Future Applications (pp. 165–208). Springer International Publishing. https://doi.org/10.1007/978-3-031-39094-4_6

Van Puyvelde, B., Hunter, C. L., Zhgamadze, M., Savant, S., Wang, Y. O., Hoedt, E., Raedschelders, K., Pope, M., Huynh, C. A., Ramanujan, V. K., Tourtellotte, W., Razavi, M., Anderson, N. L., Martens, G., Deforce, D., Fu, Q., Dhaenens, M., & Van Eyk, J. E. (2024). Acoustic ejection mass spectrometry empowers ultra-fast protein biomarker quantification. Nature Communications, 15(1), 5114. https://doi.org/10.1038/s41467-024-48563-z

Vavricka, C. J., Hasunuma, T., & Kondo, A. (2020). Dynamic Metabolomics for Engineering Biology: Accelerating Learning Cycles for Bioproduction. Trends in Biotechnology, 38(1), 68–82. https://doi.org/10.1016/j.tibtech.2019.07.009

Vieira, I. M. M., Santos, B. L. P., Ruzene, D. S., & Silva, D. P. (2021). An overview of current research and developments in biosurfactants. Journal of Industrial and Engineering Chemistry, 100, 1–18. https://doi.org/10.1016/j.jiec.2021.05.017

Villate, A., San Nicolas, M., Gallastegi, M., Aulas, P.-A., Olivares, M., Usobiaga, A., Etxebarria, N., & Aizpurua-Olaizola, O. (2021). Review: Metabolomics as a prediction tool for plants performance under environmental stress. Plant Science: An International Journal of Experimental Plant Biology, 303, 110789. https://doi.org/10.1016/j.plantsci.2020.110789

Volmer, D., & Jessome, L. L. (2006). Ion Suppression: A Major Concern in Mass Spectrometry. 24, 498–510.

Volpp, M., Fröhlich, L. P., Fischer, K., Doerr, A., Falkner, S., Hutter, F., & Daniel, C. (2019, September 25). Meta-Learning Acquisition Functions for Transfer Learning in Bayesian Optimization. International Conference on Learning Representations. https://openreview.net/forum?id=ryeYpJSKwr

Vuckovic, D. (2020). Chapter 4—Sample preparation in global metabolomics of biological fluids and tissues. In H. J. Issaq & T. D. Veenstra (Eds.), Proteomic and Metabolomic Approaches to Biomarker Discovery (Second Edition) (pp. 53–83). Academic Press. https://doi.org/10.1016/B978-0-12-818607-7.00004-9

Waller, D., Putnam, J., Steiner, J. N., Fisher, B., Burcham, G. N., Oliver, J., Smith, S. B., Erickson, R., Remek, A., & Bodoeker, N. (2023). Targeted metabolomics characterizes

metabolite occurrence and variability in stable freshwater mussel populations. Conservation Physiology, 11(1), coad040. https://doi.org/10.1093/conphys/coad040

Wang, K., & Dowling, A. W. (2022). Bayesian optimization for chemical products and functional materials. Current Opinion in Chemical Engineering, 36, 100728.

Wang, R. (Ed.). (2012). Karhunen-Loève transform and principal component analysis. In Introduction to Orthogonal Transforms: With Applications in Data Processing and Analysis (pp. 412–460). Cambridge University Press. https://doi.org/10.1017/CBO9781139015158.011

Wang, W., Rong, Z., Wang, G., Hou, Y., Yang, F., & Qiu, M. (2023). Cancer metabolites: Promising biomarkers for cancer liquid biopsy. Biomarker Research, 11(1), 66. https://doi.org/10.1186/s40364-023-00507-3

Wang, Y., Chen, T.-Y., & Vlachos, D. G. (2021). NEXTorch: A Design and Bayesian Optimization Toolkit for Chemical Sciences and Engineering. Journal of Chemical Information and Modeling, 61(11), 5312–5319. https://doi.org/10.1021/acs.jcim.1c00637

Wei, Y.-H., Wang, L.-F., & Chang, J.-S. (2004). Optimizing Iron Supplement Strategies for Enhanced Surfactin Production with Bacillus subtilis. Biotechnology Progress, 20(3), 979–983. https://doi.org/10.1021/bp030051a

Weissman, K. J. (2015). The structural biology of biosynthetic megaenzymes. Nature Chemical Biology, 11(9), 660–670. https://doi.org/10.1038/nchembio.1883

Wenzel, T. (2023). Open hardware: From DIY trend to global transformation in access to laboratory equipment. PLOS Biology, 21(1), e3001931. https://doi.org/10.1371/journal.pbio.3001931

Wieder, C., Bundy, J. G., Frainay, C., Poupin, N., Rodríguez-Mier, P., Vinson, F., Cooke, J., Lai, R. P. J., Jourdan, F., & Ebbels, T. M. D. (2022). Avoiding the Misuse of Pathway Analysis Tools in Environmental Metabolomics. Environmental Science & Technology, 56(20), 14219–14222. https://doi.org/10.1021/acs.est.2c05588

Wieder, C., Frainay, C., Poupin, N., Rodríguez-Mier, P., Vinson, F., Cooke, J., Lai, R. P., Bundy, J. G., Jourdan, F., & Ebbels, T. (2021). Pathway analysis in metabolomics:

Recommendations for the use of over-representation analysis. PLoS Computational Biology, 17(9), e1009105. https://doi.org/10.1371/journal.pcbi.1009105

Wierenga, R. P., Golas, S. M., Ho, W., Coley, C. W., & Esvelt, K. M. (2023). PyLabRobot: An open-source, hardware-agnostic interface for liquid-handling robots and accessories. Device, 1(4), 100111. https://doi.org/10.1016/j.device.2023.100111

William Allwood, J., Winder, C. L., Dunn, W. B., & Goodacre, R. (2013). Considerations in Sample Preparation, Collection, and Extraction Approaches Applied in Microbial, Plant, and Mammalian Metabolic Profiling. In J. V. Sweedler, N. W. Lutz, & R. A. Wevers (Eds.), Methodologies for Metabolomics: Experimental Strategies and Techniques (pp. 79–118). Cambridge University Press. https://doi.org/10.1017/CBO9780511996634.006

Williams, J. D., Pu, F., Sawicki, J. W., & Elsen, N. L. (2024). Ultra-high-throughput mass spectrometry in drug discovery: Fundamentals and recent advances. Expert Opinion on Drug Discovery, 19(3), 291–301. https://doi.org/10.1080/17460441.2023.2293153

Williams, P. J. H., Chagunda, I. C., & McIndoe, J. S. (2024). OptiMS: An Accessible Program for Automating Mass Spectrometry Parameter Optimization and Configuration. Journal of the American Society for Mass Spectrometry, 35(3), 449–455. https://doi.org/10.1021/jasms.3c00354

Wilm, M. S., & Mann, M. (1994). Electrospray and Taylor-Cone theory, Dole's beam of macromolecules at last? International Journal of Mass Spectrometry and Ion Processes, 136(2), 167–180. https://doi.org/10.1016/0168-1176(94)04024-9

Wilson, J. T. (2024). Stopping Bayesian Optimization with Probabilistic Regret Bounds (arXiv:2402.16811). arXiv. https://doi.org/10.48550/arXiv.2402.16811

Wilson, J. T., Hutter, F., & Deisenroth, M. P. (2018). Maximizing acquisition functions for Bayesian optimization. Proceedings of the 32nd International Conference on Neural Information Processing Systems, 9906–9917.

Wipf, D., & Nagarajan, S. (2007). A New View of Automatic Relevance Determination. Advances in Neural Information Processing Systems, 20.

https://papers.nips.cc/paper_files/paper/2007/hash/9c01802ddb981e6bcfbec0f0516b8e35-Abstract.html

Wishart, D. S. (2013). Exploring the Human Metabolome by Nuclear Magnetic Resonance Spectroscopy and Mass Spectrometry. In J. V. Sweedler, N. W. Lutz, & R. A. Wevers (Eds.), Methodologies for Metabolomics: Experimental Strategies and Techniques (pp. 3–29). Cambridge University Press. https://doi.org/10.1017/CBO9780511996634.002

Wishart, D. S. (2016). Emerging applications of metabolomics in drug discovery and precision medicine. Nature Reviews Drug Discovery, 15(7), 473–484. https://doi.org/10.1038/nrd.2016.32

Wishart, D. S. (2019). Metabolomics for Investigating Physiological and Pathophysiological Processes. Physiological Reviews, 99(4), 1819–1875. https://doi.org/10.1152/physrev.00035.2018

Worley, B., & Powers, R. (2013). Multivariate Analysis in Metabolomics. Current Metabolomics, 1(1), 92–107. https://doi.org/10.2174/2213235X11301010092

Wortel, M. T., Noor, E., Ferris, M., Bruggeman, F. J., & Liebermeister, W. (2018). Metabolic enzyme cost explains variable trade-offs between microbial growth rate and yield. PLOS Computational Biology, 14(2), e1006010. https://doi.org/10.1371/journal.pcbi.1006010

Wu, Y., & Li, L. (2016). Sample normalization methods in quantitative metabolomics. Journal of Chromatography A, 1430, 80–95. https://doi.org/10.1016/j.chroma.2015.12.007

Xia, L., & Wen, J. (2022). Available strategies for improving the biosynthesis of surfactin: A review. Critical Reviews in Biotechnology, 0(0), 1–18. https://doi.org/10.1080/07388551.2022.2095252

Xie, L., Gao, J., & Zhou, Y. J. (2024). Synthetic biology for Taxol biosynthesis and sustainable production. Trends in Biotechnology, 42(6), 674–676. https://doi.org/10.1016/j.tibtech.2024.04.001

Yoneda, T., Miyota, Y., Furuya, K., & Tsuzuki, T. (2006). Production process of surfactin (United States Patent No. US7011969B2). https://patents.google.com/patent/US7011969B2/en

Yoshida, K., Watanabe, K., Chiou, T.-Y., & Konishi, M. (2023). High throughput optimization of medium composition for Escherichia coli protein expression using deep learning and Bayesian optimization. Journal of Bioscience and Bioengineering, 135(2). https://doi.org/10.1016/j.jbiosc.2022.12.004

Yost, R. A. (2022). The triple quadrupole: Innovation, serendipity and persistence. Journal of Mass Spectrometry and Advances in the Clinical Lab, 24, 90–99. https://doi.org/10.1016/j.jmsacl.2022.05.001

Yost, R. A., & Enke, C. G. (1978). Selected ion fragmentation with a tandem quadrupole mass spectrometer. Journal of the American Chemical Society, 100(7), 2274–2275. https://doi.org/10.1021/ja00475a072

Zampieri, G., Campanaro, S., Angione, C., & Treu, L. (2023). Metatranscriptomics-guided genome-scale metabolic modeling of microbial communities. Cell Reports Methods, 3(1), 100383. https://doi.org/10.1016/j.crmeth.2022.100383

Zeng, C., Wen, B., Hou, G., Lei, L., Mei, Z., Jia, X., Chen, X., Zhu, W., Li, J., Kuang, Y., Zeng, W., Su, J., Liu, S., Peng, C., & Chen, X. (2017). Lipidomics profiling reveals the role of glycerophospholipid metabolism in psoriasis. GigaScience, 6(10), 1–11. https://doi.org/10.1093/gigascience/gix087

Zhang, A., Sun, H., Yan, G., Wang, P., & Wang, X. (2015). Metabolomics for Biomarker Discovery: Moving to the Clinic. BioMed Research International, 2015(1), 354671. https://doi.org/10.1155/2015/354671

Zhang, J. D., Xue, C., Kolachalama, V. B., & Donald, W. A. (2023). Interpretable Machine Learning on Metabolomics Data Reveals Biomarkers for Parkinson's Disease. ACS Central Science, 9(5), 1035–1045. https://doi.org/10.1021/acscentsci.2c01468

Zhang, J., Petersen, S. D., Radivojevic, T., Ramirez, A., Pérez-Manríquez, A., Abeliuk, E., Sánchez, B. J., Costello, Z., Chen, Y., Fero, M. J., Martin, H. G., Nielsen, J., Keasling, J. D., & Jensen, M. K. (2020). Combining mechanistic and machine learning models for

predictive engineering and optimization of tryptophan metabolism. Nature Communications, 11(1), Article 1. https://doi.org/10.1038/s41467-020-17910-1

Zhou, J., & Yin, Y. (2016). Strategies for large-scale targeted metabolomics quantification by liquid chromatography-mass spectrometry. Analyst, 141(23), 6362–6373. https://doi.org/10.1039/C6AN01753C

Zhou, T., Reji, R., Kairon, R. S., & Chiam, K. H. (2023). A review of algorithmic approaches for cell culture media optimization. Frontiers in Bioengineering and Biotechnology, 11. https://doi.org/10.3389/fbioe.2023.1195294

Zhou, Z., Luo, M., Zhang, H., Yin, Y., Cai, Y., & Zhu, Z.-J. (2022). Metabolite annotation from knowns to unknowns through knowledge-guided multi-layer metabolic networking. Nature Communications, 13(1), 6656. https://doi.org/10.1038/s41467-022-34537-6

Zhu, M., Wang, Q., Mu, H., Han, F., Wang, Y., & Dai, X. (2023). A fitness trade-off between growth and survival governed by Spo0A-mediated proteome allocation constraints in Bacillus subtilis. Science Advances, 9(39), eadg9733. https://doi.org/10.1126/sciadv.adg9733

Zouari, R., Ellouze-Chaabouni, S., & Ghribi-Aydi, D. (2014). Optimization of Bacillus subtilis SPB1 Biosurfactant Production Under Solid-state Fermentation Using By-products of a Traditional Olive Mill Factory. Achievements in the Life Sciences, 8(2), 162–169. https://doi.org/10.1016/j.als.2015.04.007

Zubarev, R. A., & Makarov, A. (2013). Orbitrap Mass Spectrometry. Analytical Chemistry, 85(11), 5288–5296. https://doi.org/10.1021/ac4001223

Zumpano, R., Del Giudice, A., Resta, S., D'Annibale, A., Sciubba, F., Mura, F., Parisi, G., di Gregorio, M. C., & Galantini, L. (2024). Sodium lauryl ether sulfates, pivotal surfactants for formulations: Rationalization of their assembly properties. Colloids and Surfaces A: Physicochemical and Engineering Aspects, 686, 133375. https://doi.org/10.1016/j.colsurfa.2024.133375

Appendix A

A.1. Thin layer chromatography

For the thin layer chromatography protocol, plastic plates coated with silica were used as stationary phase. The 20 x 20 cm plates are cut in strips for manipulation and allow for testing of more solvent combinations. The strips are carefully marked around 1 cm from the bottom with a graphite pencil. This mark corresponds to the starting point of the sample. Caution should be exercised to keep the integrity of the silica coat. Then, a small amount of sample (~1ul), supernatant in our case, is deposited in the marked site using a capillary tube. The drop is allowed to dry. In parallel, a glass chromatography chamber is prepared by transferring a solvent mix to test until the level is just below 1 cm, corresponding to plate mark. The chamber is closed to allow the solvent vapours to equilibrate inside the chamber atmosphere. When ready, the chamber is quickly open, the plates inserted carefully with tweezers inside the chamber and closed again for the run. The solvent ascends the silica coat by capillary action, dragging the sample and performing the separation into compounds of different polarities present in the sample. After the solvent has run for a few minutes and is in level below the end of the plate, the plate is carefully removed, and it is left to dry. For the staining solutions, ninhydrin and rhodamine 6G are prepared. The ninhydrin solution is transferred to a clean glass sprayer and a homogeneous layer is sprayed over the plate. The plate is left to dry and then moved into a hot plate or oven to reveal a yellowish spot on compounds with amino acid residues. For rhodamine 6G, a similar protocol is used, but heating is not necessary, revealing pinkish spots over a general class of compounds. Finally, for iodine staining, the plates are inserted in a dedicated chamber, where a few balls of iodine are deposited in the bottom. The chamber should be always kept close as much as possible and inside a hood, with appropriated chemical personal protection elements, since the purple vapour emanating from the iodine balls is highly reactive with any organic molecule, including those present in skin. Similar considerations need to be taken for ninhydrin and rhodamine 6G, specially avoiding contact with skin. After staining, spots may be marked for displacement calculation and obtention of retention factors.

A.1.1 Early attempts of quantification of surfactin using thin layer chromatography

At the start of the project, we explored the idea of developing a visual method to quantify surfactin via thin layer chromatography (TLC). This method on separation of compounds relying on interactions with a thin silica phase (stationary phase) over time after while dragged by a solvent (mobile phase). However, the method was unsuccessful, and readings (distinctive spots after staining) were available only for very high surfactin concentrations (>100 mM), on a scale not realistic for surfactin production with bacteria. In detail, the protocol involved taking 1ul from a supernatant sample, deposit it at the marked end of a silica-coated plate and then move the plate to a TLC chamber, a glass chamber where the solvent (or mobile phase) is at the bottom. When the plate is inserted, the solvent ascends the plate by capillary forces and the sample is dragged up the plate. In that process, different compounds in the sample ascend with different velocities, depending on the affinity with the silica (stationary phase), and thus, compounds can be separated depending on their polarity.

Several solvent mixes were tested for the mobile phase, including Chloroform : Methanol : Water (CMW) on ratios 1:3:1 and 20:9:1, and Acetonitrile : Water ratios including 1:3, 1:1 and 3:1. Together with the supernatant sample from an overnight M9 medium culture, a 1ul drop of a surfactin standard (5 mM) was added in a secondary starting point next to the first sample. The plates were developed and dried. After that, the plate was stained by either ninhydrin (amino acids detection), rhodamine 6G (general, fatty acids) or iodine (general).

However, none of the plates showed a clear spot pattern along the sample line, with sporadic spots such as the one shown in the circle on Figure A1, even after repeating the protocol and changing the staining. Indeed, rhodamine 6G staining was difficult to achieve for the sample and the selected standard. Only after performing runs changing the standard concentration to 100 mM, a blurred pink is visible, but no separation is apparent. Since this dye can fluorescence under an appropriate UV source, images were taking in a transilluminator to see any spot pattern that might be

present. This plate, for CMW 20:9:1, shows a single fluorescent spot (Figure A2) and no other spots can be determined.

Figure A1. A TLC run of a *Bacillus subtilis* supernatant sample using a solvent mix of Chloroform: Methanol: Water 20:9:1. The shown plate was stained with iodine. The circle indicates the unique spot that was identified after the run.

Figure A2. Imaging of plate after staining with Rhodamine G. The picture was taken in a transilluminator with UV filter. The blob on top is the surfactin standard 100 mM. For the *Bacillus* supernatant drops (different sample volume), no spots are observed.

Additional problems were found. An unexpected event happened when trying to heat the silica plate after the ninhydrin staining (Figure A3). One of the plates was not completely dried and after heating it, bubbles formed between the silica and the plastic, destroying the samples and generating an unusual pattern in the remaining plate. This observation is important, since increasing drying times would be detrimental for streamlining whole plate quantification using a TLC method, if ever a method would work.

Figure A3. Insufficient drying and quick heating can lead to bubbling and destruction of silica coated layer.

In the literature, there are only 2 reports employing classic TLC for surfactin quantification (Dlamini et al., 2020; Barale et al., 2022), but these achieved inconsistent results. There are other reports employing high performance TLC (HPTLC), which requires a specialised machine, where surfactin quantification can be performed more reliably (Geissler et al., 2017). However, the necessary equipment is not available in campus, as far as it is known. After these results, we decided to abandon this protocol and move forward to mass spectrometry quantitation.

A.2. Development of a 3D-printed electroporation tip for Opentrons P1000 pipette and building an open-source glass slide smearer.

In addition to the 3D-printed models used directly in the thesis protocols, I have worked on two other models that could have interesting applications for general experiments. The first is an electroporation tip for the Opentrons P100 pipette (Figure A4). Essentially, it is a tip with an attached banana plug receiver, allowing it to be

connected to an electroporation source, which often uses banana plug cables. The plugs are then connected to the electrodes at the tip, which are held in place solely by material support and the space occupied by the copper end of the cable. The separation between the electrodes is 3 mm, and the size of the electrode-holding arm was designed to fit wells in a 48-well plate.

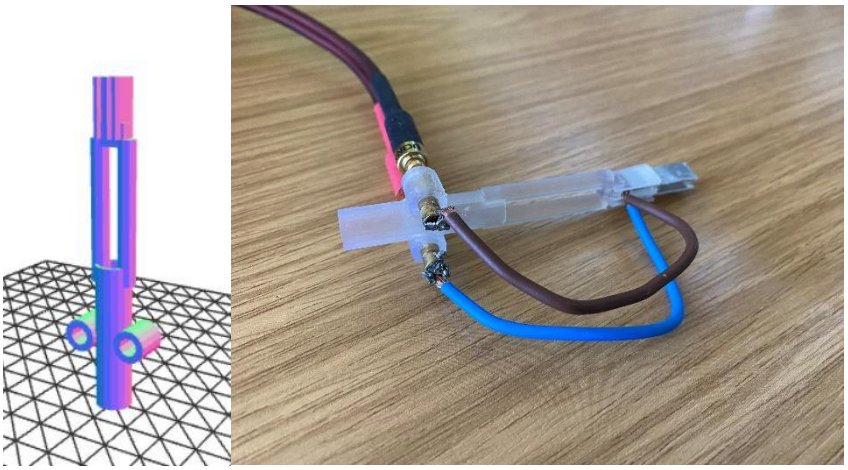


Figure A4. An Opentrons tip to perform electroporation on 48-well plates.

The tip has not been tested properly due to problems with the aluminium plates, but it is reported here for future purposes.

The second model was not designed from scratch, but rather it takes an open-source model from the web. It is a glass slide smearer (Figure A5), designed to generate consistent blood smears, and provided by the Prakash Lab at Stanford University (Nowak et al., 2023). The smearer is basically a support slide, that moves in a linear motion against to the secondary glass slide, which makes the smear. The linear motion is possible by a constant force spring connected to a syringe body, and the movement is damped by the available airspace in the syringe. This air volume is fine-tuned by the valve. Although purposed for blood smears, it can be used for preparation of bacterial slides for microscopy.

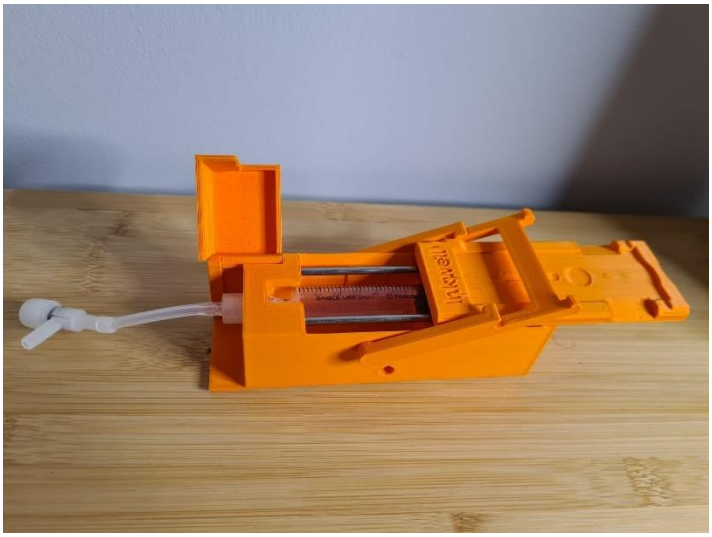


Figure A5. The built glass slide smearer. The 3D print design was provided by the Prakash Lab at Stanford University and reconstructed using preprint instructions.

A.3. Mathematical explanation for the sigmoidal shape in ordered samples.

Here, I record a simple explanation of why ordered samples by titre reassembles a sigmoidal shape when plotted. It is phenomenon that has some occurrence in the biological and chemical literature, but as far as I know, it is not often discussed. As described for Figure 4.7, the sigmoid shape can be used to detect linearity of the function respect to the variables. If a function depends on only one variable and this variable is sampled from a normal distribution, after evaluating the function on these samples and order the samples, the plot should the quantile function of the sampling distribution of the one variable, i.e., the probit function (Figure A6).

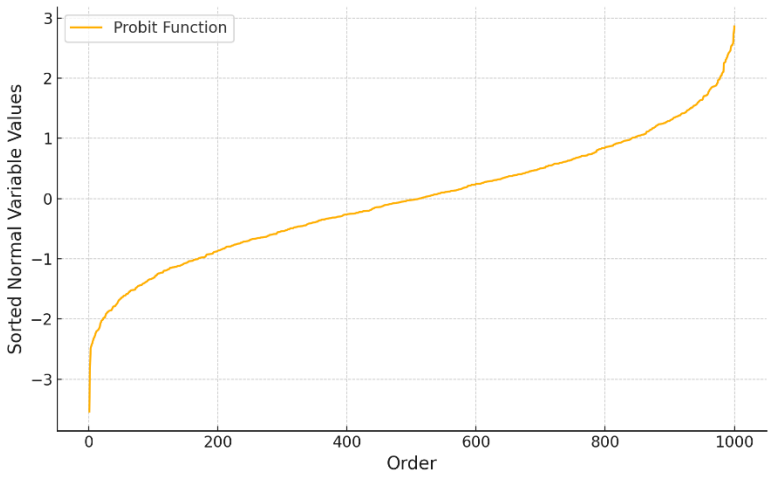


Figure A6. The probit function or quantile function of the normal distribution.

Now if the functions depend linearly on many variables, the same situation occurs, since a sum of normal random variables is again a normal random variable (Petrov, 1975), as the plot with the normal variable sampling and reorder will give a probit function.

There are two ways that this shape can be modified: sampling the variables with a different distribution that normal (Figure A7) or non-linearity respect to the variables (Figure A8). In our case, since we use Sobol sampling for the space filling design, and the number of combinations is not big, the tails of sigmoid curve can be lost (Figure A7, left). In the limit of the number of samples, both Sobol and random sampling converges to the probit (as expected)

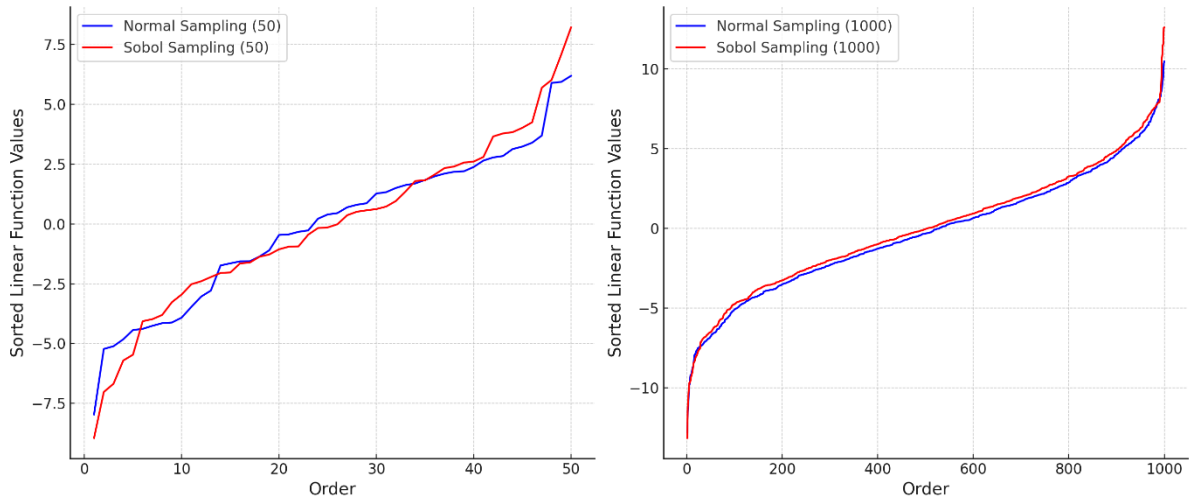


Figure A7. Plots of values of a linear function over 5 variables, in increasing order, using normal sampling or Sobol sampling.

In the case that the function is not linear over the parameters, like for example, using sine and exponential function, the shape of the ordered plot can change drastically, especially in the tails (Figure A8). They can become more steep or gentler depending on the nature of the function.

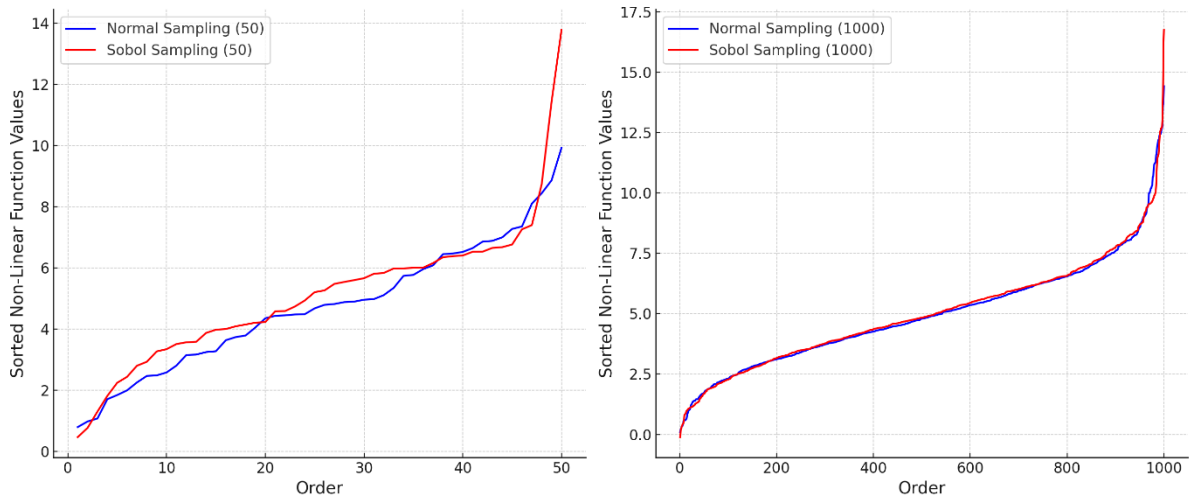


Figure A8. Plot of values of a non-linear function over 5 variables, in increasing order, using normal sampling or Sobol Sampling

As a literature example, in Zhang et al. 2020, this sigmoidal curve appears when ordering strains engineered strains by GFP synthesis rate. In this case, the underlying synthesis rate is expected to be non-linear over a big number of variables, since it depends on the gene circuit state and the strain state. The shape reassembles the plot in Figure A8, with gentler tails in the sigmodal shape, but a proper test is needed to see how they different they are.

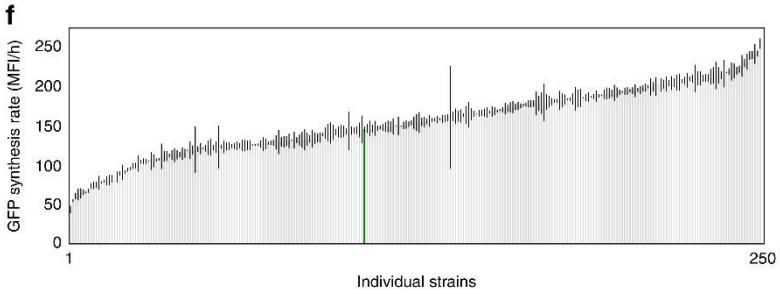


Figure A9. Ordered GFP synthesis rate vs individual strains, extracted from Zhang et al. 2020.

A.4. List of materials for building the robotic arm

- RC Servos: The robot uses 3 RC servos (13kg torque) for the wrist, elbow, and base joints. Servo horns are provided.
- 20KG Servo: Specifically for the shoulder joint, a 20kg torque servo should be used to sustain the weight of the arm. Servo horns are provided.

- Micro Servo: A micro servo is used for the gripper.
- Arduino Board: The design works with an Arduino Uno board, although other boards can be used with the appropriate modifications. Servo horns are provided.
- Servo Driver Board: A 16-channel, 12-bit PWM servo motor driver IIC module was used, which is compatible with Arduino boards.
- Potentiometers: 10k Ω potentiometers for the mini controller and analogue input.
- Gripper Gears: It is important to get a box of gears for DIY projects (64 pieces), as sizes may vary slightly in the final model. Three gears were employed.
- Larger Set of Screws: M2-M4 screws of varying lengths, typically sold as a box set.
- Small Set of Screws: M1-M1.6 screws, often sold as a box of screws typically used for repairing watches and clocks.
- Wire and Connectors: Dupont wires for board connections, and 28 AWG electrical wires for other connections.
- T-Plugs
- Adjustable Power Supply: It should reach 6V for proper functioning.
- Wire/Shrink Wrap: Usually included in wire packs.
- Rubber Band
- Gripper Foam Pad
- Power Switch: A 6A/250V, 10A/125V boat rocker switch was used.
- Controller Push Button: A box of tactile push buttons of different sizes and designs is sufficient for the purpose.

Appendix B: Supplementary Figures and Tables

Supplementary Figures

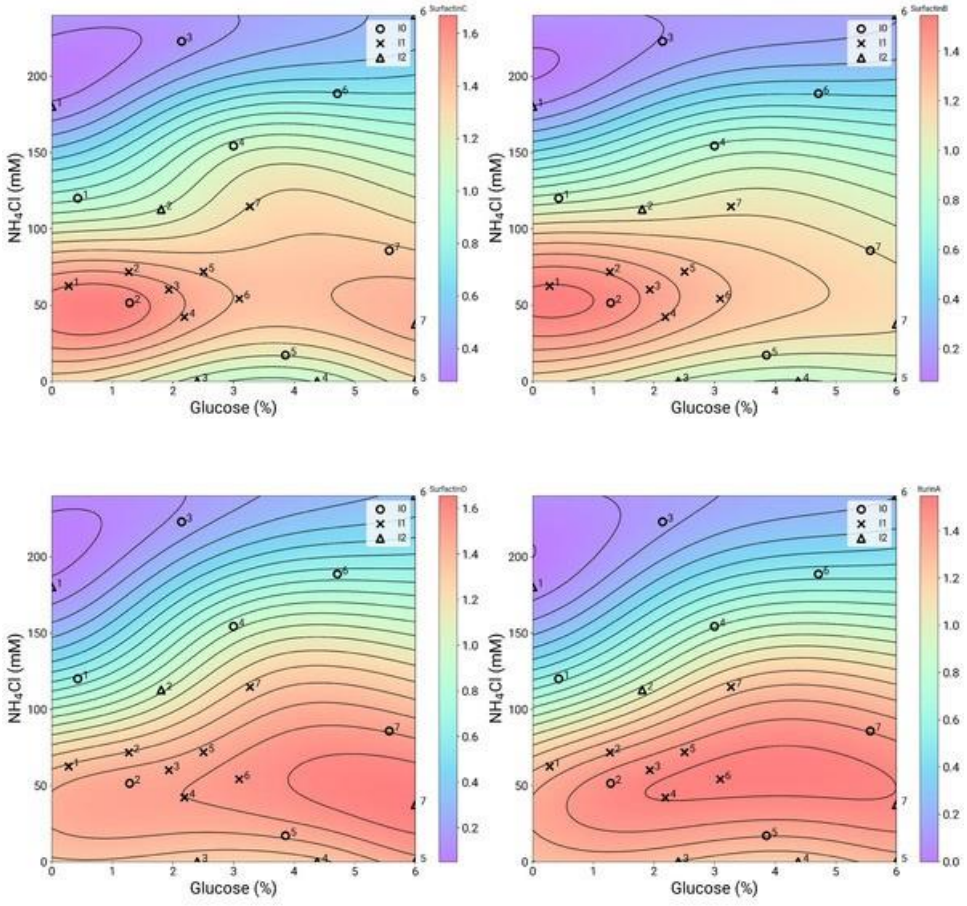


Figure B.1. Surfaces for the 4 lipopeptides that were simultaneously measures using the flow injection MS method. Samples from all iterations are depicted in the surfaces. Colour indicates predicted titre by the Gaussian process regression model.

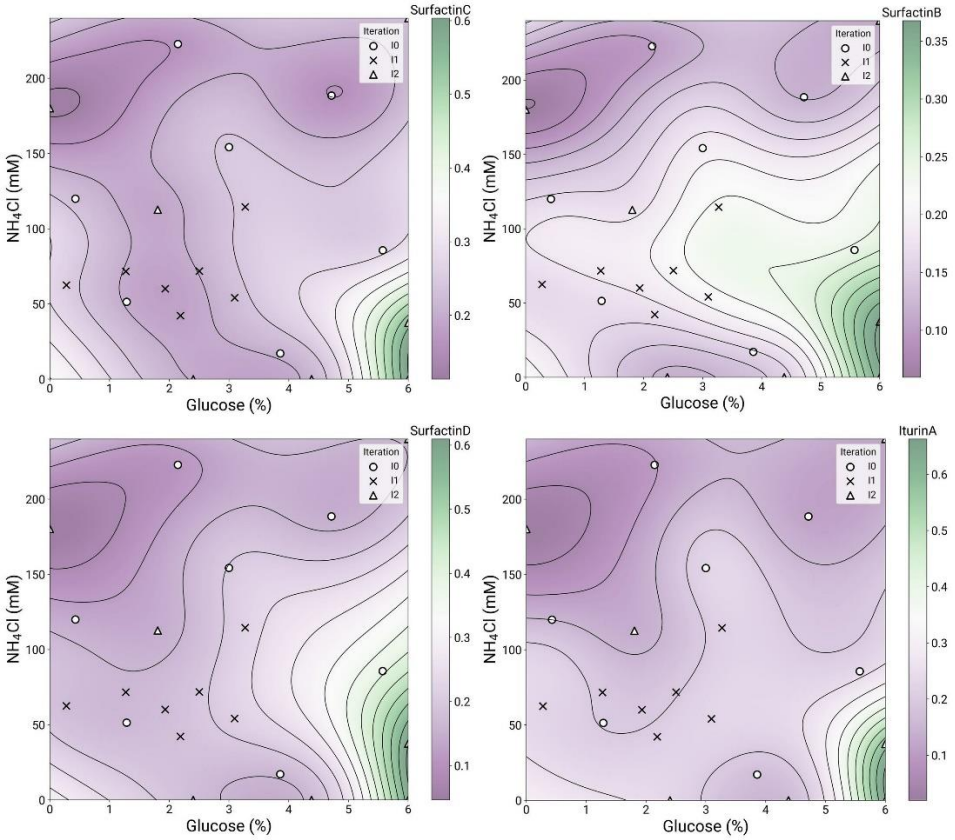


Figure B.2. Uncertainty surfaces for the 4 lipopeptides that were simultaneously measures using the flow injection MS method. Samples from all iterations are depicted in the surfaces. Colour indicates predicted titre by the Gaussian process regression model.

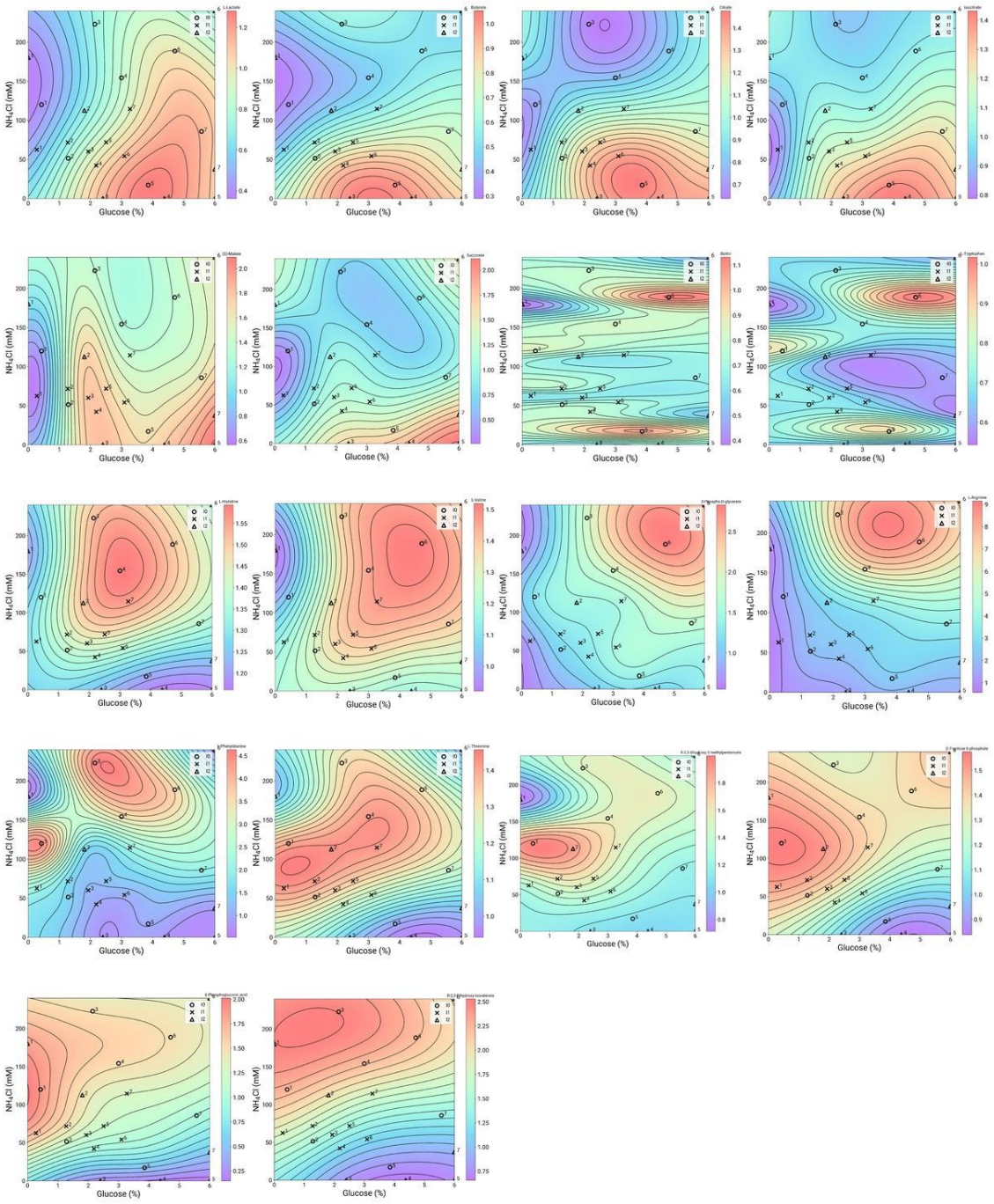


Figure B.3. Surfaces for 18 carbon/TCA-related compounds that were measured outside of the loop using the flow injection MS method. Samples from all iterations are depicted in the surfaces. Colour indicates predicted titre by the Gaussian process regression model.

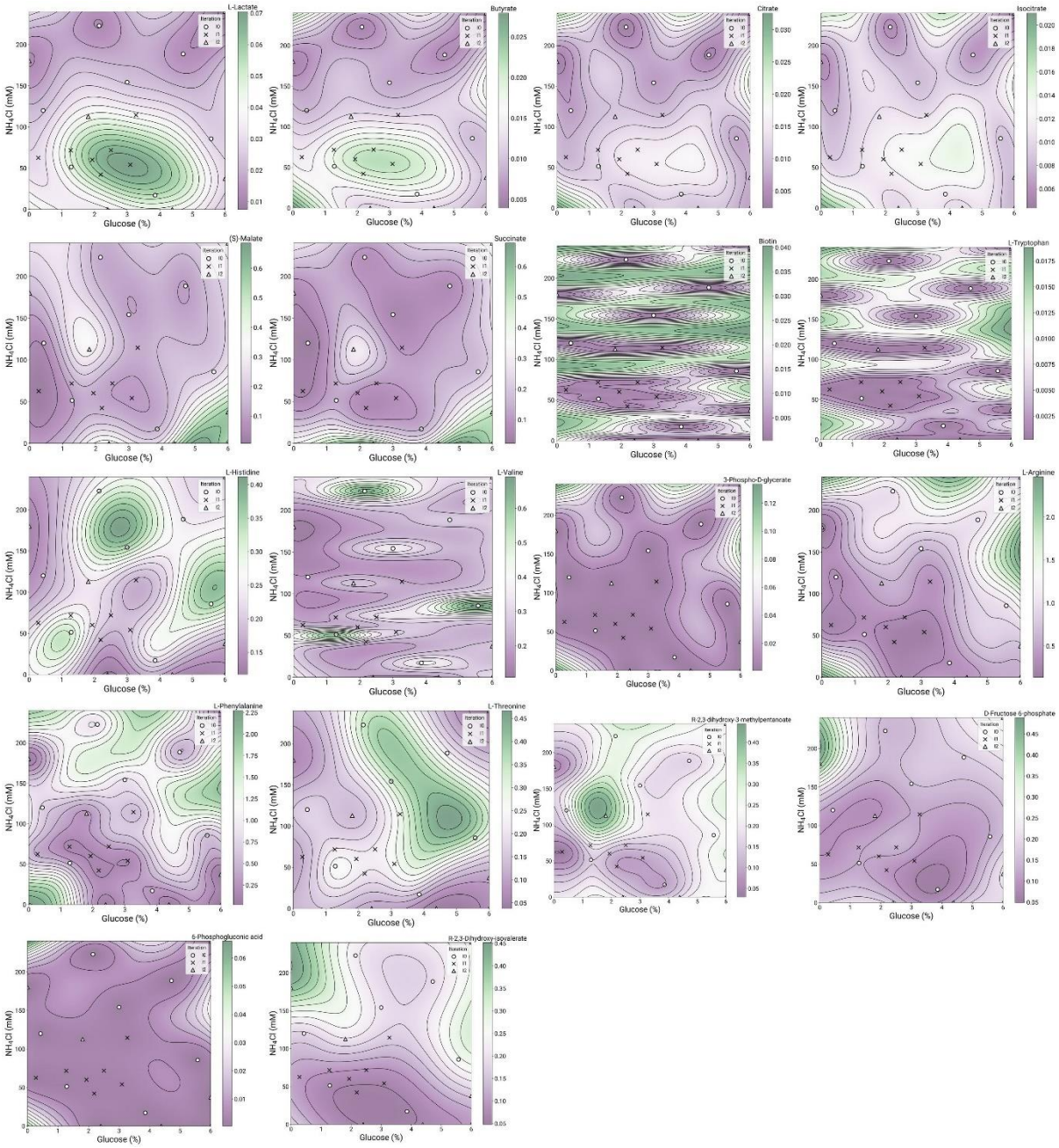


Figure B.4. Uncertainty surfaces for 18 carbon/TCA-related compounds that were measured outloop using the flow injection MS method. Samples from all iterations are depicted in the surfaces. Colour indicates predicted titre by the Gaussian process regression model.

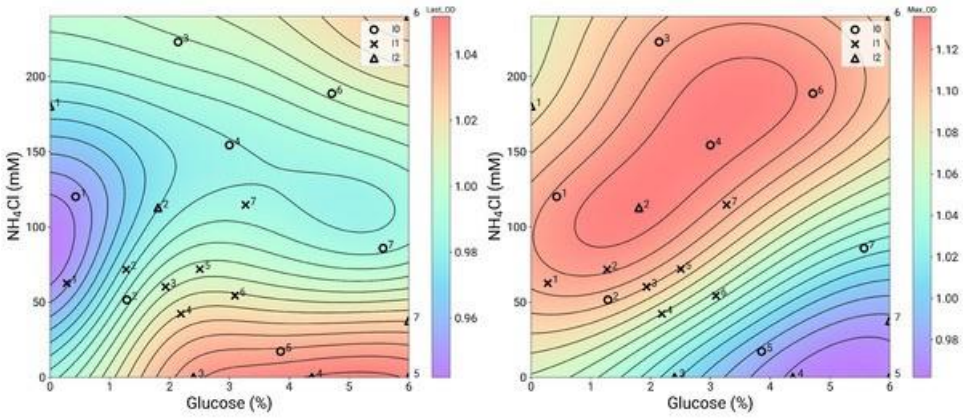


Figure B.5. Surfaces for 2 growth measurements that were obtained from the plate reader experiment. Samples from all iterations are depicted in the surfaces. Colour indicates predicted titre by the Gaussian process regression model.

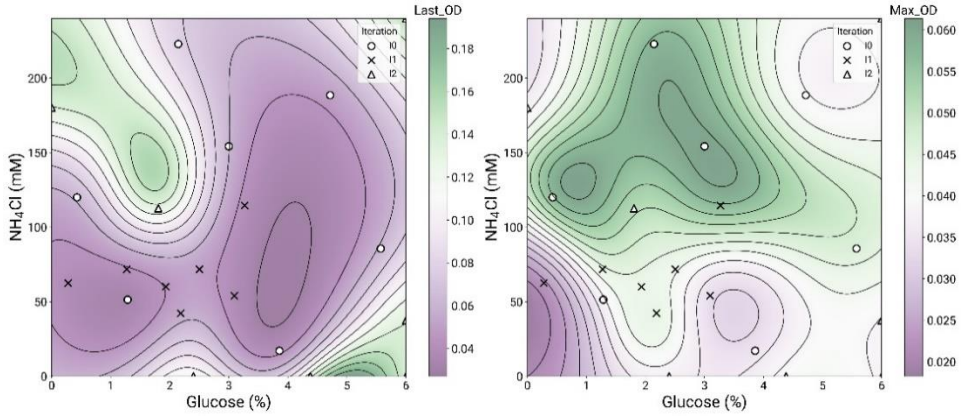


Figure B.6. Uncertainty surfaces for 2 growth measurements that were obtained from the plate reader experiment. Samples from all iterations are depicted in the surfaces. Colour indicates predicted titre by the Gaussian process regression model.

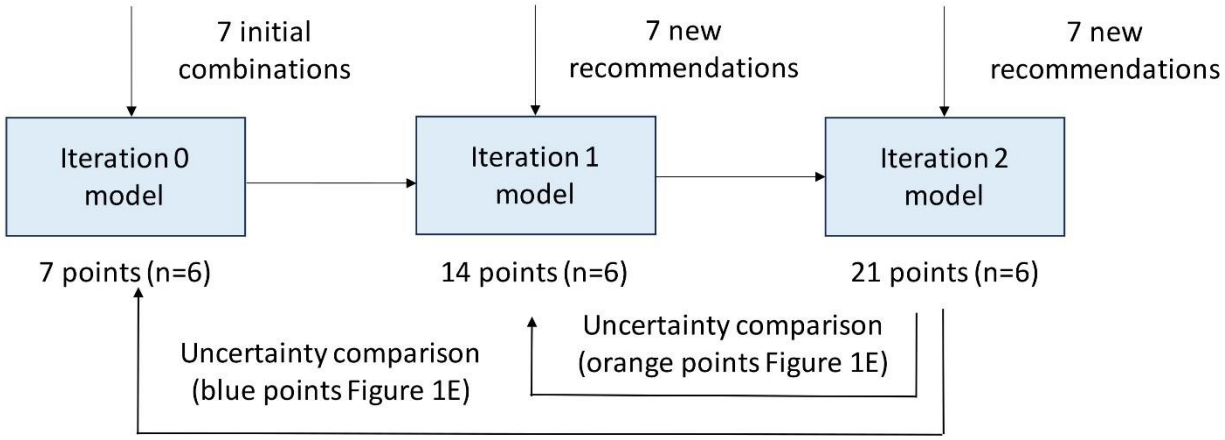


Figure B.7. Diagram showing what data serves as input to the model in each iteration and how the predicted uncertainty in the model is compared. The model is being

updated after each iteration, augmenting the available information and therefore reducing the predicted uncertainty for simulated points in the models.

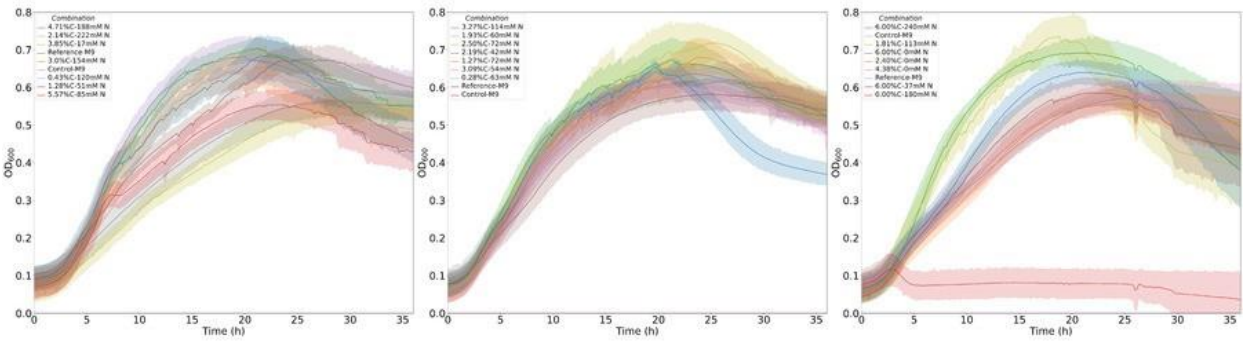


Figure B.8. Growth curves for the microplates. From left to right, it is depicted the growth curves for different C/N combinations in Iteration 0, Iteration 1 and Iteration 2. The trend colour is given by a specific carbon and nitrogen concentration, as shown in the legend. The shadow in each trend corresponds to the confidence band calculated from the 6 biological replicates.

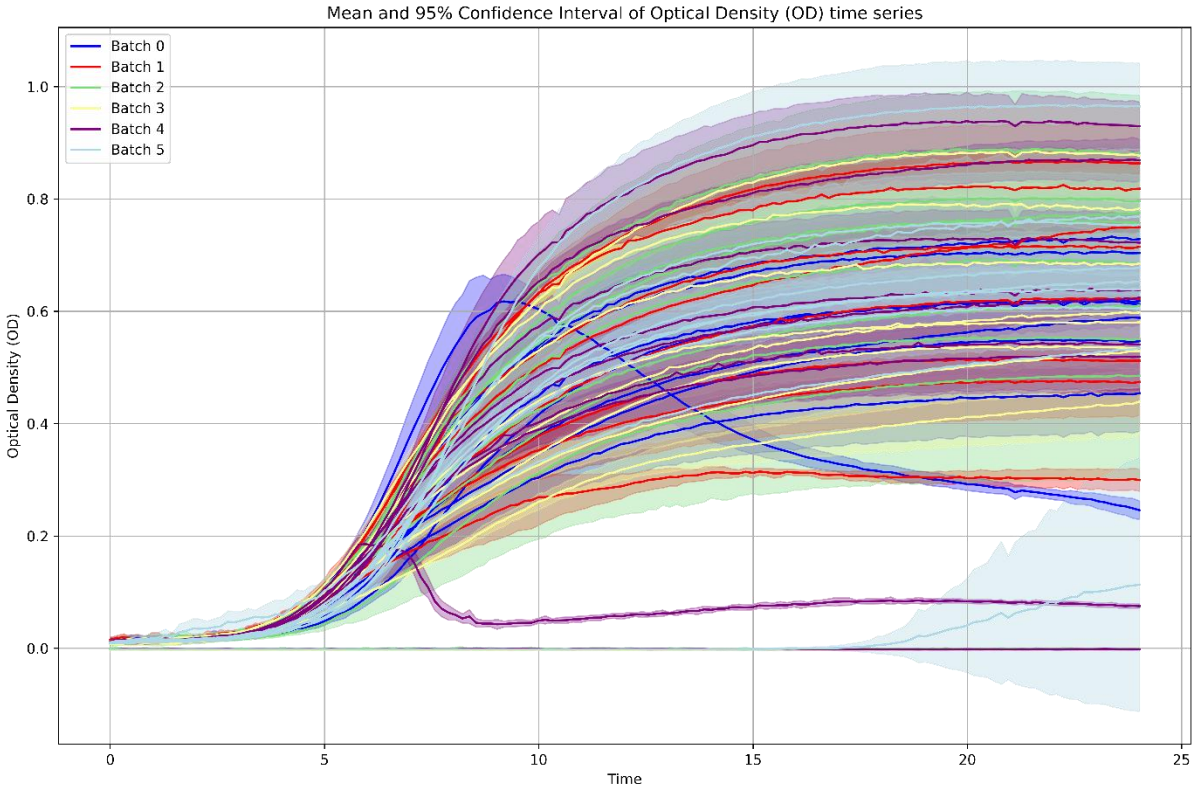


Figure B.9. OD curves for the space-filling design experiment. The colour of the curves depends of the batch the sample come from.

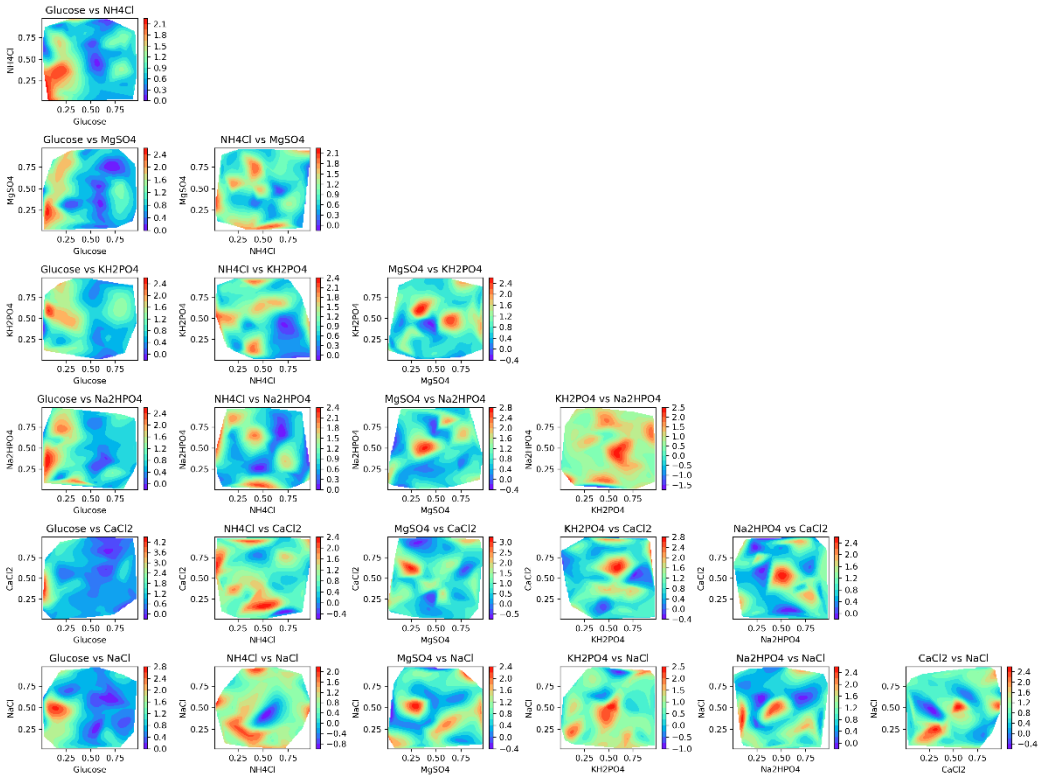


Figure B.10. Contour plot of Surfactin B titres against pair of components concentrations.

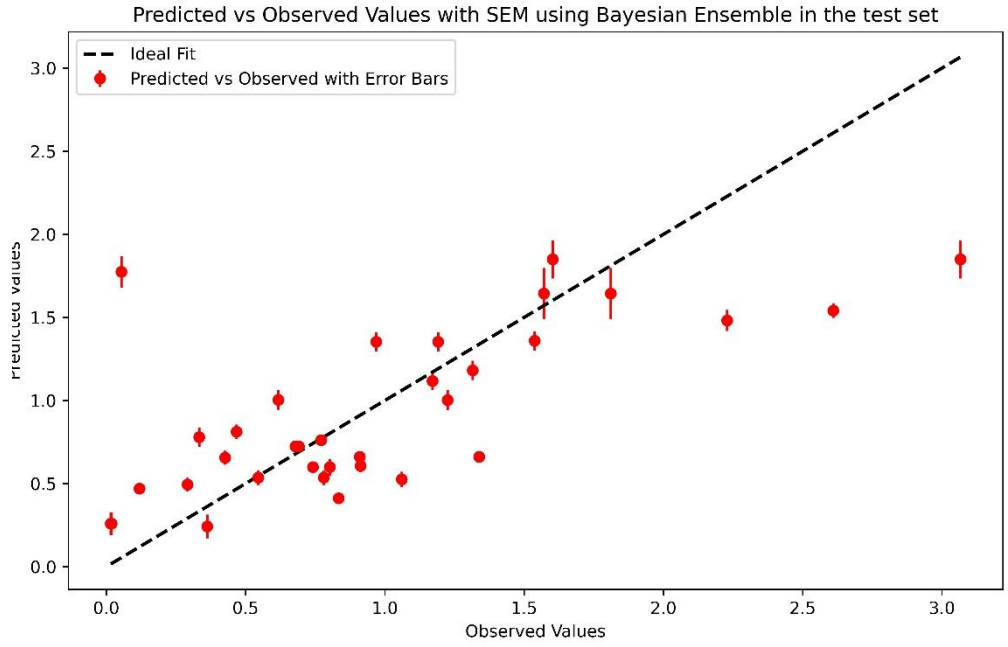


Figure B.11. Results of the Bayesian averaging ensemble prediction on test set for Surfactin B

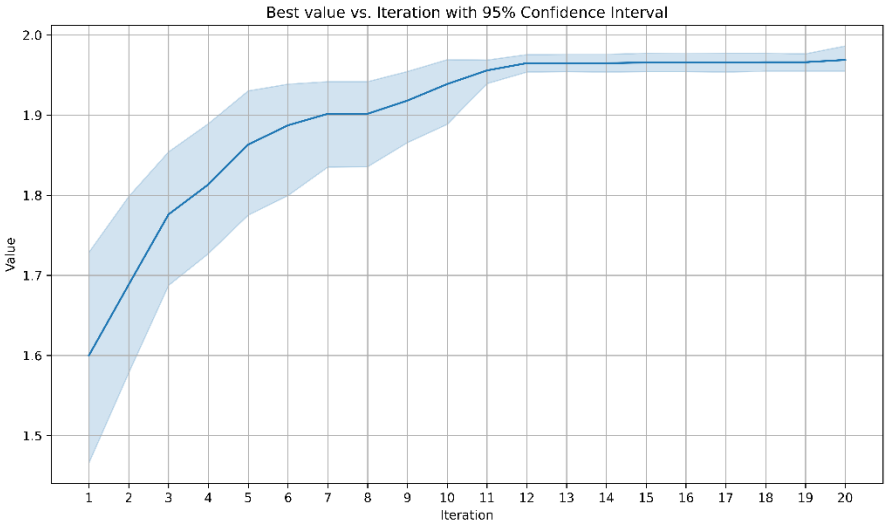


Figure B.12. Single objective optimisation of Surfactin B using the surrogate data. 7 initial samples and 7 samples per batch were employed.

Supplementary Tables

Table BT1. Parameters used on the triple quadrupole mass spectrometer (QqQ-MS) runs using the developed flow injection method.

Parameter	Value
Spray Voltage	**Static**
Positive Ion (V)	**3500**
Negative Ion (V)	**3500**
Sheath Gas (Arb)	**35**
Aux Gas (Arb)	**5**
Sweep Gas (Arb)	**0**
Ion transfer Tube Temp (°C)	**325**
Vaporizer Temperature (°C)	**275**

Table BT2. Precursor and product masses used for multiple reaction monitoring (MRM) in the QqQ-MS

Molecule	Polarity	Precursor mass (m/z)	Product mass (m/z)	Collision Energy (V)	Dwell time (ms)
Lipopeptides					
SurfactinB	Positive	1008.2	685	20	6.581
SurfactinC	Positive	1022.3	685	20	6.581
IturinA	Positive	1044.3	391	35	6.581
SurfactinD	Positive	1058.2	685	20	6.581
Central metabolism					
L-Valine	Positive	118.086	57.054	29.94	6.581
L-Valine	Positive	118.086	72	11.49	6.581
L-Threonine	Positive	120.066	74	11.4	6.581
L-Threonine	Positive	120.066	103	18.44	6.581
L-Histidine	Positive	156.077	93	23.62	6.581
L-Histidine	Positive	156.077	110.054	15.19	6.581
L-Phenylalanine	Positive	166.086	77	39.84	6.581
L-Phenylalanine	Positive	166.086	120.054	14.06	6.581
L-Arginine	Positive	175.119	70.071	23.7	6.581
L-Arginine	Positive	175.119	158.054	12.71	6.581
L-Tryptophan	Positive	205.097	146.071	18.18	6.581
L-Tryptophan	Positive	205.097	188.071	10.35	6.581
L-Lactate	Negative	89.024	45.125	12.2	6.581

L-Lactate	Negative	89.024	71.012	9.21	6.581
Butyrate	Negative	89.1	43.1	14	6.581
Succinate	Negative	117.019	73.113	11.74	6.581
Succinate	Negative	117.019	98.827	7.86	6.581
(S)-Malate	Negative	133.014	70.988	13.34	6.581
(S)-Malate	Negative	133.014	114.929	11.15	6.581
R-2,3-Dihydroxy-isovalerate	Negative	133.051	133.051	0	6.581
R-2,3-dihydroxy-3-methylpentanoate	Negative	147.066	147.066	0	6.581
3-Phospho-D-glycerate	Negative	184.986	123.107	17.01	6.581
3-Phospho-D-glycerate	Negative	184.986	166.917	9.04	6.581
Citrate	Negative	191.02	86.845	17.89	6.581
Isocitrate	Negative	191.02	116.958	14.81	6.581
Isocitrate	Negative	191.02	172.929	7.86	6.581
Biotin	Negative	245.1	227	14	6.581
D-Fructose 6-phosphate	Negative	259.022	138.929	16.5	6.581
D-Fructose 6-phosphate	Negative	259.022	168.708	8.71	6.581
6-Phosphogluconic acid	Negative	275.017	195.143	21.68	6.581
6-Phosphogluconic acid	Negative	275.017	256.899	12.83	6.581

Table BT3. Estimation of growth rates and maximum OD for OFAT experiments. For biphasic growth, only the first calculated growth rate is shown. Mean and standard deviation over replicates is reported.

Glucose variable, fixed ammonium chloride concentration.

Treatment	Max OD mean	Max OD std	Growth rate (hr^{-1}) mean	Growth rate rate (hr^{-1})std
0%	0.168	0.042	0.299	0.143
1%	0.515	0.044	0.376	0.155
2%	0.521	0.042	0.335	0.141
3%	0.525	0.046	0.291	0.119
4%	0.512	0.039	0.326	0.136
5%	0.496	0.054	0.333	0.159

6%	0.536	0.053	0.270	0.115
M9/0.4%	0.559	0.054	0.272	0.106
M9/Control	0	0	0	0

Ammonium chloride concentration variable, fixed glucose.

Treatment	Max OD mean	Max OD std	Growth rate mean	Growth rate std
0 mM	0.467467	0.052022	0.347049	0.139907
40 mM	0.544433	0.053184	0.328716	0.148633
80 mM	0.581983	0.041882	0.403632	0.159925
120 mM	0.59035	0.07058	0.466632	0.154816
160 mM	0.627233	0.05569	0.429489	0.155371
200 mM	0.658733	0.059223	0.384713	0.121643
240 mM	0.661417	0.062584	0.381513	0.132298
M9/18.7 mM	0.50335	0.066733	0.373267	0.132273
M9/Control	0	0	0	0

Table BT4. Pipetting results for normal speed test. Both P300 and P1000 testing are represented in the table.

Pipette	**Expected Volume (µL)**	**Measured Volume (µL)**
P300	30	29.5
P300	30	28.7
P300	30	28.6
P300	30	29.3
P300	30	29.3
P300	30	29.2
P300	30	28.6
P300	30	29
P300	30	29
P300	30	29
P300	30	28.9
P300	30	29
P300	30	28.3
P300	30	29.1
P300	30	28.6
P300	30	28.6
P300	30	29.1
P300	30	29
P300	150	145.4
P300	150	144.3
P300	150	144.7
P300	150	146.7

P300	150	143.9
P300	150	146.1
P300	150	146.1
P300	150	145.1
P300	150	144.4
P300	150	143.5
P300	150	146.5
P300	150	145.4
P300	150	143.7
P300	150	144.8
P300	150	144.7
P300	150	145.4
P300	150	146.4
P300	150	147.4
P300	300	290.3
P300	300	295.3
P300	300	293.3
P300	300	295.4
P300	300	294.9
P300	300	291.2
P300	300	292.9
P300	300	290.4
P300	300	297
P300	300	293
P300	300	295.5
P300	300	294.1
P300	300	294.1
P300	300	288.7
P300	300	291.5
P300	300	285.9
P300	300	289.9
P300	300	288.2
P1000	100	99
P1000	100	95
P1000	100	97
P1000	100	98
P1000	100	98
P1000	100	98
P1000	100	98
P1000	100	100
P1000	100	100
P1000	100	99
P1000	100	96

P1000	100	97
P1000	100	97
P1000	100	97
P1000	100	100
P1000	100	98
P1000	100	97
P1000	100	96
P1000	500	490
P1000	500	476
P1000	500	474
P1000	500	487
P1000	500	487
P1000	500	497
P1000	500	485
P1000	500	484
P1000	500	492
P1000	500	489
P1000	500	482
P1000	500	484
P1000	500	479
P1000	500	486
P1000	500	488
P1000	500	499
P1000	500	487
P1000	500	480
P1000	1000	961
P1000	1000	973
P1000	1000	972
P1000	1000	984
P1000	1000	986
P1000	1000	976
P1000	1000	993
P1000	1000	955
P1000	1000	957
P1000	1000	958
P1000	1000	979
P1000	1000	965
P1000	1000	982
P1000	1000	972
P1000	1000	972
P1000	1000	981
P1000	1000	957
P1000	1000	976

Table BT5. Pipetting results for slow speed test (0.5 rate). Both P300 and P1000 testing are represented in the table.

Pipette	Expected Volume (µL)	Measured Volume (µL)
P300	30	29.1
P300	30	29.5
P300	30	29.1
P300	30	29.1
P300	30	28.9
P300	30	29.3
P300	30	29.3
P300	30	29.3
P300	30	29.7
P300	30	29.6
P300	30	29.1
P300	30	29.5
P300	30	29.3
P300	30	29.5
P300	30	29.3
P300	30	29.4
P300	30	29.2
P300	30	29.1
P300	150	146.7
P300	150	144.4
P300	150	146.2
P300	150	147.6
P300	150	145.5
P300	150	146.9
P300	150	147.1
P300	150	149.2
P300	150	150.4
P300	150	147.7
P300	150	146.9
P300	150	147.4
P300	150	149
P300	150	145.2
P300	150	147.4
P300	150	149.7
P300	150	147.4
P300	150	146
P300	300	294.7
P300	300	299.2

P300	300	292.6
P300	300	296.9
P300	300	294.2
P300	300	299.3
P300	300	295.8
P300	300	291.9
P300	300	297
P300	300	299.2
P300	300	296.4
P300	300	294.9
P300	300	296
P300	300	297.9
P300	300	294.4
P300	300	297.4
P300	300	293.4
P300	300	293.3
P1000	100	101
P1000	100	98
P1000	100	97
P1000	100	100
P1000	100	97
P1000	100	99
P1000	100	97
P1000	100	99
P1000	100	97
P1000	100	94
P1000	100	98
P1000	100	100
P1000	100	99
P1000	100	97
P1000	100	99
P1000	100	97
P1000	100	100
P1000	100	97
P1000	500	490
P1000	500	492
P1000	500	484
P1000	500	488
P1000	500	488
P1000	500	488
P1000	500	489
P1000	500	495
P1000	500	496

P1000	500	490
P1000	500	491
P1000	500	500
P1000	500	492
P1000	500	491
P1000	500	495
P1000	500	485
P1000	500	493
P1000	500	491
P1000	1000	981
P1000	1000	988
P1000	1000	986
P1000	1000	983
P1000	1000	987
P1000	1000	986
P1000	1000	982
P1000	1000	994
P1000	1000	999
P1000	1000	990
P1000	1000	989
P1000	1000	989
P1000	1000	992
P1000	1000	998
P1000	1000	989
P1000	1000	999
P1000	1000	985
P1000	1000	979