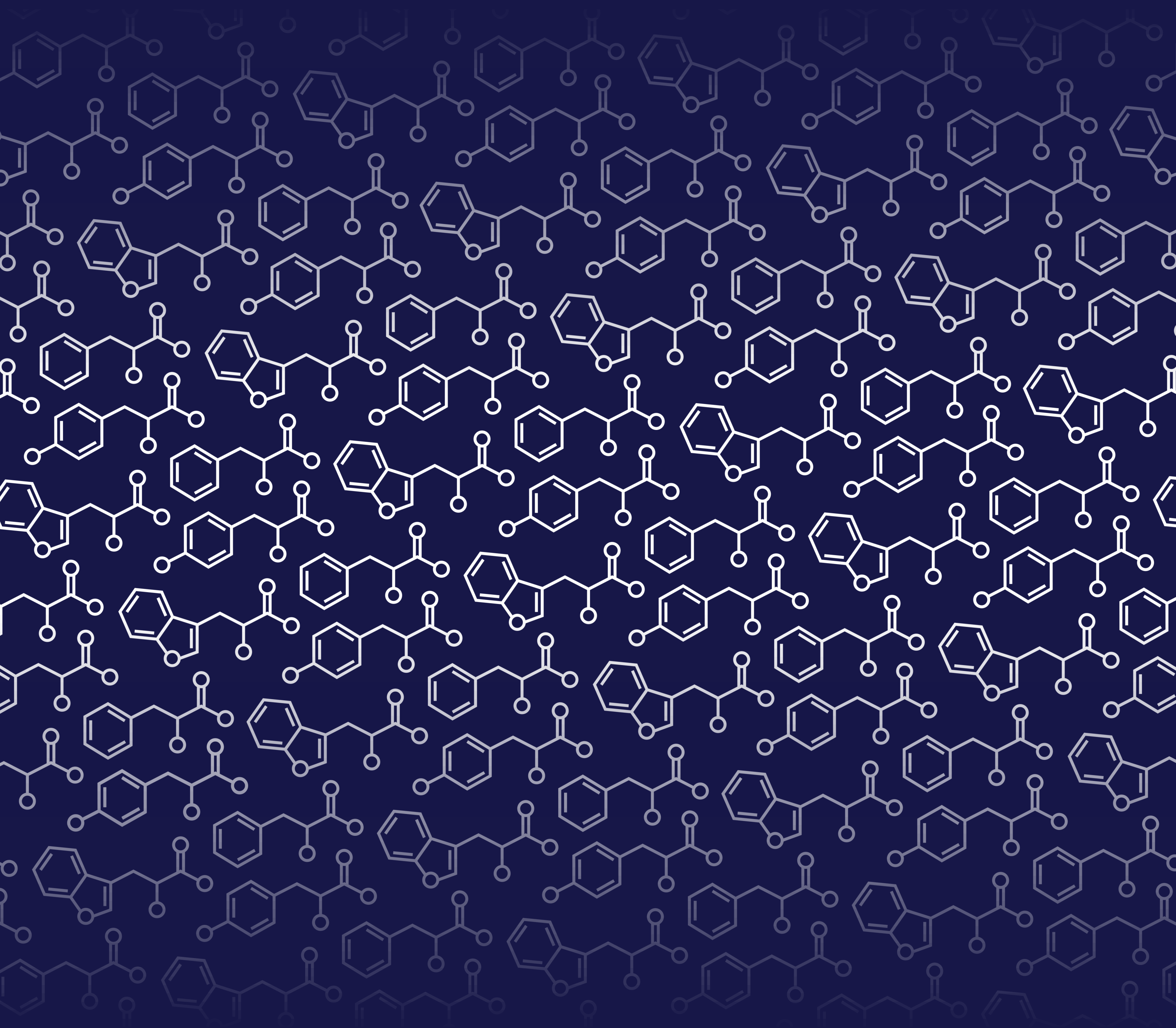
Computational tools and analytical methods for effective metabolic engineering of microbial cell factories

Taylor Parkins, Shannara Kayleigh

DTU Biosustain
The Novo Nordisk Foundation Center for Biosustainability

Computational tools and analytical methods for effective metabolic engineering of microbial cell factories



PhD thesis

Shannara Kayleigh Taylor Parkins

Computational tools and analytical methods for effective metabolic engineering of microbial cell factories

PhD thesis
September, 2023

By
Shannara Kayleigh Taylor Parkins

Published by: DTU, The Novo Nordisk Foundation Center for Biosustainability,
Building 220, Kemitorvet, 2800 Kgs. Lyngby, Denmark

Approval

This thesis is written as partial fulfilment of the requirements to obtain a PhD degree at the Technical University of Denmark (DTU), from October 2020 to September 2023. The work was carried out in the Quantitative Modelling of Cell Metabolism research group supervised by Professor Lars Nielsen, with co-supervision by Teddy Groves. The PhD study included an external research stay at the University of California San Diego (UCSD), San Diego, CA, from January to March 2023 under the supervision of Professor Bernhard Palsson.

Shannara Kayleigh Taylor Parkins

____

Signature

____

Date

Abstract

Biomanufacturing of chemical and natural products is set to improve environmental sustainability and increase reliability of the production processes. Using microbial cell factories for biomanufacturing opens up a large range of potential products, especially considering the ability to produce non-native products through metabolic engineering. While possible, metabolic engineering remains challenging due to the inherent complexity of microbial metabolism. Metabolic engineering addresses this challenge by following an iterative design-build-test-learn (DBTL) cycle, where the design and the build phase have seen the most advancements in recent years. The full cycle would benefit from high-throughput and automated strategies, which can be implemented both in the laboratory and digitally. In this thesis, two computational tools and two analytical methods were developed to improve the performance and throughput of the test and learn phases. Regarding the test phase, a high-throughput workflow for absolute quantitative proteomics was established and integrated into an automated data analysis tool. An optimised sample preparation protocol for membrane proteomics was developed, which results in a more representative membrane fraction for proteome-wide analysis. As a tool for the learn phase, a kinetic model was constructed to investigate the inner workings of metabolism through integration of proteomics, metabolomics and fluxomics data. Altogether, the computational tools and analytical methods presented here could be applied to improve future development of microbial cell factories.

Sammenfatning

Bioteknologisk produktion af kemiske og naturlige produkter vil mindske aftrykket på miljøet, samt gøre produktionen mere robust. Anvendelse af mikrobielle cellefabrikker har et stort potentiale, da det muliggør produktionen af bred vifte af kendte molekyler, men samtidigt giver det mulighed for at producere helt nye molekyler. Desværre er udvikling a cellefabrikker fortsat udfordret af mikrobernes iboende kompleksitet. Cellefabrik ingeniører håndterer denne udfordring ved at følge en iterativ design-byg-test-lær (DBTL)-cyklus, hvor design- og byg faserne har set flest fremskridt de seneste år. For at få mest mulig ud af den fulde DBTL-cyklus kan man med fordel anvende højt-kapacitetsmetoder og automatisering. Disse kan implementeres både i laboratoriet og digitalt. I denne Ph.d.-afhandling blev der udviklet to beregningsværktøjer og to analysemetoder som øger kapaciteten og kvaliteten af test- og læringsfaserne. Med hensyn til testfasen blev der etableret et høj-kapacitetsdatabehandlingsprogram til absolut kvantitativ proteom data, som blev integreret i et automatiseret dataanalyseværktøj. Der blev udviklet en optimeret prøveforberedelsesprotokol til membranproteiner, som resulterede i en mere repræsentativ membranfraktion i proteomanalysen. Som et værktøj til indlæringsfasen blev der konstrueret en kinetisk model til undersøgelse af det metaboliske netværk gennem integration af proteom-, metabolom- og fluxom-data. Alt i alt kan de her viste beregningsværktøjer og analysemetoder anvendes til at forbedre den fremtidige udvikling af mikrobielle cellefabrikker.

Publications

Included in this thesis

- Chapter 2: **Shannara Kayleigh Taylor Parkins**, Nicolás Gurdo, Tune Wulff, Pablo Iván Nikel and Lars Keld Nielsen, Automated data analysis for absolute quantitative proteomics - a benchmarking study in *Escherichia coli.* (Under review)
- Chapter 3: **Shannara Kayleigh Taylor Parkins**, Nicolás Gurdo, Elli Dertili, Martina Fricano, Pablo Iván Nikel and Lars Keld Nielsen, High-throughput quantitative membrane proteomics of gram-negative bacteria - optimising the sample preparation. (Ready for submission)
- Chapter 4: **Shannara Kayleigh Taylor Parkins**, Nicholas Luke Cowie, Jorge Carrasco Muriel, Teddy Groves and Lars Keld Nielsen, Investigation of the aromatic amino acid biosynthesis in Escherichia coli – applications of a kinetic model. (Manuscript in preparation)
- Chapter 4 - epilogue: Pascal Aldo Pieters, Se Hyeuk Kim, Christina Lenhard, Zofia Dorota Jarczynska, **Shannara Kayleigh Taylor Parkins**, Marta Reventós Montané, Peter Ehlert Jensen, Suresh Sudarsan, Emre Özdemir, Lei Yang and Lars Keld Nielsen, Multi-omics analysis of an *Escherichia coli* L-tyrosine overproducer. (Manuscript in preparation)

Not included in this thesis

- Gurdo, N., **Taylor Parkins, S. K.**, Fricano, M., Wulff, T., Nielsen, L. K., and Nikel, P. I. (2023), Protocol for absolute quantification of proteins in gram-negative bacteria based on QconCAT-based labeled peptides, *STAR Protocols*, *4*(1),102060. doi:10.1016/j.xpro.2023.102060 (Published)
- Gurdo, N., Mohamed, E. T., Johnsen, J., Tagliani, T., O'Connell, G. W., **Taylor Parkins, S.K.**, Nielsen, L.K., Feist, A. M. and Nikel, P. I., Engineering the native metabolism of Pseudomonas putida for adaptation to multi-substrate environments. (Manuscript in preparation)

Acknowledgements

First of all I would like to thank my supervisor Professor Lars Nielsen for this opportunity and his supervision through these three years. I had no idea what I said yes to after our one-hour interview, but it turned out to be exactly what I needed. I am so glad you let me follow my own path during this PhD and you were always there whenever I needed your guidance. Next, I would like to thank my co-supervisor Teddy Groves for his supervision and commitment. I learned a lot about statistics and writing from you and I hope you learned a little about biology from me in return. I would also like to thank Tune Wulff for being a mentor to me for more than just proteomics. I always enjoyed our chats and working together in general.

Thank you to Professor Pablo Nikel for allowing me to collaborate with him and his amazing PhD student Nicolás Gurdo. It was a fantastic collaboration, which I enjoyed every minute of. Thank you Nico, for always being there for me, especially for those long hours in the lab. Next, I would like to thank Lizzie for all her support in- and outside of the lab. I enjoyed working with you very much over the past years. To everyone in the Tyrosine Strain Design biofoundry project, thank you for collaborating with me. It has been one of my favourite projects and something I would like to pursue after my PhD as well. I would like to thank my two amazing students, Elli and Martina, whom I had the pleasure of supervising for their master thesis projects. A huge thank you to Rebeca for being our PhD coordinator, but also my friend. Next, I would like to thank Professor Bernhard Palsson for allowing me to visit his research group at UCSD for my external stay.

I would like to thank past and present members of the Quantitative Modelling of Cell Metabolism research group for making our office such a welcoming environment: Nick, Christina, Viktor, Areti, Marina, Jorge, Erin, Hossein, Sergi and Victor. It fills me with joy when I look back on our shared lunches, coffee chats and brainstorming sessions. Next, a big thank you to the people at UCSD for making my external stay a lot more enjoyable: Kevin, Arjun, Heera, Amy and Annie.

A special thank you to all the amazing women I have had the absolute pleasure of meeting during my PhD and who continue to inspire and motivate me every day: Christina, Rebeca, Erin, Heera, Lizzie, Elli, Martina and Ling.

I would like to thank my family for always believing in me, even when they do not understand much of what I am actually researching. You have always been my safe space and I am so glad to share these special moments with you. Last, but certainly not least, thank you to my loving husband Matthew. You really do make everything better, including this PhD journey. I may have been able to do it alone, but I will choose to embark on any journey with you again and again.

Contents

List of Figures

List of Tables

Chapter 1.

General introduction

1.1. General introduction

Microbial cell factories constitute a potential solution for improving environmental sustainability and increasing reliability of production processes otherwise based on fossil fuels or plant production [1]. Biomanufacturing using microbial cell factories offers extreme product diversity, ranging from biofuels [2] to natural products, such as pharmaceuticals, flavourings and pigments [3]. Despite the clear potential, application of microbial cell factories poses challenges due to the inherent complexity of microbial metabolism. Metabolic engineering addresses this by following an iterative design-build-test-learn (DBTL) cycle (Figure 1) [4]. Typically, several iterations are required in order to achieve a significant improvement in product titres and overall performance. The majority of this optimisation process is based on a trial-and-error approach and is often time consuming and labour intensive. For example, achieving industrial-scale microbial production of artemisin and 1,3-propanediol required 150 and 175 person-years, respectively [5]. Decreasing the duration of DBTL cycle iterations is a worthwhile goal, as it would increase the efficiency and cost-effectiveness of the construction and optimisation of microbial cell factories.

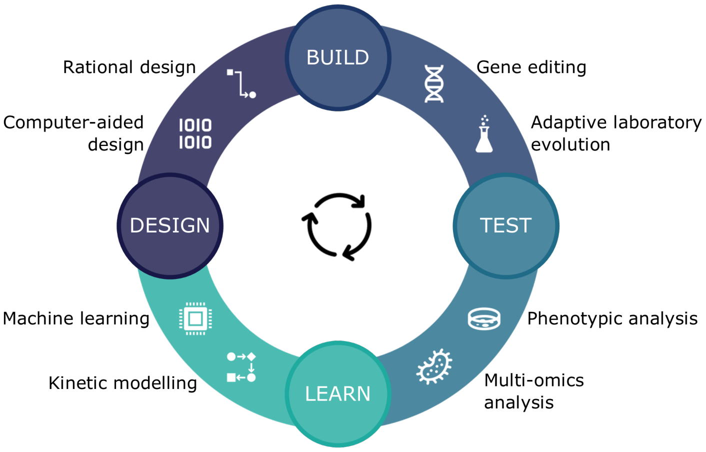


Figure 1.1: Overview of the Design-Build-Test-Learn cycle commonly applied to metabolic engineering of microbial cell factories. Typically multiple iterations are performed for a specific development. Two examples are given per phase here, however, more tools and approaches exist for each phase.

The overall DBTL cycle duration has been greatly reduced by implementation of more high-throughput and partly automated methods [6, 7], while several improvements have also been made in specific cycle phases. The design phase has seen advancement through guidance from genome-scale metabolic models, however, the design of microbial cell factories is often still a one-off, manual process based on the expertise of the researchers involved [4]. A large step forward in the build phase came from the introduction and availability of the CRISPR/Cas9 system for conventional host organisms, such as *Escherichia coli* and *Saccharomyces cerevisiae* [8]. In recent years, the focus in the test phase has shifted towards omics techniques due to increased prevalence of accessible methods to acquire omics data sets [6]. The learn phase has consistently been underrepresented during microbial cell factory development following the DBTL cycle, yet this phase holds the most potential to accelerate the overall development [4].

The emergence of biofoundries brings a high-throughput platform for the complete DBTL cycle including multiple iterations (see Box 1.1 for the case study) [9, 10]. Laboratory automation is key to a successful biofoundry, where liquid-handling robots, efficient cultivation devices and high-throughput analytical equipment are the building blocks. Additionally, collection and processing of multi-omics data in an automated and high-throughput manner is important. Genomics and transcriptomics analyses are performed on large scale and cover most of the cellular genome and transcriptome, while analysis of the proteome lags behind [6, 11]. The proteome provides a middle ground between the genotype and phenotype, allowing for a more detailed, grey-box approach to metabolic engineering. Proteins play a major role in many biological processes and accurate, high-throughput protein quantification is therefore crucial for adequate metabolic engineering of microbial cell factories [11].

Box 1.1: Biofoundry case study.

The biofoundry capacity at the Novo Nordisk Foundation Center for Biosustainability (NNF CfB) is applied in an ongoing study to develop an L-tyrosine-overproducing *Escherichia coli* strain. L-tyrosine is a major building block for a variety of natural products, notably products following from the L-DOPA (L-3,4-dihydroxyphenylalanine) node. In the study, rational design is combined with computational tools, including iModulon analysis [12] and kinetic modelling [13], with transcriptomics, proteomics and metabolomics data as input. In total, two rounds of the DBTL cycle will be performed to obtain an *E. coli* L-tyrosine overproducer strain, which will be used in further biofoundry projects as base strain in order to produce a natural product, such as melanin.

In combination with biofoundries and high-throughput multi-omics data generation, computational biology tools within the learn phase pave the way towards efficient development of microbial cell factories [14, 15]. A strong and established learn phase serves a dual purpose: integration of the larger amounts of multi-omics data and more effective information extraction from this data. One approach to multi-omics data integration is kinetic modelling, which combines proteomics, metabolomics and fluxomics data in order to quantify cellular metabolism [13, 16]. Despite their extensive development, cost and relatively small scale, kinetic models provide a targeted and detailed tool for investigating the inner workings of metabolic pathways. Not many pathways have been covered besides central carbon metabolism, hence the potential of kinetic modelling as a tool in the learn phase for the improvement of specific product formation has yet to be realised.

In this thesis, the overall objective was to develop a range of computational tools and analytical methods in order to support effective metabolic engineering of microbial cell factories. The test phase and in particular proteomics analysis is lacking an automated, high-throughput approach and a large part of the current thesis was dedicated to addressing this need. Chapter 2 covers an extensive benchmarking study with the aim of identifying an optimal workflow for proteome-wide absolute quantification. Data-independent acquisition combined with library-free analysis, a standard-free quantification approach and recent protein inference algorithms resulted in the highest number of absolutely quantified proteins for *E. coli.* This complete workflow and all other evaluated workflows were made available in *autoprot*, an automated pipeline tool for absolute quantitative proteomics.

Since membrane proteins are historically underrepresented in proteomics analysis, the work in Chapter 3 proposes an optimised sample preparation protocol for membrane protein extraction from gram-negative bacteria. Additional detergents improved solubilisation and increased representation of membrane proteins within a proteome-wide analysis of *E. coli* and *Pseudomonas putida*. Regarding the learn phase, considerable progress is still required to properly integrate and interpret the increasing amount of omics data generated through metabolic engineering. An established approach to multi-omics data integration is kinetic modelling, which was applied to L-tyrosine overproduction in *E. coli* covered in Chapter 4. The aromatic amino acid biosynthesis pathway was implemented as kinetic model and fitted to experimental data of metabolite concentrations and reaction fluxes. Common metabolic engineering targets for L-tyrosine overproduction were assessed and incorporated in an *E. coli* strain as part of the biofoundry project (see Box 1). Further research will be necessary to evaluate the kinetic model performance and overall success of the metabolic engineering approach. All in all, multiple computational tools and analytical methods, specifically for proteomics analysis and multi-omics data integration through kinetic modelling, were established and are ready to be applied to the development of microbial cell factories.

All additional supplementary materials, which were too large to include in the written thesis, were deposited in a private repository accessed on `https://gitfront.io/r/user-8292993/xzqTPUFmaVbZ/PhDthesisSUP/`. The following additional supplementary materials were deposited: Chapter 2: Table S2.4 - *Escherichia coli* absolute quantitative proteomics data set; Chapter 3: Table S3.1 - *Escherichia coli* membrane protein annotation, Table S3.2 - *Pseudomonas putid*a membrane protein annotation, Table S3.4 - *Escherichia coli* semi-absolute quantitative proteomics, Table S3.5 - *Pseudomonas putida* semi-absolute quantitative proteomics; Chapter 4: Kinetic model input files.

References

[1] M. Eisenstein. "Living factories of the future". In: *Nature* 531.7594 (2016), pp. 401–403. DOI: 10.1038/531401a.

[2] R. Kumar and P. Kumar. "Future microbial applications for bioenergy production: a perspective". In: *Frontiers in microbiology* 8 (2017), p. 450. DOI: 10.3389/fmicb.2017.00450.

[3] X. Liu, W. Ding, and H. Jiang. "Engineering microbial cell factories for the production of plant natural products: from design principles to industrial-scale production". In: *Microbial Cell Factories* 16.1 (2017), p. 125. DOI: 10.1186/s12934-017-0732-7.

[4] J. Nielsen and J. D. Keasling. "Engineering Cellular Metabolism". In: *Cell* 164.6 (2016), pp. 1185–1197. DOI: 10.1016/j.cell.2016.02.004.

[5] C. E. Hodgman and M. C. Jewett. "Cell-free synthetic biology: Thinking outside the cell". In: *Metabolic Engineering* 14.3 (2012), pp. 261–269. DOI: 10.1016/j.ymben.2011.09.002.

[6] E. Marcellin and L.K. Nielsen. "Advances in analytical tools for high throughput strain engineering". In: *Current Opinion in Biotechnology* 54 (2018), pp. 33–40. DOI: 10.1016/j.copbio.2018.01.027.

[7] M. D. Leavell, A. H. Singh, and B. B. Kaufmann-Malaga. "High-throughput screening for improved microbial cell factories, perspective and promise". In: *Current Opinion in Biotechnology* 62 (2020), pp. 22–28. DOI: 10.1016/j.copbio.2019.07.002.

[8] T. Jakočiūnas, M. K. Jensen, and J. D. Keasling. "CRISPR/Cas9 advances engineering of microbial cell factories". In: *Metabolic Engineering* 34 (2016), pp. 44–59. DOI: 10.1016/j.ymben.2015.12.003.

[9] M. B. Holowko, E. K. Frow, J. C. Reid, M. Rourke, and C. E. Vickers. "Building a biofoundry". In: *Synthetic Biology* 6.1 (2020), ysaa026. DOI: 10.1093/synbio/ysaa026.

[10] J. Tellechea-Luzardo, I. Otero-Muras, A. Goñi-Moreno, and P. Carbonell. "Fast biofoundries: coping with the challenges of biomanufacturing". In: *Trends in Biotechnology* 40.7 (2022), pp. 831–842. DOI: 10.1016/j.tibtech.2021.12.006.

[11] F. Calderón-Celis, J. R. Encinar, and A. Sanz-Medel. "Standardization approaches in absolute quantitative proteomics with mass spectrometry". In: *Mass Spectrometry Reviews* 37 (2018), pp. 715–737. DOI: 10.1002/mas.21542.

[12] A. V. Sastry, Y. Gao, R. Szubin, Y. Hefner, S. Xu, D. Kim, K. S. Choudhary, L. Yang, Z. A. King, and B. O. Palsson. "The Escherichia coli transcriptome mostly consists of independently regulated modules". In: *Nature Communications* 10.1 (2019), p. 5536. DOI: 10.1038/s41467-019-13483-w.

[13] P. A. Saa and L. K. Nielsen. "Formulation, construction and analysis of kinetic models of metabolism: A review of modelling frameworks". In: *Biotechnology Advances* 35.8 (2017), pp. 981–1003. DOI: 10.1016/j.biotechadv.2017.09.005.

[14] A. J. Lopatkin and J. J. Collins. "Predictive biology: modelling, understanding and harnessing microbial complexity". In: *Nature Reviews Microbiology* 18.9 (2020), pp. 507–520. DOI: `10.1038/s41579-020-0372-5`.

[15] P. Carbonell, R. Le Feuvre, E. Takano, and N. S. Scrutton. "In silico design and automated learning to boost next-generation smart biomanufacturing". In: *Synthetic Biology* 5.1 (2020), ysaa020. DOI: `10.1093/synbio/ysaa020`.

[16] S. Volkova, M. R. A. Matos, M. Mattanovich, and I. Marín de Mas. "Metabolic Modelling as a Framework for Metabolomics Data Integration and Analysis". In: *Metabolites* 10.8 (2020), p. 303. DOI: `10.3390/metabo10080303`.

Chapter 2.

Automated data analysis for absolute quantitative proteomics - a benchmarking study in *Escherichia coli*

Shannara Kayleigh Taylor Parkins, Nicolás Gurdo, Tune Wulff, Pablo Iván Nikel and Lars Keld Nielsen

Absolute quantification of proteins in an accurate and precise manner is desirable for a range of applications: from drug development to systems biology. However, this remains challenging and many strategies exist for absolute quantitative proteomics, without consensus on an optimal workflow. Here, we developed a versatile tool, *autoprot*, for proteome-wide, automated data analysis of raw mass spectrometry files with intracellular protein concentrations as result. The *autoprot* tool allowed for an extensive benchmarking study performed on *Escherichia coli* proteomics data, where the optimal workflow identified and absolutely quantified 2,802 unique proteins. Application of a standard-free quantification approach using the TPA (total protein approach) method, library-free DIA (data-independent acquisition) data analysis using DIA-NN and the LFAQ protein inference algorithm, all increased the quantity, precision and accuracy of the final absolute proteomics data set relative to more traditional strategies. These results highlight the advantages of current state-of-the-art methods and, combined with the *autoprot* tool, provide an improved workflow for absolute quantitative proteomics. Data are available via ProteomeXchange with identifier PXD043377.

2.1. Introduction

Quantitative proteomics is an important part of modern analytical biochemistry that connects genomic perturbations with phenotypic changes. Absolute rather than relative quantitative proteomics is on the rise, since it is required for certain applications, such as comparison of different proteins and different data sets, biomarker quantification for drug development or appropriate input for computational models of cell metabolism [1, 2]. Ultimately, accurate proteome-wide absolute quantification would be desirable; however this remains difficult due to the limitations of current protein identification and quantification methods. Although mass spectrometry (MS)-based methods have led to high-throughput quantification of higher quality and of more proteins, challenges remain with the conversion of MS signals into protein concentrations [1, 3].

A variable relationship exists between the measured intensity from MS analysis and the absolute abundance of a peptide due to incomplete proteolysis, inefficient ionisation, or suppression from co-eluting peptides [1]. A common approach to account for this variability is the addition of internal standards to the biological samples. These internal standards should mimic the proteins of interest as closely as possible to increase the accuracy of the final quantification. Internal standards can be either labelled or unlabelled, where labelling allows the use of endogenous proteins as the internal standard. While stable isotope labelling (SIL) is most prevalent, all labelling substantially increases the production cost of the internal standard [4]. Different variants of standards are available, ranging from short peptides (AQUA [5]) to concatenated peptides (QconCATs [6]) as well as partial proteins (QPrEST [7, 8]) and full-length proteins (PSAQ [9]). While the addition of protein standards provides more sequence coverage and takes proteolysis efficiency into account, peptide internal standards are more cost effective to produce than proteins [9, 10].

Inference of protein intensities based on properties of the corresponding peptides also provides a way to compensate for the variable signal response factor. Multiple algorithms are available, each exhibiting distinct approaches to protein inference, such as TopN [11, 12, 13], iBAQ [14], NSAF [15], SCAMPI [16], APEX [17], emPAI [18], MaxLFQ [19], xTop [20], and LFAQ [21]. Most of these algorithms calculate the intensity of a protein by a normalised average, based on a selection of identified peptides of the corresponding protein. Older algorithms are quite simple, e.g. TopN and iBAQ, whereas newer algorithms, such as xTop and LFAQ, take more complex peptide properties into account. The performance of these algorithms has been assessed in benchmarking studies previously [22, 23, 24, 25, 26], although mostly on artificial samples where two or more proteomes are mixed with known ratios.

There is no consensus on the optimal approach for absolute quantification of proteins [1, 27, 28]. This may be due to differences between organisms or growth conditions, which limit any one-size-fits-all approach. Proteome-wide absolute quantification can be performed with a labelled standard, a label-free standard, or a standard-free approach.

Using labelled, e.g. SIL AQUA peptides, or label-free standards, e.g. UPS2 protein mix (Proteomics Dynamic Range Standard Set, Merck), the proteome composition is calculated by linear regression on the log-transformed intensity and known concentration of the standards. The total protein approach (TPA) [29, 30] is a standard-free approach, where the proteome composition is calculated based on the ratio of individual protein intensity to the sum of all protein intensities. While implementation of internal standards is expensive and time consuming in terms of production, sample preparation and data analysis, it is perceived as a more accurate method over the standard-free approach.

The full workflow of absolute quantitative proteomics has been performed on multiple microbes and under a variety of conditions (Table 2.1). Combinations of different protein inference algorithms and absolute quantification approaches have been benchmarked [22, 23, 31], where the most common combination for high-throughput application is the TopN or iBAQ algorithm combined with the UPS2 protein mix [31, 32, 33, 34]. While data-dependent acquisition (DDA) and data-independent acquisition (DIA) can both be used for proteome-wide absolute quantification [35, 36], DIA-based workflows dominate the field, especially SWATH-MS DIA [37]. For data processing, a high-throughput, standardised and automated workflow to retrieve protein concentrations from raw spectral data has not yet been fully developed.

The aim of this study was to benchmark several methods for each step in the full absolute quantitative proteomics workflow, in order to determine the optimal combination. We deployed three approaches for absolute quantification of *E. coli* proteins during growth on a range of substrates: a labelled standard approach with 19 *Escherichia coli* QconCATs [38], a label-free standard approach with the UPS2 protein mix, and a standard-free approach with the TPA method. Both DDA and DIA methods were operated for MS acquisition, where the appropriate software tool, ProteomeDiscoverer (Thermo Fisher Scientific) for DDA data and Spectronaut (commercial, Biognosys) or DIA-NN (open source, [39]) for DIA data, was applied for identification and precursor quantification. Six algorithms were assessed for the inference of peptide to protein intensities: TopN, iBAQ, APEX, NSAF, xTop and LFAQ. To address the need for standardisation, we developed an automated data analysis workflow called *autoprot*. This novel tool automates the complete workflow for absolute quantitative proteomics, starting from raw MS files and producing estimated protein concentrations using a configurable combination of quantification approaches, analysis software and protein inference algorithms (Figure 2.1).

Table 2.1: Comparison of proteome-wide absolute quantification workflows from literature. An elaboration on the calculation of intracellular protein concentrations is provided in the Discussions section.

Organism	Identification and relative quantification	Protein inference algorithms	Absolute quantification approach	Intracellular concentrations	Reference
E. coli	DDA (MaxQuant)	iBAQ, APEX and emPAI	Label-free (UPS2) and standard-free (TPA)	Cellular protein density (with total protein and biomass concentrations)	[22]
E. coli, *A. gossypii*, *S. pombe*, *M. pneumonia*, *L. interrogans* and *D. melanogaster*	DDA (Progenesis + MASCOT)	TopN, iBAQ, APEX and emPAI	Labelled (AQUA), label-free (UPS2) and standard-free (TPA)	Not applied	[23]
M. tuberculosis	Library-based DIA (OpenSWATH)	TopN, iBAQ, APEX and NSAF	Labelled (AQUA) and label-free (UPS2)	Cellular protein density (constant) and cell volume	[40]
E. coli	DDA (Progenesis + MASCOT)	iBAQ	Labelled (AQUA)	Cellular protein density (with internal standards) and cell volume)	[41]
S. cerevisiae	DDA (MaxQuant)	iBAQ	Labelled and label-free (UPS2+SILAC)	Cellular protein density (with total protein and biomass concentrations)	[32]
E. coli	DDA (PD)	Top3	Label-free (UPS2)	Cellular protein density (constant)	[33]
S. cerevisiae	DDA (MaxQuant)	iBAQ	Labelled and label-free (UPS2+SILAC) and standard-free (TPA)	Not applied	[31]
E. coli	Library-based DIA (OpenSWATH)	TopN, iBAQ and xTop	Labelled (AQUA) and standard-free (TPA)	Cellular protein density (constant) and ribosome profiling	[20]
C. autoethanogenum	Library-based DIA (Skyline)	TopN and iBAQ	Labelled (PSAQ)	Cellular protein density (with internal standards)	[42]
S. cerevisiae	DDA (MaxQuant)	iBAQ	Labelled and label-free (UPS2+SILAC)	Cellular protein density (with total protein and biomass concentrations)	[34]
E. coli	DDA (PD), library-based and library-free DIA (SN and DIA-NN)	TopN, iBAQ, APEX, NSAF, xTop and LFAQ	Labelled (QconCAT), label-free (UPS2) and standard-free (TPA)	Cellular protein density (with total protein and biomass concentrations)	This study

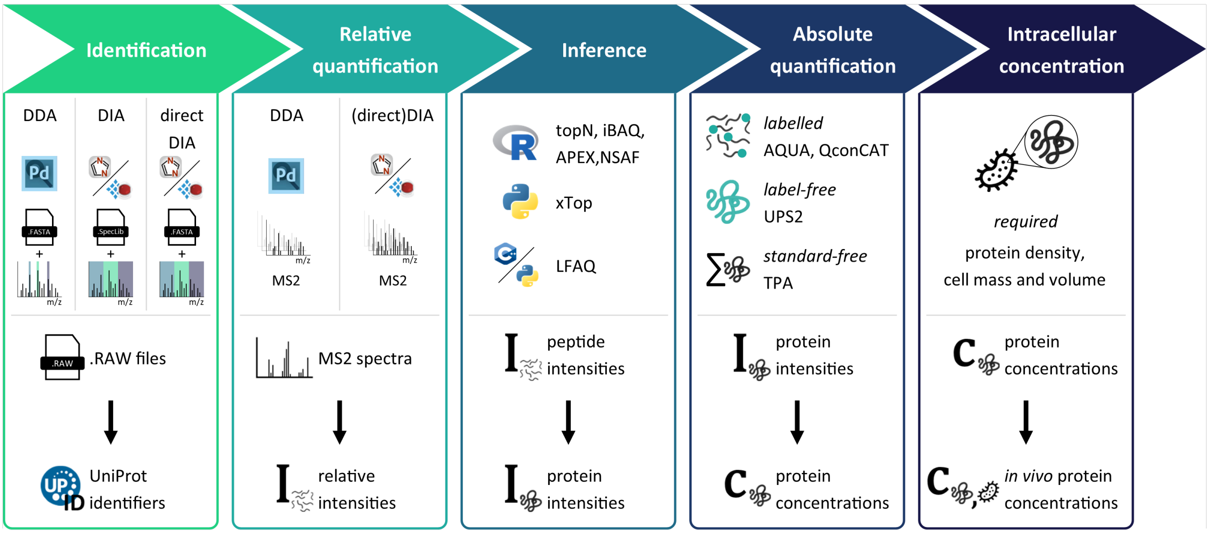


Figure 2.1: Overview of the *autoprot* pipeline. In the identification and relative quantification steps, peptides are identified and quantified by DDA, library-based DIA or library-free DIA data analysis from DDA or DIA raw MS files. The DIA data analysis is performed with either Spectronaut or DIA-NN. Next, relative protein intensities are inferred from the relative peptide intensities using a protein inference algorithm. The implemented protein inference algorithms are TopN, iBAQ, APEX, NSAF, xTop and LFAQ. In the absolute quantification step, the protein concentrations, i.e. proteome composition, are calculated with a labelled, label-free or standard-free quantification approach. Intracellular protein concentrations are calculated from the proteome composition using cellular protein density values.

2.2. Material and methods

2.2.1. Bacterial strains and culture conditions

Escherichia coli MG1655 was cultivated in lysogeny broth (LB) medium, M9 minimal medium with 4 g/L glucose or M9 minimal medium with 2 g/L glycerol, to increase proteome diversity within the experiment. Each culture flask was inoculated from LB agar plates with colonies of the corresponding organism and grown overnight into stationary phase, where the optical density (OD) at 600 nm was noted as OD_{600}^{max}. Culture flasks with fresh media were inoculated from the corresponding stationary culture and grown to 75% of OD_{600}^{max}. This transfer and subsequent growth step was repeated twice in order to minimise the lag phase in the final culture. The final culture flask was inoculated from the corresponding transfer culture and grown to 25% of OD_{600}^{max} before sampling for proteomics. Either 1 or 2 mL was sampled from the cultures grown in LB and M9 minimal glycerol medium or M9 minimal glucose medium, respectively. Samples were immediately centrifuged at 15,000 g at -5°C for 5 minutes, the supernatant was discarded and the remaining cell pellets were kept at -80°C until further processing.

2.2.2. Proteomics sample preparation

Sample preparation for proteomics analysis was performed as described previously by Kozaeva *et al.* [43] and any modifications are mentioned below. Cell pellets of *E. coli* were lysed in 6 M guanidinium·HCl, 5 mM *tris*(2-carboxyethyl)phosphine, 10 mM chloroacetamide and 100 mM Tris·HCl (pH = 8.5), disrupted mechanically and heated to 99°C. After centrifugation, the cell-free lysates were diluted with 50 mM ammonium bicarbonate and subjected to bicinchoninic acid (BCA) assay to estimate protein concentrations. Trypsin and LysC digestion mix (Promega) was added to 20 µg protein of each sample and incubated for 8 hours. Trifluoroacetic acid was added to halt digestion and the samples were desalted using C18 resin (Empore, 3M) before HPLC-MS analysis.

The *E. coli* QconCAT proteins were designed by Batth *et al.* [38] and the corresponding expression plasmids (Addgene) were used to produce 18 QconCAT proteins. An additional QconCAT protein containing peptides from aromatic amino acid biosynthesis enzymes was designed by Batth *et al.* (2014) [38]. Since the corresponding expression plasmid was not available, the plasmid was designed following Gurdo *et al.* [44] and ordered separately from IDT. All SIL QconCAT proteins were expressed, purified and quantified as described by Gurdo *et al.* [44]. The QconCAT proteins were labelled with ^{13}C(6)-L-arginine and ^{13}C(6)-L-lysine to ensure solely SIL QconCAT peptides after tryptic digestion. During sample preparation, QconCAT proteins in varying amounts were added to the samples preceding digestion (see Table S2.1 for full details). Since the *E. coli* QconCAT proteins were not developed in-house, a dilution series was performed to identify peptides with a consistent response factor on the current HPLC-MS setup. Other QconCAT peptides were considered non-representative. The dilution series was executed in duplicate with the DDA method and spanned six different concentration levels for each of the QconCAT proteins. Non-representative and duplicate QconCAT peptides were removed during data analysis.

One tube of the UPS2 proteomics dynamic range standard set (Sigma Aldrich) was dissolved in 40 µL of lysis buffer (6 M guanidinium·HCl, 5 mM *tris*(2-carboxyethyl) phosphine, 10 mM chloroacetamide and 100 mM Tris·HCl (pH = 8.5)) and 20 µL of this UPS2 protein mix was added to each sample preceding digestion (see Table S2.2).

2.2.3. HPLC-MS analysis

HPLC-MS analysis of the samples was performed on an Orbitrap Exploris 480 instrument (Thermo Fisher Scientific) preceded by an EASY-nLC 1200 HPLC system (Thermo Fisher Scientific). For each sample, 1 µg of peptides was captured on a 2-cm C18 trap column (Thermo Fisher 164946). Subsequently separation was executed using a 70 minute gradient from 8% (v/v) to 48% (v/v) of acetonitrile in 0.1% (v/v) formic acid on a 15-cm C18 reverse-phase analytical column (Thermo EasySpray ES904) at a flow rate of 250 nL/min. The mass spectrometer was operated in either data-dependent or data-independent acquisition mode with the specific settings listed below.

DDA
For data-dependent acquisition, the mass spectrometer was run with a DD-MS2 method preceded by the FAIMS Pro Interface (Thermo Fisher Scientific) with alternating CV of -50 V or -70 V. Full MS1 spectra were collected at a resolution of 60,000 and scan range of 375-1,500 m/z, with the maximum injection time set to auto, an intensity threshold of $5.0 \cdot 10^3$, and a dynamic exclusion at 45 s. MS2 spectra were obtained at a resolution of 15,000, an isolation window of 1.6 m/z, the HCD collision energy set to 28%, the maximum injection time set to auto.

DIA
For data-independent acquisition, the mass spectrometer was run with the HRMS1 method as previously described [45] preceded by the FAIMS Pro Interface (Thermo Fisher Scientific) with a compensation voltage (CV) of -45 V, and any modifications are mentioned below. Full MS1 spectra were collected at a resolution of 120,000 and scan range of 400-1,000 m/z, with the maximum injection time set to auto. MS2 spectra were obtained at a resolution of 60,000, with the maximum injection time set to auto and the collision energy set to 32. Each cycle consisted of three DIA experiments each covering a range of 200 m/z with a window size of 6 m/z and a 1 m/z overlap, while a full MS scan was obtained in between experiments.

2.2.4. Protein identification and relative quantification

For all analyses, sequence identification was performed using a protein database consisting of the *E. coli* (UP000000625) reference proteome. The UPS2 protein sequences were appended to the protein database for data analysis of samples including the UPS2 protein mix. The specific settings for DDA and DIA data analysis can be found in the Supplementary materials section.

2.2.5. Protein intensity inference

The protein identification and relative quantification steps generated a precursor-based results table, which served as input for the inference of protein intensities. For result tables from samples including QconCAT proteins, the data corresponding to precursors with ^{13}C(6)-labelling was removed and stored in a separate table for further analysis, in advance of protein inference. The precursor-based results table was converted into the OpenSWATH format [46] to increase accessibility and comparability.

The TopN, iBAQ, APEX and NSAF protein inference algorithms were accessed through the aLFQ R package [47]. The precursor-based input was imported as OpenSWATH MS file type, with the FDR cut-off set to 1%, and "remove decoys" enabled. The protein inference function was executed for each algorithm using the following settings: the top 5 transitions were summed with strictness set to loose, both precursor types and modification types were summed to retrieve peptide intensities. For the TopN algorithm, both Top all and Top3 were used with the peptide summary set to mean and the peptide strictness set to loose.

The same protein database from the protein identification and relative quantification was used for the iBAQ, APEX, and NSAF algorithms.

Before protein inference using the xTop [20] and LFAQ [21] algorithms, precursor intensities were summed based on identical sequence to obtain peptide intensities and filtered with an FDR cut-off of 1%, similar to the aLFQ R package import functionality. The Python-based version of the xTop algorithm was used with the default settings. Since the xTop algorithm only considers peptides identified in three or more samples, peptides identified in fewer than three samples were filtered out in advance. A custom Python script was developed to interact with the LFAQ algorithm from the command line. The LFAQ input was split per sample and the LFAQ output per sample was combined afterwards, since the LFAQ algorithm only performs protein inference for one sample at a time. Only the protein intensities were used for further analyses; however the LFAQ algorithm only works when protein concentrations were calculated as well. To circumvent this issue, a file with random protein identifiers, drawn from the protein database, and random concentrations was constructed and used as additional input without affecting the calculated protein intensities.

2.2.6. Proteome-wide absolute quantification

Proteome-wide absolute quantification was carried out using three approaches: linear regression with either labelled or label-free internal standards, or the standard-free TPA method [29, 30]. A custom Python script was developed to carry out all three quantification approaches. The calculated protein intensities from any of the inference algorithms were used as input, together with a table recording total protein mass per sample. For analyses using internal standards, a table with either peptide sequences or protein identifiers and known concentrations was provided as additional input. The protein database was added as input for analyses using the TPA method. A detailed description of the quantification approaches can be found in the Supplementary materials section.

The estimated proteome composition values were converted into intracellular protein concentrations using the cellular protein density, which was calculated for each sample using total protein and biomass measurements and set value for the cell volume of $3.9 \cdot 10^{-15}$ L for *E. coli* MG1655 [48]. The cellular protein density can be taken from literature or calculated from total protein concentrations, cell volume and biomass measurements, thus *autoprot* allows for a specific cellular protein density value for each sample as input.

2.2.7. *E. coli* subunit stoichiometries

Expected subunit stoichiometries of *E. coli* protein complexes were taken from the Complex Portal database [49]. For the purpose of developing the metric, only protein complexes with two subunits in equimolar stoichiometries were kept (see Table S2.3 for the full list of protein complexes). The goodness-of-fit between all experimental and theoretical, i.e. equimolar, subunit concentrations found in a specific analysis was determined as the inverse of the mean square error (MSE).

2.2.8. Construction of automated pipeline - *autoprot*

To increase efficiency and accessibility, the full data analysis for proteome-wide absolute quantification was developed into an automated pipeline. The *autoprot* pipeline is based on PowerShell and can be executed with a single command. The input to *autoprot* consists of the raw MS files, the acquisition mode as string, the absolute quantification approach as string, the experiment name as string and the protein database. Additionally, a spectral library file is required as input for library-based analyses and a table with internal standard concentrations is required as input for analyses including absolute quantification using internal standards. The output is a single protein-based table for each inference algorithm applied, recording the estimated protein composition values. The full documentation for *autoprot* can be found in the Supplementary materials section.

2.2.9. Data availability

The mass spectrometry proteomics data of the optimal workflow have been deposited to the ProteomeXchange Consortium (`http://proteomecentral.proteomexchange.org`) via the PRIDE [50] partner repository with the data set identifier PXD043377. The *autoprot* pipeline code including examples of input and output files have been made available on GitHub (`https://github.com/biosustain/autoprot`). The proteome-wide quantification results obtained with the optimal workflow can be found in Table S2.4.

2.3. Results

2.3.1. Experimental overview

Samples of the gram-negative bacterium *E. coli* were used in the current study, where three biological replicates of each condition provided a data set of appropriate size for extensive benchmarking (Figure 2.2A). To ensure sufficient differences of biological significance in the proteome compositions, *E. coli* was grown under three different conditions: rich medium and minimal medium with either glucose (glycolytic growth regime) or glycerol (gluconeogenic growth regime). The total of 9 biological samples were each split into three technical replicates, one for each quantification approach. Both a DDA and a DIA method were performed during HPLC-MS analysis for all 27 samples, i.e., a total of 54 injections, to assess the impact of acquisition mode on the absolute protein quantification. The influence of spectral library usage and protein inference algorithm implementation was evaluated by altering steps in the data analysis of the raw MS files.

Overall, the experimental data were highly reproducible with a median Pearson correlation coefficient of 0.96 for biological replicates (Figure 2.2B). As expected, the data of minimal medium cultivations with either carbon source shared more similarities, compared to data of rich medium cultivations. A maximum of 2,802 unique *E. coli* proteins (64% of the 4,403 annotated *E. coli* proteins) were identified and relatively quantified, where 74% of identified proteins were detected with three or more peptides (Figure 2.2C).

Protein-specific coefficient of variation (CV) values were calculated for the three biological replicates per condition, as a measure of precision of the biological experiment (Figure 2.2D). Although these CV values were acceptable for further analysis, manual operation of cultivation and sampling likely decreased the experimental precision, which could be improved upon by laboratory automation [51]. An overall strong correlation between estimated protein concentrations and expected equimolar stoichiometries of protein complexes was established, which indicates high accuracy of the experimental data (Figure 2.2E).

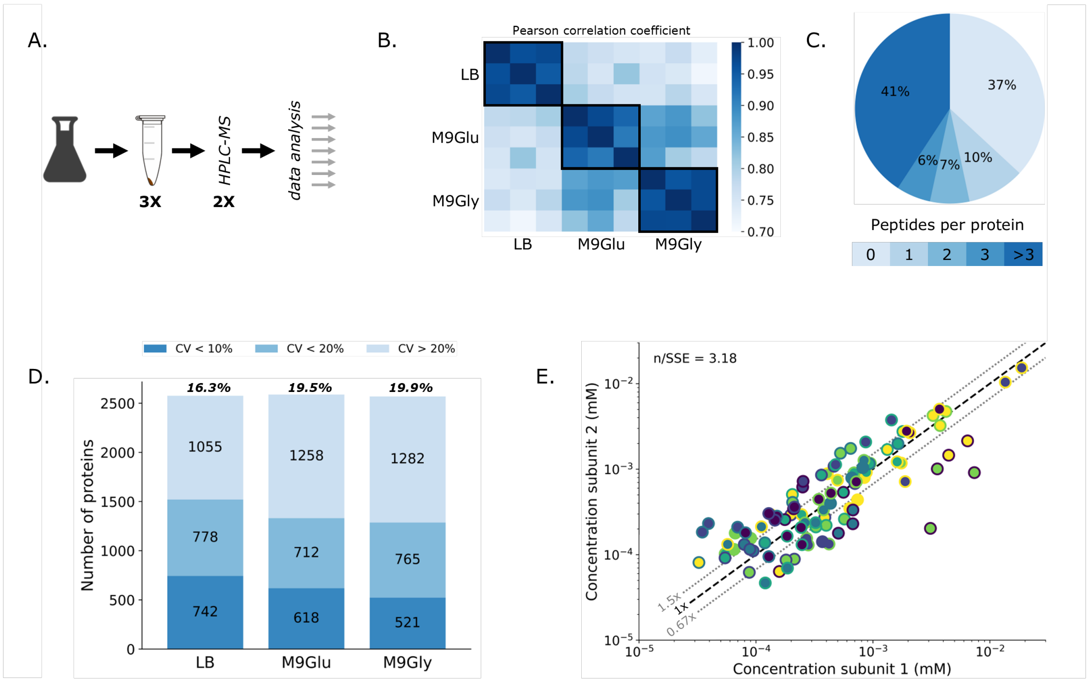


Figure 2.2: Experimental overview of the benchmarking study. All graphs are based on the dataset generated with a workflow which combined standard-free, library-free DIA data analysis using DIA-NN with either the xTop or the LFAQ protein inference algorithm. (A) Workflow deployed to generate a total of 54 HPLC-MS injections. (B) Pearson correlation coefficients of protein concentrations in all 9 samples compared against each other. Black boxes indicate comparisons between biological replicates. (C) Distribution of detectable peptides per protein compared to the total protein database. (D) Distribution of coefficient of variation (CV) values for each condition with the median value displayed on top. (E) Correlation between protein concentrations of equimolar protein complexes. Data points of the same colour scheme represent the same protein complex under different experimental conditions.

2.3.2. Workflow development

We developed an automated tool for absolute quantitative proteomics, *autoprot*, in order to benchmark various methods and identify the optimal workflow for a specific experiment (Figure 2.1). The default settings were used for each tool and only adapted when necessary, e.g. in order to accommodate ^{13}C(6)-labelling or to increase comparability. To improve compatibility between the software tools and protein inference algorithms applied, a universal format for the precursor- and protein-based tables was used.

The format was adapted from the OpenSWATH format [46] and all outputs were converted into this format. Future workflows based on tools from different sources would benefit greatly from a standardised format for proteomics data.

Both high precision and high accuracy are relevant for absolute quantitative proteomics, while a high number of unique proteins is also desirable in order to increase the information gained from a specific experiment. For this benchmarking study, two performance metrics were developed: one reflecting precision and one reflecting accuracy. The precision metric was based on the inverse of the 1000th CV value of a workflow data set in favour of the median CV value to avoid penalisation of higher detection rates, following Equation 2.1.

$$\tau = \frac{1}{CV_{1000}} \tag{2.1}$$

In Equation 2.1, τ is the precision of a given workflow data set and CV_{1000} is the 1000th CV value of a given workflow data set. For the accuracy metric, the goodness-of-fit of estimated protein concentrations from a workflow data set to the expected equimolar stoichiometries of certain protein complexes was calculated as the inverse of the mean square error (MSE), following Equation 2.2.

$$\eta = \frac{N}{SSE} \tag{2.2}$$

In Equation 2.2, η is the accuracy of a given workflow data set, N is the total number of quantified protein complexes, and SSE is the sum of squares error.

2.3.3. Full comparison

For the benchmarking, all possible combinations of different quantification approaches, acquisition modes, and protein inference algorithms were applied to establish the optimal workflow for absolute quantitative proteomics of the current experimental data set. The τ precision and η accuracy metrics were determined for each workflow and the relative comparison of all workflows highlights the challenges of balancing both precision and accuracy within one specific workflow (Figure 2.3).

The largest number of unique proteins was found using standard-free and library-free DIA data analysis with DIA-NN and either the LFAQ or the xTop inference algorithm, while the most precise data was found using standard-free DDA data analysis using xTop (Figure 2.3A). The difference is either a 20% increase in quantity or a 146% increase in τ precision and the choice of workflow would highly depend on the experimental objective. For proteome-wide analysis, the former would be more desirable, since an increase in quantity, i.e. a higher total number of proteins, provides more information. The workflows resulting in the highest number of quantified proteins all incorporated the LFAQ protein inference algorithm. The additional proteins quantified by a workflow applying LFAQ compared to the other protein inference algorithms were mostly in the lower concentration range, when quantified by otherwise identical workflows (Figure S2.1A).

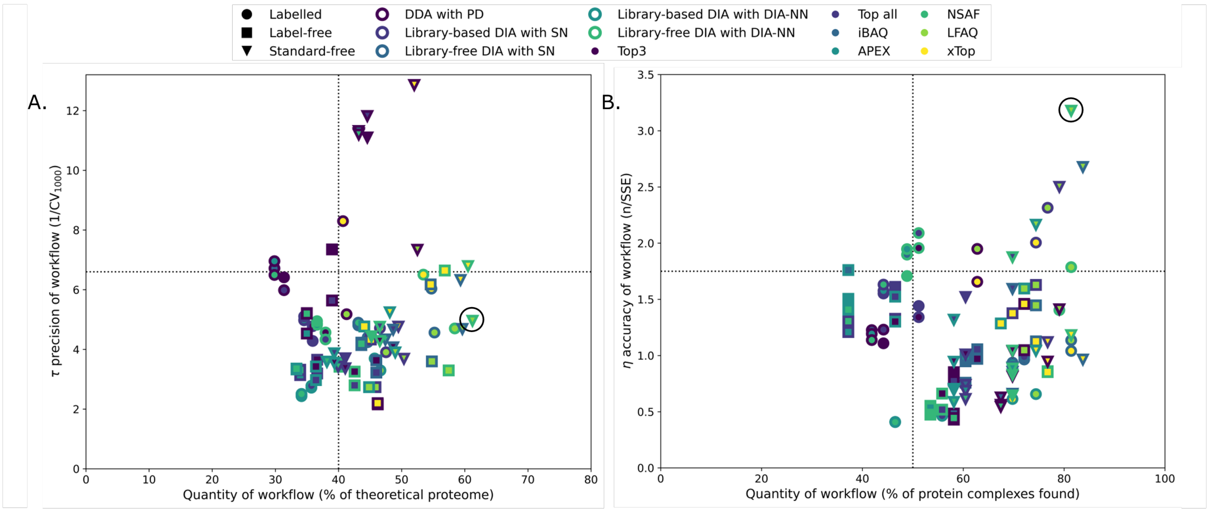


Figure 2.3: Full comparison of the data sets generated using all 105 different workflows available through the *autoprot* pipeline. The optimal workflow, standard-free, library-free DIA data analysis using DIA-NN and LFAQ, is highlighted with a circle. (A) Full comparison of the τ precision metric. The quantity values were determined as the percentage of unique *E. coli* proteins absolutely quantified compared to the total number of *E. coli* proteins in the protein database. The τ precision was determined as the inverse of the 1000$^{\text{th}}$ coefficient of variation (CV) value of a workflow data set. (B) Full comparison of the η accuracy metric. The quantity values were determined as the percentage of *E. coli* protein complexes for which both subunits were absolutely quantified compared to the total number of protein complexes. The η accuracy was determined as the inverse of the mean square error (MSE) of the subunit concentrations fit to the theoretical, i.e. equimolar concentrations, normalised to the total number of protein complexes quantified.

Proteins with low abundance are challenging to identify and subsequently quantify, however they often perform a crucial biological function, e.g. transporter proteins [52].

For the accuracy metric, the optimal workflow combined standard-free and library-free DIA data analysis with DIA-NN and LFAQ, which produced the highest η accuracy and quantified the second highest number of protein complexes (Figure 2.3B). An additional quantified protein complex was obtained by applying a workflow that combines standard-free, library-free DIA data analysis with Spectronaut and LFAQ, however this led to a 19% decrease in η accuracy. The data set produced by implementing the optimal workflow consisted of 2,802 unique *E. coli* proteins with a median CV value of 25%. Ultimately, the biological accuracy, here reflected by the η accuracy metric, and the quantity are the most imperative for proteome-wide analysis and subsequent applications.

To assess the full capability of the *autoprot* pipeline, the raw MS files of 7 calibration samples from Mori *et al.* [20] were used as input and the resulting data set was compared to the processed data set of Mori *et al.* (Figure S2.2). The authors developed and deployed an extensive spectral library for the proteome-wide analysis of *E. coli* samples. Their workflow combined library-based DIA data analysis using OpenSWATH with the xTop protein inference algorithm and a standard-free quantification approach [20].

When the original data set was compared to the data set generated using the *autoprot* pipeline with standard-free and library-free DIA data analysis using DIA-NN and xTop, the reproduced data set achieved a 9% increase in the number of quantified proteins while maintaining an equivalent precision and accuracy for shared proteins (Figure S2.2A and S2.2B). The proteome composition of the reproduced data set was highly similar to the original data set for shared proteins (Figure S2.2C) and the additional proteins quantified resided in the lower concentration range (Figure S2.2D). When the xTop algorithm was swapped for LFAQ for the reproduction, the number of quantified proteins increased by 11% and the precision dropped by 8% for shared proteins (Figure S2.2E and S2.2F). The raw MS format was different than used in the current study, .raw from Thermo Fisher Scientific Orbitrap Exploris 480 instrument compared to .wiff from Sciex TripleTOF5600 instrument, nonetheless data analysis with *autoprot* was easily performed due to the input format flexibility of DIA-NN.

2.3.4. Quantification approach comparison

To facilitate absolute quantification, three different approaches were implemented in the *autoprot* pipeline: application of ^{13}C(6)-labelled QconCAT proteins, label-free UPS2 protein mix or standard-free TPA method. For both the τ precision and the η accuracy, the workflows incorporating the standard-free TPA method achieved the highest quantity (proteins or protein complexes quantified) overall (x-axis, Figure 2.4A and 2.4D). This may be due to the substantial portion of the total injected protein mass occupied by the internal standards for the other two approaches. Both the *E. coli* QconCAT proteins and the UPS2 protein mix occupied approximately 20% of the total protein mass injected in the mass spectrometer for the current study, similar to application of the UPS2 protein mix in other studies [47, 40]. Since internal standards are usually added in high abundance, the increased number of proteins with high concentrations may repress the MS signals of endogenous proteins with low abundance. The additional proteins gained from switching to a workflow applying the standard-free TPA method instead of applying the labelled or label-free method were indeed residing in the lower concentration range (Figure S2.1B and S2.1C).

The sections of high τ precision and high η accuracy were also dominated by workflows incorporating the standard-free TPA method (Figure 2.4A and 2.4D). The linear regression performed within the labelled and label-free approaches solely estimates the concentrations of proteins for which an internal standard is not available, thus the estimates are as reliable as the regression fit. The *E. coli* QconCAT proteins and UPS2 protein mix carried considerable noise which decreased the regression fit (Figure S2.3). For the *E. coli* QconCAT proteins, the noise was likely due to contaminant proteins from the background strain used for production, i.e. *E. coli* BL21. A potential solution would incorporate a smaller set of QconCAT proteins optimised for the specific HPLC-MS setup and produced in a different organism, so that background contaminant proteins may be filtered out during data analysis.

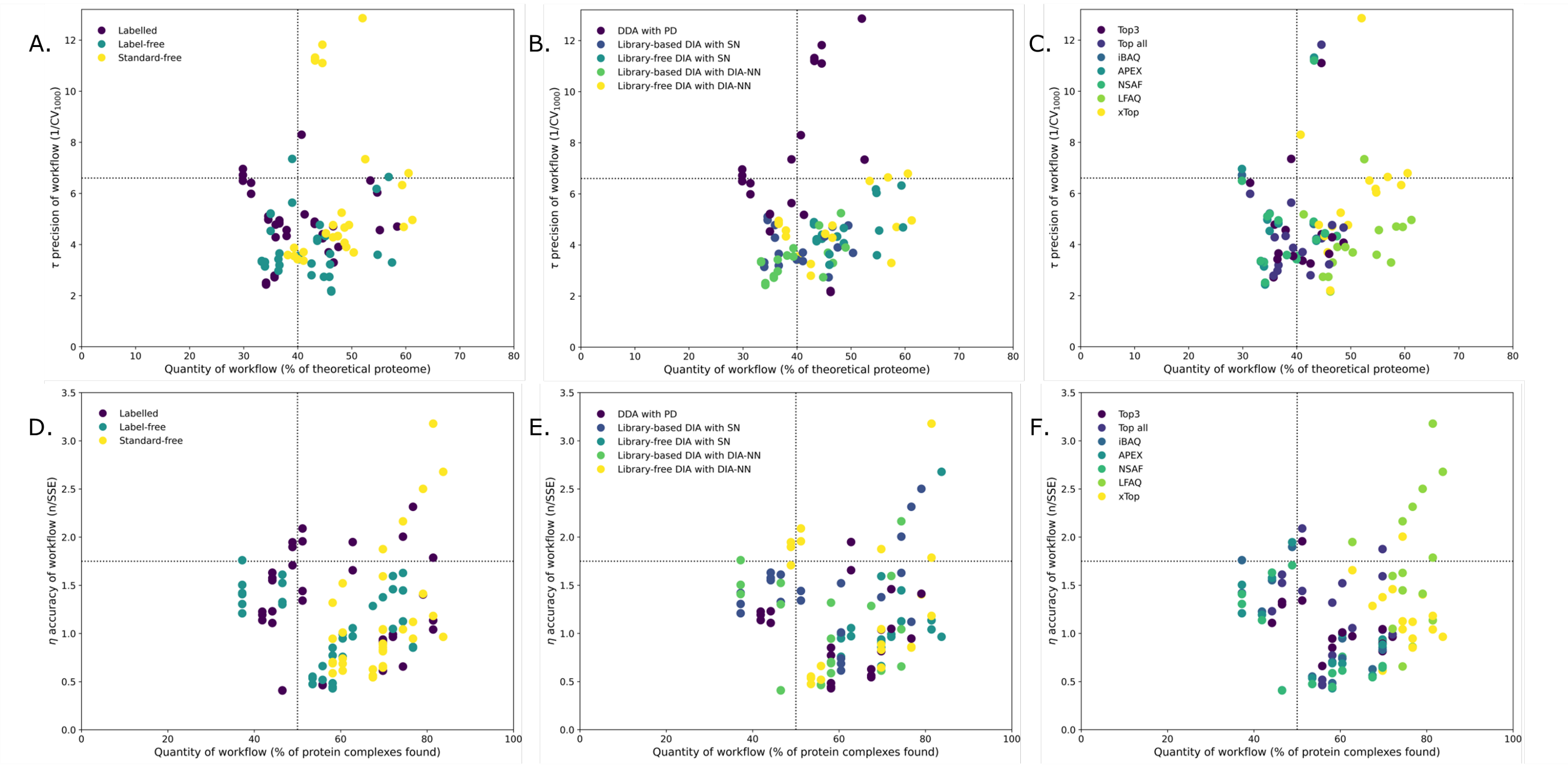


Figure 2.4: Detailed comparison of the data sets generated using all 105 different workflows available through the *autoprot* pipeline. For the τ precision, the quantity values were determined as the percentage of unique *E. coli* proteins absolutely quantified compared to the total number of *E. coli* proteins in the protein database. The τ precision was determined as the inverse of the 1000th coefficient of variation (CV) value of a workflow data set. For the η accuracy, the quantity values were determined as the percentage of *E. coli* protein complexes for which both subunits were absolutely quantified compared to the total number of protein complexes. The η accuracy was determined as the inverse of the mean square error (MSE) of the subunit concentrations fit to the theoretical, i.e. equimolar concentrations, normalised to the total number of protein complexes quantified. (A) τ precision comparison for the absolute quantification approaches. (B) τ precision comparison for the acquisition modes. (C) τ precision comparison for the protein inference algorithms. (D) η accuracy comparison for the absolute quantification approaches. (E) η accuracy comparison for the acquisition modes. (F) η accuracy comparison for the protein inference algorithms.

Notably, the proteins for which an internal standard was available were more precisely quantified, i.e. had lower median CV values compared to proteins for which not internal standard was available (Figure S2.4). The noise of the UPS2 protein mix stemmed from the limited number of concentration levels available within the mix, i.e. 6 levels. Multiple proteins in the UPS2 protein mix have the same spike-in concentration, however their signal response factor is quite different which poses challenges for a good regression fit (Figure S2.3B). Remarkably, the application of library-free DIA data analysis and xTop or LFAQ in combination with the label-free approach enabled protein quantification of all 6 levels within the UPS2 protein mix, compared to the 4 levels quantified in previous studies [23, 40] using identical total spike-in amounts.

As expected, the total protein mass equalled 1 µg for all workflows implementing a standard-free quantification approach, since the TPA method normalises to unity (Figure S2.5A, S2.5D and S2.5G). The additional protein mass of internal standards within the labelled and label-free approaches was reflected in the total protein mass exceeding 1 µg for most samples (Figure S2.5). This effect was less pronounced for the DDA data analysis workflows, due to fewer proteins and thus less protein mass quantified with DDA methods. For the label-free approach using the UPS2 protein mix, the total protein mass with and without the internal standards could be determined separately and showed an average 20% increase when internal standards were accounted for as anticipated (Figure S2.5C, S2.5F and S2.5I). Overall, the total protein mass per sample resided between 1 and 1.5 µg for workflows applying a quantification approach including internal standards.

2.3.5. Acquisition mode comparison

The *autoprot* pipeline covers DDA and DIA data analysis, with the latter supported through library-based and library-free approaches. To broaden the application of the *autoprot* pipeline, DIA-NN was added for DIA data analysis to enable the full pipeline to be run using only open-source software. The five most precise workflows, i.e. producing high τ precision values, were based on DDA analysis (Figure 2.4B). Relatively low median CV values are expected for proteins quantified in DDA analysis, due to the highly confident identification from the single-peptide MS2 fragmentation. The high τ precision values are however offset by a decrease in the number of quantified proteins and a considerable loss in η accuracy values. DDA methods would be more appropriate for experimental analyses concerning a smaller number of higher abundance proteins, e.g. production of specific proteins, compared to proteome-wide analyses.

The workflows applying DIA compared to DDA data analysis achieved higher numbers of unique proteins as well as protein complexes (Figure 2.4B and 2.4E). The workflows deploying library-free DIA data analysis attained the highest number of quantified proteins, regardless of using Spectronaut or DIA-NN. Since neither DIA analysis nor library-free data analysis select for specific peptides, more proteins are identified and subsequently quantified compared to DDA analysis and library-based data analysis, respectively.

It should be noted that library-based DIA data analyses are limited by the applied spectral library, i.e. a specific protein could only be identified and quantified if it was present in the spectral library. The spectral libraries of the current study were based on DDA data of the corresponding samples, whereas spectral libraries are more commonly based on a large number of samples including fractionated samples to increase the proteome coverage [53]. Deployment of the spectral library developed by Mori *et al.* [20] did not produce higher τ precision or η accuracy compared to the other library-based workflows applied to samples of the current study (Figure S2.6), despite achieving higher quantities for both the current samples and the calibration samples of Mori *et al.* (Figure S2.2A).

2.3.6. Protein inference algorithm comparison

The older TopN, iBAQ, APEX and NSAF protein inference algorithms have been applied repeatedly [23, 40, 42] and these combined with the newer xTop and LFAQ protein inference algorithms were incorporated in the *autoprot* pipeline for benchmarking purposes. Workflows applying xTop or LFAQ quantified more proteins (Figure 2.4C) as well as more protein complexes (Figure 2.4F). The aLFQ R package used to access the TopN, iBAQ, APEX and NSAF algorithms has a build-in filter which reduces the total number of quantified proteins. The iBAQ, APEX and NSAF also require additional information, e.g. a protein database to calculate the theoretical number of tryptic peptides per protein, which often further limits the total number of quantified proteins [14, 15, 17].

Whereas workflows incorporating the xTop algorithm were generally more precise (Figure 2.4C), the workflows incorporating the LFAQ algorithm were more accurate (Figure 2.4F). This reflects deliberate design objectives in the two algorithms. The xTop algorithm quantifies proteins based on the most consistent corresponding peptide through all samples [20], thus achieving low median CV values, i.e. high τ precision. In contrast, the LFAQ algorithm focuses on rectifying response factors, thus achieving more accurate determination of abundances and higher η accuracy. With both a high number of quantified proteins and high η accuracy, the LFAQ algorithm would be the most appropriate for proteome-wide analyses, whereas the xTop algorithm may be used when higher precision is required, depending on the experimental objective. One noteworthy disadvantage of the xTop and LFAQ protein inference algorithms is the need for lengthy preprocessing of the input data, which reduces the usability. While the aLFQ R package containing the TopN, iBAQ, APEX and NSAF algorithms does include convenient import functionality, the *autoprot* pipeline now enables easy and flexible usage of all six algorithms.

2.4. Discussions

In this study we developed a versatile tool, *autoprot*, for automated analysis of data from absolute quantitative proteomics experiments. The pipeline was deployed to benchmark 105 different workflows and the resulting proteomics data sets were assessed based on quantity, precision and accuracy. This benchmarking study is the most extensive to date in any organism, as most other studies apply different techniques in only one of the workflow steps (Table 2.1, Figure 2.1). While other studies did provide a description of the full workflow with some inherent flexibility [40, 42, 54], the *autoprot* pipeline provides automated data analysis of raw MS files all the way through to proteome-wide absolute quantification in a high-throughput manner. Additionally, *autoprot* is an adaptable tool which allows for the determination of an optimal workflow for any given experimental objective.

The optimal workflow in this study combined library-free DIA data analysis using DIA-NN and LFAQ with a standard-free quantification approach and led to the identification and absolute quantification of 2,802 unique *E. coli* proteins over three experimental conditions. Despite the extensive range of experimental conditions in previous studies of *E. coli*, the current number of absolutely quantified proteins is comparable or higher than the number from the previous studies [20, 41]. We observed that two factors led to decreased filtering and therefore a greater number of quantified proteins, especially low abundance proteins: the use of DIA rather than DDA methods and the use of library-free rather than library-based analyses. When *autoprot* was deployed for analysis of the raw MS files from Mori *et al.* [20], it demonstrated the advantages of library-free methods combined with the recent analysis software DIA-NN. By leveraging functionalities such as advanced neural networks and match-between-runs strategies [39], the resulting precision and accuracy were similar to the original analysis, while the number of quantified proteins was greater. The ability to reprocess older data sets enables increased information gathering, and widens the application range of the *autoprot* pipeline substantially.

The prevailing approach for absolute quantification of proteins is the addition of internal standards, which provides reliability and accuracy for the corresponding endogenous proteins [1, 2]. Within this study, the addition of labelled internal standards produced a consistently higher precision for absolute quantification of the corresponding endogenous proteins. For proteome-wide analysis however, the overall precision was lower when a labelled rather than a standard-free quantification approach was deployed. The proteome-wide quantification with labelled and label-free approaches assumes that the correction factor derived from the internal standards is representative of the whole proteome. This application of internal standards as the internal distributor of the proteome composition was surpassed by the standard-free approach in both precision and accuracy. Additionally, the standard-free approach led to a greater number of quantified proteins, because the entire acquisition capacity could be dedicated to measuring the endogenous proteins rather than a large fraction occupied by internal standards. The high costs associated with internal standards are an added disadvantage and lower the potential throughput, especially for proteome-wide analyses [1, 28].

Apart from increased precision for a subset of proteins, internal standards can provide a rescaling factor for the calculation of intracellular protein concentrations. While the proteome composition derived with the standard-free approach always sums to 1 due to underlying assumptions [29, 30], the labelled and label-free approaches result in differing total protein mass values. Based on the exactly-known mass of the internal standards, the total injected mass could be calculated and used for rescaling of the proteome composition as a more reliable method compared to total protein measurements such as the BCA assay. Still, studies reporting calculated total protein mass per sample applied a reversed strategy of assuming the injected mass from total protein measurements as known, and justified the applied data analysis workflow based on the comparison [31, 40, 42]. This strategy relies on the assumption that the entirety of the injected protein mass cannot be detected, which would in fact be the case when internal standards were added to the samples. Afterwards, a constant value for the cellular protein density, i.e. total protein mass per cell volume, was commonly applied directly to calculate intracellular protein concentrations (Table 2.1). When applicable, the lengthy process of ribosome profiling could offer an accurate and reliable method for determination of the rescaling factor [20, 55]. If the addition of internal standards serves as no more than a poor internal distributor, the application of standard-free approaches for proteome-wide analyses would be preferable, especially for high-throughput experimental studies.

The current benchmark for protein identification is the application of spectral libraries which provides reliable precision, even though spectral libraries are time consuming and resource intensive to produce [53]. Since the development of spectral libraries typically relies on DDA methods, a fraction of proteins, mostly low abundance proteins, goes undetected even after careful fractionation of the samples. These proteins remain undetected during the subsequent DIA analysis, because the deployed library-based methods can only identify and thus quantify proteins which are present in the spectral library. In recent years, software for library-free identification and relative quantification, e.g. DIA-NN, has advanced to the point where the *in silico* generated spectral library can outperform library-based analysis [39, 56, 57]. The 2,802 unique *E. coli* proteins quantified with the optimal, library-free workflow exceeded the 2,770 unique *E. coli* proteins included in the spectral library of Mori *et al.* [20], which was generated using samples of 34 diverse growth conditions and 13 fractions of a pooled sample. Notably, the reproduction of the calibration data set from Mori *et al.* [20] showed the increased performance which can be achieved by deploying *in silico* spectral library generation with DIA-NN. When the corresponding spectral library was deployed, it was enhanced by the algorithms within DIA-NN to produce similar results to the library-free data set and thus increased quantities compared to the original data set. While the application of spectral libraries often leads to increased precision of quantified proteins [56, 57], it severely limits the throughput due to the costs of development for each new experimental study. It should be investigated extensively whether an experimental spectral library is crucial to achieve the experimental objective or if library-free methods would be more appropriate, for example in case of proteome-wide analysis.

Most studies with absolute quantification workflows deployed one or multiple of the older protein inference algorithm: TopN, iBAQ, APEX and NSAF (Table 2.1). When included in the comparison, the newer algorithms, xTop and LFAQ, always outperformed the older algorithms [20, 21, 26]. In this study, xTop and LFAQ also outperformed the older algorithms and workflows incorporating xTop achieved higher precision, while workflows incorporating LFAQ achieved both higher quantities and higher accuracy. Notably, the LFAQ protein inference algorithm may overestimate the concentration of low-abundance proteins (Figure S2.5F). Further investigation would be required in order to determine the exact biases of certain protein inference algorithms. Currently the iBAQ algorithm is still the most commonly used (Table 2.1), however due to their better performance and ease-of-use, the xTop and LFAQ algorithms should be considered for future proteome-wide analyses.

Although proteome composition values can be used as an absolute quantification, additional calculations are required to determine intracellular protein concentrations. Multiple strategies have been applied (Table 2.1), which converge towards obtaining the cellular protein density as total protein mass per cell volume. The total protein mass per biomass unit is commonly determined by measuring the biomass and total protein separately with, for example, an optical density measurement and the BCA assay. Cell size and volume are commonly measured and calculated using microscopy and image analysis. These measurements are not without error, since not all protein mass is extracted, in particular membrane proteins [52] and ribosomes [22], and cell volumes are highly variable across experimental conditions [58, 59]. Several studies opted for estimated values from literature instead (Table 2.1) to circumvent laborious measurements, e.g. values from Milo [60]. Regardless of the deployed strategy, a constant error is introduced when proteome composition values are converted into intracellular protein concentrations. This should not be ignored and where possible accounted for as described extensively by Mori *et al.* [20].

2.5. Conclusions

Through our benchmarking study and automated data analysis tool, we identified an optimal workflow for absolute quantitative proteomics in *E. coli*. The optimal workflow exploits advantages of the current state-of-the-art software and algorithms, which resulted in an unprecedentedly high number of identified and quantified proteins. All workflows applied in this study are available in the extensive *autoprot* tool, allowing for benchmarking studies in other organisms. We expect *autoprot* to provide an excellent platform for automated absolute quantitative proteomics, also for reprocessing older data sets.

References

[1] F. Calderón-Celis, J. R. Encinar, and A. Sanz-Medel. “Standardization approaches in absolute quantitative proteomics with mass spectrometry”. In: *Mass Spectrometry Reviews* 37 (2018), pp. 715–737. DOI: 10.1002/mas.21542.

[2] J. R. Wiśniewski. “Dilemmas With Absolute Quantification of Pharmacologically Relevant Proteins Using Mass Spectrometry”. In: *Journal of Pharmaceutical Sciences* 110.1 (2021), pp. 17–21. DOI: 10.1016/j.xphs.2020.10.034.

[3] M. Blein-Nicolas and M. Zivy. “Thousand and one ways to quantify and compare protein abundances in label-free bottom-up proteomics”. In: *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics* 1864.8 (2016), pp. 883–895. DOI: 10.1016/j.bbapap.2016.02.019.

[4] H. Li, J. Han, J. Pan, T. Liu, C. E. Parker, and C. H. Borchers. “Current trends in quantitative proteomics - an update”. In: *Journal of Mass Spectrometry* 52.5 (2017), pp. 319–341. DOI: 10.1002/jms.3932.

[5] S. A. Gerber, J. Rush, O. Stemman, M. W. Kirschner, and S. P. Gygi. “Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS”. In: *PNAS* 100.12 (2003), pp. 6940–6945. DOI: 10.1073/pnas.0832254100.

[6] R. J. Beynon, M. K. Doherty, J. M. Pratt, and S. J. Gaskell. “Multiplexed absolute quantification in proteomics using artificial QCAT proteins of concatenated signature peptides”. In: *Nature Methods* 2.8 (2005), pp. 587–589. DOI: 10.1038/NMETH774.

[7] M. Zeiler, W. L. Straube, E. Lundberg, M. Uhlen, and M. Mann. “A Protein Epitope Signature Tag (PrEST) Library Allows SILAC-based Absolute Quantification and Multiplexed Determination of Protein Copy Numbers in Cell Lines”. In: *Molecular and Cellular Proteomics* 11.3 (2012), O111.009613. DOI: 10.1074/mcp.O111.009613.

[8] F. Edfors, B. Forsström, H. Vunk, D. Kotol, C. Fredolini, G. Maddalo, A. S. Svensson, T. Boström, H. Tegel, P. Nilsson, J. M. Schwenk, and M. Uhlen. “Screening a Resource of Recombinant Protein Fragments for Targeted Proteomics”. In: *Journal of Proteome Research* 18 (2019), pp. 2706–2718. DOI: 10.1021/acs.jproteome.8b00924.

[9] V. Brun, A. Dupuis, A. Adrait, M. Marcellin, D. Thomas, M. Court, F. Vandenesch, and J. Garin. “Isotope-labeled Protein Standards: Toward Absolute Quantitative Proteomics”. In: *Molecular and Cellular Proteomics* 6 (12 2007), pp. 2139–2149. DOI: 10.1074/mcp.M700163-MCP200.

[10] H. A. Feteisi, B. Achour, J. Barber, and A. Rostami-Hodjegan. “Choice of LC-MS Methods for the Absolute Quantification of Drug-Metabolizing Enzymes and Transporters in Human Tissue: a Comparative Cost Analysis”. In: *The AAPS Journal* 17 (2015), pp. 438–446. DOI: 10.1208/s12248-014-9712-6.

[11] J. C. Silva, M. V. Gorenstein, G. Z. Li, J. P. C. Vissers, and S. J. Geromanos. “Absolute Quantification of Protein by LCMSE: A Virtue of Parallel MS Acquisition”. In: *Molecular and Cellular Proteomics* 5.1 (2006), pp. 144–156. DOI: 10.1074/mcp.M500230-MCP200.

[12] J. Malmström, M. Beck, A. Schmidt, V. Lange, E. W. Deutsch, and R. Aebersold. "Proteome-wide cellular protein concentrations of the human pathogen *Leptospira interrogans*". In: *Nature* 460 (2009), pp. 762–765. DOI: `10.1038/nature08184`.

[13] C. Ludwig, M. Claassen, A. Schmidt, and R. Aebersold. "Estimation of Absolute Protein Quantities of Unlabeled Samples by Selected Reaction Monitoring Mass Spectrometry". In: *Molecular and Cellular Proteomics* 11.3 (2012), p. M111.013987. DOI: `10.1074/mcp.M111.013987`.

[14] B. Schwanhäusser, D. Busse, N. Li, G. Dittmar, J. Schuchhardt, J. Wolf, W. Chen, and M. Selbach. "Global quantification of mammalian gene expression control". In: *Nature* 473 (2011), pp. 337–342. DOI: `10.1038/nature10098`.

[15] B. Zybailov, A. L. Mosley, M. E. Sardiu, M. K. Coleman, L. Florens, and M. P. Washburn. "Statistical Analysis of Membrane Proteome Expression Changes in *Saccharomyces cerevisiae*". In: *Journal of Proteome Research* 5 (2006), pp. 2339–2347. DOI: `10.1021/pr060161n`.

[16] S. Gerster, T. Kwon, C. Ludwig, M. Matondo, C. Vogel, E. M. Marcotte, R. Aebersold, and P. Bühlmann. "Statistical Approach to Protein Quantification". In: *Molecular and Cellular Proteomics* 13.2 (2014), pp. 666–677. DOI: `10.1074/mcp.M112.025445`.

[17] P. Lu, C. Vogel, R. Wang, X. Yao, and E. M. Marcotte. "Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation". In: *Nature Biotechnology* 25.1 (2007), pp. 117–124. DOI: `10.1038/nbt1270`.

[18] Y. Ishihama, Y. Oda, T. Tabata, T. Sato, T. Nagasu, J. Rappsilber, and M. Mann. "Exponentially Modified Protein Abundance Index (emPAI) for Estimation of Absolute Protein Amount in Proteomics by the Number of Sequenced Peptides per Protein". In: *Molecular and Cellular Proteomics* 4.9 (2005), pp. 1265–1272. DOI: `10.1074/mcp.M500061-MCP200`.

[19] J. Cox, M. Y. Hein, C. A. Luber, I. Paron, N. Nagaraj, and M. Mann. "Accurate Proteome-wide Label-free Quantification by Delayed Normalization and Maximal Peptide Ratio Extraction, Termed MaxLFQ". In: *Molecular and Cellular Proteomics* 13.9 (2014), pp. 2513–2526. DOI: `10.1074/mcp.M113.031591`.

[20] M. Mori, Z. Zhang, A. Banaei-Esfahani, J. B. Lalanne, H. Okano, B. C. Collins, A. Schmidt, O. T. Schubert, D. S. Lee, G. W. Li, R. Aebersold, T. Hwa, and C. Ludwig. "From coarse to fine: The absolute *Escherichia coli* proteome under diverse growth conditions". In: *Molecular Systems Biology* 17.5 (2021), e9536. DOI: `10.15252/msb.20209536`.

[21] C. Chang, Z. Gao, W. Ying, Y. Fu, Y. Zhao, S. Wu, M. Li, G. Wang, X. Qian, Y. Zhu, and F. He. "LFAQ: Toward Unbiased Label-Free Absolute Protein Quantfication by Predicting Peptide Quantitative Factors". In: *Analytical Chemistry* 91 (2019), pp. 1335–1343. DOI: `10.1021/acs.analchem.8b03267`.

[22] L. Arike, K. Valgepea, L. Peil, R. Nahku, K. Adamberg, and R. Vilu. "Comparison and applications of label-free absolute proteome quantification methods on

Escherichia coli". In: *Journal of Proteomics* 75 (2012), pp. 5437–5448. DOI: 10.1016/j.jprot.2012.06.020.

[23] E. Ahrné, L. Molzahn, T. Glatter, and A. Schmidt. "Critical assessment of proteome-wide label-free absolute abundance estimation strategies". In: *Proteomics* 13 (2013), pp. 2567–2578. DOI: 10.1002/pmic.201300135.

[24] T. Shalit, D. Elinger, A. Savidor, A. Gabashvili, and Y. Levin. "MS1-Based Label-Free Proteomics Using a Quadrople Orbitrap Mass Spectrometer". In: *Journal of Proteome Research* 14 (2015), pp. 1979–1986. DOI: 10.1021/pr501045t.

[25] J. A. Bubis, L. I. Levitsky, M. V. Ivanov, I. A. Tarasova, and M. V. Gorshkov. "Comparative evaluation of label-free quantification methods for shotgun proteomics". In: *Rapid Communications in Mass Spectrometry* 31 (2017), pp. 606–612. DOI: 10.1002/rcm.7829.

[26] L. Zhao, X. Cong, L. Zhai, H. Hu, J. Y. Xu, W. Zhao, M. Zhu, M. Tan, and B. C. Ye. "Comparative evaluation of label-free quantification strategies". In: *Journal of Proteomics* 215 (2020), p. 103669. DOI: 10.1016/j.jprot.2020.103669.

[27] C. Lindemann, N. Thomanek, F. Hundt, T. Lerari, H. E. Meyer, D. Wolters, and K. Marcus. "Strategies in relative and absolute quantitative mass spectrometry based proteomics". In: *Biological Chemistry* 398.5-6 (2017), pp. 687–699. DOI: 10.1515/hsz-2017-0104.

[28] J. A. Ankney, A. Muneer, and X. Chen. "Relative and Absolute Quantification in Mass Spectrometry-Based Proteomics". In: *Annual Review of Analytical Chemistry* 11 (2018), pp. 49–77. DOI: 10.1146/annurev-anchem-061516-045357.

[29] J. R. Wiśniewski, P. Ostasiewicz, K. Duś, D. F. Zielińska, F. Gnad, and M. Mann. "Extensive quantitative remodeling of the proteome between normal colon tissue and adenocarcinoma". In: *Molecular Systems Biology* 8.1 (2012), p. 611. DOI: 10.1038/msb.2012.44.

[30] J. R. Wiśniewski, M. Y. Hein, J. Cox, and M. Mann. "A "Proteomic Ruler" for Protein Copy Number and Concentration Estimation without Spike-in Standards". In: *Molecular and Cellular Proteomics* 13.12 (2014), pp. 3497–3506. DOI: 10.1074/mcp.M113.037309.

[31] B. J. Sánchez, P. J. Lahtvee, K. Campbell, S. Kasvandik, R. Yu, I. Domenzain, A. Zelezniak, and J. Nielsen. "Benchmarking accuracy and precision of intensity-based absolute quantification of protein abundance in *Saccharomyces cerevisiae*". In: *Proteomics* 21.6 (2021), pp. 1–5. DOI: 10.1002/pmic.202000093.

[32] P. J. Lahtvee, B. J. Sánchez, A. Smialowska, S. Kasvandik, I. E. Elsemman, F. Gatto, and J. Nielsen. "Absolute Quantification of Protein and mRNA Abundances Demonstrate Variability in Gene-Specific Translation Efficiency in Yeast". In: *Cell Systems* 4 (2017), pp. 495–504. DOI: 10.1016/j.cels.2017.03.003.

[33] D. Heckmann, A. Campeau, C. J. Lloyd, P. V. Phaneuf, Y. Hefner, M. Carrillo-Terrazas, A. M. Feist, D. J. Gonzalez, and B. O. Palsson. "Kinetic profiling of metabolic specialists demonstrates stability and consistency of in vivo enzyme turnover numbers". In: *PNAS* 117.37 (2020), pp. 23182–23190. DOI: 10.1073/pnas.200156211.

[34] J. Xia, B. J. Sánchez, Y. Chen, K. Campbell, S. Kasvandik, and J. Nielsen. "Proteome allocations change linearly with the specific growth rate of *Saccharomyces cerevisiae* under glucose limitation". In: *Nature Communications* 13 (2022), p. 2819. DOI: 10.1038/s41467-022-30513-2.

[35] H. L. Röst, L. Malmström, and R. Aebersold. "Reproducible quantitative proteotype data matrices for systems biology". In: *Molecular Biology of the Cell* 26 (2015), pp. 3926–3931. DOI: 10.1091/mbc.E15-07-0507.

[36] O. T. Schubert, H. L. Röst, B. C. Collins, G. Rosenberger, and R. Aebersold. "Quantitative proteomics: challenges and opportunities in basic and applied research". In: *Nature Protocols* 12.7 (2017), pp. 1289–1294. DOI: 10.1038/nprot.2017.040.

[37] C. Ludwig, L. Gillet, G. Rosenberger, S. Amon, B. C. Collins, and R. Aebersold. "Data-independent acquistion-based SWATH-MS for quantitative proteomics: a tutorial". In: *Molecular Systems Biology* 14.8 (2018), e8126. DOI: 10.15252/msb.20178126.

[38] T. S. Batth, P. Singh, V. R. Ramakrishnan, M. M. L. Sousa, L. J. G. Chan, H. M. Tran, E. G. Luning, E. H. Y. Pan, K. M. Vuu, J. D. Keasling, P. D. Adams, and C. J. Petzold. "A targeted proteomics toolkit for high-throughput absolute quantification of *Escherichia coli* proteins". In: *Metabolic Engineering* 26 (2014), pp. 48–56. DOI: 10.1016/j.ymben.2014.08.004.

[39] V. Demichev, C. B. Messner, S. I. Vernardis, K. S. Lilley, and M. Ralser. "DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput". In: *Nature Methods* 17 (2020), pp. 41–44. DOI: 10.1038/s41592-019-0638-x.

[40] O. T. Schubert, C. Ludwig, M. Kogadeeva, M. Zimmermann, G. Rosenberger, M. Gengenbacher, L. C. Gillet, B. C. Collins, H. L. Röst, S. H. E. Kaufmann, and U. Sauer. "Absolute Proteome Composition and Dynamics during Dormancy and Resuscitation of *Mycobacterium tuberculosis*". In: *Cell Host and Microbe* 18.1 (2015), pp. 96–108. DOI: 10.1016/j.chom.2015.06.001.

[41] A. Schmidt, K. Kochanowski, S. Vedelaar, E. Ahrné, B. Volkmer, L. Callipo, K. Knoops, M. Bauer, R. Aebersold, and M. Heinemann. "The quantitative and condition-dependent Escherichia coli proteome". In: *Nature Biotechnology* 34.1 (2016), pp. 104–110. DOI: 10.1038/nbt.3418.

[42] K. Valgepea, G. Talbo, N. Takemori, A. Takemori, C. Ludwig, V. Mahamkali, A. P. Mueller, R. Tappe; M. Köpke, S. D. Simpson, L. K. Nielsen, and E. Marcellin. "Absolute Proteome Quantification in the Gas-Fermenting Acetogen *Clostridium autoethaneogenum*". In: *Genomics and Proteomics* 7.2 (2022), pp. 1–19. DOI: 10.1128/msystems.00026-22.

[43] E. Kozaeva, S. Volkova, M. R. A. Matos, M. P. Mezzina, T. Wulff, D. C. Volke, L. K. Nielsen, and P. I. Nikel. "Model-guided dynamic control of essential metabolic nodes boosts acetyl-coenzyme A–dependent bioproduction in rewired Pseudomonas putida". In: *Metabolic Engineering* 67 (2021), pp. 373–386. DOI: 10.1016/j.ymben.2021.07.014.

[44] N. Gurdo, S. K. Taylor Parkins, M. Fricano, T. Wulff, L. K. Nielsen, and P. I. Nikel. "Protocol for absolute quantification of proteins in Gram-negative bacteria based on QconCAT-based labeled peptides". In: *STAR Protocols* 4.1 (2023), p. 102060. DOI: 10.1016/j.xpro.2023.102060.

[45] Y Xuan, N. W. Bateman, S. Gallien, S. Goetze, Y. Zhou, P. Navarro, M. Hu, N. Parikh, B. L. Hood, K. A. Conrads, C. Loosse, R. B. Kitata, S. R. Piersma, D. Chiasserini, H. Zhu, G. Hou, M. Tahir, A. Macklin, A. Khoo, X. Sun, B. Crossett, A. Sickmann, Y. Chen, C. R. Jimenez, H. Zhou, S. Liu, M. R. Larsen, T. Kislinger, Z. Chen, B. L. Parker, S. J. Cordwell, B. Wollscheid, and T. P. Conrads. "Standardization and harmonization of distributed multi-center proteotype analysis supporting precision medicine studies". In: *Nature Communications* 11.1 (2020), p. 5248. DOI: 10.1038/s41467-020-18904-9.

[46] H. L. Röst, G. Rosenberger, P. Navarro, L. Gillet, S. M. Miladinović, O. T. Schubert, W. Wolski, B. C. Collins, J. Malmström, L. Malmström, and R. Aebersold. "OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data". In: *Nature Biotechnology* 32.3 (2014), pp. 219–223. DOI: 10.1038/nbt.2841.

[47] G Rosenberger, C Ludwig, H. L. Röst, R. Aebersold, and L. Malmström. "aLFQ: an R-package for estimating absolute protein quantities from label-free LC-MS/MS proteomics data". In: *Bioinformatics* 30.17 (2014), pp. 2511–2513. DOI: 10.1093/bioinformatics/btu200.

[48] B. Volkmer and M. Heinemann. "Condition-Dependent Cell Volume and Concentration of Escherichia coli to Facilitate Data Conversion for Systems Biology Modeling". In: *PLoS ONE* 6.7 (2011), e23126. DOI: 10.1371/journal.pone.0023126.

[49] B. H. M. Meldal, L. Perfetto, C. Combe, T. Lubiana, J. V. Ferreira Cavalcante, H. Bye-A-Jee, A. Waagmeester, N. del-Toro, A. Shrivastava, E. Barrera, E. Wong, B. Mlecnik, G. Bindea, K. Panneerselvam, E. Willighagen, J. Rappsilber, P. Porras, H. Hermjakob, and S. Orchard. "Complex Portal 2022: new curation frontiers". In: *Nucleic Acids Research* 50.D1 (2021), pp. D578–D586. DOI: 10.1093/nar/gkab991.

[50] Y. Perez-Riverol, J. Bai, C. Bandla, D. García-Seisdedos, S. Hewapathirana, S. Kamatchinathan, D. J. Kundu, A. Prakash, A. Frericks-Zipper, M. Eisenacher, M. Walzer, S. Wang, A. Brazma, and J. A. Vizcaíno. "The PRIDE database resources in 2022: a hub for mass spectrometry-based proteomics evidences". In: *Nucleic Acids Research* 50.D1 (2022), pp. D543–D552. DOI: 10.1093/nar/gkab1038.

[51] Y. Chen, J. M. Guenther, J. W. Gin, L. J. G. Chan, Z. Costello, T. L. Ogorzalek, H. M. Tran, J. M. Blake-Hedges, J. D. Keasling, P. D. Adams, H. García Martín, N. J. Hillson, and C. J. Petzold. "Automated "Cells-To-Peptides" Sample Preparation Workflow for High-Throughput Quantitative Proteomic Assays for Microbes". In: *Journal of Proteome Research* 18 (2019), pp. 3752–3761. DOI: 10.1021/acs.jproteome.9b00455.

[52] J. N. Savas, B. D. Stein, C. C. Wu, and J. R. Yates. “Mass spectrometry accelerates membrane protein analysis”. In: *Trends in Biochemical Sciences* 36.7 (2011), pp. 388–396. DOI: `10.1016/j.tibs.2011.04.005`.

[53] O. T. Schubert, L. C. Gillet, B. C. Collins, P. Navarro, G. Rosenberger, W. E. Wolski, H. Lam, D. Amodei, P. Mallick, B. MacLean, and R. Aebersold. “Building high-quality assay libraries for targeted analysis of SWATH MS data”. In: *Nature Protocols* 10.3 (2015), pp. 426–441. DOI: `10.1038/nprot.2015.015`.

[54] B. J. Sánchez, C. Zhang, A. Nilsson, P. J. Lahtvee, E. J. Kerkhoven, and J. Nielsen. “Improving the phenotype predictions of a yeast genome-scale metabolic model by incorporating enzymatic constraints”. In: *Molecular Systems Biology* 13.8 (2017), p. 935. DOI: `10.15252/msb.20167411`.

[55] G. Li, D. Burkhardt, C. Gross, and J. S. Weissman. “Quantifying Absolute Protein Synthesis Rates Reveals Principles Underlying Allocation of Cellular Resources”. In: *Cell* 157.3 (2014), pp. 624–635. DOI: `10.1016/j.cell.2014.02.033`.

[56] Y. Yang, X. Liu, C. Shen, Y. Lin, P. Yang, and L. Qiao. “In silico spectral libraries by deep learning facilitate data-independent acquisition proteomics”. In: *Nature Communications* 11.1 (2020), p. 146. DOI: `10.1038/s41467-019-13866-z`.

[57] R. Lou, Y. Cao, S. Li, X. Lang, Y. Li, Y. Zhang, and W. Shui. “Benchmarking commonly used software suites and analysis workflows for DIA proteomics and phosphoproteomics”. In: *Nature Communications* 14.1 (2023), p. 94. DOI: `10.1038/s41467-022-35740-1`.

[58] M. Basan, M. Zhu, X. Dai, M. Warren, D. Sévin, Y. Wang, and T. Hwa. “Inflating bacterial cells by increased protein synthesis”. In: *Molecular Systems Biology* 11.10 (2015), p. 836. DOI: `10.15252/msb.20156178`.

[59] F. Si, D. Li, S. E. Cox, J. T. Sauls, O. Azizi, C. Sou, A. B. Schwartz, M. J. Erickstad, Y. Jun, X. Li, and S. Jun. “Invariance of Initiation Mass and Predictability of Cell Size in Escherichia coli”. In: *Current Biology* 27.9 (2017), pp. 1278–1287. DOI: `10.1016/j.cub.2017.03.022`.

[60] R. Milo. “What is the total number of protein molecules per cell volume? A call to rethink some published values”. In: *BioEssays* 35.12 (2013), pp. 1050–1055. DOI: `10.1002/bies.201300066`.

Supplementary materials

Extended experimental methods

Protein identification and relative quantification - DDA data analysis
The raw files from the DDA method were analysed using Proteome Discoverer v2.4 (Thermo Fisher Scientific) with the following settings: carbamidomethylation of cysteine residues and oxidation of methionine residues as dynamic modifications, precursor mass tolerance set to 10 ppm, fragment mass tolerance at 0.02 Da, trypsin (full) as digestion enzyme, the maximum missed cleavage sites set to 2, peptide length set to a minimum of 6 and maximum of 144, and the false discovery rate (FDR) set to 0.1%. For the data analysis of samples including QconCAT proteins, $^{13}C(6)$ labelling of arginine and lysine was added as dynamic modification. Only peptides unique to one protein were used in further analysis and the normalisation between samples was based on total peptide amounts.

Protein identification and relative quantification - DIA data analysis
For the data analysis of raw files from the DIA method, either Spectronaut v17 (17.3.230224. 55965; Biognosys AG) or DIA-NN v1.8 [1] were used with either a library-based or library-free approach. Only peptides unique to one protein were used in further analysis. All spectral libraries were constructed from the corresponding DDA raw files and protein database, using the Pulsar search engine available in Spectronaut with the following settings: Trypsin/P as cleavage rule, peptide length set to a minimum of 7 and maximum of 52, allowing 2 missed cleavages, carbamidomethylation of cysteine residues as a fixed modification, acetylation of the protein N-term and oxidation of methionine residues as variable modifications, and allowing for both b and y ions. When analysing samples including QconCAT proteins, a channel with the labels "Arg6" and "Lys6" and a channel without labels were added.

Spectronaut was applied for library-based analyses with the following settings: the MZ extraction strategy set to maximum intensity, the precursor FDR cut-off set to 1%, single hit definition by stripped sequence, quantification on MS2 level using the sum of the precursor quantities, allowing for minimum 1 and maximum 5 precursors in TopN strategy, and the workflow fallback option set to spike-in. For library-free analyses with Spectronaut, the settings for library construction were identical to the Pulsar search engine settings. Additionally, the workflow was set as "directDIA+ (Deep)", missing channels were generated *in silico*, the peptide FDR set to 0.01%, and the fragment ion selection strategy was intensity based.

Library-based analyses with DIA-NN were performed with the high precision and robust LC quantification modes enabled, the smart profiling and the heuristic protein inference activated and the FDR cut-off set at 1%. For the additional *in silico* digest feature, acetylation of the protein N-term and oxidation of methionine residues were set as variable modifications, and the cleavage specificity was set to "K*,R*" (Trypsin/P). When samples including QconCAT proteins were analysed, two channels were added to the settings: one without labels and one with $^{13}C(6)$ labels for arginine and lysine.

Fixed modifications for ^{13}C(6) labelling of arginine and lysine were also added and peak translation was activated. DIA-NN was used for library-free analyses with identical settings as the library-based DIA-NN analyses, and the following additions: generate spectral library, predictor, FASTA search and match-between-runs (MBR) enabled, minimum 200 and maximum 1,800 m/z for fragment exclusion, minimum 7 and maximum 30 for peptide length, minimum 300 and maximum 1,800 m/z for precursor exclusion, minimum 1 and maximum 4 for precursor charge, and the maximum missed cleavages set to 1. Specifically for library-free DIA-NN analyses applied to samples including QconCAT proteins, the library construction and subsequent analysis using this library had to be performed as two distinct processes due to limitations of DIA-NN. The library construction was performed using the same settings as the library-free analyses, however with re-analysis disabled. For the subsequent identification and relative quantification, the settings were identical to the library-based analysis of samples including QconCAT proteins.

Proteome-wide absolute quantification
For analysis of samples including QconCAT proteins, the ratio between the intensity of a peptide from a QconCAT protein and the corresponding endogenous peptide was used as a basis for absolute quantification. The concentrations of proteins for which both an unlabelled (light) peptide and the corresponding ^{13}C(6) labelled (heavy) peptide were identified and relatively quantified were estimated according to Equation S2.1.

$$c_{prot,end} = \frac{\sum c_{pep,end}}{N_{pep,end}}, \quad c_{pep,end} = \frac{I_{pep,end}}{I_{pep,QconCAT}} \cdot c_{pep,QconCAT} \tag{S2.1}$$

In Equation S2.1, $c_{prot,end}$ is the estimated concentration of the endogenous protein, $c_{pep,end}$ is the estimated concentration of an endogenous peptide, $N_{end,pep}$ is the number of endogenous peptides for which a ^{13}C(6)-labelled QconCAT peptide was identified, for a particular endogenous protein, $I_{pep,end}$ is the intensity of the endogenous peptide (light peptide), $I_{pep,QconCAT}$ is the intensity of the ^{13}C(6)-labelled QconCAT peptide (heavy peptide), and $c_{pep,QconCAT}$ is the concentration of the QconCAT peptide.

The proteins for which peptides were included in the QconCAT proteins were used to achieve the proteome-wide absolute quantification. For analysis of samples including the UPS2 protein mix, all UPS2 proteins which were identified and relatively quantified in the samples, were used in the proteome-wide absolute quantification. Linear regression was performed on log10-transformed protein intensities and log10-transformed protein concentrations. The parameters of the linear regression were used to estimate concentration for each protein using Equation S2.2, except for the proteins which were previously absolutely quantified by internal standards.

$$c_{prot} = 10^{\,a \cdot log_{10}(I_{prot}) + b} \tag{S2.2}$$

In Equation S2.2, c_{prot} is the estimated protein concentration, a is the slope of the linear regression, I_{prot} is the protein intensity, and b is the intercept of the linear regression.

For analyses using the standard-free approach, the protein concentrations are estimated following equation S2.3, adapted from Wiśniewski *et al.* (2014) [2].

$$c_{prot,i} \;=\; \frac{I_{prot,i}}{\sum(I_{prot} \cdot MW_{prot})} \tag{S2.3}$$

In Equation S2.3, $c_{prot,i}$ is the estimated protein concentration, $I_{prot,i}$ is the protein intensity, and MW_{prot} is the protein molecular weight. The sum of all protein intensities multiplied by the corresponding protein molecular weight is determined before calculation of individual protein concentrations. The protein molecular weights were calculated using the protein sequences from the protein database and the ProtParam module of the BioPython Python package [3].

autoprot **documentation**

The *autoprot* pipeline allows for absolute quantification of proteins from raw mass spectrometry (MS) files in an automated manner. The pipeline covers data analysis from both DIA and DDA methods, where a fully open-source option is available for DIA methods. Raw data from labelled, label-free and standard-free approaches can be analysed with the pipeline. The normalisation of peptide intensities into protein intensities is performed with seven different algorithms to identify the optimal algorithm for the current experiment. The incorporated algorithms are Top3 [4], Top all [4], iBAQ [5], APEX [6], NSAF [7], LFAQ [8], and xTop [9].

Install
The required files can be downloaded from this GitHub repository with the following command:

```
git clone git@github.com:biosustain/autoprot.git
```

Due to the many available options, the *autoprot* pipeline depends on a number of different software and packages. A list of all dependencies and their corresponding, tested version is provided below. The `autoprot.ps1` script and multiple other scripts or executables have to be added to the PATH variable for the *autoprot* pipeline to work properly. While file paths can be added to the PATH variable through the command line, on Windows one can also add to the PATH variable through the graphical user interface (GUI).

To test if the *autoprot* pipeline is set up properly, the files in `Examples\Input` can be used in combination with raw MS files of the standard-free DIA analysis (place the 9 .raw files in the `Examples\Input` folder first) for a test run with the following command:

```
autoprot.ps1 -osDIA -mode "directDIA" -approach "free" -InputDir
"$PSScriptRoot\..\Examples\Input" -ExpName "test_run" -fasta
"$PSScriptRoot\..\Examples\Input\URF_UP000000625_E_coli.fasta"
-totalProt "$PSScriptRoot\..\Examples\Input\CPD_example.csv"
```

The output files can be verified with the files in `Examples\Output`.

Dependencies

Name	Version	Source
PowerShell 7	7.2.4	`https://learn.microsoft.com/en-us/powershell/scripting/install/installing-powershell-on-windows?view=powershell-7.2\#installing-the-msi-package` (Windows operating system has PowerShell 5.1 as default, however PowerShell 7.2 (or higher) is required alongside the default, so that additional functions can be accessed. The whole pipeline runs on 7.2 or up.)
Python	3.8.8 (or higher)	`https://www.anaconda.com/` (Including numpy==1.20.1, pandas==1.2.4, statsmodels==0.12.2, matplotlib==3.3.4, Biopython==1.78. Add location of python.exe to PATH variable.)
Spectronaut	17 (17.3.230224.55965)	`https://biognosys.com/software/spectronaut/` (Commercially available. Add location of spectronaut.exe to PATH variable.)
DIA-NN	1.8	`https://github.com/vdemichev/DiaNN` (Open source.)
Proteome Discoverer	2.4.1.15	`https://www.thermofisher.com/dk/en/home/industrial/mass-spectrometry/liquid-chromatography-mass-spectrometry-lc-ms/lc-ms-software/multi-omics-data-analysis/proteome-discoverer-software.html` (Commercially available. Not actually part of the pipeline, since no command line tool is available.)
R (Rscript)	4.1.1 (or higher)	`https://cran.r-project.org/bin/windows/base/` (Add location of rscript.exe to PATH variable.)
aLFQ	1.3.5	`https://github.com/aLFQ/aLFQ` (Add package to R.)
xTop	1.2	`https://gitlab.com/mm87/xtop` (Add location of xTop_pipeline.py to PATH variable.)
LFAQ	1.0.0	`https://github.com/LFAQ/LFAQ` (Add location of LFAQ executables to PATH variable.)

Usage

The `autoprot.ps1` script can be executed in PowerShell 7 (when added to the PATH variable) as follows:

```
autoprot.ps1 [args]
```

To access the *autoprot* help from the command line in PowerShell 7:

```
Get-Help autoprot.ps1 -Full
```

When the `autoprot.ps1` script is located on a drive with restricted access, e.g. a network drive, and cannot be executed, the following command can provide access for execution:

```
Set-ExecutionPolicy -ExecutionPolicy Bypass -Scope Process
```

The available arguments are:

`-osDIA` [flag] enables the open-source option for DIA analysis, which uses DIA-NN instead of Spectronaut.

`-mode` [string] **mandatory** specify the acquisition mode as "DDA", "DIA" or "directDIA".

`-approach` [string] **mandatory** specify the quantification approach as "label", "unlabel" or "free".

`-InputDir` [directory] **mandatory** specify the input directory containing all input files with raw MS spectra. The output directory will be located in the input directory after the run.

`-ExpName` [string] **mandatory** specify the name of the experiment.

`-fasta` [file] **mandatory** specify the FASTA file with the proteome sequences.

`-totalProt` [file] **mandatory** specify the file with the cellular protein density values for each sample.

`-DDAresultsFile` [file] (mandatory for "DDA" mode) specify the file with the Proteome Discoverer Peptide Groups results.

`-SpecLib` [file] (mandatory for "DIA" mode) specify the file with the spectral library for the "DIA" mode.

`-BGSfasta` [file] (mandatory for "directDIA" mode with Spectronaut) specify the FASTA file in .BGSfasta format, which is required for the "directDIA" mode using Spectronaut (commercial).

`-ISconc` [file] (mandatory for "label" and "unlabel" approaches) specify the file with the absolute concentrations of each standard peptide ("label" approach) or protein ("unlabel" approach).

Specific input data

Ensure that the FASTA file with the proteome sequences follows the official UniProt configuration for the headers. An example FASTA file can be found in `Examples\Input\URF_UP000000625_E_coli.fasta`.

All workflows in `DIA` and `directDIA` mode can be initialised from .RAW files (Thermo Fisher Scientific instrument specific - please open an issue if another type is required in combination with Spectronaut) using either Spectronaut (commercial; Biognosys AG, Schlieren, Switzerland) or DIA-NN (open source; [1]). Any workflow in `DDA` mode can be initialised from the `PeptideGroups.csv` output file of Proteome Discoverer (Thermo Fisher Scientific, Waltham, MA, USA). How to get the `PeptideGroups.csv` file with Proteome Discoverer results: Open the .PDRESULTS file of the study in Proteome Discoverer, click on "File" -> "Export" -> "To Microsoft Excel", select "Peptide Groups" from the drop-down menu for level 1 and click on "Export". Open the resulting file in Microsoft Excel and save as a .CSV file with the name `PeptideGroups`.

For a workflow in `directDIA` mode using Spectronaut (commercial; Biognosys AG, Schlieren, Switzerland), a BGSfasta version of the fasta file is required. This BGSfasta version can be obtained by loading the fasta file with the proteome sequences in Spectronaut (commercial; Biognosys AG, Schlieren, Switzerland) as a protein database. Then, the BGSfasta version of the fasta file should be in the folder `$HOME\Databases\Spectronaut\`.

The *autoprot* pipeline has two custom input files which are described below.

Cellular protein density

The table with cellular protein density for each sample should have the following headers: `Sample` [string] with the name of each sample which should be the same as the names of the .RAW files and `CPD` [float] with the cellular protein density of each sample in g/L. An example file for the cellular protein density table can be found in `Examples\Input\CPD_example.csv`.

Sample	**CPD**
sample1	<float>
sample2	<float>
...	...

Internal standard concentration

For the `label` approach, the table with the concentration for each internal standard should be peptide-based (for example AQUA or QconCAT peptides) with the following headers: `FullPeptideName`[string] with the peptide sequence, `ProteinName` [string] with the UniProt identifier of the corresponding protein (should be identical to the identifiers in the fasta file with the proteome sequences), `Concentration` [float] with the spiked-in concentration of each internal standard peptide into the sample in fmol/µg whole cell lysate (total protein extracted). An example file for the peptide-based internal standard concentration table can be found in `Examples\Input\ISconc_peptides_example.csv`.

FullPeptideName	**ProteinName**	**Concentration**
sequence1	UniProt ID1	<float>
sequence2	UniProt ID2	<float>
...	...	...

For the `unlabel` approach, the table with the concentration for each internal standard should be protein-based (for example UPS2 protein mix) with the following headers: `ProteinName` [string] with the UniProt identifier of the corresponding protein (should be identical to the identifiers in the fasta file with the proteome sequences), `Concentration` [float] with the spiked-in concentration of each internal standard peptide into the sample in fmol/µg whole cell lysate (total protein extracted). An example file for the peptide-based internal standard concentration table can be found in `Examples\Input\ISconc_proteins_example.csv`.

ProteinName	**Concentration**
UniProt ID1	<float>
UniProt ID2	<float>
...	...

Output data

The output directory will be located in the input directory after the run and will contain seven files with a protein concentration table, one for each algorithm. The protein concentration table has the following headers: `ProteinName` [string] with the UniProt identifier of the corresponding protein (identical to the identifiers in the fasta file with the proteome sequences), `sample_conc(fmol/µg)_X` [float] with the protein concentration in sample X in fmol/µg whole cell lysate (total protein extracted) for each sample, `invivo_conc(mM)_X` [float] with the intracellular protein concentration in sample X in mM (millimol/liter) for each sample. Example files for the protein-based results table can be found in the `Examples\Output` folder.

ProteinName	**sample_conc(fmol/µg)_X**	**invivo_conc(mM)_X**	...
UniProt ID1	<float>	<float>	...
UniProt ID2	<float>	<float>	...
...	...	...	...

Intermediate files

All intermediate output files of the *autoprot* pipeline will be located in `intermediate_results` in the output directory. Of particular interest, the linear regression plots of the proteome absolute quantification for the `labelled` or `unlabel` approach will be located in `intermediate_results\Absolute_quantification\LR_plots`.

Analysis settings

Currently, only 13C(6) labelling of arginine (Arg6) and lysine (Lys6) residues is allowed for the `label` approach, which are incorporated into the DIA analysis settings of the `directDIA` mode. However, the `label` approach is peptide-based, thus both methods using AQUA peptides or QconCAT proteins are supported. The `unlabel` approach is protein-based and allows for any protein to be used as internal standard, e.g. UPS2 protein kit.

The DIA analysis settings for both Spectronaut and DIA-NN include quantification on MS2 level. Specifically for the `directDIA` mode, the DIA analysis settings include the Trypsin/P cleavage rule (digestion with Trypsin/Lys-C mix) and the following modifications: Carbamidomehtyl (C), Acetyl (Protein N-term), and Oxidation (M). The exact settings can be found in the corresponding DIA analysis settings file in `Scripts\DIA_analysis`. DIA-NN uses config files which can be viewed using any text editor, while Spectronaut uses property files which can be viewed by importing the file into Spectronaut in the Settings tab.

Copyright

- Copyright (c) 2023, Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark.
- Free software distributed under the GNU General Public License 3.0.

Supplementary figures

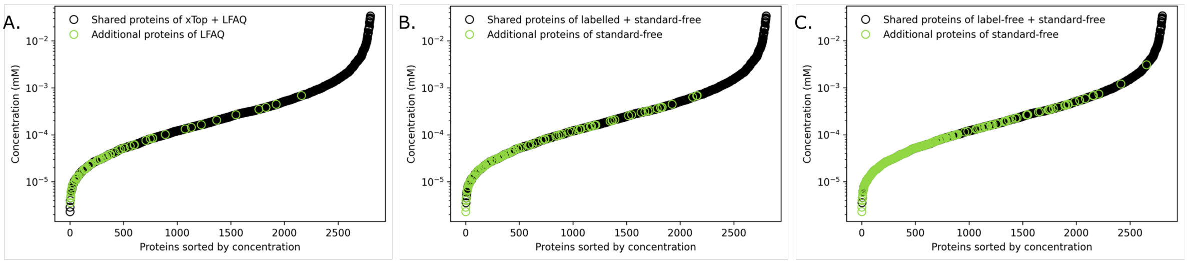


Figure S2.1: Comparison of the optimal workflow, which deployed standard-free, library-free DIA data analysis using DIA-NN and LFAQ, to other workflows in terms of protein concentration range. (A) Comparison to the workflow which deployed standard-free, library-free DIA data analysis using DIA-NN and xTop. (B) Comparison to the workflow which combined absolute quantification using the labelled approach (QconCAT proteins) with library-free DIA data analysis using DIA-NN and LFAQ. (C) Comparison to the workflow which combined absolute quantification using the label-free approach (UPS2 protein mix) with library-free DIA data analysis using DIA-NN and LFAQ.

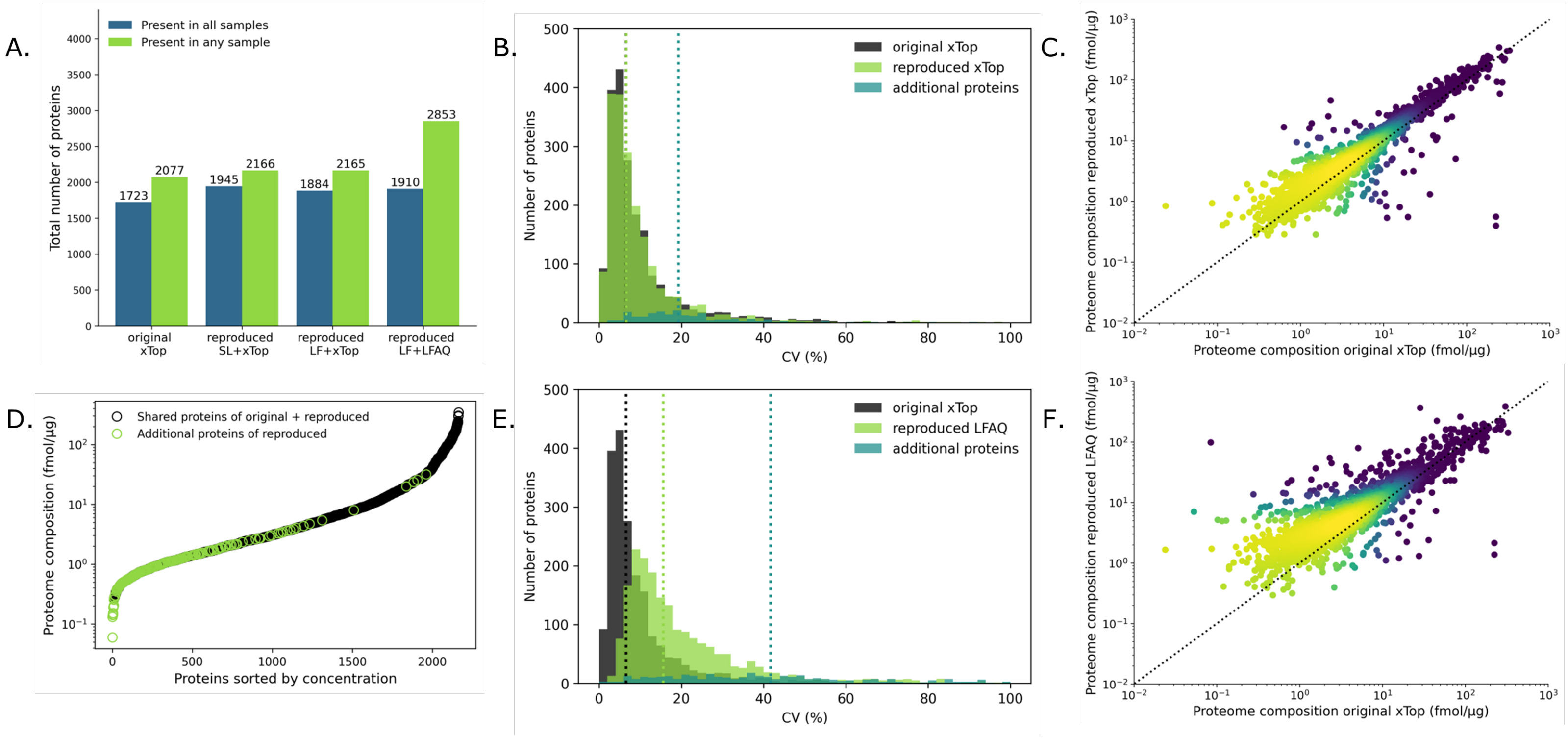


Figure S2.2: Comparison of the processed absolute quantitative proteomics data of the seven calibration samples of Mori *et al.* (2021) to the absolute quantitative proteomics data generated with autoprot and the raw MS files of the seven calibration samples of Mori *et al.* (2021) [9]. The optimal workflows were used for the reproduction, which deployed standard-free, library-free DIA data analysis using DIA-NN and xTop or LFAQ. (A) Total number of proteins absolutely quantified in all seven samples or in any sample. (B) Distribution of coefficient of variation (CV) values of shared proteins between original and xTop reproduction, and of additional proteins. The median CV values are displayed as dotted lines with the same colour as the corresponding distribution. (C) Scatter of proteome composition of shared proteins between the original and the xTop reproduction. (D) Comparison of original and xTop reproduction in terms of protein concentration range. (E) Distribution of coefficient of variation (CV) values of shared proteins between original and LFAQ reproduction, and of additional proteins. The median CV values are displayed as dotted lines with the same colour as the corresponding distribution. (F) Scatter of proteome composition of shared proteins between the original and the LFAQ reproduction.

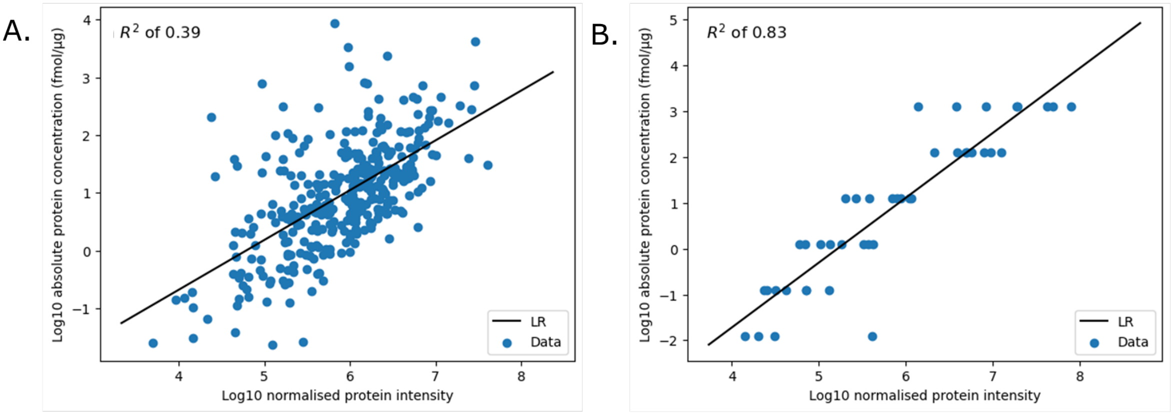


Figure S2.3: Linear regression of log-normalised protein intensities to log-normalised known protein concentrations used for absolute quantification approaches which rely on internal standards. All data stemmed from the third biological replicate of the rich media (LB) condition and was produced with a workflow deploying library-free DIA data analysis using DIA-NN and LFAQ. (A) Linear regression of *Escherichia coli* QconCAT proteins. (B) Linear regression of UPS2 protein mix.

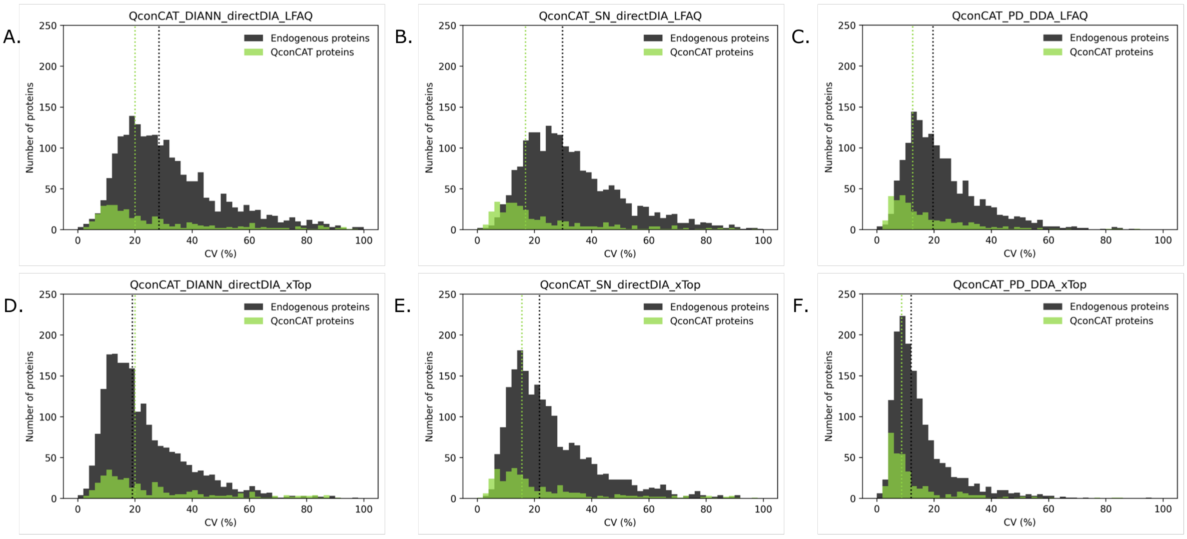


Figure S2.4: Distribution of coefficient of variation (CV) values from the data sets of workflows which incorporated *Escherichia coli* QconCAT proteins (labelled approach) for absolute quantification. The median CV value of the QconCAT proteins and the endogenous proteins is displayed as a blue and grey dotted line, respectively. (A) CV distribution of a workflow which combined library-free DIA data analysis using DIA-NN and LFAQ. (B) CV distribution of a workflow which combined library-free DIA data analysis using Spectronaut and LFAQ. (C) CV distribution of a workflow which combined DDA data analysis LFAQ. (D) CV distribution of a workflow which combined library-free DIA data analysis using DIA-NN and xTop. (E) CV distribution of a workflow which combined library-free DIA data analysis using Spectronaut and xTop. (F) CV distribution of a workflow which combined DDA data analysis xTop.

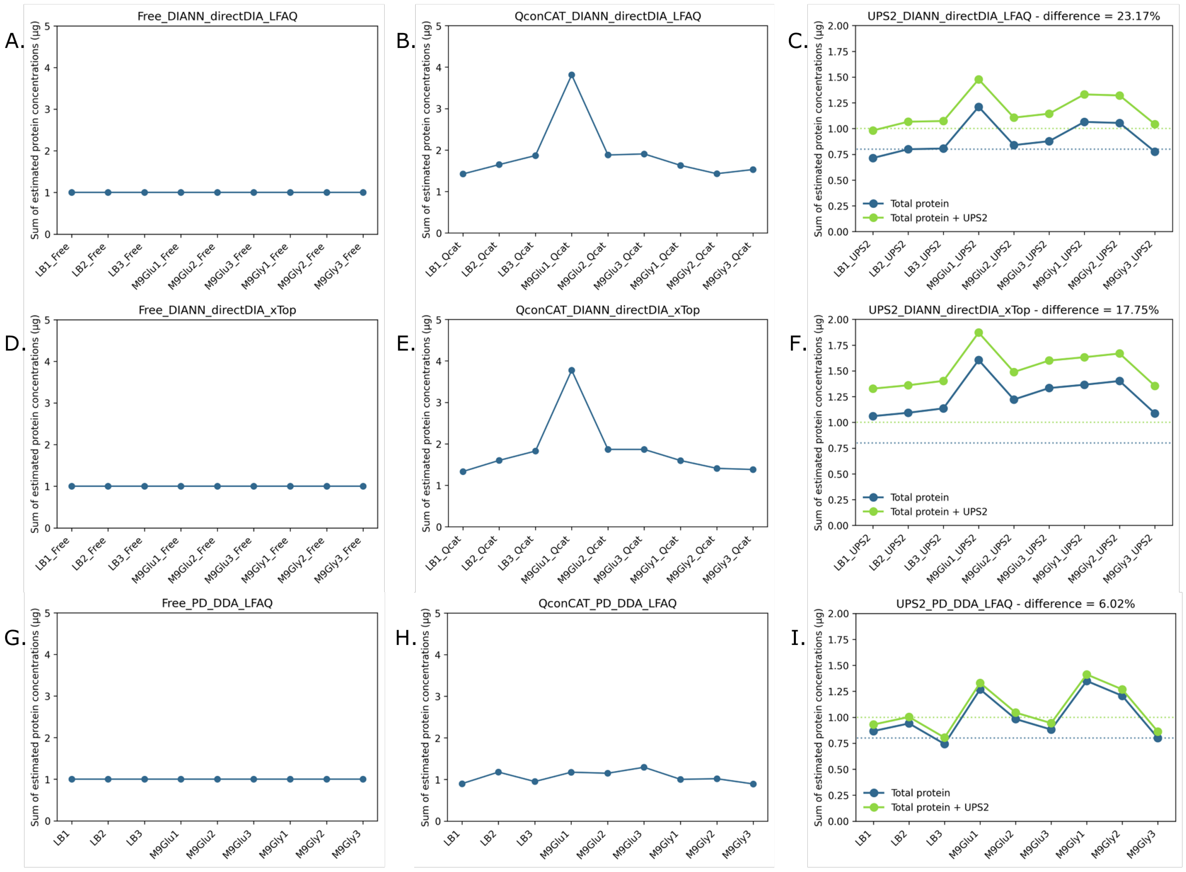


Figure S2.5: Total protein mass across all samples determined by summing all quantified proteins in a sample. (A) Total protein mass of a workflow which combined the standard-free quantification approach with library-free DIA data analysis using DIA-NN and LFAQ. (B) Total protein mass of a workflow which combined the labelled quantification approach with library-free DIA data analysis using DIA-NN and LFAQ. (C) Total protein mass of a workflow which combined the label-free quantification approach with library-free DIA data analysis using DIA-NN and LFAQ. (D) Total protein mass of a workflow which combined the standard-free quantification approach with library-free DIA data analysis using DIA-NN and xTop. (E) Total protein mass of a workflow which combined the labelled quantification approach with library-free DIA data analysis using DIA-NN and xTop. (F) Total protein mass of a workflow which combined the label-free quantification approach with library-free DIA data analysis using DIA-NN and xTop. (G) Total protein mass of a workflow which combined the standard-free quantification approach with DDA data analysis using Proteome Discoverer and LFAQ. (H) Total protein mass of a workflow which combined the labelled quantification approach with DDA data analysis using Proteome Discoverer and LFAQ. (I) Total protein mass of a workflow which combined the label-free quantification approach with DDA data analysis using Proteome Discoverer and LFAQ.

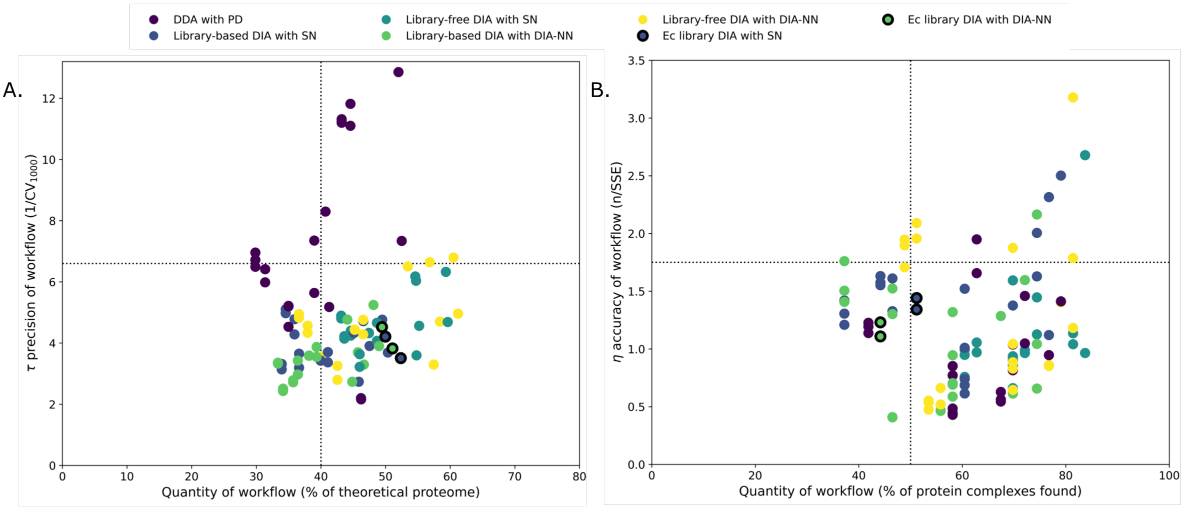


Figure S2.6: Full comparison of acquisition modes of the data sets generated using all 105 different workflows available through the autoprot pipeline with four additional workflows employing the spectral library of Mori *et al.* [9]. The data sets of the four additional workflows are highlighted with black circles. (A) τ precision comparison for the acquisition modes. The quantity values were determined as the percentage of unique *E. coli* proteins absolutely quantified compared to the total number of *E. coli* proteins in the protein database. The τ precision was determined as the inverse of the 1000th coefficient of variation (CV) value of a workflow data set. (B) η accuracy comparison for the acquisition modes. The quantity values were determined as the percentage of *E. coli* protein complexes for which both subunits were absolutely quantified compared to the total number of protein complexes. The η accuracy was determined as the inverse of the squared sum of errors (SSE) of the subunit concentrations fit to the theoretical, i.e. equimolar concentrations, normalised to the total number of protein complexes quantified.

Supplementary tables

Table S2.1: Overview of spiked-in QconCAT proteins including concentration in fmol/µg.

Peptide	Protein	Concentration	QconCAT
VAADFLAK	P00509	200	1
SVFDTLATAAK	P00805	200	1
LSEALEQVR	P00893	200	1
LDAFLASIR	P00894	200	1
LGLIEVQAPILSR	P00963	200	1
STVEAIWAGIK	P00963	200	1
FNDLLATLK	P04693	200	1
VATIQTLGGSGALK	P04693	200	1
QVSFALR	P05791	200	1
VVAGDVVVIR	P05791	200	1
AAAPAFSEESIR	P08142	200	1
DIVLAIIGK	P0A6A6	200	1
SVDGIQVGEGR	P0AB80	200	1
IAVYSSLIK	P0AC38	200	1
ISDIPEFVR	P0AC38	200	1
LSIVDVNAGAQPLYNQQK	P22106	200	1
ANVLQSSILWR	P30125	200	1
LSDAEVDELFALVK	P30126	200	1
TGFGAHLFNDWR	P30126	200	1
GLANPLTVTR	P00903	225	2
TSPITHNGEGVFR	P00903	225	2
APFSAFLR	P05041	225	2
LEQGAILSLSPER	P05041	225	2
FGEIEEVELGR	P06959	225	2
DGIASAILPGR	P08192	225	2
FQIVSESPR	P08192	225	2
NIGIYPEIK	P09394	225	2
LGVLGFEVDHER	P0A6A3	225	2
GAIFGLTR	P0A6F3	225	2
ASLILTK	P0A6I3	225	2
IPYIISIAGSVAVGK	P0A6I3	225	2
EWSFISSSLVK	P0A6I6	225	2
IFANPEEK	P0A6I9	225	2
LAVADDVIDNNGAPDAIASDVAR	P0A6I9	225	2
AWQVSDAVK	P0A9J4	225	2
QGHEVQGWLR	P0A9J4	225	2
LNAPVDEQGR	P0A9M8	225	2
TGGDAPDQTTTIVR	P0A9M8	225	2
AAATQHNLEVLASR	P0ABQ0	225	2
VIPVVEAIAQR	P0AC13	225	2
YFIEQIAR	P0AC13	225	2
FNSPWVR	P0AC16	225	2
VAEEVAELLLAR	P0AC16	225	2
WLDAPAAAALTK	P0AFC0	225	2
VQLLGSGSILR	P0AFG8	225	2
GLLLDEWR	P10908	225	2

Table S2.1: Overview of spiked-in QconCAT proteins including concentration in fmol/µg - continued.

Peptide	Protein	Concentration	QconCAT
TSNGWGVAGELNWQDLLR	P10908	225	2
GLVNATGPWVK	P13035	225	2
VSQWLVEYTQQR	P13035	225	2
IAEAAVVGIPHNIK	P27550	225	2
LVITSDEGVR	P27550	225	2
NFLAETGDIR	P31057	225	2
ADDIQIR	P31663	225	2
DLDEIITIAGQELNEK	P31663	225	2
QILNWAEAHPR	P07000	250	3
ALGVGEVK	P0A6Q3	250	3
LIYTASDLK	P0A6Q3	250	3
VLDFEEGR	P0A6Q6	250	3
SQLDWLVPHQANLR	P0A6R0	250	3
AFTDFFAR	P0A722	250	3
GTVQGGGLTK	P0A722	250	3
AVGPYVVTK	P0A953	250	3
ALIGFAGPR	P0A9Q5	250	3
ASTPLGVGGFGAAR	P0AAI5	250	3
TIFGEAASR	P0AAI5	250	3
ITFNAPTVPVVNNVDVK	P0AAI9	250	3
QLYNPVQWTK	P0AAI9	250	3
AIVGGIAR	P0ABD5	250	3
LIDSIIPEPLGGAHR	P0ABD5	250	3
AFIEVGQK	P0ABD8	250	3
SFASLSLR	P0ACU5	250	3
AGILAQVPAGR	P0AEK2	250	3
AGLIGFSK	P0AEK2	250	3
ASLEANVR	P0AEK4	250	3
VNAISAGPIR	P0AEK4	250	3
TLVLVIGESTQR	P0CB39	250	3
ATDLVLAFLPFEK	P10441	250	3
DSWYDLDAHYDALETR	P21515	250	3
FWQFYPR	P21515	250	3
TNVDLQIR	P24182	250	3
EELAQHQEGVLDIIQR	P30845	250	3
AEDFNGWLLNR	P76472	250	3
SNAASLYGWDILLAGTAWPGK	P76472	250	3
FVSGDEFANWLNQHR	P76473	250	3
VVESSSYYGK	P77398	250	3
APQAEVLTPGYK	P77690	250	3
IYTDVR	P77757	250	3
LWHEPVSPR	Q47377	250	3
DVIAEPYR	P00547	400	4
FFAALAR	P00561	400	4
GELVVLGR	P00561	400	4
ILDEAGLPGELR	P00562	400	4
NNAGETVLLGR	P00562	400	4

Table S2.1: Overview of spiked-in QconCAT proteins including concentration in fmol/µg - continued.

Peptide	Protein	Concentration	QconCAT
FLVAQSGVLITQVR	P00861	400	4
LDFVTR	P00934	400	4
VVILYPR	P00934	400	4
AAAGISETLLR	P00935	400	4
VPGTIGFATVR	P04036	400	4
LGVTTSWFDPLIGADIVK	P06721	400	4
VAEWLAEHPQVAR	P06721	400	4
GFDDSFLAPHSR	P07623	400	4
GFLAEVFGILAR	P08660	400	4
LNEGLVITQGFIGSENK	P08660	400	4
LDIAWK	P0A6K1	400	4
ELGLVATDTLR	P0A6L2	400	4
LFVEPNPIPVK	P0A6L2	400	4
SPFVTSGIR	P0A825	400	4
INDNQVIEGAESR	P0A9D8	400	4
VGINELLR	P0A9D8	400	4
ELTPAAVTGTLTTPVGR	P0A9Q9	400	4
ESGWQGYWIDAASSLR	P0A9Q9	400	4
GALDDEQLK	P0A9T0	400	4
AQVLALLEK	P0AED7	400	4
GELDFTASLR	P0AGB0	400	4
SGWLLYGR	P0AGB0	400	4
GVVGLFPANR	P13009	400	4
YVAGVLGPTNR	P13009	400	4
DNLTYIADK	P23256	400	4
GVAGLINAIR	P23256	400	4
DLLNVPSNYK	P23721	400	4
NIGPAGLTIVIVR	P23721	400	4
SWFAFALQK	P25665	400	4
EAYELVAPILTK	P00350	75	5
NLALNIESR	P00350	75	5
AAIAAAQANPNAK	P00363	75	5
LGSNSLAELVVFGR	P00363	75	5
HVLLLSR	P00864	75	5
LPVEFVPVR	P00864	75	5
AAEQYVIDEYNK	P07658	75	5
TDWQIISEIATR	P07658	75	5
GAVASLTSVAK	P09373	75	5
YPQLTIR	P09373	75	5
IIASVAEK	P0A7Z0	75	5
VSTEVDAR	P0A867	75	5
ILDWYK	P0A870	75	5
LASTWQGIR	P0A870	75	5
TWFELAPK	P0A8Q0	75	5
VLAFAQSFIGR	P0A8Q3	75	5
AAALAAADAR	P0A9Q7	75	5
LLAWLETLK	P0A9Q7	75	5

Table S2.1: Overview of spiked-in QconCAT proteins including concentration in fmol/µg - continued.

Peptide	Protein	Concentration	QconCAT
ALHLVDNTDIAR	P0AAK1	75	5
APPAPPAPAR	P0AAK1	75	5
DNLAPDLSYR	P0AC47	75	5
HVDPAAAIQQGK	P0AC47	75	5
LQPDEGVDIQVLNK	P0AC53	75	5
WAGVPFYLR	P0AC53	75	5
DQLTVTVK	P16431	75	5
VPAAVWGER	P16431	75	5
HADILLFTGAVTR	P16433	75	5
WLEAENDPR	P16433	75	5
VFLNIGDK	P21599	75	5
TEEQLANIAR	P27302	75	5
TVIGFGSPNK	P33570	75	5
VVGLELAK	P37351	75	5
WLPLPGGLR	P06282	150	6
LPIDLSQLK	P0A6B7	150	6
NAGIEVVEALK	P0A6H8	150	6
FQDLQR	P0A6L9	150	6
NILVVVPSHVFGEVLR	P0A6S7	150	6
GGIEYIEVR	P0A6W9	150	6
SLADIQQR	P0A6Z1	150	6
ADLLTLR	P0A7A7	150	6
VYQLLAELITSDVR	P0A7A7	150	6
VNLVEQLESLSVTK	P0A8K1	150	6
VTDEDLVVEIPR	P0A9R4	150	6
LLYNYLVK	P0AA84	150	6
SLQFLDGTQLDFVK	P0AAC8	150	6
VNTFLANR	P0AAC8	150	6
EIIISALR	P0ABF8	150	6
FGAFLDPVADK	P0ABF8	150	6
IGSEYHELSGR	P0ABN1	150	6
SLDEAQAIK	P0ACD4	150	6
VNDEGIIEDAR	P0ACD4	150	6
SIADQIDINK	P18956	150	6
TWLVTGSPGGSR	P18956	150	6
LQYYVNTDQLVVR	P23830	150	6
VALELVK	P27247	150	6
TSGSGNPGATNVLR	P60782	150	6
IFTGDEILPALVSTLK	P76407	150	6
IINEVNGISR	P04079	125	7
WLAQGTIYPDVIESAASATGK	P04079	125	7
EFPLPTYATSGSAGLDLR	P06968	125	7
DAGVDIDAGNALVGR	P08178	125	7
SEIIDGSK	P08178	125	7
HPAFVNR	P09029	125	7
WPETALTR	P09029	125	7
SLVLDIK	P0A720	125	7

Table S2.1: Overview of spiked-in QconCAT proteins including concentration in fmol/µg - continued.

Peptide	Protein	Concentration	QconCAT
YIVIEGLEGAGK	P0A720	125	7
DLLGATNPANALAGTLR	P0A763	125	7
NVIGNIFAR	P0A763	125	7
YVDYVLGILK	P0A7D4	125	7
GEVVLGDEFSPDGSR	P0A7D7	125	7
LEFGLYK	P0A7D7	125	7
FEDFEIEGYDPHPGIK	P0A884	125	7
HIDQITTVLNQLK	P0A884	125	7
GIFSEYGLLK	P0AB89	125	7
ELQLVYNK	P0ABA6	125	7
ILEVPVGR	P0ABB0	125	7
VVNTLGAPIDGK	P0ABB0	125	7
GVQSILQR	P0ABB4	125	7
YTLAGTEVSALLGR	P0ABB4	125	7
LVPEGIEGR	P0ADG7	125	7
DVAPGEAIYITEEGQLFTR	P0AG16	125	7
DVDQGYLDFLDTLR	P0AG16	125	7
AGLVGFSVSNLR	P15254	125	7
GSPALSAFR	P15254	125	7
AGIVEFAQALSAR	P15639	125	7
GVELLSTGGTAR	P15639	125	7
IDLTIVGPEAPLVK	P15640	125	7
IFGPTAGAAQLEGSK	P15640	125	7
DIEAWLDEGR	P28248	125	7
LIQESILR	P28903	125	7
NNLGVISLNLPR	P28903	125	7
AFLGLPVGGIR	P33221	125	7
FADSESLFR	P33221	125	7
DEAVHGYYIGYK	P37146	125	7
DAFLEHAEVFGNYYGTSR	P60546	125	7
SSLIQALLK	P60546	125	7
FLEGAAR	P68699	125	7
QPDLIPLLR	P68699	125	7
IILLGAPGAGK	P69441	125	7
NGFLLDGFPR	P69441	125	7
DWADYLFR	P69924	125	7
YQTLLDSIQGR	P69924	125	7
FFTGQITAAGK	P07001	100	8
VIGYTDLPGR	P07001	100	8
QLLGASSDR	P08395	100	8
FLSAPEAVEYGLVDSILTHR	P0A6G7	100	8
SFDIYSR	P0A6G7	100	8
IYLAETPWFNISATIIR	P0A752	100	8
GYEVIVEQQIAHELQLK	P0A7B3	100	8
VLVEGLQR	P0A9M0	100	8
ASVDAILK	P0AB67	100	8
EITAEETAELLK	P0AB67	100	8

Table S2.1: Overview of spiked-in QconCAT proteins including concentration in fmol/µg - continued.

Peptide	Protein	Concentration	QconCAT
IGLNPTGR	P0AEE3	100	8
EQFGLELPR	P0AFV4	100	8
TGDLVLFR	P0AFV4	100	8
VWVLDFK	P0AG14	100	8
NFLLTEALR	P10902	100	8
SNAVDLIVSDK	P10902	100	8
HPASTLLVAGVR	P11458	100	8
LGIDPQSK	P18133	100	8
VQLSFGIGTR	P18133	100	8
SYLQTYPFIK	P18843	100	8
YGDGGTDINPLYR	P18843	100	8
TSLDDWLR	P21369	100	8
ALVVGEPTFGK	P23865	100	8
LDDVVALIK	P23865	100	8
QQASIGTQIDK	P23898	100	8
QTEVPGFYAANYR	P24555	100	8
ENHIIASGSVR	P30011	100	8
ELLEANWYR	P32664	100	8
GVQLAEFYR	P32664	100	8
VPDLTPQQATDITAFANK	P37056	100	8
AFNLDVQR	P39099	100	8
VLPAVVSVR	P39099	100	8
QGVIIPTANVR	P75867	100	8
VILVGER	P75867	100	8
LLADIDWQALVAR	P76008	100	8
LTTPNTIVLAVR	P76008	100	8
AISVTGFR	P76176	100	8
YEIHDIEGR	P76176	100	8
GTADHVGVYVGNGK	P76190	100	8
GPLTTPVGGGIR	P08200	475	9
TVTYDFER	P08200	475	9
YYQGTPSPVK	P08200	475	9
LGADGNALFR	P0A836	475	9
IGAGPWVVK	P0A836	475	9
GISYETATFPWAASGR	P0A9P0	475	9
ATVLATGGAGR	P0AC41	475	9
LPGILELSR	P0AC41	475	9
VTGQALTVNEK	P0AC41	475	9
GYEVGGTVR	P0AFG3	475	9
LNVLVNVLGK	P0AFG3	475	9
ELLEDPTR	P0AFG6	475	9
GLVTPVLR	P0AFG6	475	9
ESAPAAAAPAAQPALAAR	P0AFG6	475	9
FAALEAAGVK	P0AGE9	475	9
SLADIGEALK	P0AGE9	475	9
SGTLTYEAVK	P0AGE9	475	9
GTFANIR	P25516	475	9

Table S2.1: Overview of spiked-in QconCAT proteins including concentration in fmol/µg - continued.

Peptide	Protein	Concentration	QconCAT
VVIAESFER	P25516	475	9
FGDDEAFEENVR	P25516	475	9
SPLLTPEK	P36683	475	9
FGLSLVR	P61889	475	9
LFGVTTLDIIR	P61889	475	9
SDLFNVNAGIVK	P61889	475	9
IYAYLSR	P06999	450	10
LTQLISAAQK	P06999	450	10
FGANAILAVSLANAK	P0A6P9	450	10
GIANSILIK	P0A6P9	450	10
IQLVGDDLFVTNTK	P0A6P9	450	10
LLSNFFAQTEALAFGK	P0A6T1	450	10
AVLHVALR	P0A6T1	450	10
ATVLGHIQR	P0A796	450	10
GGTFLGSAR	P0A796	450	10
EDLVNEIK	P0A796	450	10
LTVLDSLSK	P0A799	450	10
VATEFSETAPATLK	P0A799	450	10
ASLPTIELALK	P0A799	450	10
ADAFAVIVK	P0A858	450	10
SATPAQAQAVHK	P0A858	450	10
LDLFANEK	P0A993	450	10
TLGEFIVEK	P0A993	450	10
AGIALNDNFVK	P0A9B2	450	10
GASQNIIPSSTGAAK	P0A9B2	450	10
VLDLIAHISK	P0A9B2	450	10
VPTPNVSVVDLTVR	P0A9B2	450	10
GTIDLNLPLADNLR	P0A9C9	450	10
LIVGPGAK	P0A9C9	450	10
ANEAYLQGQLGNPK	P0AB71	450	10
DSVSYGVVK	P0AB71	450	10
GVVPQLVK	P0AD61	450	10
EITSTDDFYR	P0AD61	450	10
TLNLTALYR	P21599	450	10
GLPADVVPGDILLLDDGR	P21599	450	10
ALLEFDDQEPQLQNEIR	P23538	450	10
LEFIINR	P23538	450	10
EAYAQLSADDENASFAVR	P23538	450	10
AFFANPVLTGAVDK	P37689	450	10
ILINSPK	P37689	450	10
ELPLTESLALTIDR	P62707	450	10
VIIAAHGNSLR	P62707	450	10
HYGALQGLNK	P62707	450	10
AAVLPANLIQAQR	P00350	375	11
GYTVSIFNR	P00350	375	11
LIEPLIR	P04825	375	11
TVVTAVSQAVR	P04825	375	11

Table S2.1: Overview of spiked-in QconCAT proteins including concentration in fmol/µg - continued.

Peptide	Protein	Concentration	QconCAT
APLYPDDILWNFEK	P06610	375	11
AIAQVGTISANSDETVGK	P0A6F5	375	11
LAGGVAVIK	P0A6F5	375	11
SAGGIVLTGSAAAK	P0A6F9	375	11
VGDIVIFNDGYGVK	P0A6F9	375	11
ALQAIAGPFSQVR	P0A955	375	11
VLEVTLR	P0A955	375	11
LTSENPIDLVR	P0A991	375	11
NGLLLAR	P0A9Q1	375	11
SLIGPDGEQYK	P0A9Q1	375	11
EAADIILLEK	P0ABB8	375	11
ILTGDSELVAAK	P0ABB8	375	11
ATFVVDPQGIIQAIEVTAEGIGR	P0AE08	375	11
EGEATLAPSLDLVGK	P0AE08	375	11
IEYVYQSAEQLR	P19926	375	11
NADALTLQAPAQR	P19926	375	11
LFWLSQTPFEQR	P21179	375	11
LIPEELVPVQR	P21179	375	11
NALQELIIDGIK	P24182	375	11
SGFIFIGPK	P24182	375	11
GVFNLVLGR	P25553	375	11
IVDEIGLPR	P25553	375	11
ASLSAFDYLIR	P35340	375	11
SIIVATGAK	P35340	375	11
ALVQESIYER	P37685	375	11
ETSAADVPLAIDHFR	P37685	375	11
SGSGTLTVSNTTLTQK	P39180	375	11
LVLDGIEVVGSLVGTR	P39451	375	11
VIAIDVNDEQLK	P39451	375	11
TLWVPALK	P52697	375	11
YLIAAGQK	P52697	375	11
DLDVTATNR	P61517	375	11
TAIVEGLAQR	P63284	375	11
LGAAVATLK	P77674	375	11
FGLIPEFIGR	P0A6H1	500	12
ATGLDALFDATIK	P0A6K6	500	12
FGASSLLASLLK	P0A6L0	500	12
IATDPFVGNLTFFR	P0A6M8	500	12
YLGGEELTEAEIK	P0A6M8	500	12
FEVGEGIEK	P0A6P1	500	12
IGVLVAAK	P0A6P1	500	12
LESLVEDLVNR	P0A6Y8	500	12
GISLLDAFGAANDVLK	P0A9A6	500	12
LDEFETVGNTIR	P0A9A6	500	12
LIDGTVFDSSVAR	P0A9L3	500	12
VINQGEGAIPAR	P0A9L3	500	12
TLAEGQNVEFEIQDGQK	P0A9Y6	500	12

Table S2.1: Overview of spiked-in QconCAT proteins including concentration in fmol/µg - continued.

Peptide	Protein	Concentration	QconCAT
ALGANLVLTEGAK	P0ABK5	500	12
IQGIGAGFIPANLDLK	P0ABK5	500	12
IALESVLLGDK	P0ABP8	500	12
AAQEEEFSLELR	P0ABS1	500	12
AVQLGGVALGTTQVINSK	P0ABT2	500	12
DDDTADILTAASR	P0ABT2	500	12
EGDAVQLVGFGTFK	P0ACF0	500	12
IAAANVPAFVSGK	P0ACF0	500	12
AETLYYIVK	P0ACJ8	500	12
AGENVGVLLR	P0CE47	500	12
ADDFIEALFAR	P10121	500	12
LVIPPELAYGK	P45523	500	12
VTDAEIAEVLAR	P63284	500	12
YWDVELR	P69908	500	12
LEEDDVATPGEGQVLLR	P76113	500	12
VVGVAGGAEK	P76113	500	12
AEGALLINAVGVDDVK	P76621	500	12
LLELTFTEQTTK	P76621	500	12
TPAQIVIR	Q46857	500	12
IIDDAVAR	P04079	50	13
VSQAFTVFLPVR	P04079	50	13
FNWETLIASR	P06715	50	13
FINELLPVIDSLDR	P09372	50	13
VANLEAQLAEAQTR	P09372	50	13
LSNAQVIDVTK	P0A6W5	50	13
QNLISVNSPIAR	P0A6W5	50	13
GGYFPVPPVDSAQDIR	P0A9C5	50	13
IPVVSSPK	P0A9C5	50	13
AAFDDAIAAR	P0ABC7	50	13
NISDDLR	P0AC59	50	13
GGLDPLLK	P0AC62	50	13
FAYVDILQNPDIR	P0AC69	50	13
ADVAPSNLAIVGR	P0AEP3	50	13
YVLSADIWPLLAK	P0AEP3	50	13
AVGDSLEAQQYGIAFPK	P0AEQ3	50	13
APVAGALLIWDK	P0AES0	50	13
TWAWETAFDQIR	P0AES0	50	13
VSEISIVGR	P0AES4	50	13
VVSLIVPR	P0AES4	50	13
LELVIQR	P0AES6	50	13
HPDTLLAAR	P17169	50	13
ILVPLTIEDAQIR	P22256	50	13
SIAGGFPLAGVTGR	P22256	50	13
AFTGVGGTPLFIEK	P23893	50	13
EETFGPLAPLFR	P25526	50	13
FFSIDGTK	P31120	50	13
ILVIAADER	P31658	50	13

Table S2.1: Overview of spiked-in QconCAT proteins including concentration in fmol/µg - continued.

Peptide	Protein	Concentration	QconCAT
AEAAEINLR	P33195	50	13
LTDILAAGLQQK	P33195	50	13
FGALEYR	P37747	50	13
YQGIPVGGYTK	P37747	50	13
AAVLLADSFK	P63224	50	13
SGDWAYR	P77454	50	13
LLGQTSVDR	P03023	350	14
FGAIGIGSR	P03817	350	14
VIASELGEER	P08997	350	14
VIDGQINLR	P08997	350	14
ILDVLIPPAK	P0A6H5	350	14
LLIEEEAAK	P0A6H5	350	14
TEELLTLPANEVLWR	P0A6Y5	350	14
AQALWTR	P0A6Z3	350	14
GPVATVLVR	P0A705	350	14
TSLLDYIR	P0A705	350	14
FLEEGDK	P0A707	350	14
LTGLEGEQLGIVSLR	P0A707	350	14
LGIPYVFK	P0A715	350	14
TGAVINVK	P0A715	350	14
LYTSLGDAAVGR	P0A717	350	14
VVADFLSSVGVDR	P0A717	350	14
AEIVASFER	P0A7A9	350	14
ESGALFVDR	P0A7A9	350	14
FFINPTGR	P0A817	350	14
NIVAAGLADR	P0A817	350	14
EQEVLVR	P0A9M0	350	14
NPLFLLDEIDK	P0A9M0	350	14
SLDDFLIK	P0ACF8	350	14
YSYVDENGETK	P0ACF8	350	14
YPDLQIIGGNVATAAGAR	P0ADG7	350	14
TWEEIPALDK	P0AEX9	350	14
LSGNTASGLPAR	P12281	350	14
TILEELGEIAFWK	P12281	350	14
LPGLYYIETDSTGER	P37647	350	14
TFYYWR	P37647	350	14
SAEILDR	P37769	350	14
GALIDSQAAIEALK	P52643	350	14
TAGVIGTGK	P52643	350	14
EVIYVDSPSK	P60785	350	14
FLFDEYVR	P67910	350	14
TVSVGGEVGK	Q46938	350	14
FNDGDFLR	P00490	300	15
VFLFGAK	P00490	300	15
GDISEFAPR	P05055	300	15
AVITGDVTQIDLPR	P0A9K3	300	15
VLEQSAESVPEYGK	P0A9K3	300	15

Table S2.1: Overview of spiked-in QconCAT proteins including concentration in fmol/µg - continued.

Peptide	Protein	Concentration	QconCAT
GQGIVLNEPSVVAIR	P0A9X4	300	15
NLAEGVPR	P0A9X4	300	15
EITLEAAR	P0AFF6	300	15
GAQLFVTR	P0AFF6	300	15
VSVSIFGR	P0AFG0	300	15
GYAGDTATTSEIK	P0AFH8	300	15
LLADDIVPSR	P0AFH8	300	15
TTGISVSTR	P0AFK0	300	15
TVQAALDIAR	P0AFK0	300	15
AEITLDYQLK	P0C0L2	300	15
VDAGFAITK	P0C0L2	300	15
IIAGNIINFSR	P0C0S1	300	15
SLVNTYQEILK	P15288	300	15
ADGIGSLLPAAR	P21165	300	15
ALQLGIEASNINPK	P21165	300	15
TFLDTR	P23843	300	15
AAADEWDER	P25738	300	15
ETGQSFLDNILSR	P27298	300	15
YSDVGLVTPLR	P31677	300	15
LAGITVPDR	P33599	300	15
LLASAAQENEPFWR	P37095	300	15
LYTSSWR	P45577	300	15
QLLNEFPELK	P63020	300	15
DFGSVDNFK	P00448	425	16
EFWNVVNWDEAAAR	P00448	425	16
AAADLISR	P00452	425	16
DAPDYQYLAAR	P00452	425	16
ELANVQDLTVR	P02925	425	16
VIELQGIAGTSAAR	P02925	425	16
ISGDYAYGWLR	P07012	425	16
SYVLDDSR	P07012	425	16
ASAQLETIK	P08839	425	16
EENPFLGWR	P08839	425	16
LIDSQDVETR	P0A7E5	425	16
AGAWYSYK	P0A7G6	425	16
GYVPASTR	P0A7Z4	425	16
EIEEGLINNQILDVR	P0A800	425	16
TTVIALR	P0A800	425	16
INVFDR	P0A805	425	16
QLASVTVEDSR	P0A805	425	16
LGIQAFEPVLIEGK	P0A8T7	425	16
SVITVGPYLR	P0A8T7	425	16
LGDLPTSGQIR	P0A8V2	425	16
YDLSAVGR	P0A8V2	425	16
ELTLWESR	P0ACA3	425	16
AIVEAAGLK	P0AF93	425	16
VLDLASPIGR	P0AG30	425	16

Table S2.1: Overview of spiked-in QconCAT proteins including concentration in fmol/µg - continued.

Peptide	Protein	Concentration	QconCAT
DWQPEVK	P0AG86	425	16
LDLDTASSQLADDVYEVVLR	P0AG86	425	16
GQFAAVPLNILGDK	P23721	425	16
YGVIYAGAQK	P23721	425	16
AIGINFIDTYIR	P28304	425	16
QSGFLDPVNYR	P37194	425	16
SVDDVIAQVK	P60651	425	16
YDAVIALGTVIR	P61714	425	16
TQVNNAVSVDEK	P69797	425	16
ILEVSLDR	P77243	425	16
FTAAEFR	P00956	175	17
TGDIGLFR	P00957	175	17
IDDIDLNLEDFVQR	P00959	175	17
SDEVLSDR	P00961	175	17
VIPATILGIQSDR	P00961	175	17
SVEENLALFEK	P00962	175	17
TALYSWLFAR	P04805	175	17
APLVEELYR	P06612	175	17
SDAYFVLR	P06612	175	17
GFLQTLAR	P07118	175	17
TAISDLEVENR	P07118	175	17
VAVATIGAVLPGDFK	P07395	175	17
NWVSPVDAIVER	P07813	175	17
NYTIGDVIAR	P07813	175	17
SQAIEGLVK	P0A850	175	17
GNFDLEGLER	P0A853	175	17
FNPLDR	P0A855	175	17
LAYVTFESGR	P0A855	175	17
AQTFTLVAK	P0A862	175	17
IYTFGPTFR	P0A8M0	175	17
TNLIGAVAR	P0A8M0	175	17
VFEINR	P0A8N3	175	17
GIPTLLLFK	P0AA25	175	17
QYATTYINIVGK	P0ADG4	175	17
LANAVGIGAVK	P11875	175	17
LGLDTLGIETVER	P11875	175	17
AVEGSSFPQVALLVR	P16659	175	17
TGDIVEYLVK	P16659	175	17
FGFLLDALK	P21889	175	17
LIVVTGLSGSGK	P0A698	325	18
NEETLEPVPYFQK	P0A8F0	325	18
YQGEYVAGLAVK	P0A8G7	325	18
AVVTPLPR	P0A941	325	18
LLIDDLSFSIPK	P0A9W3	325	18
NISLSFFPGAK	P0A9W3	325	18
NADLPLAQAAIDR	P0AC02	325	18
GAWVAVVNR	P0AC03	325	18

Table S2.1: Overview of spiked-in QconCAT proteins including concentration in fmol/µg - continued.

Peptide	Protein	Concentration	QconCAT
SGDTLSAISK	P0ADE6	325	18
VLLGVAGFQAR	P0ADE8	325	18
AGLNEIR	P0ADS6	325	18
SVSLGVAQPDAYK	P0ADS6	325	18
DDTWVTLR	P0ADU5	325	18
ISDDLYVFK	P0ADU5	325	18
IVTGVTASQALLDEAVR	P0AFP6	325	18
NAIDWLSR	P0AGE6	325	18
NDIEGLATLFSNHIPDYR	P21367	325	18
NNVLALGDLAK	P21367	325	18
LEQELVLLAQR	P23839	325	18
YLETYFR	P23839	325	18
NLGLDVLDVVNLR	P25906	325	18
LGYQVVAVSGR	P26646	325	18
VVALGSAAQDK	P27250	325	18
ELVHNVALR	P32132	325	18
LDYVIPSR	P32132	325	18
ADQINEAYER	P75691	325	18
GADVLVLTSGQTDNK	P75694	325	18
FNAIGEAVK	P76177	325	18
VVADLYK	P76177	325	18
AESWLPAWPHLGGK	P76347	325	18
TVTASLANNGASDNK	P76347	325	18
TIPLQDQDTR	P77395	325	18
YQGLVAQIELLK	P77395	325	18
LVFITDTVQPVINQGGTK	P77717	325	18
AISLSGEAQSGR	P77804	325	18
ALAFAQR	P0A6D5	275	19
EADFDILEWR	P05194	275	19
EEWAGFR	P0A6E1	275	19
EIHQDLNGQLTAR	P00888	275	19
EVAHIIIDATNEPSQVISEIR	P0A6E1	275	19
EVLEALANER	P0A6D7	275	19
FADVLEK	P0A6D3	275	19
IAVDAIR	P00887	275	19
ILLIGAGGASR	P15770	275	19
LILPLAIGK	P07639	275	19
SLLLVDSALR	P00895	275	19
TIIGFGSPNK	P27302	275	19

Table S2.2: Overview of spiked-in UPS2 protein mix including concentration in fmol/µg.

Protein	Concentration	Protein	Concentration
P00915ups	1250	P01008ups	1.25
P00918ups	1250	P61769ups	1.25
P01031ups	1250	P55957ups	1.25
P69905ups	1250	O76070ups	1.25
P68871ups	1250	P08263ups	1.25
P41159ups	1250	P01344ups	1.25
P02768ups	1250	P01127ups	1.25
P62988ups	1250	P10599ups	1.25
P04040ups	125	P99999ups	0.125
P00167ups	125	P06396ups	0.125
P01133ups	125	P09211ups	0.125
P02144ups	125	P01112ups	0.125
P15559ups	125	P01579ups	0.125
P62937ups	125	P02787ups	0.125
Q06830ups	125	O00762ups	0.125
P63165ups	125	P51965ups	0.125
P00709ups	12.5	P08758ups	0.0125
P06732ups	12.5	P02741ups	0.0125
P12081ups	12.5	P05413ups	0.0125
P61626ups	12.5	P10145ups	0.0125
Q15843ups	12.5	P02788ups	0.0125
P02753ups	12.5	P10636-8ups	0.0125
P16083ups	12.5	P00441ups	0.0125
P63279ups	12.5	P01375ups	0.0125

Table S2.3: Overview of *Escherichia coli* protein complexes.

Protein complex	Subunit 1	Protein 1	Subunit 2	Protein 2
p-aminobenzoyl-glutamate hydrolase complex	abgB	P76052	abgA	P77357
Ribonucleoside-diphosphate reductase 1 complex	nrdA	P00452	nrdB	P69924
Chemotaxis phosphorelay complex CheY-CheZ	cheY	P0AE67	cheZ	P0A9H9
Succinyl-CoA synthetase	sucD	P0AGE9	sucC	P0A836
Glutamate synthase [NADPH] complex	gltD	P09832	gltB	P09831
HypAB Ni-hydrogenase maturation complex	hypB	P0AAN3	hypA	P0A700
Ribonucleoside-diphosphate reductase 2 complex	nrdF	P37146	nrdE	P39452
Imidazole glycerol phosphate synthase complex	hisH	P60595	hisF	P60664
3-isopropylmalate dehydratase complex	leuD	P30126	leuC	P0A6A6
Glycyl-tRNA synthetase complex	glyQ	P00960	glyS	P00961
Aminodeoxychorismate synthase complex	pabA	P00903	pabB	P05041

Table S2.3: Overview of *Escherichia coli* protein complexes - continued.

Protein complex	Subunit 1	Protein 1	Subunit 2	Protein 2
tRNA uridine 5-carboxymethylamino-methyl modification complex	mnmG	P0A6U3	mnmE	P25522
Sulfate adenylyltransferase complex	cysD	P21156	cysN	P23845
Glutathione/cysteine ABC exporter complex	cydC	P23886	cydD	P29018
Glucose-specific enzyme II complex	crr	P69783	ptsG	P69786
TrpAB tryptophan synthase complex	trpA	P0A877	trpB	P0A879
Anthranilate synthase complex	trpE	P00895	trpGD	P00904
NapAB nitrate reductase complex	napB	P0ABL3	napA	P33937
Acetolactate synthase II complex	ilvG	P0DP90	ilvM	P0ADG1
Acetolactate synthase I complex	ilvN	P0ADF8	ilvB	P08142
Acetolactate synthase III complex	ilvI	P00893	ilvH	P00894
fadBA fatty acid oxidation complex, aerobic conditions	fadA	P21151	fadB	P21177
fadJI fatty acid oxidation complex, anaerobic conditions	fadJ	P77399	fadI	P76503
L-tartrate dehydratase complex	ttdA	P05847	ttdB	P0AC35
NAD-dependent dihydropyrimidine dehydrogenase complex	preT	P76440	preA	P25889
NAD(P) transhydrogenase complex	pntB	P0AB67	pntA	P07001
Enoyl CoA hydratase/isomerase complex	paaF	P76082	paaG	P77467
entAE 2,3-dihydroxybenzoate-AMP ligase complex	entA	P15047	entE	P10378
entBE aryl carrier complex	entB	P0ADI4	entE	P10378
GyrA-GyrB DNA Gyrase complex	gyrB	P0AES6	gyrA	P0AES4
Ethanolamine ammonia-lyase complex	eutC	P19636	eutB	P0AEJ6
Molybdopterin-synthase adenylyltransferase complex	moaD	P30748	moeB	P12282
Molybdopterin synthase	moaD	P30748	moaE	P30749
Aspartate carbamoyltransferase complex	pyrB	P0A786	pyrI	P0A7F3
dnaB-dnaC complex	dnaB	P0ACB0	dnaC	P0AEF0
Carbamoyl phosphate synthetase complex	carA	P0A6F1	carB	P00968
dnaA-diaA complex	diaA	P66817	dnaA	P03004
BtuCD complex	btuD	P06611	btuC	P06609
ThiF-ThiS complex	thiF	P30138	thiS	O32583
thiG-thiH thiazole phosphate synthase complex	thiH	P30140	thiG	P30139
iscS-tusA cysteine desulfurase complex	iscS	P0A6B7	tusA	P0A890
iscS-iscU iron-sulfur cluster assembly complex	iscS	P0A6B7	iscU	P0ACD4
tRNA-specific 2-thiouridylase tusE-mnmA complex	tusE	P0AB18	mnmA	P25745
UvrAB DNA damage sensor complex	uvrA	P0A698	uvrB	P0A8F8
Holo-translocon SecYEG-SecDF-YajC-YidC complex	secF	P0AG93	secD	P0AG90
FtsEX ABC cell division complex	ftsX	P0AC30	ftsE	P0A9R7

References

[1] V. Demichev, C. B. Messner, S. I. Vernardis, K. S. Lilley, and M. Ralser. “DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput”. In: *Nature Methods* 17 (2020), pp. 41–44. DOI: 10.1038/s41592-019-0638-x.

[2] J. R. Wiśniewski, M. Y. Hein, J. Cox, and M. Mann. “A "Proteomic Ruler" for Protein Copy Number and Concentration Estimation without Spike-in Standards”. In: *Molecular and Cellular Proteomics* 13.12 (2014), pp. 3497–3506. DOI: 10.1074/mcp.M113.037309.

[3] P. J. A. Cock, T. Antao, J. T. Chang, B. A. Chapman, C. J. Cox, A. Dalke, I. Friedberg, T. Hamelryck, F. Kauff, B. Wilczynski, and M. J. L. de Hoon. “Biopython: freely available Python tools for computational molecular biology and bioinformatics”. In: *Bioinformatics* 25.11 (2009), pp. 1422–1423. DOI: 10.1093/bioinformatics/btp163.

[4] J. C. Silva, M. V. Gorenstein, G. Z. Li, J. P. C. Vissers, and S. J. Geromanos. “Absolute Quantification of Protein by LCMSE: A Virtue of Parallel MS Acquisition”. In: *Molecular and Cellular Proteomics* 5.1 (2006), pp. 144–156. DOI: 10.1074/mcp.M500230-MCP200.

[5] B. Schwanhäusser, D. Busse, N. Li, G. Dittmar, J. Schuchhardt, J. Wolf, W. Chen, and M. Selbach. “Global quantification of mammalian gene expression control”. In: *Nature* 473 (2011), pp. 337–342. DOI: 10.1038/nature10098.

[6] P. Lu, C. Vogel, R. Wang, X. Yao, and E. M. Marcotte. “Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation”. In: *Nature Biotechnology* 25.1 (2007), pp. 117–124. DOI: 10.1038/nbt1270.

[7] B. Zybailov, A. L. Mosley, M. E. Sardiu, M. K. Coleman, L. Florens, and M. P. Washburn. “Statistical Analysis of Membrane Proteome Expression Changes in *Saccharomyces cerevisiae*”. In: *Journal of Proteome Research* 5 (2006), pp. 2339–2347. DOI: 10.1021/pr060161n.

[8] Y. Chen, J. M. Guenther, J. W. Gin, L. J. G. Chan, Z. Costello, T. L. Ogorzalek, H. M. Tran, J. M. Blake-Hedges, J. D. Keasling, P. D. Adams, H. García Martín, N. J. Hillson, and C. J. Petzold. “Automated "Cells-To-Peptides" Sample Preparation Workflow for High-Throughput Quantitative Proteomic Assays for Microbes”. In: *Journal of Proteome Research* 18 (2019), pp. 3752–3761. DOI: 10.1021/acs.jproteome.9b00455.

[9] M. Mori, Z. Zhang, A. Banaei-Esfahani, J. B. Lalanne, H. Okano, B. C. Collins, A. Schmidt, O. T. Schubert, D. S. Lee, G. W. Li, R. Aebersold, T. Hwa, and C. Ludwig. “From coarse to fine: The absolute *Escherichia coli* proteome under diverse growth conditions”. In: *Molecular Systems Biology* 17.5 (2021), e9536. DOI: 10.15252/msb.20209536.

Chapter 3.

High-throughput quantitative membrane proteomics of gram-negative bacteria - optimising the sample preparation

Shannara Kayleigh Taylor Parkins, Nicolás Gurdo, Elli Dertili, Martina Fricano, Pablo Iván Nikel and Lars Keld Nielsen

The quantification of membrane proteins is essential for several applications, such as increased understanding of cellular processes, biomarker validation and product transport optimisation. Recent years have seen considerable advancements in the quantification of cytosolic proteins, yet similar advancements have not occurred in quantitative membrane proteomics, especially for gram-negative bacteria. This is due to the multi-layered and mostly hydrophobic cell envelope of gram-negative bacteria, which hinders efficient solubilisation of the membrane proteins embedded within and results in many membrane proteins evading adequate detection. Here, we developed a sample preparation protocol combining multiple detergents for enhanced solubilisation with DIA analysis, which resulted in 794 quantified *E. coli* membrane proteins (56% of theoretical membrane proteome) and 485 quantified *P.putida* membrane proteins (61% of theoretical membrane proteome). The detergent protocol outperformed both our standard reference protocol and a protocol based on membrane protein enrichment through ultracentrifugation, while maintaining satisfactory precision and throughput. The optimal method, i.e. the detergent protocol with DIA analysis, produced a high number of both cytosolic and membrane proteins, thereby allowing for proteome-wide quantitative analysis with a stronger representation of the membrane proteome.

3.1. Introduction

Membrane proteins play a crucial role in many cellular processes, including energy conversion, signal transduction and metabolite transport. Systems biology aims to understand these cellular processes by means of computational and mathematical analyses and thus heavily relies on quantitative data. Despite many advancements in quantitative proteomics, quantitative membrane proteomics still lags behind, especially absolute quantification of membrane proteins [1, 2]. Workflows from quantitative proteomics must be adapted to accommodate membrane proteins during both sample preparation and HPLC-MS analysis.

Membrane protein extraction remains challenging due to the subcellular location, hydrophobic nature and presumed low abundance of membrane proteins. The cell envelope of gram-negative bacteria consists of multiple layers: the outer membrane, the periplasmic space and the inner membrane. Within these layers reside lipoproteins and peripheral membrane proteins which are anchored in the membrane, as well as integral membrane proteins which are embedded within the membrane [3]. Due to the attachments, membrane proteins are often physical inaccessible and difficult to release from the cell envelope during sample preparation. Since the membranes consist of phospholipids, membrane proteins are mostly hydrophobic which also challenges solubilisation during sample preparation. Membranes have a finite surface area which presumably results in low abundances of membrane proteins relative to cytosolic proteins. Even after enrichment, membrane proteins often seem to dwell around the lower limits of detection of HPLC-MS analysis [1, 4].

Due to the more complex cell envelope structure of gram-negative bacteria, previous membrane proteomics studies of these micro-organisms focus on a specific section or protein family only. Since inner membrane vesicles are relatively easy to prepare, multiple studies have only reported on the inner membrane proteome instead of the full membrane proteome [4, 5]. Studies on the quantification rather than merely identification of membrane proteins in gram-negative bacteria are even more limited due to inadequate methods [1, 6]. Common sample preparation protocols incorporate detergents for increased solubilisation, enrichment by ultracentrifugation, or a combination of these methods [6, 7], which introduces additional noise to the quantification. A disadvantage of deploying detergents is the required clean-up after digestion in order to avoid interference during HPLC-MS analysis. Ultracentrifugation enables strong enrichment of membrane proteins and thus increased detection, yet it greatly reduces the throughput and reproducibility of the method [1]. An ideal method would combine high throughput and reproducibility with high coverage in the quantification of all types of membrane proteins.

In this study, we compared several sample preparation protocols for quantification of the *Escherichia coli* and *Pseudomonas putida* membrane proteome. Each protocol incorporated a different strategy to aid the solubilisation or enrichment of membrane proteins by adding additional steps on top of a reference protocol.

In one protocol, an enzyme mix was added to further degrade cell debris after cell lysis in order to expose more proteins from the membranes. Additional detergents with differing chemical properties to the reference protocol detergent were added to the cell lysis buffer to investigate the extent of solubilisation for the second protocol. Lastly, a traditional ultracentrifugation protocol was applied where all sample fractions were subjected to HPLC-MS analysis to investigate the extent of membrane protein enrichment. Although the ultracentrifugation protocol resulted in a large enrichment of membrane proteins, the protocol deploying additional detergents provided the highest number of quantified membrane proteins with comparable precision to the reference protocol. The optimal method combined the detergent protocol with DIA analysis and resulted in 794 quantified membrane proteins (56% of theoretical membrane proteome) and 485 quantified membrane proteins (61% of theoretical membrane proteome) for *E. coli* and *P. putida*, respectively.

3.2. Material and methods

3.2.1. Bacterial strains and culture conditions

Both *E. coli* MG1655 and *P. putida* KT2440 were cultivated in M9 minimal medium with 4 g/L glucose. Each culture flask was inoculated from lysogeny broth (LB) agar plates with a colony from the corresponding organism and grown overnight into stationary phase. Culture flasks with fresh media were inoculated with the corresponding overnight culture and grown until mid-exponential phase prior to sampling. Either 1 mL or 5 mL was sampled depending on the intended sample preparation protocol, since the ultracentrifugation protocol required more starting material. Samples were immediately centrifuged at 15,000 g at -5°C for 5 minutes, the supernatant was discarded and the remaining cell pellets were frozen at -80°C until further processing.

3.2.2. Proteomics sample preparation

The standard sample preparation protocol for proteomics analysis, here the reference protocol, was performed as previously described in (REF Chapter1) and additional steps from the protocols focused on membrane proteomics analysis are mentioned below. All protocols were performed in triplicate and can be found in the Supplementary materials section.

Enzymatic protocol
Three different enzymes were deployed to increase the degradation of the cell envelope: lipase and phospholipase C to degrade phospholipids and lysozyme to degrade peptidoglycan. Lipase from *Candida rugosa* (Sigma Aldrich) and phospholipase C from *Clostridium perfringens* (Sigma Aldrich) were dissolved and diluted in HEPES buffer to reach a concentration of 0.08 U/µL and 0.05 U/µL, respectively. Lysozyme from chicken egg white (Sigma Aldrich) was dissolved and diluted in EDTA buffer to reach a concentration of 0.08 U/µL. After cell lysis, samples were spun down rather than centrifuged at high speed to keep cell debris in solution to subject it to further degradation.

For each sample, 10 µL of each enzyme solution was added to 20 µg of protein and incubated for 1 hour at 37°C. The incubation conditions for the enzymatic degradation were partly taken from Perczyk and Broniatowski [8]. Subsequent protein digestion and desalting was performed as in the reference protocol.

Detergent protocol
To increase the solubilisation of membrane proteins during cell lysis, 1.2% (w/v) NP-40 or 1.2% (w/v) Triton X-100 was added as additional detergent to the lysis buffer of the reference protocol, namely 6 M guanidinium·HCl, 5 mM *tris*(2-carboxyethyl)phosphine, 10 mM chloroacetamide and 100 mM Tris·HCl (pH = 8.5). The SP4 clean-up protocol [9], which uses glass beads for protein immobilisation, was applied to remove these detergents prior to HPLC-MS analysis, since NP-40 and Triton X-100 are known to cause interference during HPLC-MS analysis. The SP4 clean-up was performed after cell lysis and protein concentration determination and the subsequent protein digestion was executed with shaking at 1,000 rpm, since the glass beads were not removed beforehand as recommended by the authors [9]. The glass beads were pelleted by centrifugation and the supernatant including the peptides was desalted as in the reference protocol.

Ultracentrifugation protocol
For sample fractionation and membrane protein enrichment, two rounds of ultracentrifugation were performed based on the method described previously by Roma-Rodrigues *et al.* [10] and any modifications are mentioned below. Cell lysis was performed following the reference protocol with larger cell pellets, total OD of 5 instead of 1, and all resulting protein mass was subjected to ultracentrifugation. The samples were diluted in EDTA buffer up to a volume of 1.5 mL to allow for ultracentrifugation in tubes with a 4 mL maximum capacity. After the first round of ultracentrifugation, the pellet was resuspended in 200 µL lysis buffer and diluted in EDTA buffer up to a volume of 1.5 mL. The supernatants of both ultracentrifugation rounds, i.e. supernatant and wash samples, were kept and analysed alongside the pellet sample. The final pellet after ultracentrifugation was resuspended in 200 µL lysis buffer. The protein in both the supernatant and the wash samples was concentrated using trichloroacetic acid (TCA) precipitation [11]. Briefly, the protein was precipitated using 100% TCA, incubated on ice and subsequently washed with ice-cold acetone. After overnight incubation at -20°C, the acetone was removed, the protein resuspended in 200 µL lysis buffer and incubated at 60°C to aid solubilisation. All samples, supernatant, wash and pellet, were diluted with 200 µL 50 mM ammonium bicarbonate and subjected to protein concentration determination. Protein digestion and desalting were performed following the reference protocol.

3.2.3. HPLC-MS analysis

HPLC-MS analysis of the samples was performed on an Orbitrap Exploris 480 instrument (Thermo Fisher Scientific) preceded by an EASY-nLC 1200 HPLC system (Thermo Fisher Scientific). For each sample, 1 µg of peptides was captured on a 2-cm C18 trap column (Thermo Fisher 164946).

Subsequently separation was executed using a 70 minute gradient from 8% (v/v) to 48% (v/v) of acetonitrile in 0.1% (v/v) formic acid on a 15-cm C18 reverse-phase analytical column (Thermo EasySpray ES904) at a flow rate of 250 nL/min. The mass spectrometer was operated in either data-dependent or data-independent acquisition mode with the specific settings listed below.

DDA
For data-dependent acquisition (DDA), the mass spectrometer was run with a DD-MS2 method preceded by the FAIMS Pro Interface (Thermo Fisher Scientific) with alternating CV of -50 V or -70 V. Full MS1 spectra were collected at a resolution of 60,000 and scan range of 375-1,500 m/z, with the maximum injection time set to auto, an intensity threshold of $5.0 \cdot 10^3$, and a dynamic exclusion at 45 s. MS2 spectra were obtained at a resolution of 15,000, an isolation window of 1.6 m/z, the HCD collision energy set to 28%, the maximum injection time set to auto.

DIA
For data-independent acquisition (DIA), the mass spectrometer was run with the HRMS1 method as previously described [12] preceded by the FAIMS Pro Interface (Thermo Fisher Scientific) with a compensation voltage (CV) of -45 V, and any modifications are mentioned below. Full MS1 spectra were collected at a resolution of 120,000 and scan range of 400-1,000 m/z, with the maximum injection time set to auto. MS2 spectra were obtained at a resolution of 60,000, with the maximum injection time set to auto, and the collision energy set to 32. Each cycle consisted of three DIA experiments each covering a range of 200 m/z with a window size of 6 m/z and a 1 m/z overlap, while a full MS scan was obtained in between experiments.

3.2.4. Protein identification and quantification

For all analyses, sequence identification was performed using a protein database consisting of the *E. coli* (UP000000625) or the *P. putida* (UP000000556) reference proteome. The annotation of membrane proteins for both organisms was based on the subcellular location tags of each protein in the corresponding reference proteome (see Table S3.1 and S3.2 for full details). The full data analysis workflow was executed with the *autoprot* pipeline (REF Chapter1) for both DDA and DIA analyses to obtain proteome composition values for all samples. For DDA data analysis, the raw MS files were first processed with Proteome Discoverer v2.4 (Thermo Fisher Scientific) and the peptide-based results table was exported. Next, the *autoprot* pipeline was deployed with the following settings: the approach set to "free" for standard-free quantification approach, the mode set to "DDA" for DDA data analysis, the corresponding reference proteome as "fasta", the peptide-based results table as "DDAresultsFile" and xTop [13] as protein inference algorithm. Since the raw MS file of replicate 1 of the cytosolic fraction of the *E. coli* ultracentrifugation sample contained substantially less total signal (Figure S3.1A), the sample was excluded from further analysis.

For DIA data analysis, the *autoprot* pipeline was deployed with the following settings: the "osDIA" flag set for using DIA-NN v1.8 [14], the approach set to "free" for standard-free quantification approach, the mode set to "directDIA" for library-free DIA data analysis, the corresponding reference proteome as "fasta" and xTop [13] as protein inference algorithm.

3.2.5. Development of membrane peptide mix

The membrane peptide mix was constructed using *in silico* tryptic digestion and peptide detection prediction based on strongly expressed membrane proteins. Transcriptomics data from the reference condition of the PRECISE database [15], i.e. *E. coli* MG1655 grown on M9 minimal media with 2 g/L glucose, was used to identify the highest 25% of expressed membrane proteins. Tryptic peptides of these membrane proteins were obtained by *in silico* digestion with trypsin, while allowing for 0 missed cleavages and a peptide length between 9 and 15 amino acids. Only unique peptides were subjected to the peptide detectability prediction using DeepMSPeptide [16] and peptides with a detection probability above 90% were selected subsequently. The resulting list of membrane peptides contained 284 peptides from 110 membrane proteins (see Table S3.3 for full details). The membrane peptide mix was ordered from JPT as an unlabelled SpikeMix peptide pool and subjected to HPLC-MS analysis using the DDA method described above. Out of the 24 membrane peptides, 228 were detected and included in the final membrane peptide mix used here as a metric in the sample preparation protocol comparison.

3.2.6. Data availability

The mass spectrometry proteomics data will be deposited to the ProteomeXchange Consortium (`http://proteomecentral.proteomexchange.org`) via the PRIDE [17] partner repository upon journal submission. The proteome-wide quantification results obtained with the detergent protocol for sample preparation and DIA analysis can be found in Table S3.4 and S3.5 for *E. coli* and *P. putida*, respectively.

3.3. Results

3.3.1. Development of protocols

For fair comparison of the sample preparation protocols, all samples were taken from one cultivation, i.e. from the same biological replicate of the respective organism, to limit variance which did not stem from differences between the protocols. The three sample preparation protocols which focused on membrane proteomics analysis were based on the reference protocol and include additional steps targeting membrane protein solubilisation or enrichment (Figure 3.1). The enzymatic protocol aimed to decompose the remaining phospholipids and peptidoglycan in the cell debris after the cell lysis step, in order to expose more membrane proteins to solubilisation. To this end, the samples were incubated with a mix of lipase, phospholipase C and lysozyme before the digestion step. Based on previous protocols [7], NP-40 and Triton X-100 are effective non-ionic detergents for membrane protein extraction.

These detergents form micelles which mimic the phospholipid environment of the cell membrane, potentially increasing the solubilisation of membrane proteins during cell lysis. The detergent protocol used an enhanced lysis buffer during cell lysis, with either NP-40 or Triton X-100 in addition to guanidinium·HCl found in the reference protocol. The SP4 clean-up method [9] was required for this protocol to properly remove the detergents which would otherwise contaminate and interfere during HPLC-MS analysis. For the ultracentrifugation protocol, two rounds of ultracentrifugation were included to extract the majority of cytosolic proteins in order to enrich the membrane protein fraction, based on previous protocols [2, 10]. The enriched membrane proteins were expected in the pellet sample, however the two supernatants of the ultracentrifugation workflow were precipitated and processed further alongside the other samples to assess the enrichment efficiency. Due to the limited capacity of the ultracentrifuge and the extent of the additional steps, the ultracentrifugation protocol was low-throughput compared to the other protocols.

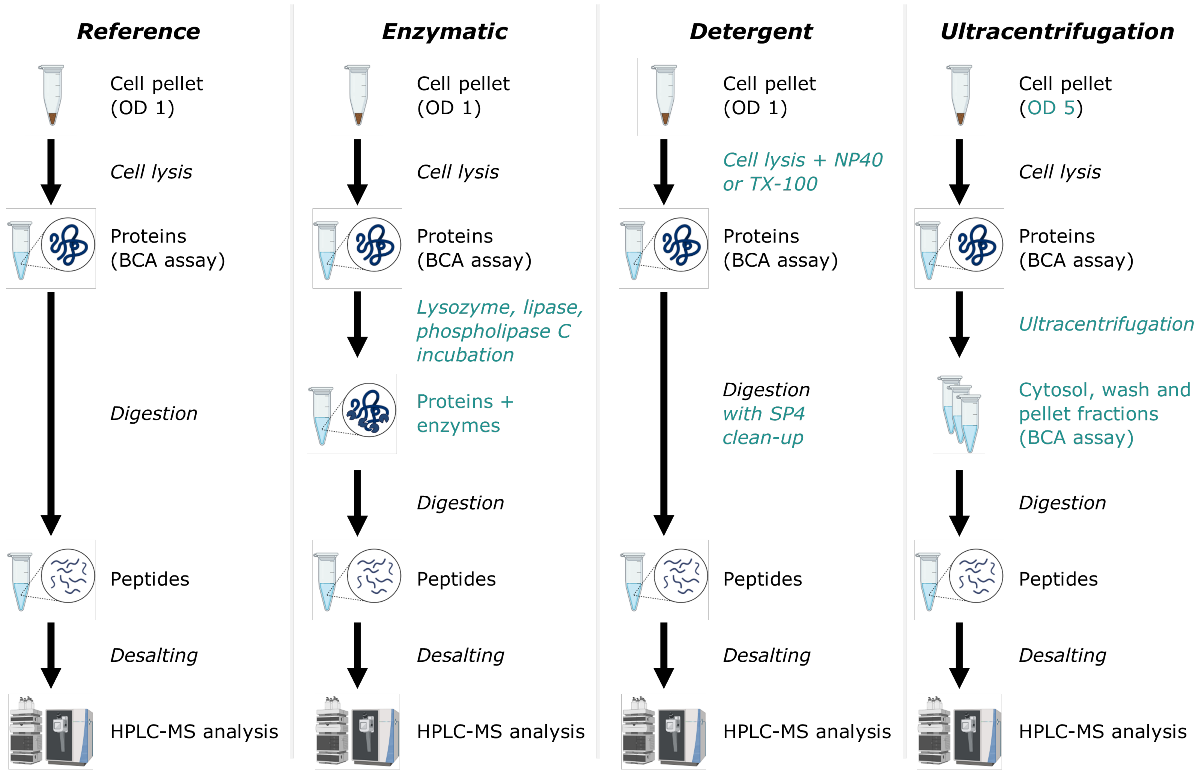


Figure 3.1: Overview of the sample preparation protocols. The reference protocol is a standard quantitative proteomics protocol, with no additional steps to increase membrane protein quantification. The enzymatic protocol contains an additional incubation step with an enzyme mix after the cell lysis. The enzyme mix consisted of lipase, phospholipase C and lysozyme to further degrade cell debris in order to expose more membrane proteins. The detergent protocol deployed NP-40 or Triton X-100 as additional detergent during cell lysis to aid in solubilisation of membrane proteins. The additional detergents require subsequent removal with the SP4 clean-up, which is applied in combination with protein digestion. The ultracentrifugation protocol includes two rounds of ultracentrifugation after cell lysis for membrane protein enrichment. Icons created with BioRender.com.

Previous membrane proteomics studies have been performed with DDA methods for HPLC-MS analysis, yet DIA methods have quickly taken over the proteomics field, showing increased performance compared with DDA [18]. Hence, the current study deployed both DDA and DIA analysis to investigate the difference between acquisition modes for membrane proteomics analysis specifically. The experimental data of all sample preparation protocols and for both organisms was highly consistent with a median Pearson correlation of 0.96 between technical replicates (Figure S3.1). Notably, the pellet sample of the ultracentrifugation protocol differed the most from other samples for the membrane protein enrichment, as expected. Based on the substantially lower total MS signal of replicate 1 of the DDA analysis after the ultracentrifugation protocol compared to the other replicates, replicate 1 was excluded from further analyses.

3.3.2. Overall comparison

Both the number of quantified membrane proteins and the precision, here the inverse of the median coefficient of variation (CV) value, are relevant for membrane proteomics analysis and were thus used to determine the optimal sample preparation protocol and acquisition mode (Figure 3.2). As expected, the data from DDA analyses resulted in higher precision, while the data from DIA analyses resulted in substantially greater numbers of quantified membrane and cytosolic proteins. Even for DIA analysis, the median CV value was a modest 20-25% and we expect most groups will select DIA analysis for proteomics analysis of the cell envelope.

The detergent protocol using either NP-40 or Triton X-100 produced the highest number of quantified membrane proteins for both organisms, with NP-40 slightly outperforming Triton X-100 for *E. coli* membrane proteins. For *E. coli*, 794 membrane proteins (56% of theoretical membrane proteome) were quantified applying the NP-40 detergent protocol in combination with DIA analysis, while 485 membrane proteins (61% of theoretical membrane proteome) were quantified for *P. putida.* Since the whole theoretical membrane proteome is not expected to be expressed under one experimental condition, transcriptomics data from the reference condition of the PRECISE database [15] was used determine the expressed membrane proteome of *E. coli* under comparable experimental conditions. Out of 1,431 theoretical *E. coli* membrane proteins, 1,314 membrane proteins had non-zero expression levels, thus 60% of the expressed membrane proteome of *E. coli* was quantified using the optimal method. Out of 2,972 theoretical cytosolic proteins, 2,555 cytosolic proteins had non-zero expression levels and 73% of the expressed cytosolic proteome of *E. coli* was therefore quantified using the optimal method. The expression levels and proteome composition values correlated well for both membrane and cytosolic proteins (Figure S3.2), even though the carbon source differed from 2 g/L to 4 g/L of glucose between experimental conditions of the transcriptomics and proteomics analysis.

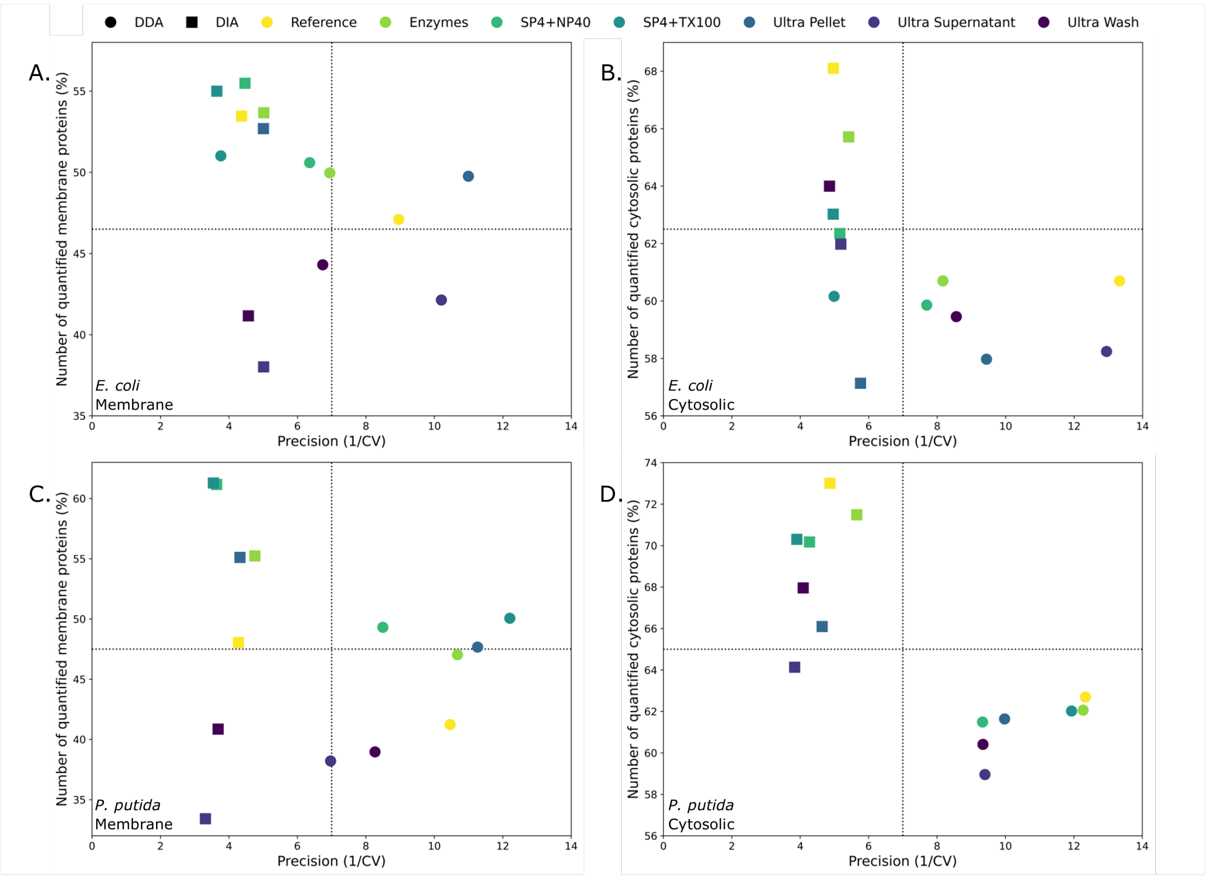


Figure 3.2: Comparison of sample preparation protocols and acquisition modes in terms of quantity and precision. The quantity is defined as a percentage of the theoretical number of either membrane or cytosolic proteins. The theoretical proteome of *E. coli* consists of 1,431 membrane proteins and 2,972 cytosolic proteins, while the theoretical *P. putida* proteome consists of 793 membrane proteins and 4,743 cytosolic proteins. The precision is defined as the inverse of the median coefficient of variation (CV) value. (A) Comparison for *E. coli* membrane proteins. (B) Comparison for *E. coli* cytosolic proteins. (C) Comparison for *P. putida* membrane proteins. (D) Comparison for *P. putida* cytosolic proteins.

To obtain a metric which could capture the membrane protein extraction efficiency, a mix of *E. coli* membrane peptides with a high probability of detection within the biological samples was developed. Using the transcriptomics data from the reference condition of the PRECISE database [15], the highest 25% of expressed membrane proteins was determined and subjected to *in silico* tryptic digestion and peptide detection prediction. The detection of 228 membrane peptides of the resulting mix was verified by HPLC-MS and used for the metric. The detergent protocol combined with DIA analysis yielded the identification of 75% of the membrane peptides (Figure S3.3) and confirmed the optimal method for membrane protein extraction.

Although the highest number of quantified cytosolic proteins was obtained with the reference protocol, a comparable number was achieved with the detergent protocol with either NP-40 or Triton X-100.

The CV values and concentration ranges were comparable between membrane and cytosolic proteins for the detergent protocol with DIA analysis applied to either organism (Figure S3.4 and S3.5). The comparable concentration range for membrane and cytosolic proteins was interesting in particular, since membrane proteins are assumed to be present in low abundance relative to cytosolic proteins due to cell surface constraints [19]. For the purpose of proteome-wide analysis with a focus on maximising the number of quantified membrane proteins, the detergent protocol combined with DIA analysis would provide optimal results without compromising the quantification of cytosolic proteins.

3.3.3. Composition analysis

To assess the extent of membrane protein solubilisation and detection, the proteome composition values of all samples were estimated using the standard-free total protein approach (TPA) method [20, 21] (Figure 3.3). An increased membrane protein fraction was indeed found for all developed protocols compared to the reference protocol. Notably, the substantially increased membrane fraction of the membrane protein enrichment through the ultracentrifugation protocol did not result in a higher number of quantified membrane proteins compared to the other protocols. The increase in membrane fraction for the detergent protocol could result from either the greater solubilisation of membrane proteins through the micelles of the additional detergent or a certain selectivity for membrane proteins with the glass beads used during the SP4 clean-up. Compared with the reference protocol, there was an overall minimal shift in proteome composition values for both membrane and cytosolic proteins (Figure S3.6), which suggests that the glass beads are not selective for specific proteins.

The proteome composition values of the membrane protein fraction grouped by subcellular location confirmed the extent of sample fractionation from the ultracentrifugation protocol (Figure 3.4). As expected, the pellet sample captured mostly membrane proteins associated with the inner and outer membrane, while the supernatant captured mostly periplasmic membrane proteins alongside the cytosolic proteins. The detergent protocol with either NP-40 or Triton X-100 also quantified a large fraction of membrane proteins associated with the inner and outer membrane, which supports the proposition that the detergents improve the solubilisation of membrane proteins. It should be noted that the subcellular location annotation of membrane proteins is less extensive for *P. putida* than for *E. coli*, which resulted in a larger "other" fraction of membrane proteins. Although the proteome composition values showed substantial differences between protocols, the number of quantified membrane proteins annotated with subcellular location showed a similar distribution for all protocols (Figure S3.7). Measurements of protein concentration were performed before and after the ultracentrifugation rounds for the ultracentrifugation protocol, which resulted in a comparable total protein amount of the unfractionated and fractionated samples of both organisms (Figure S3.8A and S3.8D). When sample fractions were compared for membrane and cytosolic proteins separately, the results confirmed that ultracentrifugation enriches for membrane and cytosolic proteins in the pellet and supernatant, respectively (Figure S3.8B, S3.8C, S3.8E and S3.8F).

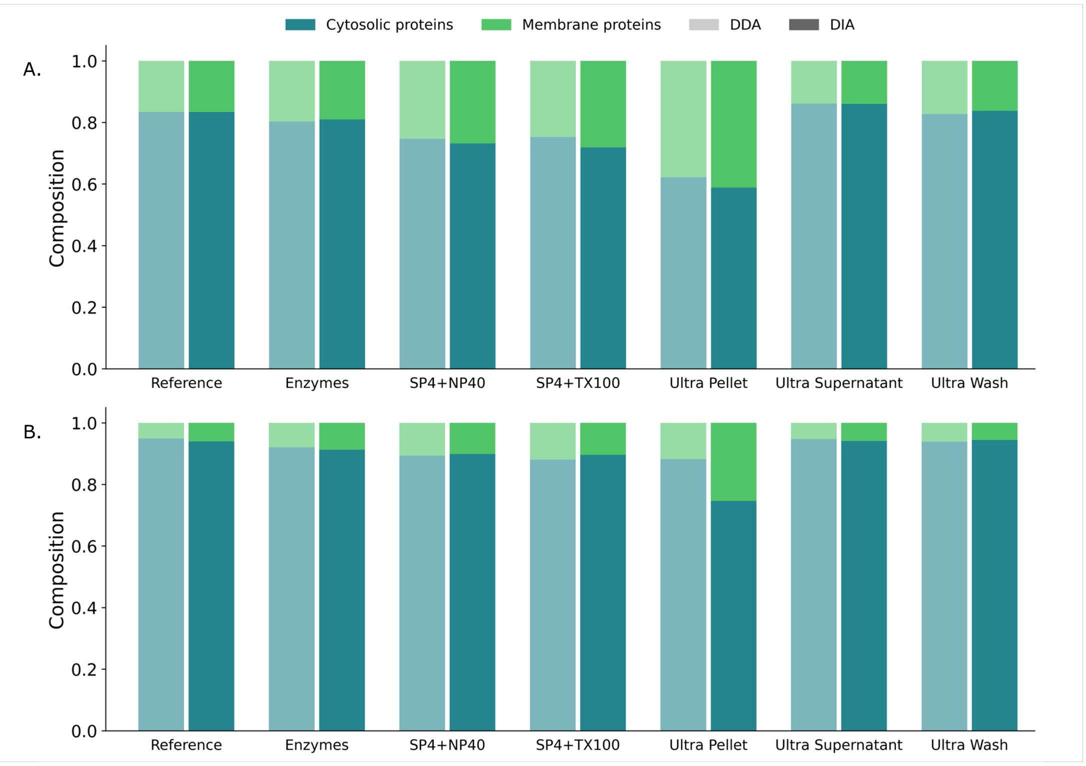


Figure 3.3: Proteome composition analysis for comparison of sample preparation protocols. (A) Proteome composition of *E. coli*. The membrane fraction occupies 27% of the total proteome composition for the detergent protocol. (B) Proteome composition of *P. putida*. The membrane fraction occupies 10% of the total proteome composition for the detergent protocol.

Further investigation of the *E. coli* quantified membrane proteome using proteomaps [22, 23] showed that the membrane proteome is mostly allocated to transport across the cell membrane, energy metabolism and cell structure compared with the overall proteome (Figure 3.5). As expected, the outer membrane porins, ATP synthase subunits and substrate importers represent the largest fraction of the quantified membrane proteome (Figure S3.9).

3.4. Discussions

In this study we developed a high-throughput sample preparation protocol for quantitative membrane proteomics of gram-negative bacteria, specifically *E. coli* and *P. putida*. Out of the investigated protocols, the detergent protocol proved optimal for increasing both the number of quantified membrane proteins and the size of the membrane fraction within the proteome, while maintaining comparable precision to the reference protocol. The addition of non-ionic detergents such as NP-40 and Triton X-100 is a proven method for enhanced solubilisation of membrane proteins, particularly for integral membrane proteins residing within the inner and outer membrane [6, 7].

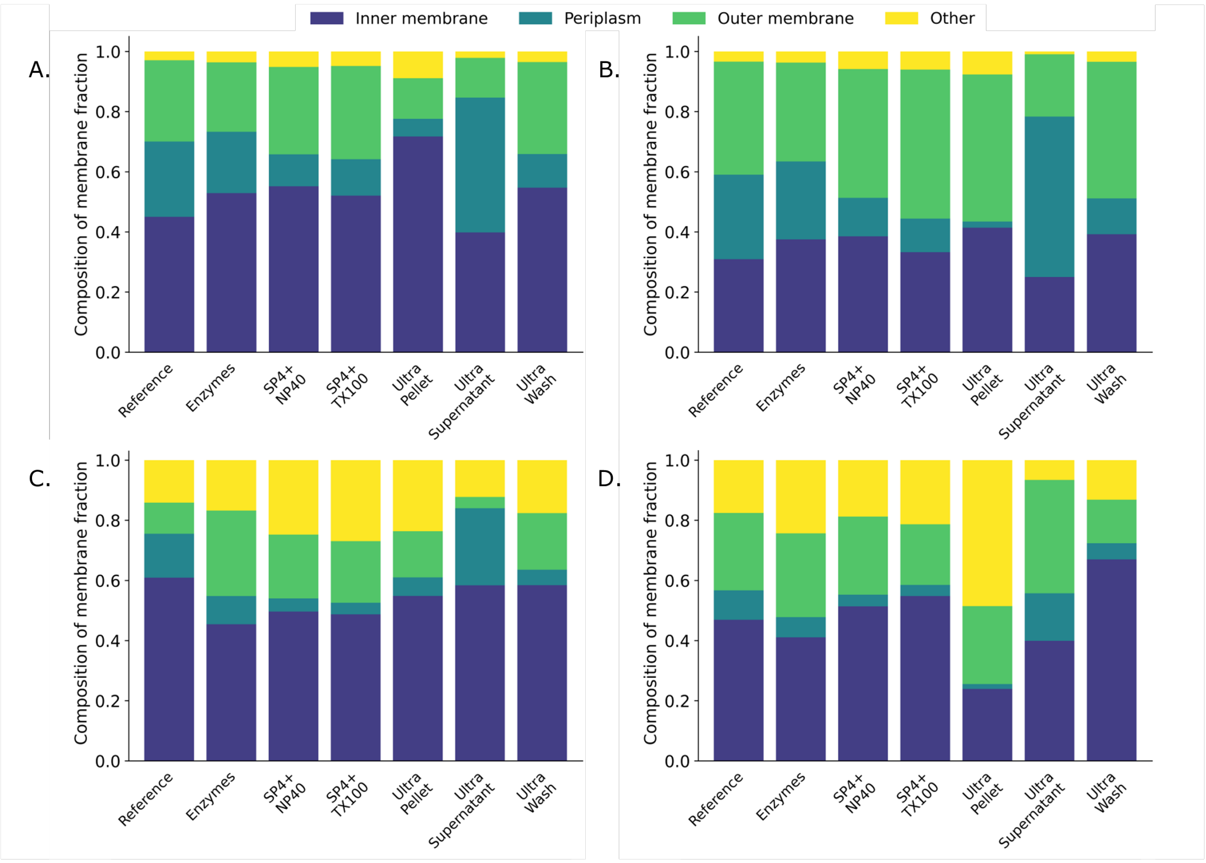


Figure 3.4: Membrane proteome composition analysis based on subcellular location for comparison of sample preparation protocols. (A) Membrane proteome composition of *E. coli* DDA data. (B) Membrane proteome composition of *E. coli* DIA data. (C) Membrane proteome composition of *P. putida* DDA data. (D) Membrane proteome composition of *P. putida* DIA data.

The enzymatic protocol did result in a larger number of quantified membrane proteins compared with the reference protocol, thus further degradation of cell debris seemingly exposed more membrane proteins to solubilisation. A combination of the enzymatic and detergent protocols could allow for increased solubilisation of a larger number of exposed membrane proteins, however verifying this would require additional investigation. It is unlikely that any one protocol would be able to capture the full membrane proteome due to the considerable differences between membrane proteins, e.g. subcellular location, number of transmembrane regions, and abundance levels [4, 24]. Nevertheless, the current optimal method, the detergent protocol combined with DIA analysis, resulted in the highest number of quantified membrane proteins for both *E. coli* (56% of theoretical membrane proteome) [5, 24, 25] and *P. putida* (61% of theoretical membrane proteome) [10]. The optimal method and the reference protocol quantified a similar number of cytosolic proteins, thus the optimal method could be appropriate for proteome-wide analysis with an increased representation of membrane proteins.

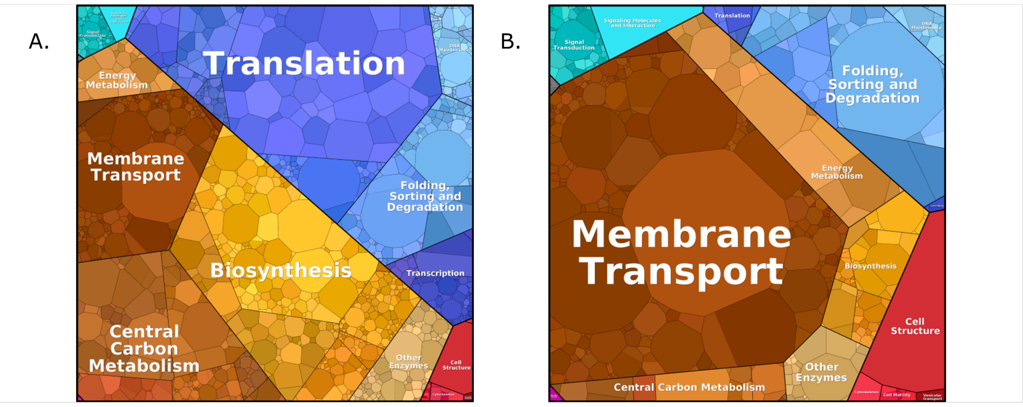


Figure 3.5: Proteomaps of the total proteome and the membrane proteome of *E. coli*. (A) Proteomap of the total proteome with level 2 annotation. Out of 2,646 proteins submitted, 1,800 were supported by the *E. coli* treemap. (B) Proteomap of the membrane proteome with level 2 annotation. Out of 794 submitted proteins, 533 were supported by the *E. coli* treemap.

Many ultracentrifugation protocols for bacterial membrane protein enrichment exist [2, 7, 10], yet no study has analysed all sample sections in order to investigate the sample fractionation efficiency and the extent of the enrichment. Membrane protein enrichment was achieved within the pellet sample using the current ultracentrifugation protocol, however the cytosolic fraction still occupied most of the proteome composition. The ultracentrifugation heavily selected for integral membrane proteins associated with the inner or outer membrane, thus the pellet sample was unrepresentative of the full membrane proteome. The precision of the ultracentrifugation protocol was similar to the detergent and reference protocol, despite expectations of introducing substantial variance [1, 2]. Most importantly, in comparison with the detergent protocol, the membrane protein enrichment of the ultracentrifugation protocol did not result in greater numbers of quantified membrane proteins in the pellet sample, which is the only sample section analysed in other studies [2, 10]. We concluded that, due to improved capacity and sensitivity of current HPLC-MS analyses, membrane protein enrichment through ultracentrifugation is no longer essential in order to identify and quantify a substantial fraction of the membrane proteome.

Membrane proteins encompass 20 to 30% of all encoded genes for various organisms [26], where *E. coli* and *P. putida* membrane proteins cover 33% and 14% of the corresponding proteomes, respectively. The lower theoretical fraction of membrane proteins in the *P. putida* proteome could be due to lacking annotation or could reflect biological differences between the two bacteria, since *P. putida* is known for its extensive repertoire of metabolic activities [27]. In terms of proteome composition, membrane proteins occupied 27% and 10% for *E. coli* and *P. putida* for the optimal method, respectively. Moreover, the concentration range of membrane and cytosolic proteins were comparable for both organisms, which contradicts the assumption of membrane proteins mostly having lower abundances than cytosolic proteins [1, 28]. The substantial membrane fraction of the total proteome composition has not been experimentally verified before, yet has been hypothesised by simulations with resource allocation models [29, 30].

Membrane proteins are assumed to crowd the membrane surface of prokaryotes [19] and quantitative measurements of the membrane proteome within the total proteome can provide new information on cell membrane occupation.

Absolute quantitative membrane proteomics data is sparse [1] and no absolute quantification of either the *E. coli* or the *P. putida* membrane proteome exists to date. Here we provided a semi-absolute quantification of both membrane proteomes through proteome composition values resulting from the TPA quantification method [20, 21]. To reach an absolute quantification in terms of intracellular protein concentrations, or cell surface densities for membrane proteins, additional calculations are necessary which depend on multiple parameters and assumptions. For cytosolic proteins, the inference of intracellular concentrations from proteome composition values requires the cellular protein density, i.e. grams of protein per litres of cell volume [13]. For membrane proteins, combining the cellular protein density in grams of protein per cell and the cell surface in µm^2 results in the cell surface densities [1]. The assumption of efficient extraction is applied for these calculations and is deemed fair for cytosolic proteins [13]. However, this assumption does not hold true for most membrane proteins and a correction factor is crucial, albeit challenging to determine [1, 2]. An enrichment factor should be calculated alongside whenever membrane protein enrichment is applied [2], such as with the ultracentrifugation protocol. The application of the detergent protocol provides a representative membrane proteome analysis without the need for enrichment and eliminates the need for enrichment factor calculation. Further research is vital for accurate absolute quantitative membrane proteomics in the future and we hope that the optimal method presented here paves the way.

3.5. Conclusions

Comparative analysis based on number of quantified membrane proteins and precision determined that the detergent protocol combined with DIA analysis is the optimal method for quantitative membrane proteomics in gram-negative bacteria. Additional detergents allowed for enhanced solubilisation of membrane proteins, resulting in higher numbers of quantified membrane proteins and increased membrane fractions of the proteome composition compared to the reference protocol, while maintaining similar precision. With a high number of both cytosolic and membrane proteins quantified, the optimal method allows for proteome-wide quantitative analysis with a stronger representation of the membrane proteome.

References

[1] C. Trötschel and A. Poetsch. “Current approaches and challenges in targeted absolute quantification of membrane proteins”. In: *Proteomics* 15.5-6 (2015), pp. 915–929. DOI: 10.1002/pmic.201400427.

[2] M. Antelo-Varela, J. Bartel, A. Quesada-Ganuza, K. Appel, M. Bernal-Cabas, T. Sura, A. Otto, M. Rasmussen, J. M. van Dijl, A. Nielsen, S. Maaß, and D. Becher. “Ariadne’s Thread in the Analytical Labyrinth of Membrane Proteins: Integration of Targeted and Shotgun Proteomics for Global Absolute Quantification of Membrane Proteins”. In: *Analytical Chemistry* 91.18 (2019), pp. 11972–11980. DOI: 10.1021/acs.analchem.9b02869.

[3] A. Poetsch and D. Wolters. “Bacterial membrane proteomics”. In: *Proteomics* 8.19 (2008), pp. 4100–4122. DOI: 10.1002/pmic.200800273.

[4] A. Bernsel and D. O. Daley. “Exploring the inner membrane proteome of Escherichia coli: which proteins are eluding detection and why?” In: *Trends in Microbiology* 17.10 (2009), pp. 444–449. DOI: 10.1016/j.tim.2009.07.005.

[5] M. Papanastasiou, G. Orfanoudaki, M. Koukaki, N. Kountourakis, M. F. Sardis, M. Aivaliotis, S. Karamanou, and A. Economou. “The Escherichia coli Peripheral Inner Membrane Proteome”. In: *Molecular and Cellular Proteomics* 12.3 (2013), pp. 599–610. DOI: 10.1074/mcp.M112.024711.

[6] K.C. Tsolis and A. Economou. “Chapter Two - Quantitative Proteomics of the E. coli Membranome”. In: *Proteomics in Biology, Part B*. Ed. by A. K. Shukla. Vol. 586. Methods in Enzymology. Academic Press, 2017, pp. 15–36. DOI: 10.1016/bs.mie.2016.09.026.

[7] S. M. Smith. “Strategies for the Purification of Membrane Proteins”. In: *Protein Chromatography: Methods and Protocols*. Ed. by D. Walls and S. T. Loughran. Springer New York, 2017, pp. 389–400. ISBN: 978-1-4939-6412-3. DOI: 10.1007/978-1-4939-6412-3.

[8] Paulina Perczyk and Marcin Broniatowski. “Simultaneous action of microbial phospholipase C and lipase on model bacterial membranes – Modeling the processes crucial for bioaugmentation”. In: *Biochimica et Biophysica Acta (BBA) - Biomembranes* 1863.7 (2021), p. 183620. DOI: https://doi.org/10.1016/j.bbamem.2021.183620.

[9] H. E. Johnston, K. Yadav, J. M. Kirkpatrick, G. S. Biggs, D. Oxley, H. B. Kramer, and R. S. Samant. “Solvent Precipitation SP3 (SP4) Enhances Recovery for Proteomics Sample Preparation without Magnetic Beads”. In: *Analytical Chemistry* 94.29 (2022), pp. 10320–10328. DOI: 10.1021/acs.analchem.1c04200.

[10] C. Roma-Rodrigues, P. M. Santos, D. Benndorf, E. Rapp, and I. Sá-Correia. “Response of Pseudomonas putida KT2440 to phenol at the level of membrane proteome”. In: *Journal of Proteomics* 73.8 (2010), pp. 1461–1478. DOI: 10.1016/j.jprot.2010.02.003.

[11] L. Koontz. “Chapter One - TCA Precipitation”. In: *Laboratory Methods in Enzymology: Protein Part C*. Ed. by J. Lorsch. Vol. 541. Methods in Enzymology. Academic Press, 2014, pp. 3–10. DOI: 10.1016/B978-0-12-420119-4.00001-X.

[12] Y Xuan, N. W. Bateman, S. Gallien, S. Goetze, Y. Zhou, P. Navarro, M. Hu, N. Parikh, B. L. Hood, K. A. Conrads, C. Loosse, R. B. Kitata, S. R. Piersma, D. Chiasserini, H. Zhu, G. Hou, M. Tahir, A. Macklin, A. Khoo, X. Sun, B. Crossett, A. Sickmann, Y. Chen, C. R. Jimenez, H. Zhou, S. Liu, M. R. Larsen, T. Kislinger, Z. Chen, B. L. Parker, S. J. Cordwell, B. Wollscheid, and T. P. Conrads. "Standardization and harmonization of distributed multi-center proteotype analysis supporting precision medicine studies". In: *Nature Communications* 11.1 (2020), p. 5248. DOI: 10.1038/s41467-020-18904-9.

[13] M. Mori, Z. Zhang, A. Banaei-Esfahani, J. B. Lalanne, H. Okano, B. C. Collins, A. Schmidt, O. T. Schubert, D. S. Lee, G. W. Li, R. Aebersold, T. Hwa, and C. Ludwig. "From coarse to fine: The absolute *Escherichia coli* proteome under diverse growth conditions". In: *Molecular Systems Biology* 17.5 (2021), e9536. DOI: 10.15252/msb.20209536.

[14] V. Demichev, C. B. Messner, S. I. Vernardis, K. S. Lilley, and M. Ralser. "DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput". In: *Nature Methods* 17 (2020), pp. 41–44. DOI: 10.1038/s41592-019-0638-x.

[15] A. V. Sastry, Y. Gao, R. Szubin, Y. Hefner, S. Xu, D. Kim, K. S. Choudhary, L. Yang, Z. A. King, and B. O. Palsson. "The Escherichia coli transcriptome mostly consists of independently regulated modules". In: *Nature Communications* 10.1 (2019), p. 5536. DOI: 10.1038/s41467-019-13483-w.

[16] G. Serrano, E. Guruceaga, and V. Segura. "DeepMSPeptide: peptide detectability prediction using deep learning". In: *Bioinformatics* 36.4 (2019), pp. 1279–1280. DOI: 10.1093/bioinformatics/btz708.

[17] Y. Perez-Riverol, J. Bai, C. Bandla, D. García-Seisdedos, S. Hewapathirana, S. Kamatchinathan, D. J. Kundu, A. Prakash, A. Frericks-Zipper, M. Eisenacher, M. Walzer, S. Wang, A. Brazma, and J. A. Vizcaíno. "The PRIDE database resources in 2022: a hub for mass spectrometry-based proteomics evidences". In: *Nucleic Acids Research* 50.D1 (2022), pp. D543–D552. DOI: 10.1093/nar/gkab1038.

[18] O. T. Schubert, H. L. Röst, B. C. Collins, G. Rosenberger, and R. Aebersold. "Quantitative proteomics: challenges and opportunities in basic and applied research". In: *Nature Protocols* 12.7 (2017), pp. 1289–1294. DOI: 10.1038/nprot.2017.040.

[19] R. Phillips and R. Milo. "A feeling for the numbers in biology". In: *PNAS* 106.51 (2009), pp. 21465–21471. DOI: 10.1073/pnas.0907732106.

[20] J. R. Wiśniewski, P. Ostasiewicz, K. Duś, D. F. Zielińska, F. Gnad, and M. Mann. "Extensive quantitative remodeling of the proteome between normal colon tissue and adenocarcinoma". In: *Molecular Systems Biology* 8.1 (2012), p. 611. DOI: 10.1038/msb.2012.44.

[21] J. R. Wiśniewski, M. Y. Hein, J. Cox, and M. Mann. "A "Proteomic Ruler" for Protein Copy Number and Concentration Estimation without Spike-in Standards".

In: *Molecular and Cellular Proteomics* 13.12 (2014), pp. 3497–3506. DOI: 10.1074/mcp.M113.037309.

[22] A. Otto, J. Bernhardt, H. Meyer, M. Schaffer, F. A. Herbst, J. Siebourg, U. Mäder, M. Lalk, M. Hecker, and D. Becher. “Systems-wide temporal proteomic profiling in glucose-starved Bacillus subtilis”. In: *Nature Communications* 1.1 (2010), p. 137. DOI: 10.1038/ncomms1137.

[23] W. Liebermeister, E. Noor, A. Flamholz, D. Davidi, J. Bernhardt, and R. Milo. “Visual account of protein investment in cellular functions”. In: *PNAS* 111.23 (2014), pp. 8488–8493. DOI: 10.1073/pnas.1314810111.

[24] J. H. Weiner and L. Li. “Proteome of the Escherichia coli envelope and technological challenges in membrane proteome analysis”. In: *Biochimica et Biophysica Acta (BBA) - Biomembranes* 1778.9 (2008), pp. 1698–1713. DOI: 10.1016/j.bbamem.2007.07.020.

[25] A. Sueki, F. Stein, M. M. Savitski, J. Selkrig, and A. Typas. “Systematic Localization of Escherichia coli Membrane Proteins”. In: *mSystems* 5.2 (2020), e00808–19. DOI: 10.1128/msystems.00808-19.

[26] E. Wallin and G. Von Heijne. “Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms”. In: *Protein Science* 7.4 (1998), pp. 1029–1038. DOI: 10.1002/pro.5560070420.

[27] E. Belda, R. G. A. van Heck, M. J. Lopez-Sanchez, S. Cruveiller, V. Barbe, C. Fraser, H. P. Klenk, J. Petersen, A. Morgat, P. I. Nikel, D. Vallenet, Z. Rouy, A. Sekowska, V. A. P. Martins dos Santos, V. de Lorenzo, A. Danchin, and C. Médigue. “The revisited genome of Pseudomonas putida KT2440 enlightens its value as a robust metabolic chassis”. In: *Environmental Microbiology* 18.10 (2016), pp. 3403–3424. DOI: 10.1111/1462-2920.13230.

[28] J. R. Wiśniewski. “Dilemmas With Absolute Quantification of Pharmacologically Relevant Proteins Using Mass Spectrometry”. In: *Journal of Pharmaceutical Sciences* 110.1 (2021), pp. 17–21. DOI: 10.1016/j.xphs.2020.10.034.

[29] K. Zhuang, G. N. Vemuri, and R. Mahadevan. “Economics of membrane occupancy and respiro-fermentation”. In: *Molecular Systems Biology* 7.1 (2011), p. 500. DOI: 10.1038/msb.2011.34.

[30] A. Regueira, J. M. Lema, and M. Mauricio-Iglesias. “Microbial inefficient substrate use through the perspective of resource allocation models”. In: *Current Opinion in Biotechnology* 67.1 (2021), pp. 130–140. DOI: 10.1016/j.copbio.2021.01.015.

Supplementary materials

Extended experimental methods

Proteomics sample preparation - Reference protocol

Cell lysis

Preparation: Set Thermo mixer to 99°C. Make lysis buffer according to number of samples. After making lysis buffer, transfer the buffer to Eppendorf tubes and place in Thermo mixer at 99°C. Take samples from -80°C freezer and put on ice.

1. Add 2 (3-mm zirconium oxide) beads to the cell pellet (while on ice).
2. Add 50 µL of hot lysis buffer (99°C) to cell pellet.
3. 5 minutes in Tissuelyser at 25 Hz.
4. Boil in heat block (99°C) for 10 minutes while shaking/mixing at 1,800 rpm.
5. Centrifuge at 15,000 g for 10 minutes.
6. Transfer 50 µL of supernatant to a new Eppendorf tube and add same volume of 50 mM Ambic (ammonium bicarbonate).
7. Vortex mix the samples.

Perform total protein measurement using BCA assay with micro BCA Protein Assay Kit (Thermo Scientific, product number 23235).

Protein digestion with trypsin and Lys-C mixture

1. Take 20 µg of protein per sample and dilute in 50 mM Ambic until a volume of 100 µL.
2. Add 20 µL of 0.1 µg/µL of trypsin and Lys-C mixture to the samples in Eppendorf tubes.
3. Incubate in Thermomixer (8 hours at 37°C at 600 rpm, followed by xxx hours at 4°C at 600 rpm).

Stagetipping

1. Make 10% TFA from 100% TFA: Add 900 µL Milli Q water and 100 µL 100% TFA to 1.5 mL Eppendorf tube.
2. Stop reaction by adding 10 µL 10% TFA.
3. Vortex mix the samples.
4. Centrifuge the samples at 14,000 g for 15 minutes.

To the following steps: Never let them dry longer than 2 minutes! (Max capacity around 20 µg, loading more you will elute more hydrophobic peptides).

Activate the C18 filters with:

- 1x 20 µL MeOH (1 minute at 1,000g at 20°C).
- 1x 20 µL buffer B (80% CH3CN, 0.1% FA) (1 minute at 1,000 g at 20°C).
- 2x 20 µL buffer A´ (3% CH3CN, 1% TFA) (1 minute at 1,000 g at 20°C).
- Spin samples through C18 column (1,500 g, 1.5 minutes at 20°C, depending on volume).
- Wash 2x with 20 µL buffer A (0.1% FA) (1 minute at 1,000 g at 20°C). Do not let C18 dry!!

Change plate to elution plate.

- Elute with 2x 20 µL buffer B (1 minute at 1,000g).
- Reduce ACN concentration to 5 µL by running on Eppendorf concentrator (Program V-AQ at room temperature). If in 96 well plate for 1.5 hours.
- Add 40 µL 0.1% FA in water and keep at 4°C.

Proteomics sample preparation - Enzymatic protocol

Cell lysis
Preparation: Set Thermo mixer to 99°C. Make lysis buffer according to number of samples. After making lysis buffer, transfer the buffer to Eppendorf tubes and place in Thermo mixer at 99°C. Take samples from -80°C freezer and put on ice.

1. Add 2 (3-mm zirconium oxide) beads to the cell pellet (while on ice).
2. Add 50 µL of hot lysis buffer (99°C) to cell pellet.
3. 5 minutes in Tissuelyser at 25 Hz.
4. Boil in heat block (99°C) for 10 minutes while shaking/mixing at 1,800 rpm.
5. Spin down the samples.
6. Transfer 50 µL of supernatant to a new Eppendorf tube and add same volume of 50 mM Ambic (ammonium bicarbonate).
7. Vortex mix the samples.

Perform total protein measurement using BCA assay with micro BCA Protein Assay Kit (Thermo Scientific, product number 23235).

Enzymatic digest with lipase, phospholipase C, and lysozyme

1. Take 20 µg of protein per sample and dilute in 50 mM Ambic until a volume of 100 µL.
2. Add 10 µL of each enzyme solution to the samples.
3. Incubate for 1 hour at 37°C in the Thermomixer.

Protein digestion with trypsin and Lys-C mixture

1. Add 20 µL of 0.1 µg/µL of trypsin and Lys-C mixture to the samples in Eppendorf tubes.
2. Incubate in Thermomixer (8 hours at 37°C at 600 rpm, followed by xxx hours at 4°C at 600 rpm).

Stagetipping

1. Make 10% TFA from 100% TFA: Add 900 µL Milli Q water and 100 µL 100% TFA to 1.5 mL Eppendorf tube.
2. Stop reaction by adding 10 µL 10% TFA.
3. Vortex mix the samples.
4. Centrifuge the samples at 14,000 g for 15 minutes.

To the following steps: Never let them dry longer than 2 minutes! (Max capacity around 20 µg, loading more you will elute more hydrophobic peptides).

Activate the C18 filters with:

- 1x 20 µL MeOH (1 minute at 1,000g at 20°C).
- 1x 20 µL buffer B (80% CH3CN, 0.1% FA) (1 minute at 1,000 g at 20°C).
- 2x 20 µL buffer A´ (3% CH3CN, 1% TFA) (1 minute at 1,000 g at 20°C).
- Spin samples through C18 column (1,500 g, 1.5 minutes at 20°C, depending on volume).
- Wash 2x with 20 µL buffer A (0.1% FA) (1 minute at 1,000 g at 20°C). Do not let C18 dry!!

Change plate to elution plate.

- Elute with 2x 20 µL buffer B (1 minute at 1,000g).
- Reduce ACN concentration to 5 µL by running on Eppendorf concentrator (Program V-AQ at room temperature). If in 96 well plate for 1.5 hours.
- Add 40 µL 0.1% FA in water and keep at 4°C.

Proteomics sample preparation - Detergent protocol

Cell lysis
Preparation: Set Thermo mixer to 99°C. Make lysis buffer according to number of samples. After making lysis buffer, transfer the buffer to Eppendorf tubes and place in Thermo mixer at 99°C. Take samples from -80°C freezer and put on ice.

1. Add 2 (3-mm zirconium oxide) beads to the cell pellet (while on ice).
2. Add 50 µL of hot lysis buffer (99°C) to cell pellet.
3. 5 minutes in Tissuelyser at 25 Hz.
4. Boil in heat block (99°C) for 10 minutes while shaking/mixing at 1,800 rpm.
5. Centrifuge at 15,000 g for 10 minutes.
6. Transfer 50 µL of supernatant to a new Eppendorf tube and add same volume of 50 mM Ambic (ammonium bicarbonate).
7. Vortex mix the samples.

Perform total protein measurement using BCA assay with micro BCA Protein Assay Kit (Thermo Scientific, product number 23235).

Glass beads solution preparation

- 9–13 µm glass spheres/beads. Glass beads broadly improved recovery, digestion efficiency and reproducibility, but are not required.
- Suspend 100 mg in 1 mL of Ultrapure water, vortex until suspended fully, and pellet at > 500g for 1 minute. Of note: approximately 50% of the beads are buoyant, and will not pellet, and should be removed over the course of these wash steps. Additionally, small amounts of metal in the beads can be removed by magnet or acid wash but had no effect on the performance of the beads. Larger scale preps are possible but may require additional washes due to buoyant beads.
- Resuspend, vortex and wash with > 1 mL of: 100% acetonitrile (ACN) (1×), 50 mM AmBic (1×), and Ultrapure water (> 2×) ensuring no unpelleted beads remain.
- Resuspend beads in 900 µL acetonitrile to 50 mg/mL (recommended). Avoids protein dilution from beads in water. Ensures uniform bead dispersion.
- Dilute beads to 2.5 µg/µL by adding 50 µL of the 50 mg/mL beads solution to 950 µL acetonitrile. Enough for 10 samples. Dilute beads to at least 2.5× [protein] (so bead:protein is 10:1 from 4 volumes of bead–ACN suspension.

SP4 clean-up protocol
The recommendations are the following:

- Liquids should be kept low in the tube, with losses/contamination possible from tube walls/lid.

- Pipette ACN directly into the sample to ensure rapid mixing, but do not touch the ACN–sample mix with the tip.
- Use the tube hinge to orientate the location of the pellet (fixed angle rotors). Initially orientate the tube hinge inwards during the pellet precipitation and turn 180°after 2.5 min will give a denser pellet and less risk of loss from fragile wall adhesion.
- During wash removals, avoid touching the tube walls with the tip as precipitation may occur on them, pipette slowly and avoid agitating the pellet.
- If adding beads, ensure they maintain a uniform suspension in water/ACN by pipetting up and down at least once between additions.

1. Aliquot 20 µg of protein into an Eppendorf tube and dilute in 50 mM Ambic until a volume of 20 µL.
2. Add 80 µL of 2.5 µg/µL ACN:bead suspension and ensure complete mixing. Without pipette mixing, e.g., by consistent ACN addition.
3. Centrifuge for 5 minutes at 15,000 g.
4. Remove supernatant by pipetting slowly and remove a consistent volume of 90–95%. Avoid disturbing beads/pellet.
5. Wash with 80% ethanol, with a volume of 180 µL. Pipette gently down the side opposite the hinge/pellet to avoid disturbance, do not vortex/resuspend.
6. Centrifuge for 2 minutes at 15,000 g and remove 90–95% of wash.
7. Repeat wash steps 5 and 6 twice, for a total of 3 washes.
8. Remove >= 95% of final wash. For larger volumes a final 2 minute spin will help with removal of excess wash.
9. Add 100 µL of 50 mM AmBic to bead pellets of SP4 clean-up protocol samples. No mixing!

Protein digestion with trypsin and Lys-C mixture

1. Add 20 µL of 0.1 µg/µL of trypsin and Lys-C mixture to the samples in Eppendorf tubes.
2. Incubate in Thermomixer (Use program 8 in Thermomixer; 8 hours at 37°C at 1,000 rpm, followed by xxx hours at 4°C at 1,000 rpm).

Stagetipping

1. Make 10% TFA from 100% TFA: Add 900 µL Milli Q water and 100 µL 100% TFA to 1.5 mL Eppendorf tube.
2. Stop reaction by adding 10 µL 10% TFA.
3. Vortex mix the samples.

4. Centrifuge the samples at 14,000 g for 15 minutes.

To the following steps: Never let them dry longer than 2 minutes! (Max capacity around 20 µg, loading more you will elute more hydrophobic peptides).

Activate the C18 filters with:

- 1x 20 µL MeOH (1 minute at 1,000g at 20°C).
- 1x 20 µL buffer B (80% CH3CN, 0.1% FA) (1 minute at 1,000 g at 20°C).
- 2x 20 µL buffer A´ (3% CH3CN, 1% TFA) (1 minute at 1,000 g at 20°C).
- Spin samples through C18 column (1,500 g, 1.5 minutes at 20°C, depending on volume).
- Wash 2x with 20 µL buffer A (0.1% FA) (1 minute at 1,000 g at 20°C). Do not let C18 dry!!

Change plate to elution plate.

- Elute with 2x 20 µL buffer B (1 minute at 1,000g).
- Reduce ACN concentration to 5 µL by running on Eppendorf concentrator (Program V-AQ at room temperature). If in 96 well plate for 1.5 hours.
- Add 40 µL 0.1% FA in water and keep at 4°C.

Proteomics sample preparation - Ultracentrifugation protocol

Cell lysis
Preparation: Set Thermo mixer to 99°C. Make lysis buffer according to number of samples. After making lysis buffer, transfer the buffer to Eppendorf tubes and place in Thermo mixer at 99°C. Take samples from -80°C freezer and put on ice.

1. Add 2 (3-mm zirconium oxide) beads to the cell pellet (while on ice).
2. Add 100 µL of hot lysis buffer (99°C) to cell pellet.
3. 5 minutes in Tissuelyser at 25 Hz.
4. Boil in heat block (99°C) for 10 minutes while shaking/mixing at 1,800 rpm.
5. Spin down the samples.
6. Transfer 100 µL of supernatant to a new Eppendorf tube and add same volume of 50 mM Ambic (ammonium bicarbonate).
7. Vortex mix the samples.

Perform total protein measurement using BCA assay with micro BCA Protein Assay Kit (Thermo Scientific, product number 23235).

Membrane protein extraction and fractionation

1. Set the ultracentrifuge to 4°C.
2. Transfer all of the whole cell extract as starting material to the ultracentrifuge tube.
3. Adjust the volume up to 1.5 mL with Tris EDTA buffer (10 mM EDTA, 20 mM Tris-HCl, pH 7.5).
4. Ultracentrifuge at 100,000 g and 4°C for 90 minutes to fractionate the proteins.
5. Transfer the supernatant containing the soluble, cytosolic proteins to new Eppendorf tubes.
6. Wash step: Resuspend the pellet in 200 µL of lysis buffer.
7. Adjust the volume up to 1.5 mL with Tris EDTA buffer (10 mM EDTA, 20 mM Tris-HCl, pH 7.5).
8. Transfer to an Eppendorf tube.
9. Incubate the solution at 4°C for 30 minutes on a bench rotator to resuspend.
10. Transfer to the ultracentrifuge tube.
11. Ultracentrifuge at 100,000 g and 4°C for 60 minutes to wash the membrane proteins.
12. Transfer the supernatant containing the soluble proteins to new Eppendorf tubes.

13. Resuspend the pellet in 200 µL lysis buffer, vortex and transfer to new Eppendorf tubes.
14. Incubate the solution at 4°C for 30 minutes on a bench rotator to solubilise the membrane proteins.
15. Add 200 µL of 50 mM Ambic only to the pellet fraction samples.

Precipitation of membrane proteins in both supernatants

1. Thaw samples and keep on ice.
2. Prepare 100% trichloroacetic acid (TCA).
3. Add 225 µL 100% TCA to each tube to reach a final TCA concentration of 15%. The sample should turn milky white if proteins are present.
4. Turn on the centrifuge on 4°C.
5. Vortex well and incubate on ice for 1 hour.
6. Centrifuge the samples at 4°C at 14,000 g for 60 minutes.
7. Discard the supernatant properly. (First with 1000 µL pipette then with 10 µL pipette).
8. Add 800 µL (use more if necessary) of -20°C acetone to the pellet.
9. Vortex and then sonicate (3x 10 seconds) the samples. The pellet will not completely be resuspended. The proteins will not dissolve in the acetone, they will only be suspended.
10. Incubate overnight at -20°C.
11. Turn the centrifuge on 4°C.
12. Turn the Thermomixer on 60°C.
13. Centrifuge at 14,000 g at 4°C for 30 minutes.
14. Discard the supernatant.
15. Leave to air dry for around 30 minutes.
16. Prepare Guanidinium HCl lysis buffer.
17. Add 100 µL Guanidinium HCl cell lysis buffer and resuspend by pipetting up and down three times.
18. Place in Thermomixer 600 rpm at 60°C for 30 minutes (2x 15 minutes and vortex in between).
19. (If pellet is still visible: add 50 µL Guanidinium HCl cell lysis buffer, vortex and place in Thermomixer again for 30 minutes.)

20. Add xxx µL (100 µL; same amount as total of Guanidinium HCl cell lysis buffer) of 50 mM Ambic added to each sample. Whirl thoroughly!

Perform total protein measurement using BCA assay with micro BCA Protein Assay Kit (Thermo Scientific, product number 23235).

Protein digestion with trypsin and Lys-C mixture

1. Take 20 µg of protein per sample and dilute in 50 mM Ambic until a volume of 100 µL.
2. Add 20 µL of 0.1 µg/µL of trypsin and Lys-C mixture to the samples in Eppendorf tubes.
3. Incubate in Thermomixer (8 hours at 37°C at 600 rpm, followed by xxx hours at 4°C at 600 rpm).

Stagetipping

1. Make 10% TFA from 100% TFA: Add 900 µL Milli Q water and 100 µL 100% TFA to 1.5 mL Eppendorf tube.
2. Stop reaction by adding 10 µL 10% TFA.
3. Vortex mix the samples.
4. Centrifuge the samples at 14,000 g for 15 minutes.

To the following steps: Never let them dry longer than 2 minutes! (Max capacity around 20 µg, loading more you will elute more hydrophobic peptides).

Activate the C18 filters with:

- 1x 20 µL MeOH (1 minute at 1,000g at 20°C).
- 1x 20 µL buffer B (80% CH3CN, 0.1% FA) (1 minute at 1,000 g at 20°C).
- 2x 20 µL buffer A´ (3% CH3CN, 1% TFA) (1 minute at 1,000 g at 20°C).
- Spin samples through C18 column (1,500 g, 1.5 minutes at 20°C, depending on volume).
- Wash 2x with 20 µL buffer A (0.1% FA) (1 minute at 1,000 g at 20°C). Do not let C18 dry!!

Change plate to elution plate.

- Elute with 2x 20 µL buffer B (1 minute at 1,000g).
- Reduce ACN concentration to 5 µL by running on Eppendorf concentrator (Program V-AQ at room temperature). If in 96 well plate for 1.5 hours.
- Add 40 µL 0.1% FA in water and keep at 4°C.

Supplementary figures

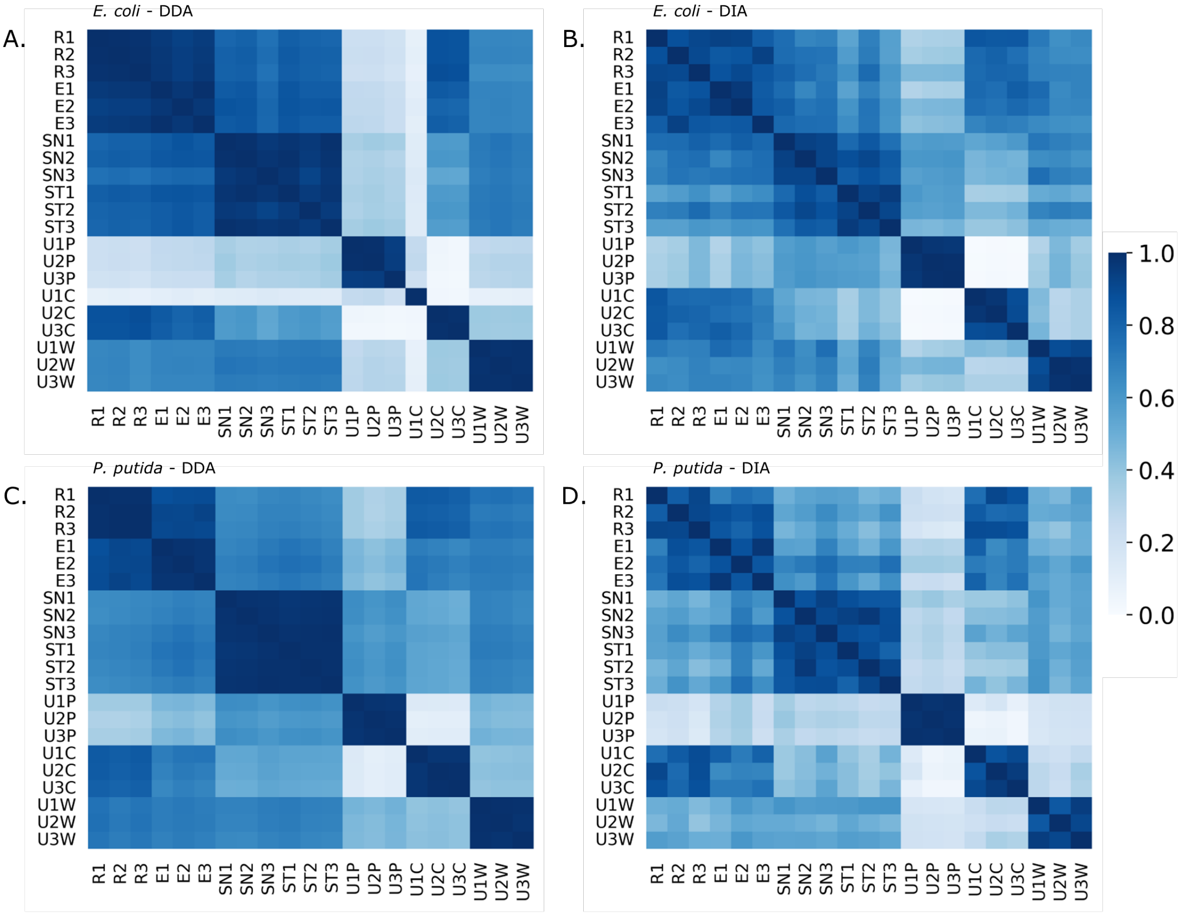


Figure S3.1: Pearson correlation coefficients of proteome composition values in all samples compared against each other. (A) Pearson correlation coefficients of *E. coli* data obtained with DDA analysis. Based on the substantially lower total MS signal of replicate 1 of the DDA analysis after the ultracentrifugation protocol compared to the other replicates, replicate 1 was excluded from further analyses. (B) Pearson correlation coefficients of *E. coli* data obtained with DIA analysis. (C) Pearson correlation coefficients of *P. putida* data obtained with DDA analysis. (D) Pearson correlation coefficients of *P. putida* data obtained with DIA analysis.

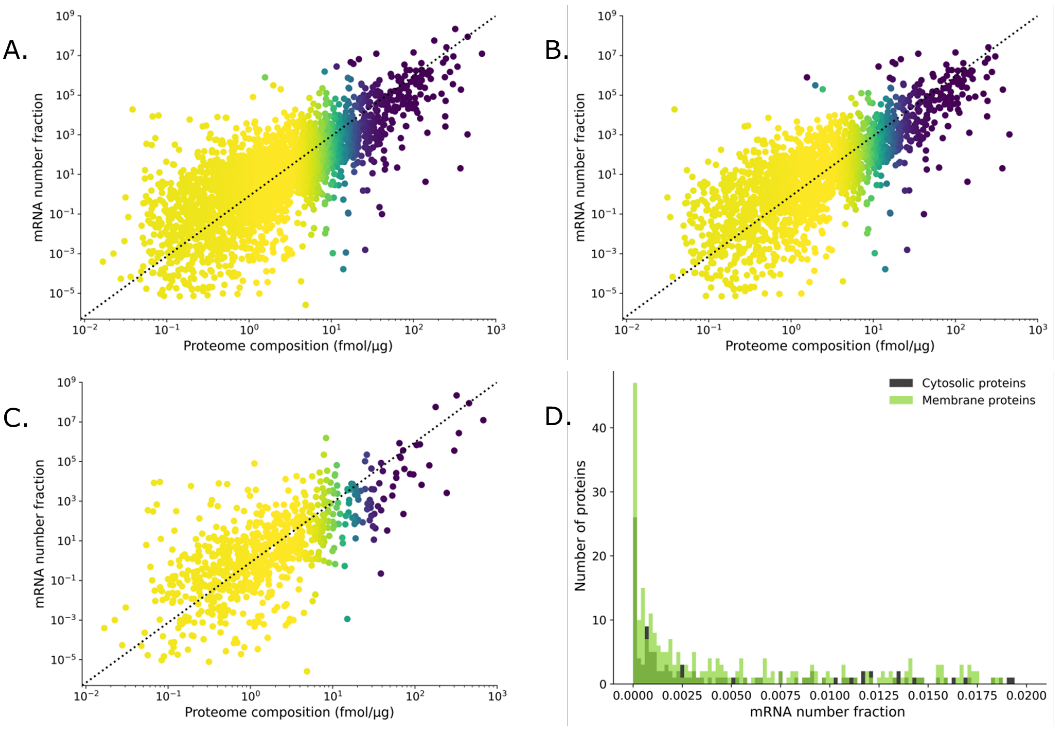


Figure S3.2: Correlation between proteome composition values produced with the detergent protocol and DIA analysis, and mRNA number fraction values (number of mRNA molecules per total number of mRNA molecules) extracted from the *E. coli* reference condition of the PRECISE database [1]. The density is indicated by the colour map, where light (yellow) is more dense than dark (blue). (A) Correlation of all proteins. (B) Correlation of cytosolic proteins. (C) Correlation of membrane proteins. (D) Distribution of mRNA number fraction values of expressed proteins which did not show up in the protein quantification.

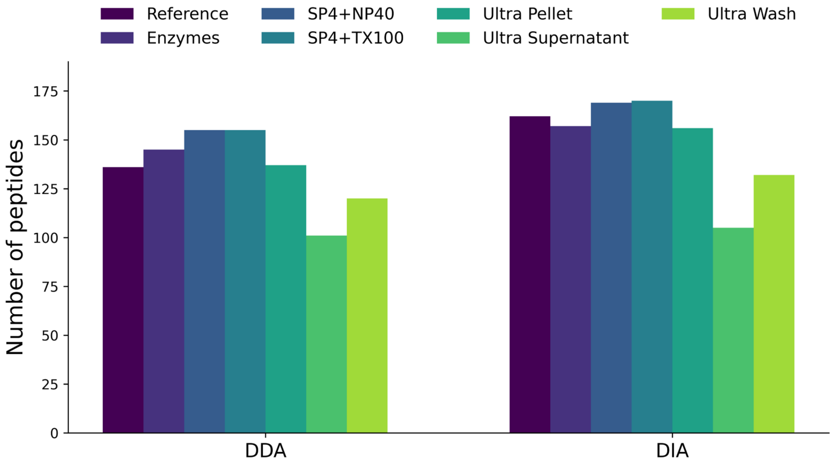


Figure S3.3: Comparison of sample preparation protocols and acquisition modes in regards to the *E. coli* membrane peptide mix consisting of 228 peptides. The highest number of peptides (170) was identified by deploying the detergent protocol in combination with DIA analysis.

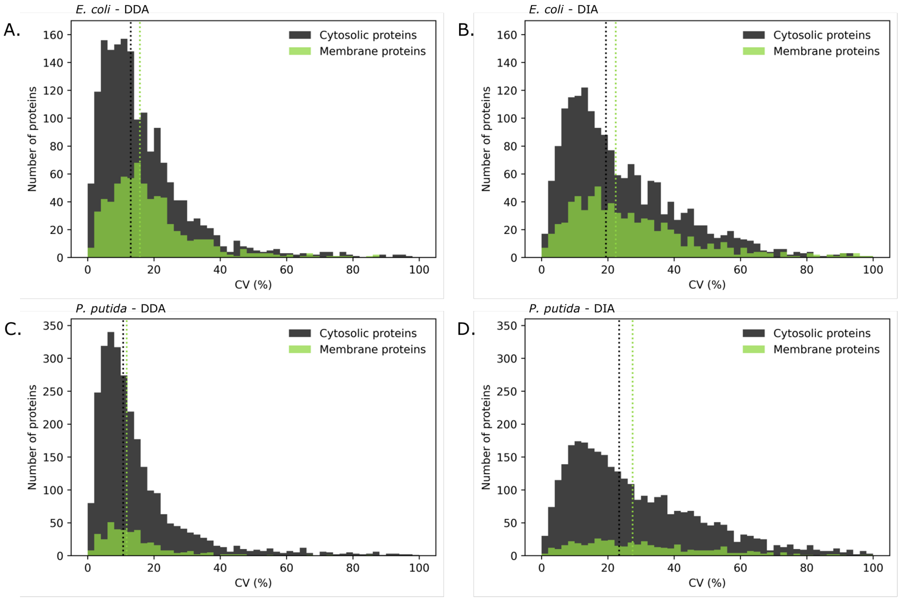


Figure S3.4: Distribution of coefficient of variation (CV) values from the samples of the detergent protocol (NP-40), with membrane proteins in comparison to cytosolic proteins. (A) CV distribution of *E. coli* data obtained with DDA analysis. (B) CV distribution of *E. coli* data obtained with DIA analysis. (C) CV distribution of *P. putida* data obtained with DDA analysis. (D) CV distribution of *P. putida* data obtained with DIA analysis.

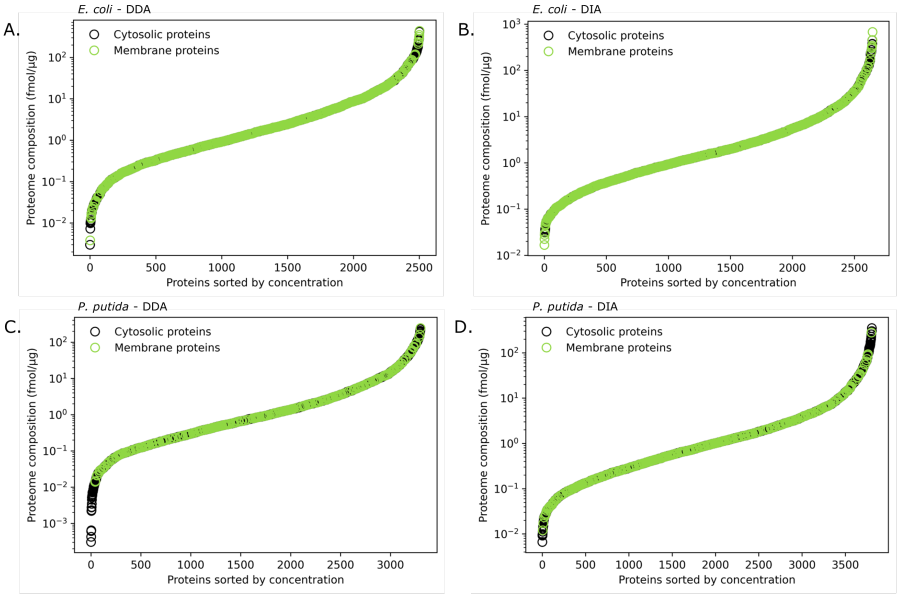


Figure S3.5: Concentration range from the samples of the detergent protocol (NP-40), with membrane proteins in comparison to cytosolic proteins. (A) Concentration range of *E. coli* data obtained with DDA analysis. (B) Concentration range of *E. coli* data obtained with DIA analysis. (C) Concentration range of *P. putida* data obtained with DDA analysis. (D) Concentration range of *P. putida* data obtained with DIA analysis.

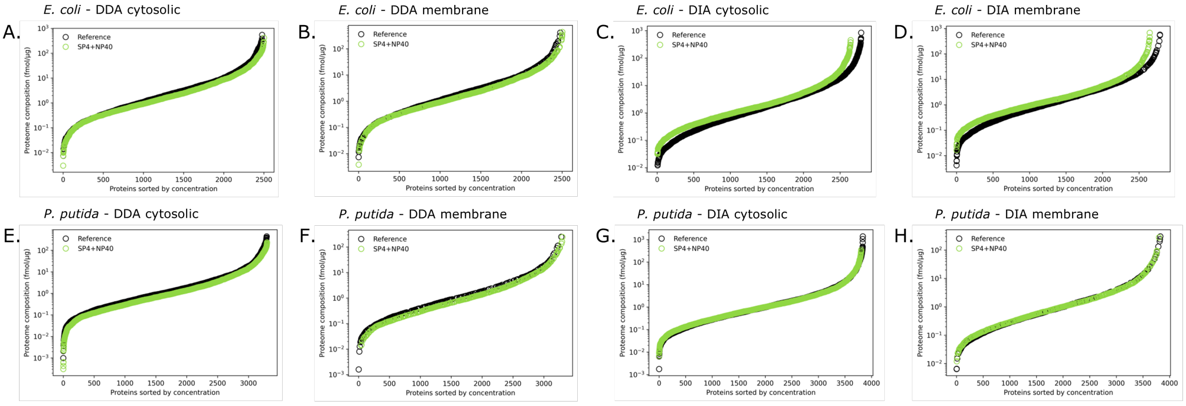


Figure S3.6: Concentration range from the samples of the detergent protocol (NP-40) in comparison to the reference protocol. (A) Concentration range of *E. coli* cytosolic protein data obtained with DDA analysis. (B) Concentration range of *E. coli* membrane protein data obtained with DDA analysis. (C) Concentration range of *E. coli* cytosolic protein data obtained with DIA analysis. (D) Concentration range of *E. coli* membrane protein data obtained with DIA analysis. (E) Concentration range of *P. putida* cytosolic protein data obtained with DDA analysis. (F) Concentration range of *P. putida* membrane protein data obtained with DDA analysis. (G) Concentration range of *P. putida* cytosolic protein data obtained with DIA analysis. (H) Concentration range of *P. putida* membrane protein data obtained with DIA analysis.

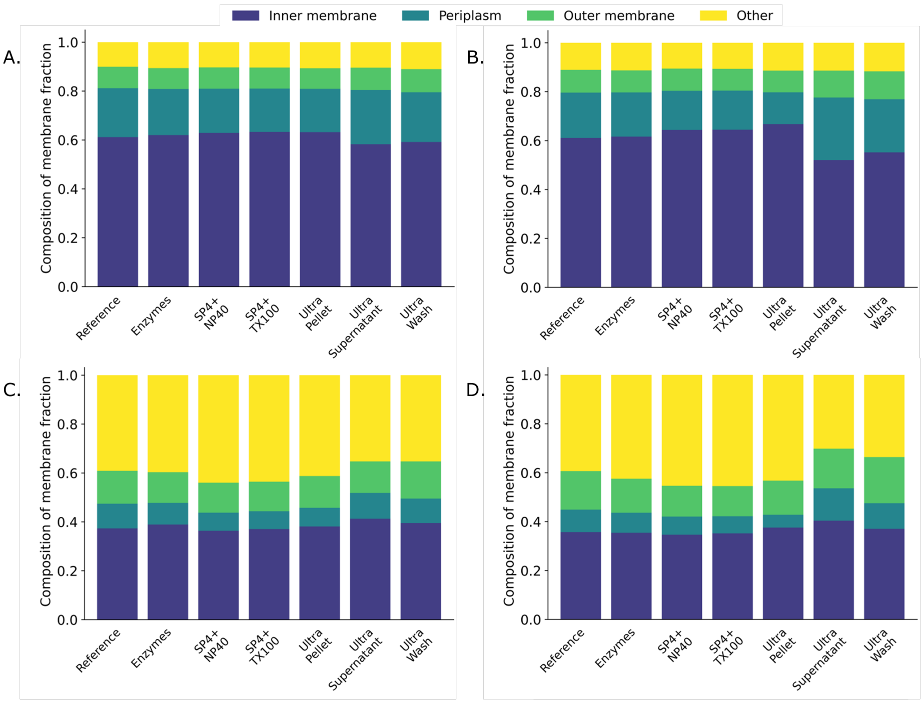


Figure S3.7: Membrane proteome fraction analysis based on subcellular location for comparison of sample preparation protocols. (A) Membrane proteome fraction analysis of *E. coli* DDA data. (B) Membrane proteome fraction analysis of *E. coli* DIA data. (C) Membrane proteome fraction analysis of *P. putida* DDA data. (D) Membrane proteome fraction analysis of *P. putida* DIA data.

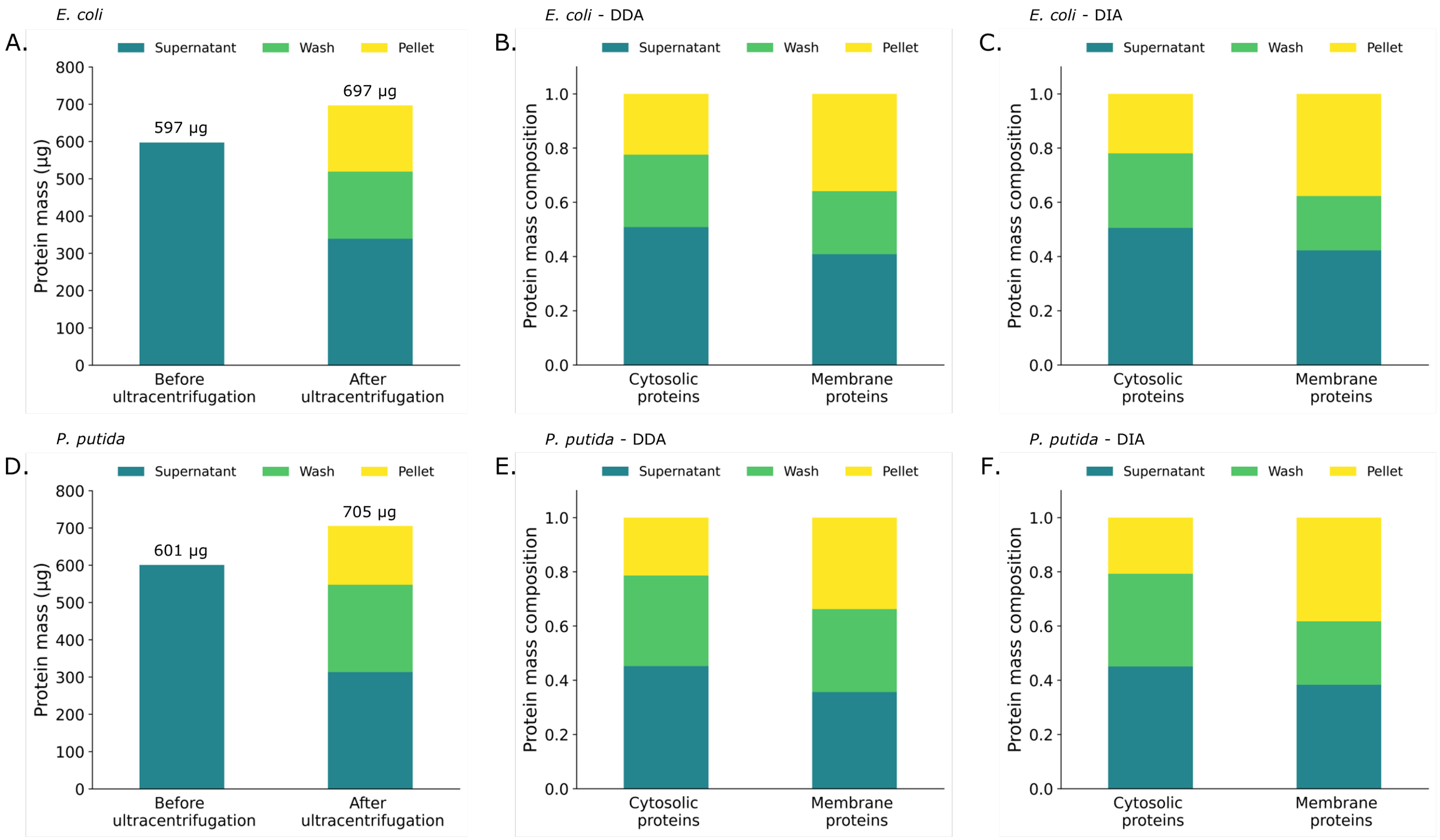


Figure S3.8: Protein mass and proteome composition analysis of the samples from the ultracentrifugation protocol, namely the supernatant, wash and pellet samples. (A) Protein mass analysis of the *E. coli* samples. (B) Proteome composition analysis in terms of fractionation of cytosolic and membrane proteins within the supernatant, wash and pellet samples of *E. coli* analysed with the DDA method. (C) Proteome composition analysis in terms of fractionation of cytosolic and membrane proteins within the supernatant, wash and pellet samples of *E. coli* analysed with the DIA method. (D) Protein mass analysis of the *P. putida* samples. (E) Proteome composition analysis in terms of fractionation of cytosolic and membrane proteins within the supernatant, wash and pellet samples of *P. putida* analysed with the DDA method. (F) Proteome composition analysis in terms of fractionation of cytosolic and membrane proteins within the supernatant, wash and pellet samples of *P. putida* analysed with the DIA method.

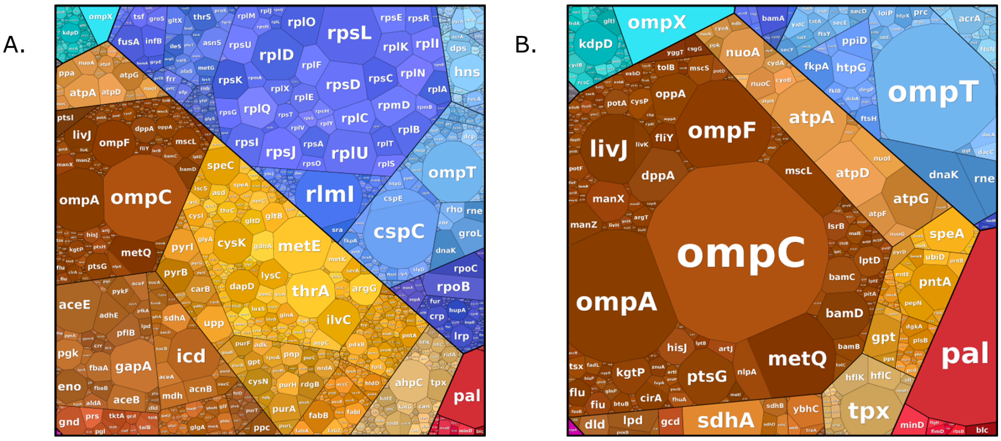


Figure S3.9: Proteomaps of the total proteome and the membrane proteome of *E. coli*. (A) Proteomap of the total proteome with level 5 annotation. Out of 2,646 proteins submitted, 1,800 were supported by the *E. coli* treemap. (B) Proteomap of the membrane proteome with level 5 annotation. Out of 794 submitted proteins, 533 were supported by the *E. coli* treemap.

Supplementary tables

Table S3.3: Overview of *E. coli* membrane peptide mix.

Peptide	Protein	Gene	Description
DNIFGSIEEDAQR	P0ABF1	pcnB	Poly(A) polymerase I
LATLLNDIPPAR	P0ABF1	pcnB	Poly(A) polymerase I
FYDASSYAGK	P19934	tolA	Tol-Pal system protein TolA
IPKPPSQAVYEVFK	P19934	tolA	Tol-Pal system protein TolA
VIPLPDEYK	P0AEH1	rseP	Regulator of sigma-E protease RseP
QYGPFNAIVEATDK	P0AEH1	rseP	Regulator of sigma-E protease RseP
GFELPEDVGR	P69931	hda	DnaA regulatory inactivator Hda
TLFMTLDQLDR	P69931	hda	DnaA regulatory inactivator Hda
NVPFESGIDSVGSAR	P0AFC3	nuoA	NADH-quinone oxidoreductase subunit A
LVLLSDPVTMAR	P62517	mdoH	Glucans biosynthesis glucosyltransferase H
IGALDWTPAR	P0AFC3	nuoA	NADH-quinone oxidoreductase subunit A
MNPETNSIANR	P0AFC3	nuoA	NADH-quinone oxidoreductase subunit A
GHIEAATAFGFTR	P52094	hisQ	Histidine transport system permease protein HisQ
YALPGIGNNWQVILK	P52094	hisQ	Histidine transport system permease protein HisQ
VIQGIGGAMMMPVAR	P31474	hsrA	Probable transport protein HsrA
FMLQDLLGVVSPGLK	P15078	cstA	Peptide transporter CstA
LIFTGGVAK	P0AB01	elyC	Envelope biogenesis factor ElyC
FALVVVINDPQAGK	P0AD68	ftsI	Peptidoglycan D,D-transpeptidase FtsI
VAWLQVISPDMLVK	P0AD68	ftsI	Peptidoglycan D,D-transpeptidase FtsI
EVHDAGGISVGDR	P0AD68	ftsI	Peptidoglycan D,D-transpeptidase FtsI
VQNALQSVPGVTQAR	Q59385	copA	Copper-exporting P-type ATPase
TGTLTEGKPQVVAVK	Q59385	copA	Copper-exporting P-type ATPase
EYQVVIDPQR	P38054	cusA	Cation efflux system protein CusA
LIMSVPEVAR	P38054	cusA	Cation efflux system protein CusA
GSSLVTGIVDAR	P23930	lnt	Apolipoprotein N-acyltransferase
YDTYNTIITLGK	P23930	lnt	Apolipoprotein N-acyltransferase
SIGPWQHFQMAR	P23930	lnt	Apolipoprotein N-acyltransferase
DFAGAQNLLAK	P45464	lpoA	Penicillin-binding protein activator LpoA
ALIAQEPLLGAK	P45464	lpoA	Penicillin-binding protein activator LpoA
TIQQGFEAAK	P45464	lpoA	Penicillin-binding protein activator LpoA
QLAHGDFGPSFK	P0AFH2	oppB	Oligopeptide transport system permease protein OppB
QVAILAVAGAEK	P0ABA0	atpF	ATP synthase subunit b
DQEVHAVVQDWK	P46889	ftsK	DNA translocase FtsK
MDDDEEITYTAR	P46889	ftsK	DNA translocase FtsK
EGIGPQLPRPK	P46889	ftsK	DNA translocase FtsK
LMVALDVGGAIK	P0A935	mltA	Membrane-bound lytic murein transglycosylase A
LTGEESDTLR	P06282	cdh	CDP-diacylglycerol pyrophosphatase
IAAEEIENLLLR	P10378	entE	Enterobactin synthase component E
MGDYIVAYALPDDVK	P15877	gcd	Quinoprotein glucose dehydrogenase
TGNIFVLDR	P15877	gcd	Quinoprotein glucose dehydrogenase
EVLQQFDDQSR	P0AFA7	nhaB	Na(+)/H(+) antiporter NhaB
AAMESGAITLK	P0AFA7	nhaB	Na(+)/H(+) antiporter NhaB
QMQMNAQAEK	P0AC02	bamD	Outer membrane protein assembly factor BamD
NIDWYAFSLPK	P08369	creD	Inner membrane protein CreD
ADNPFDLLLPAAMAK	P0AC23	focA	Probable formate transporter 1

Table S3.3: Overview of *E. coli* membrane peptide mix - continued.

Peptide	Protein	Gene	Description
IIAGNIINFSR	P0C0S1	mscS	Small-conductance mechanosensitive channel
QGQEIIAGNFR	P0AEQ6	glnP	Glutamine transport system permease protein GlnP
TTLTDLTHSLK	P23837	phoQ	Sensor protein PhoQ
TPLAVLQSTLR	P23837	phoQ	Sensor protein PhoQ
EQQLLAEAR	P0ABU7	exbB	Biopolymer transport protein ExbB
AAFDDAIAAR	P0ABC7	hflK	Modulator of FtsH protease HflK
AQTILEAQGEVAR	P0ABC7	hflK	Modulator of FtsH protease HflK
GGNLMVLPLDQMLK	P0ABC7	hflK	Modulator of FtsH protease HflK
VATISGSDAK	P09127	hemX	Protein HemX
LLVAAQAVPR	P09127	hemX	Protein HemX
QALENVSTWVR	P09127	hemX	Protein HemX
SVDTPVIGLK	P0ABJ9	cydA	Cytochrome bd-I ubiquinol oxidase subunit 1
AYSLLEQLR	P0ABJ9	cydA	Cytochrome bd-I ubiquinol oxidase subunit 1
YHFEQSSTTTQPAR	P0ABJ9	cydA	Cytochrome bd-I ubiquinol oxidase subunit 1
GIALYYGGR	P0AFB1	nlpI	Lipoprotein NlpI
LAVANNVHNFVEHR	P0AFB1	nlpI	Lipoprotein NlpI
AIVEAAHEFGR	P07001	pntA	NAD(P) transhydrogenase subunit alpha
AEMELFAAQAK	P07001	pntA	NAD(P) transhydrogenase subunit alpha
SLTNVNAQFQR	P60752	msbA	ATP-dependent lipid A-core flippase
AIQAALDELQK	P60752	msbA	ATP-dependent lipid A-core flippase
LPGGQLEQAR	P0ABI4	corA	Magnesium transport protein CorA
MIFIVFHGK	P33607	nuoL	NADH-quinone oxidoreductase subunit L
AQSVVDYLISK	P0A910	ompA	Outer membrane protein A
IGSDAYNQGLSER	P0A910	ompA	Outer membrane protein A
GVVGQVVAVAK	P16926	mreC	Cell shape-determining protein MreC
FPEGYPVAVVSSVK	P16926	mreC	Cell shape-determining protein MreC
GDTPEVLHFDISSR	P0AFX7	rseA	Anti-sigma-E factor RseA
INAMLQDYELQR	P0AFX7	rseA	Anti-sigma-E factor RseA
DQFVQPVVK	P0ADB1	osmE	Osmotically-inducible putative lipoprotein OsmE
EILVEHYDNIEQK	P0C0L7	proP	Proline/betaine transporter
IDDIDHEIADLQAK	P0C0L7	proP	Proline/betaine transporter
EYAVQQNINILR	P0AG90	secD	Protein translocase subunit SecD
NIAQWLQENGITR	Q46833	yghE	Putative type II secretion system L-type protein YghE
GQFENAFNSER	P0ADY1	ppiD	Periplasmic chaperone PpiD
LIDEALLDQYAR	P0ADY1	ppiD	Periplasmic chaperone PpiD
ALDAYYALQQK	P0ADY1	ppiD	Periplasmic chaperone PpiD
MNIFEQTPPNR	P0AAA1	yagU	Inner membrane protein YagU
HQAIQALMR	P76278	yebZ	Inner membrane protein YebZ
YYGVAYGYSK	P69805	manZ	PTS system mannose-specific EIID component
VAINFFYYPDDK	P77538	yfhR	Uncharacterized protein YfhR
VLPLTGVVSPR	P0ADE4	tamA	Translocation and assembly module subunit TamA
WESPVGPIK	P0ADE4	tamA	Translocation and assembly module subunit TamA
FLENAMYASR	P67244	yqhA	UPF0114 protein YqhA
LIAETAPDANNLLR	P0AG00	wzzE	ECA polysaccharide chain length modulation protein
QYVAFASQR	P0AG00	wzzE	ECA polysaccharide chain length modulation protein
IAEQHNISR	P0AG00	wzzE	ECA polysaccharide chain length modulation protein

Table S3.3: Overview of *E. coli* membrane peptide mix - continued.

Peptide	Protein	Gene	Description
LPFDDGVMSQYK	P76507	yfdI	Uncharacterized protein YfdI
YHATYFGSYLYMK	P76507	yfdI	Uncharacterized protein YfdI
FNAFGDGVAQLGR	P46130	ybhC	Putative acyl-CoA thioester hydrolase YbhC
GAVVFDNTEFR	P46130	ybhC	Putative acyl-CoA thioester hydrolase YbhC
QLQQNATQAEVNR	P64429	ypfJ	Uncharacterized protein YpfJ
VAPITTGDVVLQSAR	P0ADC3	lolC	Lipoprotein-releasing system transmembrane protein LolC
SVAVGVMLGIDPAQK	P0ADC3	lolC	Lipoprotein-releasing system transmembrane protein LolC
AQYDTVLANEVTAR	P02930	tolC	Outer membrane protein TolC
TDKPQPVNALLK	P02930	tolC	Outer membrane protein TolC
QNLLDIESLK	P76372	wzzB	Chain length determinant protein
VDDLDIHAYR	P76372	wzzB	Chain length determinant protein
ELLTNDPFSSR	P75818	ybjP	Uncharacterized lipoprotein YbjP
YLGGSVHATAGTLR	P75818	ybjP	Uncharacterized lipoprotein YbjP
DLNNESTPMAFENIK	P64451	ydcL	Uncharacterized lipoprotein YdcL
SVAAFAAVDQQGIER	P76221	ydjZ	TVP38/TMEM64 family inner membrane protein YdjZ
HGLLDAQEYQR	P0AD27	yejM	Inner membrane protein YejM
AAGNVDDQINR	P0AD27	yejM	Inner membrane protein YejM
LDNTVVIITAGR	P0AD27	yejM	Inner membrane protein YejM
VIEQGNVMVLGGDMR	P37624	rbbA	Ribosome-associated ATPase
FGSFVAVDHVNFR	P37624	rbbA	Ribosome-associated ATPase
IEISSALNSTDMR	P39838	rcsD	Phosphotransferase RcsD
GVFAMLNLVPGK	P39838	rcsD	Phosphotransferase RcsD
SPVEPVQSTAPQPK	P69411	rcsF	Outer membrane lipoprotein RcsF
IYLVDQNDR	P77774	bamB	Outer membrane protein assembly factor BamB
QAQQLAEQQR	P29131	ftsN	Cell division protein FtsN
QLLEQMQADMR	P29131	ftsN	Cell division protein FtsN
TSQAAPVQAQPR	P29131	ftsN	Cell division protein FtsN
QVVATATFR	P06971	fhuA	Ferrichrome outer membrane transporter/phage receptor
IYDDAAVER	P31554	lptD	LPS-assembly protein LptD
FQLTPVDVITAIK	P31224	acrB	Multidrug efflux pump subunit AcrB
STGEAMELMEQLASK	P31224	acrB	Multidrug efflux pump subunit AcrB
LMLDSAGSVAFFR	P0A6H8	clsA	Cardiolipin synthase A
ALFATGNFEDVR	P0A940	bamA	Outer membrane protein assembly factor BamA
VQSMPEINDADK	P0A940	bamA	Outer membrane protein assembly factor BamA
LGFFETVDTDTQR	P0A940	bamA	Outer membrane protein assembly factor BamA
DPQLQVVTNK	P0AEE1	dcrB	Protein DcrB
MQQLDSIISAK	P0AEE1	dcrB	Protein DcrB
DTIGDIIILPR	P0A7A7	plsB	Glycerol-3-phosphate acyltransferase
EQAVLMTYYR	P0A7A7	plsB	Glycerol-3-phosphate acyltransferase
MTDLTAQEPAWQTR	P33599	nuoC	NADH-quinone oxidoreductase subunit C/D
IYDLVEAITGFR	P33599	nuoC	NADH-quinone oxidoreductase subunit C/D
FGPDAFTVQATR	P33599	nuoC	NADH-quinone oxidoreductase subunit C/D
SQGVAAYGAK	P33599	nuoC	NADH-quinone oxidoreductase subunit C/D

Table S3.3: Overview of *E. coli* membrane peptide mix - continued.

Peptide	Protein	Gene	Description
QSVDQPVQTGYK	P0ABB0	atpA	ATP synthase subunit alpha
IGSFEAALLAYVDR	P0ABB0	atpA	ATP synthase subunit alpha
LESLVEDLVNR	P0A6Y8	dnaK	Chaperone protein DnaK
ASSGLNEDEIQK	P0A6Y8	dnaK	Chaperone protein DnaK
GYASLDYNFK	P60785	lepA	Elongation factor 4
EVIYVDSPSK	P60785	lepA	Elongation factor 4
VLVVGGSQGAR	P17443	murG	UDP-N-acetylglucosamine–N-acetylmuramyl-(pentapeptide) pyrophosphoryl-undecaprenol N-acetylglucosamine transferase
ILNQTMPQVAAK	P17443	murG	UDP-N-acetylglucosamine–N-acetylmuramyl-(pentapeptide) pyrophosphoryl-undecaprenol N-acetylglucosamine transferase
AAIAAAQANPNAK	P00363	frdA	Fumarate reductase flavoprotein subunit
VINQLTGGLAGMAK	P0A9P0	lpdA	Dihydrolipoyl dehydrogenase
YDAVLVAIGR	P0A9P0	lpdA	Dihydrolipoyl dehydrogenase
GVHEGHVAAEVIAGK	P0A9P0	lpdA	Dihydrolipoyl dehydrogenase
GGVNDLESVVK	P0ABH0	ftsA	Cell division protein FtsA
ESHLNGEAEVEK	P0ABH0	ftsA	Cell division protein FtsA
IITNLTEGASR	P10121	ftsY	Signal recognition particle receptor FtsY
LFHEAVGLTGITLTK	P10121	ftsY	Signal recognition particle receptor FtsY
ADDFIEALFAR	P10121	ftsY	Signal recognition particle receptor FtsY
YSDHIALPVEIEK	P0A6Z3	htpG	Chaperone protein HtpG
EILQDSTVTR	P0A6Z3	htpG	Chaperone protein HtpG
YIFELNPDHVLVK	P0A6Z3	htpG	Chaperone protein HtpG
LDDIAAWTYR	P33602	nuoG	NADH-quinone oxidoreductase subunit G
GADVGITMIAR	P33602	nuoG	NADH-quinone oxidoreductase subunit G
VVDGAMQIIGR	P0AFL6	ppx	Exopolyphosphatase
LIFMGVEHTQPEK	P0AFL6	ppx	Exopolyphosphatase
IQGWNVAMGASGTIK	P0AFL6	ppx	Exopolyphosphatase
VLDTTMQMYEQWR	P0AFL6	ppx	Exopolyphosphatase
GILNALYEAK	P23830	pssA	CDP-diacylglycerol–serine O-phosphatidyltransferase
LQYYVNTDQLVVR	P23830	pssA	CDP-diacylglycerol–serine O-phosphatidyltransferase
DGYAIGTGR	P07000	pldB	Lysophospholipase L2
ALPFAINVLTHSR	P07000	pldB	Lysophospholipase L2
AVIDGEPITER	P77611	rsxC	Ion-translocating oxidoreductase complex subunit C
TAVEAAIAR	P77611	rsxC	Ion-translocating oxidoreductase complex subunit C
LVQQWLDILGIDK	P31060	modF	ABC transporter ATP-binding protein ModF
LVLEVQQQLGGGIVR	P0ABB4	atpD	ATP synthase subunit beta
IMQQLAEEGK	P07109	hisP	Histidine transport ATP-binding protein HisP
TMVVVTHEMGFAR	P07109	hisP	Histidine transport ATP-binding protein HisP
IAEDGNPQVLIK	P10346	glnQ	Glutamine transport ATP-binding protein GlnQ
QYSGVNVLK	P77257	lsrA	Autoinducer 2 import ATP-binding protein LsrA
VMLQAYDEGR	P0ABA6	atpG	ATP synthase gamma chain
LSVYDNLMAVLQIR	P0A9V1	lptB	Lipopolysaccharide export system ATP-binding protein LptB

Table S3.3: Overview of *E. coli* membrane peptide mix - continued.

Peptide	Protein	Gene	Description
ANELMEEFHIEHLR	P0A9V1	lptB	Lipopolysaccharide export system ATP-binding protein LptB
MEQSVANLVDMR	P63386	mlaF	Intermembrane phospholipid transport system ATP-binding protein MlaF
NEAELNALWDSK	P23865	prc	Tail-specific protease
LDDVVALIK	P23865	prc	Tail-specific protease
GDMLSMEDVLEILR	P0AEZ3	minD	Septum site-determining protein MinD
LLGEERPFR	P0AEZ3	minD	Septum site-determining protein MinD
DMMLLDALIQLK	P07014	sdhB	Succinate dehydrogenase iron-sulfur subunit
MQDYTLEADEGR	P07014	sdhB	Succinate dehydrogenase iron-sulfur subunit
QGQDLMVQVVK	P0A9J0	rng	Ribonuclease G
FLVYASPAVAEALK	P0A9J0	rng	Ribonuclease G
ILAQSIEVYQR	P10408	secA	Protein translocase subunit SecA
GVHVVTVNDYLAQR	P10408	secA	Protein translocase subunit SecA
TFAHVDPVK	P0AC41	sdhA	Succinate dehydrogenase flavoprotein subunit
LGGNSLLDLVVFGR	P0AC41	sdhA	Succinate dehydrogenase flavoprotein subunit
VLEQIAAQMR	P0AFI2	parC	DNA topoisomerase 4 subunit A
EVAQAAIALIDQPK	P0AFI2	parC	DNA topoisomerase 4 subunit A
IVYAMSELGLNASAK	P0AFI2	parC	DNA topoisomerase 4 subunit A
YAHIGTGNFNEK	P0A7B1	ppk	Polyphosphate kinase
QLSVNQQNWLR	P0A7B1	ppk	Polyphosphate kinase
DMPNALVEVLR	P0A7B1	ppk	Polyphosphate kinase
IMLVGGGNIGAGLAR	P0AGI8	trkA	Trk system potassium uptake protein TrkA
IEQGDHVIMFLTDK	P0AGI8	trkA	Trk system potassium uptake protein TrkA
GIAIIHQELALVK	P37388	xylG	Xylose import ATP-binding protein XylG
AFFDSNNDDMLVK	P37751	wbbK	Putative glycosyltransferase WbbK
NTAVFVQQFWMK	P37751	wbbK	Putative glycosyltransferase WbbK
TFNEFFVRPLR	P0A8K1	psd	Phosphatidylserine decarboxylase proenzyme
VNLVEQLESLSVTK	P0A8K1	psd	Phosphatidylserine decarboxylase proenzyme
IVSGDDVDLNR	P0AAB4	ubiD	3-octaprenyl-4-hydroxybenzoate carboxy-lyase
LNDMVNWGR	P0AFC7	nuoB	NADH-quinone oxidoreductase subunit B
LYDQMLEPK	P0AFC7	nuoB	NADH-quinone oxidoreductase subunit B
AISLSGEAQSGR	P77804	ydgA	Protein YdgA
QQVEGASAMGQMFR	P77804	ydgA	Protein YdgA
LIHGQVATR	P69797	manX	PTS system mannose-specific EIIAB component
ITSVNVGGMAFR	P69797	manX	PTS system mannose-specific EIIAB component
LVSVHINDEPWTAWK	P04825	pepN	Aminopeptidase N
EVALELYVDR	P04825	pepN	Aminopeptidase N
WDAAQSLLATYIK	P04825	pepN	Aminopeptidase N
DVVIISTLR	P06149	dld	Quinone-dependent D-lactate dehydrogenase
EQMLELLQQR	P06149	dld	Quinone-dependent D-lactate dehydrogenase
SDIALIQIQNPK	P0C0V0	degP	Periplasmic serine endoprotease DegP
DVPPGNMFR	C1P605	azuC	Uncharacterized protein AzuC
TFMSISGAQIR	P27278	nadR	Trifunctional NAD biosynthesis/regulator protein NadR
IALGHAQYIDFAVK	P27278	nadR	Trifunctional NAD biosynthesis/regulator protein NadR

Table S3.3: Overview of *E. coli* membrane peptide mix - continued.

Peptide	Protein	Gene	Description
EHPFVQALIDEYR	P27278	nadR	Trifunctional NAD biosynthesis/regulator protein NadR
AGYNLVASATEGQMR	P0AEB2	dacA	D-alanyl-D-alanine carboxypeptidase DacA
LVESQGGEIIFNGQR	P75796	gsiA	Glutathione import ATP-binding protein GsiA
AAFDFAVEHQSVER	P0ABA4	atpH	ATP synthase subunit delta
WQDMLAFAAEVTK	P0ABA4	atpH	ATP synthase subunit delta
AGDMVIDGSVR	P0ABA4	atpH	ATP synthase subunit delta
LTDMVTVGK	P08506	dacC	D-alanyl-D-alanine carboxypeptidase DacC
ALIHDVPEEYAIHK	P08506	dacC	D-alanyl-D-alanine carboxypeptidase DacC
FVLEFMGEVNR	P16676	cysA	Sulfate/thiosulfate import ATP-binding protein CysA
LFVGLQHAR	P16676	cysA	Sulfate/thiosulfate import ATP-binding protein CysA
APVILAVNK	P06616	era	GTPase Era

References

[1] A. V. Sastry, Y. Gao, R. Szubin, Y. Hefner, S. Xu, D. Kim, K. S. Choudhary, L. Yang, Z. A. King, and B. O. Palsson. “The Escherichia coli transcriptome mostly consists of independently regulated modules”. In: *Nature Communications* 10.1 (2019), p. 5536. DOI: `10.1038/s41467-019-13483-w`.

Chapter 4.

Investigation of the aromatic amino acid biosynthesis in *Escherichia coli* - applications of a kinetic model

Shannara Kayleigh Taylor Parkins, Nicholas Luke Cowie, Jorge Carrasco Muriel, Teddy Groves and Lars Keld Nielsen

Aromatic amino acids are central precursors to a variety of natural products with important applications such as food additives and pharmaceuticals. The *Escherichia coli* aromatic amino acid biosynthesis pathway converts phosphoenolpyruvate and erythrose 4-phosphate into the three aromatic amino acids, L-phenylalanine, L-tyrosine and L-tryptophan. We are interested in developing an L-tyrosine chassis strain for use as a base strain in natural product development. After numerous metabolic engineering efforts, challenges remain for the development of an *E. coli* L-tyrosine overproducer due to the tightly regulated pathway. Here, a kinetic model covering the reactions involved in *E. coli* aromatic amino acid biosynthesis was constructed. The kinetic model takes into account mass conservation, allosteric regulations, and thermodynamic constraints in order to realistically simulate enzymatic fluxes and steady state metabolite concentrations. A multi-omics data set of five *E. coli* strains producing varying titers of L-tyrosine, as well as a specification of prior distributions based on detailed literature research, were used to fit the kinetic model. This resulted in a feasible parameterisation of the kinetic model whose statistical properties reflect the information contained in the omics dataset, as well as prior information. Simulated and experimental values of metabolite concentrations and reaction fluxes were in agreement. Additionally, proposed metabolic engineering targets including overexpression of multiple enzymes and feedback-resistant versions of the DDPA (3-deoxy-7-phosphoheptulonate synthase) *aroG* and the CHORM (chorismate mutase) and PPND (prephenate dehydrogenase) *tyrA* enzymes, aligned with experimentally tested metabolic engineering strategies for L-tyrosine overproduction. Furthermore, a *ptsHIcrr* knockout strain was combined with constitutive expression of feedback-resistant versions of *aroG* and *tyrA*, and a codon-optimised version of *aroB*. The engineered strain displayed significantly higher protein levels of the three gene targets compared to the base strain. These protein changes should lead to higher activity in the aromatic amino acid biosynthesis pathway, which will be confirmed with further research. The kinetic model will be an advantageous tool for further metabolic engineering and the construction of *E. coli* L-tyrosine overproducers.

4.1. Introduction

The aromatic amino acids, L-phenylalanine, L-tyrosine and L-tryptophan, are precursors for many natural products. These aromatic amino acid derivatives are in high demand, yet chemical synthesis or extraction from natural sources is often costly and time consuming. Microbial production has the potential for higher yields in shorter time frames [1]. Microbial aromatic amino acid biosynthesis is however tightly regulated, which results in great difficulties with metabolic engineering of this pathway towards industrially relevant titres of the derivatives [2]. Specifically L-tyrosine overproduction in the gram-negative bacterium *Escherichia coli* has been the target of numerous metabolic engineering efforts [3, 4]. Addition of feedback-resistant versions of two key enzymes increased L-tyrosine titres substantially and demonstrated the extent of allosteric regulation within the pathway. Nevertheless, multiple allosteric regulations remain and an appropriate strategy for balancing L-phenylalanine production with L-tyrosine overproduction was not identified. This highlights the need for alternative tools to aid the rational metabolic engineering for L-tyrosine overproduction.

Kinetic modelling connects the proteome with the fluxome and metabolome in a mechanistically detailed way, which has a strong advantage over constraint-based modelling when dealing with highly regulated pathways [5]. The application of kinetic models provides a tool to investigate the intricacies of metabolic pathways including thermodynamic constraints, enzymatic mechanisms and allosteric regulations [6]. For *E. coli*, multiple kinetic models of (parts of) the central carbon metabolism exist, where each kinetic model implements a different approach in order to represent metabolism [7]. The first *E. coli* kinetic model [8] was followed up by increasingly larger kinetic models [9, 10, 11] and the latest kinetic models cover all of central carbon metabolism [12, 13, 14]. A trade-off exists between the size and complexity of a kinetic model and the choice of appropriate boundaries will depend on the objective of the kinetic model. Recently, the first kinetic model of *E. coli* central carbon metabolism and aromatic amino acid biosynthesis was developed and applied to the optimisation of aromatic amino acid titres [15]. This kinetic model includes the required enzymatic reactions and associated regulatory network, however it was not trained on any experimental data, therefore model parameterisation is a potential area for further improvement.

Here, we constructed a kinetic model of the aromatic amino acid biosynthesis in *E. coli* (Figure 4.1). The kinetic model takes into account mass conservation, allosteric regulation and thermodynamic constraints, in order to infer reaction fluxes and metabolite concentrations. A multi-omics data set of five *E. coli* strains, one wild type and four mutant strains [16, 17, 18, 19, 20, 21], producing varying concentrations of intracellular L-tyrosine was used to fit the kinetic model. This resulted in a feasible parameterisation of the kinetic model whose statistical properties reflect the information contained in the multi-omics data set, as well as prior information. The simulated and experimental values of reaction fluxes and metabolite concentrations aligned, granting a platform for detailed analysis of the pathway.

For L-tyrosine production specifically, our results suggested overexpression of most enzymes and implementing feedback-resistance of multiple enzymes as metabolic engineering targets beneficial for overproduction. Feedback-resistant versions of two enzymes were incorporated in an *E. coli* chassis strain to improve L-tyrosine production as ongoing study. Overall, the kinetic model will be an advantageous tool for metabolic engineering and the construction of *E. coli* aromatic amino acid overproducers, in particular L-tyrosine.

4.2. Material and methods

4.2.1. Kinetic model structure

The boundaries of the kinetic model were set to encompass the entire aromatic amino acid biosynthesis pathway, starting at phosphoenolpyruvate and erythrose 4-phosphate and ending in the three aromatic amino acids. While phosphoenolpyruvate and erythrose 4-phosphate were set as unbalanced boundary metabolites, the aromatic amino acids were incorporated as balanced metabolites to allow dynamic evaluation of the metabolite concentrations and regulations. Accordingly, a grouped reaction estimating all reactions and transport consuming the metabolite was set for each aromatic amino acid. Multiple allosteric regulations play a key role in the functionality of the pathway in *E. coli* (Figure 4.1): allosteric inhibition of the three isozymes catalysing the first reaction of the pathway, DDPA (3-deoxy-7-phosphoheptulonate synthase), each inhibited by one of the aromatic amino acids, and allosteric inhibition of the first enzyme of the corresponding branch by each aromatic amino acid, the CHORM (chorismate mutase), PPNDH (prephenate dehydratase) and PPND (prephenate dehydrogenase) reactions for L-phenylalanine and L-tyrosine, and the ANS (anthranilate synthase) and ANPRT (anthranilate phosphoribosyltransferase) reactions for L-tryptophan. These allosteric inhibitions were all included in the kinetic model. The pathway structure, reaction stoichiometry and reaction reversibility were obtained from the EcoCyc database [22], while the prior distributions for the kinetic parameters were obtained from the EcoCyc [22], BRENDA [23] and Sabio-RK [24] databases. An overview of the literature research on the reactions of the aromatic amino acid biosynthesis in *E. coli* can be found in the Supplementary materials section.

The KALE multi-omics data set includes fluxomics, endo-metabolomics and proteomics data of five *E. coli* strains, one wild type and four mutant strains [16, 17, 18, 19, 20, 21], which were used as the experimental values to fit the kinetic model. The strains included in the experimental data set for model fitting, hereafter referred to as experiments, were the eWT01 strain (evolved wild type strain, replicate 1), ePgi04 (evolved Δ*pgi* strain, replicate 4), ePtsHIcrr03 (evolved Δ*ptsHIcrr* strain, replicate 3), eSdhCB03 (evolved Δ*sdhCB* strain, replicate 3) and eTpiA02 (evolved Δ*tpiA* strain, replicate 2). These strains were selected in order to maximise differences between intracellular concentrations of the aromatic amino acids, specifically for L-tyrosine. All experimental values were taken from samples of the mid-exponential phase, thus assumed to be pseudo steady-state measurements.

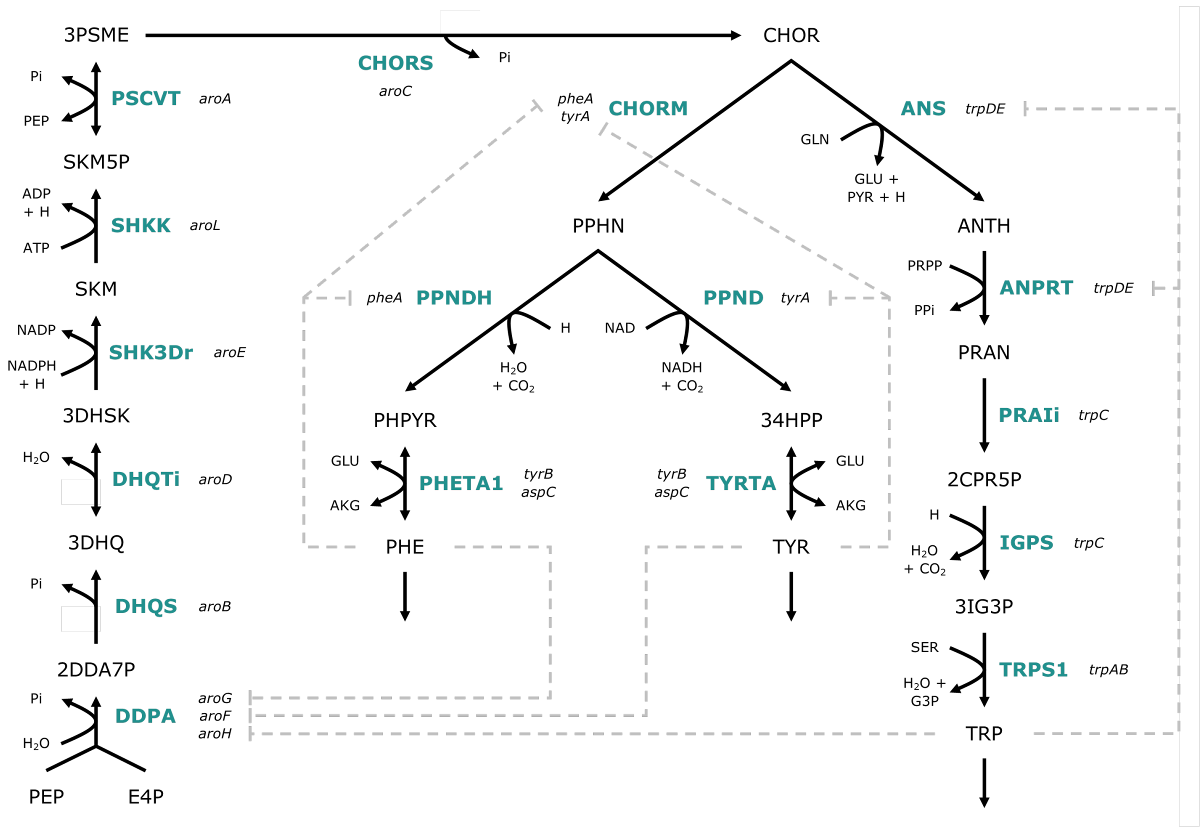


Figure 4.1: Metabolic map of the aromatic amino acid biosynthesis in *Escherichia coli* which is incorporated in the current kinetic model. The aromatic amino acid biosynthesis consists of the shikimate pathway towards chorismate after which the pathway splits into three branches, each producing one of the aromatic amino acids. All reactions are in bold teal with the corresponding gene names in italics and allosteric inhibitions are represented by grey dashed lines. For each aromatic amino acid, a single reaction combining all metabolic reactions and transport consuming the metabolite was included in the kinetic model. PEP: phosphoenolpyruvate; E4P: erythrose 4-phosphate; DDPA: 3-deoxy-7-phosphoheptulonate synthase; 2DDA7P: 2-dehydro-3-deoxyarabino-heptulosonate 7-phosphate; DHQS: 3-dehydroquinate synthase; 3DHQ: 3-dehydroquinate; DHQTi: 3-dehydroquinate dehydratase; 3DHSK: 3-dehydroshikimate; SHK3Dr: shikimate dehydrogenase; SKM: shikimate; SHKK: shikimate kinase; SKM5P: shikimate 5-phosphate; PSCVT: 3-phosphoshikimate 1-carboxyvinyltransferase; 3PSME: 5-O-(1-carboxyvinyl)-3-phosphoshikimate; CHORS: chorismate mutase; CHOR: chorismate; CHORM: chorismate mutase; PPHN: prephenate; PPNDH: prephenate dehydratase; PHPYR: phenylpyruvate; PHETA1: L-phenylalanine transaminase; PHE: L-phenylalanine; PPND: prephenate dehydrogenase; 34HPP: 3-(4-hydroxyphenyl)pyruvate; TYRTA: L-tyrosine aminotransferase; TYR: L-tyrosine; ANS: anthranilate synthase; ANTH: anthranilate; ANPRT: anthranilate phosphoribosyltransferase; PRAN: N-(5-phosphoribosyl)anthranilate; PRAIi: phosphoribosylanthranilate isomerase; 2CPR5P: 1-(2-carboxyphenylamino)-1-deoxyribulose 5-phosphate; IGPS: indole 3-glycerol-phosphate synthase; 3IG3P: (3-indolyl)-glycerol 3-phosphate; TRPS1: L-tryptophan synthase; TRP: L-tryptophan; H_2O: water; Pi: phosphate; NADP(H): nicotinamide adenine dinucleotide phosphate; ATP: adenosine triphosphate; ADP: adenosine diphosphate; CO_2: carbon dioxide; GLU: L-glutamate; AKG: 2-oxoglutarate; NAD(H): nicotinamide adenine dinucleotide; GLN: L-glutamine; PYR: pyruvate; PRPP: 5-phosphoribose 1-diphosphate; PPi: diphosphate; SER: L-serine; G3P: glyceraldehyde 3-phosphate.

To obtain more accurate values for the enzyme concentrations from the proteomics data, the raw MS files of Heckmann *et al.* [21] were reprocessed using *autoprot* REF chapter 1. A workflow using Proteome Discoverer for DDA data analysis, xTop [25] for protein inference and the TPA method [26, 27] as standard-free quantification approach was employed for the reprocessing. The original data analysis relied on the UPS2 protein mix as unlabelled internal standard for absolute quantification, however a maximum of four UPS2 proteins was identified and quantified in the samples, with only two UPS2 proteins in most samples. Even though the standard-free quantification approach is not ideal for DDA data analysis (REF chapter 1), it was deemed more suitable than the unlabelled quantification approach in this case, due to the scarcity of quantified UPS2 proteins. The enzyme concentrations from the reprocessed data can be found in Table S4.1.

The Maud kinetic modelling framework (`https://pypi.org/project/maud-metabolic-models`) was used to construct and fit the kinetic model to the experimental data set. Maud combines a mechanistically and thermodynamically realistic kinetic model with a plausible measurement model and prior model, resulting in a statistically sound basis for kinetic parameter estimation and prediction of behaviour in new experimental conditions. The prior model consists mostly of curated and marginally independent distributions, which were set based on a detailed literature search. The full prior specification can be found in Table S4.2. Independent log-normal prior distributions were used with location parameters set based on literature search and scale parameters set to 0.2. Since the log-normal distribution has constant coefficient of variation and we judged that the proportional accuracy of the literature-based estimates were about the same, it was appropriate to use the same value despite the wide range of location parameters. Transfer constants between the active and inactive form of allosteric enzymes are generally unknown and were therefore set to 1.0 with an uncertainty of 0.5 on natural logarithmic scale to allow shifting of the equilibrium to either enzyme form. Reversible and irreversible reactions were modelled using modular rate laws [28]. All pathway intermediates, including the aromatic amino acids, were set as balanced metabolites, while phosphoenolpyruvate, erythrose 4-phosphate and all co-factors were set as unbalanced metabolites. Prior distributions are set for the intracellular concentrations of unbalanced metabolites based on the experimental data, while the intracellular concentrations of unbalanced metabolites are simulated during fitting of the kinetic model. In order to model prior information about our system's thermodynamic properties, a multivariate normal distribution was used. The mean vector and covariance matrix for this distribution were derived from eQuilibrator [29, 30] using the eQuilibrator API package.

The Maud kinetic modelling framework models enzyme-catalysed fluxes according to Equation 4.1, making it possible to disambiguate the different factors contributing to flux regulation.

$$flux = [E] * k_{cat} * reversibility * saturation * allostery \tag{4.1}$$

In Equation 4.1, the $flux$ is calculated in (mol/L)/s, $[E]$ is the enzyme concentration in mol/L, k_{cat} is the catalytic rate constant in 1/s, *saturation* is the saturation factor calculated following Equation 4.2, *reversibility* is the reversibility factor calculated following Equation 4.3, *allostery* is the allostery factor calculated following Equation 4.4.

$$saturation \ = \prod_{substrates} \frac{x_s}{K_{M,s}} \ * \ free\ enzyme\ ratio \tag{4.2}$$

In Equation 4.2, *saturation* is the saturation factor representing the fraction of the enzyme bound to the complete set of substrates required for the reaction, x_s is the substrate concentration, $K_{M,s}$ is the Michaelis constant of the substrate and the $free\ enzyme\ ratio$ is the ratio of free enzyme to bound enzyme as defined in Liebermeister *et al.* [28].

$$\begin{aligned} reversibility \ &= \ 1 \ - \ e^{\frac{\Delta_r G' \ + \ RT \ * \ S^T ln(x)}{RT}}, \\ \Delta_r G' \ &= \ S^T \Delta_f G' \ + \ nF\psi \end{aligned} \tag{4.3}$$

In Equation 4.3, *reversibility* is the reversibility factor, $\Delta_r G'$ is the Gibbs free energy of the reaction, R is the gas constant, T is the temperature in Kelvin, S is the stoichiometric matrix, x is the vector of metabolite concentrations, $\Delta_f G'$ is the vector of metabolite Gibbs free energy of formation, n is the transported charge, F is the Faraday constant and ψ is the membrane potential.

$$\begin{aligned} allostery \ &= \ \frac{1}{1 \ + \ tc_r \ * \ (free\ enzyme\ ratio_r \ * \ \frac{Q_{tense}}{Q_{relaxed}})^{sb}}, \\ Q_{tense} \ &= \ 1 \ + \sum_{inhibitor} \frac{x_i}{dc_{r,i}}, \\ Q_{relaxed} \ &= \ 1 \ + \sum_{activator} \frac{x_a}{dc_{r,a}} \end{aligned} \tag{4.4}$$

In Equation 4.4 following Popova and Sel'kov [31], *allostery* is the allostery factor, tc_r is the transfer constant, $free\ enzyme\ ratio$ is the ratio of free enzyme to bound enzyme, Q_{tense} is the fraction of enzyme bound to the inhibitor, $Q_{relaxed}$ is the fraction of unbound (active) enzyme, sb is the number of subunits of the enzyme, x_i is the inhibitor concentration and x_a is the activator concentration.

4.2.2. Maximum a posteriori probability estimation

One of the functionalities of the Maud is to estimate the kinetic model parameters that achieve the maximum a posteriori probability (MAP), specifically finding the parameter configuration that maximises the log posterior density, taking into account both the prior distribution and likelihood function. The following settings were applied for the optimisation using the L-BFGS algorithm [32]: the number of iterations before termination set to 3,000, the step size of the first iteration set to $1 \cdot 10^{-4}$, the convergence tolerance set to $1 \cdot 10^{-12}$ and the number of update vectors used for Hessian approximation set to 10.

It is possible to obtain an approximation to the posterior distribution by finding the MAP and inspecting its gradients as described in Kass *et al.* [33] and MacKay [34]. While this approximation is less accurate compared with Markov chain Monte Carlo (MCMC) methods, used for sampling of posterior distributions in Maud, it is much less computationally expensive. In addition, when combined with well chosen priors, MAP estimation can produce better predictions than maximum likelihood estimation due to the regularising effect of the prior distribution [35].

In the current model specification it is possible to define a set of parameters that do not result in a steady state. This can occur with successive irreversible reactions when the rate of the first is higher than the V_{max} of the second. To avoid these parameter configurations, non-steady states were penalised using equation whereby the $\frac{dC}{dt}$ vector decreases the log-probability of the optimiser the further the state of solutions are from steady-state.

$$S \; * \; f_v(x_t, \theta) \;\sim\; N(0, \epsilon C) \tag{4.5}$$

In Equation 4.5, S is the stoichiometric matrix, x_t is the vector of metabolite concentrations, ϵ is the tolerance factor set to $1 \cdot 10^{-10}$ and C is the vector of balanced metabolite concentrations.

4.2.3. Metabolic control analysis

After optimisation of the kinetic model, the MAP estimate was used as the reference state for metabolic control analysis [36, 37, 38]. Maud includes a metabolic control analysis, in particular the calculation of flux response coefficients to changes in enzyme concentrations. The metabolic control analysis were applied to the current kinetic model with the following settings: the time point set to $1 \cdot 10^8$ seconds, the maximum number of steps set to $1 \cdot 10^6$ and both the relative and absolute tolerance set to $1 \cdot 10^{-9}$. For a simplified investigation of feedback-resistance of the DDPA *aroG* isozyme and the CHORM and PPND *tyrA* enzyme, the corresponding allosteric regulations were removed from the kinetic model and metabolic control analysis was repeated for the ePtsHIcrr03 experiment.

4.3. Results

Using the Maud kinetic modelling framework, a kinetic model of aromatic amino acid biosynthesis in *E. coli* was constructed. The optimisation functionality was deployed to estimate the MAP which best fit the experimental data set, while maintaining consistency with our curated prior model. The parameterisation and the simulation results including reaction fluxes and enzyme and metabolite concentrations can be found in Table S4.2 - S4.5. After evaluation of the kinetic model fit to the experimental data set, the kinetic model parameterisation was used to analyse the aromatic amino acid biosynthesis pathway through regulatory decomposition and metabolic control analysis.

4.3.1. Fit of the kinetic model parameterisation to experimental data

Overall, the simulated concentrations coincided with the experimentally measured concentrations of balanced metabolites (Figure 4.2). Apart from the shikimate concentration for the ePtsHIcrr03 experiment, all observed metabolite concentrations were within the bounds of the corresponding posterior predictive distributions. For both L-phenylalanine and L-tyrosine, the observed and simulated concentrations completely aligned, while the L-tryptophan observed concentrations were underestimated repeatedly by the simulated concentrations. The simulated reaction fluxes were in agreement with the experimentally measured fluxes (Figure S4.1). Simulated fluxes of the L-tryptophan-producing branch did display an underestimation of measured fluxes for all experiments, which conforms to the underestimated L-tryptophan concentrations. Overall, the fit to the data was satisfactory. In our judgement the issues mentioned above were due to the lack of experimental measurements for intermediate metabolite concentrations. For example, additional experimental measurements of 34HPP (3-(4-hydroxyphenyl)pyruvate) and ANTH (anthranilate) concentrations could elucidate the reaction fluxes through the corresponding branches more quickly during kinetic model fitting.

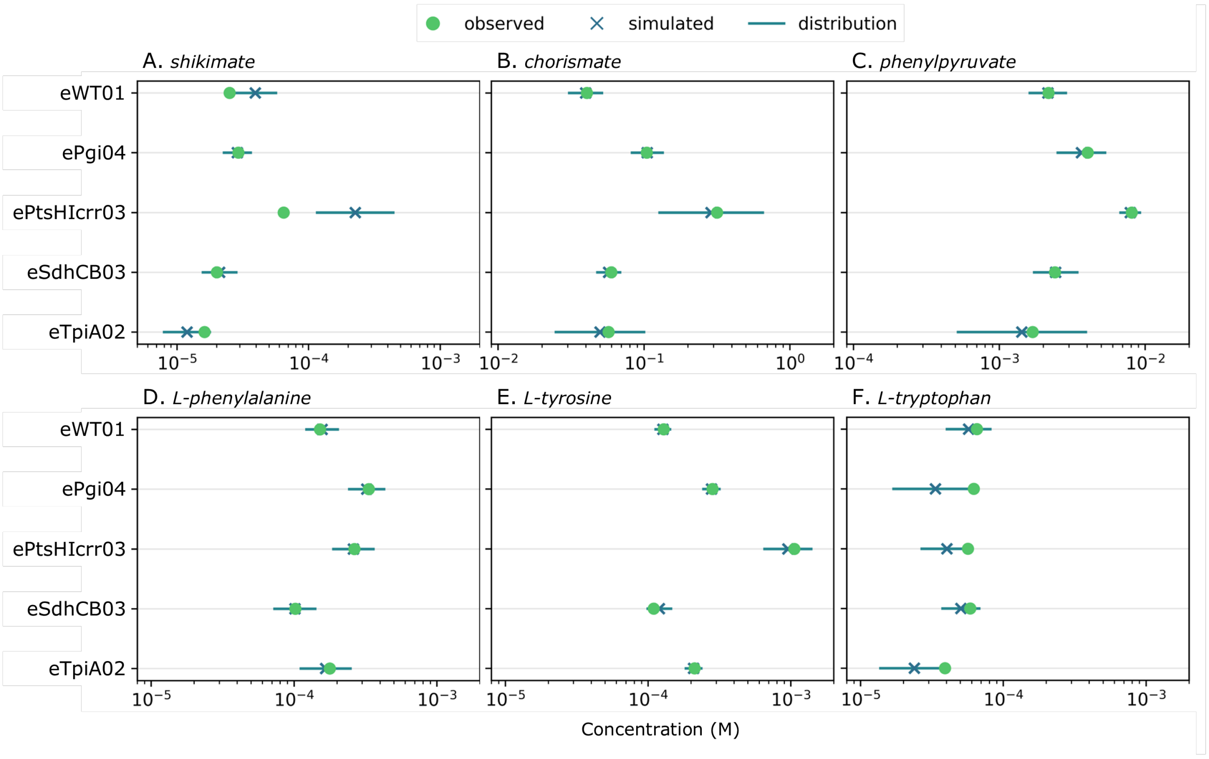


Figure 4.2: Comparison of experimental and simulated intracellular concentrations of experimentally measured, balanced metabolites. Green circles represent the experimentally observed concentrations, while blue crosses and lines represent the simulated concentration (mean) and the 95% quantile of the posterior predictive distribution, respectively. (A) Comparison of shikimate. (B) Comparison of chorismate. (C) Comparison of phenylpyruvate. (D) Comparison of L-phenylalanine. (E) Comparison of L-tyrosine. (F) Comparison of L-tryptophan.

4.3.2. Regulatory decomposition of the pathway

Maud allows for regulatory decomposition of the simulated reaction fluxes of the pathway. The enzyme concentration, k_{cat} value (rate constant of catalytic conversion), reversibility of the enzymatic reaction, enzyme saturation and allosteric effects combined determine the reaction flux (Equation 1).

Focusing on the production of L-tyrosine, the enzymes in the shikimate pathway leading up to chorismate display lower estimated saturation levels compared to the enzymes of the L-tyrosine-producing branch for the eWT01 experiment (Figure 4.3A). The intermediate metabolites of the L-tyrosine-producing branch (chorismate, prephenate and 3-(4-hydroxyphenyl)pyruvate) were present in relatively higher concentrations than the shikimate pathway intermediates (Table S4.4), likely causing this general shift in enzyme saturation. Only four enzymes in the L-tyrosine production pathway catalyse reversible reactions: DHQTi (3-dehydroquinate dehydratase), PSCVT (3-phosphoshikimate 1-carboxyvinyltransferase), TYRTA (L-tyrosine aminotransferase) *tyrB* and *aspC* isozymes. PSCVT and the TYRTA isozymes has estimated reversibility factors close to 1, thus were modelled as operating far from equilibrium. The L-phenylalanine-sensitive isozyme of DDPA (3-deoxy-7-phosphoheptulonate synthase; *aroG*) is known to carry approximately 80% of the DDPA reaction flux in wild type *E. coli* [39] and this was reflected in the high enzyme concentration (Table S4.3) and low allostery factor, which indicates high inhibition of the isozyme by L-phenylalanine. For the CHORM (chorismate mutase) *pheA* and *tyrA* isozymes, the *tyrA* isozyme appears to carry most of the CHORM reaction flux, where the higher catalytic efficiency and higher enzyme concentration (Table S4.3) outweighed the stronger inhibition by L-tyrosine of the *tyrA* compared to the *pheA* isozyme.

In contrast to the eWT01 experiment, the enzymes of the shikimate pathway produced higher saturation factors for the ePtsHIcrr03 experiment (Figure 4.3B), which was likely due to higher substrate and lower enzyme concentrations (Table S4.3 and S4.4). The increased PEP concentration led to higher concentrations of shikimate pathway intermediates, while redirection of protein resources led to lower enzyme concentrations for the ePtsHIcrr03 compared with the eWT01 strain [18, 21]. The higher concentrations of intermediate metabolites resulted in higher product concentrations for the reversible enzymatic reactions which led to generally lower reversibility factors. For the ePtsHIcrr03 experiment, the enzyme concentrations of the DDPA isozymes evened out (Table S4.3) and the L-tyrosine concentration was substantially higher than the L-phenylalanine concentration, which resulted in a lower allostery factor and stronger inhibition of the L-tyrosine-sensitive *aroF* isozyme. The allosteric inhibitory impact of the higher L-tyrosine concentration on the CHORM *tyrA* isozyme and PPND (prephenate dehydrogenase) enzyme was negated by the high enzyme saturation factors and thus low free enzyme ratios. Lower free enzyme ratios limits the allosteric impact (Equation 4.4) and resulted in the comparable allosteric factors of CHORM *tyrA* isozyme and PPND between the eWT01 and ePtsHIcrr03 experiments.

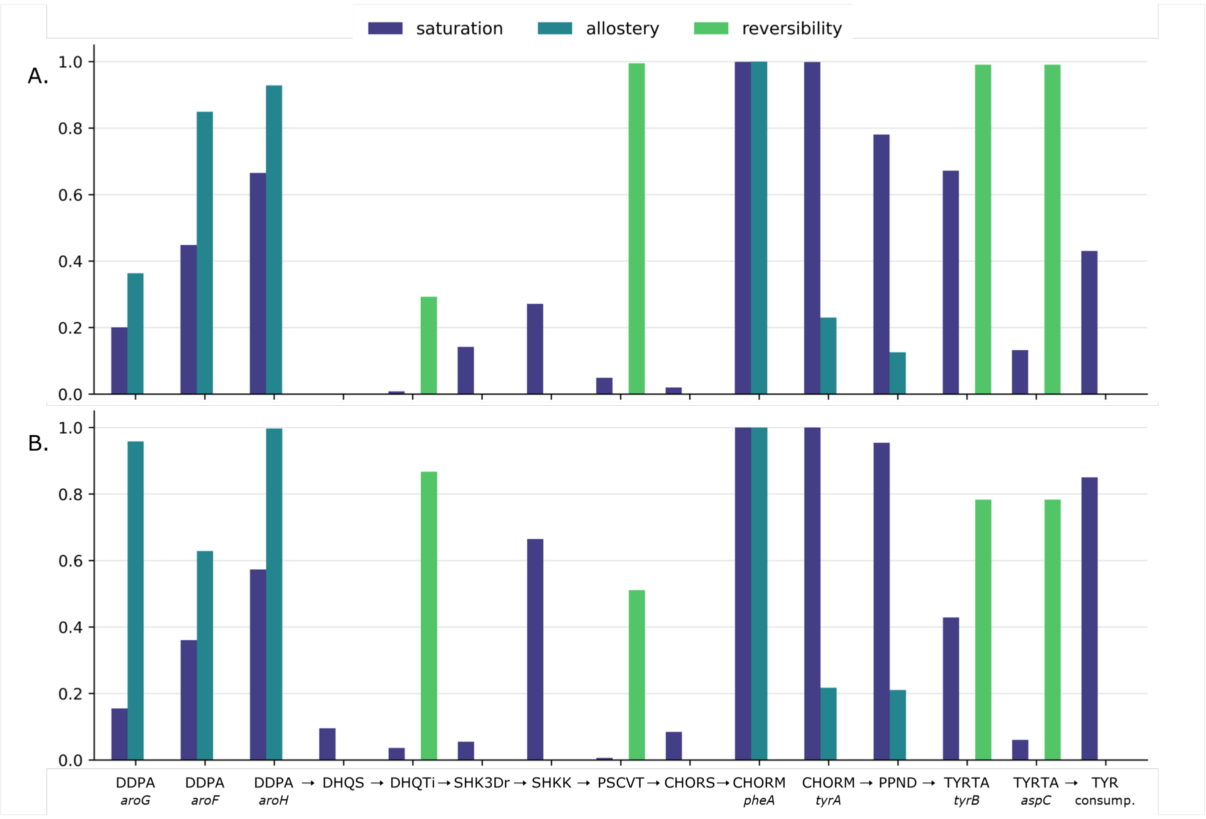


Figure 4.3: Regulatory decomposition of the enzymatic reactions leading towards L-tyrosine production. While a saturation value could be calculated for all enzymes, only the DHQTi (3-dehydroquinate dehydratase), PSCVT (3-phosphoshikimate 1-carboxyvinyltransferase) and TYRTA (tyrosine aminotransferase) reactions are reversible and only the DDPA (3-deoxy-7-phosphoheptulonate synthase) and CHORM (chorismate mutase) enzymes are regulated by allosteric inhibitions. (A) Regulatory decomposition for the eWT01 experiment. (B) Regulatory decomposition for the ePtsHIcrr03 experiment.

4.3.3. Metabolic control analysis of the pathway

Investigating the best engineering targets could be achieved by performing a metabolic control analysis on the allosteric parameters to determine the most sensitive enzymes. Using the previous estimated MAP, metabolic control analysis was performed to investigate the intricacies of the aromatic amino acid biosynthesis pathway. The resulting flux response factors to changes in enzyme concentrations was focused on the L-tyrosine-consuming reaction to explore beneficial alterations for L-tyrosine production (Figure 4.4). Generally and as expected, an increase in concentration of the shikimate pathway and L-tyrosine producing branch and a decrease in concentration of the other enzymes would positively impact the L-tyrosine-consuming reaction flux. The increase in enzyme concentrations would have the greatest impact for the ePtsHIcrr03 experiment, since the higher concentrations of intermediate metabolites (Table S4.4) in combination with the higher enzyme concentrations would lead to a substantial flux increase.

The flux response factor of changes in CHORM *pheA* isozyme and PPNDH (prephenate dehydrogenase; *pheA*) and in CHORM *tyrA* and PPND (*tyrA*) differed, since the bifunctionality of the *pheA* and *tyrA* isozymes could not be modelled within the framework. Nevertheless, the flux response factors agreed in absolute terms: a decrease in *pheA* and an increase in *tyrA* enzyme concentration would be advantageous towards L-tyrosine production.

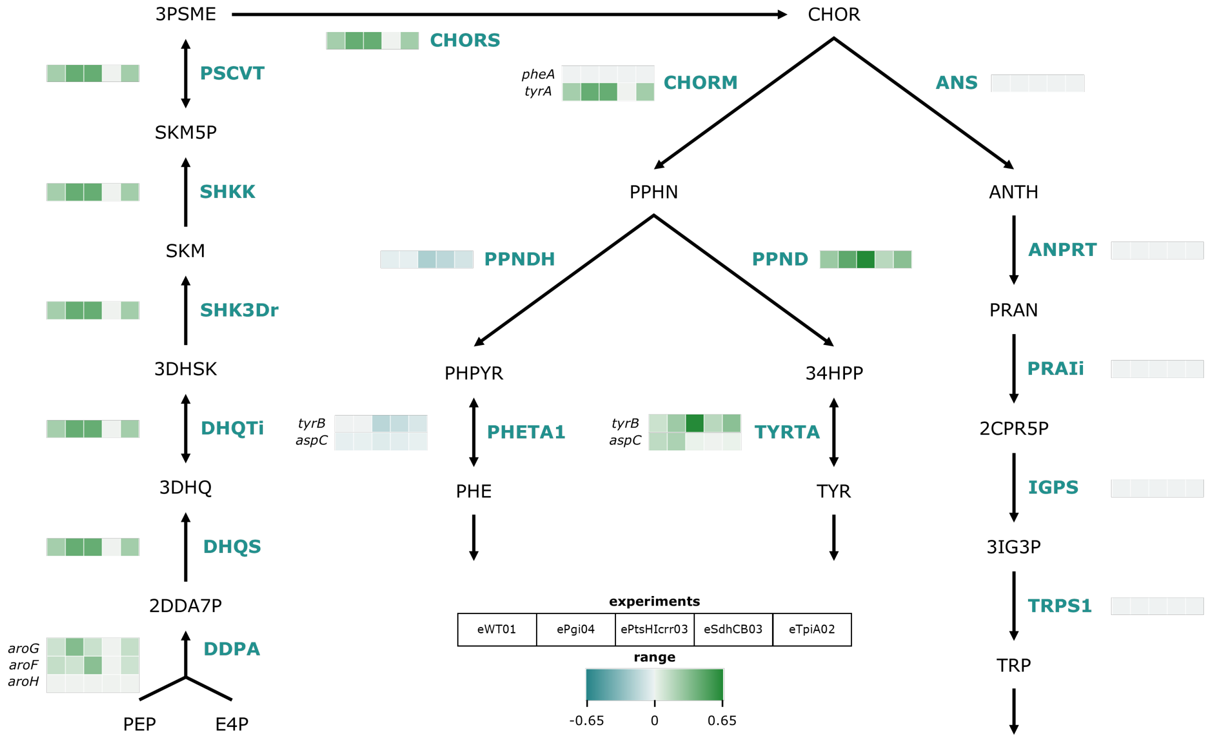


Figure 4.4: Metabolic control analysis of the aromatic amino acid biosynthesis where the L-tyrosine combined consumption reaction was set as target. The flux response coefficients of changes in enzyme concentrations in mol/L enzyme per (mol/L)/s were plotted for all five experiments for each enzyme.

Our training dataset consisted of native enzymes whereas in production strains the following metabolic engineering is usually performed: feedback-resistance of the DDPA *aroG* isozyme and the CHORM and PPND *tyrA* enzyme. We simulated these feedback-resistant enzymes in our model to investigate the impact on L-tyrosine production, here the L-tyrosine-consuming reaction, for the ePtsHIcrr03 experiment (Figure S4.2). Removing these allosteric inhibitions resulted in a substantial increase in flux through the shikimate pathway and L-tyrosine branch, assuming constant enzyme concentrations. Increased concentrations of the DDPA *aroG* isozyme were no longer required for increased flux of the L-tyrosine-consuming reaction when the allosteric inhibition acting on the enzyme was removed. This confirmed the strong impact of the allosteric inhibition by L-phenylalanine on the DDPA *aroG* isozyme, which was identified in the regulatory decomposition.

By removal of the allosteric inhibition by L-tyrosine on both the CHORM and PPND *tyrA* enzyme, an increase in enzyme concentration would result in a higher increase in flux of the L-tyrosine-consuming reaction. Both feedback-resistance and overexpression for higher enzyme concentrations of the *tyrA* enzyme would be required to maximise a beneficial metabolic engineering strategy based on the current metabolic control analysis. Additionally, the feedback-resistant versions of the DDPA *aroG* isozyme and the CHORM and PPND *tyrA* bifunctional enzyme were incorporated in the ePtsHIcrr03 strain to further optimise L-tyrosine production (see Epilogue).

4.4. Discussions

In this study, we developed a kinetic model of the *E. coli* aromatic amino acid biosynthesis using the Maud kinetic modelling framework. The resulting parameterisation after MAP estimation produced simulated metabolite concentrations and reaction fluxes which aligned with the experimental values. Even though Bayesian inference inherently limits overfitting to experimental data [35], validation using a comparison of simulated values to experimental values which the kinetic model was fitted to is not entirely appropriate. It would be more suitable for future kinetic model evaluations to fully leverage the benefits of Bayesian inference by sampling from the posterior distribution using MCMC methods. Afterwards, out-of-sample predictions could be used to properly evaluate the model through cross-validation. This approach was not included in the current study due to time limitations, since MCMC sampling is computationally expensive [40, 41, 42], especially in cases like Bayesian kinetic models that require solutions to large systems of non-linear differential equations.

Through regulatory decomposition and metabolic control analysis, the inner workings of the aromatic amino acid biosynthesis were examined with a focus on L-tyrosine production. This resulted in several insights on the allosteric regulations, where the DDPA *aroG* isozyme and the CHORM and PPND *tyrA* enzyme were mostly affected by allosteric inhibition through L-phenylalanine and L-tyrosine, respectively. However, L-phenylalanine concentrations varied little between experiments and L-tyrosine concentrations only differed substantially for the ePtsHIcrr03 experiment compared to the other experiments within the current experimental data set. The kinetic model fit and quality of the resulting parameterisation would benefit greatly from a more informative experimental data set, which includes more direct changes within the pathway, e.g. overexpression of pathway enzymes or incorporation of feedback-resistant versions of strongly inhibited enzymes. Larger differences between experimental values would contain more information about the pathway dynamics which could in turn be learned by the kinetic model and lead to an improved parameterisation representative of the pathway.

The main suggestions for the metabolic engineering of *E. coli* L-tyrosine overproducers resulting from the current analysis of the aromatic amino acid biosynthesis pathway were the overexpression of the shikimate pathway enzymes and the CHORM and PPND *tyrA* enzyme, and the integration of feedback-resistant versions of the DDPA *aroG* and CHORM and PPND *tyrA* enzymes.

The suggested approach aligns with the metabolic engineering strategy of Juminaga *et al.* [43], who constructed an L-tyrosine overproducing strain with L-tyrosine titres over 2 g/L, and with strategies of multiple other studies [44, 45] and of another kinetic modelling study [15]. Interestingly, metabolic control analysis results suggested that overexpression of the DDPA *aroG* isozyme, a common metabolic engineering approach, would no longer be required after addition of a feedback-resistant version, alleviating demand for protein resources which could be needed elsewhere in the pathway. With an improved kinetic model parameterisation from fitting to a more representative experimental data set, a metabolic engineering strategy balancing L-tyrosine overproduction with moderate L-phenylalanine and L-tryptophan intracellular concentrations could be tested and evaluated *in silico*. Such a strategy would be desirable, since moderate levels of L-phenylalanine and L-tryptophan are required to sustain prototrophic growth, which is occasionally eliminated by deletion of the CHORM and PPNDH *pheA* enzyme to redirect pathway intermediates to L-tyrosine production [46, 47]. Although the focus of the current study was on L-tyrosine overproduction, the overproduction of L-phenylalanine and L-tryptophan could also be investigated using the current aromatic amino acid kinetic model, since the specific pathways to all three aromatic amino acids were implemented in full detail including allosteric regulations.

Within iterative metabolic engineering of microbial cell factories, kinetic modelling provides a tool for multi-omics data integration to analyse a set of previous strain designs in order to inform the next set of strain designs. While kinetic models provide a more detailed and nuanced approach to *in silico* evaluation of strain designs compared with constraint-based modelling [5], kinetic models require an extensive number of input parameters and are more time consuming to develop [6]. The application of Bayesian inference to kinetic modelling introduces an increased flexibility of the input [48, 49], yet the scope is still limited to a set of relatively well-characterised microbes. Overall, kinetic modelling can present a highly comprehensive insight in a specific metabolic pathway to guide metabolic engineering efforts, which would be most appropriate for a targeted experimental objective with a specific product in mind.

4.5. Conclusions

Kinetic model development of *E. coli* aromatic amino acid biosynthesis provided a detailed insight into the intricacies of the pathway. Both the DDPA *aroG* isozyme and the CHORM and PPND *tyrA* enzyme were impacted heavily through allosteric inhibition by L-phenylalanine and L-tyrosine, respectively. Feedback-resistant versions of both enzymes, together with overexpression of most enzymes in the pathway towards L-tyrosine, was proposed as a potential metabolic engineering strategy for L-tyrosine overproduction. The aromatic amino acid biosynthesis model will, especially with further benefit from Bayesian inference methods, provide an advantageous tool for metabolic engineering of *E. coli* aromatic amino acid overproducers.

References

[1] N. J. H. Averesch and J. O. Krömer. "Metabolic Engineering of the Shikimate Pathway for Production of Aromatics and Derived Compounds — Present and Future Strain Construction Strategies". In: *Frontiers in Bioengineering and Biotechnology* 6 (2018), p. 32. DOI: 10.3389/fbioe.2018.00032.

[2] M. Cao, M. Gao, M. Suástegui, Y. Mei, and Z. Shao. "Building microbial factories for the production of aromatic amino acid pathway derivatives: From commodity chemicals to plant-sourced natural products". In: *Metabolic Engineering* 58 (2020), pp. 94–132. DOI: 10.1016/j.ymben.2019.08.008.

[3] M. I. Chávez-Béjar, J. L. Báez-Viveros, A. Martínez, F. Bolívar, and G. Gosset. "Biotechnological production of l-tyrosine and derived compounds". In: *Process Biochemistry* 47.7 (2012), pp. 1017–1026. DOI: 10.1016/j.procbio.2012.04.005.

[4] Z. Li, H. Wang, D. Ding, Y. Liu, H. Fang, Z. Chang, T. Chen, and D. Zhang. "Metabolic engineering of Escherichia coli for production of chemicals derived from the shikimate pathway". In: *Journal of Industrial Microbiology and Biotechnology* 47.6–7 (2020), pp. 525–535. DOI: 10.1007/s10295-020-02288-2.

[5] R. P. van Rosmalen, R. W. Smith, V. A. P. Martins dos Santos, C. Fleck, and M. Suarez-Diez. "Model reduction of genome-scale metabolic models as a basis for targeted kinetic models". In: *Metabolic Engineering* 64 (2021), pp. 74–84. DOI: 10.1016/j.ymben.2021.01.008.

[6] P. A. Saa and L. K. Nielsen. "Formulation, construction and analysis of kinetic models of metabolism: A review of modelling frameworks". In: *Biotechnology Advances* 35.8 (2017), pp. 981–1003. DOI: 10.1016/j.biotechadv.2017.09.005.

[7] D. Shepelin, D. Machado, L. K. Nielsen, and M. J. Herrgård. "Benchmarking kinetic models of Escherichia coli metabolism". In: *bioRxiv* (2020). DOI: 10.1101/2020.01.16.908921.

[8] C. Chassagnole, N. Noisommit-Rizzi, J. W. Schmid, K. Mauch, and M. Reuss. "Dynamic modeling of the central carbon metabolism of Escherichia coli". In: *Biotechnology and Bioengineering* 79.1 (2002), pp. 53–73. DOI: 10.1002/bit.10288.

[9] H. Link, K. Kochanowski, and U. Sauer. "Systematic identification of allosteric protein-metabolite interactions that control enzyme activity in vivo". In: *Nature Biotechnology* 31.4 (2013), pp. 357–361. DOI: 10.1038/nbt.2489.

[10] D. Christodoulou, H. Link, T. Fuhrer, K. Kochanowski, L. Gerosa, and U. Sauer. "Reserve Flux Capacity in the Pentose Phosphate Pathway Enables Escherichia coli's Rapid Response to Oxidative Stress". In: *Cell Systems* 6.5 (2018), pp. 569–578. DOI: 10.1016/j.cels.2018.04.009.

[11] H. Kurata and Y. Sugimoto. "Improved kinetic model of Escherichia coli central carbon metabolism in batch and continuous cultures". In: *Journal of Bioscience and Bioengineering* 125.2 (2018), pp. 251–257. DOI: 10.1016/j.jbiosc.2017.09.005.

[12] A. Khodayari and C. D. Maranas. "A genome-scale Escherichia coli kinetic metabolic model k-ecoli457 satisfying flux data for multiple mutant strains". In: *Nature Communications* 7.1 (2016), p. 13806. DOI: 10.1038/ncomms13806.

[13] P. Millard, K. Smallbone, and P. Mendes. “Metabolic regulation is sufficient for global and robust coordination of glucose uptake, catabolism, energy production and growth in Escherichia coli”. In: *PLoS Computational Biology* 13.2 (2017), e1005396. DOI: `10.1371/journal.pcbi.1005396`.

[14] C. J. Foster, S. Gopalakrishnan, M. R. Antoniewicz, and C. D. Maranas. “From Escherichia coli mutant 13C labeling data to a core kinetic model: A kinetic model parameterization pipeline”. In: *PLoS Computational Biology* 15.9 (2019), e1007319. DOI: `10.1371/journal.pcbi.1007319`.

[15] A. Fonseca and I. Rocha. “Strain optimization for aromatic amino acids using an Escherichia coli kinetic model”. In: *IFAC-PapersOnLine* 55.7 (2022), pp. 691–696. DOI: `10.1016/j.ifacol.2022.07.524`.

[16] D. McCloskey, S. Xu, T. E. Sandberg, E. Brunk, Y. Hefner, R. Szubin, A. M. Feist, and B. O. Palsson. “Evolution of gene knockout strains of E. coli reveal regulatory architectures governed by metabolism”. In: *Nature Communications* 9.1 (2018), p. 3796. DOI: `10.1038/s41467-018-06219-9`.

[17] D. McCloskey, S. Xu, T. E. Sandberg, E. Brunk, Y. Hefner, R. Szubin, A. M. Feist, and B. O. Palsson. “Multiple Optimal Phenotypes Overcome Redox and Glycolytic Intermediate Metabolite Imbalances in Escherichia coli pgi Knockout Evolutions”. In: *Applied and Environmental Microbiology* 84.19 (2018), e00823–18. DOI: `10.1128/AEM.00823-18`.

[18] D. McCloskey, S. Xu, T. E. Sandberg, E. Brunk, Y. Hefner, R. Szubin, A. M. Feist, and B. O. Palsson. “Adaptive laboratory evolution resolves energy depletion to maintain high aromatic metabolite phenotypes in Escherichia coli strains lacking the Phosphotransferase System”. In: *Metabolic Engineering* 48 (2018), pp. 233–242. DOI: `10.1016/j.ymben.2018.06.005`.

[19] D. McCloskey, S. Xu, T. E. Sandberg, E. Brunk, Y. Hefner, R. Szubin, A. M. Feist, and B. O. Palsson. “Growth Adaptation of gnd and sdhCB Escherichia coli Deletion Strains Diverges From a Similar Initial Perturbation of the Transcriptome”. In: *Frontiers in Microbiology* 9 (2018), p. 1793. DOI: `10.3389/fmicb.2018.01793`.

[20] D. McCloskey, S. Xu, T. E. Sandberg, E. Brunk, Y. Hefner, R. Szubin, A. M. Feist, and B. O. Palsson. “Adaptation to the coupling of glycolysis to toxic methylglyoxal production in tpiA deletion strains of Escherichia coli requires synchronized and counterintuitive genetic changes”. In: *Metabolic Engineering* 48 (2018), pp. 82–93. DOI: `10.1016/j.ymben.2018.05.012`.

[21] D. Heckmann, A. Campeau, C. J. Lloyd, P. V. Phaneuf, Y. Hefner, M. Carrillo-Terrazas, A. M. Feist, D. J. Gonzalez, and B. O. Palsson. “Kinetic profiling of metabolic specialists demonstrates stability and consistency of in vivo enzyme turnover numbers”. In: *PNAS* 117.37 (2020), pp. 23182–23190. DOI: `10.1073/pnas.200156211`.

[22] I. M. Keseler, S. Gama-Castro, A. Mackie, R. Billington, C. Bonavides-Martínez, R. Caspi, A. Kothari, M. Krummenacker, P. E. Midford, L Muñiz-Rascado, W K Ong, S Paley, A Santos-Zavaleta, P Subhraveti, V. H. Tierrafría, A. J. Wolfe, J. Collado-Vides, I. T. Paulsen, and P. D. Karp. “The EcoCyc Database in 2021”.

In: *Frontiers in Microbiology* 12 (2021), p. 711077. DOI: 10.3389/fmicb.2021.711077.

[23] A. Chang, L. Jeske, S. Ulbrich, J. Hofmann, J. Koblitz, I. Schomburg, M. Neumann-Schaal, D. Jahn, and D. Schomburg. "BRENDA, the ELIXIR core data resource in 2021: new developments and updates". In: *Nucleic Acids Research* 49.D1 (2020), pp. D498–D508. DOI: 10.1093/nar/gkaa1025.

[24] U. Wittig, M. Rey, A. Weidemann, R. Kania, and W. Müller. "SABIO-RK: an updated resource for manually curated biochemical reaction kinetics". In: *Nucleic Acids Research* 46.D1 (2017), pp. D656–D660. DOI: 10.1093/nar/gkx1065.

[25] M. Mori, Z. Zhang, A. Banaei-Esfahani, J. B. Lalanne, H. Okano, B. C. Collins, A. Schmidt, O. T. Schubert, D. S. Lee, G. W. Li, R. Aebersold, T. Hwa, and C. Ludwig. "From coarse to fine: The absolute *Escherichia coli* proteome under diverse growth conditions". In: *Molecular Systems Biology* 17.5 (2021), e9536. DOI: 10.15252/msb.20209536.

[26] J. R. Wiśniewski, P. Ostasiewicz, K. Duś, D. F. Zielińska, F. Gnad, and M. Mann. "Extensive quantitative remodeling of the proteome between normal colon tissue and adenocarcinoma". In: *Molecular Systems Biology* 8.1 (2012), p. 611. DOI: 10.1038/msb.2012.44.

[27] J. R. Wiśniewski, M. Y. Hein, J. Cox, and M. Mann. "A "Proteomic Ruler" for Protein Copy Number and Concentration Estimation without Spike-in Standards". In: *Molecular and Cellular Proteomics* 13.12 (2014), pp. 3497–3506. DOI: 10.1074/mcp.M113.037309.

[28] W. Liebermeister, J. Uhlendorf, and E. Klipp. "Modular rate laws for enzymatic reactions: thermodynamics, elasticities and implementation". In: *Bioinformatics* 26.12 (2010), pp. 1528–1534. DOI: 10.1093/bioinformatics/btq141.

[29] A. Flamholz, E. Noor, A. Bar-Even, and R. Milo. "eQuilibrator — the biochemical thermodynamics calculator". In: *Nucleic Acids Research* 40.D1 (2011), pp. D770–D775. DOI: 10.1093/nar/gkr874.

[30] M. E. Beber, M. G Gollub, D. Mozaffari, K. M. Shebek, A. I. Flamholz, R. Milo, and E. Noor. "eQuilibrator 3.0: a database solution for thermodynamic constant estimation". In: *Nucleic Acids Research* 50.D1 (2021), pp. D603–D609. DOI: 10.1093/nar/gkab1106.

[31] S. V. Popova and E. E. Sel'kov. "Generalization of the model by Monod, Wyman and Changeux for the case of a reversible monosubstrate reaction". In: *FEBS Letters* 53.3 (1975), pp. 269–273. DOI: 10.1016/0014-5793(75)80034-2.

[32] J. Nocedal and S. J. Wright. *Numerical optimization.* Spinger, 2006. ISBN: 9780387227429. DOI: 10.1007/b98874.

[33] R. E. Kass, L. Tierney, and J. B. Kadane. "Laplace's method in Bayesian analysis". In: *Statistical Multiple Integration.* Vol. 115. Contemporary Mathematics, 1991, pp. 89–99. ISBN: 9780821877036. DOI: 10.1090/conm/115.

[34] D. J. MacKay. "Laplace's method". In: *Information theory, inference and learning algorithms.* Cambridge university press, 2003, pp. 341–342.

[35] D. Cousineau and S. Helie. "Improving maximum likelihood estimation using prior probabilities: A tutorial on maximum a posteriori estimation and an examination of the weibull distribution". In: *Tutorials in Quantitative Methods for Psychology* 9.2 (2013), pp. 61–71. DOI: 10.20982/tqmp.09.2.p061.

[36] H. Kacser and J. A. Burns. "The control of flux". In: *Symposia of the Society for Experimental Biology* 27 (1973), pp. 65–104.

[37] R. Heinrich and T. A. Rapoport. "A linear steady-state treatment of enzymatic chains. General properties, control and effector strength". In: *European Journal of Biochemistry* 42.1 (1974), pp. 89–95. DOI: 10.1111/j.1432-1033.1974.tb03318.x.

[38] R. Steuer and B. H. Junker. "Advances in Chemical Physics". In: John Wiley & Sons, Ltd, 2009. Chap. Computational Models of Metabolism: Stability and Regulation in Metabolic Networks, pp. 105–251. ISBN: 9780470475935. DOI: 10.1002/9780470475935.ch3.

[39] D. E. Tribe, H. Camakaris, and J. Pittard. "Constitutive and Repressible Enzymes of the Common Pathway of Aromatic Biosynthesis in Escherichia coli K-12: Regulation of Enzyme Synthesis at Different Growth Rates". In: *Journal of Bacteriology* 127.3 (1976), pp. 1085–1097. DOI: 10.1128/jb.127.3.1085-1097.1976.

[40] B. Sengupta, K. J. Friston, and W. D. Penny. "Gradient-based MCMC samplers for dynamic causal modelling". In: *NeuroImage* 125 (2016), pp. 1107–1118. DOI: 10.1016/j.neuroimage.2015.07.043.

[41] G. I. Valderrama-Bahamóndez and H. Fröhlich. "MCMC techniques for parameter estimation of ODE based models in systems biology". In: *Frontiers in Applied Mathematics and Statistics* 5 (2019), p. 55. DOI: 10.3389/fams.2019.00055.

[42] P. Tsiros, F. Y. Bois, A. Dokoumetzidis, G. Tsiliki, and H. Sarimveis. "Population pharmacokinetic reanalysis of a Diazepam PBPK model: a comparison of Stan and GNU MCSim". In: *Journal of Pharmacokinetics and Pharmacodynamics* 46.2 (2019), pp. 173–192. DOI: 10.1007/s10928-019-09630-x.

[43] D. Juminaga, E. E. K. Baidoo, A. M. Redding-Johanson, T. S. Batth, H. Burd, A. Mukhopadhyay, C. J. Petzold, and J. D. Keasling. "Modular Engineering of l-Tyrosine Production in Escherichia coli". In: *Applied and Environmental Microbiology* 78.1 (2012), pp. 89–98. DOI: 10.1128/AEM.06017-11.

[44] T. Lütke-Eversloh and G. Stephanopoulos. "L-Tyrosine production by deregulated strains of Escherichia coli". In: *Applied Microbiology and Biotechnology* 75.1 (2007), pp. 103–110. DOI: 10.1007/s00253-006-0792-9.

[45] B. Kim, R. Binkley, H. U. Kim, and S. Y. Lee. "Metabolic engineering of Escherichia coli for the enhanced production of L-tyrosine". In: *Biotechnology and Bioengineering* 115.10 (2018), pp. 2554–2564. DOI: 10.1002/bit.26797.

[46] C. N. S. Santos and G. Stephanopoulos. "Melanin-Based High-Throughput Screen for l-Tyrosine Production in Escherichia coli". In: *Applied and Environmental Microbiology* 74.4 (2008), pp. 1190–1197. DOI: 10.1128/AEM.02448-07.

[47] P. Singh, T. S. Batth, D. Juminaga, R. H. Dahl, J. D. Keasling, P. D. Adams, and C. J. Petzold. "Application of targeted proteomics to metabolically engineered Escherichia coli". In: *Proteomics* 12.8 (2012), pp. 1289–1299. DOI: `10.1002/pmic.201100482`.

[48] P. A. Saa and L. K. Nielsen. "Construction of feasible and accurate kinetic models of metabolism: A Bayesian approach". In: *Scientific Reports* 6.1 (2016), p. 29635. DOI: `10.1038/srep29635`.

[49] P. C. St. John, J. Strutz, L. J. Broadbelt, Tyo K. E. J., and Bomble Y. J. "Bayesian inference of metabolic kinetics from genome-scale multiomics data". In: *PLoS Computational Biology* 15.11 (2019), e1007424. DOI: `10.1371/journal.pcbi.1007424`.

Supplementary materials

Kinetic model definition

The following supplement provides an overview of kinetic parameters and additional information for all enzymatic reactions in the aromatic amino acid biosynthesis of *Escherichia coli*. The kinetic information presented here is based on literature research including the EcoCyc [1], BRENDA [2] and Sabio-RK [3] databases. All kinetic parameters used as input for the kinetic model are indicated in green.

DDPA - 3-deoxy-7-phosphoheptulonate synthase

The DDPA (3-deoxy-7-phosphoheptulonate synthase) reaction consists of three isozymes: L-phenylalanine-sensitive DDPA (P0AB91) encoded by the *aroG* gene L-tyrosine-sensitive DDPA (P00888) encoded by the *aroF* gene, and L-tryptophan-sensitive DDPA (P00887) encoded by the *aroH* gene. DDPA Phe is an homotetrameric enzyme, DDPA Tyr and Trp are an homodimeric enzyme, and all isozymes are located in the cytosol. There is a high degree of sequence identity (41%) between the three isozymes and the polypeptides are nearly identical in size. DDPA Phe makes up about 80% of the total DDPA activity, DDPA Tyr makes up about 20%, and DDPA Trp makes up about 1% [4]. All three isozymes are metalloenzymes and require a divalent metal for catalysis and/or structural integrity: DDPA Tyr can use several metal ions, DDPA Phe uses Fe^{2+} mostly, and DDPA Trp is activated by Fe^{2+}. Certain mutations in all three isozymes lead to insensitivity towards the corresponding aromatic amino acid. The role of metal ion in DDPA is to position the amino acids with the appropriate geometry required to coordinate and activate the water molecule; the rate constant varies with the bound metal ion [5].

Reaction equation

$$1\ PEP + 1\ E4P + 1\ H_2O \rightarrow 1\ 2DDA7P + 1\ Pi$$

phosphoenolpyruvate + erythrose 4-phosphate + water →
2-dehydro-3-deoxyarabino-heptulosonate 7-phosphate + phosphate

Kinetic parameters - L-phenylalanine-sensitive DDPA

Allosteric inhibition of L-phenylalanine-sensitive DDPA by L-phenylalanine [6] and inhibition of L-phenylalanine-sensitive DDPA by L-alanine and L-dihydrophenylalanine [6]. Competitive inhibition of L-phenylalanine-sensitive DDPA by 2,3-bisphosphoglycerate, 2-phosphoglycerate (2PG; glycerate 2-phosphate), 3-methylphosphoenolpyruvate, and 3-propylphosphoenolpyruvate [7].

Parameter	Substrate	Value	Unit	Reference
k_{cat}	-	32.0	s^{-1}	[8]
k_{cat}	-	62.3	s^{-1}	[9]
k_{cat}	-	71.0	s^{-1}	[10]
K_M	PEP	0.08	mM	[7]
K_M	PEP	0.009	mM	[8]
K_M	PEP	0.035	mM	[9]
K_M	E4P	0.9	mM	[7]
K_M	E4P	0.086	mM	[8]
K_M	E4P	0.25	mM	[9]
K_M	E4P	0.021	mM	[10]
K_D	PHE	0.013	mM	[6]
K_I	2PG	1.0	mM	[7]

Kinetic parameters - L-tyrosine-sensitive DDPA
Allosteric (potential) inhibition of L-tyrosine-sensitive DDPA by L-tyrosine, noncompetitive inhibition of L-tyrosine-sensitive DDPA by phosphate and competitive for PEP (product) inhibition of L-tyrosine-sensitive DDPA by 2DDA7P (2-dehydro-3-deoxy-arabino-heptulosonate 7-phosphate) [11].

Parameter	Substrate	Value	Unit	Reference
k_{cat}	-	60.3	s^{-1}	[11]
k_{cat}	-	29.5	s^{-1}	[12]
K_M	PEP	0.00585	mM	[11]
K_M	PEP	0.013	mM	[12]
K_M	E4P	0.0965	mM	[11]
K_M	E4P	0.0814	mM	[12]
K_D	TYR	0.009	mM	[12]
K_D	TYR	0.082	mM	[6]

Kinetic parameters - L-tryptophan-sensitive DDPA
(Noncompetitive, possibly allosteric) inhibition of L-tryptophan-sensitive DDPA by L-tryptophan [13, 14] and L-tryptophan-sensitive DDPA likely follows non-Michaelis-Menten kinetics [15].

Parameter	Substrate	Value	Unit	Reference
k_{cat}	-	20.6	s^{-1}	[15]
K_M	PEP	0.0053	mM	[15]
K_M	E4P	0.076	mM	[14]
K_M	E4P	0.035	mM	[15]
K_D	TRP	0.0014	mM	[15]

DHQS - 3-dehydroquinate synthase

DHQS (3-dehydroquinate synthase; P07639) is a monomeric enzyme encoded by the *aroB* gene and is located in the cytosol. DHQS requires multiple cofactors: NAD^+ and Zn^{2+} or Co^{2+}. Co^{2+} as cofactor results in higher specific activity, Zn^{2+} is more readily available in nature.

Reaction equation

$$1\ 2DDA7P \rightarrow 1\ 3DHQ + 1\ Pi$$

2-dehydro-3-deoxyarabino-heptulosonate 7-phosphate → 3-dehydroquinate + phosphate

Kinetic parameters
Competitive inhibition by 2DDA7P (2-dehydro-3-deoxyarabino-heptulosonate 7-phosphate; substrate inhibition) and variants of 2DDA7P [16].

Parameter	Substrate	Value	Unit	Reference
k_{cat}	-	72.9	s^{-1}	[17]
k_{cat}	-	16.0	s^{-1}	[18]
K_M	2DDA7P	0.018	mM	[16]
K_M	2DDA7P	0.033	mM	[19]
K_M	2DDA7P	0.0055	mM	[17]
K_M	2DDA7P	0.0057	mM	[18]

DHQTi - 3-dehydroquinate dehydratase

DQDH (3-dehydroquinate dehydratase; P05194) is an homodimeric enzyme encoded by the *aroD* gene and is located in the cytosol.

Reaction equation

$$1\ 3DHQ \rightarrow 1\ 3DHSK + 1\ H_2O$$

3-dehydroquinate → 3-dehydroshikimate + water

Kinetic parameters
Competitive inhibition by acetate, succinate, tartrate, and chloride; inhibition by diethylpyrocarbonate, and sodium borohydride [20].

Parameter	**Substrate**	**Value**	**Unit**	**Reference**
k_{cat}	-	135.0	s^{-1}	[21]
k_{cat}	-	142.0	s^{-1}	[22]
k_{cat}	-	29.53	s^{-1}	[23]
K_M	3DHQ	0.018	mM	[20]
K_M	3DHQ	0.073	mM	[20]
K_M	3DHQ	0.44	mM	[24]
K_M	3DHQ	0.016	mM	[21]
K_M	3DHQ	0.017	mM	[22]
K_M	3DHQ	0.188	mM	[23]
K_I	acetate	102	mM	[20]
K_I	succinate	74	mM	[20]
K_I	chloride	17	mM	[20]
K_I	tartrate	21	mM	[20]

SHK3Dr - shikimate dehydrogenase

SHK3Dr (shikimate dehydrogenase; P15770) is a monomeric enzyme encoded by the *aroE* gene and is located in the cytosol. Another SHK3Dr homodimeric enzyme is encoded by the *ydiB*. The SHK3Dr from *aroE* is NADP$^+$-specific and has much higher catalytic efficiency than the SHK3Dr from *ydiB*, which has broader substrate specificity and can use either NADP$^+$ or NAD$^+$ as a co-substrate [25]. Mutant data and the results from metabolic engineering experiments strongly suggest that the SHK3Dr from *ydiB* is unable to replace the SHK3Dr from *aroE* under normal physiological conditions. An Δ*aroE* mutant is viable, but accumulates 3-dehydroshikimate in minimal medium and is able to grow on media supplemented with shikimate, L-phenylalanine, and L-tyrosine. In the "reverse" direction, AroE appears to be able to further dehydrogenate 3-dehydroshikimate to 3,5-dehydroshikimate, which can spontaneously convert to gallic acid [26]. The physiological relevance is unclear.

Reaction equation

$$1\ 3DHSK + 1\ NADPH \rightarrow 1\ SKM + 1\ NADP^+$$

$$\text{3-dehydroshikimate} + \text{NADPH} \rightarrow \text{shikimate} + \text{NADP}^+$$

Kinetic parameters

(Linear) mixed inhibition by shikimate [27].

Parameter	Substrate	Value	Unit	Reference
k_{cat}	-	236.7	s^{-1}	[25]
k_{cat}	-	8750.0	s^{-1}	[26]
k_{cat}	-	190.0	s^{-1}	[28]
k_{cat}	-	178.0	s^{-1}	[29]
K_M	3DHSK	0.11	mM	[30]
K_M	3DHSK	0.11	mM	[29]
K_M	SKM	0.065	mM	[25]
K_M	SKM	0.102	mM	[26]
K_M	SKM	0.13	mM	[28]
K_M	SKM	0.095	mM	[29]
K_M	$NADP^+$	0.056	mM	[25]
K_M	$NADP^+$	0.347	mM	[26]
K_M	$NADP^+$	0.058	mM	[28]
K_M	$NADP^+$	0.011	mM	[29]
K_M	NADPH	0.0126	mM	[29]
K_I	SKM	0.16	mM	[27]

SHKK - shikimate kinase

The SHKK (shikimate kinase) reaction consists of two isozymes: SHKK I (P0A6D7) encoded by *aroK* and SHKK II (P0A6E1) encoded by *aroL*. Both enzymes are present in the cytosol, however SHKK I has approximately 100-fold lower affinity for shikimate [31] and a three-fold lower specific activity than SHKK II [32], so SHKK II is the major isozyme. SHKK I is relatively easy to isolate and is expressed constitutively compared to aromatic amino acid biosynthesis, so possibly other, still unknown, function. SHKK I and II are both monomeric enzymes which require Mg^{2+} as a cofactor. A double mutant Δ*aroK*Δ*aroL* cannot grow on minimal medium, even when supplemented with shikimate, L-phenylalanine and L-tyrosine.

Reaction equation

$$1\ SKM + 1ATP \rightarrow 1\ SKM5P + 1\ ADP$$

$$\text{shikimate} + \text{ATP} \rightarrow \text{shikimate 5-phosphate} + \text{ADP}$$

Kinetic parameters - SHKK I

Parameter	Substrate	Value	Unit	Reference
k_{cat}	-	6.0	s^{-1}	[32]
K_M	SKM	20	mM	[31]
K_M	SKM	0.4	mM	[33]

Kinetic parameters - SHKK II
Substrate inhibition of SHKK II by shikimate [31].

Parameter	Substrate	Value	Unit	Reference
k_{cat}	-	21.8	s^{-1}	[32]
k_{cat}	-	32.0	s^{-1}	[31]
K_M	SKM	0.2	mM	[31]
K_M	ATP	0.16	mM	[31]

PSCVT - 3-phosphoshikimate 1-carboxyvinyltransferase

PSCVT (3-phosphoshikimate 1-carboxyvinyltransferase or EPSP synthase; P0A6D3) is a monomeric enzyme encoded by the *aroA* gene and is located in the cytosol. A Δ*aroA* mutants are auxotrophic for the aromatic amino acids and are unable to grow on minimal media.

Reaction equation

$$1\ SKM5P + 1\ PEP \leftrightarrow 1\ 3PSME + 1\ Pi$$

shikimate 5-phosphate + phosphoenolpyruvate ↔
5-O-(1-Carboxyvinyl)-3-phosphoshikimate + phosphate

Kinetic parameters

Activation by PEP (phosphoenolpyruvate) [34]. Competitive inhibition of PEP by glyphosate [35], inhibition by pyruvate [36] and 3-bromopyruvate [37] and product inhibition by 5-O-(1-Carboxyvinyl)-3-phosphoshikimate (3PSME or EPSP) [35].

Parameter	Substrate	Value	Unit	Reference
k_{cat}	-	56.7	s^{-1}	[34]
k_{cat}	-	17.2	s^{-1}	[35]
k_{cat}	-	41.2	s^{-1}	[38]
k_{cat}	-	50.6	s^{-1}	[39]
k_{cat}	-	8.2	s^{-1}	[40]
k_{cat}	-	26.4	s^{-1}	[41]
k_{cat}	-	14.0	s^{-1}	[42]
k_{cat}	-	40.8	s^{-1}	[43]
k_{cat}	-	30.5	s^{-1}	[44]
k_{cat}	-	46.6	s^{-1}	[45]
k_{cat}	-	43.3	s^{-1}	[46]
K_M	SKM5P	0.0032	mM	[34]
K_M	SKM5P	0.0025	mM	[35]
K_M	SKM5P	0.020	mM	[38]
K_M	SKM5P	0.0025	mM	[39]
K_M	SKM5P	0.135	mM	[40]
K_M	SKM5P	0.14	mM	[41]
K_M	SKM5P	0.060	mM	[43]
K_M	SKM5P	0.008	mM	[47]
K_M	SKM5P	0.12	mM	[44]
K_M	SKM5P	0.048	mM	[45]
K_M	SKM5P	0.09	mM	[46]
K_M	PEP	0.021	mM	[34]
K_M	PEP	0.016	mM	[35]
K_M	PEP	0.025	mM	[38]
K_M	PEP	0.022	mM	[48]
K_M	PEP	22.5	mM	[49]
K_M	PEP	0.016	mM	[39]
K_M	PEP	0.1	mM	[40]
K_M	PEP	0.1	mM	[41]
K_M	PEP	0.16	mM	[41]
K_M	PEP	0.060	mM	[43]
K_M	PEP	0.013	mM	[47]
K_M	PEP	0.088	mM	[44]
K_M	PEP	0.045	mM	[45]
K_M	PEP	0.10	mM	[46]
K_M	3PSME	0.003	mM	[35]
K_M	3PSME	0.003	mM	[39]
K_M	3PSME	0.011	mM	[40]
K_M	3PSME	0.010	mM	[47]

Parameter	Substrate	Value	Unit	Reference
K_M	Pi	2.5	mM	[35]
K_M	Pi	2.5	mM	[39]
K_M	Pi	4.6	mM	[40]
K_M	Pi	5	mM	[47]
K_I	glyphosate	0.009	mM	[35]
K_I	glyphosate	0.0015	mM	[48]
K_I	glyphosate	0.96	mM	[49]
K_I	glyphosate	0.009	mM	[39]
K_I	glyphosate	0.0012	mM	[40]
K_I	glyphosate	0.0004	mM	[43]
K_I	glyphosate	0.00013	mM	[47]
K_I	glyphosate	80.0003	mM	[45]

CHORS - chorismate synthase

CHORS (chorismate synthase; P12008) is a homotetrameric enzyme encoded by the *aroC* gene and is located in the cytosol. The flavin $FMNH_2$ is required as cofactor for CHORS; $FADH_2$ works as well but $FMNH_2$ is favoured. The CHORS enzyme is inactive under aerobic conditions, because it is oxygen sensitive. A Δ*aroC* mutant is unable to grow in minimal medium.

Reaction equation

$$1\ 3PSME \rightarrow 1\ CHOR + 1\ Pi$$

5-O-(1-Carboxyvinyl)-3-phosphoshikimate → chorismate + phosphate

Kinetic parameters

Structural (conformational) changes upon flavin and substrate binding [50]. "The line through the data points assumes dissociation of the active tetramer to two inactive dimers (K_D = 0.25 nM) on dilution" [51]. Competitive inhibition of 3PSME by 6-fluoro-3PSME variants [52].

Parameter	Substrate	Value	Unit	Reference
k_{cat}	-	16.5	s^{-1}	[51]
k_{cat}	-	29.0	s^{-1}	[53]
${k_{appt}}^{max}$	-	60.6	s^{-1}	Calculated from experimental data
K_M	3PSME	0.0013	mM	[51]

CHORM - chorismate mutase

The CHORM (chorismate mutase) reaction consists of two (iso)enzymes: PheA (P0A9J8) encoded by the *pheA* gene and TyrA (P07023) encoded by the *tyrA* gene. Both enzymes are homodimeric and are located in the cytosol. PheA is a fused (bifunctional) chorismate mutase/prephenate dehydratase and thus catalyses both reactions. The native enzyme is a dimer of identical subunits each containing a dehydratase active site, a mutase active site and an L-phenylalanine binding site. Prephenate, which is formed from chorismate, dissociates from the mutase site and equilibrates with the bulk medium before combining at the dehydratase site [54]. TyrA is a fused (bifunctional) chorismate mutase/prephenate dehydrogenase and thus catalyses both reactions. The two catalytic activities of TyrA occur in separate portions of the protein. Specifically, the chorismate mutase activity requires the amino-terminal portion of the protein, and the prephenate dehydrogenase activity is in the carboxy-terminal portion of the protein. Both PheA and TyrA are common targets for metabolic engineering to increase titers of L-phenylalanine and L-tyrosine, respectively.

Reaction equation

$$1\ CHOR \rightarrow 1\ PPHN$$

$$\text{chorismate} \rightarrow \text{prephenate}$$

Kinetic parameters - PheA

Allosteric inhibition of PheA by L-phenylalanine [55, 56] and competitive inhibition of PheA by citrate (CIT) [57] and prephenate [58, 54].

Parameter	Substrate	Value	Unit	Reference
k_{cat}	-	14.7	s^{-1}	[54]
k_{cat}	-	34.7	s^{-1}	[56]
k_{cat}	-	40.0	s^{-1}	[59]
k_{cat}	-	72.0	s^{-1}	[60]
k_{cat}	-	41.4	s^{-1}	[61]
k_{cat}	-	39.0	s^{-1}	[62]
k_{cat}	-	34.2	s^{-1}	[63]
k_{cat}	-	38.9	s^{-1}	[64]
K_M	CHOR	0.045	mM	[55]
K_M	CHOR	0.044	mM	[58]
K_M	CHOR	0.024	mM	[54]
K_M	CHOR	0.031	mM	[57]
K_M	CHOR	0.29	mM	[59]
K_M	CHOR	0.296	mM	[60]
K_M	CHOR	0.3	mM	[61]
K_M	CHOR	0.226	mM	[62]
K_M	CHOR	0.127	mM	[63]
K_M	CHOR	0.304	mM	[64]
K_D	PHE	0.02013	mM	[62]
K_I	PPHN	0.047	mM	[58]
K_I	PPHN	0.031	mM	[54]
K_I	CIT	1.01	mM	[57]

Kinetic parameters - TyrA
Allosteric inhibition of TyrA by L-tyrosine [65, 66], competitive inhibition by prephenate to chorismate [67, 68] and noncompetitive inhibition by citrate [68].

Parameter	Substrate	Value	Unit	Reference
k_{cat}	-	71.0	s^{-1}	[65]
k_{cat}	-	27.0	s^{-1}	[69]
K_M	CHOR	0.39	mM	[70]
K_M	CHOR	0.14	mM	[68]
K_M	CHOR	0.092	mM	[65]
K_M	CHOR	0.045	mM	[69]
K_D	TYR	0.01	mM	Chosen (weakly informative)

PPNDH - prephenate dehydratase

PPNHD (prephenate dehydratase; P0A9J8) is a homodimeric enzyme encoded by the *pheA* gene and is located in the cytosol. PheA is a fused (bifunctional) chorismate mutase/prephenate dehydratase and thus catalyses both reactions. The native enzyme is a dimer of identical subunits each containing a dehydratase active site, a mutase active site and a L-phenylalanine binding site. Prephenate, which is formed from chorismate, dissociates from the mutase site and equilibrates with the bulk medium before combining at the dehydratase site [54].

Reaction equation

$$1\ PPHN \rightarrow 1\ PHPYR + 1\ CO_2 + 1\ H_2O$$

prephenate → phenylpyruvate + carbon dioxide + water

Kinetic parameters

Allosteric inhibition by L-phenylalanine [55, 56], competitive inhibition by aconitate [57] and inhibition by chorismate [54].

Parameter	**Substrate**	**Value**	**Unit**	**Reference**
k_{cat}	-	12.0	s^{-1}	[54]
k_{cat}	-	21.3	s^{-1}	[56]
k_{cat}	-	26.2	s^{-1}	[62]
k_{cat}	-	32.2	s^{-1}	[63]
K_M	PPHN	1.0	mM	[55]
K_M	PPHN	0.47	mM	[54]
K_M	PPHN	0.549	mM	[62]
K_M	PPHN	0.559	mM	[63]
K_D	PHE	0.02013	mM	[62]
K_I	CHOR	26.0	mM	[54]
K_I	PHPYR	4.6	mM	[54]

PHETA1 - L-phenylalanine transaminase

The PHETA1 (L-phenylalanine transaminase) reaction consists of three isozymes: TyrB (P04693) encoded by the *tyrB* gene, AspC (P00509) encoded by the *aspC* gene and IlvE (P0AB80) encoded by the *ilvE* gene. All isozymes are located in the cytosol. TyrB and AspC are homodimeric enzymes, while IlvE is a hexameric enzyme (a dimer of trimers). TyrB, AspC and IlvE are involved in catalyzing the third step of L-phenylalanine and L-tyrosine biosynthesis: all three can contribute to the synthesis of L-phenylalanine; only TyrB and AspC contribute to the biosynthesis of L-tyrosine. Under normal physiological conditions, TyrB is the primary enzyme contributing to the synthesis of L-tyrosine and L-phenylalanine. AspC contributes to their synthesis when substrate pools are large. The contribution of IlvE to L-phenylalanine biosynthesis was demonstrated in triple mutants of *Escherichia coli* K-12 that lacked all three aminotransferases and required both L-phenylalanine and L-tyrosine for growth.

However, $\Delta tyrB$ and $\Delta aspC$ double mutants required only L-tyrosine for growth [71]. TyrB is 1000-fold more active toward aromatic substrates than AspC [72]. PLP (pyridoxal 5-phosphate) is a cofactor of TyrB, AspC (one per subunit) and IlvE. IlvE has very low activity for L-tyrosine and L-phenylalanine [73].

Reaction equation

$$1\ PHPYR + 1\ GLU \leftrightarrow 1\ PHE + 1\ AKG$$

phenylpyruvate + L-glutamate ↔ L-phenylalanine + 2-oxoglutarate

Kinetic parameters - TyrB
Inhibition of TyrB by L-tyrosine [74], L-leucine [75], and 3MOB (3-methyl-2-oxobutanoate) [76].

Parameter	**Substrate**	**Value**	**Unit**	**Reference**
k_{cat}	-	26.7	s^{-1}	[77]
k_{cat}	-	29.3	s^{-1}	[71]
k_{cat}	-	13.2	s^{-1}	[75]
k_{cat}	-	250.0	s^{-1}	[72]
k_{cat}	-	520.0	s^{-1}	[78]
k_{cat}	-	180.0	s^{-1}	[79]
K_M	PHPYR	0.012	mM	[71]
K_M	PHPYR	0.056	mM	[75]
K_M	GLU	0.28	mM	[75]
K_M	PHE	0.333	mM	[80]
K_M	PHE	0.06	mM	[75]
K_M	PHE	0.26	mM	[72]
K_M	PHE	0.56	mM	[79]
K_M	AKG	0.23	mM	[75]
K_M	AKG	1.7	mM	[72]
K_M	AKG	5.0	mM	[79]

Kinetic parameters - AspC
Allosteric inhibition of AspC by 2-methylaspartate [81] and competitive inhibition of AspC by maleate [82].

Parameter	Substrate	Value	Unit	Reference
k_{cat}	-	9.8	s^{-1}	[77]
k_{cat}	-	52.1	s^{-1}	[71]
k_{cat}	-	6.6	s^{-1}	[75]
k_{cat}	-	13.8	s^{-1}	[83]
K_M	PHPYR	3.9	mM	[71]
K_M	PHPYR	0.65	mM	[75]
K_M	GLU	0.9	mM	[75]
K_M	GLU	15	mM	[84]
K_M	GLU	0.6	mM	[82]
K_M	PHE	2.17	mM	[80]
K_M	PHE	0.55	mM	[75]
K_M	PHE	8.0	mM	[83]
K_M	AKG	0.15	mM	[75]
K_M	AKG	0.24	mM	[84]
K_M	AKG	0.47	mM	[85]
K_M	AKG	0.59	mM	[86]
K_I	maleate	5.6	mM	[85]

Kinetic parameters - IlvE

Parameter	Substrate	Value	Unit	Reference
k_{cat}	-	0.8	s^{-1}	[73]
K_M	AKG	1.28	mM	[73]

PPND - prephenate dehydrogenase

PPND (prephenate dehydrogenase; P07023) is a homodimeric enzyme encoded by the *tyrA* gene and is located in the cytosol. TyrA is a fused (bifunctional) chorismate mutase/prephenate dehydrogenase and thus catalyses both reactions. The two catalytic activities of TyrA occur in separate portions of the protein. Specifically, the chorismate mutase activity requires the amino-terminal portion of the protein, and the prephenate dehydrogenase activity is in the carboxy-terminal portion of the protein. Both PheA and TyrA are common targets for metabolic engineering to increase titers of L-phenylalanine and L-tyrosine, respectively.

Reaction equation

$$1\ PPHN + 1\ NAD^{+} \rightarrow 1\ 34HPP + 1\ CO_2 + 1\ NADH$$

prephenate + NAD^{+} → 3-(4-hydroxyphenyl)pyruvate + carbon dioxide + NADH

Kinetic parameters

Competitive/allosteric inhibition by L-tyrosine to prephenate [65, 66].

Parameter	**Substrate**	**Value**	**Unit**	**Reference**
k_{cat}	-	71.0	s^{-1}	[65]
k_{cat}	-	27.0	s^{-1}	[69]
K_M	PPHN	0.37	mM	[70]
K_M	PPHN	0.13	mM	[68]
K_M	PPHN	0.044	mM	[69]
K_M	NAD^+	0.33	mM	[70]
K_M	NAD^+	0.05	mM	[65]
K_M	NAD^+	0.103	mM	[69]
K_I	TYR	0.1	mM	[65]

TYRTA - L-tyrosine aminotransferase

The TYRTA (L-tyrosine aminotransferase) reaction consists of two isozymes: TyrB (P04693) encoded by the *tyrB* gene and AspC (P00509) encoded by the *aspC* gene. Both isozymes are homodimeric enzymes and located in the cytosol. TyrB, AspC and IlvE are involved in catalyzing the third step of L-phenylalanine and L-tyrosine biosynthesis: all three can contribute to the synthesis of L-phenylalanine; only TyrB and AspC contribute to the biosynthesis of L-tyrosine. Under normal physiological conditions, TyrB is the primary enzyme contributing to the synthesis of L-tyrosine and L-phenylalanine. AspC contributes to their synthesis when substrate pools are large. The contribution of IlvE to L-phenylalanine biosynthesis was demonstrated in triple mutants of *Escherichia coli* K-12 that lacked all three aminotransferases and required both L-phenylalanine and L-tyrosine for growth. However, Δ*tyrB* and Δ*aspC* double mutants required only L-tyrosine for growth [71]. TyrB is 1000-fold more active toward aromatic substrates than AspC [72]. PLP (pyridoxal 5-phosphate) is a cofactor of TyrB and AspC (one per subunit).

Reaction equation

$$1\ 34HPP + 1\ GLU \leftrightarrow 1\ TYR + 1\ AKG$$

3-(4-hydroxyphenyl)pyruvate + L-glutamate ↔ L-tyrosine + 2-oxoglutarate

Kinetic parameters - TyrB

Inhibition of TyrB by L-tyrosine [74], L-leucine [75], and 3MOB (3-methyl-2-oxobutanoate) [76].

Parameter	Substrate	Value	Unit	Reference
k_{cat}	-	31.1	s^{-1}	[77]
k_{cat}	-	31.5	s^{-1}	[71]
k_{cat}	-	18.3	s^{-1}	[75]
k_{cat}	-	210.0	s^{-1}	[72]
k_{cat}	-	660.0	s^{-1}	[78]
K_M	34HPP	0.013	mM	[71]
K_M	34HPP	0.032	mM	[75]
K_M	GLU	0.28	mM	[75]
K_M	TYR	0.625	mM	[80]
K_M	TYR	0.042	mM	[75]
K_M	TYR	0.32	mM	[72]
K_M	AKG	0.23	mM	[75]
K_M	AKG	1.3	mM	[72]

Kinetic parameters - AspC
Allosteric inhibition of AspC by 2-methylaspartate [81] and competitive inhibition of AspC by maleate [82].

Parameter	Substrate	Value	Unit	Reference
k_{cat}	-	10.1	s^{-1}	[77]
k_{cat}	-	88.0	s^{-1}	[71]
k_{cat}	-	6.6	s^{-1}	[75]
K_M	34HPP	3.9	mM	[71]
K_M	34HPP	0.4	mM	[75]
K_M	GLU	0.9	mM	[75]
K_M	GLU	15	mM	[84]
K_M	GLU	0.6	mM	[82]
K_M	TYR	1.43	mM	[80]
K_M	TYR	0.45	mM	[75]
K_M	AKG	0.15	mM	[75]
K_M	AKG	0.24	mM	[84]
K_M	AKG	0.47	mM	[85]
K_I	maleate	5.6	mM	[85]

ANS - anthranilate synthase
The ANS (anthranilate synthase) reaction is catalysed by the TrpDE (TrpGDE) complex, a heterotetrameric enzyme (two TrpD (P00895) and two TrpE (P00904) subunits) located in the cytosol and encoded by the *trpD* and *trpE* genes. TrpE on its own can carry out an alternate version of this reaction, using ammonium sulfate rather than glutamine as an amino donor [87]. However, TrpD dramatically increases the affinity of TrpE for glutamine over TrpE alone [88]. Mg^{2+} is preferred as cofactor, but Co^{2+} and Fe^{2+} work as well.

Both mutants are not viable on minimal medium. The complex is more thermostable for both the ANS and ANPRT reactions, than the corresponding individual components. Drawing of subunits and conformational changes in Pabst *et al.* [89].

Reaction equation

$$1\ CHOR + 1\ GLN \rightarrow 1\ ANTH + 1\ GLU + 1\ PYR$$

chorismate + L-glutamine → anthranilate + L-glutamate + pyruvate

Kinetic parameters
Competitive/Allosteric inhibition of chorismate by L-tryptophan [90, 89] and 7-mehtyl-L-tryptophan [91]. Noncompetitive inhibition with respect to ammonium sulfate or L-glutamine by L-tryptophan [90], TrpE alone is inhibited by L-tryptophan (competitive to chorismate) [87].

Parameter	**Substrate**	**Value**	**Unit**	**Reference**
k_{appt}^{max}	-	2.27	s^{-1}	Calculated from experimental data
K_M	CHOR	0.0055	mM	[88]
K_M	CHOR	0.005	mM	[89]
K_M	CHOR	0.0012	mM	[90]
K_M	GLN	0.36	mM	[90]
K_M	NH_4^+	39	mM	[88]
K_M	CHOR of TrpE	0.03	mM	[88]
K_M	NH_4^+ of TrpE	15	mM	[87]
K_M	NH_4^+ of TrpE	25	mM	[88]
K_D	TRP	0.001	mM	[89]

ANPRT - anthranilate phosphoribosyltransferase
ANPRT (anthranilate phosphoribosyltransferase; P00904) is likely a monomeric enzyme encoded by the *trpD* gene or the TrpDE complex [92] and is located in the cytosol. The phosphoribosyl transferase and anthranilate synthase contributing portions of TrpD are present in different sections of the protein. The anthranilate synthase reaction requires the amino-terminal portion of the protein, whereas the phosphoribosyltransferase reaction requires the carboxy-terminal region [92]. A Δ*trpD* mutant is not viable in minimal medium.

Reaction equation

$$1\ ANTH + 1\ PRPP \rightarrow 1\ PRAN + 1\ PPi$$

anthranilate + 5-phosphoribose 1-diphosphate → N-(5-phosphoribosyl)anthranilate + diphosphate

Kinetic parameters
Competitive inhibition of PRPP by L-tryptophan [93] or noncompetitive inhibition of PRPP by L-tryptophan [88] and of anthranilate by L-tryptophan [93].

Parameter	Substrate	Value	Unit	Reference
k_{cat}	-	6.6	s^{-1}	[92]
k_{cat}	-	4.4	s^{-1}	[93]
K_M	ANTH	0.00058	mM	[93]
K_M	PRPP	0.1	mM	[88]
K_M	PRPP	0.05	mM	[93]
K_M	PRPP of TrpD	0.2	mM	[88]
K_I	TRP	0.0005	mM	[89]
K_I	TRP	0.0005	mM	[93]

PRAIi - phosphoribosylanthranilate isomerase

PRAI (phosphoribosylanthranilate isomerase; P00909) is a monomeric enzyme encoded by the *trpC* gene and is located in the cytosol. TrpC catalyses both the PRAI and IGPS reaction and mutant complementation studies demonstrated that the two reactions occur at two distinct, non-overlapping sites on the polypeptide and that 2CPR5P is a free intermediate [94]. The amino-terminal domain carries out the synthase activity and the carboxy-terminal domain carries out the isomerase activity [95], so channelling of the intermediate substrate is not likely. TrpC is unique among the five enzymes in the tryptophan biosynthesis pathway in that it is not part of a multisubunit enzyme complex [96]. A Δ*trpC* mutant in not viable in minimal medium.

Reaction equation

$$1\ PRAN \rightarrow 1\ 2CPR5P$$

N-(5-phosphoribosyl)anthranilate → 1-(2-carboxyphenylamino)-1-deoxyribulose 5-phosphate

Kinetic parameters

Parameter	Substrate	Value	Unit	Reference
k_{cat}	-	38.9	s^{-1}	[95]
k_{cat}	-	40.0	s^{-1}	[97]
k_{cat}	-	40.0	s^{-1}	[98]
K_M	PRAN	0.005 - 0.01	mM	[94]
K_M	PRAN	0.0071	mM	[95]
K_M	PRAN	0.0049	mM	[97]
K_M	PRAN	0.0049	mM	[98]

IGPS - indole-3-glycerol-phosphate synthase

IGPS (indole-3-glycerol-phosphate synthase; P00909) is a monomeric enzyme encoded by the *trpC* gene and is located in the cytosol. TrpC catalyses both the PRAI and IGPS reaction and mutant complementation studies demonstrated that the two reactions occur at two distinct, non-overlapping sites on the polypeptide and that 2CPR5P is a free intermediate [94]. The amino-terminal domain carries out the synthase activity and the carboxy-terminal domain carries out the isomerase activity [95], so channelling of the intermediate substrate is not likely. TrpC is unique among the five enzymes in the tryptophan biosynthesis pathway in that it is not part of a multisubunit enzyme complex [96]. A Δ*trpC* mutant in not viable in minimal medium.

Reaction equation

$$12CPR5P \rightarrow 1\ 3IG3P + 1\ CO_2 + 1\ H_2O$$

1-(2-carboxyphenylamino)-1-deoxyribulose 5-phosphate → (3-indolyl)-glycerol 3-phosphate + carbon dioxide + water

Kinetic parameters

Inhibition by anthranilate and derivatives [99] and by rCdRP [100].

Parameter	Substrate	Value	Unit	Reference
k_{cat}	-	7.2	s^{-1}	[95]
k_{cat}	-	3.6	s^{-1}	[98]
k_{cat}	-	2.7	s^{-1}	[101]
K_M	2CPR5P	0.005	mM	[102]
K_M	2CPR5P	0.0012	mM	[95]
K_M	2CPR5P	0.00042	mM	[98]
K_M	2CPR5P	0.0003	mM	[101]

TRPS1 - L-tryptophan synthase

The TRPS (L-tryptophan synthase) reaction is catalysed by the TrpAB complex, a heterotetrameric enzyme (two TrpA (P0A877) subunits and one dimer of TrpB (P0A879)) located in the cytosol and encoded by the *trpA* and *trpB* genes. The overall L-tryptophan synthase reaction consists of a sequence of two partial reactions. The α subunit of the complex carries out the aldol cleavage of 3IG3P to indole and G3P. The β subunit is responsible for the synthesis of L-tryptophan from indole and L-serine. The intermediate substrate (indole) is channelled through the enzyme complex and does not appear in solution [103]. The TrpA monomer is able to catalyse the first part of the reaction, however within the physiological complex with the B subunit, the reaction rate is increased by 1-2 orders of magnitude [104, 105]. The TrpB dimer is seems to be able to catalyse the second part of the reaction [106]. PLP (pyridoxal 5-phosphate) is a cofactor of the enzyme complex, because two PLP molecules bind to the B dimer. Both Δ*trpA* and Δ*trpB* mutants are not viable on minimal medium.

Reaction equation

$$1\ 3IG3P + 1\ SER \rightarrow 1\ TRP + 1\ G3P + 1\ H_2O$$

(3-indolyl)-glycerol 3-phosphate + L-serine → L-tryptophan + glyceraldehyde 3-phosphate + water

Kinetic parameters

The partial reaction of the TrpA subunit is competitively inhibited by indolepropanol phosphate [107].

Parameter	**Substrate**	**Value**	**Unit**	**Reference**
k_{cat}	-	23.4	s^{-1}	[92]
k_{cat}	-	1.5	s^{-1}	[105]
k_{cat}	-	3.8	s^{-1}	[108]
k_{cat}	-	1.4	s^{-1}	[109]
k_{cat}	-	4.7	s^{-1}	[110]
K_M	3IG3P	0.03	mM	[105]
K_M	3IG3P	0.03	mM	[108]
K_M	3IG3P	0.069	mM	[109]
K_M	SER	14.5	mM	[111]
K_M	SER	0.34	mM	[109]

Supplementary figures

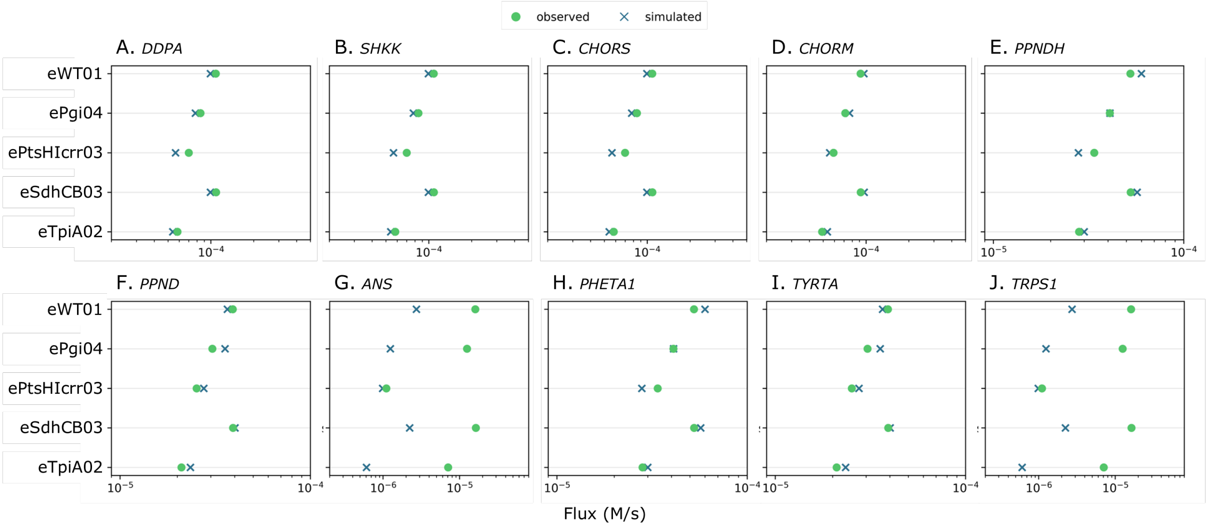


Figure S4.1: Comparison of experimental and simulated fluxes of several reactions. Green circles represent the experimentally observed fluxes, while blue crosses represent the simulated fluxes. The posterior predictive distributions of the simulated fluxes could not be displayed due to the small size caused by the small reported measurement error on the fluxes, $1 \cdot 10^{-5}$ (A) Comparison of DDPA (3-deoxy-7-phosphoheptulonate synthase). (B) Comparison of SHKK (shikimate kinase). (C) Comparison of CHORS (chorismate synthase). (D) Comparison of CHORM (chorismate mutase). (E) Comparison of PPNDH (prephenate dehydratase). (F) Comparison of PPND (prephenate dehydrogenase). (G) Comparison of ANS (anthranilate synthase). (H) Comparison of PHETA1 (L-phenylalanine transaminase). (I) Comparison of TYRTA (L-tyrosine aminotransferase). (J) Comparison of TRPS1 (L-tryptophan synthase).

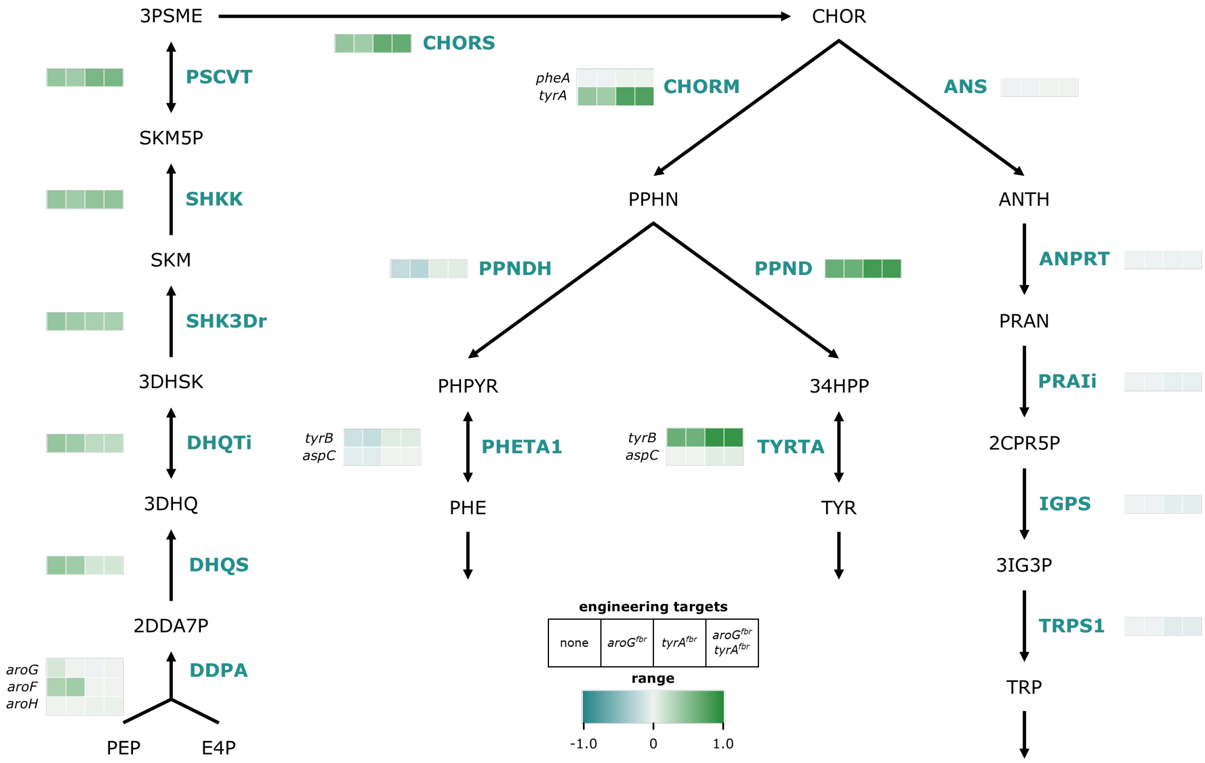


Figure S4.2: Metabolic control analysis of the aromatic amino acid biosynthesis for the ePtsHIcrr03 experiment and three common metabolic engineering targets for L-tyrosine overproduction. The targets include a feedback-resistant version of *aroG*, the L-phenylalanine-sensitive DDPA (3-deoxy-7-phosphoheptulonate synthase), a feedback-resistant version of *tyrA*, the L-tyrosine-sensitive CHORM (chorismate mutase), and the combination of the two previous targets. The L-tyrosine combined consumption reaction was set as target and the flux response coefficients of changes in enzyme concentrations in mol/L enzyme per (mol/L)/s were plotted for each enzyme.

Supplementary tables

Table S4.1: Reprocessed enzyme concentrations in mol/L of the KALE data set.

Enzyme name	eWT01	ePgi04	ePtsHIcrr03	eSdhCB03	eTpiA02
DDPA *aroG*	8.10e-6±1.42e-6	1.30e-5±2.51e-6	2.72e-6±3.07e-6	8.29e-7±2.07e-7	1.98e-6±3.06e-7
DDPA *aroF*	4.91e-6±5.57e-7	1.02e-6±1.18e-7	1.02e-6±3.33e-7	1.19e-6±3.74e-7	1.30e-6±1.78e-7
DDPA *aroH*	1.13e-6±1.03e-7	1.18e-6±4.69e-8	2.22e-6±4.15e-7	2.00e-6±3.20e-7	2.07e-6±1.89e-7
DHQS	5.55e-7±5.67e-8	6.99e-7±6.34e-8	7.00e-6±1.21e-6	5.48e-6±7.90e-7	5.80e-6±1.54e-6
DHQTi	3.91e-7±5.64e-8	4.61e-7±6.21e-8	6.20e-6±1.87e-6	7.59e-6±1.37e-6	5.98e-6±2.51e-6
SHK3Dr	1.46e-7±1.09e-8	2.33e-7±2.01e-8	4.19e-6±1.12e-6	4.44e-6±1.45e-6	4.03e-6±1.11e-6
SHKK	2.13e-7±3.33e-8	2.56e-7±7.63e-8	1.02e-7		
PSCVT	2.57e-6±4.12e-7	2.55e-6±1.31e-7	1.48e-7	3.32e-7±2.90e-7	2.34e-7
CHORS	1.40e-6±1.36e-7	1.38e-6±4.16e-8	1.86e-6±8.10e-7	1.73e-6±4.31e-7	1.37e-6±6.56e-7
CHORM *pheA*	9.23e-7±1.01e-7	6.60e-7±3.69e-8	1.60e-6±2.38e-7	1.75e-6±4.73e-7	1.82e-6±9.11e-7
CHORM *tyrA*	1.77e-6±1.21e-7	5.73e-7±7.24e-8	6.72e-7±1.14e-6	3.87e-7±2.31e-7	1.75e-6±1.59e-6
PPNDH	9.23e-7±1.01e-7	6.60e-7±3.69e-8	1.60e-6±2.38e-7	1.75e-6±4.73e-7	1.82e-6±9.11e-7
PHETA1 *tyrB*	1.69e-6±1.01e-7	1.86e-6±4.87e-8	2.84e-6±7.17e-7	3.31e-6±9.50e-7	2.85e-6±5.17e-7
PHETA1 *aspC*	1.16e-5±9.87e-7	1.35e-5±9.76e-7	1.03e-6±9.75e-7	1.01e-6±5.54e-7	1.29e-6±1.48e-6
PPND	1.77e-6±1.21e-7	5.73e-7±7.24e-8	6.72e-7±1.14e-6	3.87e-7±2.31e-7	1.75e-6±1.59e-6
TYRTA *tyrB*	1.69e-6±1.01e-7	1.86e-6±4.87e-8	2.84e-6±7.17e-7	3.31e-6±9.50e-7	2.85e-6±5.17e-7
TYRTA *aspC*	1.16e-5±9.87e-7	1.35e-5±9.76e-7	1.03e-6±9.75e-7	1.01e-6±5.54e-7	1.29e-6±1.48e-6
ANS	2.28e-6±4.06e-8	9.01e-7±6.20e-8	1.04e-6±7.60e-7	4.45e-7±1.47e-7	6.79e-7±7.74e-7
ANPRT	2.67e-6±2.33e-7	1.06e-6±8.50e-8	2.08e-6±5.24e-7	1.19e-6±2.86e-7	1.22e-6±5.03e-7
PRAIi	2.63e-6±1.12e-7	1.07e-6±1.19e-7	2.47e-6±9.38e-7	1.69e-6±4.82e-7	1.89e-6±4.15e-7
IGPS	2.63e-6±1.12e-7	1.07e-6±1.19e-7	2.47e-6±9.38e-7	1.69e-6±4.82e-7	1.89e-6±4.15e-7
TRPS1	5.15e-6±3.94e-7	2.20e-6±1.45e-7	5.53e-7±3.43e-7	1.00e-6±3.42e-7	9.47e-7±8.55e-8

Table S4.2: Prior and posterior values of kinetic parameters. The k_{cat} values are in 1/s and the K_M and K_D values are in mol/L.

Parameter	Prior mean	Scale	Posterior mean	Parameter	Prior mean	Scale	Posterior mean
k_{cat} DDPA *aroG*	62.3	0.2	62.7	K_M PPND, NAD	1.03e-4	0.2	9.86e-5
k_{cat} DDPA *aroF*	29.5	0.2	29.1	K_M PHETA1 *tyrB*, PHPYR	5.60e-5	0.2	5.62e-5
k_{cat} DDPA *aroH*	20.6	0.2	20.3	K_M PHETA1 *tyrB*, PHE	6.00e-5	0.2	6.00e-5
k_{cat} DHQS	72.9	0.2	84.0	K_M PHETA1 *tyrB*, GLU	2.80e-4	0.2	2.80e-4
k_{cat} DHQTi	142	0.2	286.5	K_M PHETA1 *tyrB*, AKG	2.30e-4	0.2	2.30e-4
k_{cat} SHK3Dr	178	0.2	242.2	K_M PHETA1 *aspC*, PHPYR	6.50e-4	0.2	6.54e-4
k_{cat} SHKK	32	0.2	241.5	K_M PHETA1 *aspC*, PHE	5.50e-4	0.2	5.50e-4
k_{cat} PSCVT	50.6	0.2	711.0	K_M PHETA1 *aspC*, GLU	9.00e-4	0.2	9.00e-4
k_{cat} CHORS	60.6	0.2	200.4	K_M PHETA1 *aspC*, AKG	1.50e-4	0.2	1.50e-4
k_{cat} CHORM *pheA*	14.7	0.2	5.5	K_M TYRTA *tyrB*, 34HPP	3.20e-5	0.2	2.95e-5
k_{cat} CHORM *tyrA*	27	0.2	79.3	K_M TYRTA *tyrB*, TYR	4.20e-5	0.2	4.23e-5
k_{cat} PPNDH	21.3	0.2	23.7	K_M TYRTA *tyrB*, GLU	2.80e-4	0.2	2.78e-4
k_{cat} PPND	27	0.2	29.1	K_M TYRTA *tyrB*, AKG	2.30e-4	0.2	2.31e-4
k_{cat} PHETA1 **tyrB**	13.2	0.2	13.8	K_M TYRTA *aspC*, 34HPP	4.00e-4	0.2	3.93e-4
k_{cat} PHETA1 *aspC*	6.6	0.2	6.6	K_M TYRTA *aspC*, TYR	4.50e-4	0.2	4.50e-4
k_{cat} TYRTA *tyrB*	18.3	0.2	20.9	K_M TYRTA *aspC*, GLU	9.00e-4	0.2	8.99e-4
k_{cat} TYRTA *aspC*	6.6	0.2	6.7	K_M TYRTA *aspC*, AKG	1.50e-4	0.2	1.50e-4
k_{cat} ANS	2.27	0.2	1.9	K_M ANS, CHOR	1.20e-6	0.2	1.21e-6
k_{cat} ANPRT	4.4	0.2	280.0	K_M ANS, GLN	3.60e-4	0.2	3.63e-4
k_{cat} PRAIi	38.9	0.2	886.7	K_M ANPRT, ANTH	5.80e-7	0.2	5.12e-7
k_{cat} IGPS	7.2	0.2	186.6	K_M ANPRT, PRPP	5.00e-5	0.2	4.40e-5
k_{cat} TRPS1	4.7	0.2	206.5	K_M PRAIi, PRAN	7.10e-6	0.2	6.27e-6

Table S4.2: Prior and posterior values of kinetic parameters - continued. The k_{cat} values are in 1/s and the K_M and K_D values are in mol/L.

Parameter	Prior mean	Scale	Posterior mean	Parameter	Prior mean	Scale	Posterior mean
k_{cat} PHE consump.	1.0	0.2	1.1	K_M IGPS, 2CPR5P	1.20e-6	0.2	1.06e-6
k_{cat} TYR consump.	1.0	0.2	1.1	K_M TRPS1, 3IG3P	6.90e-5	0.2	6.10e-5
k_{cat} TRP consump.	1.0	0.2	0.6	K_M TRPS1, SER	3.40e-4	0.2	3.25e-4
K_M DDPA *aroG*, PEP	3.50e-5	0.2	3.44e-5	K_M PHE consump., PHE	2.06e-4	0.5	9.42e-5
K_M DDPA *aroG*, E4P	2.50e-4	0.2	2.48e-4	K_M TYR consump., TYR	3.58e-4	0.5	1.68e-4
K_M DDPA *aroF*, PEP	1.30e-5	0.2	1.31e-5	K_M TRP consump., TRP	5.64e-5	0.5	8.74e-4
K_M DDPA *aroF*, E4P	8.14e-5	0.2	8.21e-5	K_D DDPA *aroG*, PHE	1.30e-5	0.5	1.42e-5
K_M DDPA *aroH*, PEP	5.30e-6	0.2	5.25e-6	K_D DDPA *aroF*, TYR	9.00e-6	0.2	8.90e-6
K_M DDPA *aroH*, E4P	3.50e-5	0.2	3.47e-5	K_D DDPA *aroH*, TRP	1.40e-6	0.2	1.41e-6
K_M DHQS, 2DDA7P	5.50e-6	0.2	4.86e-6	K_D CHORM *pheA*, PHE	2.01e-5	0.2	1.98e-5
K_M DHQTi, 3DHQ	1.70e-5	0.2	1.53e-5	K_D CHORM *tyrA*, TYR	1.00e-4	0.2	9.34e-8
K_M DHQTi, 3DHSK	4.00e-6	0.2	4.30e-6	K_D PPNDH *pheA*, PHE	2.01e-5	0.2	2.13e-5
K_M SHK3Dr, 3DHSK	1.10e-4	0.2	8.87e-5	K_D PPND *tyrA*, TYR	1.00e-4	0.2	1.04e-7
K_M SHK3Dr, NADPH	1.26e-5	0.2	1.15e-5	K_D ANS, TRP	1.00e-6	0.2	1.00e-6
K_M SHKK, SKM	2.00e-4	0.2	1.02e-4	K_D ANPRT, TRP	1.00e-6	0.2	1.11e-6
K_M SHKK, ATP	1.60e-4	0.2	1.50e-4	$transfer\ constant$ DDPA *aroG*	1	0.5	0.97
K_M PSCVT, SKM5P	2.50e-6	0.2	2.29e-6	$transfer\ constant$ DDPA *aroF*	1	0.5	1.03
K_M PSCVT, PEP	1.60e-5	0.2	1.57e-5	$transfer\ constant$ DDPA *aroH*	1	0.5	0.97
K_M PSCVT, 3PSME	3.00e-6	0.2	3.00e-6	$transfer\ constant$ CHORM *pheA*	1	0.5	1.06
K_M PSCVT, Pi	2.50e-3	0.2	2.54e-3	$transfer\ constant$ CHORM *tyrA*	1	0.5	1.24
K_M CHORS, 3PSME	1.30e-6	0.2	1.12e-6	$transfer\ constant$ PPNDH	1	0.5	0.65
K_M CHORM *pheA*, CHOR	3.10e-5	0.2	3.17e-5	$transfer\ constant$ PPND	1	0.5	0.77

Table S4.2: Prior and posterior values of kinetic parameters - continued. The k_{cat} values are in 1/s and the K_M and K_D values are in mol/L.

Parameter	Prior mean	Scale	Posterior mean	Parameter	Prior mean	Scale	Posterior mean
K_M CHORM *tyrA*, CHOR	4.50e-5	0.2	4.82e-5	$transfer\ constant$ ANS	1	0.5	1.00
K_M PPNDH, PPHN	4.70e-4	0.2	4.33e-4	$transfer\ constant$ ANPRT	1	0.5	0.51
K_M PPND, PPHN	4.40e-5	0.2	4.22e-5				

Table S4.3: Simulated enzyme concentrations in mol/L.

Enzymes	eWT01	ePgi04	ePtsHIcrr03	eSdhCB03	eTpiA02
DDPA *aroG*	8.15e-6	1.30e-5	2.70e-6	8.32e-7	1.96e-6
DDPA *aroF*	4.33e-6	1.02e-6	1.01e-6	1.22e-6	1.30e-6
DDPA *aroH*	1.13e-6	1.21e-6	2.13e-6	2.01e-6	2.06e-6
DHQS	1.77e-2	1.83e-5	7.07e-6	8.60e-6	5.90e-6
DHQTi	1.51e-4	8.53e-6	6.32e-6	7.67e-6	6.20e-6
SHK3Dr	2.89e-6	2.84e-5	4.25e-6	4.56e-6	4.08e-6
SHKK	1.52e-6	1.51e-6	3.52e-7	2.47e-6	2.23e-6
PSCVT	2.87e-6	2.58e-6	2.53e-4	1.15e-3	9.80e-5
CHORS	2.53e-5	3.62e-6	3.34e-6	1.00e-4	2.28e-6
CHORM *pheA*	1.09e-7	1.30e-7	8.92e-7	1.33e-6	1.22e-6
CHORM *tyrA*	5.28e-6	3.15e-6	2.94e-6	2.77e-6	3.97e-6
PPNDH	4.13e-5	4.75e-4	3.73e-6	3.55e-5	1.21e-5
PPND	1.28e-5	2.04e-4	4.72e-6	1.36e-5	5.48e-6
PHETA1 *tyrB*	1.69e-6	1.81e-6	2.24e-6	4.25e-6	4.13e-6
PHETA1 *aspC*	1.43e-5	8.80e-6	1.00e-6	1.04e-6	1.30e-6
TYRTA *tyrB*	1.70e-6	2.15e-6	3.89e-6	6.48e-6	3.41e-6
TYRTA *aspC*	1.48e-5	1.69e-5	1.04e-6	1.01e-6	1.29e-6
ANS	1.49e-6	6.72e-7	5.32e-7	1.19e-6	3.30e-7
ANPRT	2.54e-6	1.16e-6	1.50e-6	4.60e-7	1.02e-6
PRAIi	2.65e-6	1.08e-6	4.27e-6	1.96e-6	2.10e-6
IGPS	2.65e-6	1.08e-6	4.30e-6	1.98e-6	2.11e-6
TRPS1	5.69e-6	2.23e-6	6.01e-7	1.14e-6	1.56e-6

Table S4.4: Simulated metabolite concentrations in mol/L.

Metabolite	eWT01	ePgi04	ePtsHIcrr03	eSdhCB03	eTpiA02
PEP	2.42e-4	1.96e-4	1.18e-3	3.70e-4	4.10e-4
E4P	7.36e-5	2.68e-4	4.71e-5	5.20e-4	5.89e-5
2DDA7P	3.25e-10	2.58e-7	5.11e-7	7.74e-7	5.95e-7
3DHQ	7.10e-7	7.02e-7	2.18e-6	3.11e-6	1.38e-6
3DHSK	2.09e-5	1.39e-6	1.20e-5	1.23e-5	6.54e-6
SKM	3.94e-5	2.87e-5	2.26e-4	2.11e-5	1.19e-5
SKM5P	1.55e-7	1.49e-7	1.50e-9	4.70e-10	5.06e-9
3PSME	2.24e-8	1.35e-7	1.03e-7	5.60e-9	1.51e-7
CHOR	3.98e-2	1.05e-1	2.88e-1	5.73e-2	4.99e-2
PPHN	1.83e-4	1.83e-5	1.93e-3	1.49e-4	3.43e-4
PHPYR	2.15e-3	3.65e-3	7.90e-3	2.44e-3	1.43e-3
34HPP	6.49e-5	4.49e-5	3.16e-5	1.32e-5	1.65e-5
ANTH	2.27e-8	1.46e-8	1.04e-8	8.79e-8	6.29e-9
PRAN	7.28e-9	8.17e-9	1.64e-9	8.08e-9	2.04e-9
2CPR5P	5.87e-9	6.59e-9	1.31e-9	6.45e-9	1.64e-9
3IG3P	1.95e-7	2.30e-7	6.64e-7	9.15e-7	1.52e-7
PHE	1.57e-4	3.23e-4	2.61e-4	1.01e-4	1.67e-4
TYR	1.27e-4	2.78e-4	9.53e-4	1.20e-4	2.08e-4
TRP	5.71e-5	3.35e-5	4.03e-5	5.04e-5	2.38e-5
Pi	9.66e-3	9.53e-3	9.59e-3	1.00e-2	9.16e-3
ATP	5.50e-3	4.69e-3	4.07e-3	5.43e-3	3.73e-3
ADP	6.11e-4	9.87e-4	8.87e-4	5.61e-4	8.83e-4
NADP	8.39e-4	1.50e-3	1.06e-3	1.21e-3	1.91e-3
NADPH	3.32e-5	3.16e-5	9.69e-6	3.24e-5	4.46e-5
NAD	2.38e-3	2.86e-3	3.84e-3	2.87e-3	3.29e-3
NADH	3.00e-6	3.46e-6	2.48e-6	2.67e-6	4.04e-6
CO_2	1.00e-4	1.00e-4	1.00e-4	1.00e-4	1.00e-4
GLU	4.62e-2	8.08e-2	6.21e-2	5.02e-2	8.63e-2
AKG	3.54e-4	6.32e-4	7.03e-4	1.67e-4	5.90e-4
GLN	7.68e-3	1.06e-2	1.15e-2	1.43e-2	8.83e-3
PRPP	1.16e-4	1.15e-4	1.15e-4	1.24e-4	1.15e-4
PYR	1.00e-4	1.00e-4	1.00e-4	1.00e-4	1.00e-4
SER	8.67e-4	8.36e-4	9.22e-4	5.80e-4	1.02e-3
G3P	1.00e-4	1.00e-4	1.00e-4	1.00e-4	1.00e-4
PPi	1.00e-4	1.00e-4	1.00e-4	1.00e-4	1.00e-4

Table S4.5: Simulated reaction fluxes in (mol/L)/s.

Reaction	eWT01	ePgi04	ePtsHIcrr03	eSdhCB03	eTpiA02
DDPA	9.93e-5	7.78e-5	5.65e-5	9.93e-5	5.40e-5
DHQS	9.93e-5	7.78e-5	5.65e-5	9.93e-5	5.40e-5
DHQTi	9.93e-5	7.78e-5	5.65e-5	9.93e-5	5.40e-5
SHK3Dr	9.93e-5	7.78e-5	5.65e-5	9.93e-5	5.40e-5
SHKK	9.93e-5	7.78e-5	5.65e-5	9.93e-5	5.40e-5
PSCVT	9.93e-5	7.78e-5	5.65e-5	9.93e-5	5.40e-5
CHORS	9.93e-5	7.78e-5	5.65e-5	9.93e-5	5.40e-5
CHORM	9.66e-5	7.66e-5	5.55e-5	9.71e-5	5.34e-5
PPNDH	6.00e-5	4.10e-5	2.79e-5	5.69e-5	3.00e-5
PPND	3.66e-5	3.56e-5	2.75e-5	4.01e-5	2.34e-5
PHETA1	6.00e-5	4.10e-5	2.79e-5	5.69e-5	3.00e-5
TYRTA	3.66e-5	3.56e-5	2.75e-5	4.01e-5	2.34e-5
ANS	2.72e-6	1.24e-6	9.88e-7	2.23e-6	6.07e-7
ANPRT	2.72e-6	1.24e-6	9.88e-7	2.23e-6	6.07e-7
PRAIi	2.72e-6	1.24e-6	9.88e-7	2.23e-6	6.07e-7
IGPS	2.72e-6	1.24e-6	9.88e-7	2.23e-6	6.07e-7
TRPS1	2.72e-6	1.24e-6	9.88e-7	2.23e-6	6.07e-7
PHE consump.	6.00e-5	4.10e-5	2.79e-5	5.69e-5	3.00e-5
TYR consump.	3.66e-5	3.56e-5	2.75e-5	4.01e-5	2.34e-5
TRP consump.	2.72e-6	1.24e-6	9.88e-7	2.23e-6	6.07e-7

References

[1] I. M. Keseler, S. Gama-Castro, A. Mackie, R. Billington, C. Bonavides-Martínez, R. Caspi, A. Kothari, M. Krummenacker, P. E. Midford, L Muñiz-Rascado, W K Ong, S Paley, A Santos-Zavaleta, P Subhraveti, V. H. Tierrafría, A. J. Wolfe, J. Collado-Vides, I. T. Paulsen, and P. D. Karp. “The EcoCyc Database in 2021”. In: *Frontiers in Microbiology* 12 (2021), p. 711077. DOI: 10.3389/fmicb.2021.711077.

[2] A. Chang, L. Jeske, S. Ulbrich, J. Hofmann, J. Koblitz, I. Schomburg, M. Neumann-Schaal, D. Jahn, and D. Schomburg. “BRENDA, the ELIXIR core data resource in 2021: new developments and updates”. In: *Nucleic Acids Research* 49.D1 (2020), pp. D498–D508. DOI: 10.1093/nar/gkaa1025.

[3] U. Wittig, M. Rey, A. Weidemann, R. Kania, and W. Müller. “SABIO-RK: an updated resource for manually curated biochemical reaction kinetics”. In: *Nucleic Acids Research* 46.D1 (2017), pp. D656–D660. DOI: 10.1093/nar/gkx1065.

[4] D. E. Tribe, H. Camakaris, and J. Pittard. “Constitutive and Repressible Enzymes of the Common Pathway of Aromatic Biosynthesis in Escherichia coli K-12: Regulation of Enzyme Synthesis at Different Growth Rates”. In: *Journal of Bacteriology* 127.3 (1976), pp. 1085–1097. DOI: 10.1128/jb.127.3.1085-1097.1976.

[5] C. Furdui, L. Zhou, R. W. Woodard, and K. S. Anderson. “Insights into the Mechanism of 3-Deoxy-D-arabino-heptulosonate 7-Phosphate Synthase (Phe) from Escherichia coli Using a Transient Kinetic Analysis”. In: *Journal of Biological Chemistry* 279.44 (2004), pp. 45618–5625. DOI: 10.1074/jbc.M404753200.

[6] R. J. McCandliss, M. D. Poling, and K. M. Herrmann. “3-Deoxy-D-arabino-heptulosonate 7-phosphate synthase. Purification and molecular characterization of the phenylalanine-sensitive isoenzyme from Escherichia coli”. In: *Journal of Biological Chemistry* 253.12 (1978), pp. 4259–4265. DOI: 10.1016/S0021-9258(17)34713-0.

[7] R. J. Simpson and B. E. Davidson. “Studies on 3-Deoxy-d-arabino heptulosonate-7-phosphate Synthetase(phe) from Escherichia coli K12. 2. Kinetic Properties”. In: *European Journal of Biochemistry* 70.2 (1976), pp. 501–507. DOI: 10.1111/j.1432-1033.1976.tb11041.x.

[8] A. K. Sundaram, D. L. Howe, G. Y. Sheflyan, and R. W. Woodard. “Probing the potential metal binding site in Escherichia coli 3-deoxy-d-arabino-heptulosonate 7-phosphate synthase (phenylalanine-sensitive)”. In: *FEBS Letters* 441.2 (1998), pp. 195–199. DOI: 10.1016/S0014-5793(98)01545-2.

[9] J. Xu, C. Hu, S. Shen, W. Wang, P. Jiang, and W. Huang. “Requirement of the N-terminus for dimer formation of phenylalanine-sensitive 3-deoxy-D-arabino-heptulosonate synthase AroG of Escherichia coli”. In: *Journal of Basic Microbiology* 44.5 (2004), pp. 400–406. DOI: 10.1002/jobm.200410396.

[10] R. M. Williamson, A. L. Pietersma, G. B. Jameson, and E. J. Parker. “Stereospecific deuteration of 2-deoxyerythrose 4-phosphate using 3-deoxy-D-arabino-heptulosonate

7-phosphate synthase". In: *Bioorganic and Medicinal Chemistry Letters* 15.9 (2005), pp. 2339–2342. DOI: 10.1016/j.bmcl.2005.02.080.

[11] R. Schoner and K. M. Herrmann. "3-Deoxy-D-arabino-heptulosonate 7-Phosphate Synthase. Purification, Properties, and Kinetics of the Tyrosine-Sensitive Isoenzyme from Escherichia coli". In: *Journal of Biological Chemistry* 251.18 (1976), pp. 5440–5447. DOI: 10.1016/S0021-9258(17)33079-X.

[12] C. A. Ramilo and J. N. S. Evans. "Overexpression, Purification, and Characterization of Tyrosine-Sensitive 3-Deoxy-d-arabino-heptulosonic Acid 7-Phosphate Synthase fromEscherichia coli". In: *Protein Expression and Purification* 9.2 (1997), pp. 253–261. DOI: 10.1006/prep.1996.0690.

[13] J. M. Ray and R. Bauerle. "Purification and Properties of Tryptophan-Sensitive 3-Deoxy-Darabino-Heptulosonate-7-Phosphate Synthase from Escherichia coli". In: *Journal of Bacteriology* 173.6 (1991), pp. 1894–1901. DOI: 10.1128/jb.173.6.1894-1901.1991.

[14] J. Camakaris and J. Pittard. "Purification and Properties of 3-Deoxy-D-Arabino-heptulosonic Acid-7-Phosphate Synthetase (trp) from Escherichia coli". In: *Journal of Bacteriology* 120.2 (1974), pp. 590–597. DOI: 10.1128/jb.120.2.590-597.1974.

[15] J. P. Akowski and R. Bauerle. "Steady-State Kinetics and Inhibitor Binding of 3-Deoxy-D-arabino-heptulosonate-7-phosphate Synthase (Tryptophan sensitive) from Escherichia coli". In: *Biochemistry* 36.50 (1997), pp. 15817–15822. DOI: 10.1021/bi971135t.

[16] S. Myrvold, L. M. Reimer, D. L. Pompliano, and J. W. Frost. "Chemical inhibition of dehydroquinate synthase". In: *Journal of the American Chemical Society* 111.5 (1989), pp. 1861–1866. DOI: 10.1021/ja00187a048.

[17] S. L. Bender, S. Mehdi, and J. R Knowles. "Dehydroquinate synthase: the role of divalent metal cations and of nicotinamide adenine dinucleotide in catalysis". In: *Biochemistry* 28.19 (1989), pp. 7555–7560. DOI: 10.1021/bi00445a009.

[18] L. Negron and E. J Parker. "Fluorinated substrates result in variable leakage of a reaction intermediate during catalysis by dehydroquinate synthase". In: *Organic and Biomolecular Chemistry* 9.8 (2011), pp. 2861–2867. DOI: 10.1039/C0OB01141J.

[19] U. S. Maitra and D. B. Sprinson. "5-Dehydro-3-deoxy-D-arabino-heptulosonic Acid 7-Phosphate. An Intermediate in the 3-Dehydroquinate Synthase Reaction". In: *Journal of Biological Chemistry* 253.15 (1978), pp. 5426–5430. DOI: 10.1016/S0021-9258(17)30389-7.

[20] S. Chaudhuri, J. M. Lambert, L. A. McColl, and J. R. Coggins. "Purification and characterization of 3-dehydroquinase from Escherichia coli". In: *Biochemical Journal* 239.3 (1986), pp. 699–704. DOI: 10.1042/bj2390699.

[21] C. Kleanthous, R. Deka, K. Davis, S. M. Kelly, A. Cooper, S. E. Harding, N. C. Price, A. R. Hawkins, and J. R. Coggins. "A comparison of the enzymological and biophysical properties of two distinct classes of dehydroquinase enzymes". In: *Biochemical Journal* 282.3 (1992), pp. 687–695. DOI: 10.1042/bj2820687.

[22] A. P. Leech, R. James, J. R. Coggins, and C. Kleanthous. "Mutagenesis of Active Site Residues in Type I Dehydroquinase from Escherichia coli: STALLED

CATALYSIS IN A HISTIDINE TO ALANINE MUTANT". In: *Journal of Biological Chemistry* 270.43 (1995), pp. 25827–25836. DOI: 10.1074/jbc.270.43.25827.

[23] C. Liu, Y. M. Liu, C. Y. Sun Q. L. andJiang, and S. J. Liu. "Unraveling the kinetic diversity of microbial 3-dehydroquinate dehydratases of shikimate pathway". In: *AMB Express* 5.1 (2015), p. 7. DOI: 10.1186/s13568-014-0087-y.

[24] S. Mituhashi and B. D. Davis. "Aromatic biosynthesis: XII. Conversion of 5-dehydroquinic acid to 5-dehydroshikimic acid by 5-dehydroquinase". In: *Biochimica et Biophysica Acta (BBA)* 15.1 (1954), pp. 54–61. DOI: 10.1016/0006-3002(54)90093-1.

[25] G. Michel, A. W. Roszak, V. Sauvé, J. Maclean, A. Matte, J. R. Coggins, M. Cygler, and A. J. Lapthorn. "Structures of Shikimate Dehydrogenase AroE and Its Paralog YdiB: A COMMON STRUCTURAL FRAMEWORK FOR DIFFERENT ACTIVITIES". In: *Journal of Biological Chemistry* 278.21 (2003), pp. 19463–19472. DOI: 10.1074/jbc.M300794200.

[26] R. M. Muir, A. M. Ibáñez, S. L. Uratsu, E. S. Ingham, C. A. Leslie, G. H. McGranahan, N. Batra, S. Goyal, J. Joseph, E. D. Jemmis, and A. M. Dandekar. "Mechanism of gallic acid biosynthesis in bacteria (Escherichia coli) and walnut (Juglans regia)". In: *Plant Molecular Biology* 75.6 (2011), pp. 555–565. DOI: 10.1007/s11103-011-9739-3.

[27] K. A. Dell and J. W. Frost. "Identification and Removal of Impediments to Biocatalytic Synthesis of Aromatics from D-Glucose: Rate-Limiting Enzymes in the Common Pathway of Aromatic Amino Acid Biosynthesis". In: *Journal of the American Chemical Society* 115.24 (1993), pp. 11581–11589. DOI: 10.1021/ja00077a065.

[28] F. García-Guevara, I. Bravo, C. Martínez-Anaya, and L. Segovia. "Cofactor specificity switch in Shikimate dehydrogenase by rational design and consensus engineering". In: *Protein Engineering, Design and Selection* 30.8 (2017), pp. 533–541. DOI: 10.1093/protein/gzx031.

[29] M. Noble, Y. Sinha, A. Kolupaev, O. Demin, D. Earnshaw, F. Tobin, J. West, J. D. Martin, C. Qiu, W. S. Liu, W. E. DeWolf, D. Tew, and I. I. Goryanin. "The kinetic model of the shikimate pathway as a tool to optimize enzyme assays for high-throughput screening". In: *Biotechnology and Bioengineering* 95.4 (2006), pp. 560–571. DOI: 10.1002/bit.20772.

[30] K. M. Draths, D. R. Knop, and J. W. Frost. "Shikimic acid and quinic acid: replacing isolation from plant sources with recombinant microbial biocatalysis". In: *Journal of the American Chemical Society* 121.7 (1999), pp. 1603–1604. DOI: 10.1021/ja9830243.

[31] R. C. DeFeyter and J. Pittard. "Purification and Properties of Shikimate Kinase II from Escherichia coli K-12". In: *Journal of Bacteriology* 165.1 (1986), pp. 331–333. DOI: 10.1128/jb.165.1.331-333.1986.

[32] D. Ding, Y. Liu, Y. Xu, P. Zheng, H. Li, D. Zhang, and J. Sun. "Improving the Production of L-Phenylalanine by Identifying Key Enzymes Through Multi-Enzyme

Reaction System in Vitro". In: *Scientific Reports* 6.1 (2016), p. 32208. DOI: 10.1038/srep32208.

[33] G. Durante-Rodríguez, J. M. Mancheño, G. Rivas, C. Alfonso, J. L. García, E. Diaz, and M. Carmona. "Identification of a missing link in the evolution of an enzyme into a transcriptional regulator". In: *PLoS ONE* 8.3 (2013), e57518. DOI: 10.1371/journal.pone.0057518.

[34] K. J. Gruys, M. C. Walker, and J. A. Sikorski. "Substrate synergism and the steady-state kinetic reaction mechanism for EPSP synthase from Escherichia coli". In: *Biochemistry* 31.24 (1992), pp. 5534–5544. DOI: 10.1021/bi00139a016.

[35] K. Duncan, A. Lewendon, and J. R Coggins. "The purification of 5-enolpyruvylshikimate 3-phosphate synthase from an overproducing strain of Escherichia coli". In: *FEBS letters* 165.1 (1984), pp. 121–127. DOI: 10.1016/0014-5793(84)80027-7.

[36] Q. K. Huynh. "Inactivation of 5-Enolpyruvylshikimate 3-Phosphate Synthase by its Substrate Analogue Pyruvate in the Presence of Sodium Cyanoborohydride". In: *Biochemical and Biophysical Research Communications* 185.1 (1992), pp. 317–322. DOI: 10.1016/S0006-291X(05)90002-8.

[37] Q. K. Huynh. "5-Enolpyruvylshikimate-3-phosphate synthase from Escherichia coli—The substrate analogue bromopyruvate inactivates the enzyme by modifying Cys-408 and Lys-411". In: *Archives of Biochemistry and Biophysics* 284.2 (1991), pp. 407–412. DOI: 10.1016/0003-9861(91)90316-B.

[38] Q. K. Huynh. "Reaction of 5-enol-pyruvoylshikimate-3-phosphate synthase with diethyl pyrocarbonate: Evidence for an essential histidine residue". In: *Archives of Biochemistry and Biophysics* 258.1 (1987), pp. 233–239. DOI: 10.1016/0003-9861(87)90340-7.

[39] A. Lewendon and J. R. Coggins. "3-Phosphoshikimate 1-carboxyvinyltransferase from Escherichia coli". In: *Metabolism of Aromatic Amino Acids and Amines*. Vol. 142. Methods in Enzymology. Academic Press, 1987, pp. 342–348. DOI: 10.1016/S0076-6879(87)42045-4.

[40] W. A. Shuttleworth and J. N. S. Evans. "The H385N Mutant of 5-Enolpyruvylshikimate-3-phosphate Synthase: Kinetics, Fluorescence, and Nuclear Magnetic Resonance Studies". In: *Archives of Biochemistry and Biophysics* 334.1 (1996), pp. 37–42. DOI: 10.1006/abbi.1996.0426.

[41] K. Haghani, A. Hatef Salmanian, B. Ranjbar, K. Zakikhan-Alang, and K. Khajeh. "Comparative studies of wild type Escherichia coli 5-enolpyruvylshikimate 3-phosphate synthase with three glyphosate-insensitive mutated forms: Activity, stability and structural characterization". In: *Biochimica et Biophysica Acta (BBA)* 1784.9 (2008), pp. 1167–1175. DOI: 10.1016/j.bbapap.2007.07.021.

[42] P. J. Berti and P. Chindemi. "Catalytic Residues and an Electrostatic Sandwich That Promote Enolpyruvyl Shikimate 3-Phosphate Synthase (AroA) Catalysis". In: *Biochemistry* 48.17 (2009), pp. 3699–3707. DOI: 10.1021/bi802251s.

[43] M. L. Healy-Fried, T. Funke, M. A. Priestman, H. Han, and E. Schönbrunn. "Structural basis of glyphosate tolerance resulting from mutations of Pro101 in Escherichia coli 5-enolpyruvylshikimate-3-phosphate synthase". In: *Journal of*

Biological Chemistry 282.45 (2007), pp. 32949–32955. DOI: 10.1074/jbc.M705624200.

[44] S. Eschenburg, M. L. Healy, M A. Priestman, G. H. Lushington, and E. Schönbrunn. “How the mutation glycine96 to alanine confers glyphosate insensitivity to 5-enolpyruvyl shikimate-3-phosphate synthase from Escherichia coli”. In: *Planta* 216.1 (2002), pp. 129–135. DOI: 10.1007/s00425-002-0908-0.

[45] T. Funke, Y. Yang, H. Han, M. Healy-Fried, S. Olesen, A. Becker, and E. Schönbrunn. “Structural basis of glyphosate resistance resulting from the double mutation Thr97 Ile and Pro101 Ser in 5-enolpyruvylshikimate-3-phosphate synthase from Escherichia coli”. In: *Journal of Biological Chemistry* 284.15 (2009), pp. 9854–9860. DOI: 10.1074/jbc.M809771200.

[46] M. A. Priestman, M. L. Healy, T. Funke, A. Becker, and E. Schönbrunn. “Molecular basis for the glyphosate-insensitivity of the reaction of 5-enolpyruvylshikimate 3-phosphate synthase with shikimate”. In: *FEBS Letters* 579.25 (2005), pp. 5773–5780. DOI: 10.1016/j.febslet.2005.09.066.

[47] W. A. Shuttleworth, M. E. Pohl, G. L. Helms, D. L. Jakeman, and J. N. S. Evans. “Site-Directed Mutagenesis of Putative Active Site Residues of 5-Enolpyruvylshikimate-3-phosphate Synthase”. In: *Biochemistry* 38.1 (1999), pp. 296–302. DOI: 10.1021/bi9815142.

[48] M. He, Y. F. Nie, and P. Xu. “A T42M substitution in bacterial 5-enolpyruvylshikimate-3-phosphate synthase (EPSPS) generates enzymes with increased resistance to glyphosate”. In: *Bioscience, Biotechnology and Biochemistry* 67.6 (2003), pp. 1405–1409. DOI: 10.1271/bbb.67.1405.

[49] M. He, Z. Y. Yang, Y. F. Nie, J. Wang, and P. Xu. “A new type of class I bacterial 5-enopyruvylshikimate-3-phosphate synthase mutants with enhanced tolerance to glyphosate”. In: *Biochimica et Biophysica Acta (BBA)* 1568.1 (2001), pp. 1–6. DOI: 10.1016/S0304-4165(01)00181-7.

[50] P. Macheroux, E. Schönbrunn, D. I. Svergun, V. V. Volkov, M. H. J. Koch, S. Bornemann, and R. N. F. Thorneley. “Evidence for a major structural change in Escherichia coli chorismate synthase induced by flavin and substrate binding”. In: *Biochemical Journal* 335.2 (1998), pp. 319–327. DOI: 10.1042/bj3350319.

[51] M. K. Ramjee, J. R. Coggins, and R. N. F. Thorneley. “A Continuous, Anaerobic Spectrophotometric Assay for Chorismate Synthase Activity That Utilizes Photoreduced Flavin Mononucleotide”. In: *Analytical Biochemistry* 220.1 (1994), pp. 137–141. DOI: 10.1006/abio.1994.1309.

[52] A. Osborne, R. N. F. Thorneley, C. Abell, and S. Bornemann. “Studies with Substrate and Cofactor Analogues Provide Evidence for a Radical Mechanism in the Chorismate Synthase Reaction”. In: *Journal of Biological Chemistry* 275.46 (2000), pp. 35825–35830. DOI: 10.1074/jbc.M005796200.

[53] S. Bornemann, D. J. Lowe, and R. N. F. Thorneley. “The transient kinetics of Escherichia coli chorismate synthase: substrate consumption, product formation, phosphate dissociation, and characterization of a flavin intermediate”. In: *Biochemistry* 35.30 (1996), pp. 9907–9916. DOI: 10.1021/bi952958q.

[54] R. G. Duggleby, M. K. Sneddon, and J. F. Morrison. “Chorismate mutase-prephenate dehydratase from Escherichia coli: active sites of a bifunctional enzyme”. In: *Biochemistry* 17.8 (1978), pp. 1548–1554. DOI: 10.1021/bi00601a030.

[55] T. A. A. Dopheide, P. Crewther, and B. E Davidson. “Chorismate Mutase-Prephenate Dehydratase from Escherichia coli K-12: II. KINETIC PROPERTIES”. In: *Journal of Biological Chemistry* 247.14 (1972), pp. 4447–4452. DOI: 10.1016/S0021-9258(19)45005-9.

[56] G. S. Baldwin and B. E. Davidson. “A Kinetic and Structural Comparison of Chorismate Mutase / Prephenate Dehydratase from Mutant Strains of Escherichia co/i K12 Defective in the PheA Gene”. In: *Archives of Biochemistry and Biophysics* 211.1 (1981), pp. 66–75. DOI: 10.1016/0003-9861(81)90430-6.

[57] G. S. Baldwin and B. E. Davidson. “Kinetic studies on the mechanism of chorismate mutase/prephenate dehydratase from Escherichia coli K12”. In: *Biochimica et Biophysica Acta (BBA)* 742.2 (1983), pp. 374–383. DOI: 10.1016/0167-4838(83)90324-2.

[58] M. J. H. Gething and B. E. Davidson. “Chorismate Mutase/Prephenate Dehydratase from Escherichia coli K12: Modification with 5,5’-Dithio-bis(2-nitrobenzoic acid)”. In: *European Journal of Biochemistry* 78.1 (1977), pp. 103–110. DOI: 10.1111/j.1432-1033.1977.tb11718.x.

[59] J. Stewart, D. B. Wilson, and B. Ganem. “A genetically engineered monofunctional chorismate mutase”. In: *Journal of the American Chemical Society* 112.11 (1990), pp. 4582–4584. DOI: 10.1021/ja00167a088.

[60] D. R. Liu, S. T. Cload, R. M. Pastor, and P. G. Schultz. “Analysis of active site residues in Escherichia coli chorismate mutase by site-directed mutagenesis”. In: *Journal of the American Chemical Society* 118.7 (1996), pp. 1789–1790. DOI: 10.1021/ja953151o.

[61] S. Zhang, P. Kongsaeree, J. Clardy, D. B. Wilson, and B. Ganem. “Site-directed mutagenesis of monofunctional chorismate mutase engineered from the E. coli P-protein”. In: *Bioorganic and Medicinal Chemistry* 4.7 (1996), pp. 1015–1020. DOI: 10.1016/0968-0896(96)00099-5.

[62] S. Zhang, G. Pohnert, P. Kongsaeree, D. B. Wilson, J. Clardy, and B. Ganem. “Chorismate mutase-prephenate dehydratase from Escherichia coli: study of catalytic and regulatory domains using genetically engineered proteins”. In: *Journal of Biological Chemistry* 273.11 (1998), pp. 6248–6253. DOI: 10.1074/jbc.273.11.6248.

[63] S. Zhang, D. B. Wilson, and B. Ganem. “Probing the catalytic mechanism of prephenate dehydratase by site-directed mutagenesis of the Escherichia coli P-protein dehydratase domain”. In: *Biochemistry* 39.16 (2000), pp. 4722–4728. DOI: 10.1021/bi9926680.

[64] J. K. Lassila, J. R. Keeffe, P. Oelschlaeger, and S. L. Mayo. “Computationally designed variants of Escherichia coli chorismate mutase show altered catalytic activity”. In: *Protein Engineering, Design and Selection* 18.4 (2005), pp. 161–163. DOI: 10.1093/protein/gzi015.

[65] G. S. Hudson, G. J. Howlett, and B. E. Davidson. “The binding of tyrosine and NAD+ to chorismate mutase/prephenate dehydrogenase from Escherichia coli K12 and the effects of these ligands on the activity and self-association of the enzyme. Analysis in terms of a model.” In: *Journal of Biological Chemistry* 258.5 (1983), pp. 3114–3120. DOI: `10.1016/S0021-9258(18)32838-2`.

[66] T. Lütke-Eversloh and G. Stephanopoulos. “Feedback Inhibition of Chorismate Mutase/Prephenate Dehydrogenase (TyrA) of Escherichia coli: Generation and Characterization of Tyrosine-Insensitive Mutants”. In: *Applied and Environmental Microbiology* 71.11 (2005), pp. 7224–7228. DOI: `10.1128/AEM.71.11.7224-7228.2005`.

[67] G. L. E. Koch, D. C. Shaw, and F. Gibson. “Studies on the relationship between the active sites of chorismate mutase-prephenate dehydrogenase from Escherichia coli or Aerobacter aerogenes”. In: *Biochimica et Biophysica Acta (BBA)* 258.3 (1972), pp. 719–730. DOI: `10.1016/0005-2744(72)90173-8`.

[68] J. I. Rood, B. Perrot, E. Heyde, and J. F Morrison. “Characterization of monofunctional chorismate mutase/prephenate dehydrogenase enzymes obtained via mutagenesis of recombinant plasmids in vitro”. In: *European Journal of Biochemistry* 124.3 (1982), pp. 513–519. DOI: `10.1111/j.1432-1033.1982.tb06623.x`.

[69] D. Christendat, V. C. Saridakis, and J. L. Turnbull. “Use of Site-Directed Mutagenesis To Identify Residues Specific for Each Reaction Catalyzed by Chorismate Mutase- Prephenate Dehydrogenase from Escherichia coli”. In: *Biochemistry* 37.45 (1998), pp. 15703–15712. DOI: `10.1021/bi981412b`.

[70] G. L. E. Koch, D. C. Shaw, and F. Gibson. “The purification and characterisation of chorismate mutase-prephenate dehydrogenase from Escherichia coli K12”. In: *Biochimica et Biophysica Acta (BBA)* 229.3 (1971), pp. 795–804. DOI: `10.1016/0005-2795(71)90298-4`.

[71] D. H. Gelfand and R. A. Steinberg. “Escherichia coli mutants deficient in the aspartate and aromatic amino acid aminotransferases”. In: *Journal of Bacteriology* 130.1 (1977), pp. 429–440. DOI: `10.1128/jb.130.1.429-440.1977`.

[72] H. Hayashi, K. Inoue, T. Nagata, S. Kuramitsu, and H. Kagamiyama. “Escherichia coli aromatic amino acid aminotransferase: characterization and comparison with aspartate aminotransferase”. In: *Biochemistry* 32.45 (1993), pp. 12229–12239. DOI: `10.1021/bi00096a036`.

[73] F. C. Lee-Peng, M. A. Hermodson, and G. B. Kohlhaw. “Transaminase B from Escherichia coli: quaternary structure, amino-terminal sequence, substrate specificity, and absence of a separate valine-alpha-ketoglutarate activity”. In: *Journal of Bacteriology* 139.2 (1979), pp. 339–345. DOI: `10.1128/jb.139.2.339-345.1979`.

[74] R. H. Collier and G. Kohlhaw. “Nonidentity of the aspartate and the aromatic aminotransferase components of transaminase A in Escherichia coli”. In: *Journal of Bacteriology* 112.1 (1972), pp. 365–371. DOI: `10.1128/jb.112.1.365-371.1972`.

[75] J. T. Powell and J. F. Morrison. "Role of the Escherichia coli aromatic amino acid aminotransferase in leucine biosynthesis". In: *Journal of Bacteriology* 136.1 (1978), pp. 1–4. DOI: 10.1128/jb.136.1.1-4.1978.

[76] N. B. Vartak, B. M. Wang, and C. M. Berg. "A functional leuABCD operon is required for leucine synthesis by the tyrosine-repressible transaminase in Escherichia coli K-12". In: *Journal of Bacteriology* 173.12 (1991), pp. 3864–3871. DOI: 10.1128/jb.173.12.3864-3871.1991.

[77] C. Mavrides and W. Orr. "Multiple forms of plurispecific aromatic:2-oxoglutarate (oxaloacetate) aminotransferase (transaminase A) in Escherichia coli and selective repression by l-tyrosine". In: *Biochimica et Biophysica Acta (BBA)* 336.1 (1974), pp. 70–78. DOI: 10.1016/0005-2795(74)90385-7.

[78] J. J. Onuffer, B. T. Ton, I. Klement, and J. F. Kirsch. "The use of natural and unnatural amino acid substrates to define the substrate specificity differences of Escherichia coli aspartate and tyrosine aminotransferases". In: *Protein Science* 4.9 (1995), pp. 1743–1749. DOI: 10.1002/pro.5560040909.

[79] T. N. Luong and J. F. Kirsch. "A Continuous Coupled Spectrophotometric Assay for Tyrosine Aminotransferase Activity with Aromatic and Other Nonpolar Amino Acids". In: *Analytical Biochemistry* 253.1 (1997), pp. 46–49. DOI: 10.1006/abio.1997.2344.

[80] C. Mavrides and W. Orr. "Multispecific aspartate and aromatic amino acid aminotransferases in Escherichia coli". In: *Journal of Biological Chemistry* 250.11 (1975), pp. 4128–4133. DOI: 10.1016/S0021-9258(19)41395-1.

[81] A. Okamoto, T. Higuchi, K. Hirotsu, S. Kuramitsu, and H. Kagamiyama. "X-Ray Crystallographic Study of Pyridoxal 5'-Phosphate-Type Aspartate Aminotransferases from Escherichia coli in Open and Closed Form". In: *Journal of Biochemistry* 116.1 (1994), pp. 95–107. DOI: 10.1093/oxfordjournals.jbchem.a124509.

[82] I. Miyahara, K. Hirotsu, H. Hayashi, and H. Kagamiyama. "X-Ray Crystallographic Study of Pyridoxamine 5'-Phosphate-Type Aspartate Aminotransferases from Escherichia coli in Three Forms". In: *Journal of Biochemistry* 116.5 (1994), pp. 1001–1012. DOI: 10.1093/oxfordjournals.jbchem.a124620.

[83] Q. Han, J. Fang, and J. Li. "Kynurenine aminotransferase and glutamine transaminase K of Escherichia coli: identity with aspartate aminotransferase". In: *Biochemical Journal* 360.3 (2001), pp. 617–623. DOI: 10.1042/bj3600617.

[84] T. Yagi, H. Kagamiyama, M. Nozaki, and K. Soda. "Glutamate-aspartate transaminase from microorganisms". In: *Glutamate, Glutamine, Glutathione, and Related Compounds*. Vol. 113. Methods in Enzymology. Academic Press, 1985, pp. 83–89. DOI: 10.1016/S0076-6879(85)13020-X.

[85] M. D. Toney and J. F. Kirsch. "Tyrosine 70 fine-tunes the catalytic efficiency of aspartate aminotransferase". In: *Biochemistry* 30.30 (1991), pp. 7456–7461. DOI: 10.1021/bi00244a013.

[86] E. Deu, K. A. Koch, and J. F. Kirsch. "The role of the conserved Lys68*: Glu265 intersubunit salt bridge in aspartate aminotransferase kinetics: multiple forced

covariant amino acid substitutions in natural variants". In: *Protein Science* 11.5 (2002), pp. 1062–1073. DOI: 10.1110/ps.0200902.

[87] J. Ito, E. C. Cox, and C. Yanofsky. "Anthranilate synthetase, an enzyme specified by the tryptophan operon of Escherichia coli: purification and characterization of component I". In: *Journal of Bacteriology* 97.2 (1969), pp. 725–733. DOI: 10.1128/jb.97.2.725-733.1969.

[88] J. Ito and C. Yanofsky. "Anthranilate synthetase, an enzyme specified by the tryptophan operon of Escherichia coli: Comparative studies on the complex and the subunits". In: *Journal of Bacteriology* 97.2 (1969), pp. 734–742. DOI: 10.1128/jb.97.2.734-742.1969.

[89] M. J. Pabst, J. C. Kuhn, and R. L. Somerville. "Feedback Regulation in the Anthranilate Aggregate from Wild Type and Mutant Strains of Escherichia coli". In: *Journal of Biological Chemistry* 248.3 (1973), pp. 901–914. DOI: 10.1016/S0021-9258(19)44352-4.

[90] T. I. Baker and I. P. Crawford. "Anthranilate Synthetase: PARTIAL PURIFICATION AND SOME KINETIC STUDIES ON THE ENZYME FROM ESCHERICHIA COLI". In: *Journal of Biological Chemistry* 241.23 (1966), pp. 5577–5584. DOI: 10.1016/S0021-9258(18)96383-0.

[91] W. A. Held and O. H. Smith. "Mechanism of 3-methylanthranilic acid derepression of the tryptophan operon in Escherichia coli". In: *Journal of Bacteriology* 101.1 (1970), pp. 209–217. DOI: 10.1128/jb.101.1.209-217.1970.

[92] E. N. Jackson and C. Yanofsky. "Localization of two functions of the phosphoribosyl anthranilate transferase of Escherichia coli to distinct regions of the polypeptide chain". In: *Journal of Bacteriology* 117.2 (1974), pp. 502–508. DOI: 10.1128/jb.117.2.502-508.1974.

[93] J. E. Gonzalez and R. L. Somerville. "The anthranilate aggregate of Escherichia coli: kinetics of inhibition by tryptophan of phosphoribosyltransferase". In: *Biochemistry and Cell Biology* 64.7 (1986), pp. 681–691. DOI: 10.1139/o86-094.

[94] T. E. Creighton. "N-(5'-Phosphoribosyl)anthranilate isomerase–indol-3-ylglycerol phosphate synthetase of tryptophan biosynthesis. Relationship between the two activities of the enzyme from Escherichia coli". In: *Biochemical Journal* 120.4 (1970), pp. 699–707. DOI: 10.1042/bj1200699.

[95] M. Eberhard and K. Kirschner. "Modification of a catalytically important residue of indoleglycerol-phosphate synthase from Escherichia coli". In: *FEBS Letters* 245.1 (1989), pp. 219–222. DOI: 10.1016/0014-5793(89)80225-X.

[96] G. E. Christie and T. Platt. "Gene structure in the tryptophan operon of Escherichia coli: Nucleotide sequence of trpC and the flanking intercistronic regions". In: *Journal of Molecular Biology* 142.4 (1980), pp. 519–530. DOI: 10.1016/0022-2836(80)90261-2.

[97] U. Hommel, A. Lustig, and K. Kirschner. "Purification and characterization of yeast anthranilate phosphoribosyltransferase". In: *European Journal of Biochemistry* 180.1 (1989), pp. 33–40. DOI: 10.1111/j.1432-1033.1989.tb14611.x.

[98] M. Eberhard, M. Tsai-Pflugfelder, K. Bolewska, U. Hommel, and K. Kirschner. "Indoleglycerol phosphate synthase-phosphoribosyl anthranilate isomerase: comparison of the bifunctional enzyme from Escherichia coli with engineered monofunctional domains". In: *Biochemistry* 34.16 (1995), pp. 5419–5428. DOI: 10.1021/bi00016a013.

[99] O. H. Smith and C. Yanofsky. "Enzymes involved in the biosynthesis of tryptophan". In: vol. 5. Methods in Enzymology. Academic Press, 1962, pp. 794–806. DOI: 10.1016/S0076-6879(62)05315-X.

[100] J. P. Priestle, M. G. Grütter, J. L. White, M. G. Vincent, M. Kania, E. Wilson, T. S. Jardetzky, K. Kirschner, and J. N. Jansonius. "Three-dimensional structure of the bifunctional enzyme N-(5'-phosphoribosyl) anthranilate isomerase-indole-3-glycerol-phosphate synthase from Escherichia coli". In: *Proceedings of the National Academy of Sciences* 84.16 (1987), pp. 5690–5694. DOI: 10.1073/pnas.84.16.5690.

[101] B. Darimont, C. Stehlin, H. Szadkoski, and K. Kirscner. "Mutational analysis of the active site of indoleglycerol phosphate synthase from Escherichia coli". In: *Protein science* 7.5 (1998), pp. 1221–1232. DOI: 10.1002/pro.5560070518.

[102] T. E. Creighton and C. Yanofsky. "Indole-3-glycerol Phosphate Synthetase of Escherichia coli, an Enzyme of the Tryptophan Operon". In: *Journal of Biological Chemistry* 241.20 (1966), pp. 4616–4624. DOI: 10.1016/S0021-9258(18)99693-6.

[103] E. W. Miles, S. Rhee, and D. R. Davies. "The molecular basis of substrate channeling". In: *Journal of Biological Chemistry* 274.18 (1999), pp. 12193–12196. DOI: 10.1074/jbc.274.18.12193.

[104] W. K. Lim, H. J. Shin, D. L. Milton, and J. K. Hardman. "Relative activities and stabilities of mutant Escherichia coli tryptophan synthase alpha subunits". In: *Journal of Bacteriology* 173.6 (1991), pp. 1886–1893. DOI: 10.1128/jb.173.6.1886-1893.1991.

[105] K. Kirschner, A. N. Lane, and A. W. M. Strasser. "Reciprocal communication between the lyase and synthase active sites of the tryptophan synthase bienzyme complex". In: *Biochemistry* 30.2 (1991), pp. 472–478. DOI: 10.1021/bi00216a024.

[106] M. Kaufmann, T. Schwarz, R. Jaenicke, K. D. Schnackerz, H. E. Meyer, and P Bartholmes. "Limited Proteolysis of the beta2-Dimer of Tryptophan Synthase Yields an Enzymatically Active Derivative That Binds alpha-Subunits". In: *Biochemistry* 30.17 (1991), pp. 4173–4179. DOI: 10.1021/bi00231a010.

[107] K. Kirschner, R. L. Wiskocil, M. Foehn, and L. Rezeau. "The Tryptophan Synthase from Escherichia coli: An Improved Purification Procedure for the alpha-Subunit and Binding Studies with Substrate Analogues". In: *European Journal of Biochemistry* 60.2 (1975), pp. 513–523. DOI: 10.1111/j.1432-1033.1975.tb21030.x.

[108] U. Banik, D. M. Zhu, P. B. Chock, and E. W. Miles. "The tryptophan synthase alpha 2 beta 2 complex: kinetic studies with a mutant enzyme (beta K87T) to provide evidence for allosteric activation by an aminoacrylate intermediate". In: *Biochemistry* 34.39 (1995), pp. 12704–12711. DOI: 10.1021/bi00039a029.

[109] A. N. Lane and K. Kirschner. “Mechanism of the physiological reaction catalyzed by tryptophan synthase from Escherichia coli”. In: *Biochemistry* 30.2 (1991), pp. 479–484. DOI: 10.1021/bi00216a025.

[110] K. S. Anderson, A. Y. Kim, J. M. Quillen, E. Sayers, X. J. Yang, and E. W. Miles. “Kinetic characterization of channel impaired mutants of tryptophan synthase”. In: *Journal of Biological Chemistry* 270.50 (1995), pp. 29936–29944. DOI: 10.1074/jbc.270.50.29936.

[111] I. P. Crawford and J. Ito. “Serine Deamination by the B Protein of Escherichia coli Tryptophan Synthetase”. In: *Proceedings of the National Academy of Sciences of the United States of America* 51.3 (1964), pp. 390–397. DOI: 10.1073/pnas.51.3.390.

Chapter 5.

General conclusions and recommendations

5.1. General conclusions and recommendations

The overall objective of this thesis was to develop computational tools and analytical methods for effective metabolic engineering of microbial cell factories. There is still room for improvement within the DBTL cycle, particularly in the form of high-throughput, automated methods. Here, an automated data analysis pipeline, *autoprot*, was developed for high-throughput absolute quantification of proteins in order to obtain accurate and extensive omics data for subsequent applications. The proteome-wide analysis was improved further by implementation of an optimised sample preparation protocol, which increased the representation of membrane proteins. For integration of proteomics, metabolomics and fluxomics data, a kinetic model of the aromatic amino acid biosynthesis in *E. coli* was constructed. The kinetic model can guide strain design for following DBTL cycle iterations. Altogether, two computational tools, i.e. *autoprot* for automated proteomics data analysis and the *E. coli* aromatic amino acid biosynthesis kinetic model, and two analytical methods, i.e. the full workflow for absolute quantitative proteomics and the optimised protocol for quantitative membrane proteomics, were developed as part of the thesis. The main conclusions and recommendations of the thesis are summarised below.

5.2. Conclusions

A workflow combining library-free DIA data analysis with a standard-free quantification approach and recent protein inference algorithms is optimal for proteome-wide absolute quantification.

Using *autoprot* to benchmark 105 different workflows for absolute quantitative proteomics, an optimal workflow was identified for the current *E. coli* data set. This workflow implemented library-free DIA data analysis with recent protein inference algorithms xTop and LFAQ, which consistently resulted in high numbers of quantified proteins. By carefully selecting the peptides used for quantification, xTop favoured slightly higher precision, which is ideal for comparative studies. In contrast, LFAQ uses machine learning to improve the estimated abundance and hence delivers superior accuracy. The TPA method was deployed as standard-free quantification approach, which achieved comparable precision and slightly higher accuracy with a much higher accessibility than other quantification approaches. From the biofoundry perspective, standard-free quantification approaches are especially beneficial due to the low costs and high throughput compared with quantification approaches using internal standards.

Over 25% of the quantified *E. coli* proteome mass is occupied by membrane proteins.

Membrane proteins account for 33% of the annotated proteome in *E. coli* in terms of number of proteins, which has been known for over two decades. Through semi-absolute protein quantification, 27% of the quantified *E. coli* proteome composition (in fmol/µg) was found to be occupied by membrane proteins.

This substantial membrane fraction was not experimentally demonstrated before and highlights the limited general knowledge about the bacterial membrane. Increasing the number of membrane proteins within a proteome-wide quantitative analysis is essential to obtain a representative proteomics data set.

Automated data analysis enables high-throughput characterisation of microbial cell factories.
Shown in this thesis through proteomics analysis, a computational tool for automated data analysis allowed the benchmarking of a sizeable amount of workflows in a matter of days. Similarly, the tool can be applied to proteomics data analysis of a large number of samples through the execution of a single command line. The future of metabolic engineering for microbial cell factory development lies with biofoundries and high-throughput, automated methods including data analysis approaches for characterisation during the test phase.

Kinetic modelling allows for effective learning from integrated multi-omics data.
The kinetic model covering *E. coli* aromatic amino acid biosynthesis combined proteomics, metabolomics and fluxomics data to investigate the intricacies of the pathway. The collective learning through fitting of the kinetic model to the experimental data allowed for a detailed insight into allosteric regulations that strongly affect the production objectives. The resulting information could be used in strain designs for further iteration of the DBTL cycle within the development of an *E. coli* aromatic amino acid overproducer.

A combination of multiple detergents during cell lysis improves solubilisation of membrane proteins.
A sample preparation protocol implementing multiple detergents for increased solubilisation achieved higher numbers of quantified membrane proteins than a conventional protocol incorporating ultracentrifugation for enrichment. Either NP-40 or Triton X-100 in combination with a high concentration of guanidinium·HCl outperformed all other sample preparation protocols, despite the required clean-up method before HPLC-MS analysis. This optimal method resulted in quantification of 56% and 61% of the theoretical membrane proteome of *E. coli* and *P. putida*, respectively.

Deregulation of key enzymes is essential for L-tyrosine overproduction in *E. coli.*
Application of the *E. coli* aromatic amino acid biosynthesis kinetic model suggested the deregulation of the DDPA *aroG* and CHORM and PPND *tyrA* enzymes as beneficial metabolic engineering targets for the overproduction of L-tyrosine. This suggestion aligned with previous experimental studies and was implemented in an *E. coli* strain which lacks the phosphotransferase system for glucose uptake. Multi-omics data will be collected for the engineered strain and used to refit the kinetic model in order to inform further strain designs.

5.3. Recommendations

Appropriate determination of intracellular protein concentrations should be carefully considered.

The main result produced by the *autoprot* tool are proteome composition values, i.e. fmol of a specific protein per µg of total protein mass. The conversion of proteome composition values to intracellular concentrations relies on the cellular protein density in grams of protein per litre of cell volume. The cellular protein density is usually calculated from measured biomass and protein concentrations, which introduce uncertainty through measurement errors. This uncertainty should be taken into account through, e.g. the use of a scaling factor as discussed by Mori *et al.* [1].

A representative approach should be developed in order to estimate extraction efficiency of membrane proteins.

Accurate membrane proteomics data is mostly limited by the extraction efficiency during cell lysis. The hydrophobic nature of membrane proteins inhibits proper solubilisation, which was partly alleviated by the addition of multiple detergents in the current study. Although the extraction efficiency is challenging to estimate, it would nonetheless be worthwhile to attempt to do so and thereby produce a correction factor for absolute quantification. A representative approach would be required and could involve, for example, tagging of specific membrane proteins where the abundance of the tags can be measured independently as shown in Antelo-Varela *et al.* [2].

The full potential of Bayesian inference should be leveraged to obtain an improved kinetic model fit.

Currently, a MAP estimate of the kinetic model was used for pathway investigation instead of sampling from the posterior distributions due to time constraints. Since posterior approximations derived from this approach are generally not fully accurate, sampling of posterior distributions should be a priority for future applications of the kinetic model. Afterwards, out-of-sample predictions could be used to properly evaluate the model through cross-validation. Fully leveraging of the benefits of Bayesian inference should improve the kinetic model performance substantially and yield additional insights [3, 4].

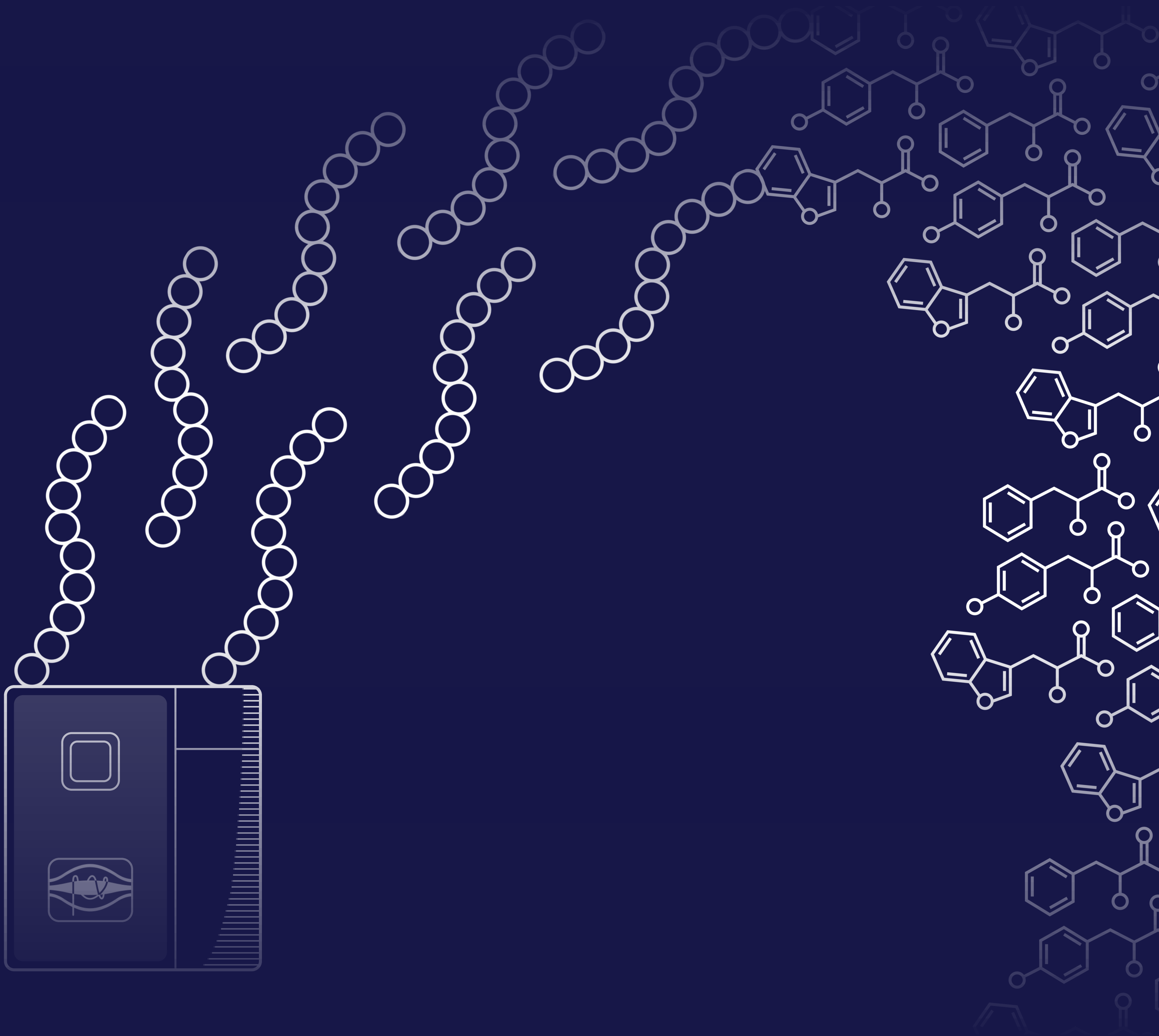
References

[1] M. Mori, Z. Zhang, A. Banaei-Esfahani, J. B. Lalanne, H. Okano, B. C. Collins, A. Schmidt, O. T. Schubert, D. S. Lee, G. W. Li, R. Aebersold, T. Hwa, and C. Ludwig. "From coarse to fine: The absolute *Escherichia coli* proteome under diverse growth conditions". In: *Molecular Systems Biology* 17.5 (2021), e9536. DOI: `10.15252/msb.20209536`.

[2] M. Antelo-Varela, J. Bartel, A. Quesada-Ganuza, K. Appel, M. Bernal-Cabas, T. Sura, A. Otto, M. Rasmussen, J. M. van Dijl, A. Nielsen, S. Maaß, and D. Becher. "Ariadne's Thread in the Analytical Labyrinth of Membrane Proteins: Integration of Targeted and Shotgun Proteomics for Global Absolute Quantification of Membrane Proteins". In: *Analytical Chemistry* 91.18 (2019), pp. 11972–11980. DOI: `10.1021/acs.analchem.9b02869`.

[3] P. A. Saa and L. K. Nielsen. "Construction of feasible and accurate kinetic models of metabolism: A Bayesian approach". In: *Scientific Reports* 6.1 (2016), p. 29635. DOI: `10.1038/srep29635`.

[4] P. A. Saa and L. K. Nielsen. "Formulation, construction and analysis of kinetic models of metabolism: A review of modelling frameworks". In: *Biotechnology Advances* 35.8 (2017), pp. 981–1003. DOI: `10.1016/j.biotechadv.2017.09.005`.



DTU Biosustain
Building 220, Kemitorvet
2800 Kgs. Lyngby, Denmark