universität
wien

DISSERTATION / DOCTORAL THESIS

Titel der Disseratation / Title of the Doctoral Thesis

„Metabolic Modeling in Practice: Advancing Biotechnological Production and Metabolome Data Analysis“

verfasst von / submitted by

Dipl.-Ing. Mathias Gotsmy, BSc

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Doktor der Naturwissenschaften (Dr. rer. nat.)

Wien, 2023 / Vienna, 2023

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	UA 796 605 419
Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet:	Doktoratsstudium Chemie
Betreut von / Supervisor:	Univ.-Prof. Dipl.-Ing. Dr. Jürgen Zanghellini

Acknowledgements

I express my deepest gratitude to my supervisor, Jürgen Zanghellini, for convincing me to do a PhD in his group and for his patience after I got cold feet. It undoubtedly was the right decision to join your team. Your scientific diligence, likeability and love for celebrations truly foster a great work environment. Your quick and constructive feedback is always appreciated.

I thank my collaborators from the University of Vienna and enGenes for trusting the predictions of my simulations. You are crazy.

I thank my colleagues for countless entertaining discussions about science and everything else and for joining me during lunch breaks to get some Hofer hummus.

I am grateful to Radhakrishnan Mahadevan who hosted me at the University of Toronto and all my colleagues there who made me feel at home far away from Austria.

Sincere thanks go out to all my friends in Austria and scattered around Europe. I will visit you all.

Last but not least, I am extremely grateful to my family, especially my parents, Heide and Klaus, for their support over the years and for my brother, Raphael, who is always ready to join me on my travels.

Abstract

Metabolic models serve a dual purpose in understanding biological systems: (1) decoding mechanistic relationships after experimental data acquisition; and (2) acting as predictive tools for experimental design. In my thesis, I showcase both aspects in two key applications demonstrating the versatility of metabolic modeling: (1) normalization of the finger sweat metabolome measurements to enable a quantitative analysis for clinical applications; (2) designing an optimal industrial production process for plasmid DNA production.

Up to date, finger sweat normalization has been a challenge as the sweat rate of participants cannot be controlled for and is hard to measure directly. As a case study on caffeine was conducted by my experimental collaborators, I developed a metabolic model that included the absorption, conversion, and elimination of caffeine in the human body as well as a term representing the mechanism of sweating. By fitting the experimental data onto the developed model, we were able to estimate personalized kinetic constants and showed that they shift little over time.

In a follow-up study, I further improved the goodness of normalization by adding a previously published statistical normalization method on top of the metabolic model. Simulations and case studies of the combined model showed promising results for the quantification of time series measurements of biomarkers in the finger sweat and other body fluids with size effects.

In the plasmid DNA production project, I used metabolic models for medium design of an industrial production process. Counterintuitively, I found that the partial removal of an essential medium component, namely sulfate, can lead to improved productivity and specific yield. The optimal concentration of sulfate in the medium was predicted with dynamic simulations using a genome-scale metabolic model of *Escherichia coli.* Validation experiments conducted by experimental collaborators indeed confirmed the theoretical predictions. We hypothesize that this strategy has high future potential as it is predictions are easily convertible to other biomolecule production processes.

In conclusion, my thesis demonstrates the multifaceted utility of (dynamic) metabolic models in elucidating and predicting biological phenomena. Spanning scientific disciplines, from analytical chemistry to biotechnology, these models offer invaluable insights and hold the key to transformative advancements.

Kurzfassung

Metabolische Modelle dienen einer zweifachen Aufgabe zum Verständnis von biologischen Systemen: (1) sie können, nach experimenteller Datenerfassung, mechanistische Zusammenhänge entschlüsseln und (2) sie können Vorhersagen für die Versuchsplanung treffen. In meiner Dissertation demonstriere ich beide Anwendungsmöglichkeiten von metabolischen Modellen: (1) die Normalisierung von Fingerschweißmetabolommessungen zur quantitative Auswertung für klinische Studien; (2) das Design einer optimalen industriellen Plasmid-DNA-Produktionsfermentation.

Bisher war die Fingerschweißnormalisierung ein ungelöstes Problem da die Fingerschweißproduktionsrate der Probanden nicht beeinflusst und schwerlich direkt gemessen werden kann. Da meine experimentellen Kooperationspartner in ihrer Fallstudie Coffein maßen, entwickelte ich ein metabolisches Modell das die Aufnahme, die Umwandlung und den Abbau sowie Ausscheidung von Coffein im Schweiß des menschlichen Körpers abbildet. Anschließend wurden die Parameter des entwickelten Modells auf die experimentellen Messdaten angepasst und damit konnten wir individuelle kinetische Konstanten des Coffeinmetabolismus eruieren. Außerdem zeigten wir, dass diese individuellen Konstanten geringe zeitliche Variabilität aufweisen.

In einer Folgestudie verbesserte ich die Normalisierungsqualität zusätzlich indem ich eine bereits veröffentlichte statistische Normalisierungsmethode in meinem metabolischen Modell inkludierte. Simulationen und Fallstudien des kombinierten Modells überzeugten mit guter Normalisierungs- und Quantifizierungsqualität von Zeitverlaufsmessungen von Biomarkern in Fingerschweiß und anderen Körperflüssigkeiten.

Im Plasmid-DNA-Produktionsprojekt verwendete ich metabolische Modelle für das Wachstumsmediumdesign eines industriellen Produktionsprozesses. Kontraintuitiverweise fand ich heraus, dass die teilweise Entfernung einer wesentlichen Mediumkomponente, nämlich Sulfat, zu einer verbesserten Produktivität und spezifischen Ausbeute führen kann. Die optimale Sulfatkonzentration im Medium wurde mit dynamischen Simulationen unter Verwendung eines metabolischen Modells von *Escherichia coli* im Genommaßstab vorhergesagt. Von experimentellen Kooperationspartnern durchgeführte Validierungsexperimente bestätigten tatsächlich die theoretischen Vorhersagen. Wir gehen davon aus, dass diese Strategie ein hohes Zukunftspotenzial hat, da sich die Prognosen leicht auf andere Biomolekülproduktionsprozesse übertragen lassen.

Zusammenfassend zeigt meine Dissertation den vielfältigen Nutzen (dynamischer) metabolischer Modelle bei der Aufklärung und Vorhersage biologischer Phänomene. Metabolische Modelle bieten unschätzbare Erkenntnisse und sind der Schlüssel zu transformativen Fortschritten in verschiedensten Wissenschaftsbereichen, von der analytischen Chemie bis zur Biotechnologie.

List of Symbols

Symbol	Name
C_G	glucose concentration in feed
$\mathbf{c}$	cost vector
D	bioreactor dilution rate
E	enzyme concentration
ES	enzyme-substrate complex concentration
f_{avail}	bioavailability of external metabolite
F	bioreactor feed rate
F_G	bioreactor glucose feed rate
I	inhibitor
k	kinetic rate constant
M	metabolite concentration
M_{dose}	amount of external metabolite absorbed
n_{M}	number of metabolites
n_{R}	number of reactions
r_F	scalar feed rate
S	substrate/sulfate concentration
$\mathbf{S}$	stoichiometric matrix
t	time
V	bioreactor volume
V_{dist}	volume of distribution
$v, \mathbf{v}$	flux (vector)
$\mathbf{v}_{\text{lb}}, \mathbf{v}_{\text{ub}}$	lower and upper flux bounds
X	biomass concentration
$\mathbf{z}$	state variable vector
$Y_{X/G}$	biomass yield
$Y_{X/G}^{\text{app}}$	apparent biomass yield
$Y_{P/G}$	product yield
γ	glucose uptake rate
γ_M	glucose to maintenance rate
γ_P	glucose to product rate
γ_X	glucose to biomass rate
$\boldsymbol{\theta}$	parameter vector
μ	growth rate
π	product synthesis rate
Φ	bioreactor dilution

Contents

1. Introduction

1.1. Systems Biology

Cell biology is an incredibly complex research discipline due to an intricate interplay of molecular reactions governed by biological regulation that have the power to shape cells, organisms, and by extension whole ecosystems. Traditionally, scientists investigate life in a reductionist way, i.e., they try to confront the challenge of immense complexity by focusing on smaller and smaller conceptual subunits. Although this focus has lead to many discoveries, it runs the risk of losing sight on the bigger picture, namely that in reality, these conceptual subunits do not exist in isolation but are influencing each other all the time. Therefore, a shift in paradigm has taken place where instead of the traditional reductionist approach, a holistic view on a biological system is emphasized. This newly established field of science is aptly named systems biology [1]. Systems biology fundamentally claims that only by knowing the isolated behaviour in combination with systematic relationships between organs, cells, organelles, proteins, and metabolites, one is able to see, understand, and describe the full picture of cell biology [2].

The establishment of systems biology was critically facilitated by the advent of high-throughput sequencing and analytical chemistry methods. These methods greatly improve our qualitative and quantitative knowledge on genes, proteins, and molecules in cells which facilitates holistic analyses as proposed by systems biology [1]. For example, with the measured data scientists can develop mathematical models to describe cellular processes [2]. Typically, these models describe signal transduction, gene regulations, or metabolic conversions [1]. Moreover, they can focus on subsystems of an organism [3] or try to provide an as broad as possible overview [4]. Figure 1.1 shows the systems biology cycle, where scientists use high-throughput analytical chemistry to inform computer-based models which further explain and predict cellular processes.

My doctoral work sits on the "models & data analysis" third of the systems biology cycle (Figure 1.1), where I used analytical data provided by the scientific community and experimental collaborators to get insight into biological processes. More precisely, during my work, I focused on metabolic conversion networks of variable organisms and sizes, thus in the following sections, I describe the principles of metabolic modeling in more detail.

1.2. Metabolic Models

As the name suggests, metabolic models describe the interactions and conversions of metabolites. Metabolites are small molecules including carbohydrates, amino acids, lipids, peptides, nucleic acids, organic acids, vitamins, and thiols [6]. Metabolic models, for example, can describe the uptake of metabolites into an organism or cell, their conversion from one metabolite into another, and their final excretion or respiration [1]. A metabolic model can be mathematically represented as a network, where metabolites are nodes, and chemical reactions are edges that connect two (or more) metabolites which can be transformed into each other, e.g., by enzymatic catalysis (Figure 1.2).

The creation of a metabolic model typically starts with sequencing a genome of an organism of interest. Functional protein genes in the sequencing data can subsequently be annotated

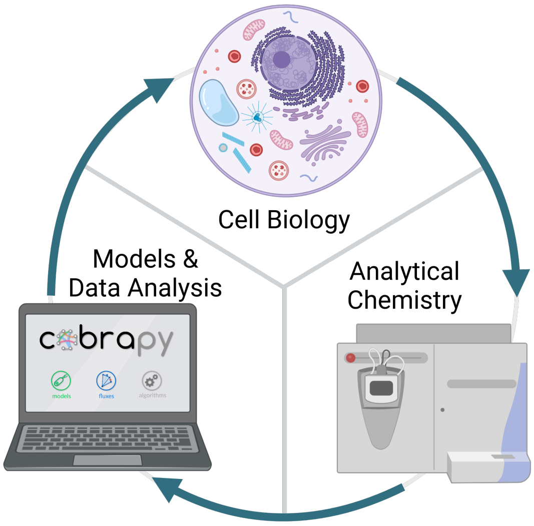


Figure 1.1.: The systems biology cycle [5]. Analytical chemistry comprises most importantly genomics, transcriptomics, proteomics, and metabolomics.

[2]. Using databases like KEGG [7] and BioCyc [8] the annotated genes can be mapped to a metabolic network. Historically, metabolic networks mapped only subsets of the metabolism of a cell, for example, the central carbon metabolism. However, as experimental methods and algorithms improved, it became feasible to create genome scale metabolic model (GSMM) with thousands of metabolites, for example for *Escherichia coli* [9] or *Homo sapiens* [4]. After their reconstruction, metabolic models can be used for computer simulations to study the relationships between metabolites and other properties of the metabolic network [2].

1.2.1. Mathematical Definition

Mathematically, metabolic networks of any size can be represented as a stoichiometric matrix ($\mathbf{S}$) [10]. For example,

$$\mathbf{S} = \begin{array}{c} \\ \mathrm{M}_1 \\ \mathrm{M}_2 \\ \mathrm{M}_3 \\ \mathrm{M}_4 \end{array} \begin{array}{c} \begin{array}{cccccc} \mathrm{R}_1 & \mathrm{R}_2 & \mathrm{R}_3 & \mathrm{R}_1^{\mathrm{ex}} & \mathrm{R}_2^{\mathrm{ex}} & \mathrm{R}_3^{\mathrm{ex}} \end{array} \\ \left(\begin{array}{ccc:ccc} -1 & 0 & 0 & 1 & 0 & 0 \\ 1 & -1 & -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & -1 & 0 \\ 0 & 0 & 1 & 0 & 0 & -1 \end{array} \right) \end{array} \tag{1.1a}$$

describes the metabolic model shown in Figure 1.2. The rows of $\mathbf{S}$ represent the metabolites, and the columns the reactions in the model. Thus $\mathbf{S}$ has the shape number of metabolites times number of reactions ($n_{\mathrm{M}} \times n_{\mathrm{R}}$). The value of the entries represents the chemical net stoichiometry of each reaction. The reactions of the model can be split into two categories as indicated by the dashed line in Equation (1.1a). Reactions to its left represent mass balanced internal conversions of metabolites. In contrast, reactions to the right of the dashed line are exchange reactions that represent fluxes over the system boundary. In this example, the system boundary is the cell membrane and the exchange reactions correspond to uptake and excretion of metabolites from the cell. As we consider only internal metabolites (for now) exchange reactions are not mass

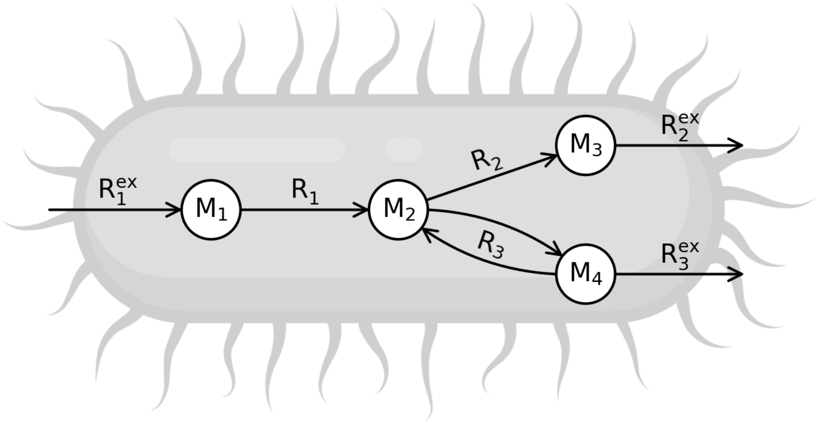


Figure 1.2.: Network representation of a simple example metabolic model. Generally, internal metabolic reactions ($\mathrm{R}_1, \mathrm{R}_2, \mathrm{R}_3$) are set up to be stoichiometrically accurate and mass balanced. R_3 is an example of a reversible reaction.

balanced. To relate the stoichiometry of a metabolic model to the rate of reactions, we can associate a flux vector ($\mathbf{v}$) of the length n_R to $\mathbf{S}$. Together, they describe the time derivatives of the concentration vector ($\mathbf{z}$),

$$\mathbf{S}\mathbf{v} - \mu\mathbf{z} = \dot{\mathbf{z}}. \tag{1.1b}$$

However, in many analysis methods a quasi steady state is assumed [11] and dilution by growth ($\mu\mathbf{z}$) is neglected [12], thus Equation (1.1b) simplifies to

$$\mathbf{S}\mathbf{v} = \mathbf{0}. \tag{1.1c}$$

To further constrain the metabolic network, lower and upper flux bounds can be defined ($\mathbf{v}_\mathrm{lb}$ and $\mathbf{v}_\mathrm{ub}$, respectively). They can be used to set realistic ranges to fluxes, either defined by experimental measurements, enzyme kinetics, thermodynamic considerations, or (non-) reversibility information [1, 13, 14]. Note that in the simple example metabolic network (Figure 1.2), reaction R_3 is reversible. In this case (and without the knowledge of any other information constraining our system), the flux bounds can be written as

$$\begin{pmatrix}\mathbf{v}_\mathrm{lb} & \mathbf{v}_\mathrm{ub}\end{pmatrix} = \begin{pmatrix} 0 & 0 & -\infty & 0 & 0 & 0 \\ \infty & \infty & \infty & \infty & \infty & \infty \end{pmatrix}^\mathrm{T}. \tag{1.1d}$$

As infinite values are computationally hard to deal with, in practice unconstrained fluxes of reaction i are usually set to $\begin{pmatrix}v_\mathrm{lb} & v_\mathrm{ub}\end{pmatrix}_i = \begin{pmatrix}-10^3 & 10^3\end{pmatrix}\,\mathrm{mmol\,g^{-1}\,h^{-1}}$ instead of $\begin{pmatrix}-\infty & \infty\end{pmatrix}$ as these values are still far beyond biological limitations.

To efficiently store and share metabolic models, researchers developed the systems biology markup language (SBML) [15] that is compatible with many computational analysis and simulation packages, e.g. COBRA toolbox [16], CobraPy [17], CNAPy [18], ecmtool [19], DFBAlab [20], and many more. SBML models can be uploaded to and distributed from online repositories such as BIGG [21] and BioModels [22].

1.2.2. Flux Balance Analysis

Flux balance analysis (FBA) is a fundamental computational method used to investigate metabolic models in form of stoichiometric matrices [11]. It is a biased method, i.e., in order to calculate FBA, we need to define an objective function (typically one reaction or a linear combination of reactions in the metabolic network ($\mathbf{c}^{\mathrm{T}}\mathbf{v}$) which is subsequently optimized,

$$\text{maximize} \quad \mathbf{c}^{\mathrm{T}}\mathbf{v} \tag{1.2a}$$

subject to steady state conditions and flux bounds,

$$\begin{aligned} &\mathbf{S}\mathbf{v} = \mathbf{0} \\ &\mathbf{v}_{\mathrm{lb}} \leq \mathbf{v} \leq \mathbf{v}_{\mathrm{ub}} \end{aligned} \tag{1.2b}$$

where the cost vector ($\mathbf{c}$) is a vector of zeros except of the objective reaction(s). As the FBA optimization problem takes the form of a linear program (LP), it can be solved with free or commercial LP solvers like GLPK [23], CPLEX [24], or Gurobi [25]. One important property of FBA solutions is that, although a single optimal value for $\mathbf{c}^{\mathrm{T}}\mathbf{v}$ is found (if the problem is feasible), uniqueness of $\mathbf{v}$ is not guaranteed. Especially, in a typically underconstrained GSMM the non-uniqueness of $\mathbf{v}$ is the norm rather than the exception [26].

The definition of the objective function ($\mathbf{c}^{\mathrm{T}}\mathbf{v}$) and the upper and lower flux bounds ($\mathbf{v}_{\mathrm{ub}}$, $\mathbf{v}_{\mathrm{lb}}$) are crucial for the predictive power of FBA. For microorganisms, the objective function is often a biomass reaction, that combines several intracellular metabolites in an artificial biomass metabolite [27]. Metabolic models for FBA are typically set up such that $\dot{\mathbf{z}}$ as well as $\mathbf{v}$ are normalized by the biomass, i.e., they both have the unit $\mathrm{mmol\,g^{-1}\,h^{-1}}$ [1] and that the molecular mass of the artificial biomass metabolite sums up to $1\,\mathrm{g\,mol^{-1}}$. If this is the case, the rate (i.e., the flux) of the biomass reaction can be biologically interpreted as the growth rate of an organism. The values of the flux bounds ($\mathbf{v}_{\mathrm{ub}}$, $\mathbf{v}_{\mathrm{lb}}$) can be set to measured values which is typically done for fluxes of carbon uptake. However, for the bulk of fluxes (especially many intracellular ones) only (ir-)reversibility information exists and the fluxes are left unconstrained otherwise.

FBA is a powerful tool to explore the solution space of metabolic models, and to predict hard to measure (often intracellular) flux rates. Therefore, many extensions to FBA have been developed. For my thesis, the most important extensions are parsimonious flux balance analysis (pFBA) [28], lexicographic FBA [29], and dynamic flux balance analysis (dFBA) [30]. However, many more methods exist, for example, enzyme constrained FBA [31], thermodynamically constrained FBA [14], or membrane constrained FBA [32].

1.2.2.1. Parsimonious Flux Balance Analysis

pFBA tries to solve the problem of non-uniqueness of the solution by introducing a second optimization step where the sum of enzymes needed to support a flux distribution is minimized [28]. However, as enzyme data often is missing or unavailable, the minimization of the sum of all fluxes in the metabolic model is used as a proxy [28]. pFBA is a two step optimization algorithm,

Step 1:

$$\text{maximize} \quad \mathbf{c}^{\mathrm{T}}\mathbf{v} = g_1 \tag{1.3a}$$

subject to

$$\begin{aligned} &\mathbf{S}\mathbf{v} = 0 \\ &\mathbf{v}_{\mathrm{lb}} \leq \mathbf{v} \leq \mathbf{v}_{\mathrm{ub}}, \end{aligned} \tag{1.3b}$$

Step 2:

$$\text{minimize} \quad ||\mathbf{v}|| \tag{1.3c}$$

subject to

$$\begin{aligned} \mathbf{S}\mathbf{v} &= \mathbf{0} \\ \mathbf{v}_{\text{lb}} \leq \mathbf{v} &\leq \mathbf{v}_{\text{ub}} \\ \mathbf{c}^{\text{T}}\mathbf{v} &= g_1 \end{aligned} \tag{1.3d}$$

Due to its conceptual as well as computational conciseness, pFBA is a popular method for creating unique FBA solutions for all values of $\mathbf{v}$.

1.2.2.2. Lexicographic Flux Balance Analysis

Lexicographic FBA tries to solve the problem of non-uniqueness of the solution by introducing a hierarchical optimization strategy [29].

In order to do so, first, an ordered priority list of fluxes of interested has to be compiled [20]. A typical representation of such a priority list is shown in Table 1.1. The top flux on the list is optimized first, subsequently its optimal value is added as a constraint and the next flux can be optimized.

For the k^{th} reaction of interest ($k \in \{1, 2, 3, ..., K\}$)

$$\text{optimize} \quad v_k = g_k \tag{1.4a}$$

subject to

$$\begin{aligned} \mathbf{S}\mathbf{v} &= 0 \\ \mathbf{v}_{\text{lb}} \leq \mathbf{v} &\leq \mathbf{v}_{\text{ub}} \end{aligned} \tag{1.4b}$$

and subject to $k - 1$ constrained fluxes, for $l \in \{1, 2, 3, ..., k - 1\}$ previous optimization steps

$$v_l = g_l. \tag{1.4c}$$

This is done for all fluxes of interest on the priority list [29] and unique values for these fluxes are obtained. Note that the list has to be manually curated, which requires prior knowledge of the metabolic network [20], however, it may give more accurate solutions to fluxes of interest than

Order (k)	Direction	Reaction
1	max	biomass
2	min	exchange A
3	min	exchange B
K	...	...

Table 1.1.: Typical lexicographic priority list. The main objective often is the biomass reaction (i.e., growth), further reactions of interest often comprise exchange (i.e., uptake) rates and/or production rates [20].

the more generic pFBA approach. Contrary to pFBA where only 2 LP optimization steps are necessary to obtain an unique solution of all fluxes in a system, here K optimization steps are required to obtain K unique fluxes. Depending on the value of K, this property of lexicographic FBA can significantly slow down computations, especially when it is combined with more complex methods like dFBA.

1.2.3. Dynamic Models

FBA methods are a powerful tool to investigate steady state solutions of metabolic fluxes. However, in order to simulate scenarios, such as depletion of nutrient sources due to uptake of metabolites, or biotechnological production processes, a steady state cannot be assumed for all metabolites in a metabolic model. As soon as this is the case, dynamic modeling is required. Mathematically, we can (re)sort the stoichiometric matrix into two parts

$$\mathbf{S} = \begin{pmatrix} \mathbf{S}^{\mathrm{ss}} \\ \mathbf{S}^{\mathrm{dy}} \end{pmatrix} \tag{1.5a}$$

where $\mathbf{S}^{\mathrm{ss}}$ describes metabolites in a steady state

$$\mathbf{S}^{\mathrm{ss}}\mathbf{v} = \dot{\mathbf{z}}^{\mathrm{ss}} = \mathbf{0}, \tag{1.5b}$$

and $\mathbf{S}^{\mathrm{dy}}$ describes the dynamic part of the metabolic model

$$\mathbf{S}^{\mathrm{dy}}\mathbf{v} = \dot{\mathbf{z}}^{\mathrm{dy}} \neq \mathbf{0}. \tag{1.5c}$$

Note, that in some models all metabolites concentrations can change, therefore, $\mathbf{S}^{\mathrm{ss}}$ may be an empty matrix. In dynamic modeling, $\mathbf{z}(t)$ is called the state variable vector and in order to calculate it over time, one has to find a function (f) that describes the time derivative of these state variables ($\dot{\mathbf{z}}$)

$$\dot{\mathbf{z}} = f(\mathbf{z}(t), t; \boldsymbol{\theta}) \tag{1.5d}$$

where the right hand side can be a function of the time (t), the state variables themselves ($\mathbf{z}$), or other parameters ($\boldsymbol{\theta}$). In many cases, these dynamic models are based on knowledge of enzyme kinetics (e.g., the Michaelis Menten model [33]) or physical knowledge (e.g., the Bateman function [34]). In relatively simple cases, it is possible to solve the postulated differential equations analytically, however, in more complex scenarios this is not feasible and numerical approaches have to be applied. In contrast to many well established methods for obtaining the stoichiometric matrix of an organism, defining a dynamic model and estimating its parameters is a difficult challenge [2]. Therefore, compared to $n_{\mathrm{M}} > 1000$ metabolites modeled with GSMM FBA, the number of state variables in dynamic models usually does not exceed 20 [1]. This means that dynamic models are typically more coarse-grained than FBA models. More concretely, to reduce the number of metabolites present, one can lump multiple reactions into one, or omit less relevant metabolic pathways. Despite the challenges, recently researches were able to develop a dynamic model of the central carbon metabolism of *E. coli* [35].

In the subsequent sections, I present several small dynamic models. The concepts they establish may seem basic but are central to the more complex models I have developed throughout my doctoral research.

1.2.3.1. Michaelis Menten Model

As the name suggests the Michaelis Menten model was named after Leonor Michaelis and Maud Menten who published their model on enzyme kinetics in 1913 [33]. Due to its almost universal

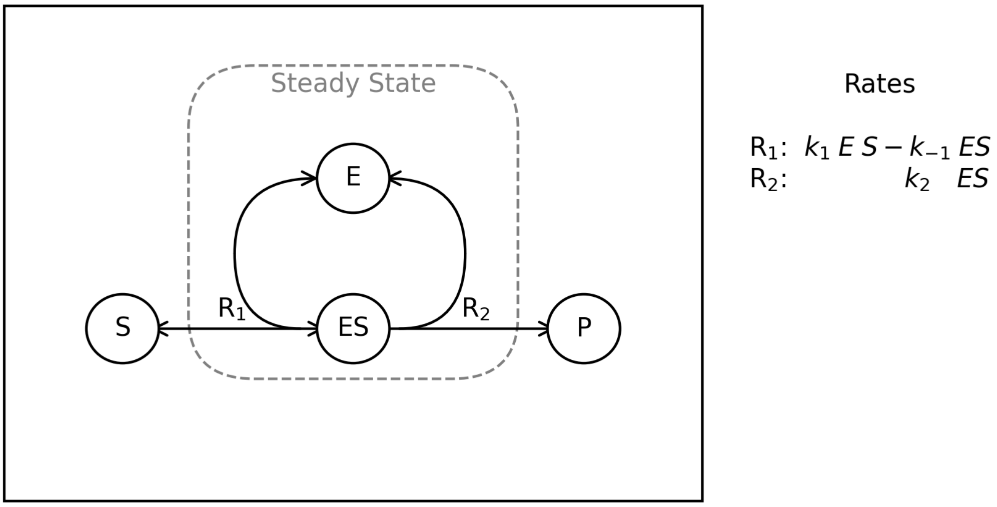


Figure 1.3.: Metabolic network of the Michaelis Menten model. E and ES are assumed to be in a steady state. R$_1$ is reversible, whereas R$_2$ is not.

applicability it has been used in countless studies [36]. The Michelis Menten model describes the enzymatic conversion of a substrate into a product.[1] Free enzyme (E) reacts with the substrate molecule (S) into an enzyme-substrate complex (ES) and, finally, the product (P) and free enzyme (Figure 1.3),

$$\mathrm{E} + \mathrm{S} \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} \mathrm{ES} \xrightarrow{k_2} \mathrm{E} + \mathrm{P}. \tag{1.6a}$$

In the Michaelis Menten model, first order kinetics are assumed. Thus the time derivatives of the state variables read,

$$\dot{\mathbf{z}} = \mathbf{S}\mathbf{v} = \begin{pmatrix} -1 & 1 \\ 1 & -1 \\ -1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} k_1 E\; S - k_{-1} ES \\ k_2\;\; ES \end{pmatrix}. \tag{1.6b}$$

As we are interested in the product formation rate ($\dot{P}$), we can consult the constructed metabolic model and find that

$$\dot{P}(t) = v_2 = k_2 ES. \tag{1.6c}$$

Next, we need to consider which molecules are in a steady state. As substrate is usually significantly more abundant than enzyme ($S \gg E_{\mathrm{tot}}$), the steady state assumption can be made for the enzyme (E) and the enzyme-substrate complex (ES) concentrations [36],

$$\dot{E} = 0, \quad \dot{ES} = 0, \tag{1.6d}$$

[1] Attentive readers may point out that enzymes are not metabolites, however, in this model we can disregard the chemical difference.

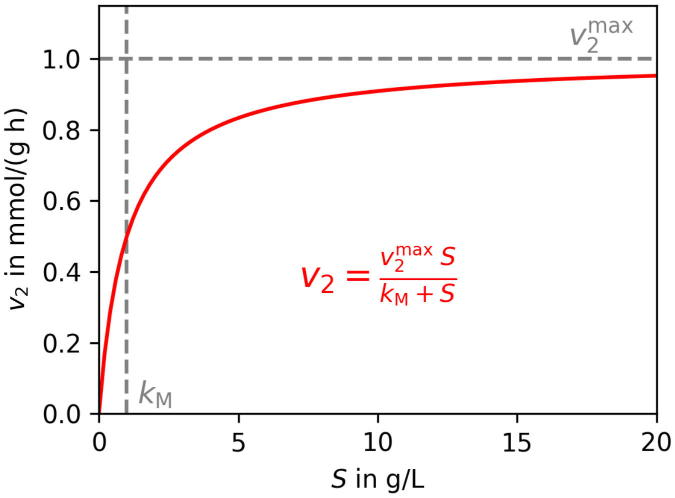


Figure 1.4.: Relationship of flux rate (v) as a function of substrate concentration (S) in form of Michaelis Menten kinetics plotted for the parameters $k_\mathrm{M} = 1\,\mathrm{g\,L^{-1}}$ and $v_2^{\max} = 1\,\mathrm{mmol\,g^{-1}\,h^{-1}}$.

and following, the total enzyme concentration (E_tot) is constant,

$$E_\mathrm{tot} = ES + E = \mathrm{const.} \tag{1.6e}$$

Therefore, the ordered stoichiometric matrix $\mathbf{z} = (E, ES, S, P)^\mathrm{T}$ can be written as

$$\mathbf{S} = \begin{pmatrix} \mathbf{S}^\mathrm{ss} \\ \mathbf{S}^\mathrm{dy} \end{pmatrix} = \left(\begin{array}{cc} -1 & 1 \\ 1 & -1 \\ \hdashline -1 & 0 \\ 0 & 1 \end{array}\right) \tag{1.6f}$$

where the dashed line symbolizes the border between $\mathbf{S}^\mathrm{ss}$ and $\mathbf{S}^\mathrm{dy}$. However, this expression is not helpful to parameterize the model as ES is extremely challenging to measure experimentally. Luckily, by combining Equations (1.6d), (1.6e), and (1.6b), we can substitute ES,

$$0 = k_1(E_\mathrm{tot} - ES)S - (k_{-1} + k_2)ES \tag{1.6g}$$
$$k_1 S\; E_\mathrm{tot} = k_1 ES\; S + (k_{-1} + k_2)ES \tag{1.6h}$$
$$S\; E_\mathrm{tot} = ES(k_\mathrm{M} + S) \tag{1.6i}$$

and get the commonly known Michaelis Menten equation,

$$v_2 = \frac{v_2^{\max}\; S}{k_\mathrm{M} + S} \tag{1.6j}$$

where

$$\begin{aligned} v_2^{\max} &= k_2\; E_\mathrm{tot} \\ k_\mathrm{M} &= \frac{k_{-1} + k_2}{k_1}. \end{aligned} \tag{1.6k}$$

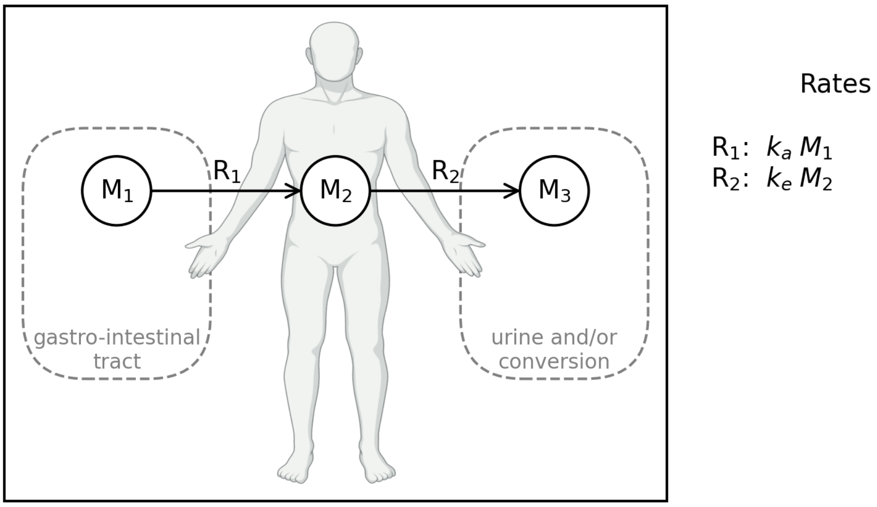


Figure 1.5.: The pharmacological interpretation of the Bateman function. Uptake (R_1) and excretion (R_2) rates of the internal metabolite (M_2) are defined by first order kinetic constants (k_a and k_e, respectively).

Note that for experimental parameterization of $v_2^{\max}$ and k_M in Equation (1.6j), only initial values of v_2 are measured. I.e., measurements are done in a timescale where $S \approx$ const. However, in a modeling context the Michaelis Menten equation is used at all time points. The relationship of substrate concentration and product formation rate (Equation (1.6j)) is illustrated in Figure 1.4. Note that k_M defines the value of S where $v_2^{\max}/2$ is reached.

The Michaelis Menten model has virtually been an enzyme kinetic standard since its establishment [36]. Many extensions to the model have been proposed, for example with competitive and non-competitive inhibitions [37], or by the addition of temperature in the enzyme kinetics [38].

1.2.3.2. Bateman Function

The Bateman function is named after Harry Bateman who in 1910 first published a general analytical solution to the differential equations for a chain of exponential decays [34]. Although it was originally derived within the context of radioactivity, it has been extensively used in pharmacology [39]. A pharmacological intuition of the model is given Figure 1.5 where three metabolites ($\mathrm{M}_1, \mathrm{M}_2, \mathrm{M}_3$) are converted into each other by means of first order kinetics ($\mathrm{R}_1, \mathrm{R}_2$). As no steady state can be assumed, the stoichiometric matrix of the model reads

$$\mathbf{S} = \mathbf{S}^{\mathrm{dy}} = \begin{pmatrix} -1 & 0 \\ 1 & -1 \\ 0 & 1 \end{pmatrix} \tag{1.7a}$$

which combined with the first order kinetics flux vector gives

$$\dot{\mathbf{z}} = \mathbf{S}\mathbf{v} = \begin{pmatrix} -1 & 0 \\ 1 & -1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} k_a\ M_1 \\ k_e\ M_2 \end{pmatrix}. \tag{1.7b}$$

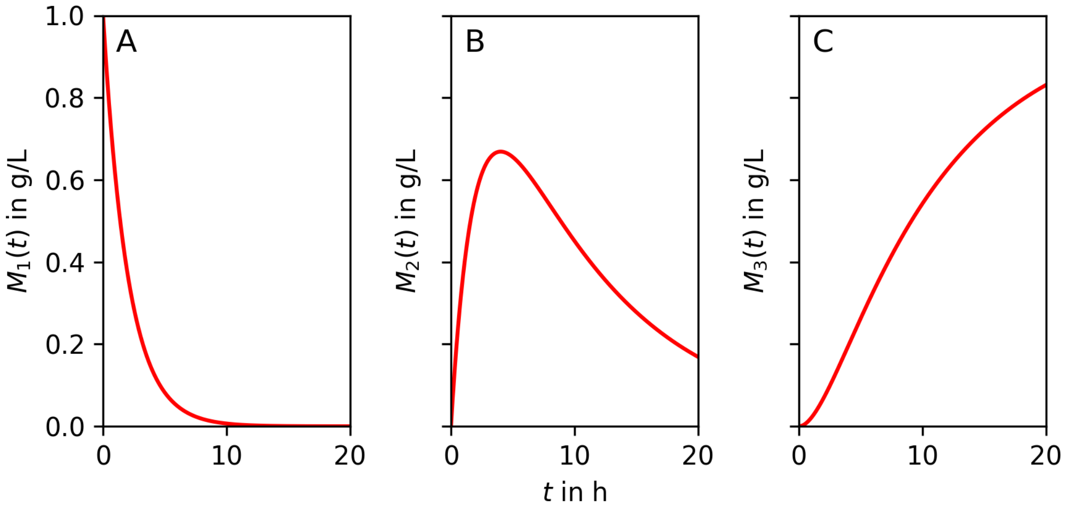


Figure 1.6.: Analytically evaluated time series of the Bateman Function for M_1, M_2, and M_3 (Panels 1-3, respectively). The parameters were set to $M_1(0) = 1\,\mathrm{g\,L^{-1}}, k_a = 0.5\,\mathrm{h^{-1}}$, and $k_e = 0.1\,\mathrm{h^{-1}}$.

With the assumption of $M_1(0)$ being the only non-zero initial state variable, the analytical integration of the differential equation system results in the general solution

$$\mathbf{z}(t) = M_1(0) \begin{pmatrix} e^{-k_a t} \\ k_a/(k_a - k_e)\left(e^{-k_e t} - e^{-k_a t}\right) \\ k_a/(k_a - k_e)\left(1 - k_e/k_a - e^{-k_e t} + k_e/k_a e^{-k_a t}\right) \end{pmatrix}. \tag{1.7c}$$

In the context of pharmacology, scientists are usually only interested in the concentration over time (i.e., the pharmacokinetics) of M_2, thus the general solution is often reduced to

$$M_2(t) = M_1(0) k_a/(k_a - k_e)\left(e^{-k_e t} - e^{-k_a t}\right) \tag{1.7d}$$

which itself is also known as the Bateman Function [39].

Figure 1.6 exemplifies the concentration time series of $\mathrm{M}_1, \mathrm{M}_2$, and M_3. Concentration curves as modeled for $M_2(t)$ (panel B) are commonly seen when xenobiotic metabolites are measured in the blood of a patient after their ingestion. e.g., caffeine [40], ephedrine [41], epicatechin [42], diphenhydramine [43], ibuprofen [44], cannabinoids [45], and many more. Therefore, in pharmacology, the Bateman function is used to describe the pharmokinetics of xenobiotic metabolites. Notably, both reactions of the model are not necessarily chemical transformations but rather complicated biochemical reactions across multiple compartments that are lumped into single process. M_1 can be understood as a xenobiotic metabolite (i.e., a food or drug compound) in the stomach of a patient after ingestion at the initial time point ($t = 0$). M_2 is located in the internal compartment of the model, i.e., the internal fluids of the human body. Consequently, M_3 is the eliminated xenobiotic metabolite. Typically elimination happens by dialysis in the kidneys (i.e., another transport reaction as illustrated in Figure 1.5) or chemical transformation

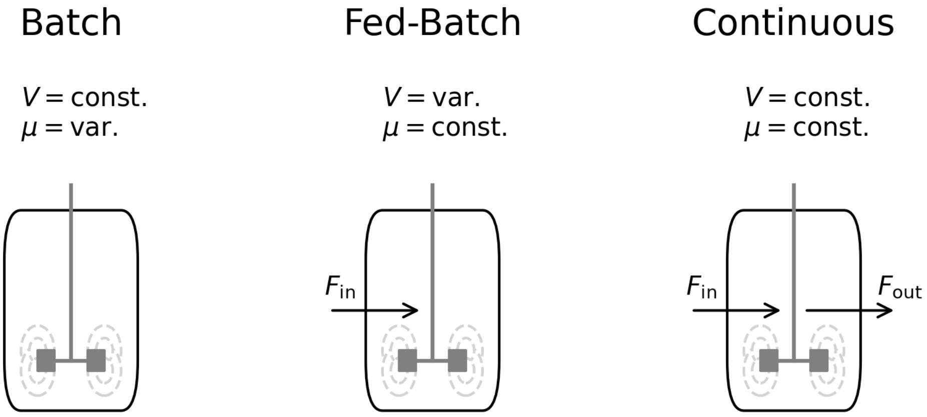


Figure 1.7.: Most commonly used process types for biotechnological production.

to another metabolite via liver enzymes (e.g., via cytochrome P450 group enzymes). Curiously in this context, the value $M_1(0)$ has concrete biological meaning as it is composed out of the volume of distribution (V_{dist}) the dose of an external metabolite ingested (e.g., a drug compound, M_{dose}) and the bioavailability of this compound (f_{avail})

$$M_1(0) = \frac{M_{\text{dose}}\, f_{\text{avail}}}{V_{\text{dist}}} \quad [39]. \tag{1.7e}$$

The values of $M_{\text{dose}}, f_{\text{avail}}, V_{\text{dist}}, k_a$, and k_e can vary significantly between different xenobiotics (e.g., drugs), but also between individual patients. Population averages of these parameters are often measured in clinical studies, as they are critical for drug administration dosage and frequency. However, individual differences are often overlooked and only trough the relatively recent advent of personal health has gained more momentum [46].

1.2.4. Dynamic Process Models

Dynamic metabolic models are a great way to not only simulate cell biology, but also whole biotechnological production processes. Figure 1.7 gives an overview of the three most popular process types: batch, fed-batch, continuous culture. Fed-batch process models are especially relevant for biotechnology as many important drugs are produced with fed-batch cultures (e.g., insulin or plasmid DNA) [47–50].Fed-batch processes are especially popular because setting their feed rate gives the process engineer additional control [48]. Moreover, fed-batch processes usually exhibit improved productivity and yields compared to batch [48].

1.2.4.1. Fed-Batch Process Models

In order to simulate (fed-batch) process, we have to extend our dynamic metabolic model. A schematic of a fed-batch model is given in Figure 1.8, where we have 4 state variables: the bioreactor volume (V) and the concentrations of glucose (G), biomass (X), and product (P)

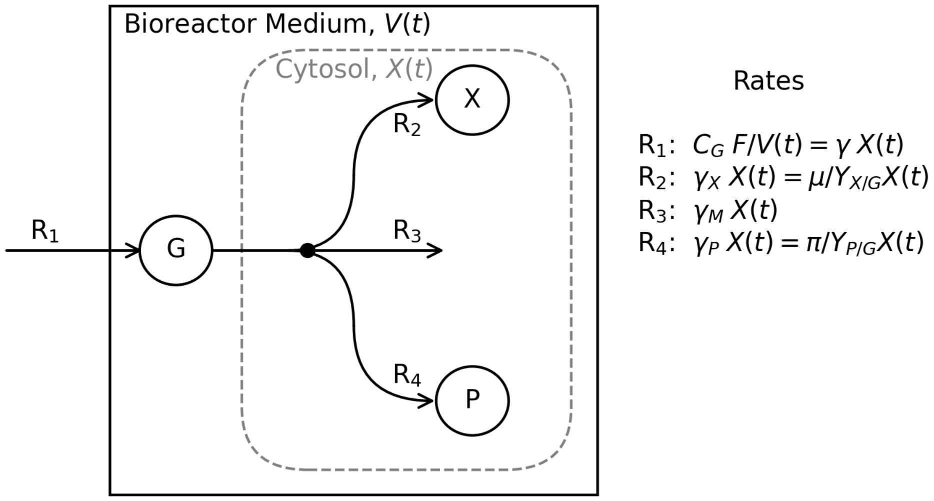

Bioreactor Medium, $V(t)$
Cytosol, $X(t)$
R$_1$
G
R$_2$
X
R$_3$
R$_4$
P
Rates
R$_1$: $C_G\, F/V(t) = \gamma\, X(t)$
R$_2$: $\gamma_X\, X(t) = \mu/Y_{X/G} X(t)$
R$_3$: $\gamma_M\, X(t)$
R$_4$: $\gamma_P\, X(t) = \pi/Y_{P/G} X(t)$

Figure 1.8.: Schematic of a fed-batch bioreactor model. The reactor volume (V) increases with the rate F, biomass grows with the rate $\mu = \gamma_X Y_{X/G}$, and the product is secreted with the rate $\pi = \gamma_P Y_{P/G}$. The rate γ_M represents the maintenance energy requirement of already existing biomass to live.

in the bioreactor. Glucose is fed with the rate $F_G = F\ C_G$ where F is the feed rate and C_G is the glucose concentration in the feed medium. Once in the reactor, glucose is taken up by the cells with the rate γ. As the glucose is the sole source of energy, and assuming that there is no overflow, we can divide γ into three parts supporting the maintenance energy of living cells (γ_M), the biomass growth (γ_X), and the product formation (γ_P). All three of these single step reactions are in reality multiple reactions lumped together. To adjust for stoichiometry a yield parameter (Y) is multiplied (Figure 1.8). We can now set up the stoichiometric matrix and the flux vector of the metabolites $\mathbf{z}(t) = \begin{pmatrix} G & X & P \end{pmatrix}^{\mathrm{T}}$ as

$$\mathbf{S}\mathbf{v} = \begin{pmatrix} \mathbf{S}^{\mathrm{ss}} \\ \mathbf{S}^{\mathrm{dy}} \end{pmatrix} \mathbf{v} = \left(\begin{array}{cccc} 1 & -1 & -1 & -1 \\ \hdashline 0 & Y_{X/G} & 0 & 0 \\ 0 & 0 & 0 & Y_{P/G} \end{array} \right) \begin{pmatrix} C_G F/V \\ \gamma_X X \\ \gamma_M X \\ \gamma_P X \end{pmatrix}. \tag{1.8a}$$

Note that a fed-batch is typically limited by the fed glucose, so we can assume that G is in a steady state [51]

$$G(t) = 0 \text{ and } \dot{G}(t) = 0. \tag{1.8b}$$

Including the dilution of metabolites by the feed rate (F) the time derivatives of the variables read

$$\begin{aligned}\dot{\mathbf{z}} &= \left(\begin{array}{cccc} 1 & -1 & -1 & -1 \\ \hdashline 0 & Y_{X/G} & 0 & 0 \\ 0 & 0 & 0 & Y_{P/G} \end{array}\right) \begin{pmatrix} C_G F/V \\ \gamma_X X \\ \gamma_M X \\ \gamma_P X \end{pmatrix} - \frac{F}{V}\mathbf{z} \\ \dot{V} &= F. \end{aligned} \tag{1.8c}$$

As a consequence of Equation (1.8b) γ can be calculated as

$$\gamma = \frac{F C_G}{V X}. \tag{1.8d}$$

Moreover, we define the growth rate (μ) as

$$\mu = \gamma_X Y_{X/G} = \gamma\, Y^{\text{app}}_{X/G} \tag{1.8e}$$

where the apparent growth yield ($Y^{\text{app}}_{X/G}$) is derived from the Pirt equation [52]

$$\gamma = \gamma_M + \gamma_P + \gamma_X \tag{1.8f}$$

therefore substitution gives

$$Y^{\text{app}}_{X/G} = \frac{\mu}{\gamma_M + \gamma_P + \gamma_X}. \tag{1.8g}$$

Subsequently, while we assume $C_G, Y_{X/G}, Y_{P/G}, \gamma_M$, and γ_P to be constant, we choose F and solve the differential equations analytically. In the following paragraphs, I will present the solution to the process model for the exponential and linear fed-batch case.

Exponential Fed-Batch An exponential feeding rate is the most common form of a fed-batch [48]. It is popular as the internal flux rates are in a steady state throughout the process, i.e., the process engineer can choose a constant μ (granted $\mu \leq \mu_{\max}$, the maximum growth rate of the organism). The feed rate calculates from the chosen μ as

$$F(t) = F_0 e^{\mu t} = \frac{X_0 \mu e^{\mu t}}{C_G Y^{\text{app}}_{X/G}} \quad [49], \tag{1.9a}$$

and the analytical solution to the differential Equation (1.8c) (e.g., computed with SymPy [53]) gives

$$\begin{pmatrix} G(t) \\ X(t) \\ P(t) \\ V(t) \end{pmatrix} = \begin{pmatrix} 0 \\ \Phi(t)^{-1} X_0 e^{\mu t} \\ \Phi(t)^{-1} X_0 (e^{\mu t} - 1)\pi/\mu \\ V_0 + F_0/\mu\, (e^{\mu t} - 1) \end{pmatrix} \tag{1.9b}$$

where

$$\begin{aligned} \Phi(t) &= V(t)/V_0 \\ P_0 &= 0. \end{aligned} \tag{1.9c}$$

Linear Fed-Batch In a linear fed-batch, the feed rate is constant, i.e., the reactor volume ($V(t)$) rises linearly with the rate r_F [48],

$$F(t) = r_F = \text{const.} \tag{1.10a}$$

To always fulfill the condition $G(t) = 0$ the maximal possible feeding rate can be calculated subject to the maximum possible growth rate of a given organism (μ_{max}),

$$r_{F,\text{ max}} = X_0 / C_G(\gamma_M + \gamma_P + \mu_\text{max}/Y_{X/G}). \tag{1.10b}$$

When substituting Equation (1.10a) in Equation (1.8c) and solving the differential equation, the analytical solution reads

$$\begin{pmatrix} G(t) \\ X(t) \\ P(t) \\ V(t) \end{pmatrix} = \left(\begin{array}{c} 0 \\ \dfrac{e^{\bar{\Gamma} t}}{V(t)} \left[X_0 V_0 \Gamma + F_G (e^{\bar{\Gamma} t} - 1) \right] \\ \dfrac{e^{-\bar{\Gamma} t} \pi}{V(t) \Gamma \bar{\Gamma}} \left[-X_0 V_0 \Gamma + F_G (1 - e^{\bar{\Gamma} t}) + \Gamma \left(F_G Y_{X/G} t + X_0 V_0 \right) e^{\bar{\Gamma} t} \right] \\ V_0 + r_F t \end{array} \right) \tag{1.10c}$$

where

$$\begin{aligned} P_0 &= 0 \\ \Gamma &= \gamma_M + \gamma_P \\ \bar{\Gamma} &= Y_{X/G} \Gamma. \end{aligned} \tag{1.10d}$$

Comparison of Exponential and Linear Fed-Batch Figure 1.9 illustrates the analytical solution of an exponential and linear fed-batch production process. Given that the assumptions we made are correct, two observations become apparent: firstly, the final amount of biomass is higher for an exponential fed-batch than for a linear one, although for approximately the first two thirds of the process the linear feed produces more biomass. Secondly, throughout the process the amount of product is higher in a linear fed-batch. Therefore, with the given assumptions linear fed-batch always outperforms an exponential one. Another disadvantage of exponential fed-batches is the fact that much the biomass and product are synthesised in the end of the process. At that time other limitations (e.g., oxygen or high cell densities) may occur that reduce the well-being of the cells and, thus, the theoretical yields are hard to reach experimentally.

On another note, some research suggests that the assumption of constant γ_M and γ_P may not be valid throughout the whole process [54]. Especially the definition of γ_P is dependent on the actual product of interest and its properties. However, this can be reflected in the process model. For example, a function for $\gamma_P(t, \mathbf{z}(t); \boldsymbol{\theta})$ depending on time, state variables or any set of parameters, respectively, can be plugged into Equation (1.8c). One caveat is that analytical solutions to such complex models become complex or are not obtainable at all. In such a case, numerical integrators (e.g., SciPy's `solve_ivp` [55]) have to be used to solve the differential equations.

1.2.4.2. Dynamic Flux Balance Analysis

dFBA essentially tries to combine the concepts for FBA and dynamic (process) models by overcoming the knowledge gap of many reaction rates [30]. More concretely, we extend our

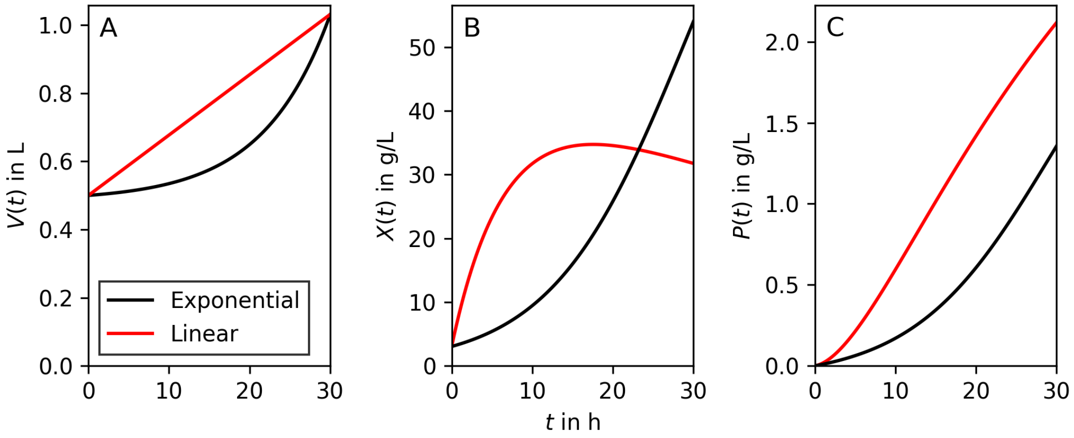


Figure 1.9.: Analytical solutions to an exponential (black) and linear (red) fed-batch process. A list of the parameters is given in Supplementary Table B.1. The values for μ for the exponential fed-batch and $\mu_{\max}$ for the linear fed-batch were chosen in a way that the start and end volumes are the same (panel A). All other parameters are the same for both models.

previous definition of the dynamic model differential equation (Equation (1.5d)) by splitting them into two parts,

$$\dot{\mathbf{z}} = \begin{pmatrix} \mathbf{f}_K(\mathbf{z}(t), t; \boldsymbol{\theta}) \\ \mathbf{f}_U \end{pmatrix} \tag{1.11a}$$

where we can describe K rates with known dynamic equations ($\mathbf{f}_K$), however, equations $\mathbf{f}_U$ for U rates of interest remain unknown. For example, $\mathbf{f}_K$ here may include the uptake rate of glucose which can be modeled by a Michaelis Menten kinetic (Section 1.2.3.1) or estimated over the feed rate (Section 1.2.4.1). $\mathbf{f}_U$, for example, may comprise the growth rate or other less studied essential nutrient uptake rates [20, 30]. dFBA promises to close the knowledge gap as we can use FBA to calculate the undefined time derivatives. In order to do so, we have to append the previous definition of FBA (Section 1.2.2) by including external metabolites. A illustration of the appended simple example model is shown in Figure 1.10. Its stoichiometric matrix can be written as,

$$\mathbf{S} = \begin{array}{c} \\ \mathrm{M}_1 \\ \mathrm{M}_2 \\ \mathrm{M}_3 \\ \mathrm{M}_4 \\ \mathrm{M}_1^{\mathrm{ex}} \\ \mathrm{M}_3^{\mathrm{ex}} \\ \mathrm{M}_4^{\mathrm{ex}} \end{array} \begin{array}{c} \begin{array}{cccccc} \mathrm{R}_1 & \mathrm{R}_2 & \mathrm{R}_3 & \mathrm{R}_1^{\mathrm{ex}} & \mathrm{R}_2^{\mathrm{ex}} & \mathrm{R}_3^{\mathrm{ex}} \end{array} \\ \left(\begin{array}{ccc:ccc} -1 & 0 & 0 & 1 & 0 & 0 \\ 1 & -1 & -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & -1 & 0 \\ 0 & 0 & 1 & 0 & 0 & -1 \\ \hdashline & & & -1 & 0 & 0 \\ & \mathbf{0} & & 0 & 1 & 0 \\ & & & 0 & 0 & 1 \end{array}\right) \end{array}, \tag{1.11b}$$

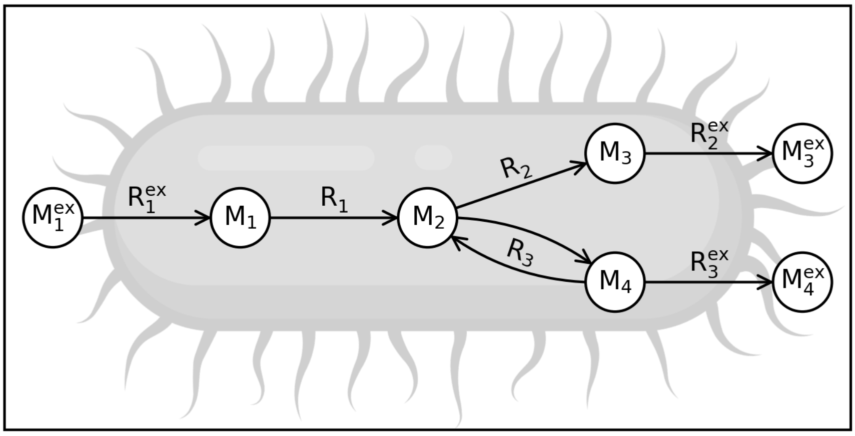


Figure 1.10.: Illustration of simple example network extended by external metabolites ($\mathrm{M}_i^{\mathrm{ex}}$).

where the horizontal dashed line indicates the border between internal (i.e., intracellular) and external metabolites and the vertical dashed line indicates the border between internal and exchange reactions. Critically for dFBA, we assume that, while internal metabolites ($\mathbf{S}^{\mathrm{in}}$) are in a steady state, external metabolites ($\mathbf{S}^{\mathrm{ex}}$), are not [30]. Additionally we can subdivide the flux vector ($\mathbf{v}$) in an internal (R_i) and exchange ($\mathrm{R}_i^{\mathrm{ex}}$) part,

$$\mathbf{S}\mathbf{v} = \begin{pmatrix} \mathbf{S}^{\mathrm{ss}} \\ \mathbf{S}^{\mathrm{dy}} \end{pmatrix} \begin{pmatrix} \mathbf{v}^{\mathrm{in}} \\ \mathbf{v}^{\mathrm{ex}} \end{pmatrix} = \begin{pmatrix} \mathbf{S}^{\mathrm{in}} & \\ \mathbf{0} & \mathbf{S}^{\mathrm{ex}} \end{pmatrix} \begin{pmatrix} \mathbf{v}^{\mathrm{in}} \\ \mathbf{v}^{\mathrm{ex}} \end{pmatrix} \tag{1.11c}$$

$$\mathbf{S}^{\mathrm{in}}\mathbf{v} = \mathbf{0} \tag{1.11d}$$

$$\mathbf{S}^{\mathrm{ex}}\mathbf{v}^{\mathrm{ex}} = \dot{\mathbf{z}}/X(t). \tag{1.11e}$$

In dynamic models, the state variable are normalized by volume, however, in FBA they are normalized by biomass. To translate the state variables from FBA to the dynamic normalization, one has to divide by the biomass concentration ($X(t)$) as done in Equation (1.11e) [1]. Subsequently FBA can be performed by optimization of the objective function

$$\text{maximize} \quad \mathbf{c}^{\mathrm{T}}\mathbf{v}. \tag{1.11f}$$

subject to

$$\begin{aligned} &\mathbf{S}\mathbf{v} = \mathbf{0} \\ &\mathbf{v}_{\mathrm{lb}} \leq \mathbf{v} \leq \mathbf{v}_{\mathrm{ub}} \end{aligned} \tag{1.11g}$$

and subject to K known dynamic exchange rates (Equation (1.11a)),

$$\mathbf{S}_K^{\mathrm{ex}}\mathbf{v}_K^{\mathrm{ex}} = \mathbf{f}_K(\mathbf{z}(t), t; \boldsymbol{\theta})/X(t). \tag{1.11h}$$

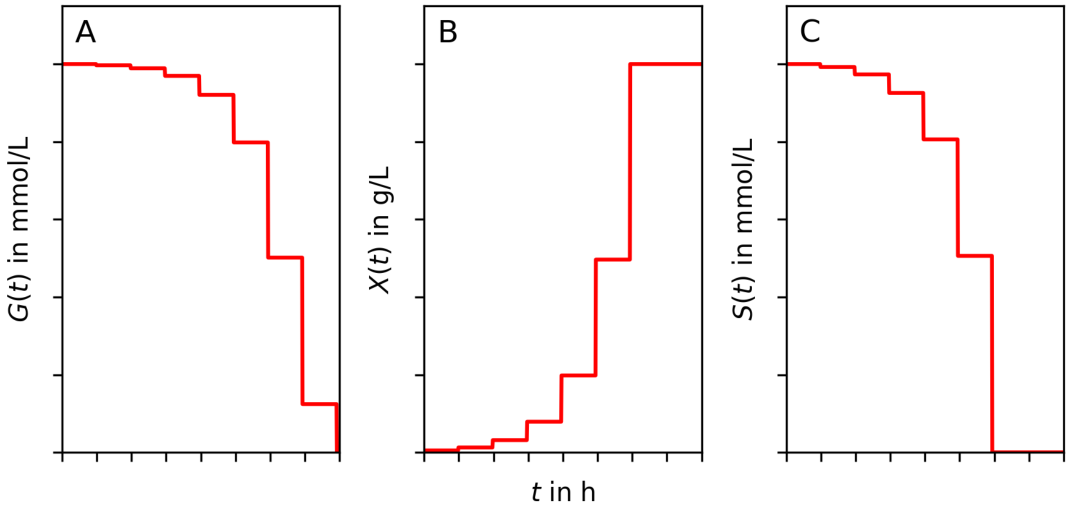


Figure 1.11.: dFBA simulation of a batch bioprocess, where the glucose uptake ($\dot{G}/X$) was calculated from a Michaelis Menten model and the rates for biomass growth ($\dot{X}/X$)and sulfate uptake ($\dot{S}/X$) were calculated with lexicographic FBA for every finite element. The lexicographic order was: 1. maximize biomass production, 2. minimize sulfate uptake. The rates and concentrations are assumed constant within one finite element of the simulation.

Typically, lexicographic [29] or parsimonious [56] FBA are performed as unique solutions to ($\mathbf{v}$) are required in dFBA [20]. The optimized values of $\mathbf{v}^{\text{ex}}$ are subsequently inserted into the differential equation of the dynamic model (Equation (1.11a)) as

$$\dot{\mathbf{z}} = \begin{pmatrix} \mathbf{f}_K(\mathbf{z}(t), t; \boldsymbol{\theta}) \\ \mathbf{S}_U^{\text{ex}} \mathbf{v}_U^{\text{ex}} X(t) \end{pmatrix} . \tag{1.11i}$$

Although the internal metabolites of the metabolic model are assumed to be in a steady state, the fluxes may change with time. Therefore, to calculated the change of concentration of external metabolites, step-wise numerical integrators (e.g., SciPy's `solve_ivp` [55]) are required [20, 30]. There, $\mathbf{S}^{\text{ex}}\mathbf{v}^{\text{ex}}$ is calculated with FBA at every finite element and subsequently used as an input for the integrators.

An example dFBA simulation is shown in Figure 1.11, where the concentration time series of glucose (G), biomass (X), and sulfate (S) in a batch bioprocess are plotted. While the time derivatives of glucose are calculated as Michaelis Menten kinetic, the rates of biomass growth and sulfate uptake were estimated via lexicographic FBA from the *E. coli* *i*ML1515 GSMM [9]. dFBA is a well established method that has shown to replicate biological phenomena, e.g., diauxic growth [30] or switching from aerobic to anaerobic respiration [29]. Moreover, as already implied in Figure 1.11, it is an useful tool to simulate medium component concentrations in bioreactors of bioprocess.

1.2.5. Process Optimization

Fed-batch processes are wide spread in the biotechnological industry, for example, for the production of pharmaceutical drugs. However, many processes are performed in sub-optimal conditions as classical design of experiment approaches involve trial and error and often are not able to cover the whole solution space. Improving sub-optimal processes is of high relevance as this could significantly reduce costs of production and subsequently lower drug prices for patients.

Dynamic metabolic models can be used to systematically sample the solution space for optimal process designs as computer simulations are cheap and comparatively easy to perform. They can provide suggestions for setting control variables, for example, the feed rate [57, 58] or the switch time of two-stage processes [59]. It is crucial to initially collect knowledge on the process of interest as several design choices need to be made for the construction of dynamic bioprocess models. A bioprocess engineer should consider following points:

Process Type Firstly, one has to decide which process type (batch, fed-batch, continuous, or other, Figure 1.7) to optimize. While fed-batch processes are currently the most popular in industry [48], batch or continuous processes may have additional advantages in certain settings. The choice of process type directly influences the properties of control variables to set. While in a fed-batch the feed rate can be constant or change over time, in a continuous process realistically only one dilution rate can be set. In contrast in a batch process the substrate uptake cannot be directly controlled at all.

Kinetic Functions Secondly, one has to find functions that model dynamic relationships of the state variables. For some variables these relationships are well explored, e.g., the Michaelis Menten kinetic for glucose uptake. However, for many product formation rates, finding realistic empirical relationships is a challenging task. In some cases Michaelis Menten-like formalae for product synthesis rates (π) are used, e.g.,

$$\pi(t) = \frac{\pi_{\max}\mu(t)}{k_\pi + \mu(t)} \quad [54] \tag{1.12a}$$

or

$$\pi(t) = \frac{\pi_{\max}\gamma(t)}{k_\pi + \gamma(t)} \quad [60]. \tag{1.12b}$$

Additionally, in some cases where inhibiting substances are present, inhibition terms ($k_I/(k_I + I)$) need to be added to the equation of product formation [60, 61]. The choice of kinetics describing state variables of interest and the quality of their parameterisation is essential for the predictive power of process models [62].

Model Type Thirdly, depending on the amount of kinetic information available, one can choose to either construct the dynamic process model entirely out of kinetic relationships [54, 59, 61] or employ dFBA which draws additional information from the incorporated stoichiometric matrix [29, 56, 63]. Originally, the disadvantage of dFBA was its slowness compared to kinetic models as multiple LPs have to be solved for each finite element. However recently, dFBA was implemented as an nonlinear optimal control problem in order to efficiently optimize process design [56].

Target Metric Typically in biotechnology, engineers aim to optimize their process in respect to one of the TRY metrics [64]. TRY stands for titer (i.e., concentration, $\mathrm{g\,L^{-1}}$), rate (i.e., productivity, $\mathrm{g\,h^{-1}}$), or yield ($\mathrm{g\,g^{-1}}$ product/substrate). In reality, one has to chose a target metric that fits their process of interest, depending on the exact product and its downstream processing. Therefore, even more target metrics have been employed, e.g., volumetric productivity ($\mathrm{g\,h^{-1}\,L^{-1}}$), specific yield ($\mathrm{g\,g^{-1}}$ product/ biomass), or even economically justified combinations of metrics [65].

1.3. Plasmid DNA Production

Together with small molecules and proteins, plasmid DNA (pDNA) is an important product of biotechnology [66]. The bulk of pDNA currently produced is either for transfection of mammalian cells for gene therapy [67] or for the production of mRNA vaccines [68]. Additionally, several DNA vaccines are in development to combat vexing diseases like HIV, HPV, Ebola, Zika and many more [69]. Meanwhile, a number of DNA vaccines have been approved worldwide for veterinary applications, for example against West Nile virus infection of horses [70] or melanoma in dogs [71]. The recent popularity of DNA vaccines is due to their many advantages, most importantly the ease of production, development and transport [66]. However, relatively high amounts of pDNA and are needed per vaccination shot and, therefore, the optimization of its production is of strong economic interest.

To increase pDNA production, several studies have been conducted. An overview of their results is depicted in Figure 1.12. Most studies used common laboratory *E. coli* strains as starting point for establishing potent production organisms [72]. Improvements were made by screening of favorable strains or via metabolic engineering through the introduction and knocking-out of genes to antibiotic-free selection systems and other highly optimized production strains [72]. Apart from maximizing the productivity three prerequisites are required for the design of pDNA production process.

1. pDNA can be present in an open circular, linear, or supercoild form. However, the supercoild form is of the most importance as it is generally considered more favorable for transfection of mammalian cells [73, 74].

2. Historically, a detrimental loss of plasmid during the production process was mitigated by the introduction of antibiotic resistance selection systems. However, using these systems comes with two main disadvantages. Firstly, cells are forced to shift metabolic resources from the production of pDNA to the production of antibiotic resistance proteins [75]. Secondly, antibiotic resistance genes may be detrimental for patients and, therefore, they have to be removed in the downstream processing. Studies showed that these disadvantages can be alleviated by renounce antibiotic selection systems all together [76].

3. Finally, European Medicines Agency (EMA) and U.S. Food and Drug Administration (FDA) require production with defined growth medium for pharmaceutical safety [66].

Conceptually, most strategies for increasing the pDNA production can be put into three categories: (1) reducing the growth rate, (2) ensuring a constant supply of precursor metabolites, and (3) optimising the plasmid itself. However, the exact method of how (1) or (2) is achieved differs widely between studies. In the following paragraphs I will present an overview of published approaches.

Growth Rate Reduction Studies on continuous cultures showed that a low growth rate increases the specific productivity of pDNA [98]. To achieve the same effect in batch fermentations, scientists designed a medium that releases glucose enzymatically and thus limits the glucose

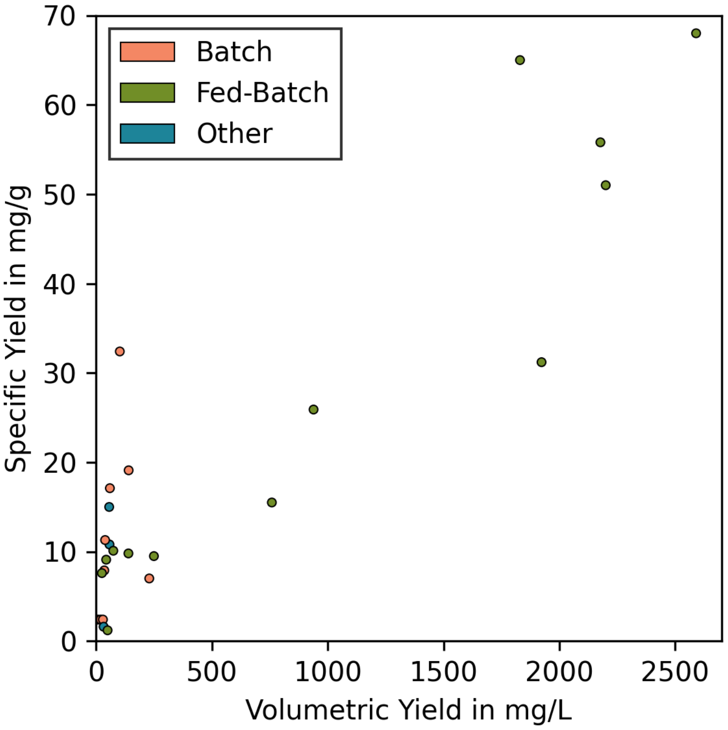


Figure 1.12.: Overview of reported volumetric and specific yields of pDNA production studies [50, 76–84, 84–93, 93–97].

uptake rate and subsequently growth of bacteria [78]. An alternative method for reduction of the glucose uptake rate was achieved by knocking-out the main uptake pathway of glucose increasing pDNA productivity [90]. Instead of limiting the carbon uptake, also a oxygen limitation proved to have a beneficial effect [79, 99].

Precursor Supplementation One example for ensuring constant pDNA precursor supplementation is the deletion of pyruvate kinase which forces metabolization of glucose over the pentose phosphate pathway [75, 86, 100]. Alternatively, scientists used of stoichiometric models to optimize the growth medium [85]. Moreover, one study emphasized the importance of aromatic amino acids (Phe, Tyr, Trp) in the medium for redirecting molecules to the nucleotide synthesis pathways [87]. Additionally, the effect of amount and type of nitrogen source in the growth medium have been shown to impact pDNA production [82]. As adding pDNA precursor molecules directly to the growth medium may be very costly, economical aspects need to be kept in mind [83].

Plasmid Optimization Further potential for optimization is the pDNA itself, for example, the reduction of its size since longer pDNA has been linked to lower volumetric yields [101]. Other approaches involved heat induced origins of replication that increase the plasmid copy number at higher temperatures than 37 °C [50, 96]. However, higher temperatures come with physiological trade-offs and therefore the amplitude and timing of heat induction is of importance [102].

1.4. Finger Sweat Analysis

As established in Section 1.2 ”Metabolic Models“, we can gain significant biological insight through measurements of biological systems with high-throughput analytical chemistry methods [1]. Here I focus on the analysis of metabolite measurements, which is aptly called metabolomics [103]. Typically, metabolomic data is obtained by analysis of biological material with liquid chromatography-mass spectrometry (LC-MS). The origin of the biological samples can be very diverse, for example, cell supernatants or cell lysates for single cell cultures [104]. However, for multicellular organisms like humans, biofluids like blood [105], urine [106], saliva [107], tears [108], or sweat [109] can be investigated.

The most classical approach to obtain metabolome information is the measurement of blood, which has been used for centuries. Blood takes a very prominent role in animal bio-fluids since it is present in all organs. It is, therefore, relatively simple to generalize from blood measurements to the whole organism [105]. Moreover, many drugs are injected into the blood stream which makes it an obvious choice to measure the metabolic behaviour of a drug. However, it has two major drawbacks. Firstly, blood not only contains many metabolites, but also proteins and (blood) cells in a high concentration. Since the latter two hamper metabolite measurements they have to be removed before LC-MS analysis [110]. Secondly, drawing blood is a cumbersome procedure for the patient and, additionally, in many cases requires qualified personnel which impairs measurement during real life settings [111].

Historically, body fluids other than blood were not easy to sample in large enough quantities for accurate analysis. This changed with the emergence of more and more sensitive LC-MS machines. With them, it is possible to qualitatively and quantitatively identify metabolites in concentrations as low as $\mathrm{pg\,\mu L^{-1}}$. This development rendered the analysis of metabolites in sweat feasible [111, 112]. Recent studies found that there is a plethora of metabolites to be found in sweat (Figure 1.13) and hypothesize that sweat is a promising matrix for biomarker discovery [113].

One advantage of finger sweat analysis compared to blood is the ease of sampling. There are different sampling protocols which can be as simple as holding a filter paper between fingertips which can easily done from home without trained personnel [111, 114, 115]. This facilitates the study of metabolic time courses within the real life settings of patients. A flow chart of finger sweat sampling is given in Figure 1.14. Moreover, finger sweat analysis has been of particular interest for forensic scientists for two reasons. Firstly, finger sweat can be sampled and analysed from finger prints after the sample donor already left the location [112]. Secondly it is possible to differentiate if a patient ingested a substance (e.g., a narcotic) or just came in touch with it by looking at the presence of metabolic degradation products of a substance of interest [112].

For forensic scientists often a qualitative answer to their questions is enough, however, for many medical applications a quantitative estimation of measured metabolites is of importance. In principle it is possible to get this information with LC-MS analysis since it can be set up as a quantitative method with calibration curves [103]. However, in combination with the previously mentioned sweat sampling techniques the recovery of quantitative data gets more complex. The reason for this is our inability to control the amount a patient sweats at a given time point. Finger tips have one of the highest densities of eccrine sweat glands on the human body [116]. The sweat flux, however, is highly variable, depending not only on interindividual differences, but also on temperature, humidity, exercise and further factors. Even with conservative estimates a variability of sweat flux on the finger tips between 0.05 and $1\,\mathrm{mg\,cm^{-2}\,min^{-1}}$ needs to be accounted for [116–120]. This variability has to be expected and measurements of sweat metabolites have to be corrected accordingly. The correction of finger sweat metabolome measurements for the respective sampled sweat volume is from now on referred to as normalization.

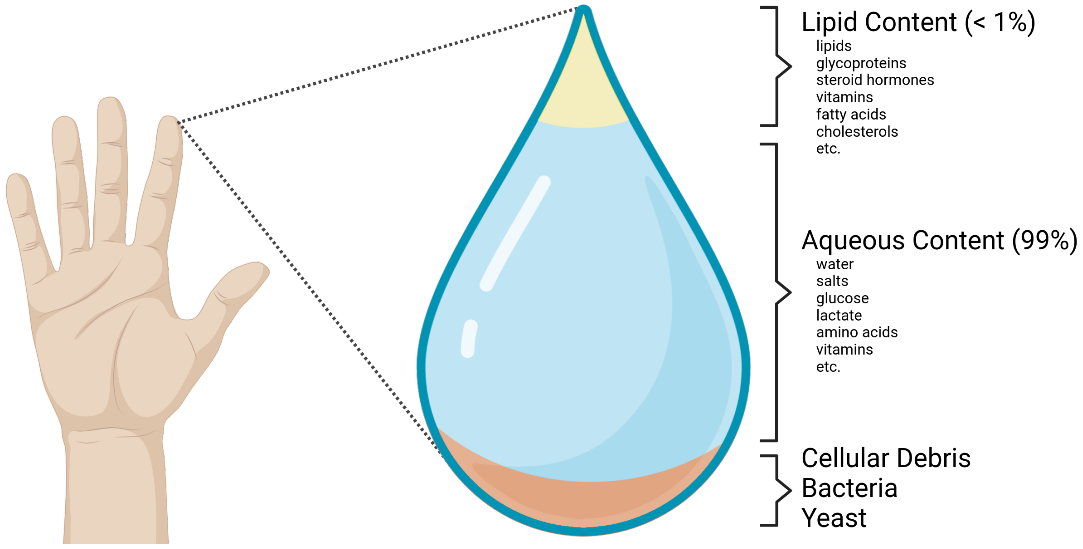


Figure 1.13.: (Finger) sweat is a valuable source of metabolites and potential biomarkers [113].

A method of normalization is to use capillary forces and microfluidics to get exact volumes of sweat already show to work in several studies [121, 122]. This, however, needs a relatively large sample area and more sophisticated sample methods. Therefore, it cannot easily be adopted for short interval finger sweat analysis. The problem for quantification of metabolite concentrations in the finger sweat has not been addressed up to date [112, 114].

1.4.1. Finger Sweat Models

As the structure of human sweat gland is very complex [123] sophisticated microfluidic models have been developed to simulate the concentration time series of metabolites during their secretion [124]. Moreover, the mode of partitioning (i.e., the passing of metabolites from the cytosol into the sweat) may change depending on the metabolite. While for many external metabolites (e.g., drugs or food compounds) a passive partitioning (via diffusion) is expected, for more common molecules in the cell active partitioning (e.g., Na^+ or Cl^-) or even active production by the sweat gland (e.g. lactate) are possible [124]. Therefore, while the concentration of some metabolites measured in sweat may correlate with concentrations in the blood, this must not mean that it can be generally assumed [124].

1.4.2. Size Effect Normalization Methods

Although, before this thesis, the post-measurement normalization of finger sweat has been overlooked, similar variability exist for other biofluid measurements. For example, in urine, the concentrations of metabolites heavily depends on the amount of liquid a patient drinks before the study. These variability is usually referred to as size effect, which need to be corrected with size effect normalization methods for (semi-)quantitative analysis [125]. Several methods for size effect normalization have been published [126].

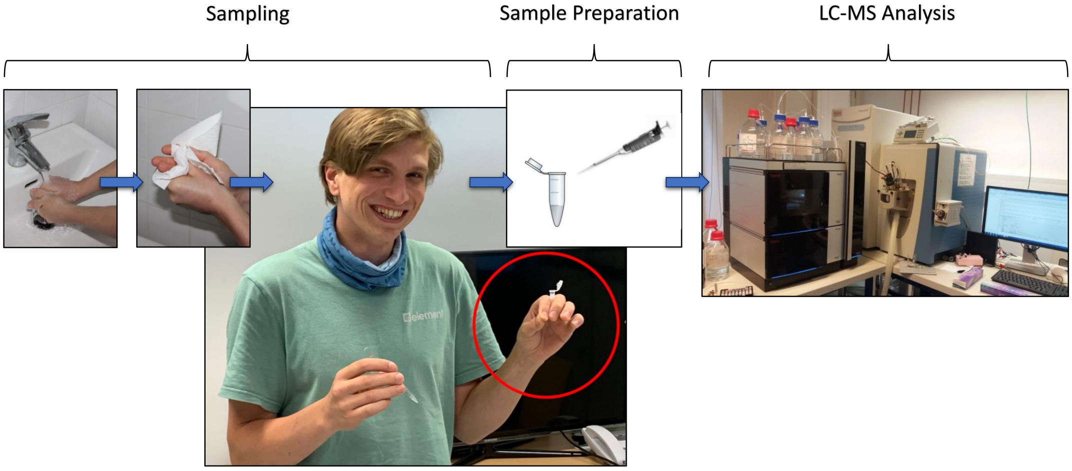


Figure 1.14.: Schematic figure of sweat sampling as described in [111]. First the study participants wash their hands, then they press their index finger and thumb on a filter paper for one minute. The filter paper can be collected and submitted to an analytical laboratory, where the metabolites in the filter paper are resuspended and measured with LC-MS [111].

Internal Standard Normalization One example is the internal standard normalization (IS). One can define one (or, theoretically, multiple) internal biomarker which concentration fluctuates little over time (e.g., creatinine in urine or blood) [126]. All measured metabolite abundances in one sample can then be divided (i.e., normalized) by the amount of the internal standard. IS has two critical disadvantages: (1) IS hinges on the assumption that fluctuations do not happen at all which is unlikely to be true [127]. (2) after normalization with IS, all measurements are biased by the measurement error of the internal standard [126].

Total Sum of Signal Normalization If information on internal standards are not available, one can try total sum of signal normalization (TSN) [125, 127]. With TSN all metabolite abundances in one sample are divided by the total sum of all metabolite abundances in one sample. TSN is also sometimes referred to as MSTUS [128]. Scientists have argued against TSN as the assumption that the sum of all metabolite abundances is constant over time is probably incorrect, e.g., when a patient has a meal we would expect higher total amounts of metabolites in the blood [129].

Probabilistc Quotient Normalization Alternatively, probabilistic quotient normalization (PQN) assumes that the average quotient of abundance of many metabolites between two samples is approximately 1 [129]. Comparative studies have found that PQN is a good option for size effect normalization [125, 128].

Many other normalization methods have been developed so far [125, 128, 130], however, they all work by dividing measured abundances by a non-physical normalization constant, thus rendering absolute quantification impossible.

2. Research Objectives

My research can be divided into two major projects: (1) finger sweat metabolome measurement normalization and (2) plasmid DNA production process optimization. The objectives of each project are described in the following paragraphs.

2.1. Finger Sweat Normalization

Measurements of the finger sweat metabolome bear large potential for clinical studies and personalized medicine, due to the plethora of metabolites present. To fully harness its potential, quantitative results on metabolite concentrations have to be obtained. This, however, remains very tricky, due to an ever-changing rate of sweating on the finger tips of patients. There are so many factors influencing the sweat rate (e.g., temperature, stress, food, drugs, disease, etc.) reported that it is virtually impossible to control for them all. Moreover, one cannot directly measure the finger sweat rate without drastically complicating the sampling procedure (Figure 1.14) which is one of the main advantages of finger sweat measurements to begin with. Therefore, the finger sweat metabolome data have to be normalized in respect to the sweat rate after measurements are performed. In my research, I tried to answer the question: can we find a suitable normalization strategy for finger sweat measurements to enable quantification?

2.2. Plasmid DNA Production Optimization

The industrial production is a key step in the manufacturing process of RNA or DNA vaccines or pharmaceuticals for gene therapy. Therefore, its production within EMA and FDA guidelines is essential. In this project, I used genome scale metabolic models of *E. coli* to comprehensively examine the addition and removal of various nutrients in the growth medium to answer the question: how can we optimize the growth medium to enhance the productivity of an industrial pDNA production system?

3. Results

This cumulative PhD thesis contains three scientific articles that have been published in international research journals.

3.1. Publication I: Finger sweat analysis enables short interval metabolic biomonitoring in humans

Julia Brunmair†, Mathias Gotsmy†, Laura Niederstaetter, Benjamin Neuditschko, Andrea Bileck, Astrid Slany, Max Lennart Feuerstein, Clemens Langbauer, Lukas Janker, Jürgen Zanghellini, Samuel M. Meier-Menches and Christopher Gerner. *Nature Communications*, 2021, 12(1), 5993.

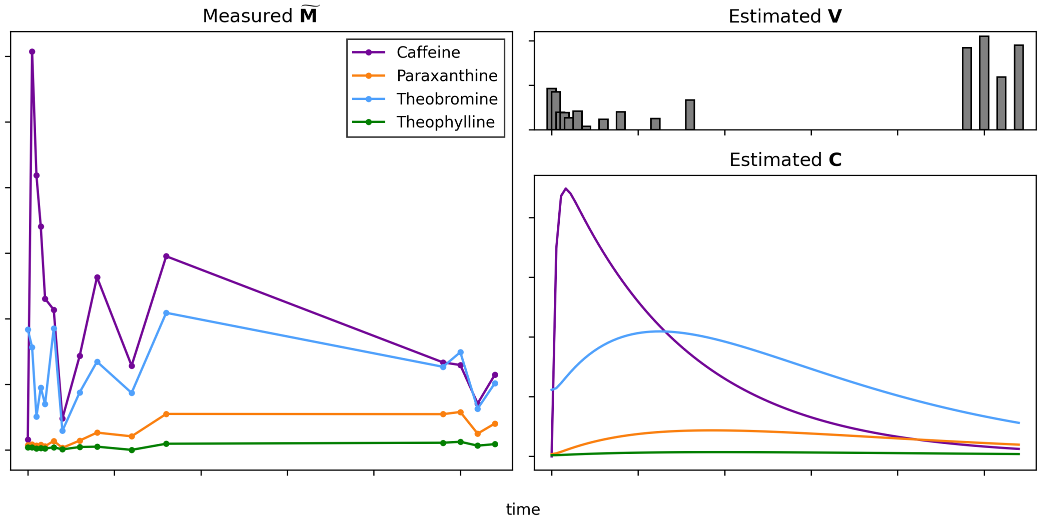


Figure 3.1.: Graphical summary of my first publication [131]. With our model we can estimate $\mathbf{C}(t)V(t)$ (right hand side panels) from the measured $\widetilde{\mathbf{M}}(t)$ (left hand side panel).

In this study we described a novel analytical method for personalized medicine. We analysed metabolites (e.g., potential biomarkers or xenobiotics) in the composition of sweat from the finger tips. In a case study, my experimental collaborators were able to detect caffeine and its degradation metabolites in the finger sweat of patients that previously ingested either coffee or purified caffeine. Although we were initially able to show great qualitative results, quantitative analysis was obstructed by an unknown sweat rate.

To address this problem, I developed a metabolic model comprised of caffeine and its three major degradation products, paraxanthine, theobromine, and theophylline. Their concentration over time can be described by a series of first order decays (i.e., a variation of the Bateman function, Section 1.2.3.2). Furthermore, I discovered that the measured amount of each molecule ($\widetilde{\mathbf{M}}$) at a time point (t) is equals to its concentration ($\mathbf{C}$) times the sweat volume (V),

$$\widetilde{\mathbf{M}}(t) = \mathbf{C}(t)V(t) \tag{3.1}$$

where, critically, $\widetilde{\mathbf{M}}$ and $\mathbf{C}$ are vectors, but V is a scalar variable. The resulting dynamic model contains eight kinetic constants describing $\mathbf{C}$ and $V(t)$ for every sample time point as parameters. As in this equation system the number of parameters is smaller than the number of measured points $\widetilde{\mathbf{M}}$, we were able to fit the experimental data ($\widetilde{\mathbf{M}}$) onto the model.

Thus, we estimated personalized kinetic first-order degradation constants as well as sweat volumes at every time point. An example for one measured finger sweat time series is given in Figure 3.1, where on the left panel the raw experimental data is plotted and on the right panels the estimated sweat volumes and concentrations can be seen.

Moreover, we were able to show that the kinetic constants are specific to individuals, and change little over time.

To summarize, in this publication we show that we can measure and quantitatively estimate biomarkers from finger sweat. Appendix A.1 includes the reprints from the original study published in *Nature Communications* [131].

3.2. Publication II: Probabilistic quotient's work and pharmacokinetics' contribution: countering size effect in metabolic time series measurements.

Mathias Gotsmy†, Julia Brunmair†, Christoph Büschl, Christopher Gerner, and Jürgen Zanghellini. *BMC Bioinformatics*, 2022, 23(1), 379.

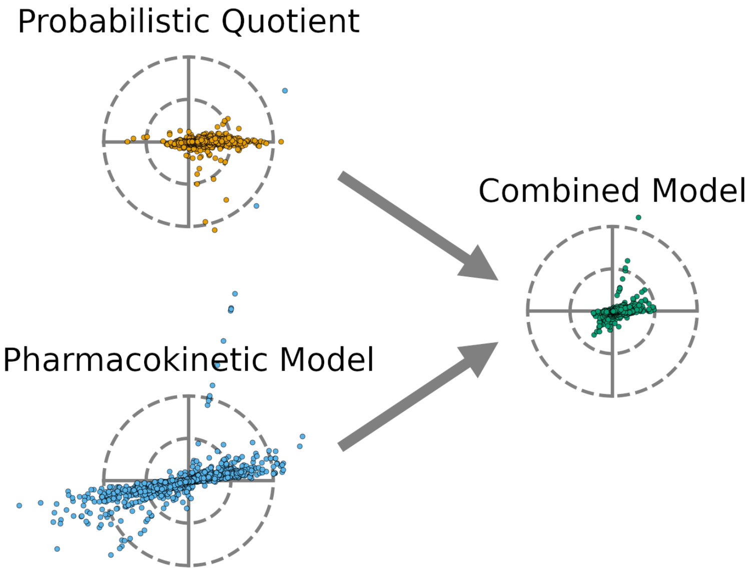


Figure 3.2.: Graphical summary of my second publication [132]. Combining PKM and PQN improves the goodness of normalization significantly.

In this publication we tried to improve on the normalization method we introduced in the previous paper [131] which is hereinafter referred to as PKM. Critically, PKM only relies on the (targeted) measurements of four metabolites in the sweat. However, as the finger sweat measurements are set up, many more untargeted compounds are detected which may hold potential information on the sweat rate and, thus, could improve the normalization.

Therefore, we combined a popular method for untargeted effect size normalization, namely PQN (Section 1.4.2), with our previously developed PKM. We simulated finger sweat measurement data and assessed the goodness of normalization of the different strategies. Moreover, I streamlined the coding pipeline for size effect normalization by compiling a new Python package `size_effect_normalization` which can be downloaded over GitHub [133].

With this novel method, were able to show that the combined model significantly outperformed standalone PQN and PKM. A representation of reduction of relative and absolute error by combining both models is shown in Figure 3.2. Moreover, we were able to show the our combined model is able to overcome high fractions of noise in the measurement data.

Finally, we concluded that the presented model further improves goodness of normalization, and is a important step towards quantification of metabolite concentrations in the finger sweat. Appendix A.2 includes the reprints from the original study published in *BMC Bioinformatics* [132].

3.3. Publication III: Sulfate limitation increases specific plasmid DNA yield and productivity in *E. coli* fed-batch processes.

Mathias Gotsmy, Florian Strobl, Florian Weiß, Petra Gruber, Barbara Kraus, Juergen Mairhofer, and Jürgen Zanghellini. *BMC Microbial Cell Factories*, 2023, 22(1), 242.

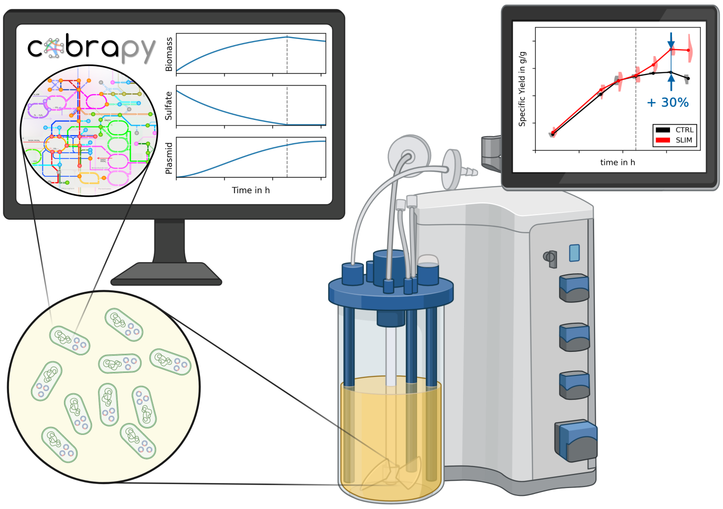


Figure 3.3.: Graphical abstract of my third publication [134]. With bioprocess simulations we were able to predict an optimized growth-decoupled fed-batch process that improved the specific pDNA yield by 33 %.

The regulatory standards for pDNA production in gene therapy are very strict, thus a defined medium has to be used during its production. However, defined media reduce the productivity. In this study, we tried to restore the productivity by improving the pDNA production process via optimization of a defined growth medium.

Counterintuitively, our simulations predicted that not the addition but the depletion of certain medium components can increase the productivity. This is due to a effect called growth-decoupling, where cells are artificially stopped in their growth and metabolic resources are freed up for production. Typically, growth-decoupling is triggered by inducible promoters that activate sophisticated regulatory processes that have to be genetically introduced into a host organism. However, FBA simulations we performed showed that this is is not necessary and the depletion of certain medium components can have the same effect. As a proof of concept, we chose sulfate from the list of potential decoupling components and developed a 3-stage fed-batch process (1. batch, 2. fed-batch with cell growth, 3. fed-batch without cell growth). Moreover, we predicted the optimal time point of switching with dFBA simulations (Figure 3.3).

Validation experiment results showed that growth-decoupling can be triggered by sulfate depletion as predicted. Moreover, our optimized 3-stage process significantly increased the specific pDNA yield by 33 % and the productivity by 13 %. Additionally, we showed that decoupling via sulfate limitation has potential to improve the performance of many biomolecule production processes. Appendix A.3 includes the reprints from the original study published in *BMC Microbial Cell Factories* [134].

4. Conclusion and Outlook

4.1. Finger Sweat Normalization

Comparative Study Studies have shown that for many metabolites there is significant correlation between their blood concentration and their abundance in (finger) sweat [123, and references therein]. However, this may not be true for all. It is known that for some endobiotic metabolites their concentration varies between blood and sweat [123, 124]. This may be due to active partition that occurs in the sweat glands, where, for example, ions are reabsorbed into the cells [124]. Unfortunately, we were not able to conduct a comparative study between finger sweat and blood concentrations of caffeine at this point, due to its organisational complexity (e.g., approval by the ethics committee is harder to obtain for blood studies compared to finger sweat). However, such a study would be immensely beneficial to further inform the normalization model and to increase the confidence in finger sweat measurements.

Improving Statistical Normalization In the second study on finger sweat normalization we combined our pharmacokinetic model with the statistical normalization method (PQN). However, as size effect normalization is not a problem unique to our group, there are many bright minds that try to develop more sophisticated normalization strategies, e.g., NOREVA [135]. Historically, statistical normalization methods were not informed by time, which may be beneficial when normalizing time-series data. However, recently NOREVA was updated explicitly to include time-series data in their normalization strategy [136]. I coded my `size_effect_normalization` Python package in a way that makes it very easy to switch the statistical normalization part from PQN to another one [133]. In a future study, it may be able to further improve the normalization procedure by including NOREVA (or other methods) into the model. A sensitivity analysis could give insight which statistical normalization approach is most suitable for finger sweat data normalization.

Personalized Medicine By modeling the amount of metabolite measured in the finger sweat with PKM [131] as well as combined PKM and PQN models [132], we were able to estimate personalized kinetic constants of xenobiotic degradation and elimination. For example, the compound of our case study, caffeine, is metabolized by the cytochrome P450 enzyme CYP1A2 [137]. Moreover, a large number of popular drugs is metabolized by the same family of enzymes [46]. As we found long lasting individual differences in the kinetic constants of degradation of caffeine [131], there exist also significant differences in the speed of metabolization of drugs [46]. For example, 20 % of the Asian population is considered a poor metabolizer of drugs dependent on CYP2C19 [46]. As illustrated in Figure 4.1, this can lead to tremendous differences in intracellular concentrations of affected drugs in patients. Unfortunately, in many cases individual characteristics in metabolization speed are not tested for, however, with the easy and cheap method of finger sweat sampling, this may change in the future.

Clinical Studies To receive approval from authorities like FDA and EMA, pharmacokinetic/pharmacological data of a new drug have to be measured [138]. Important parameters for clinicians

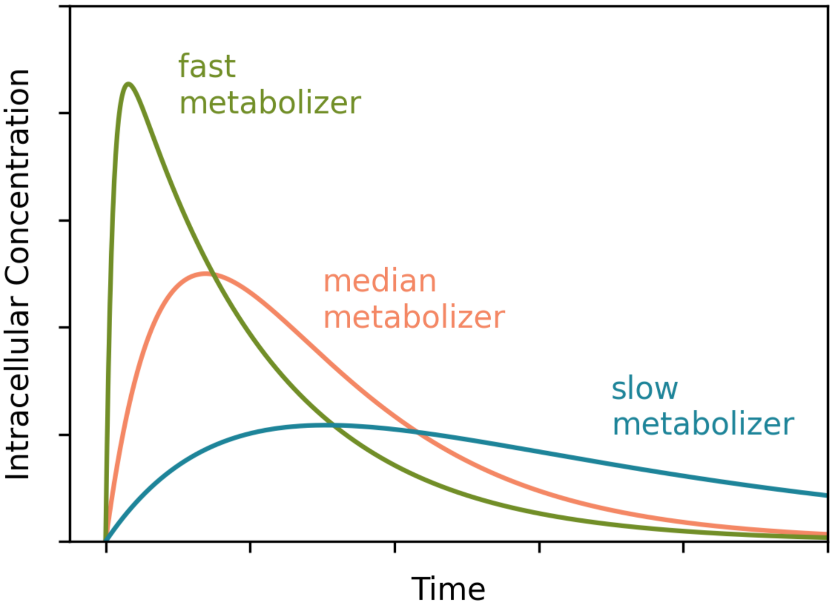


Figure 4.1.: Significant fractions of the population are fast or slow metabolizers of common drugs. The speed of metabolization (i.e., k_e in Equation 1.7d) can have a critical impact of intracellular concentration of a drug [46]. Note that only k_e was varied here.

are the maximum concentration after application and the rate of elimination [139]. After further refinement and validation of the normalisation strategy, we argue that we will be able to support or in some cases even replace cumbersome measurements of blood for clinical studies. This would lead to a significant increase in the quality of life for clinical study participants. This may especially be feasible for phase 3 clinical studies, where basic pharmacological parameters are already known and a big cohort needs to be tested [140]. Furthermore, finger sweat sample could be conducted more frequently which would allow more precise insights into the modes of action of a drug in the human body.

Conclusion To summarize, I was able to develop a specialized normalization method for finger sweat metabolome data that uses pharmacokinetic as well as statistical information. This method allows the elevate finger sweat metabolomics from a purely qualitative to a quantitative method.

4.2. Plasmid DNA Production Optimization

Ease of Implementation Historically, biotechnological production improvements were made by modifying internal reactions of a production organism. This can, for example, be achieved by knocking-out (e.g., OptKnock [141]) or modulating the size of fluxes (e.g., MoVE [142]). The disadvantage of these methods is, however, (1) that they require cumbersome genetic engineering of the production strains. Moreover, (2) unless genetic controls are introduced (e.g., replacing

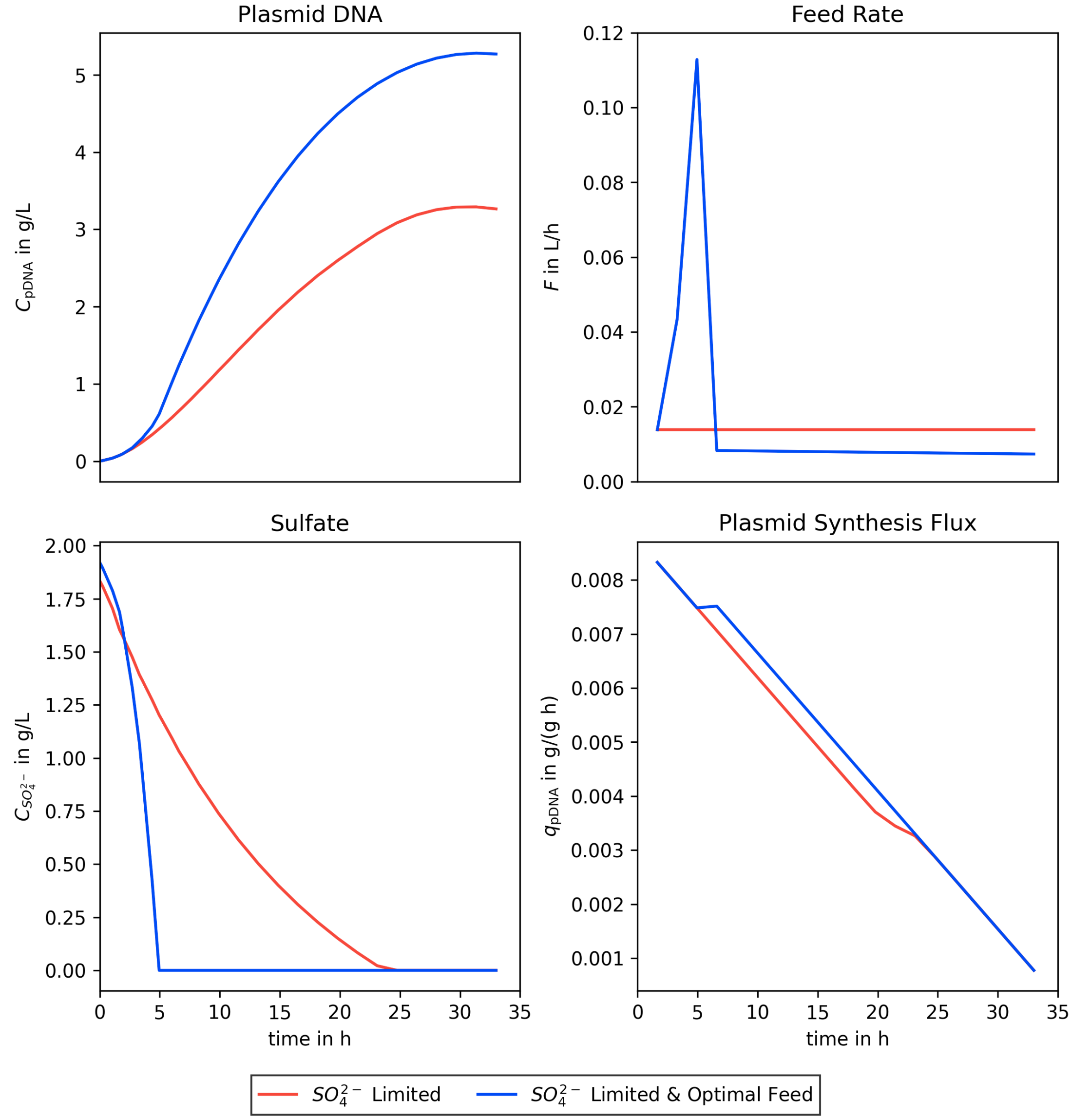


Figure 4.2.: Preliminary results for optimizing the initial sulfate concentration as well as the feed rate (blue) compared to optimal process from publication 3 (red). The method is adapted from the implementation of dFBA as an optimal control problem [56]. Clearly, several challenges are unsolved, e.g., what are realistic upper and lower bounds of the feed rate? or how long can the starvation phase realistically be?

promoters with inducible promoters [143], or RNA interference [144]), these changes are static, i.e., the cannot be easily changed throughout a production process. Finally, (3) all genetic modifications to the production strain may be very specific to one product, so much of the work done for strain design is not easily transferable to other production processes.

In my third publication on plasmid production, I demonstrate the effectiveness of sulfate starvation. In future applications, it may overcome all three disadvantages [134]. (1) no genetic changes in the production organism have to be performed, only the sulfate concentration in the growth medium is optimized. (2) theoretically it should be possible to restore a growth by adding sulfate after a starvation phase. This could be used to effectively extend production processes as shown in a Master's thesis of our group [145]. However, experimental validation of this theory has not yet been conducted. (3) given that no sulfate is required in the biosynthesis pathway of a biomolecule, sulfate starvation as a method for process optimization should in theory be applicable to any biomolecule [134]. Therefore, we think that this method can be easily adapted to many other processes.

Biological Explanation In my publication, I assume that the production rate of pDNA increases upon starvation of sulfate. This assumption has been validated. However, we do not know how exactly internal flux distributions change. Moreover, we know that some byproducts like acetate are produced. However, simulations showed that not all of the excess glucose taken up is converted to acetate right away. Therefore, insight into intracellular biological changes during the starvation phase warrants further investigations.

Feed Rate Optimization When thinking of process optimization from a process control point of view there are two control variables that are easy to set: (1) the initial concentration of sulfate (i.e., the time point of sulfate starvation) and (2) the feed rate. Therefore, parallel optimization of (1) and (2) is a logical step to further increase productivity in a follow-up study. Figure 4.2 shows preliminary results of such an optimisation. However, the increase in control variables that can be set makes the process design significantly more delicate. On one hand, optimizers (here, Ipopt [146]) become less stable, and more attention has to be given to, e.g., parameter initialization. On the other hand, one has to ensure that realistic biological boundaries are defined, which is easier said than done, as there is often little information (e.g., on the interplay between feed rate and production rate) available.

Other Decoupling Agents In publication 3 we selected sulfate as our decoupling agent of choice from a list of potential candidates. Therefore, another follow-up study could investigate how different decoupling agents (e.g., magnesium, potassium or trace element ions) would affect the productivity. Studies that focused on the internal metabolites found significant differences in ATP concentrations depending on the decoupling agents [147]. Moreover, alternative decoupling agents could be used to improve the production of biomolecules that contain sulfur and thus cannot be targeted with our previously published sulfate limitation strategy.

Conclusion To summarize, I was able to obtain valuable predictions on the metabolic fluxes with and without the presence of sulfate and, subsequently, could use this information for process optimization. My experimental collaborators were able to validate the predictions I made. I think that sulfate limitation is a promising idea, that can and will be applied to several processes in the future.

5. Bibliography

[1] Steffen Klamt, Oliver Hädicke, and Axel von Kamp. Stoichiometric and constraint-based analysis of biochemical reaction networks. *Large-scale networks in engineering and life sciences*, pages 263–316, 2014.

[2] Markus W Covert, Christophe H Schilling, Iman Famili, Jeremy S Edwards, Igor I Goryanin, Evgeni Selkov, and Bernhard O Palsson. Metabolic modeling of microbial strains in silico. *Trends in biochemical sciences*, 26(3):179–186, 2001.

[3] Ali Khodayari, Ali R Zomorrodi, James C Liao, and Costas D Maranas. A kinetic model of escherichia coli core metabolism satisfying multiple sets of mutant flux data. *Metabolic engineering*, 25:50–62, 2014.

[4] Elizabeth Brunk, Swagatika Sahoo, Daniel C Zielinski, Ali Altunkaya, Andreas Dräger, Nathan Mih, Francesco Gatto, Avlant Nilsson, German Andres Preciat Gonzalez, Maike Kathrin Aurich, et al. Recon3d enables a three-dimensional view of gene variation in human metabolism. *Nature biotechnology*, 36(3): 272–281, 2018.

[5] Hiroaki Kitano. Systems biology: a brief overview. *science*, 295(5560):1662–1664, 2002.

[6] Aihua Zhang, Hui Sun, Ping Wang, Ying Han, and Xijun Wang. Recent and potential developments of biofluid analyses in metabolomics. *Journal of proteomics*, 75(4):1079–1088, 2012.

[7] Minoru Kanehisa. The kegg database. In *'In silico'simulation of biological processes: Novartis Foundation Symposium 247*, volume 247, pages 91–103. Wiley Online Library, 2002.

[8] Peter D Karp, Richard Billington, Ron Caspi, Carol A Fulcher, Mario Latendresse, Anamika Kothari, Ingrid M Keseler, Markus Krummenacker, Peter E Midford, Quang Ong, et al. The biocyc collection of microbial genomes and metabolic pathways. *Briefings in bioinformatics*, 20(4):1085–1093, 2019.

[9] Jonathan M Monk, Colton J Lloyd, Elizabeth Brunk, Nathan Mih, Anand Sastry, Zachary King, Rikiya Takeuchi, Wataru Nomura, Zhen Zhang, Hirotada Mori, et al. i ml1515, a knowledgebase that computes escherichia coli traits. *Nature biotechnology*, 35(10):904–908, 2017.

[10] Bernhard Ø Palsson. *Systems biology: properties of reconstructed networks*. Cambridge university press, 2006.

[11] Jeffrey D Orth, Ines Thiele, and Bernhard Ø Palsson. What is flux balance analysis? *Nature biotechnology*, 28(3):245–248, 2010.

[12] Agustín Gabriel Yabo, Andrea de Martino, Andrea Weisse, Andreas Kremling, Anne Goelzer, Benjamin Mauroy, Christophe Goupil, Cyril Karamaoun, Daan de Groot, Dafni Giannari, David Lacoste, David Tourigny, Diana Széliová, Diego A. Oyarzun, Elad Noor, Elena Pascual Garcia, Eric Herbert, Felipe Scott, Frédérique Noël, Gabriele Micali, Hadrien Delattre, Herbert Sauro, Hidde De jong, Hollie J. Hindley, Hugo Dourado, Jacopo Grilli, Marcelo Rivas-Astroza, Marco Cosentino Lagomarsino, Markus Köbi, Mattia Corigliano, Meike Wortel, Ohad Golan, Olivier Rivoire, Orkun S Soyer, Pranas Grigaitis, Robert West, Steffen Waldherr, and Wolfram Liebermeister. *Economic Principles in Cell Biology*. The Economic Cell Collective, July 2023. doi: 10.5281/zenodo.8156386. URL https://hal.inrae.fr/hal-04172118.

[13] Diana Széliová, Dmytro Iurashev, David E Ruckerbauer, Gunda Koellensperger, Nicole Borth, Michael Melcher, and Jürgen Zanghellini. Error propagation in constraint-based modeling of chinese hamster ovary cells. *Biotechnology Journal*, 2020.

[14] Hong Qian and Daniel A Beard. Thermodynamics of stoichiometric biochemical networks in living systems far from equilibrium. *Biophysical chemistry*, 114(2-3):213–220, 2005.

[15] Sarah M Keating, Dagmar Waltemath, Matthias König, Fengkai Zhang, Andreas Dräger, Claudine Chaouiya, Frank T Bergmann, Andrew Finney, Colin S Gillespie, Tomáš Helikar, et al. Sbml level 3: an extensible format for the exchange and reuse of biological models. *Molecular systems biology*, 16(8):e9110, 2020.

[16] Laurent Heirendt, Sylvain Arreckx, Thomas Pfau, Sebastián N Mendoza, Anne Richelle, Almut Heinken, Hulda S Haraldsdóttir, Jacek Wachowiak, Sarah M Keating, Vanja Vlasov, et al. Creation and analysis of biochemical constraint-based models using the cobra toolbox v. 3.0. *Nature protocols*, 14(3):639–702, 2019.

[17] Ali Ebrahim, Joshua A Lerman, Bernhard O Palsson, and Daniel R Hyduke. Cobrapy: constraints-based reconstruction and analysis for python. *BMC systems biology*, 7:1–6, 2013.

[18] Sven Thiele, Axel von Kamp, Pavlos Stephanos Bekiaris, Philipp Schneider, and Steffen Klamt. Cnapy: a cellnetanalyzer gui in python for analyzing and designing metabolic networks. *Bioinformatics*, 38(5): 1467–1469, 2022.

[19] Bianca Buchner, Tom J Clement, Daan H de Groot, and Jürgen Zanghellini. ecmtool: fast and memory-efficient enumeration of elementary conversion modes. *Bioinformatics*, 39(3):btad095, 2023.

[20] Jose A Gomez, Kai Höffner, and Paul I Barton. Dfbalab: a fast and reliable matlab code for dynamic flux balance analysis. *BMC bioinformatics*, 15(1):1–10, 2014.

[21] Zachary A King, Justin Lu, Andreas Dräger, Philip Miller, Stephen Federowicz, Joshua A Lerman, Ali Ebrahim, Bernhard O Palsson, and Nathan E Lewis. Bigg models: A platform for integrating, standardizing and sharing genome-scale models. *Nucleic acids research*, 44(D1):D515–D522, 2016.

[22] Rahuman S Malik-Sheriff, Mihai Glont, Tung VN Nguyen, Krishna Tiwari, Matthew G Roberts, Ashley Xavier, Manh T Vu, Jinghao Men, Matthieu Maire, Sarubini Kananathan, et al. Biomodels—15 years of sharing computational models in life science. *Nucleic acids research*, 48(D1):D407–D415, 2020.

[23] Andrew Makhorin. The gnu linear programming kit (glpk). GNU Software Foundation, 2000. URL `https://www.gnu.org/software/glpk/glpk.html`.

[24] IBM ILOG Cplex. V12. 1: User's manual for cplex. *International Business Machines Corporation*, 46(53): 157, 2009.

[25] Gurobi Optimization, LLC. Gurobi Optimizer Reference Manual, 2023. URL `https://www.gurobi.com`.

[26] Steven M Kelk, Brett G Olivier, Leen Stougie, and Frank J Bruggeman. Optimal flux spaces of genome-scale stoichiometric models are determined by a few subnetworks. *Scientific reports*, 2(1):580, 2012.

[27] Jeremy S Edwards, Rafael U Ibarra, and Bernhard O Palsson. In silico predictions of escherichia coli metabolic capabilities are consistent with experimental data. *Nature biotechnology*, 19(2):125–130, 2001.

[28] Nathan E Lewis, Kim K Hixson, Tom M Conrad, Joshua A Lerman, Pep Charusanti, Ashoka D Polpitiya, Joshua N Adkins, Gunnar Schramm, Samuel O Purvine, Daniel Lopez-Ferrer, et al. Omic data from evolved e. coli are consistent with computed optimal growth from genome-scale models. *Molecular systems biology*, 6(1):390, 2010.

[29] Kai Höffner, Stuart M Harwood, and Paul I Barton. A reliable simulator for dynamic flux balance analysis. *Biotechnology and bioengineering*, 110(3):792–802, 2013.

[30] Radhakrishnan Mahadevan, Jeremy S Edwards, and Francis J Doyle III. Dynamic flux balance analysis of diauxic growth in escherichia coli. *Biophysical journal*, 83(3):1331–1340, 2002.

[31] Benjamín J Sánchez, Cheng Zhang, Avlant Nilsson, Petri-Jaan Lahtvee, Eduard J Kerkhoven, and Jens Nielsen. Improving the phenotype predictions of a yeast genome-scale metabolic model by incorporating enzymatic constraints. *Molecular systems biology*, 13(8):935, 2017.

[32] Kai Zhuang, Goutham N Vemuri, and Radhakrishnan Mahadevan. Economics of membrane occupancy and respiro-fermentation. *Molecular systems biology*, 7(1):500, 2011.

[33] Leonor Michaelis, Maud L Menten, et al. Die kinetik der invertinwirkung. *Biochem. z*, 49(333-369):352, 1913.

[34] Harry Bateman. The solution of a system of differential equations occurring in the theory of radioactive transformations. In *Proc. Cambridge Philos. Soc.*, volume 15, pages 423–427, 1910.

[35] Ali Khodayari and Costas D Maranas. A genome-scale escherichia coli kinetic metabolic model k-ecoli457 satisfying flux data for multiple mutant strains. *Nature communications*, 7(1):13806, 2016.

[36] Athel Cornish-Bowden. One hundred years of michaelis–menten kinetics. *Perspectives in Science*, 4:3–9, 2015.

[37] Richard B Brandt, Jerome E Laux, and Steven W Yates. Calculation of inhibitor ki and inhibitor type from the concentration of inhibitor for 50% inhibition for michaelis-menten enzymes. *Biochemical medicine and metabolic biology*, 37(3):344–349, 1987.

[38] Roy M Daniel, Michael J Danson, Robert Eisenthal, Charles K Lee, and Michelle E Peterson. The effect of temperature on enzyme activity: new insights and their implications. *Extremophiles*, 12:51–59, 2008.

[39] E R Garrett. The Bateman function revisited: a critical reevaluation of the quantitative expressions to characterize concentrations in the one compartment body model as a function of time with first-order invasion and first-order elimination. *Journal of pharmacokinetics and biopharmaceutics*, 22(2):103–128, 1994.

[40] Monika Gajewska, Alicia Paini, JV Sala Benito, Julien Burton, Andrew Worth, Chiara Urani, Heiko Briesen, and K-W Schramm. In vitro-to-in vivo correlation of the skin penetration, liver clearance and hepatotoxicity of caffeine. *Food and Chemical Toxicology*, 75:39–49, 2015.

[41] Christine A Haller, Peyton Jacob III, and Neal L Benowitz. Enhanced stimulant and metabolic effects of combined ephedrine and caffeine. *Clinical Pharmacology & Therapeutics*, 75(4):259–273, 2004.

[42] M Richelle, I Tavazzi, M Enslen, and EA Offord. Plasma kinetics in man of epicatechin from black chocolate. *European journal of clinical nutrition*, 53(1):22–26, 1999.

[43] Cathy K Gelotte, Brenda A Zimmerman, and Gary A Thompson. Single-dose pharmacokinetic study of diphenhydramine hcl in children and adolescents. *Clinical pharmacology in drug development*, 7(4):400–407, 2018.

[44] Karan Agrawal, Rémy Bosviel, Brian D Piccolo, and John W Newman. Oral ibuprofen differentially affects plasma and sweat lipid mediator profiles in healthy adult males. *Prostaglandins & other lipid mediators*, 137:1–8, 2018.

[45] Ana P Pérez-Acevedo, Roberta Pacifici, Giulio Mannocchi, Massimo Gottardi, Lourdes Poyatos, Esther Papaseit, Clara Pérez-Mañá, Soraya Martin, Francesco P Busardò, Simona Pichini, et al. Disposition of cannabinoids and their metabolites in serum, oral fluid, sweat patch and urine from healthy individuals treated with pharmaceutical preparations of medical cannabis. *Phytotherapy Research*, 35(3):1646–1657, 2021.

[46] Tom Lynch and AMY Price. The effect of cytochrome p450 metabolism on drug response, interactions, and adverse effects. *American family physician*, 76(3):391–396, 2007.

[47] Satish Babu Kaki, A Naga Prasad, Anjani Devi Chintagunta, Vijaya Ramu Dirisala, NS Sampath Kumar, SJK Naidu, and B Ramesh. Industrial scale production of recombinant human insulin using escherichia coli bl-21. *Iranian Journal of Science and Technology, Transactions A: Science*, 46(2):373–383, 2022.

[48] Henry C Lim and Hwa Sung Shin. *Fed-batch cultures: principles and applications of semi-batch bioreactors*. Cambridge University Press, 2013.

[49] Sibel Öztürk, Pinar Calik, and Tunçer H Özdamar. Fed-batch biomolecule production by bacillus subtilis: a state of the art review. *Trends in biotechnology*, 34(4):329, 2016.

[50] Aaron E Carnes, Jeremy M Luke, Justin M Vincent, Angela Schukar, Sheryl Anderson, Clague P Hodgson, and James A Williams. Plasmid dna fermentation strain and process-specific effects on vector yield, quality, and transgene expression. *Biotechnology and bioengineering*, 108(2):354–363, 2011.

[51] DJ Korz, U Rinas, K Hellmuth, EA Sanders, and W-D Deckwer. Simple fed-batch technique for high cell density cultivation of escherichia coli. *Journal of biotechnology*, 39(1):59–65, 1995.

[52] S John Pirt. Maintenance energy: a general model for energy-limited and energy-sufficient growth. *Archives of microbiology*, 133:300–302, 1982.

[53] Aaron Meurer, Christopher P. Smith, Mateusz Paprocki, Ondřej Čertík, Sergey B. Kirpichev, Matthew Rocklin, AMiT Kumar, Sergiu Ivanov, Jason K. Moore, Sartaj Singh, Thilina Rathnayake, Sean Vig, Brian E. Granger, Richard P. Muller, Francesco Bonazzi, Harsh Gupta, Shivam Vats, Fredrik Johansson, Fabian Pedregosa, Matthew J. Curry, Andy R. Terrel, Štěpán Roučka, Ashutosh Saboo, Isuru Fernando, Sumith Kulal, Robert Cimrman, and Anthony Scopatz. Sympy: symbolic computing in python. *PeerJ Computer Science*, 3:e103, January 2017. ISSN 2376-5992. doi: 10.7717/peerj-cs.103. URL `https://doi.org/10.7717/peerj-cs.103`.

[54] Michael Maurer, Manfred Kühleitner, Brigitte Gasser, and Diethard Mattanovich. Versatile modeling and optimization of fed batch processes for the production of secreted heterologous proteins with pichia pastoris. *Microbial cell factories*, 5(1):1–10, 2006.

[55] Pauli Virtanen, Ralf Gommers, Travis E Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, et al. Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature methods*, 17(3):261–272, 2020.

[56] Rafael D de Oliveira, Galo AC Le Roux, and Radhakrishnan Mahadevan. Nonlinear programming reformulation of dynamic flux balance analysis models. *Computers & Chemical Engineering*, 170:108101, 2023.

[57] Sebastián Espinel-Ríos, Bruno Morabito, Johannes Pohlodek, Katja Bettenbrock, Steffen Klamt, and Rolf Findeisen. Towards a modeling, optimization and predictive control framework for fed-batch metabolic cybergenetics. *arXiv preprint arXiv:2302.02177*, 2023.

[58] JM Modak, HC Lim, and YJ Tayeb. General characteristics of optimal feed rate profiles for various fed-batch fermentation processes. *Biotechnology and bioengineering*, 28(9):1396–1407, 1986.

[59] Steffen Klamt, Radhakrishnan Mahadevan, and Oliver Hädicke. When do two-stage processes outperform one-stage processes? *Biotechnology journal*, 13(2):1700539, 2018.

[60] Julian Kager, Johanna Bartlechner, Christoph Herwig, and Stefan Jakubek. Direct control of recombinant protein production rates in e. coli fed-batch processes by nonlinear feedback linearization. *Chemical Engineering Research and Design*, 182:290–304, 2022.

[61] Marta B Lopes, Gabriel Martins, and Cecília RC Calado. Kinetic modeling of plasmid bioproduction in escherichia coli dh5α cultures over different carbon-source compositions. *Journal of Biotechnology*, 186: 38–48, 2014.

[62] B Bayer, B Sissolak, M Duerkop, M Von Stosch, and G Striedner. The shortcomings of accurate rate estimations in cultivation processes and a solution for precise and robust process modeling. *Bioprocess and biosystems engineering*, 43:169–178, 2020.

[63] Kaushik Raj, Naveen Venayak, and Radhakrishnan Mahadevan. Novel two-stage processes for optimal chemical production in microbes. *Metabolic Engineering*, 62:186–197, 2020.

[64] Jens Nielsen and Jay D Keasling. Engineering cellular metabolism. *Cell*, 164(6):1185–1197, 2016.

[65] Kai Zhuang, Laurence Yang, William R Cluett, and Radhakrishnan Mahadevan. Dynamic strain scanning optimization: an efficient strain design strategy for balanced yield, titer, and productivity. dyssco strategy for strain design. *BMC biotechnology*, 13:1–15, 2013.

[66] Juergen Mairhofer and Alvaro R Lara. Advances in host and vector development for the production of plasmid dna vaccines. In *Cancer Vaccines*, pages 505–541. Springer, 2014.

[67] Emilia Sokołowska and Agnieszka Urszula Błachnio-Zabielska. A critical review of electroporation as a plasmid delivery system in mouse skeletal muscle. *International journal of molecular sciences*, 20(11):2776, 2019.

[68] European Medicines Agency. Comirnaty assessment report. `https://www.ema.europa.eu/en/documents/assessment-report/comirnaty-epar-public-assessment-report_en.pdf`, 2020. [Online; accessed 28-July-2022].

[69] Ebony N Gary and David B Weiner. Dna vaccines: prime time is now. *Current opinion in immunology*, 65: 21–27, 2020.

[70] F Ghaffarifar. Plasmid dna vaccines: where are we now. *Drugs Today*, 54(5):315–333, 2018.

[71] MacKenzie A Pellin. The use of oncept melanoma vaccine in veterinary patients: A review of the literature. *Veterinary Sciences*, 9(11):597, 2022.

[72] Diana M Bower and Kristala LJ Prather. Engineering of bacterial strains and vectors for the production of plasmid dna. *Applied microbiology and biotechnology*, 82(5):805–813, 2009.

[73] J-Y Cherng, NME Schuurmans-Nieuwenbroek, W Jiskoot, H Talsma, NJ Zuidam, WE Hennink, and DJA Crommelin. Effect of dna topology on the transfection efficiency of poly ((2-dimethylamino) ethyl methacrylate)–plasmid complexes. *Journal of controlled release*, 60(2-3):343–353, 1999.

[74] Lionel Cupillard, Véronique Juillard, Sophie Latour, Guy Colombet, N Cachet, S Richard, S Blanchard, and Laurent Fischer. Impact of plasmid supercoiling on the efficacy of a rabies dna vaccine to protect cats. *Vaccine*, 23(16):1910–1916, 2005.

[75] Drew S Cunningham, Richard R Koepsel, Mohammad M Ataai, and Michael M Domach. Factors affecting plasmid production in escherichia coli from a resource allocation standpoint. *Microbial cell factories*, 8(1): 1–17, 2009.

[76] Jürgen Mairhofer, Monika Cserjan-Puschmann, Gerald Striedner, Katharina Nöbauer, Ebrahim Razzazi-Fazeli, and Reingard Grabherr. Marker-free plasmids for gene therapeutic applications—lack of antibiotic resistance gene substantially improves the manufacturing process. *Journal of biotechnology*, 146(3):130–137, 2010.

[77] Zhi-Nan Xu, Wen-he Shen, Hao Chen, and Pei-lin Cen. Effects of medium composition on the production of plasmid dna vector potentially for human gene therapy. *Journal of Zhejiang University. Science. B*, 6(5): 396, 2005.

[78] Janet Galindo, Blanca L Barrón, and Alvaro R Lara. Improved production of large plasmid dna by enzyme-controlled glucose release. *Annals of Microbiology*, 66(3):1337–1342, 2016.

[79] Karim E Jaén, Daniela Velazquez, Frank Delvigne, Juan-Carlos Sigala, and Alvaro R Lara. Engineering e. coli for improved microaerobic pdna production. *Bioprocess and biosystems engineering*, 42(9):1457–1466, 2019.

[80] Tania E Pablos, René Soto, Eugenio Meza Mora, Sylvie Le Borgne, Octavio T Ramírez, Guillermo Gosset, and Alvaro R Lara. Enhanced production of plasmid dna by engineered escherichia coli strains. *Journal of biotechnology*, 158(4):211–214, 2012.

[81] José T Cortés, Noemí Flores, Francisco Bolívar, Alvaro R Lara, and Octavio T Ramírez. Physiological effects of ph gradients on escherichia coli during plasmid dna production. *Biotechnology and bioengineering*, 113(3):598–611, 2016.

[82] Fabiola Islas-Lugo, Jesus Vega-Estrada, Christian Ariel Alvis, Jaime Ortega-Lopez, and Maria del Carmen Montes-Horcasitas. Developing strategies to increase plasmid dna production in escherichia coli dh5α using batch culture. *Journal of biotechnology*, 233:66–73, 2016.

[83] Michael K Danquah and Gareth M Forde. Growth medium selection and its economic impact on plasmid dna production. *Journal of bioscience and bioengineering*, 104(6):490–497, 2007.

[84] Alison Dorward, Ronan D O'Kennedy, Olusegun Folarin, John M Ward, and Eli Keshavarz-Moore. The role of amino acids in the amplification and quality of dna vectors for industrial applications. *Biotechnology Progress*, 35(6):e2883, 2019.

[85] Zhijun Wang, Guowei Le, Yonghui Shi, and Grzegorz Wegrzyn. Medium design for plasmid dna production based on stoichiometric model. *Process Biochemistry*, 36(11):1085–1093, 2001.

[86] Geisa AL Gonçalves, Duarte MF Prazeres, Gabriel A Monteiro, and Kristala LJ Prather. De novo creation of mg1655-derived e. coli strains specifically designed for plasmid dna production. *Applied microbiology and biotechnology*, 97(2):611–620, 2013.

[87] LM Martins, AQ Pedro, D Oppolzer, F Sousa, JA Queiroz, and LA Passarinha. Enhanced biosynthesis of plasmid dna from escherichia coli vh33 using box–behnken design associated to aromatic amino acids pathway. *Biochemical Engineering Journal*, 98:117–126, 2015.

[88] Aurora García-Rendón, Rodolfo Munguía-Soto, Rosa M Montesinos-Cisneros, Roberto Guzman, and Armando Tejeda-Mansir. Performance analysis of exponential-fed perfusion cultures for pdna vaccines production. *Journal of Chemical Technology & Biotechnology*, 92(2):342–349, 2017.

[89] Mitzi de la Cruz, Elisa A Ramírez, Juan-Carlos Sigala, José Utrilla, and Alvaro R Lara. Plasmid dna production in proteome-reduced escherichia coli. *Microorganisms*, 8(9):1444, 2020.

[90] René Soto, Luis Caspeta, Blanca Barrón, Guillermo Gosset, Octavio T Ramírez, and Alvaro R Lara. High cell-density cultivation in batch mode for plasmid dna production by a metabolically engineered e. coli strain with minimized overflow metabolism. *Biochemical engineering journal*, 56(3):165–171, 2011.

[91] Ronan D O'Kennedy, John M Ward, and Eli Keshavarz-Moore. Effects of fermentation strategy on the characteristics of plasmid dna production. *Biotechnology and applied biochemistry*, 37(1):83–90, 2003.

[92] Kevin O'Mahony, Ruth Freitag, Frank Hilbrig, Patrick Müller, and Ivo Schumacher. Strategies for high titre plasmid dna production in escherichia coli dh5α. *Process Biochemistry*, 42(7):1039–1049, 2007.

[93] Geisa AL Goncalves, Kristala LJ Prather, Gabriel A Monteiro, Aaron E Carnes, and Duarte MF Prazeres. Plasmid dna production with escherichia coli galg20, a pgi-gene knockout strain: Fermentation strategies and impact on downstream processing. *Journal of biotechnology*, 186:119–127, 2014.

[94] Je-Nie Phue, Sang Jun Lee, Loc Trinh, and Joseph Shiloach. Modified escherichia coli b (bl21), a superior producer of plasmid dna compared with escherichia coli k (dh5α). *Biotechnology and bioengineering*, 101(4): 831–836, 2008.

[95] Fernando Grijalva-Hernández, Jesús Vega-Estrada, Montserrat Escobar-Rosales, Jaime Ortega-López, Ricardo Aguilar-López, Alvaro R Lara, and Ma del Carmen Montes-Horcasitas. High kanamycin concentration as another stress factor additional to temperature to increase pdna production in e. coli dh5α batch and fed-batch cultures. *Microorganisms*, 7(12):711, 2019.

[96] James A Williams, Jeremy Luke, Sarah Langtry, Sheryl Anderson, Clague P Hodgson, and Aaron E Carnes. Generic plasmid dna production platform incorporating low metabolic burden seed-stock and fed-batch fermentation processes. *Biotechnology and bioengineering*, 103(6):1129–1143, 2009.

[97] Jared Nelson, Stephen Rodriguez, Neil Finlayson, Jim Williams, and Aaron Carnes. Antibiotic-free production of a herpes simplex virus 2 dna vaccine in a high yield cgmp process. *Human Vaccines & Immunotherapeutics*, 9(10):2211–2215, 2013.

[98] Martin Wunderlich, Hilal Taymaz-Nikerel, Guillermo Gosset, Octavio T Ramírez, and Alvaro R Lara. Effect of growth rate on plasmid dna production and metabolic performance of engineered escherichia coli strains. *Journal of bioscience and bioengineering*, 117(3):336–342, 2014.

[99] Alvaro R Lara, Karim E Jaén, Olusegun Folarin, Eli Keshavarz-Moore, and Jochen Büchs. Effect of the oxygen transfer rate on oxygen-limited production of plasmid dna by escherichia coli. *Biochemical Engineering Journal*, 150:107303, 2019.

[100] Drew S Cunningham, Zhu Liu, Nathan Domagalski, Richard R Koepsel, Mohammad M Ataai, and Michael M Domach. Pyruvate kinase-deficient escherichia coli exhibits increased plasmid copy number and cyclic amp levels. *Journal of bacteriology*, 191(9):3041–3049, 2009.

[101] Alison Kay, Ronan O'Kennedy, John Ward, and Eli Keshavarz-Moore. Impact of plasmid size on cellular oxygen demand in escherichia coli. *Biotechnology and applied biochemistry*, 38(1):1–7, 2003.

[102] Karim E Jaén, Alvaro R Lara, and Octavio T Ramírez. Effect of heating rate on pdna production by e. coli. *Biochemical engineering journal*, 79:230–238, 2013.

[103] Xiaojing Liu and Jason W Locasale. Metabolomics: a primer. *Trends in biochemical sciences*, 42(4):274–284, 2017.

[104] Anna Artati, Cornelia Prehn, and Jerzy Adamski. Lc-ms/ms-based metabolomics for cell cultures. *Cell-Based Assays Using iPSCs for Drug Development and Testing*, pages 119–130, 2019.

[105] Natalie J Serkova, Theodore J Standiford, and Kathleen A Stringer. The emerging field of quantitative blood metabolomics for biomarker discovery in critical illnesses. *American journal of respiratory and critical care medicine*, 184(6):647–655, 2011.

[106] Aihua Zhang, Hui Sun, Xiuhong Wu, and Xijun Wang. Urine metabolomics. *Clinica Chimica Acta*, 414: 65–69, 2012.

[107] Aihua Zhang, Hui Sun, and Xijun Wang. Saliva metabolomics opens door to biomarker discovery, disease diagnosis, and treatment. *Applied biochemistry and biotechnology*, 168:1718–1727, 2012.

[108] Julia Brunmair, Andrea Bileck, Doreen Schmidl, Gerhard Hagn, Samuel M Meier-Menches, Nikolaus Hommer, Andreas Schlatter, Christopher Gerner, and Gerhard Garhöfer. Metabolic phenotyping of tear fluid as a prognostic tool for personalised medicine exemplified by t2dm patients. *EPMA Journal*, 13(1): 107–123, 2022.

[109] Sean W Harshman, Andrew B Browder, Christina N Davidson, Rhonda L Pitsch, Kraig E Strayer, Nicole M Schaeublin, Mandy S Phelps, Maegan L O'Connor, Nicholas S Mackowski, Kristyn N Barrett, et al. The impact of nutritional supplementation on sweat metabolomic content: a proof-of-concept study. *Frontiers in chemistry*, 9:255, 2021.

[110] Stephen J Bruce, Isabelle Tavazzi, Véronique Parisod, Serge Rezzi, Sunil Kochhar, and Philippe A Guy. Investigation of human blood plasma sample preparation for performing metabolomics using ultrahigh performance liquid chromatography/mass spectrometry. *Analytical chemistry*, 81(9):3285–3296, 2009.

[111] Julia Brunmair, Andrea Bileck, Thomas Stimpfl, Florian Raible, Giorgia Del Favero, Samuel M Meier-Menches, and Christopher Gerner. Metabo-tip: a metabolomics platform for lifestyle monitoring supporting the development of novel strategies in predictive, preventive and personalised medicine. *EPMA Journal*, 12 (2):141–153, 2021.

[112] Min Jang, Catia Costa, J Bunch, B Gibson, M Ismail, Vladimir Palitsin, Rebecca Webb, M Hudson, and MJ Bailey. On the relevance of cocaine detection in a fingerprint. *Scientific reports*, 10(1):1–7, 2020.

[113] Joy N Hussain, Nitin Mantri, and Marc M Cohen. Working up a good sweat–the challenges of standardising sweat collection for metabolomics analysis. *The Clinical Biochemist Reviews*, 38(1):13, 2017.

[114] Joanna Czerwinska, Min Jang, Catia Costa, Mark C Parkin, Claire George, Andrew T Kicman, Melanie J Bailey, Paul I Dargan, and Vincenzo Abbate. Detection of mephedrone and its metabolites in fingerprints from a controlled human administration study by liquid chromatography-tandem mass spectrometry and paper spray-mass spectrometry. *Analyst*, 145(8):3038–3048, 2020.

[115] Kenji Kuwayama, Tadashi Yamamuro, Kenji Tsujikawa, Hajime Miyaguchi, Tatsuyuki Kanamori, Yuko T Iwata, and Hiroyuki Inoue. Time-course measurements of drugs and metabolites transferred from fingertips after drug administration: usefulness of fingerprints for drug testing. *Forensic Toxicology*, 32(2):235–242, 2014.

[116] Nigel AS Taylor and Christiano A Machado-Moreira. Regional variations in transepidermal water loss, eccrine sweat gland density, sweat secretion rates and electrolyte composition in resting and exercising humans. *Extreme physiology & medicine*, 2(1):4, 2013.

[117] Hideo Ando and Ryo Noguchi. Dependence of palmar sweating response and central nervous system activity on the frequency of whole-body vibration. *Scandinavian journal of work, environment & health*, pages 216–219, 2003.

[118] Christiano A Machado-Moreira, Joanne N Caldwell, Igor B Mekjavic, and Nigel AS Taylor. Sweat secretion from palmar and dorsal surfaces of the hands during passive and active heating. *Aviation, space, and environmental medicine*, 79(11):1034–1040, 2008.

[119] Hnin Yin Yin Nyein, Mallika Bariya, Brandon Tran, Christine Heera Ahn, Brenden Janatpour Brown, Wenbo Ji, Noelle Davis, and Ali Javey. A wearable patch for continuous analysis of thermoregulatory sweat at rest. *Nature communications*, 12(1):1–13, 2021.

[120] Bowen Zhong, Kai Jiang, Lili Wang, and Guozhen Shen. Wearable sweat loss measuring devices: From the role of sweat loss to advanced mechanisms and designs. *Advanced Science*, page 2103257, 2021.

[121] Jungil Choi, Amay J Bandodkar, Jonathan T Reeder, Tyler R Ray, Amelia Turnquist, Sung Bong Kim, Nathaniel Nyberg, Aurélie Hourlier-Fargette, Jeffrey B Model, Alexander J Aranyosi, et al. Soft, skin-integrated multifunctional microfluidic systems for accurate colorimetric analysis of sweat biomarkers and temperature. *ACS sensors*, 4(2):379–388, 2019.

[122] Sung Bong Kim, Jahyun Koo, Jangryeol Yoon, Aurélie Hourlier-Fargette, Boram Lee, Shulin Chen, Seongbin Jo, Jungil Choi, Yong Suk Oh, Geumbee Lee, et al. Soft, skin-interfaced microfluidic systems with integrated enzymatic assays for measuring the concentration of ammonia and ethanol in sweat. *Lab on a Chip*, 20(1): 84–92, 2020.

[123] Lindsay B Baker. Physiology of sweat gland function: The roles of sweating and sweat composition in human health. *Temperature*, 6(3):211–259, 2019.

[124] Z Sonner, E Wilder, J Heikenfeld, G Kasting, F Beyette, D Swaile, F Sherman, J Joyce, J Hagen, N Kelley-Loughnane, et al. The microfluidics of the eccrine sweat gland, including biomarker partitioning, transport, and biosensing implications. *Biomicrofluidics*, 9(3):031301, 2015.

[125] P Filzmoser and B Walczak. What can go wrong at the data normalization step for identification of biomarkers? *Journal of Chromatography A*, 1362:194–205, 2014.

[126] Yiman Wu and Liang Li. Sample normalization methods in quantitative metabolomics. *Journal of Chromatography A*, 1430:80–95, 2016.

[127] Bethanne M Warrack, Serhiy Hnatyshyn, Karl-Heinz Ott, Michael D Reily, Mark Sanders, Haiying Zhang, and Dieter M Drexler. Normalization strategies for metabonomic analysis of urine samples. *Journal of Chromatography B*, 877(5-6):547–552, 2009.

[128] Bo Li, Jing Tang, Qingxia Yang, Xuejiao Cui, Shuang Li, Sijie Chen, Quanxing Cao, Weiwei Xue, Na Chen, and Feng Zhu. Performance evaluation and online realization of data-driven normalization methods used in lc/ms based untargeted metabolomics analysis. *Scientific reports*, 6(1):1–13, 2016.

[129] Frank Dieterle, Alfred Ross, Götz Schlotterbeck, and Hans Senn. Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. application in 1h nmr metabonomics. *Analytical chemistry*, 78(13):4281–4290, 2006.

[130] Yuliya V Karpievitch, Sonja B Nikolic, Richard Wilson, James E Sharman, and Lindsay M Edwards. Metabolomics data normalization with eigenms. *PloS one*, 9(12):e116221, 2014.

[131] Julia Brunmair, Mathias Gotsmy, Laura Niederstaetter, Benjamin Neuditschko, Andrea Bileck, Astrid Slany, Max Lennart Feuerstein, Clemens Langbauer, Lukas Janker, Jürgen Zanghellini, et al. Finger sweat analysis enables short interval metabolic biomonitoring in humans. *Nature Communications*, 12(1):5993, 2021.

[132] Mathias Gotsmy, Julia Brunmair, Christoph Büschl, Christopher Gerner, and Jürgen Zanghellini. Probabilistic quotient's work and pharmacokinetics' contribution: countering size effect in metabolic time series measurements. *BMC bioinformatics*, 23(1):379, 2022.

[133] Mathias Gotsmy. Size effect normalization python package. `https://github.com/Gotsmy/sweat_normalization`, 2022.

[134] Mathias Gotsmy, Florian Strobl, Florian Weiss, Petra Gruber, Barbara Kraus, Juergen Mairhofer, and Juergen Zanghellini. Sulfate limitation increases specific plasmid dna yield in e. coli fed-batch processes. *BMC Microbial Cell Factories*, 22(1):242, 2023.

[135] Qingxia Yang, Yunxia Wang, Ying Zhang, Fengcheng Li, Weiqi Xia, Ying Zhou, Yunqing Qiu, Honglin Li, and Feng Zhu. Noreva: enhanced normalization and evaluation of time-course and multi-class metabolomic data. *Nucleic Acids Research*, 48(W1):W436–W448, 2020.

[136] Jianbo Fu, Ying Zhang, Yunxia Wang, Hongning Zhang, Jin Liu, Jing Tang, Qingxia Yang, Huaicheng Sun, Wenqi Qiu, Yinghui Ma, et al. Optimization of metabolomic data processing using noreva. *Nature Protocols*, 17(1):129–151, 2022.

[137] Marilyn C Cornelis, Tim Kacprowski, Cristina Menni, Stefan Gustafsson, Edward Pivin, Jerzy Adamski, Anna Artati, Chin B Eap, Georg Ehret, Nele Friedrich, et al. Genome-wide association study of caffeine metabolites provides new insights to caffeine metabolism and dietary caffeine-consumption behavior. *Human molecular genetics*, 25(24):5472–5482, 2016.

[138] Anselm Jorda and Markus Zeitlinger. Preclinical pharmacokinetic/pharmacodynamic studies and clinical trials in the drug development process of ema-approved antibacterial agents: a review. *Clinical Pharmacokinetics*, 59:1071–1084, 2020.

[139] Jonathan A Bernstein. Azelastine hydrochloride: a review of pharmacology, pharmacokinetics, clinical efficacy and tolerability. *Current medical research and opinion*, 23(10):2441–2452, 2007.

[140] Tom Brody. Chapter 2 - clinical trial design. In Tom Brody, editor, *Clinical Trials (Second Edition)*, pages 31–68. Academic Press, Boston, second edition edition, 2016. ISBN 978-0-12-804217-5. doi: https://doi.org/10.1016/B978-0-12-804217-5.00002-3. URL `https://www.sciencedirect.com/science/article/pii/B9780128042175000023`.

[141] Anthony P Burgard, Priti Pharkya, and Costas D Maranas. Optknock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. *Biotechnology and bioengineering*, 84(6):647–657, 2003.

[142] Naveen Venayak, Axel von Kamp, Steffen Klamt, and Radhakrishnan Mahadevan. MoVE identifies metabolic valves to switch between phenotypic states. *Nature Communications*, 9(1):5332, 2018. doi: 10.1038/s41467-018-07719-4.

[143] Song Jiao, Xu Li, Huimin Yu, Huan Yang, Xue Li, and Zhongyao Shen. In situ enhancement of surfactin biosynthesis in bacillus subtilis using novel artificial inducible promoters. *Biotechnology and Bioengineering*, 114(4):832–842, 2017.

[144] Mattia Matasci, David L Hacker, Lucia Baldi, and Florian M Wurm. Recombinant therapeutic protein production in cultivated mammalian cells: current status and future prospects. *Drug discovery today: technologies*, 5(2-3):e37–e42, 2008.

[145] Guido Schlögel. *Designing an optimal operation profile of a fed-batch process with induced growth arrest.* Fachhochschule Wiener Neustadt, 2022.

[146] Andreas Wächter and Lorenz T Biegler. On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. *Mathematical programming*, 106:25–57, 2006.

[147] Kento Tokuyama, Yoshihiro Toya, Fumio Matsuda, Brady F Cress, Mattheos AG Koffas, and Hiroshi Shimizu. Magnesium starvation improves production of malonyl-coa-derived metabolites in escherichia coli. *Metabolic engineering*, 52:215–223, 2019.

A. Appendix I: Publications

A.1. Publication I: Finger sweat analysis enables short interval metabolic biomonitoring in humans

Julia Brunmair†, Mathias Gotsmy†, Laura Niederstaetter, Benjamin Neuditschko, Andrea Bileck, Astrid Slany, Max Lennart Feuerstein, Clemens Langbauer, Lukas Janker, Jürgen Zanghellini, Samuel M. Meier-Menches and Christopher Gerner

My role was shared first author. I conceived the idea of pharmacokinetic finger sweat normalization, implemented the code, drafted sections on mathematical modeling, and compiled Figure 5.

nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-021-26245-4 OPEN

Finger sweat analysis enables short interval metabolic biomonitoring in humans

Julia Brunmair[1,4], Mathias Gotsmy[1,4], Laura Niederstaetter[1], Benjamin Neuditschko[1,2], Andrea Bileck[1,3], Astrid Slany[1], Max Lennart Feuerstein[1], Clemens Langbauer[1], Lukas Janker[1,3], Jürgen Zanghellini[1], Samuel M. Meier-Menches[1,2,3] & Christopher Gerner[1,3]✉

Metabolic biomonitoring in humans is typically based on the sampling of blood, plasma or urine. Although established in the clinical routine, these sampling procedures are often associated with a variety of compliance issues, which are impeding time-course studies. Here, we show that the metabolic profiling of the minute amounts of sweat sampled from fingertips addresses this challenge. Sweat sampling from fingertips is non-invasive, robust and can be accomplished repeatedly by untrained personnel. The sweat matrix represents a rich source for metabolic phenotyping. We confirm the feasibility of short interval sampling of sweat from the fingertips in time-course studies involving the consumption of coffee or the ingestion of a caffeine capsule after a fasting interval, in which we successfully monitor all known caffeine metabolites as well as endogenous metabolic responses. Fluctuations in the rate of sweat production are accounted for by mathematical modelling to reveal individual rates of caffeine uptake, metabolism and clearance. To conclude, metabotyping using sweat from fingertips combined with mathematical network modelling shows promise for broad applications in precision medicine by enabling the assessment of dynamic metabolic patterns, which may overcome the limitations of purely compositional biomarkers.

[1] Department of Analytical Chemistry, Faculty of Chemistry, University of Vienna, Vienna, Austria. [2] Department of Inorganic Chemistry, Faculty of Chemistry, University of Vienna, Vienna, Austria. [3] Joint Metabolome Facility, University and Medical University of Vienna, Vienna, Austria. [4] These authors contributed equally: Julia Brunmair, Mathias Gotsmy. ✉email: christopher.gerner@univie.ac.at

Metabolic phenotyping seeks to identify biomarkers for diagnosis, prognosis or therapy and holds great promise to improve clinical practice and especially, precision medicine[1,2]. Despite considerable progress with respect to the sensitive and parallel analysis of metabolites in metabolomics/metabonomics studies[3–7] and by mass spectrometry (MS)[8,9], the successful implementation of metabolites as biomarkers in the clinical setting still represents a major challenge[10–12]. This is illustrated by the strong individual and physiological background variability[2] and individual differences in ADME properties, the latter impacting significantly on drug responses[13,14]. To the best of our knowledge, current techniques of metabolic phenotyping are largely focussed on generating static diagnostic pictures because the commonly used biological fluids (e.g. plasma, urine)[15–17] or tissues do not routinely allow for time-course studies. The implementation of dynamic metabolic responses as a biomarker strategy may be desirable, but requires a considerable number of data points on a single individual. Clearly, a non-invasive method from an alternative biological fluid is required to enable frequent sampling of the same individual in order to obtain dynamic metabolic patterns in the frame of metabolic phenotyping.

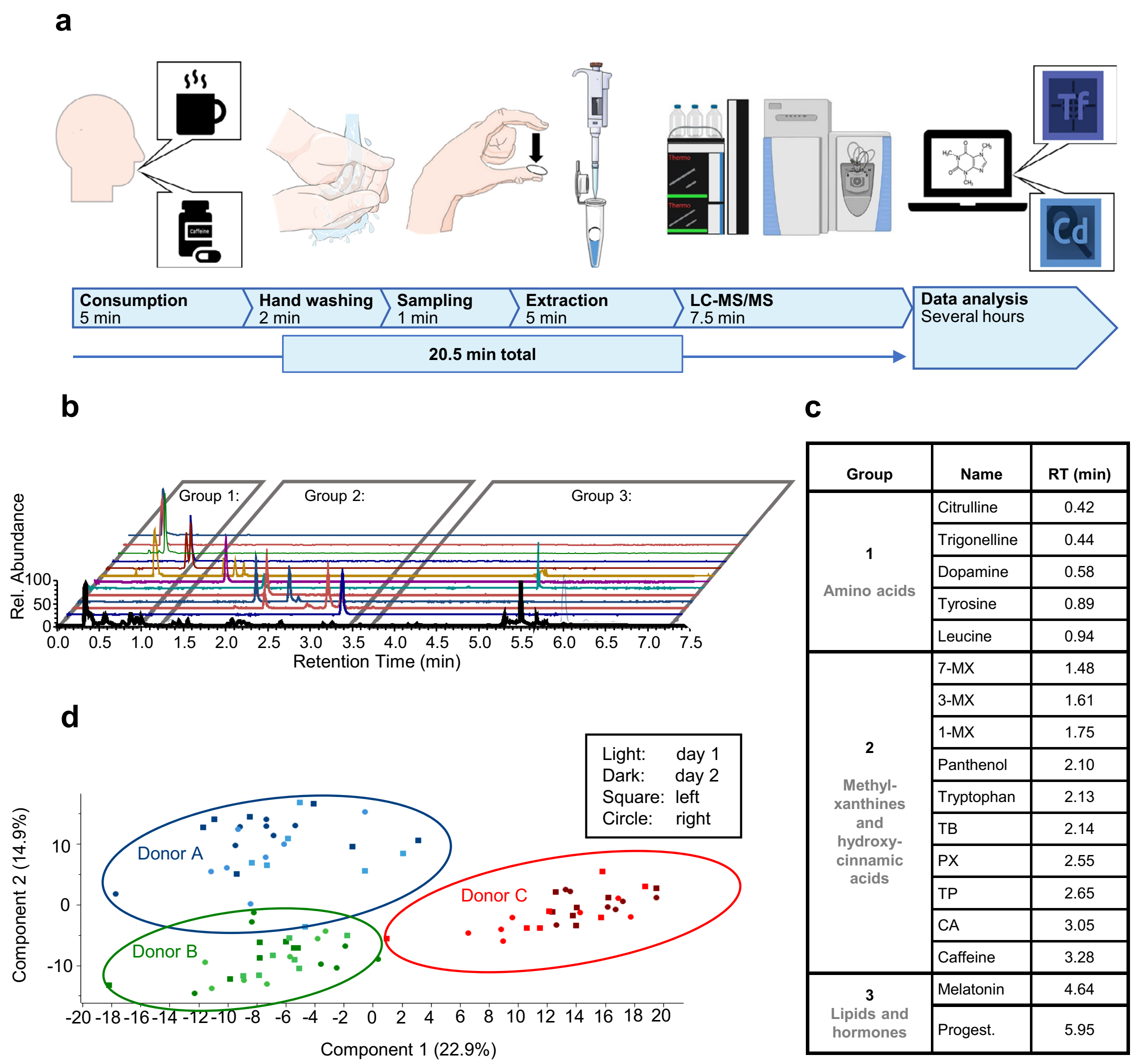
While fingerprints—the pattern of the ridge details left on a surface—have been used for the identification of individuals since the late 19th century[18], their relevance for detecting metabolites, as well as drugs and their metabolites has only recently been discovered[19,20]. While drug substances detected in the fingerprint may originate from accidental dermal contact, the detection of drug-specific metabolites implies that the drug was ingested, metabolised and subsequently excreted from sweat glands. Thus, we hypothesised that sweat from the skin surface may represent a promising source for metabolic biomonitoring. Sweat is a hypotonic, slightly acidic biofluid secreted by the eccrine, apocrine and apoeccrine glands located on the skin surface[21,22]. Eccrine sweat from the fingertips is mainly composed of water (~99%), but contains electrolytes, urea, lactate, amino acids, metal ions[23,24] and a variety of endogenous metabolites, including peptides, organic acids, carbohydrates, lipids, lipid-derived metabolites, as well as xenobiotics[21,22,25–27]. Sweat composition is highly dynamic, changes significantly with pathological states and may reveal habits of diet, metabolic conditions or use of drugs and supplements[17,24,28]. In fact, the analysis of sweat has already been reported to assess individual metabolic characteristics[29,30]. Clinical assays based on the analysis of sweat exist and include the screening of newborn children for elevated chloride and sodium levels to confirm cystic fibrosis via pilocarpine stimulated iontophoresis or forensic and criminal investigations to test for illicit drug use[17,22,31–33]. Furthermore, it has already been successfully demonstrated that the analysis of proteins contained in sweat enables not only the diagnosis of active tuberculosis but can also be used to screen for lung cancer[16,34,35], highlighting the potential of sweat analysis for precision medicine[36]. Real-time monitoring of biomarkers was demonstrated with wearable sweat sensors for uric acid and tyrosine[37], interleukin-6 and cortisol[38] or electrolytes such as sodium, ammonium ions and lactate[39].

However, these studies typically assessed a small number of metabolites and relied on elaborate methods to collect sweat, including sweat patches or artificially forcing sweat production[17,22,30]. This was necessary because the detection methods required relatively large absolute amounts of these metabolites. It is known that eccrine glands on the fingertips produce sweat at a rate of 50–500 nL cm^{-2} min^{-1} [40]. Thus, the analysis of metabolites from sweat of the fingertips may be achieved with sufficiently sensitive instrumentation, for example MS[41]. Sample collection using sweat from fingertips requires no patient pre-treatment or trained personnel, is safe and fast. Upon optimising the entire workflow for the analysis of sweat from the fingertips, we analysed 1792 samples from 40 participants, which underlines its potential as a high-throughput metabolic technology. Proof-of-principle studies based on the consumption of coffee or ingestion of a caffeine capsule were designed to assess metabolic time-series of each participant and provided evidence of the feasibility of this approach. Fluctuations in the rate of sweat production were accounted for by mathematical modelling of the conversion of xenobiotics to their catabolic products (e.g. caffeine to paraxanthine). In this study, we show that metabolic phenotyping using sweat from fingertips combined with mathematical network modelling may have far reaching relevance for precision medicine, because it allows to obtain dynamic metabolic responses of individuals.

Results

Sweat from the fingertips is a rich source for metabolic phenotyping. A straight-forward workflow was established for sampling and processing sweat samples from fingertips. In short, hands are washed without soap and dried with a disposable paper towel prior to each sampling time-point. For sweat collection, a circular sampling unit standardised to 1.15 cm diameter was then held between thumb and index finger for 1 min and was transferred with clean tweezers into an empty tube for storage (Fig. 1a). The metabolites were extracted from the sampling units using aqueous conditions and the resulting solution was directly introduced into the liquid chromatography-mass spectrometry (LC-MS) system for analysis. Sample collection and processing required ~13 min per sample. Sampling can be performed by untrained personnel in a highly frequent manner and the non-invasive nature of the sampling facilitates patient compliance. Data acquisition requires a further 7.5 min, which gives a total of ~20 min for the entire workflow per sample.

Based on the known rates of sweat production in eccrine glands on the fingertips[29,40], the median sweat volume collected using this method can be estimated at around 200–2000 nL (2 min × 2 cm^2 × 50–500 nL min^{-1} cm^{-2}) sweat per sample. High-resolution MS using a Q Exactive HF orbitrap hyphenated with an ultrahigh-performance liquid chromatography (UHPLC) system proved suitable for metabolic phenotyping from sweat samples (see methods). Initially, three participants were sampled multiple times in an observational study in order to evaluate the metabolic profile obtained from sweat of the fingertips of each individual. In detail, the participants collected sweat samples seven times per day at different intervals on 2 consecutive days and using both hands (see methods, study A). A total of 250 metabolites were identified and verified by external standards (Supplementary Data 1). Actually, many known as well as previously unknown endogenous and exogenous metabolites were identified in the sweat samples with high confidence (Fig. 1b, c). We detected not only a number of amino acid-related metabolites (e.g. tyrosine, leucine or citrulline), but also hormones (e.g. melatonin or progesterone). Newly identified metabolites include dopamine, progesterone and melatonin amongst others. Interestingly, we observed many coffee-derived metabolites, including caffeine and the related dimethyl– and methylxanthines. Principal component analysis (PCA) using those metabolites revealed that the samples clustered according to individuals (Fig. 1d). This indicated that the molecular composition of sweat associated with a given individual dominated the variances derived from multiple sampling. Interestingly, the principal components were strongly determined by the endogenous metabolites histamine, tryptophan, tyrosine and arginine (Supplementary Fig. 1). Moreover, we did not find notable



Group	Name	RT (min)
1 Amino acids	Citrulline	0.42
	Trigonelline	0.44
	Dopamine	0.58
	Tyrosine	0.89
	Leucine	0.94
2 Methyl-xanthines and hydroxy-cinnamic acids	7-MX	1.48
	3-MX	1.61
	1-MX	1.75
	Panthenol	2.10
	Tryptophan	2.13
	TB	2.14
	PX	2.55
	TP	2.65
	CA	3.05
	Caffeine	3.28
3 Lipids and hormones	Melatonin	4.64
	Progest.	5.95

Fig. 1 Sweat from the fingertips enables individualised metabolic biomonitoring. A straight-forward workflow for sweat sampling and processing was established and successfully applied to proof-of-principle studies to investigate caffeine metabolism in an individualised fashion. **a** Graphical summary of the workflow including consumption of a cup of coffee or a caffeine capsule, sampling sweat from fingertips, the extraction of analytes and subsequent LC-MS/MS analysis as well as data analysis with respective durations in minutes. Panel **a** was modified from Servier Medical Art, licensed under a Creative Common Attribution 3.0 Generic License (http://smart.servier.com/) and BioRender (https://biorender.com/). Tf Tracefinder Software, Cd Compound Discoverer Software (both Thermo Fisher Scientific). **b** Extracted ion chromatograms of exemplary sweat components are shown. Based on their retention time, analytes were assigned to three groups as listed in **c**. **c** Identities of sweat constituents according to order of elution. CA chlorogenic acid, MX methylxanthine, PX paraxanthine, TB theobromine, TP theophylline, Progest. progesterone. **d** Principal component analysis (PCA) of finger sweat samples derived from the left (square) and right (circle) hand of three participants is depicted before and after coffee consumption at two different days (light and dark colour). PCA was calculated with a set of 250 metabolites (Supplementary Data 1) and successfully clustered the finger sweat samples according to the participants.

differences of the sweat composition between the left and right hand from a given individual (Fig. 1d).

Sampling sweat from the fingertips is reliable and robust. Biomolecules are characterised by LC-MS according to retention time (RT), the accurate mass of the molecular ion derived from the full mass spectrum (MS1) and the fragmentation pattern determined by tandem mass spectrometry (MS2). The experimentally determined mass-to-charge ratios of 15 representative metabolites showed mass deviations below <2 ppm, which are typical for Q Exactive HF instruments (Supplementary Table 1). The coefficient of variation (CV) of the RT determined for the internal standard caffeine-(trimethyl-D9) was found to be 1% across 636 injections (Fig. 2a, see methods, study A and C). Caffeine-(trimethyl-D9) was injected with every sample at 10 pg on column. The CV of the areas under curve (AUCs) across the same sample set was 11% ($n = 636$). The CV improved slightly when considering study A only (CV = 7%, $n = 186$), but remained constant for study C (CV = 10%, $n = 450$). This indicated that the performance of the LC-MS system was robust across each sample set. MS2 spectra were of good quality and provided high matching factors, which supported the identification of previously known and newly identified metabolites found in sweat, e.g. tryptophan[42] and dopamine, respectively (Fig. 2b). Caffeine and its three main metabolites paraxanthine, theobromine and theophylline were spiked onto sampling units in the range of 1–100 pg μL^{-1}. These samples were processed according to the above-mentioned procedures and linear calibration curves were obtained with associated $R^2 > 0.997$ (Fig. 2c). At concentrations of 100 fg μL^{-1}, these molecules were still detected

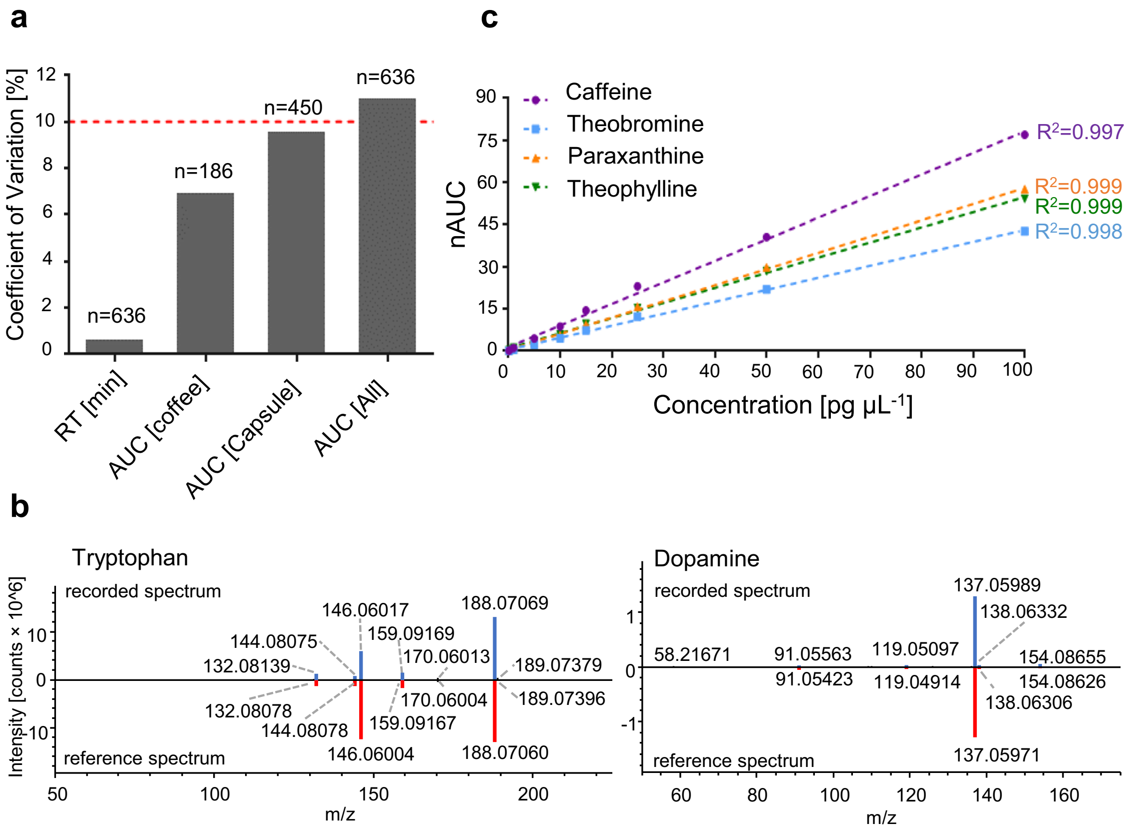


Fig. 2 LC-MS/MS analysis of metabolites from sweat of the fingertips is precise and robust. a Coefficients of variation of the retention times (RT) and areas under the curve (AUC) of a set of LC-MS/MS runs, as well as AUCs for the coffee (study B) and caffeine capsule (study C.1) intervention studies were determined for the internal standard caffeine-(trimethyl-D9). The means (boxes) and standard deviations are as follows: for the retention time 3.28 ± 0.02, for the coffee AUCs 1.80 ± 0.13 × 10^6, for the capsule AUCs 1.56 ± 0.15 × 10^6 and for all AUCs 1.63 ± 0.18 × 10^6. The dashed red line was set to 10%. **b** Head-to-tail comparison of the recorded MS2 spectrum (blue) to the reference spectrum from *mzcloud* (red) of tryptophan (left) and dopamine (right) demonstrates high spectral quality supporting reliable compound identification. **c** Calibration curves for caffeine, theobromine, paraxanthine and theophylline with respective correlation factors (R^2) are shown. nAUC normalised area under the curve.

with signal-to-noise ratios >100 on the Q Exactive HF. Comparison of a spiked and processed caffeine standard (10 pg μL^{-1}) to a directly injected caffeine standard (10 pg μL^{-1}) yielded an extraction efficiency of 93%. The lower limit of quantification (LLOQ) was determined from the calibration curves as the mean AUC plus ten times the standard deviation of caffeine and its metabolites found in blank sampling units. This resulted in a LOQ of 0.2 pg μL^{-1} for caffeine, 0.1 pg μL^{-1} for paraxanthine and 1.7 pg μL^{-1} for theobromine (see Source Data). The AUCs for theophylline in filter blanks and caffeine in tap water and paper towels were below the limit of detection (LOD), which was calculated as the mean AUC plus three times the standard deviation.

Coffee consumption revealed coffee-specific xenobiotics in finger sweat. After confirming sweat from the fingertips to contain endogenous metabolites, as well as xenobiotics mainly related to coffee consumption, we designed an intervention study with 11 participants, who consumed a standardised amount of coffee after a 12 h fasting period with regard to caffeine-containing food (see methods, study B). Two additional volunteers were sampled, who did not consume coffee, thus representing the control group. Sweat samples were collected before coffee consumption and subsequently after 15, 30, 45, 60, 90 and 120 min. Caffeine is a widely used stimulant of the central nervous system and features an excellent oral bioavailability[43,44]. Since the ingestion of an equivalent of a double espresso was already shown to have systemic effects by affecting sleep behaviour[45–47], we expected to find caffeine and related xenobiotics upon coffee consumption in sweat from the fingertips. The metabolite levels of the participants before coffee consumption (0 min) revealed negligible amounts of chlorogenic acid, trigonelline and caffeine, while the primary metabolites of caffeine showed significant background levels (e.g. paraxanthine, theobromine and theophylline). The control group featured stable metabolite levels over time with small variations probably stemming from fluctuations in the rate of sweat excretion (Supplementary Fig. 2). Strikingly, the sweat from the fingertips 15 min post consumption revealed 35 xenobiotics of 121 metabolites (29%) contained in coffee presently identified by us from aqueous extracts of the roasted coffee beans used for this study, including among others caffeine, theobromine, theophylline, paraxanthine, methylxanthines, chlorogenic acid, trigonelline, methylsuccinic acid, quinic acid and iditol (Supplementary Data 2). The AUCs of caffeine, chlorogenic acid and trigonelline increased significantly in all volunteers as early as 15 min after coffee consumption (Fig. 3a). The time-dependent sampling revealed differences in kinetic properties of the coffee-specific xenobiotics, especially regarding absorption and clearance rates. For example, the AUCs of caffeine and chlorogenic acid peaked after 15 min, followed by rapid clearance, while the AUCs of the dimethylxanthines increased steadily over time on top of a pre-existing pool (Fig. 3b). Several coffee-specific metabolites

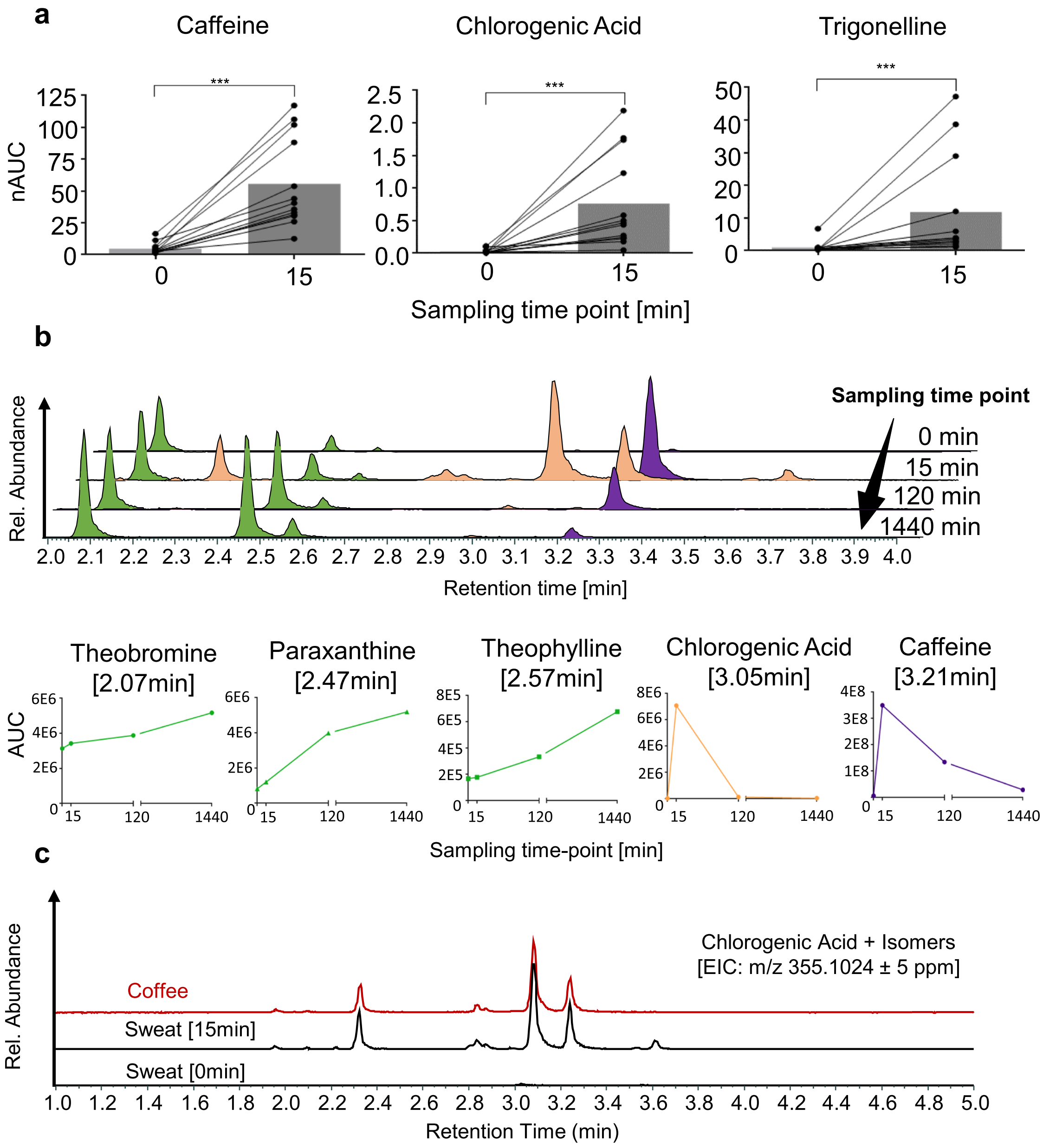
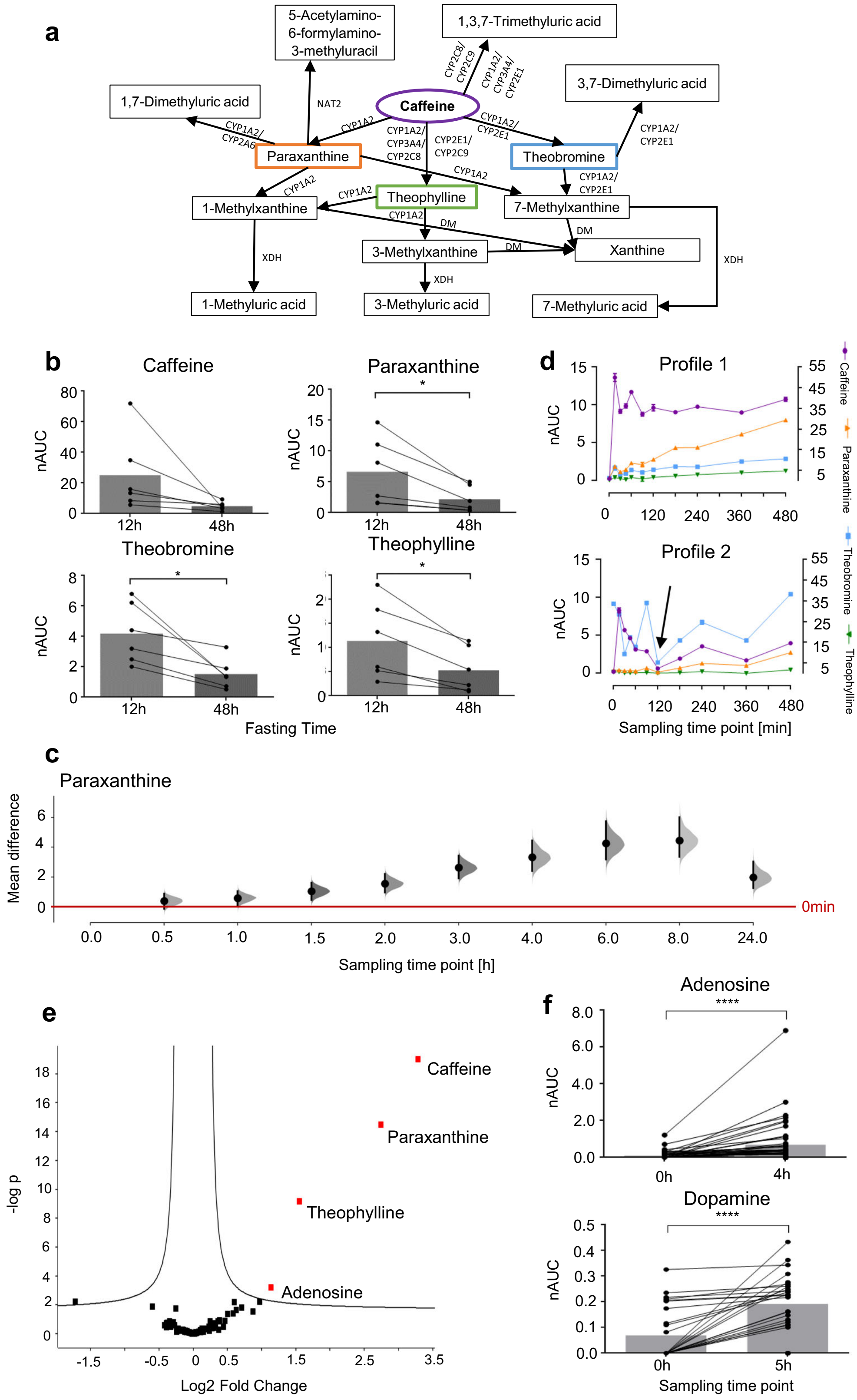


Fig. 3 Xenobiotics are detected in a time-dependent manner in sweat from the fingertips after coffee consumption. a Levels of normalised areas under the curve (nAUCs) for caffeine, chlorogenic acid and trigonelline, before (0) and 15 min (15) after coffee consumption are shown, demonstrating a significant increase in all participants ($n = 13 \times 2$ time-points) after 15 min. D'Agostino & Pearson test was performed to check normality of the data. Paired two-tailed Wilcoxon signed rank tests were performed for 13 volunteer profiles, delivering a p-value = 0.0002 for caffeine, chlorogenic acid and trigonelline. The mean nAUCs (boxes) and standard deviations are the following: for caffeine 4.8 ± 4.4 at 0 min and 56 ± 35 at 15 min, for chlorogenic acid 0.03 ± 0.04 at 0 min and 0.8 ± 0.7 at 15 min, for trigonelline 1.0 ± 1.7 at 0 min and 12 ± 16 at 15 min. **b** The temporal evolution of metabolite profiles is exemplarily shown for one participant (Volunteer 3, study A). The sampling time-point 1440 min represents the time-point before consumption on the second sampling day. Whereas caffeine (violet) and chlorogenic acid AUCs (orange) were found to increase quickly after coffee consumption followed by rapid clearance, the levels of theobromine, paraxanthine and theophylline (green) increased more slowly within the observation period. **c** Similarity of extracted ion chromatograms (EIC) of chlorogenic acid and its isomers from coffee extracts and from sweat of the fingertips 15 min after coffee consumption. The corresponding sample collected just before coffee consumption (0 min) served as negative control.

displayed a number of isomers in their extracted ion chromatograms. For example, chlorogenic acid (m/z 355.1024, RT = 3.05 min) showed at least five isomers (Fig. 3c) as verified on MS2 level. The ratio of the relative peak intensities of chlorogenic acid and its isomers was conserved when comparing coffee extracts and sweat from the fingertip. This indicated that these isomers are equally distributed into the water-soluble body compartment and are equally cleared from body on a rapid timescale. Chlorogenic acids and its isomers were not observed prior to coffee consumption. Such a comparative analysis strategy may be used to discover other xenobiotics distributed to sweat glands in a systemic fashion as indicated by the yet unidentified feature detected at m/z 337.0920 (Supplementary Fig. 3). These findings provide evidence that ingested xenobiotics may be robustly detected in the sweat from the fingertips, and their time-dependence mirrors their pharmacokinetic properties.



Finger sweat enables to elucidate individual metabolic traits. The metabolism of caffeine by different hepatic enzymes is well known[48], and the catabolic products were successfully identified in sweat from fingertips after coffee consumption (Fig. 4a, Supplementary Table 1). However, dimethyl– and methylxanthines may originate from both coffee beans and from endogenous hepatic metabolism. Additionally, we observed significant background levels of these metabolites in sweat from the fingertips before coffee consumption. In order to monitor the physiological conversion of caffeine into dimethylxanthines by hepatic enzyme

Fig. 4 Consumption of a caffeine capsule enables to elucidate individualised metabolic traits from sweat of the fingertips. a Caffeine metabolism including known metabolic routes, metabolites and related enzymes: CYP cytochromes P450, NAT2 N-acetyltransferase 2, XDH xanthine dehydrogenase, DM demethylase. These metabolites were all detected in sweat from the fingertips. **b** Six individuals participated in the coffee as well as in the caffeine capsule studies. The AUCs of caffeine and the primary metabolites are compared depending on the duration of the fasting period (12 vs 48 h, $n = 6$). Longer fasting significantly reduced the amounts of xenobiotics in sweat from the fingertips. It was tested with Kolmogorov–Smirnov test using Dallal–Wilkinson-Lilliefors *p*-value if values came from a Gaussian distribution. A two-tailed paired *t*-test (6 participants × 2 time-points) was performed for caffeine, paraxanthine, theobromine, and theophylline. **c** Shared-control plot with data from 47 volunteer profiles for paraxanthine is shown. The mean differences between the control group (time-point before consumption, red line) and each of the sampling time-points post ingestion is plotted on the *y*-axis. Paraxanthine is significantly upregulated from the sampling time-point 1.5 h on after ingesting a caffeine capsule. The effect size is presented as a bootstrap 95% confidence interval. Mean difference, lower and upper limits are provided in the Source data. **d** Exemplary metabolic profiles of two participants, demonstrating individual differences in metabolic properties regarding caffeine metabolism as exemplified by the preferential formation of paraxanthine in volunteer profile 1 in contrast to theobromine in case of volunteer profile 2. Caffeine is displayed on the right *y*-axis, while theobromine, paraxanthine and theophylline are displayed on the left *y*-axis. Error bars represent standard deviation of two technical replicates ($n = 2$) for each of the 11 time-points. Means and standard deviations can be found in the Source data. **e** Metabolic changes 4 h after consuming a caffeine capsule demonstrated with a volcano plot illustrating the similarities of metabolite regulations in 47 volunteer profiles. Next to the known caffeine metabolites, adenosine is regulated. **f** Boxplots for adenosine and dopamine before and 4 h/5 h after consuming a 200 mg caffeine capsule shown for 47 (study C.1 and C.2)/ 27 (study C.2) volunteer profiles. Normality of the data was checked with D'Agostino-Pearson test. A two-tailed Wilcoxon Signed Rank Test was performed for adenosine. A tow-tailed *t*-test was performed for dopamine. nAUC normalised area under the curve. Boxes represent the means of each time-point. All statistical test results as well as means and standard deviations can be found in the methods section.

activity, we designed a study in which participants refrained from consuming caffeine-containing products for at least 48 h before ingesting a single caffeine capsule (200 mg). The caffeine capsule and the longer fasting time were chosen to minimise background contributions from catabolic products of caffeine. Forty volunteers were enrolled in this study and sweat from the fingertips was sampled repeatedly over 27 h with up to 20 sample collections per volunteer (see methods, study C.1 and C.2). Six individuals participated in both the coffee consumption study (study B) and the caffeine capsule study (study C.1). Indeed, their prolonged fasting featured an improved baseline and revealed a significant decrease of dimethylxanthines to negligible levels after the 48 h fasting period compared to the 12 h fasting period (Fig. 4b). Ingestion of the caffeine capsule significantly increased the abundance of caffeine in sweat from fingertips in all volunteers already after 15 min, in accordance with coffee consumption. The caffeine abundance remained elevated for at least 480 min in all volunteers and returned close to baseline after 24 h (Supplementary Fig. 4). The abundance of the primary metabolite paraxanthine increased more slowly and peaked between 360 and 480 min post ingestion (Fig. 4c). Individual metabolic time-courses revealed rather striking differences regarding caffeine metabolism (Fig. 4d). For example, volunteer profile 1 displayed a sharp increase in caffeine abundance, which remained relatively constant over 480 min, while paraxanthine abundance increased steadily during this time period. In contrast, volunteer profile 2 featured a similar increase in caffeine abundance, but started with an elevated theobromine baseline, which also represented the main metabolite of caffeine. These findings suggest that sampling sweat from the fingertips may be of particular interest for characterising personalised metabolic traits. Cytochrome P450 enzymes are key players in the hepatic metabolism and several isoforms are known to process xenobiotics at different rates[48]. Thus, xenobiotics like caffeine may be subjected to variable metabolisms depending on the individual expression of these enzymes. This may reveal individual physiological responses to xenobiotic exposure that may serve as proxies for hepatic metabolic activity. Therefore, the influence of the metabolic turnover of caffeine depending on the expression of cytochrome P450 enzymes was investigated in vitro using HepG2 cells (Supplementary Information, Supplementary note 1). Indeed, we found that HepG2 cells would increase the metabolic turnover of caffeine to its primary metabolites upon chemical induction of cytochrome P450 enzymes with benzo-[a]-pyrene (Supplementary Fig. 6). Moreover, the induction of these enzymes also affected the relative ratios of the primary metabolites significantly. This supports the conclusion that the individual enzymatic activity status may modulate the formation of metabolites subsequently detected in sweat from the fingertips. Statistical analysis of the metabolites reproducibly detected in all 47 (study C.1 + C.2) or 27 (study C.2) volunteer profiles revealed the significant upregulation of caffeine, paraxanthine and theophylline, as well as adenosine 4 h post ingestion. Theophylline and paraxanthine reflected the metabolic turnover of caffeine within each volunteer profile, while adenosine was identified as an endogenous metabolite upregulated upon caffeine ingestion (Fig. 4e, f). Another endogenous metabolite, dopamine was significantly induced 5 h after consuming a caffeine capsule in 27 participants (study C.2, Fig. 4f, Supplementary Fig. 5). Adenosine and dopamine are not directly related to caffeine metabolism.

Mathematical modelling quantifies individual dynamic metabolic responses. Fluctuations in the rate of sweat excretion cause significant variance in the collected sweat volumes. This represents a fundamental challenge for the time-course analysis of sweat from the fingertips. For example, the apparent down-regulation of all analytes at 120 min in volunteer profile 2 (Fig. 4d, arrow) strongly suggests that at that time-point less sweat was collected in comparison to the adjacent measurements (see Fig. 5e, arrow). Moreover, the magnitude of this effect on the apparent concentration is unknown. We used dynamic metabolic network modelling to discern the effects of the sweat volume on the measured time-series of caffeine catabolism in the body (see methods). In brief, caffeine uptake and clearance via its major metabolic products paraxanthine, theobromine and theophylline can be described by first order kinetics (Fig. 5a)[49,50]. Due to fasting we can set the initial caffeine concentration at time 0 min to zero (Fig. 4b). Additionally, we consider the sweat volume to be a function of time, but assume that at every time-point the sweat volume is constant across all metabolites. The assumption holds if the modelled metabolites are not reabsorbed during sweating. The resulting mathematical model was fitted to each volunteer. We estimated the kinetic constants, the initial concentrations of paraxanthine, theobromine and theophylline and the sweat volumes at each time-point, as exemplified for volunteer profiles 1 and 2 (Fig. 5b, c, e, f and Supplementary Table 2). In both cases our model accurately described individual caffeine metabolisms with good accuracy (goodness of fit $R^2_{\text{adjusted}} > 0.90$). Besides the possibility to estimate the rate of sweat excretion by means of

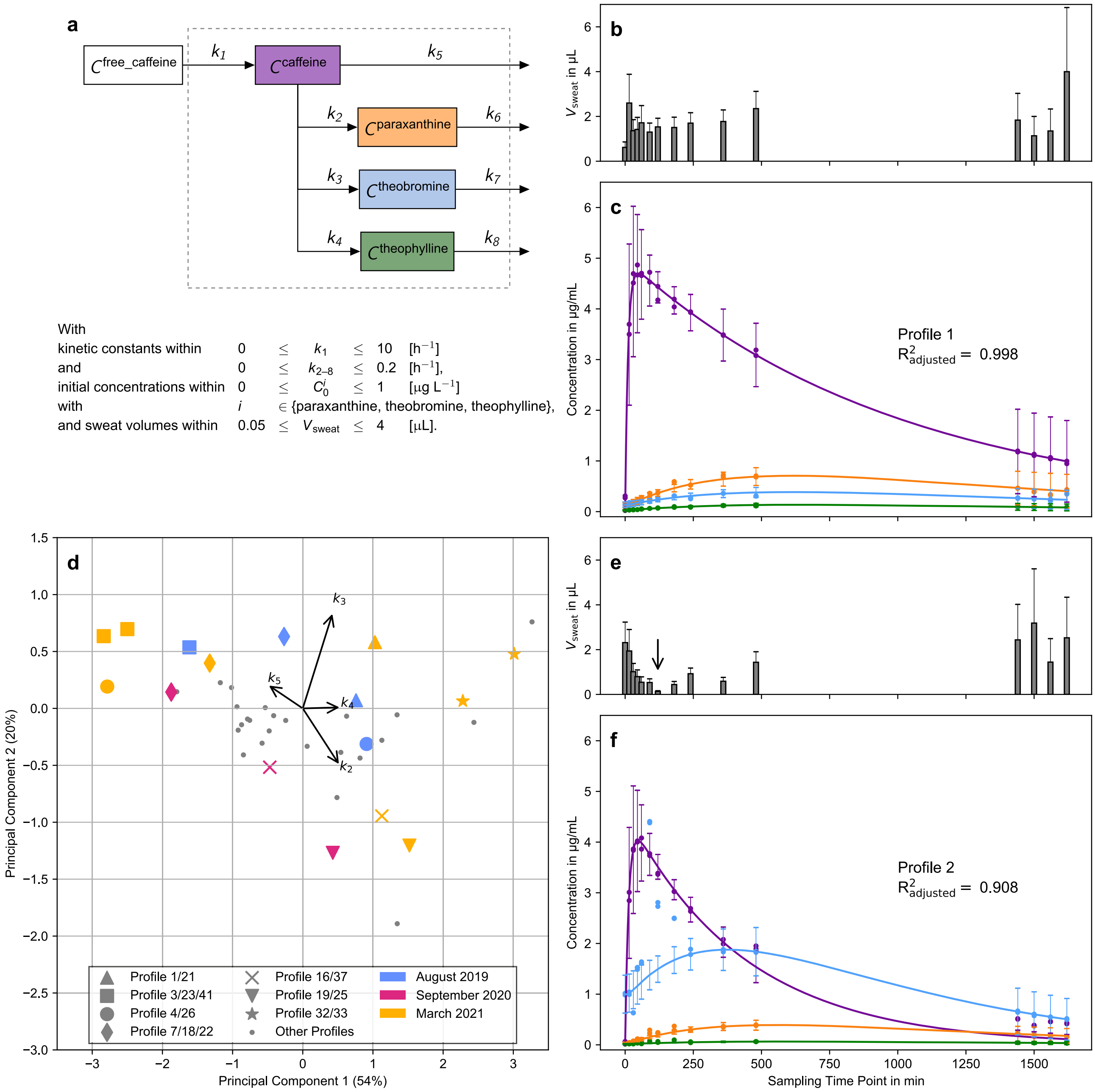


Fig. 5 Metabolic networks facilitate the discovery of dynamic metabolic patterns from individuals. a Network of caffeine and its major metabolites that was used for fitting of the concentration time-series and the constraints of fitting parameters. **b**, **e** Estimated sweat volumes (V_{sweat}) in µL of the measurements of volunteer profiles 1 and 2, respectively. Each bar represents a single V_{sweat}. The error bars show the 95% confidence intervals of the model prediction. **c**, **f** Fitted concentration time-series of caffeine, paraxanthine, theobromine and theophylline for volunteer profiles 1 and 2 (compare Fig. 4d). The lines refer to the fitted concentration and the symbols refer to the measured values ($\widetilde{\mathbf{M}}$) divided by V_{sweat} at each time-point (Eq. (1)). The error bars show the 95% confidence intervals of the model prediction ($n = 2$ technical replicates times 4 metabolites times 15 time-points per profile). The arrow marks the sweat volume of profile 2 at 120 min (see text). A visual representation of the influence of the sweat volume on the fit is shown in Supplementary Fig. 7. **d** Two-dimensional PCA plot of the fitted caffeine conversion constants. Volunteer profiles (i.e. time-series) from the same participant are plotted with large, coloured symbols whereas participants who contributed only once are marked with small grey circles. The colours represent the month of sampling.

this modelling approach, the shape of the curves visualises the dynamic metabolic patterns of each individual.

Interestingly, the kinetic constants for uptake (k_1) of caffeine is within the standard deviation, while the constants of conversion (k_2, k_3, k_4) are approximately half of the literature values of population averages for blood plasma (Supplementary Table 2)[51]. Whereas the fractional conversion of caffeine to the main metabolic product paraxanthine in volunteer profile 1 is similar to what is described as population average[51,52] we saw substantial differences for volunteer profile 2, who displayed theobromine as the main metabolic product of caffeine (Supplementary Table 3). We found individual differences to be robust over time. In Fig. 5d a two-dimensional PCA plot of the fitted conversion constants of caffeine (k_2, k_3, k_4, k_5) is shown. Individuals who generated at least two volunteer profiles (i.e. independent time-series) are marked with large symbols. Their respective colour indicates the month of sampling. Not only do two profiles of the same volunteer within one month cluster close to each other (e.g. star symbols), but also the ones that were sampled more than 1.5 years apart are in close proximity (e.g. diamond symbols). The biggest difference of volunteer profiles from one participant was found for profiles 4 and 26 (big circles). For volunteer profile 26, however, we observed an overall poor fit ($R^2_{adjusted}$ of 0.56 compared to 0.984 for profile 4). On another note, the original axes in Fig. 5d show that the catabolism of caffeine into paraxanthine (k_2) and direct elimination (k_5) is negatively correlated, whereas the catabolism of caffeine into theobromine (k_3) and theophylline (k_4) is positively correlated. This correlation is known in literature and is likely due to common hepatic cytochrome P450 enzymes catalysing the conversion of caffeine to theobromine and theophylline (Fig. 4a)[53].

Targeted assays can be established for clinical implementation. The described metabolic phenotyping approach represents a powerful discovery tool for endogenous and xenobiotic compounds found in sweat of the fingertips. In order to evaluate the feasibility of clinical implementation, we established a targeted assay for caffeine, and the primary metabolites theobromine, theophylline and paraxanthine on a triple quadrupole MS using multiple reaction monitoring (MRM, see Supplementary Information, Supplementary note 2 and Supplementary Table 4). For this purpose, five participants consumed a standardised coffee on 3 independent days after a 12 h caffeine-free fasting period and samples were collected at different time intervals in analogy to study B. The assay was validated and revealed linear ranges between 0.5 and 300 pg μL^{-1} of the respective metabolites (0.25–150 pg on column, Supplementary Fig. 9). LOD values were <0.2 pg μL^{-1} per collected sweat sample. The overall process efficiencies were generally >88% and the precision of 25 pg μL^{-1} spiked metabolite was <1% (Supplementary Table 5), while the overall CV of the AUC of caffeine 5 h after coffee consumption of all volunteers over 3 independent days was 22% (Supplementary Fig. 10). This suggests that targeted assays based on the analysis of sweat from the fingertips can be successfully established directly from metabolic phenotyping.

Discussion

The present study provides evidence that sweat from the fingertips can be used for dynamic metabolic phenotyping. The sample collection is non-invasive, safe and can be accomplished by untrained personnel, supporting patience compliance[47]. Other minimally to non-invasive approaches such as microneedle patches or sweat patches, require longer collection periods of several minutes up to days, aiming to collect sweat at a single timepoint[17,54,55]. In our approach, time-course analyses with frequent sampling can be performed due to the facile collection procedure. Our setup allows the analysis of unstimulated sweat in contrast to published approaches where sweat production was induced with pilocarpine iontophoresis (coupled with the Macroduct sweat collector) or physical exercise. Such stimuli were shown to alter the physiological sweat composition, which may introduce bias into the analysis[17,56]. The entire workflow can be accomplished within 20 min per sample, and has the potential to support large scale longitudinal metabolic studies. However, metabotyping the small amounts of sweat requires sufficiently sensitive analytical equipment. Although our approach centres on metabolic profiling using dedicated high-resolution instrumentation, we demonstrated the successful transfer to a targeted assay. Targeted MS is now routinely implemented in the clinical laboratory[57].

Sweat from the fingertips represents a rich source for metabolic phenotyping. Considering that a given metabolite may be represented in an LC-MS experiment by several features due to different adducts and charge states[58], it may be estimated that several thousand distinct metabolites can potentially be identified in sweat from the fingertips using this methodology. So far, we have verified 250 metabolites with external standards (Supplementary Data 1). The analysis is robust and sensitive with limits of detection of metabolites found in the sub-picogram range per sweat sample. Indeed, the detection limits found in this study showed improved sensitivity compared to previously used methodologies[59]. As a result, numerous endogenous metabolites were identified, which have not yet been described in sweat, including dopamine, progesterone and melatonin (Fig. 1). This highlights the potential of this approach to successfully identify low-abundant metabolites, which are challenging to detect in other biofluids due to matrix effects (e.g. melatonin in blood or plasma)[59–61]. Analysis of the area under the curve of the internal standard revealed an overall coefficient of variation of 11% across 636 samples and indicated acceptable precision (Fig. 2).

Proof-of-principle intervention studies were successfully carried out and support the applicability of the method. In two separate studies, participants were asked to consume a standardised cup of coffee or ingest a caffeine capsule after a caffeine– and theobromine-free diet for 12–72 h. After ingestion, sweat samples were collected up to 20 times within 27 h per volunteer. Sampling intervals of 15 min were feasible. Coffee consumption led to a significant upregulation of caffeine, chlorogenic acid and trigonelline within 15 min in all participants (Fig. 3, study B). This suggested a fast absorption and distribution of these xenobiotics, which also displayed distinct absorption and excretion kinetics (Fig. 3c). Altogether, 35 metabolites originating from coffee were detected in sweat from the fingertips.

The observation of significant background levels of dimethylxanthines after coffee consumption in study B pointed towards a confounding problem with respect to the origin of these caffeine metabolites. In fact, their temporal increase may have been due to their absorption from consumed coffee and hepatic caffeine metabolism. In order to resolve this question, we designed an additional study in which participants ingested a caffeine capsule (200 mg) only and adhered to a longer caffeine– and theobromine-free fasting regime. Of note, the longer fasting periods (48–72 h) significantly reduced the background levels of the primary metabolites (Fig. 4b, study C) compared to 12 h fasting (study B). Interestingly, statistical analysis of the metabolic profiling data from study C, involving the caffeine capsule, revealed a significant upregulation of caffeine and of the metabolic products theophylline and paraxanthine across all participants after 480 min (Fig. 4). Moreover, participants featured significantly increased levels of dopamine after 5 h. Being an endogenous metabolite, it is plausible to assume that this upregulation corresponded to a physiological response to caffeine ingestion. Increased dopamine levels were already observed upon caffeine[62], as well as coffee consumption by others[63]. Adenosine was significantly induced 4 h post ingestion of a caffeine capsule. Caffeine exerts most of its biological actions such as countering sleep pressure via antagonising adenosine receptors[64]. It has been demonstrated that caffeine increases plasma adenosine concentration potentially via receptor-mediated regulation of the plasma adenosine concentration[65] and this finding seems to extend to sweat from the fingertips. We have previously described individual opposing responses with regard to anti-inflammatory effects after coffee consumption[66]. Such studies required the collection of blood from volunteers and this could now be facilitated by analysing sweat from the fingertips. Adenosine is also known to be an anti-inflammatory mediator that may regulate neutrophils, macrophages and lymphocytes through interacting with surface receptors of these cells[67]. It is important

Table 1 Overview of the three studies discussed in this publication.

Study	Volunteers	Design	Fasting [hours]	Sampling time-points
A	2 males, 1 female	Observational	12	0, 15, 30, 45, 60, 90 and 120 min on 2 consecutive days
B	7 males, 6 females	Double espresso or control group	12	0, 15, 30, 45, 60, 90 and 120 min
C.1	8 males, 9 females	Caffeine capsule	48	0, 15, 30, 45, 60, 90, 120 min and 3, 4, 6, 8, 24, 25, 26 and 27 h
C.2	16 males, 11 females	Caffeine capsule	72	0, 15, 30, 45, 60, 90, 120 min and 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14 and 24 h

to note that sweat from the fingertips may not only reveal ingested xenobiotics, but also endogenously produced metabolic products and physiological responses to bioactive xenobiotics.

Individual metabolic traits were then investigated by analysing the time-dependent metabolic evolution of caffeine upon ingesting a caffeine capsule (Fig. 4d). We found that sweat from the fingertips may be successfully used for the personalised assessment of such metabolic activities. Importantly, this strategy may be extended to other xenobiotics or drugs and their causally related metabolic products in order to obtain insight into specific processes of human metabolism in an individualised manner. Moreover, by inducing cytochrome P450 enzymes in HepG2 cells in vitro (Supplementary note 1), we were able to modulate the metabolic turnover of caffeine and the formation of specific catabolic products. This suggests that the relative ratios of caffeine to its primary metabolites may reflect hepatic activity, since the physiological hepatic metabolism of caffeine relies on a similar set of enzymes as in HepG2 cells.

Variations in the sweat volume over the course of the study represented a major challenge for normalisation and quantification. Mathematical modelling overcame this issue by addressing molecular constraints of substrate-product relations of enzymatically linked metabolites. Successful modelling has two central prerequisites: firstly, the measurement of at least two metabolites with known dynamics and, secondly, a linear relationship of said metabolites to the sweat rate. Importantly, this allowed us to compute a sweat volume that is proportional to all metabolites at each time-point. This approach was capable of delivering estimates of individual rate constants for drug uptake, metabolism and clearance and therefore allows to model dynamic metabolic patterns in individuals (Fig. 5). Sampling sweat from the fingertips enables time-course studies, which are evaluated by means of conversion rates of metabolically related substance classes. Their observed robustness suggests that the development of personalised tests via finger sweat measurements is feasible. For example, caffeine elimination was shown to be a proxy for liver function[68], and we hypothesise that a future study using an experimental setup identical to the caffeine capsule study could differentiate between patients with cirrhotic and normal livers. Additionally, we argue that the method presented here provides a convenient solution to the normalisation problem of finger sweat, which previously only has been tackled by employing microcapillaries[69]. However, they require large volumes of sweat, and thus need either long sampling times or require physical exercise. Both are detrimental when measuring fast pharmacokinetics, for example, for caffeine this would circumvent the requirement of absolute quantitative information of a single measurement.

In summary, metabolic phenotyping from sweat of the fingertips in conjunction with mathematical modelling is a promising approach to obtain dynamic metabolic patterns from individuals that may overcome the limitations of conventional composition biomarkers. Further research is currently performed in order to consolidate the potential of sampling sweat from the fingertips for applications in precision medicine.

Methods

Reagents and chemicals. LC-MS grade methanol, water, acetonitrile and formic acid used during sample preparation and LC-MS/MS analysis were purchased from VWR chemicals (Vienna, AT). Xenobiotic and metabolite standards (caffeine, theobromine, theophylline, paraxanthine, 1-methylxanthine, 3-methylxanthine, 7-methylxanthine, 1-methyluric acid, 3-methyluric acid, 1,7-dimethyluric acid, 3,7-dimethyluric acid and 1,3,7-trimethyluric acid, chlorogenic acid, xanthine, 5-Acetylamino-6-formylamino-3-methyluracil, dopamine and proteinogenic amino acids) were either purchased from Sigma–Aldrich (Vienna, AT) or Honeywell Fluka (GER). Caffeine capsules were bought from Mach dich wach! GmbH (GER). Sampling units were made from filter papers (precision wipes, number = 7552, white, 11 × 21 cm, Kimtech Science, Kimberly-Clark Professional, USA) using a circular puncher of 1cm^2.

Standard solutions and calibration samples. Stock solutions of 1 mg mL^{-1} of the analytical standards and the internal deuterated standards caffeine-(trimethyl-D9) and N-acetyl-tryptophan in methanol were prepared and stored at 4 °C. For caffeine, paraxanthine, theobromine and theophylline calibration curves were generated by spiking onto sampling units with the following concentrations: 0.1, 1, 5, 10, 15, 25, 50 and 100 pg µL^{-1}. The internal deuterated standards were prepared at a concentration of 1 pg µL^{-1} in an aqueous solution containing 0.2% formic acid, which served as the extraction solution for all samples.

Cohort design. Altogether, 21 males and 19 females with ages between 20–55 years and a BMI of 21 ± 8 kg m^{-2} were enrolled in this study. Participants had different dietary habits regarding the consumption of coffee; rare to regular consumption. Prior sampling, participants were required to fast caffeinated food (e.g. chocolate) and drinks (e.g. coffee, tea and energy drinks) for a period of 12–72 h. Sweat samples from the fingertips were collected at different time intervals and in the presence or absence of an intervention (see Table 1, studies A–C). Study B involved the consumption of a standardised coffee (equivalent to a double espresso), while studies C.1 and C.2 involves the ingestion of a caffeine capsule (200 mg). Seven volunteers have participated in more than one study, which gave a total of 47 volunteer profiles for study C. It was ensured that the volunteers did not touch the prepared coffee with their fingers.

Collection of sweat from the fingertips. Sampling units of 1 cm^2 circular surface were pre-wetted with 3 µL water and provided in 0.5 mL Eppendorf tubes. For each sweat collection, volunteers cleaned their hands using warm tap water and dried them with disposable paper towels. Volunteers kept their hands open in the air at room temperature for 1 min. Then, the sampling unit was placed between thumb and index finger using a clean tweezer and held for 1 min. Sweat formation was not forced. Filters were transferred back to labelled 0.5 mL Eppendorf tubes using a clean tweezer and stored at 4 °C until sample preparation.

Sample preparation. Coffee extracts were prepared taking an aliquot of 1 mL of a 250 mL coffee cup used for study A and B, which was centrifuged for 10 min at 15000 × *g*. The supernatant was diluted 1:100, 1:1000 and 1:10000 with the extraction solution consisting of an aqueous solution of caffeine-(trimethyl-D9) (1 pg µL^{-1}) with 0.2% formic acid. The dilutions were again centrifuged before analysis by LC-MS/MS.

For the extraction of metabolites from the sampling units, 120 µL of the extraction solution consisting of an aqueous solution of caffeine-(trimethyl-D9) (1 pg µL^{-1}) with 0.2% formic acid was added into the 0.5 mL Eppendorf tube containing the sampling unit. The metabolites were extracted by pipetting up and down 15 times. The sampling unit was pelleted on the bottom of the tube and the supernatant was transferred into HPLC vials equipped with a 200 µL V-shape glass insert (both Macherey-Nagel GmbH & Co.KG) and analysed by LC-MS/MS. Additionally, 10 unused filter, 10 paper towels and 10 tap water blanks were extracted similarly to determine potential contaminants and metabolite background levels.

LC-MS/MS analysis. A Q Exactive HF (Thermo Fisher Scientific) mass spectrometer coupled to a Vanquish UHPLC System (Thermo Fisher Scientific) was employed for this study. Chromatography was performed using a Kinetex XB-C18 column (100 Å, 2.6 µm, 100 × 2.1 mm, Phenomenex Inc.). Mobile phase A consisted of water with 0.2% formic acid, mobile phase B of methanol with 0.2% formic acid and the following gradient program was run: 1–5% B in 0.3 min and then 5–40% B from 0.3–4.5 min, followed by a column washing phase of 1.4 min at 80% B and a re-equilibration phase of 1.6 min at 1% B resulting in a total runtime of 7.5 min. Flow rate was set to 500 µL min^{-1}, the column temperature to 40 °C, the injection volume was 10 µL and the injection peak was found at RT = 0.3 min. All samples were analysed in technical duplicates. An untargeted mass spectrometric approach was applied for compound identification. Electrospray ionisation was performed in positive and negative ionisation mode. MS scan range was *m/z* 100–1000 and the resolution was set to 60000 (at *m/z* 200). The four most abundant ions of the full scan were selected for HCD fragmentation applying 30 eV collision energy. Fragments were analysed at a resolution of 15000 (at *m/z* 200). Dynamic exclusion was applied for 6 s. The instrument was controlled using Xcalibur software (Thermo Fisher Scientific).

Data analysis. Raw files generated by the Q Exactive HF instrument were analysed using the Compound Discoverer Software 3.1 (Thermo Fisher Scientific). Identified compounds were manually reviewed using Xcalibur 4.0 Qual browser and Freestyle (version 1.3.115.19) (both Thermo Fisher Scientific) and the obtained MS2 spectra were compared to reference spectra, which were retrieved from *mzcloud* (Copyright © 2013–2020 HighChem LLC, Slovakia). The match factor cut-off from *mzcould* was 80, while the mass tolerances were 5 and 10 ppm on MS1 and MS2 levels, respectively. Moreover, the identity of compounds suggested by Compound Discoverer was verified by analysing purchased standards using the same LC-MS method. The Tracefinder Software 4.1 (Thermo Fisher Scientific) was used for peak integration and calculation of peak areas. The generated batch table was exported and further processed with Microsoft Excel (version 1808), GraphPad Prism (version 6.07) and the Perseus software (version 1.6.12.0)[70], the letter being used for the principal component analysis. Untargeted metabolic profiling by mass spectrometry delivered more than 50000 reproducible sweat-specific features per analysis. Microsoft PowerPoint (version 1808) was used for creating figures.

Statistical analysis. D'Agostino-Pearson tests as well as Kolmogorov–Smirnov tests with Dallal–Wilkinson–Lilliefors *p*-value were performed to test if values came from a gaussian distribution. Two-tailed, paired *t*-tests or Wilcoxon Signed Rank Tests were performed for mass spectrometry data using GraphPad Prism (Version 6.07) to evaluate the significance of the abundance increase/decrease of compounds and their metabolites. For Fig. 4b it was tested with Kolmogorov–Smirnov test using Dallal–Wilkinson–Lilliefors *p*-value if values came from a Gaussian distribution. A two-tailed paired *t*-test (6 participants × 2 time-points) for caffeine (*p*-value = 0.1033, $t = 51.990$, $df = 5$), paraxanthine (*p*-value = 0.0297, $t = 3.012$, $df = 5$), theobromine (*p*-value = 0.0203, $t = 3.353$, $df = 5$) and theophylline (*p*-value = 0.0118, $t = 3.866$, $df = 5$). Means and standard deviations are for caffeine 25 ± 25 for 12 h fasting and 4.8 ± 2.7 for 48 h fasting, for paraxanthine 6.6 ± 5.5 for 12 h fasting and 2.2 ± 2.1 for 48 h fasting, for theobromine 4.2 ± 2.0 for 12 h fasting and 1.5 ± 1.0 for 48 h fasting, for theophylline 1.1 ± 0.8 for 12 h fasting and 0.5 ± 0.5 for 48 h fasting. For Fig. 4f normality of the data was checked with D'Agostino-Pearson test. A two-tailed Wilcoxon Signed Rank Test was performed for adenosine ($n = 47$, sum of positive ranks = 1020, sum of negative ranks = −17,00, sum of signed ranks = 1003, *p*-value ≤ 0.0001). A tow-tailed *t*-test was performed for dopamine (*p*-value ≤ 0.0001, t = 5.416, df = 26). The means and standard deviations are the following: for adenosine 0.1 ± 0.2 at 0 h and 0.7 ± 1.2 at 4 h, for dopamine 0.1 ± 0.1 at 0 h and 0.2 ± 0.1 at 5 h. Volcano plots were obtained using Perseus Software[70], setting the false discovery rate (FDR) to 0.05 and the minimal fold change (s0) to 0.1. For Fig. 4e the −log *p*-value for caffeine is 19.02, for paraxanthine 14.48, for theophylline 9.16 and for adenosine 1.14. Shared-control plots were generated with an R script[71].

Mathematical modelling. The model describes the concentration time-series of the ingested free caffeine and four sweat metabolites (caffeine, paraxanthine, theobromine, theophylline) within the constraints of following assumptions (Fig. 5a):

- caffeine metabolism can be described by mass-action kinetics in a one-compartment body model[49,50],
- the uptake of external caffeine is instantaneous (i.e. no lag time between ingestion and absorption into the body),
- the steady-state volume of distribution of caffeine, paraxanthine, theobromine and theophylline is instantaneously reached and time independent[50,51],
- concentration enrichment due to an increase in the water fraction from blood to sweat and dilution through the inability of bound caffeine to diffuse cancel each other out[72],
- apparent metabolite concentrations are proportional to the sweat volume (see Supplementary Fig. 8, Eq. (1)), and finally,
- sweat volumes are time dependent, but the same for all metabolites at one time-point.

A mathematical formulation of the problem of fluctuating sweat volumes is given in Eq. (1), where $\widetilde{\mathbf{M}}(t)$ is the measured mass vector of the internal metabolites and $\mathbf{C}(t)$ is the underlying concentration vector. $V_{\text{sweat}}(t)$ is a time-dependent volume that represents the sampled sweat volume. The resulting mathematical model is explained in detail in the Supplementary Note 3: Mathematical Model. Briefly, we describe the kinetics of caffeine metabolism with a system of ordinary differential equations (Supplementary Information, Supplementary Note 3: Mathematical Model, Eq. (2)). Subsequently we connect the solution of this equation over the sweat volume to the concentrations measured in the caffeine capsule study. Our model only contains variables that are either known and are thus fixed (volume of distribution, bioavailability, and ingested dose of caffeine) or have a concrete physical meaning but are unknown and need to be fitted (kinetic parameters, initial concentrations of paraxanthine, theobromine, and theophylline, sweat volumes). It allows to estimate absolute concentrations of tri- and dimethylxanthines in the finger sweat. Note that $V_{\text{sweat}}(t)$ is not constant over time and unknown and thus a unique fitting parameter at each sampled time-point. Therefore, the number of parameters that need to be fitted for the model is equal to the number of time-points (one V_{sweat} value per time-point) plus the number of parameters of the kinetic model. This requires the simultaneous fitting of the kinetics of multiple metabolites upon assuming that at each time-point $\text{V}_{\text{sweat}}(t)$ is constant for similar metabolites (Eq. (2)). By doing so the amount of data points that can be used for fitting is multiplied by the number of metabolites while the number of parameters for $\text{V}_{\text{sweat}}(t)$ stays constant. Thus (as long as the kinetic model is not overly complex) the system is sufficiently determined and data fitting is feasible.

$$\widetilde{\mathbf{M}}(t) = V_{\text{sweat}}(t)\,\mathbf{C}(t) \tag{1}$$

$$V_{\text{sweat}} = V_{\text{sweat}}^{\text{caffeine}} = V_{\text{sweat}}^{\text{paraxanthine}} = V_{\text{sweat}}^{\text{theobromine}} = V_{\text{sweat}}^{\text{theophylline}} \tag{2}$$

Caffeine and its major catabolic products paraxanthine, theobromine and theophylline were modelled subject to the following constraints: first order kinetics for all reactions (k_1 to k_8) with $0 \le k_1 \le 10\,\text{h}^{-1}$ and $0 \le k_{2\text{-}8} \le 0.2\,\text{h}^{-1}$; initial concentration of 0 for caffeine and $0 \le C_0^i \le 1\,\mu\text{g}\,\text{L}^{-1}$ for dimethylxanthines; and variability of V_{sweat} between $0.05 \le V_{\text{sweat}}(t) \le 4\,\mu\text{L}$. Generally, literature values of kinetic constants and sweat rates (without exercise) are well within the bounds of the model[40,51,73,74]. Finally, Supplementary Eq. (15) was used to fit the experimental data of all volunteers of the caffeine capsule study normalised by the machine standard individually. Fitting was performed in Python 3.7 with the SciPy package (version 1.6.1) using the curve_fit function and the integrated trust region reflective algorithm with default numerical tolerances (10^{-8})[75]. To find optimal settings for the fitting procedure we performed a systematic investigation of the hyperparameters in the Supplementary Note 4: Sensitivity Analysis. There, our implementation of the generalised, adaptive robust-loss function[76] in combination with Monte Carlo sampling of initial parameters for 100 times and selecting the solution with the lowest overall loss resulted in the smallest errors. Therefore, the same settings were adopted for this study. Moreover, with the estimated CVs associated to the fitting procedure from the Sensitivity Analysis we calculated confidence intervals ($n = 120$, $df = 93$), which are shown as error bars in Fig. 5b, c, e, and f (and Supplementary Table 2). Finally, we performed a PCA of the standard scaled kinetic constants of caffeine degradation (k_2, k_3, k_4, k_5) of all volunteer profiles (Fig. 5d).

Programs for mathematical modelling. PCA of kinetic parameters (Fig. 5d) was performed with Python 3.7 and scikit-learn (version 0.23.2). Levene-test in sensitivity analysis was performed with Python 3.7 and scipy (version 1.6.1). The mathematical modelling and sensitivity analysis was performed with Python 3.7 heavily relying on packages scipy (version 1.6.1) and robust-loss-pytorch (version 0.0.2).

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The data supporting the findings from this this study are available within the manuscript and its supplementary information. The metabolomics datasets have been deposited in the MetaboLights repository with the accession numbers MTBLS2772 and MTBLS2776[77]. Any remaining raw data will be available from the corresponding author upon reasonable request. Source data are provided with this paper.

Code availability

The code for mathematical model fitting and sensitivity analysis is available on GitHub (https://github.com/gotsmy/finger_sweat and https://doi.org/10.5281/zenodo.5222967)[78].

Received: 7 November 2020; Accepted: 22 September 2021;
Published online: 13 October 2021

References

1. Trivedi, D. K., Hollywood, K. A. & Goodacre, R. Metabolomics for the masses: the future of metabolomics in a personalized world. *N. Horiz. Transl. Med.* **3**, 294–305 (2017).

2. Nicholson, J. K. et al. Metabolic phenotyping in clinical and surgical environments. *Nature* **491**, 384–392 (2012).
3. Wishart, D. S. et al. HMDB 4.0: the human metabolome database for 2018. *Nucleic Acids Res.* **46**, D608–D617 (2018).
4. Blanco, F. J. & Ruiz-Romero, C. Osteoarthritis: Metabolomic characterization of metabolic phenotypes in OA. *Nat. Rev. Rheumatol.* **8**, 130–132 (2012).
5. Nicholson, J. K. & Wilson, I. D. Understanding 'global' systems biology: metabonomics and the continuum of metabolism. *Nat. Rev. Drug Discov.* **2**, 668–676 (2003).
6. Holmes, E., Wilson, I. D. & Nicholson, J. K. Metabolic phenotyping in health and disease. *Cell* **134**, 714–717 (2008).
7. Assfalg, M. et al. Evidence of different metabolic phenotypes in humans. *Proc. Natl Acad. Sci. USA* **105**, 1420–1424 (2008).
8. Wood, P. L. Mass spectrometry strategies for clinical metabolomics and lipidomics in psychiatry, neurology, and neuro-oncology. *Neuropsychopharmacology* **39**, 24–33 (2014).
9. Guo, L. N. et al. Plasma metabolomic profiles enhance precision medicine for volunteers of normal health. *Proc. Natl Acad. Sci. USA* **112**, E4901–E4910 (2015).
10. Nehlig, A. Interindividual differences in caffeine metabolism and factors driving caffeine consumption. *Pharmacol. Rev.* **70**, 384–411 (2018).
11. Yang, A., Palmer, A. & de Wit, H. Genetics of caffeine consumption and responses to caffeine. *Psychopharmacology* **211**, 245–257 (2010).
12. Walsh, M. C., Nugent, A., Brennan, L. & Gibney, M. J. Understanding the metabolome—challenges for metabolomics. *Nutr. Bull.* **33**, 316–323 (2008).
13. Dorne, J. L., Walton, K. & Renwick, A. G. Human variability in xenobiotic metabolism and pathway-related uncertainty factors for chemical risk assessment: a review. *Food Chem. Toxicol.* **43**, 203–216 (2005).
14. Marchant, B. Pharmacokinetic factors influencing variability in human drug response. *Scand. J. Rheumatol.* **39**, 5–14 (1981).
15. Wishart, D. S. et al. HMDB: The human metabolome database. *Nucleic Acids Res.* **35**, D521–D526 (2007).
16. Calderon-Santiago, M. et al. Human sweat metabolomics for lung cancer screening. *Anal. Bioanal. Chem.* **407**, 5381–5392 (2015).
17. Hussain, J. N., Mantri, N. & Cohen, M. M. Working up a good sweat—the challenges of standardising sweat collection for metabolomics analysis. *Clin. Biochem. Rev.* **38**, 13–34 (2017).
18. Faulds, H. The permanence of finger-print patterns. *Nature* **98**, 388–389 (1917).
19. Jang, M. et al. On the relevance of cocaine detection in a fingerprint. *Sci. Rep.* **10**, 1974 (2020).
20. Leggett, R., Lee-Smith, E. E., Jickells, S. M. & Russell, D. A. "Intelligent" fingerprinting: simultaneous identification of drug metabolites and individuals by using antibody-functionalized nanoparticles. *Angew. Chem.* **46**, 4100–4103 (2007).
21. Mena-Bravo, A. & de Castro, M. D. L. Sweat: A sample with limited present applications and promising future in metabolomics. *J. Pharm. Biomed. Anal.* **90**, 139–147 (2014).
22. Agrawal, K., Sivamani, R. K. & Newman, J. W. Noninvasive profiling of sweat-derived lipid mediators for cutaneous research. *Ski. Res. Technol.* **25**, 3–11 (2019).
23. Agrawal, K. et al. Sweat lipid mediator profiling: a noninvasive approach for cutaneous research. *J. Lipid Res.* **58**, 188–195 (2017).
24. Jadoon, S. et al. Recent developments in sweat analysis and its applications. *Int. J. Anal. Chem.* **2015**, 164974 https://doi.org/10.1155/2015/164974 (2015).
25. Agrawal, K., Bosviel, R., Piccolo, B. D. & Newman, J. W. Oral ibuprofen differentially affects plasma and sweat lipid mediator profiles in healthy adult males. *Prostaglandins Other Lipid Mediat.* **137**, 1–8 (2018).
26. Delgado-Povedano, M. M., Calderon-Santiago, M., de Castro, M. D. L. & Priego-Capote, F. Metabolomics analysis of human sweat collected after moderate exercise. *Talanta* **177**, 47–65 (2018).
27. Delgado-Povedano, M. M., Castillo-Peinado, L. S., Calderon-Santiago, M., de Castro, M. D. L. & Priego-Capote, F. Dry sweat as sample for metabolomics analysis. *Talanta* **208**, 120428 (2020).
28. Brunmair, J. et al. Metabo-tip: a metabolomics platform for lifestyle monitoring supporting the development of novel strategies in predictive, preventive and personalised medicine. *EPMA J.* **12**, 141–153 (2021).
29. Katchman, B. A., Zhu, M. L., Christen, J. B. & Anderson, K. S. Eccrine sweat as a biofluid for profiling immune biomarkers. *Proteomics. Clin. Appl.* **12**, 1800010 (2018).
30. Harshman, S. W. et al. Metabolomic stability of exercise-induced sweat. *J. Chromatogr. B* **1126**, 121763 (2019).
31. Kuwayama, K. et al. Time-course measurements of drugs and metabolites transferred from fingertips after drug administration: Usefulness of fingerprints for drug testing. *Forensic Toxicol.* **32**, 235–242 (2014).
32. Jarmusch, A. K. et al. Initial development toward non-Invasive drug monitoring via untargeted mass spectrometric analysis of human skin. *Anal. Chem.* **91**, 8062–8069 (2019).
33. Zhou, Z., Alvarez, D., Milla, C. & Zare, R. N. Proof of concept for identifying cystic fibrosis from perspiration samples. *Proc. Natl Acad. Sci. USA* **116**, 24408–24412 (2019).
34. Adewole, O. O. et al. Proteomic profiling of eccrine sweat reveals its potential as a diagnostic biofluid for active tuberculosis. *Proteomics Clin. Appl.* **10**, 547–553 (2016).
35. Delgado-Povedano, M. D., Calderon-Santiago, M., Priego-Capote, F., Jurado-Gamez, B. & de Castro, M. D. L. Recent advances in human sweat metabolomics for lung cancer screening. *Metabolomics* **12**, 166 (2016).
36. Dutkiewicz, E. P., Lin, J. D., Tseng, T. W., Wang, Y. S. & Urban, P. L. Hydrogel micropatches for sampling and profiling skin metabolites. *Anal. Chem.* **86**, 2337–2344 (2014).
37. Yang, Y. et al. A laser-engraved wearable sensor for sensitive detection of uric acid and tyrosine in sweat. *Nat. Biotechnol.* **38**, 217–224 (2020).
38. Munje, R. D., Muthukumar, S., Jagannath, B. & Prasad, S. A new paradigm in sweat based wearable diagnostics biosensors using room temperature ionic liquids (RTILs). *Sci. Rep.* **7**, 1950 (2017).
39. Terse-Thakoor, T. et al. Thread-based multiplexed sensor patch for real-time sweat monitoring. *npj Flex. Electron.* **4**, 18 (2020).
40. Ando, H. & Noguchi, R. Dependence of palmar sweating response and central nervous system activity on the frequency of whole-body vibration. *Scand. J. Work Environ. Health* **29**, 216–219 (2003).
41. Kuwayama, K. et al. Time-course measurements of caffeine and its metabolites extracted from fingertips after coffee intake: a preliminary study for the detection of drugs from fingerprints. *Anal. Bioanal. Chem.* **405**, 3945–3952 (2013).
42. Hadorn, B., Hanimann, F., Anders, P., Curtius, H. C. & Halverson, R. Free amino-acids in human sweat from different parts of the body. *Nature* **215**, 416–417 (1967).
43. Begas, E. et al. Effects of short-term saffron (Crocus sativus L.) intake on the in vivo activities of xenobiotic metabolizing enzymes in healthy volunteers. *Food Chem. Toxicol.* **130**, 32–43 (2019).
44. Smit, H. J., Gaffan, E. A. & Rogers, P. J. Methylxanthines are the psycho-pharmacologically active constituents of chocolate. *Psychopharmacology* **176**, 412–419 (2004).
45. Landolt, H. P. Caffeine, the circadian clock, and sleep. *Science* **349**, 1289–1289 (2015).
46. Weibel, J. et al. Caffeine-dependent changes of sleep-wake regulation: evidence for adaptation after repeated intake. *Prog. Neuro Psychopharmacol. Biol. Psychiatry* **99**, 109851 (2020).
47. Lin, Y. S. et al. Daily caffeine intake induces concentration-dependent medial temporal plasticity in humans: a multimodal double-blind randomized controlled trial. *Cereb. Cortex* **31**, 3096–3106 (2021).
48. Cornelis, M. C. et al. Genome-wide association study of caffeine metabolites provides new insights to caffeine metabolism and dietary caffeine-consumption behavior. *Hum. Mol. Genet.* **25**, 5472–5482 (2016).
49. Kamimori, G. H. et al. The rate of absorption and relative bioavailability of caffeine administered in chewing gum versus capsules to normal healthy volunteers. *Int. J. Pharm.* **234**, 159–167 (2002).
50. Csajka, C., Haller, C. A., Benowitz, N. L. & Verotta, D. Mechanistic pharmacokinetic modelling of ephedrine, norephedrine and caffeine in healthy subjects. *Br. J. Clin. Pharmacol.* **59**, 335–345 (2005).
51. Bonati, M. et al. Caffeine disposition after oral doses. *Clin. Pharmacol. Ther.* **32**, 98–106 (1982).
52. Lelo, A., Miners, J. O., Robson, R. A. & Birkett, D. J. Quantitative assessment of caffeine partial clearances in man. *Br. J. Clin. Pharmacol.* **22**, 183–186 (1986).
53. Lelo, A., Birkett, D. J., Robson, R. A. & Miners, J. O. Comparative pharmacokinetics of caffeine and its primary demethylated metabolites paraxanthine, theobromine and theophylline in man. *Br. J. Clin. Pharmacol.* **22**, 177–182 (1986).
54. Elpa, D. P., Chiu, H. Y., Wu, S. P. & Urban, P. L. Skin metabolomics. *Trends Endocrinol. Metab.* **32**, 66–75 (2021).
55. Samant, P. P. & Prausnitz, M. R. Mechanisms of sampling interstitial fluid from skin using a microneedle patch. *Proc. Natl Acad. Sci. USA* **115**, 4583–4588 (2018).
56. Nunes, M. J., Cordas, C. M., Moura, J. J. G., Noronha, J. P. & Branco, L. C. Screening of potential stress biomarkers in sweat associated with sports training. *Sports Med. Open* **7**, 8 (2021).
57. Seger, C. & Salzmann, L. After another decade: LC-MS/MS became routine in clinical diagnostics. *Clin. Biochem.* **82**, 2–11 (2020).
58. Menikarachchi, L. C., Hamdalla, M. A., Hill, D. W. & Grant, D. F. Chemical structure identification in metabolomics: computational modeling of experimental features. *Comput. Struct. Biotechnol. J.* **5**, e201302005 (2013).
59. Wolrab, D., Fruhauf, P. & Gerner, C. Quantification of the neurotransmitters melatonin and N-acetyl-serotonin in human serum by supercritical fluid chromatography coupled with tandem mass spectrometry. *Anal. Chim. Acta* **937**, 168–174 (2016).

60. de Almeida, E. A. et al. Measurement of melatonin in body fluids: standards, protocols and procedures. *Childs Nerv. Syst.* **27**, 879–891 (2011).
61. Carter, M. D., Calcutt, M. W., Malow, B. A., Rose, K. L. & Hachey, D. L. Quantitation of melatonin and n-acetylserotonin in human plasma by nanoflow LC-MS/MS and electrospray LC-MS/MS. *J. Mass Spectrom.* **47**, 277–285 (2012).
62. Ferre, S. Mechanisms of the psychostimulant effects of caffeine: implications for substance use disorders. *Psychopharmacology* **233**, 1963–1979 (2016).
63. Favari, C. et al. Metabolomic changes after coffee consumption: new paths on the block. *Mol. Nutr. Food Res.* **65**, 2000875 (2021).
64. Jagannath, A. et al. Adenosine integrates light and sleep signalling for the regulation of circadian timing in mice. *Nat. Commun.* **12**, 2113 (2021).
65. Conlay, L. A., Conant, J. A., deBros, F. & Wurtman, R. Caffeine alters plasma adenosine levels. *Nature* **389**, 136 (1997).
66. Muqaku, B. et al. Coffee consumption modulates inflammatory processes in an individual fashion. *Mol. Nutr. Food Res.* **60**, 2529–2541 (2016).
67. Hasko, G. & Cronstein, B. Regulation of inflammation by adenosine. *Front. Immunol.* **4**, 85 (2013).
68. Renner, E., Wietholtz, H., Huguenin, P., Arnaud, M. J. & Preisig, R. Caffeine—a model-compound for measuring liver-function. *Hepatology* **4**, 38–46 (1984).
69. Harshman, S. W. et al. Rate normalization for sweat metabolomics biomarker discovery. *Talanta* **223**, 121797 (2021).
70. Tyanova, S. et al. The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nat. Methods* **13**, 731–740 (2016).
71. Ho, J., Tumkaya, T., Aryal, S., Choi, H. & Claridge-Chang, A. Moving beyond P values: data analysis with estimation graphics. *Nat. Methods* **16**, 565–566 (2019).
72. Sonner, Z. et al. The microfluidics of the eccrine sweat gland, including biomarker partitioning, transport, and biosensing implications. *Biomicrofluidics* **9**, 031301 (2015).
73. Taylor, N. A. & Machado-Moreira, C. A. Regional variations in transepidermal water loss, eccrine sweat gland density, sweat secretion rates and electrolyte composition in resting and exercising humans. *Extrem. Physiol. Med.* **2**, 4 (2013).
74. Grzegorzewski, J. et al. PK-DB: pharmacokinetics database for individualized and stratified computational modeling. *Nucleic Acids Res.* **49**, D1358–D1364 (2021).
75. Virtanen, P. et al. SciPy 1.0: Fundamental algorithms for scientific computing in python. *Nat. Methods* **17**, 261–272 (2020).
76. Barron, J. T. A general and adaptive robust loss function. *Proc. CVPR IEEE* 4326–4334 https://doi.org/10.1109/CVPR.2019.00446 (2019).
77. Haug, K. et al. MetaboLights: a resource evolving in response to the needs of its scientific community. *Nucleic Acids Res.* **48**, D440–D444 (2020).
78. Brunmair, J. et al. Finger sweat analysis enables short interval metabolic biomonitoring in humans. *Github* https://doi.org/10.5281/zenodo.5222967 (2021).

Acknowledgements

We acknowledge support by the Mass Spectrometry Centre of the Faculty of Chemistry, University of Vienna and the Joint Metabolome Facility, University of Vienna and Medical University of Vienna. Both facilities are members of the Vienna Life Science Instruments (VLSI). We would also like to thank the Hochschuljubiläumsstiftung (HJS) for their financial support during the course of this research project. We are grateful to Lukas Skos for help with HepG2 experiments. Figure 1 was created with the help of Servier Medical Art (https://smart.servier.com) and the schematic of the Vanquish UPLC and Q Exactive HF has been created with BioRender (www.BioRender.com).

Author contributions

J.B. performed research, interpreted data, analysed data and wrote the manuscript, M.G. performed research, analysed and interpreted data, L.N. performed research and analysed data, A.B. interpreted data and wrote the manuscript, B.N. performed research, A.S. performed research, L.J. performed research and analysed data, M.L.F. performed research and analysed data, C.L. performed research and analysed data, J.Z. analysed data and interpreted data, S.M.M. performed research, analysed, interpreted data and wrote the manuscript, C.G. conceptualised the project, interpreted data and wrote the manuscript.

Competing interests

The authors declare no competing interests.

Ethics approval

The study protocol was performed in accordance with the University of Vienna and has been approved by the ethical committee of the University of Vienna (reference number 00337). Written informed consent has been obtained from all volunteers participating in this study. The informed consent covers an information sheet about the purpose of the study, a questionnaire for minimal personal information and certificate of consent.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-021-26245-4.

Correspondence and requests for materials should be addressed to Christopher Gerner.

Peer review information *Nature Communications* thanks the anonymous reviewers for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

A.2. Publication II: Probabilistic quotient's work and pharmacokinetics' contribution: countering size effect in metabolic time series measurements.

Mathias Gotsmy†, Julia Brunmair†, Christoph Büschl, Christopher Gerner, and Jürgen Zanghellini

My role was shared first author. I conceived the idea, implemented the code, drafted the manuscript and compiled the figures.

BMC Bioinformatics

RESEARCH

Open Access

Probabilistic quotient's work and pharmacokinetics' contribution: countering size effect in metabolic time series measurements

Mathias Gotsmy[1,2†], Julia Brunmair[1†], Christoph Büschl[1], Christopher Gerner[1,3] and Jürgen Zanghellini[1*]

[†]Mathias Gotsmy and Julia Brunmair contributed equally

*Correspondence: juergen.zanghellini@univie.ac.at

[1] Department of Analytical Chemistry, Faculty of Chemistry, University of Vienna, Vienna, Austria
[2] Vienna Doctoral School in Chemistry, University of Vienna, Vienna, Austria
[3] Joint Metabolome Facility, University and Medical University of Vienna, Vienna, Austria

Abstract

Metabolomic time course analyses of biofluids are highly relevant for clinical diagnostics. However, many sampling methods suffer from unknown sample sizes, commonly known as size effects. This prevents absolute quantification of biomarkers. Recently, several mathematical post acquisition normalization methods have been developed to overcome these problems either by exploiting already known pharmacokinetic information or by statistical means. Here we present an improved normalization method, MIX, that combines the advantages of both approaches. It couples two normalization terms, one based on a pharmacokinetic model (PKM) and the other representing a popular statistical approach, probabilistic quotient normalization (PQN), in a single model. To test the performance of MIX, we generated synthetic data closely resembling real finger sweat metabolome measurements. We show that MIX normalization successfully tackles key weaknesses of the individual strategies: it (i) reduces the risk of overfitting with PKM, and (ii), contrary to PQN, it allows to compute sample volumes. Finally, we validate MIX by using real finger sweat as well as blood plasma metabolome data and demonstrate that MIX allows to better and more robustly correct for size effects. In conclusion, the MIX method improves the reliability and robustness of quantitative biomarker detection in finger sweat and other biofluids, paving the way for biomarker discovery and hypothesis generation from metabolomic time course data.

Keywords: Metabolomics, Finger Sweat, Blood Plasma, PKM, PQN

Introduction

In recent years, the analysis of the sweat metabolome has received increased attention from several fields of study [1–3]. For example, sweat has been in the focus of forensic scientists since it is possible to analyze metabolomic profiles of finger-prints that have been found (e.g., at a crime scene) [4]. Also, drug testing can easily be performed on sweat samples. One advantage of this method is to not only identify already illegal substances but their metabolic degradation products as well, thereby allowing scientist to distinguish between drug consumption and mere contact [1]. Another application of sweat metabolomics is in diagnostics for personalized medicine, where the focus is put on discerning metabolic states of the body and trying to optimize nutrition and treatment based upon information of biomarkers in sweat [5–7].

Sweat metabolomics offers several technical advantages. Firstly, sweat is a rich source of biomolecules and thus offers great potential for biomarker discovery [8, 9]. Secondly, sweat sampling is easy compared to sampling other biofluids (e.g., blood or urine). Moreover, it is non-invasive and can, in principle, be rapidly repeated.

Several sampling methods have been developed [2, 3, 9, 10]. However, most of them work in a very similar manner: a water absorbing material is put onto the skin's surface to collect sweat for some (short) time. Sweat metabolites are subsequently extracted from this material and analyzed [3, 10]. Methods differ, however, in if and how they induce sweating. Some methods induce increased sweating by physical exercise [9] or chemical stimulation [2], whereas in other studies no sweat induction is performed and the natural sweat rate is sufficient for metabolomic analysis [3, 11].

Regardless of the exact sampling method, most of the above mentioned studies suffer one major drawback. The sweat flux is highly variable, depending not only on interindividual differences but also on body location, temperature, humidity, exercise, and further factors that may change multiple times over the course of one day [12, 13]. For example, even with conservative estimates a variability of sweat flux, q_{sweat}, on the finger-tips between 0.05 and 1 mg cm^{-2} min^{-1} needs to be accounted for [13–16]. This is a major challenge for comparative or quantitative studies, which has been acknowledged by many, e.g. [1, 4, 8, 17–19], however only actively approached by few – most notably [9].

The key problem is associated to the fact that often one is interested in the true metabolite concentrations, $\mathbf{C} \in \mathbb{R}^{n_{\text{metabolites}}}$, of $n_{\text{metabolites}}$ metabolites, which is obscured by an unknown and time-dependent sweat flux. Thus, the measured metabolites' intensities are not proportional to $\mathbf{C}$ but to the metabolite mass vector, $\widetilde{\mathbf{M}} \in \mathbb{R}^{n_{\text{metabolites}}}$,

$$\widetilde{\mathbf{M}}(t) = a_{\text{sample}} \int_{t-\tau}^{t} \mathbf{C}(t')\, q_{\text{sweat}}(t')\, \mathrm{d}t'. \tag{1}$$

Here a_{sample} and τ denote the surface area of skin that is sampled and the time it takes to collect one sample, respectively. We emphasize that throughout the manuscript, the mass of a metabolite is defined as the measured abundance of the metabolite in a measured sample and neither as the molar mass or mass to charge ratio. Moreover, we acknowledge that without a calibration curve, the measured abundances have an arbitrary peak-area unit and are thus strictly neither absolute masses nor concentrations. The proportionality constant that scales measured intensities to mass units is determined by the calibration curve. The proper calibration curve is not further discussed here but is assumed to be linear and available when applicable.

Metabolic concentration shifts happen in the span of double-digit minutes to hours, whereas sampling times are usually low single-digit minutes, therefore it is possible to assume that $\mathbf{C}$ changes little over the integration time τ [20]. Thus (1) simplifies to

$$\widetilde{\mathbf{M}}(t) \approx \mathbf{C}(t)\ V(t), \tag{2a}$$

with an unknown sweat volume during sampling

$$V(t) := a_{\text{sample}} \int_{t-\tau}^{t} q_{\text{sweat}}(t')\,\mathrm{d}t', \tag{2b}$$

and the problems reads: given $\widetilde{\mathbf{M}}$, how can we compute $\mathbf{C}$ if we don't know V?

The need to calculate absolute metabolite concentrations from small biological samples of unknown volume is not unique to sweat metabolomics but known throughout untargeted metabolomics. The problem is commonly referred to as size effects [21]. For the sake of consistency with previous publications on this topic, we will use the term "size effects" throughout this publication. We emphasize that in this context, it specifically refers to perceived differences in measured abundances due to changing sample volumes and/or dilutions and not to effects of different numbers of measurements per sample, also referred to as sample size effects [22].

Three strategies have been developed to tackle size effects:

Direct sweat volume measurement. Measuring V, for instance via microfluidics [9, 23, 24], is the most straight forward method to solve (2) and typically very accurate with minimally required volumes in the range of ~ 5 to $100\,\mu$L [9, 23, 24]. However, in the case of sweat sampling, it may take quite some time, large sample areas, or increased (i.e., induced) sweating to collect enough sweat for robust volume quantification. Another alternative is the volume estimation via paired standards [25], however, such a method increases the complexity of sample preparation. Either option would impede fast and easy sample collection and analysis.

Indirect sweat volume computation. If the chemical kinetics of targeted metabolite concentrations are known, then kinetic parameters and the sweat volume at each time point can be simultaneously determined by fitting the measured mass vector to Eq. 2. Recently, we used this strategy to computationally resolve not only sample volumes in the nL to single digit μL-range but also accurately quantify personalized metabolic response patterns upon caffeine ingestion [20]. Albeit feasible for the determination of individual differences with knowledge of reaction kinetics, this method quickly becomes unconstrained when too little prior information is available. Therefore, it is not suited for the discovery of unknown reaction kinetics. Moreover, this method requires several sampling time points to allow modeling the kinetics of different metabolites, thereby decreasing the simplicity of sampling.

Statistical normalization. With this approach the aim is to normalize the mass vector by the apparent mass of a marker that scales proportionally to the sample volume so that the ratio becomes (at least approximately) independent of the sample volume. Various strategies have been developed for untargeted metabolomics; for example, normalization by total measured signal [26], and singular value decomposition-based normalization [27]. However, one of the best performing methods – probabilistic quotient normalization (PQN) – simply assumes that the median of the ratio of two apparent mass vectors is proportional to the sample volume [21, 28–30]. Although PQN does not allow one to compute sample volumes *per se*, it enables one to assess differential changes [28].

In this study, we explore the performance of three different normalization methods on synthetic data. We illustrate the disadvantages of two previously published methods only focusing on either targeted or untargeted metabolites, respectively. A third normalization method is developed by combining both strategies in a single MIX model. We show that MIX significantly outperforms its preceding normalization methods. To validate the results, we use MIX to characterize caffeine metabolization measured in the finger sweat as well as diphenhydramine metabolization measured in blood plasma.

Theory

Probabilistic quotient normalization

Definition. Probabilistic quotient normalization (PQN) assumes that for a large, untargeted set of metabolites the median metabolite concentration fold change between two samples (e.g., two measured time points, t_r and t_s) is approximately 1,

$$Q^{\mathbf{C}} = \text{median}\left\{\frac{C_j(t_r)}{C_j(t_s)}\right\} \approx 1, \quad j \in [1, n_{\text{metabolites}}]. \tag{3a}$$

Consequently, fold changes calculated from $\widetilde{\mathbf{M}}$ instead of $\mathbf{C}$ are proportional to the ratio of V,

$$Q^{\mathbf{M}} = Q^{\mathbf{C}}\frac{V(t_r)}{V(t_s)} \approx \frac{V(t_r)}{V(t_s)} \tag{3b}$$

with

$$Q^{\mathbf{M}} = \text{median}\left\{\frac{\widetilde{M}_j(t_r)}{\widetilde{M}_j(t_s)}\right\}, \quad j \in [1, n_{\text{metabolites}}]. \tag{3c}$$

In order to minimize the influence of experimental errors

$$M_j^{\text{ref}} = \text{median}\left\{\widetilde{M}_j(t_i)\right\}, \quad i \in [1, n_{\text{time points}}] \tag{4}$$

often replaces the dedicated sample in $\widetilde{M}_j(t_s)$ in the denominator of Eq. 3c [28]. Therefore, the normalization quotient by PQN is calculated as

$$Q^{\text{PQN}}(t) = \text{median}\left\{\frac{\widetilde{M}_j(t)}{M_j^{\text{ref}}}\right\}, \quad j \in [1, n_{\text{metabolites}}]. \tag{5}$$

Q^{PQN} is a relative measure and distributes around 1. In analogy to Eq. 3b, we define its relation to the (sweat) volume V^{PQN} as

$$Q^{\text{PQN}}(t) = \tfrac{V^{\text{PQN}}(t)}{V^{\text{ref}}}, \tag{6}$$

where V^{ref} denotes some unknown, time-independent reference (sweat) volume. Note that with real data only $Q^{\text{PQN}}(t)$ values can be calculated, but $V^{\text{PQN}}(t)$ as well as V^{ref} remain unknown.

Discussion. M_j^{ref} can be defined differently depending on the underlying data. However, the choice of reference is usually not critical to the outcome of PQN [28]. As no control or blank measurements are available, and the abundances of metabolites can range several orders of magnitudes, in this study, we used a metabolite-wise median reference for Q^{PQN} calculation. Moreover, PQN might be sensitive to missing values;

however, in this study, we only focused on (real and synthetic) data sets where 100% of values were present.

The biggest advantage of PQN is that no calibration curves and prior knowledge about changes over time of measured metabolites are required. Moreover, PQN is independent of the number of sample points measured in a time series. However, its major drawback is that the normalization quotient is not an absolute quantification and only shows relative changes. I.e., it does not quantify V as given in Eq. 2 directly with an absolute value but instead normalizes relative abundances between samples and time points. Another critical assumption is that sweat metabolite concentrations need to be – on average – constant over the sampled time series. Whereas this is reasonable to assume for the sweat of healthy humans [20], one has to take care when investigating disease states (for example, cystic fibrosis, which is known to alter the sweat's composition [31]).

Pharmacokinetic normalization

Definition. In the pharmacokinetic model (PKM) we assume that we know at least the functional dependence, i.e. the pharmacokinetics, but not necessarily the value of the k (pharmaco-)kinetic parameters $\boldsymbol{\theta} \in \mathbb{R}^k$ for $2 \leq \ell \leq n_{\text{metabolites}}$ metabolites. Without loss of generality we (re-)sort $\widetilde{\mathbf{M}}$ such that the first ℓ elements (collected in the vector $\widetilde{\mathbf{M}}_\ell$) correspond to metabolites with known pharmacokinetic dependence, while the remaining $n_{\text{metabolites}} - \ell$ elements (collected in the vector $\widetilde{\mathbf{M}}_{\ell+}$) correspond to metabolites with unknown kinetics. Then Eq. 2 takes the form of

$$\begin{pmatrix} \widetilde{\mathbf{M}}_\ell \quad (t) \\ \widetilde{\mathbf{M}}_{\ell+}(t) \end{pmatrix} = \begin{pmatrix} \mathbf{C}_\ell \quad (t;\boldsymbol{\theta}) \\ \mathbf{C}_{\ell+}(t) \end{pmatrix} V^{\text{PKM}}(t) \tag{7a}$$

with physically meaningful bounds;

$$V_{\text{lower bound}} \leq V^{\text{PKM}}(t) \leq V_{\text{upper bound}}, \tag{7b}$$

$$\boldsymbol{\theta}_{\text{lower bound}} \leq \quad \boldsymbol{\theta} \quad \leq \boldsymbol{\theta}_{\text{upper bound}}. \tag{7c}$$

$V^{\text{PKM}}(t)$ as well as $\boldsymbol{\theta}$ can be obtained by parametric fitting of $\widetilde{\mathbf{M}}_\ell^{\text{PKM}}(t)$. Note that this allows not only to compute absolute values of $\mathbf{C}_\ell^{\text{PKM}}(t;\boldsymbol{\theta})$ but – with $V^{\text{PKM}}(t)$ – also of all other concentrations via $\mathbf{C}_{\ell+}(t) = \widetilde{\mathbf{M}}_{\ell+}(t)/V^{\text{PKM}}(t)$.

As $V^{\text{PKM}}(t_i)$ may be different at every time step t_i, we need to know the (pharmaco-)kinetics of at least two metabolites; otherwise, the number of parameters is larger than the number of data points.

Discussion. The biggest advantage of this method is that it can implicitly estimate absolute values of V without the need for direct measurements. Therefore, sweat volumes can become smaller than the minimum required in volumetric methods, and shorter sampling times also become possible. A drawback of this method is the fact that it is only feasible if one has prior knowledge of relevant pharmacological parameters (i.e., ingested dose of metabolites of interest, volume of distribution, body mass of specimen, range of expected kinetic constants), which is limiting the approach to studies where at least two metabolites together with their pharmacokinetics are well known. Moreover, calibration curves of metabolites of interest and sufficiently many samples in a time series are required for robustly fitting the equation system. In a previously performed sensitivity analysis, an increase in the quality of fit was observed as the number of samples increased from 15 to 20 time points per measured time series [20].

Mixed normalization

Definition. The mixed normalization model (MIX) is a combination of PQN and PKM. It is designed to incorporate robust statistics of untargeted metabolomics via its PQN term as well as an absolute estimation of V via its PKM term.

Optimal parameters of MIX are found via optimization of two equations,

$$T\left[\begin{pmatrix}\widetilde{\mathbf{M}}_{\ell} \quad (t)\\ \widetilde{\mathbf{M}}_{\ell+}(t)\end{pmatrix}\right] = T\left[\begin{pmatrix}\mathbf{C}_{\ell} \quad (t;\boldsymbol{\theta})\\ \mathbf{C}_{\ell+}(t)\end{pmatrix} V^{\mathrm{MIX}}(t)\right] \tag{8a}$$

and

$$ZT\left[\mathbf{Q}^{\mathrm{PQN}}(t)\right] = ZT\left[\mathbf{V}^{\mathrm{MIX}}(t)\right] \tag{8b}$$

where additional transformations T (PKM and PQN term) and scaling Z (PQN term) can be applied to account for random and systematic errors (section "Hyperparameters") and $V^{\mathrm{MIX}}(t)$ and $\boldsymbol{\theta}$ are constrained between physically meaningful bounds,

$$V_{\text{lower bound}} \leq V^{\mathrm{MIX}}(t) \leq V_{\text{upper bound}}, \tag{8c}$$

$$\boldsymbol{\theta}_{\text{lower bound}} \leq \quad \boldsymbol{\theta} \quad \leq \boldsymbol{\theta}_{\text{upper bound}}. \tag{8d}$$

E.g. bounds for V can be calculated by Eq. 2b and minimal and maximal sweat rates from literature.

Discussion. We hypothesize that the MIX model can combine the advantages of PQN and PKM normalization models. Moreover, we believe that MIX inherits the statistical robustness of PQN while simultaneously estimating absolute values as fitted by PKM. Several prerequisites are necessary for normalization with PKM or MIX. However, if they are fulfilled, the improved goodness of normalization by using MIX instead of PKM usually does not come with an additional price as in many metabolomics studies, targeted and untargeted metabolites are measured in combination, and thus, all additional data required by MIX is already available.

Methods

Implementation

A generalized version of PKM and MIX (where an arbitrary number of independent metabolite kinetics can be modeled) was implemented as a Python class. As input it requires the number of metabolites used for kinetic modeling (ℓ), a vector of time points as well as the measured mass data ($\widetilde{\mathbf{M}}$, matrix with time points in the rows and metabolites in the columns). MIX additionally takes a $\mathbf{Q}^{\mathrm{PQN}} = [Q^{\mathrm{PQN}}(t_1), ..., Q^{\mathrm{PQN}}(t_{n_{\text{time points}}})]^{\mathrm{T}}$ vector (calculated with the PQN method from all metabolites, $n_{\text{metabolites}}$) for all time points of a time series. Upon optimization (carried out with `self.optimize_monte_carlo`, which is a wrapper for SciPy's `optimize.curve_fit` [32]) the kinetic constants and sweat volumes are optimized to the measured data by minimizing the functions listed in Eqs. 9b and 9c for PKM and MIX respectively:

$$\min(\mathcal{L}^{\text{MIX}}) = \min(\mathcal{L}^{\text{PKM}} + \mathcal{L}^{\text{PQN}}) \tag{9a}$$

where

$$\mathcal{L}^{\text{PKM}} = \sum_{i=1}^{n_{\text{time points}}} \sum_{j=1}^{n_{\text{metabolites}}} L\left[\lambda\left(T(\widetilde{M}_{ij}) - T(C_{ij}\; V_i^{\text{MIX}})\right)^2\right], \tag{9b}$$

$$\mathcal{L}^{\text{PQN}} = \sum_{i=1}^{n_{\text{time points}}} L\left[(1-\lambda)\left(ZT(\mathbf{V}^{\text{MIX}})_i - ZT(\mathbf{Q}^{\text{PQN}})_i\right)^2 \text{Var}(T(\mathbf{V}^{\text{MIX}}))\right], \tag{9c}$$

$\text{Var}(\mathbf{V})$ is the variance of $\mathbf{V}$ (which is the vector of estimated V over all time points), T is a transformation function, Z is a scaling function, and L is the loss function. The key difference between PKM and MIX is that the fitted V in MIX are biased towards relative abundances as calculated by PQN. An important additional hyperparameter of the MIX model is λ, which weights the error residuals of $\mathcal{L}^{\text{PKM}}$ and $\mathcal{L}^{\text{PQN}}$. Its calculation is discussed in section "Hyperparameters". If $\lambda = 1$, the MIX model simplifies again to a pure PKM model.

To summarize, an overview of the differences between PKM and MIX models is given in Additional file 1: Table S1 and a flow chart of data processing for MIX normalization is given in Fig. 1.

Hyperparameters

Several hyperparameters can be set for the PKM and MIX Python classes.

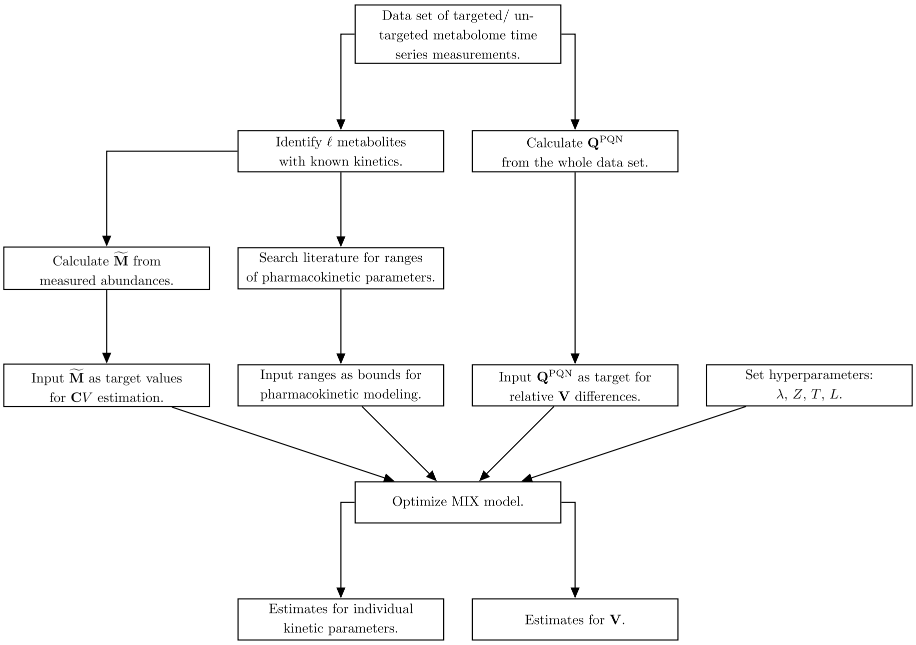

Data set of targeted/ untargeted metabolome time series measurements.
Identify ℓ metabolites with known kinetics.
Calculate $\mathbf{Q}^{\text{PQN}}$ from the whole data set.
Calculate $\widetilde{\mathbf{M}}$ from measured abundances.
Search literature for ranges of pharmacokinetic parameters.
Input $\widetilde{\mathbf{M}}$ as target values for $\mathbf{C}V$ estimation.
Input ranges as bounds for pharmacokinetic modeling.
Input $\mathbf{Q}^{\text{PQN}}$ as target for relative $\mathbf{V}$ differences.
Set hyperparameters: λ, Z, T, L.
Optimize MIX model.
Estimates for individual kinetic parameters.
Estimates for $\mathbf{V}$.

Fig. 1 Flow chart for data processing for MIX normalization

Kinetic function. Firstly, it is possible to choose the kinetic function used to calculate **C**. In this study we focused on a modified Bateman function *F*(*t*) with 5 kinetic parameters (k_a, k_e, c_0, lag, d):

$$F(t) = \begin{cases} b(t) + d & \text{if } b(t) \geq 0 \\ d & \text{if } b(t) < 0 \end{cases} \tag{10a}$$

with

$$b(t) = c_0 \, \frac{k_a}{k_e - k_a} \left(e^{-k_a (t-lag)} - e^{-k_e (t-lag)} \right). \tag{10b}$$

This function was designed to be flexible and able to represent several different metabolite consumption and production kinetics, as exemplified by Fig. 2. Intuitively, k_a and k_e correspond to kinetic constants of absorption and elimination of a metabolite of interest with the unit h^{-1}. c_0 is the total amount of a metabolite absorbed over the volume of distribution with the unit mol L^{-1}. Additionally to these parameters which are also part of the classical Batman function [33], we here introduce *lag* and *d*. The *lag* term with the unit h shifts the function along the X-axis, intuitively defining the starting time point of absorption of a metabolite of interest, whereas the *d* term with the unit mol L^{-1} shifts the function along the Y-axis.

Loss function, L. L calculates the loss value after estimation of the error residuals of the model (Eq. 9). It can be set via `self.set_loss_function` to either `cauchy_loss` or `max_cauchy_loss` (or `max_linear_loss`). In both cases the loss is calculated as a Cauchy distribution of error residuals according to SciPy [32]. The difference, however, is that `cauchy_loss` only uses the absolute error residuals, whereas `max_cauchy_loss` uses the maximum of relative and absolute error residuals (thus the word `max` is expressed in its name). The reason for its addition was that a good performance has been

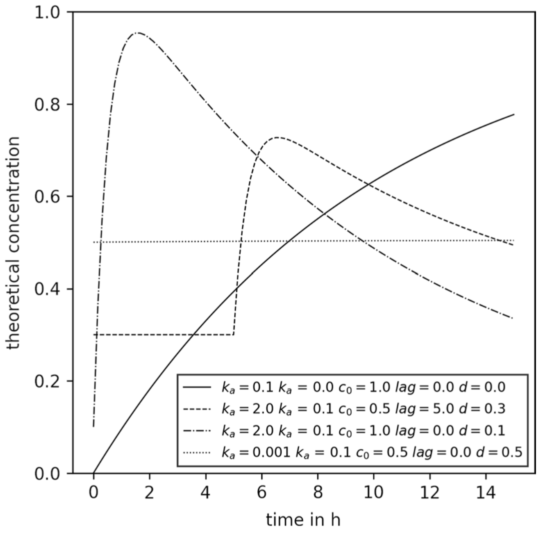


Fig. 2 Examples of concentration time series that can be modeled with the modified Bateman equation used. The legend shows the kinetic parameters used to create the respective curves. All parameters are within the bounds that were used for kinetic parameter fitting

achieved in a previous study [20]. In this study we used the `max_cauchy_loss` loss function for PKM models and `cauchy_loss` for MIX models. The choice of L is intertwined with the choice of T which becomes clear in the following paragraph.

Transformation function, T. T transforms the measured data $\widetilde{\mathbf{M}}$ as well as the calculated $\mathbf{Q}^{\mathrm{PQN}}$, $\mathbf{C}V$, and $\mathbf{V}$ before calculation of the loss (Eq. 9). Two different transformations, `none` and `log10`, can be set during initialization with the argument `trans_fun`. As originally reported [20] no transformation was done for PKM (i.e. `trans_fun='none'`),

$$T(\widetilde{\mathbf{M}}) = \widetilde{\mathbf{M}}. \tag{11a}$$

For MIX models, however, a log-transform was performed (i.e. `trans_fun='log10'`),

$$T(\widetilde{\mathbf{M}}) = \log_{10}(\widetilde{\mathbf{M}} + 10^{-8}) \tag{11b}$$

as the error on measured data is considered multiplicative [34] and the sweat volume log-normally distributed (Additional file 1: Fig. S1). To avoid problems with concentrations of the size 0 a small number (i.e., the size of optimizer precision [32]) is added.

In a sensitivity analysis study, we tested the quality of normalization of MIX with different L and T hyperparameters and concluded that a combination of `cauchy_loss` for L and `log10` for T performed best (Additional file 1: Fig. S2C, D). This is in agreement with literature where logarithmic transformations performed well in combination with PQN for size effect normalization of sweat measurements [35].

Scaling function, Z. Z describes a scaling function performed on $T(\mathbf{Q}^{\mathrm{PQN}})$ and $T(\mathbf{V})$. Scaling is performed to correct for noisy data (see Results section "In fluence of noise on PQN"). Two strategies can be set with the `scale_fun` argument during initialization of the MIX model class, `standard` or `mean`. In this study, all MIX models employ standard scaling, i. e.

$$ZT(\mathbf{Q}^{\mathrm{PQN}}) = \frac{T(\mathbf{Q}^{\mathrm{PQN}}) - \mathrm{mean}(T(\mathbf{Q}^{\mathrm{PQN}}))}{\mathrm{Std}(T(\mathbf{Q}^{\mathrm{PQN}}))}. \tag{12a}$$

We additionally implemented `mean` scaling which differs depending on the choice of T with

$$ZT(\mathbf{Q}^{\mathrm{PQN}}) = \begin{cases} T(\mathbf{Q}^{\mathrm{PQN}}) - \mathrm{mean}(T(\mathbf{Q}^{\mathrm{PQN}})) & \text{if } \texttt{trans\_fun='log10'} \\ T(\mathbf{Q}^{\mathrm{PQN}})/\mathrm{mean}(T(\mathbf{Q}^{\mathrm{PQN}})) & \text{if } \texttt{trans\_fun='none'}. \end{cases} \tag{12b}$$

Optimization strategy. The optimization of both PKM and MIX models is done with a Monte Carlo strategy where the initial parameters are sampled randomly from a uniform distribution between their bounds. Performing a sensitivity analysis, we previously showed that this method is preferable to a single fitting procedure [20]. In this study, the number of Monte Carlo replicates for model fitting was set to 100.

Weighting of MIX loss terms. A weighting constant for every measured data point can be used by the model. In a sensitivity analysis study, we found that the choice of λ is not critical to the quality of normalization as long as it is not extremely tilted to one side (i.e., λ close to 0 or 1, Additional file 1: Fig. S2A, B). Thus we propose a method where the loss terms are weighted by the number of data points fitted for each of both loss terms but not by the number of metabolites used in the calculation of each term (Additional file 1: Equation S1). For such a method the solution for λ is given by Eq. 13.

$$\lambda = \frac{1}{\ell + 1} \tag{13}$$

Full and minimal models

In this study, we differentiate between full and minimal models. With full models, we refer to pharmacokinetic normalization models (PKM or MIX) where all metabolites of a given data set are used for the pharmacokinetic normalization. This means that, for example, if $n_{\text{metabolites}} = 20$ all 20 metabolites were modeled with the modified Bateman function and thus in Eqs. 7a and 8a, $\ell = n_{\text{metabolites}}$ and $\widetilde{\mathbf{M}}_{\ell+}$ is an empty vector. On the other hand, minimal models are models where only the few known, better constrained metabolites were modeled with a kinetic function. This means that the information used for $\text{PKM}_{\text{minimal}}$ does not change upon the addition of (synthetic) metabolites. Therefore, its goodness of fit measure should stay constant within statistical variability upon change of $n_{\text{metabolites}}$. This behaviour was used to verify if the simulations worked as intended and if no biases in the random number generation existed. On the other hand, the $\text{MIX}_{\text{minimal}}$ model still gained information from the increase of $n_{\text{metabolites}}$ as the PQN part of this model was calculated with all $n_{\text{metabolites}}$. Therefore, changes in the goodness of fit measures for $\text{MIX}_{\text{minimal}}$ are expected. We emphasize that the definition of full and minimal models is specific to this particular study. Here we explicitly set $\ell = 4$, which originates from previous work where 4 targeted metabolites (caffeine, paraxanthine, theobromine, theophylline) with known kinetics were measured [20].

Synthetic data creation

Three different types of synthetic data sets were investigated. The first two types of data sets (sampled from kinetics, section "Sampled kinetics" and sampled from means and standard deviations, section "Sampled mean and standard deviation") test the behaviour of normalization models in extreme cases (either all metabolites describable by pharmacokinetics or all metabolites completely random). Finally, the third type of data set (sampled from real data, section "Sampled from real data") aims to replicate measured finger sweat data as close as possible. In sum, the performance of normalization methods on all three types of data sets can show how they behave in different situations with different amounts of describable data.

In all three cases, data creation started with a simple toy model closely resembling the concentration time series of caffeine and its degradation products (paraxanthine, theobromine, and theophylline) in the finger sweat as described elsewhere [20]. The respective parameters are listed in Additional file 1: Table S2. With them, the concentration of metabolites #1 to #4 were calculated for 20 time points (between 0 and 15 h in equidistant intervals, Fig. 3). Subsequently, new synthetic metabolite concentration time series were sampled and appended to the toy model (i.e., to the concentration vector, $\mathbf{C}(t)$). Three different synthetic data sampling strategies were tested, and their specific details are explained in the following sections. Next, sweat volumes (V) were sampled from a

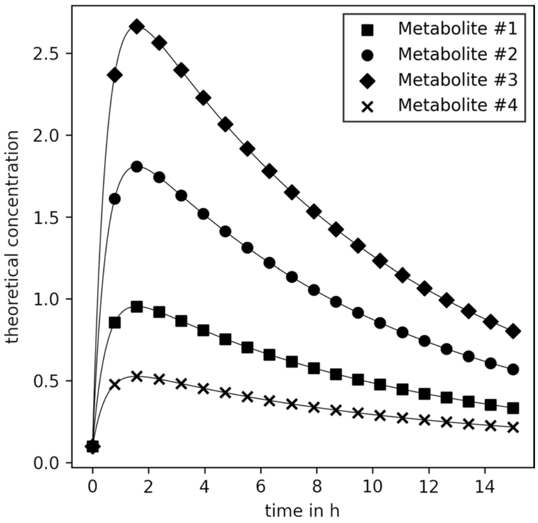


Fig. 3 Theoretical concentration **C** for the first four metabolites of the synthetic data. Kinetic parameters used for calculation are listed in Additional file 1: Table S2

log-normal distribution truncated at $(0.05 \leq V \leq 4\,\mu\mathrm{L})$ closely resembling the distribution of sweat volumes estimated in our previous publication [20], Additional file 1: Fig. S1. Finally, an experimental error (ϵ) was sampled for every metabolite and time point from a normal distribution with a coefficient of variation of 20% and the synthetic data was calculated as

$$\widetilde{\mathbf{M}}(t) = \operatorname{diag}\big(\mathbf{C}(t)\; V(t)\big)\; \boldsymbol{\epsilon}(t). \tag{14}$$

For every tested condition, 100 synthetic data replicates were generated, and the normalization models were fitted.

Sampled kinetics

In simulation v1, data was generated by sampling kinetic parameters for new metabolites from an uniform distribution. The distribution was constrained by the same bounds also used for the PKM and MIX model fitting: $(0,0,0,0)^{\mathrm{T}} \leq (k_a, k_e, c_0, lag, d)^{\mathrm{T}} \leq (3,3,5,15,3)^{\mathrm{T}}$. Subsequently the concentration time series of the synthetic metabolites were calculated according to the modified Bateman function (Eq. 10).

Sampled mean and standard deviation

Means and standard deviations of the concentration time series of metabolites were calculated from untargeted real finger sweat data (for details, see section "Real finger sweat metabolome data"). The probability density function of both can be described by a log-normal distribution (Additional file 1: Fig. S3). For the data generation of simulation v2, per added metabolite, one mean and one standard deviation were sampled from the fitted distribution and used as an input for another log-normal distribution from which a random concentration time series was subsequently sampled. This results in synthetic concentration values that behave randomly and, therefore, cannot be easily described by our pharmacokinetic models.

Sampled from real data

To get an even better approximation to real data, in simulation v3, concentration time series were directly sampled from untargeted real finger sweat data (for details, see section "Real finger sweat metabolome data"). To do so, the untargeted metabolite $\widetilde{\mathbf{M}}$ time series data set was normalized with PQN. As the number of metabolites in this data set was comparably large ($n_{\text{metabolites}} = 3446$) we could assume that the relative error (or rRMSE, for more explanation, see section "Synthetic data simulations") was negligibly small. The resulting values are, strictly speaking, fractions of concentrations. However, this does not affect the results as these values are anyways considered untargeted (i.e., no calibration curve exists) and thus relative. Therefore, the PQ normalized data set could be used as ground truth for concentration time series sampling. Subsequently, a subset of the original ground truth data was sampled for synthetic data generation.

Sampling of noisy data

We investigated the influence of background (i.e. noisy) signal on the performance on $\mathbf{Q}^{\text{PQN}}$ (and scaled and transformed variants thereof). To simulate such an environment we used data sampled from real data (section "Sampled from real data"), and applied V only to a fraction of the $\mathbf{C}$ vector,

$$\begin{pmatrix} \widetilde{\mathbf{M}}\ \ (t) \\ \widetilde{\mathbf{M}}_n(t) \end{pmatrix} = \text{diag}\begin{pmatrix} \mathbf{C}\ \ (t)V(t) \\ \mathbf{C}_n(t) \end{pmatrix} \boldsymbol{\epsilon}. \tag{15}$$

The noise fraction is given by the number of elements of $\widetilde{\mathbf{M}}$ and $\widetilde{\mathbf{M}}_n$ vectors,

$$f_n = \frac{\text{length}(\widetilde{\mathbf{M}}_n)}{\text{length}(\widetilde{\mathbf{M}}) + \text{length}(\widetilde{\mathbf{M}}_n)}, \tag{16}$$

where subscript n in $\widetilde{\mathbf{M}}_n$, $\mathbf{C}_n$, and f_n denotes them as part of the noise.

Simulations were carried out for 20 equidistant noise fractions between $0 \leq f_n \leq 0.95$ with $n_{\text{metabolites}} = 100$ and $n_{\text{time points}} = 20$ for 100 replicates. The error residuals of mean and standard scaled $\mathbf{Q}^{\text{PQN}}$ are calculated as

$$\text{Mean Scaled Error} = \sum_{i}^{n_{\text{time points}}} \left[ZT(\mathbf{Q}^{\text{PQN}})_i - ZT(\mathbf{V})_i \right] \tag{17a}$$

with Z defined as in Eq. 12b and

$$\text{Standard Scaled Error} = \sum_{i}^{n_{\text{time points}}} \left[ZT(\mathbf{Q}^{\text{PQN}})_i - ZT(\mathbf{V})_i \right] \text{Std}(T(\mathbf{V})) \tag{17b}$$

with Z defined as in Eq. 12a. For both cases T is defined as the logarithm (Eq. 11b). We point out that the multiplication with $\mathrm{Std}(T(\mathbf{V}))$ for the standard scaled error is important to make the results comparable, as otherwise the error would be biased towards the method with smaller scaled standard deviation regardless of the performance of the scaling.

Normalization model optimization

Normalizing for the sweat volume by fitting kinetics through the measured values only has a clear advantage over PQN if it is possible to infer absolute sweat volumes and concentration data. In order to be able to do that, some information about the kinetics and the starting concentrations of metabolites of interest need to be known. For example, when modeling the caffeine network in our previous publication [20], we knew that the *lag* parameter of all metabolites was 0 and that the total amount of caffeine ingested (which corresponds to c_0) was 200 mg. Moreover, we knew that caffeine and its metabolites are not synthesized by humans and implemented the same strategy into our toy model (corresponding to d). As the toy model was designed to resemble such a metabolism, we translated this information to the current study. Therefore, we assumed that the first 4 metabolites in our toy model had known c_0, *lag*, and d parameters. For their corresponding k_a and k_e and the parameters of all other metabolites the bounds were set to the same $(0, 0, 0, 0)^{\mathrm{T}} \leq (k_a, k_e, c_0, lag, d)^{\mathrm{T}} \leq (3, 3, 5, 15, 3)^{\mathrm{T}}$ used in kinetic data generation. Fig. 2 shows examples of concentration time series that can be described with the modified Bateman function and parameters within the fitting bounds.

Real finger sweat metabolome data

The real world finger sweat data was extracted from 37 time series measurements of Study C from ref. [20]. It was downloaded from MetaboLights (MTBLS2772 and MTBLS2776).

Preprocessing. The metabolome data set was split into two parts: targeted and untargeted. The targeted data (i.e., the mass time series data for caffeine, paraxanthine, theobromine, and theophylline) was directly adopted from the mathematical model developed by [36]. This data is available on GitHub (https://github.com/Gotsmy/finger_sweat).

For the untargeted metabolomics part, the raw data was converted to the mzML format with the msConvert tool of ProteoWizard (version 3.0.19228-a2fc6eda4) [37]. Subsequently, the untargeted detection of metabolites and compounds in the samples was carried out with MS-DIAL (version 4.70) [38]. A manual retention time correction was first applied with several compounds present in the majority (more than 90%) of the samples. These compounds were single chromatographic peaks with no isomeric compounds present at earlier or later retention times (m/z 697.755 at 5.57 min, m/z 564.359 at 5.10 min, m/z 520.330 at 4.85 min, m/z 476.307 at 4.58 min, m/z 415.253 at 4.28 min, m/z 371.227 at 3.95 min, m/z 327.201 at 3.56 min, m/z 283.175 at 3.13 min, m/z 239.149 at 3.63 min, m/z 166.080 at 1.69 min, m/z 159.113 at 1.19). After this, untargeted peak detection and automated alignment (after the manual alignment) were carried out with the following settings: Mass accuracy MS1 tolerance: 0.005 Da, Mass accuracy MS2 tolerance: 0.025 Da, Retention time begin: 0.5 min, Retention time end: 6 min, Execute retention time correction: yes, Minimum peak height: 1E5, Mass slice width: 0.01 Da, Smoothing method: Linear weighted moving average, Smoothing level: 3 scans, Minimum peak width: 5 scans, Alignment reference file: `C_D1_I_o_pos_ms1_1.mzML`, Retention time tolerance: 0.3 min, MS1 tolerance: 0.015 Da, Blank removal factor: 5 fold change). No blank-subtraction was carried out as the internal standard caffeine was spiked into each sample, including the blanks. Peak abundances and meta-information were exported with the `Alignment results export` functionality.

Subsequently, we excluded isomers within a *m/z* difference of less than 0.001 Da and a retention time difference of less than 0.5 min. To further reduce features that are potentially background, features with retention times after 5.5 min as well as features with minimal sample abundances of $< 5 \times$ maximum blank abundance (except for the internal standard, caffeine-D9) were excluded from the data set. This was done on a time series-wise basis. Thus the number of untargeted metabolites considered for normalization differs with a mean of 343 ± 152 for the 37 time series of interest.

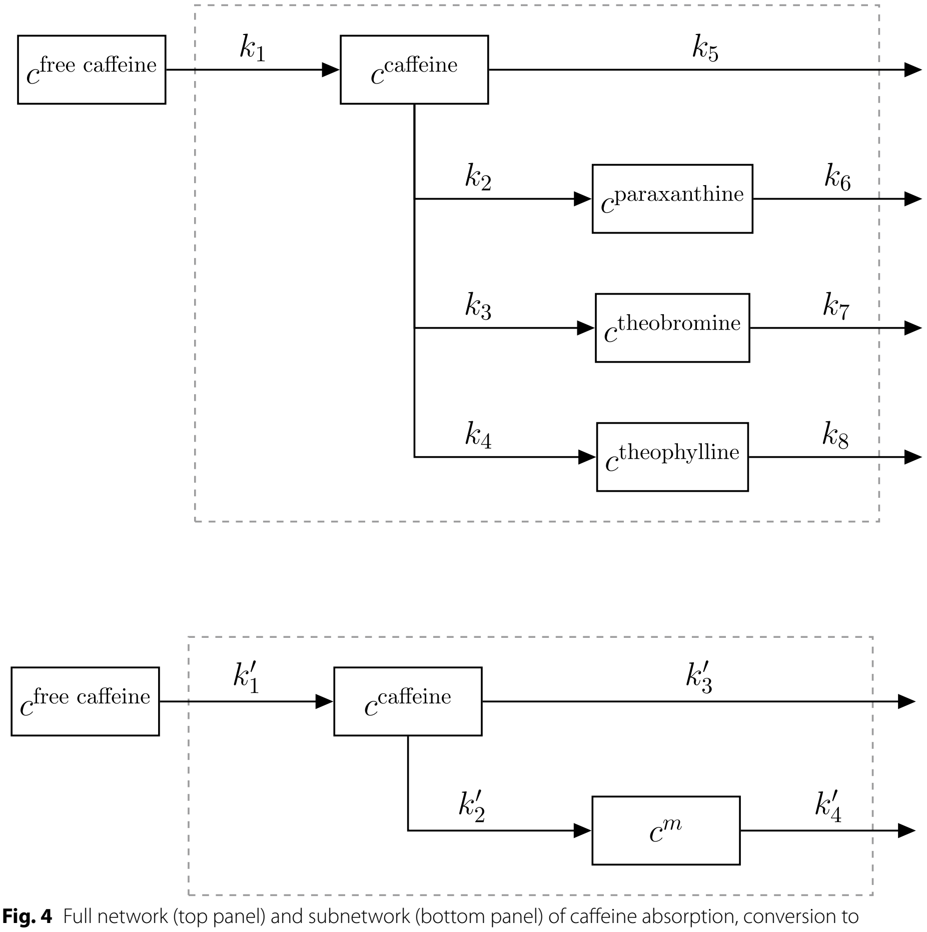


Fig. 4 Full network (top panel) and subnetwork (bottom panel) of caffeine absorption, conversion to paraxanthine, theobromine, and theophylline and their elimination. The system boundary (dashed line) represents the human body. $m \in$ {paraxanthine, theobromine, and theophylline}

Size effect normalization. In this finger sweat data set, time series of targeted as well as untargeted metabolomics, are listed. The kinetics of the four targeted metabolites (caffeine, paraxanthine, theobromine, and theophylline) are known. A reaction network of the metabolites is shown in the top panel of Fig. 4. Briefly, caffeine is first absorbed and then converted into three degradation metabolites. Additionally, all four metabolites are eliminated from the body. All kinetics can be described with first order mass action kinetics [39, 40].

In order to assess the performance of the sweat volume normalization methods, the full network was split up into three subnetworks that all contained caffeine and one degradation metabolite each (Fig. 4 bottom panel). The solution of the first order differential equations describing such network is given in Additional file 1: Eqs. S2a and S2b. Moreover, the 343 ± 152 untargeted metabolite time series were randomly split up into three (almost) equally sized batches, and each batch was assigned to one subnetwork. All three networks were subsequently separately normalized with $\mathrm{PKM}_{\mathrm{minimal}}$ and $\mathrm{MIX}_{\mathrm{minimal}}$ methods with kinetic parameters that were adjusted to the specific reaction network (Fig. 4 bottom panel). Subsequently, the kinetic constants (k'_1, k'_2, k'_3, k'_4) were estimated for 37 measured concentration time series. Fitting bounds were not changed in comparison to the original publication [20].

As all three subnetwork data sets originate from the same finger sweat measurements, the underlying kinetic constants should be exactly identical. As the kinetic constants of absorption ($k_a^{\mathrm{caf}} = k'_1$) and elimination ($k_e^{\mathrm{caf}} = k'_2 + k'_3$) of caffeine are estimated in all three subnetworks, we used their standard deviation to test the robustness of the tested normalization methods.

Real blood plasma metabolome data

In the study of Panitchpakdi et al. [41] the mass time series of the metabolome was measured in different body fluids after the uptake of diphenhydramine (DPH). Here, we focus on data measured in the blood plasma, which includes the abundances of DPH (known kinetics, calibration curve, pharmacological constants) as well as three of its metabolization products (known kinetics) and the abundances of 13526 untargeted metabolites with unknown kinetics.

Preprocessing. The data of peak areas was downloaded from the GNPS platform [42]. To reduce the number of metabolites that are potentially background and/or noise in the data set, features with minimal sample abundances of $< 5 \times$ maximum blank abundance were excluded from the data set on a time series-wise basis. Thus, the number of untargeted metabolites considered for normalization differs with a mean of 1017 ± 114 for the 10 time series of interest.

Size effect normalization. We assume that the kinetics of four metabolites (DPH, N-desmethyl-DPH, DPH N-glucuronide, and DPH N-glucose) can be described by the modified Bateman (Eq. 10). A reaction network of the metabolites is shown in Additional file 1: Fig. S4. Briefly, DPH is first absorbed and then – with unknown intermediates – converted into three degradation metabolites, which are in turn metabolized further downstream or eliminated. c_0 of DPH was calculated with pharmacological constants for bioavailability, volume of distribution, and dosage of DPH as reported in the original publication [41].

Analogously to the normalization performed on finger sweat data, the full network of four metabolites is split up into three subnetworks with only one, shared, targeted metabolite (DPH itself), one additional untargeted metabolite with known kinetic (either N-desmethyl-DPH, DPH N-glucuronide, or DPH N-glucose, Additional file 1: Fig. S5) and one third of 1017 ± 114 untargeted metabolites with unknown kinetics. To ensure better convergence during fitting of the models, the $\widetilde{\mathbf{M}}$ data was first scaled to values between 0 and 1 by dividing by its metabolite-wise maximum. This factor can be multiplied again as part of c_0 after the normalization is done. Thereafter, PKM$_{\text{minimal}}$ and MIX$_{\text{minimal}}$ models were fitted onto the scaled $\widetilde{\mathbf{M}}$ data (with $\ell = 2$) for all ten measured time series. The bounds of parameters were chosen so that previously reported estimates [41] are well within range: $0 \leq k \leq 5\,\text{h}^{-1}$ for $\{k'_1, k'_3\}$, $0 \leq k \leq 1\,\text{h}^{-1}$ for $\{k'_2, k'_4\}$, c_0^{DPH} as reported in the original publication normalized by the maximum factor, $0 \leq c_0^m \leq 300$ for $m \in$ {N-desmethyl-DPH, DPH N-glucuronide, DPH N-glucose} and $lag = d = 0$ as well as $0.01 \leq V \leq 0.03\,\text{mL}$.

As all three subnetwork data sets originate from the same plasma time series measurements, the underlying kinetic constants of DPH should be exactly identical. As the kinetic constants of absorption ($k_a^{\text{DPH}} = k'_1$) and elimination ($k_e^{\text{DPH}} = k'_2$) of DPH are estimated in all three subnetworks we used their standard deviation to test the robustness of PKM$_{\text{minimal}}$ and MIX$_{\text{minimal}}$.

Data analysis

Goodness of normalization. Two goodness of fit measures were calculated to analyze the performance of the tested methods. RMSE is the standard deviation of the residuals of a sampled sweat volume time series vector ($\mathbf{V}^{\text{true}}$) minus the fitted sweat volume vector ($\mathbf{V}^{\text{fit}}$), while rRMSE is the standard deviation of the ratio of sampled and fitted $\mathbf{V}$ vectors normalized by its mean. Intuitively, RMSE is a measure of how much absolute difference there is between the fit and a true value, rRMSE, on the other hand, gives an estimate of how good the fitted sweat volumes are relative to each other. A visual depiction of RMSE and rRMSE is shown in Additional file 1: Fig. S6 and their exact definition is given in the equations in 3.3.

Statistical analysis. The significant differences in the mean of goodness of fit measures were investigated by calculating p values with the non-parametric pairwise Wilcoxon signed-rank test [43] (SciPy's `stats.wilcoxon` function [32]). Significance levels are indicated by *, **, and *** for $p \leq 0.05$, 0.01, and 0.001 respectively.

Results

Comparison of PKM and MIX

Synthetic data simulations

In order to test the performance of different normalization models, we generated 100 synthetic data sets with three different methods (simulations v1, v2, v3) and five different $n_{\text{metabolites}}$ (4, 10, 20, 40, 60) each, where the underlying $\mathbf{C}$, V, and $\boldsymbol{\epsilon}$ values were known. Simulations v1, v2, and v3 differ in the way how $\mathbf{C}$ was generated (kinetic, random, sampled from real data set, respectively). In order to quantify the normalization model

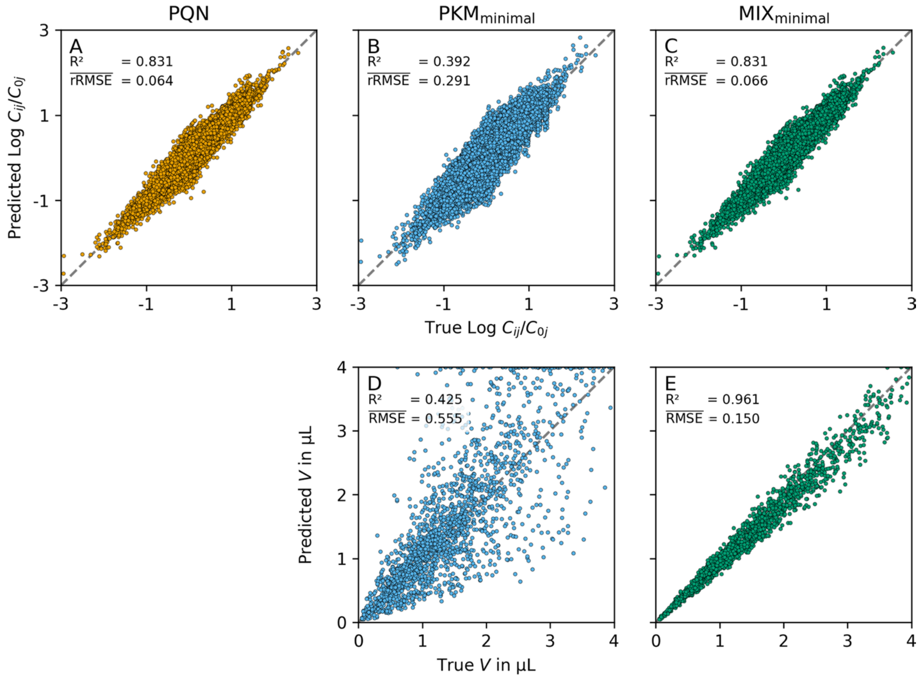


Fig. 5 Relative and absolute normalization performance. In the top row the predicted $\log_{10}(C_j(t_i;\boldsymbol{\theta})/C_j(0;\boldsymbol{\theta}))$ ($i \in \{1, ..., n_{\text{time points}}\}$, $j \in \{1, ..., n_{\text{metabolites}}\}$) are plotted as a function of the true, underlying $\log_{10}(C_j(t_i)/C_j(0))$. The bottom row shows the predicted V as a function of the true, underlying V. The columns represent different normalization models (PQN, PKM$_{\text{minimal}}$, and MIX$_{\text{minimal}}$ from left to right). As no absolute V can be calculated from PQN the bottom left plot is omitted. To illustrate the effect of different RMSE and rRMSE sizes (which both are calculated from V), we show their mean over 100 replicates in comparison to the R^2 values calculated from the points plotted. Intuitively rRMSE is a measure of good correlation on the top row whereas RMSE is a measured of good correlation on the bottom row (high R^2, low rRMSE/RMSE respectively)

performance, two measures of goodness of normalization were used for the analysis of the results: RMSE and rRMSE.

To visualize the obtained normalization performances we plotted the results for simulation v3 and $n_{\text{metabolites}} = 60$ in Fig. 5 for three normalization models (from left to right column, PQN, PKM$_{\text{minimal}}$, and MIX$_{\text{minimal}}$). The top row shows the predicted $\log_{10}(C_j(t_i;\boldsymbol{\theta})/C_j(0;\boldsymbol{\theta}))$ (i.e. the concentration of each metabolite j at each time point i divided by its concentration at time 0) as a function of the true $\log_{10}(C_j(t_i)/C_j(0))$ values. It illustrates the correlation of the relative abundances of one metabolite across all time points. Good correlations (i.e. high R^2) as seen for PQN and MIX$_{\text{minimal}}$ result in a low rRMSE measure. On the bottom row of Fig. 5 the absolute values of predicted V are plotted as a function of the true V. There it becomes evident that good correlations of absolute values result in low RMSE measures.

In the following sections, we will focus on the size of RMSE and rRMSE, respectively, as they are both calculated from the predicted V directly. Note that for PQN, no absolute V can be estimated and, therefore, no RMSE is calculated.

Influence of the number of metabolites. We tracked RMSE and rRMSE of normalization methods for different numbers of metabolites ($n_{\text{metabolites}}$) to investigate how the methods behave with different amounts of available information. An overview of their goodness of normalization measures as a function of $n_{\text{metabolites}}$ on sampled kinetic

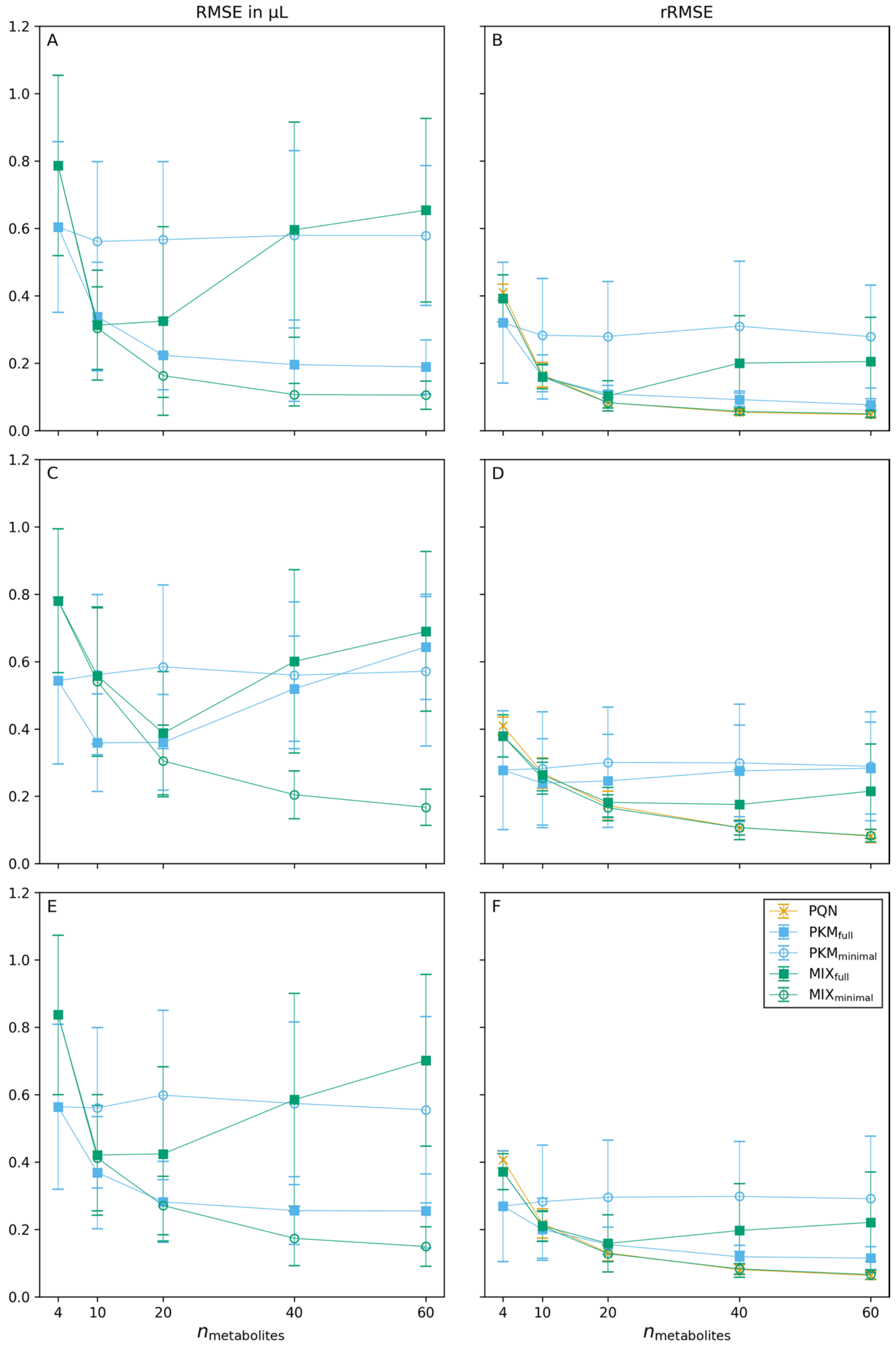


Fig. 6 Goodness of normalization measures of synthetic data simulations. The mean for 100 replicates for different sweat volume normalization models is given for RMSE (left column) and rRMSE (right column). Results for simulations v1, v2, and v3 are shown in rows one, two, and three, respectively. The error bars represent standard deviations of the replicates. For the PQN method no RMSE can be calculated

data (panels A, B), on completely random data (panels C, D), and on sampled subsets of real data (panels E, F) is given in Fig. 6.

PKM_{full} which fits a kinetic function through all possible metabolites ($\ell = n_{\text{metabolites}}$) performs well (low RMSE, low rRMSE) when the **C** data originates from a kinetic function (simulation v1, Fig. 6A, B). However, when the underlying data does not originate from kinetic time series (simulation v2, Fig. 6C, D) its performance is reduced drastically. For PKM_{full} this is resembled in an increase of RMSE (from $0.19 \pm 0.08\,\mu\text{L}$ to $0.64 \pm 0.16\,\mu\text{L}$ for $n_{\text{metabolites}} = 60$) as well as of rRMSE (from 0.08 ± 0.02 to 0.28 ± 0.14 for $n_{\text{metabolites}} = 60$).

Another observation is the behaviour of PQN. Its rRMSE approaches a value close to 0 with increasing $n_{\text{metabolites}}$, indifferently on how the underlying data was generated.

Interestingly, the results from simulation v3 lie between the results from simulations v1 and v2. This gets especially evident when comparing the performance of PKM_{full} in Fig. 6. Such a result suggests that not all of the untargeted metabolites measured are completely random, but some can be described with the modified Bateman function. This leads to the hypothesis that after sweat volume normalization, the real finger sweat data (from which values for v3 were sampled) has a high potential for discoveringunknown kinetics.

Exact numbers for RMSE and rRMSE for all normalization methods and $n_{\text{metabolites}}$ are given in Additional file 1: Tables S3 and S4 respectively. Moreover, pairwise comparisons of RMSE and rRMSE of normalization methods relative to the results from $\text{PKM}_{\text{minimal}}$ are plotted in Additional file 1: Fig. S7.

Statistical testing. As at $n_{\text{metabolites}} = 60$ the goodness of normalization measures started to flatten out, we further investigated this condition for statistical significance. We used the two-sided non-parametric Wilcoxon signed-rank test to compare pairwise differences in RMSE and rRMSE between the tested models. p-values for all combinations are given in Additional file 1: Tables S5 and S6.

As Fig. 6 already indicated, the overall best performance in RMSE as well as rRMSE is observed for the $\text{MIX}_{\text{minimal}}$ model. For $n_{\text{metabolites}} = 60$ it significantly outperforms every other method's RMSE (Fig. 7). Moreover, $\text{MIX}_{\text{minimal}}$'s performance in rRMSE is at least equal to or better than all other tested methods (Additional file 1: Table S6) with one exception: the comparison of rRMSE of $\text{MIX}_{\text{minimal}}$ and PQN in simulation v1 shows significant difference ($p = 0.0029$), however, the absolute values of rRMSE are still very

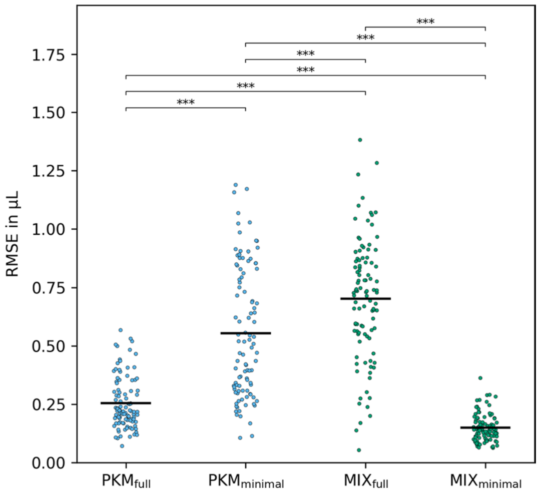


Fig. 7 RMSE measures of simulation v3 with $n_{\text{metabolites}} = 60$. The significance between the methods was calculated on 100 paired replicates with the two-sided Wilcoxon signed-rank test

similar (0.049 ± 0.010 and 0.047 ± 0.009 respectively). Compared to the previously used $\text{PKM}_{\text{minimal}}$ [20], the RMSE of $\text{MIX}_{\text{minimal}}$ improves by $73 \pm 10\,\%$, the rRMSE by $43 \pm 12\,\%$ (Additional file 1: Fig. S7). Analogously to Fig. 7 for simulation v3, the results of simulations v1 and v2 are shown in Additional file 1: Figs. S8 and S9, respectively.

The two-sided version of the Wilcoxon signed-rank test was used to test for any difference between multiple normalization methods. After it became evident that $\text{MIX}_{\text{minimal}}$ performed best, we used a one-sided version of the Wilcoxon signed-rank test to verify if RMSE and rRMSE are significantly decreased by $\text{MIX}_{\text{minimal}}$ compared to all other normalization methods. The resulting p-values are listed in Additional file 1: Table S7. Again, $\text{MIX}_{\text{minimal}}$ significantly outperformed all other tested methods in RMSE and rRMSE except for PQN in any of the simulations.

We, therefore, conclude that normalizing the sweat volume by the $\text{MIX}_{\text{minimal}}$ method reduces the error for the estimated V compared to other tested methods. Compared to PKM, $\text{MIX}_{\text{minimal}}$ has the advantage that its performance does not vary if metabolites' concentration time series can be described with a modified Bateman function (i.e., simulations v1, v2 v3 have little influence on its performance). Therefore, it is especially advantageous if this property cannot be guaranteed.

Computational performance

Analysis of metabolomics data sets is usually a computationally exhaustive process. There are several steps in (pre-)processing that need to be executed, many of them lasting for hours. Therefore, computational time can quickly stack to large numbers. Normalization models are no exception to this general rule. As $n_{\text{metabolites}}$ in a pharmacokinetic model increases, the time for optimization of pharmacokinetic models may become limiting. Therefore, we investigated the average time for one time series normalization for different methods and different numbers of metabolites.

The computational time spent for one optimization step as a function of $n_{\text{metabolites}}$ is given in Fig. 8 for simulation v3. It increased for some normalization models, however

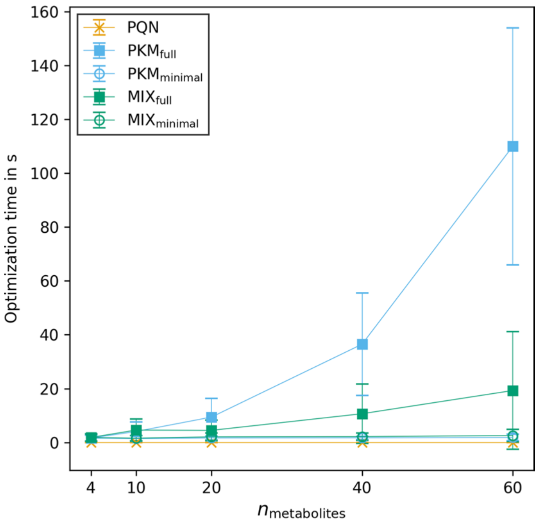


Fig. 8 Time in seconds for optimization of one normalization model in simulation v3. The error bars represent the standard deviation of normalization times between 100 replicates

not for all of them and not equally. Within the investigated range, PQN stays well under 1 second per normalization, whereas with PKM_{full} the normalization time increases drastically from $1.6 \pm 1.1\,\text{s}$ for a model with 4 metabolites to $110 \pm 44\,\text{s}$ for 60 metabolites. Similar normalization times were observed for MIX_{full} maxing out at $19 \pm 22\,\text{s}$ for $n_{\text{metabolites}} = 60$. In stark contrast to the exponential increase in computational power needed for full models are the minimal models. Their time to optimize stays nearly constant ($< 3\,\text{s}$) within the investigated metabolite range (Additional file 1: Table S8).

Here we demonstrate that $\text{MIX}_{\text{minimal}}$ is not only superior to other tested models in terms of its normalization performance but also in terms of computational feasibility. We hypothesize that even data sets with thousands of untargeted metabolites will have a minor impact on its speed.

Comparison of PQN and MIX

Influence of noise on PQN

In untargeted metabolomics, it is often difficult to distinguish between metabolites originating from the actual matrix of interest or from contamination. As PQN includes all untargeted metabolites in its calculation, metabolites stemming from contamination might become a problem as their fold change is independent of the sweat volume, which changes the underlying distributions of quotients. Therefore, we investigated the influence of different fractions of metabolites originating from contamination (i.e., noisy data). Furthermore, we tested if scaling of $\mathbf{Q}^{\text{PQN}}$ values can counteract errors introduced by noise.

Figure 9A demonstrates the problem of using the probabilistic quotient normalization on noisy raw data. The direction of size effects can still be explained when noise is present, however, absolute values of the size effects decrease. Thus, in Fig. 9A, the coefficient of variation (i.e., the standard deviation over the mean) of $\mathbf{Q}^{\text{PQN}}$ is a measure for the average value of the estimated size effect over one synthetically generated time series. As the fraction of noise (f_n, X-axis) increases the coefficient of variation decreases drastically and approaches 0 when $f_n \to 1$.

Figure 9B shows the performance of scaling methods to counteract the reduction of coefficient of variation as described above. The mean scaled error (X-axis) and standard scaled error (Y-axis) as calculated by Eq. 17 are plotted against each other. When $f_n \leq 0.05$, mean scaling outperforms standard scaling. However, thereafter the standard scaled Q^{PQN} is less erroneous than the mean scaled version.

When incorporating $\mathbf{Q}^{\text{PQN}}$ values to the MIX model, it is important to correct for errors introduced by noise. As this result shows that standard scaling reduces the detrimental effect of noise on the calculation of $\mathbf{Q}^{\text{PQN}}$, we used standard scaling throughout the study for MIX normalization. Moreover, this result underlines the good performance of standard scaling in biological data sets [44].

Synthetic data simulations with noise

The synthetic data used for the analysis of section "Comparison of PKM and MIX" did not contain any metabolites that are classified as noise, i.e., their $\widetilde{\mathbf{M}}$ is not influenced by size effects (Eq. 15). This, however, is not necessarily a realistic assumption as there

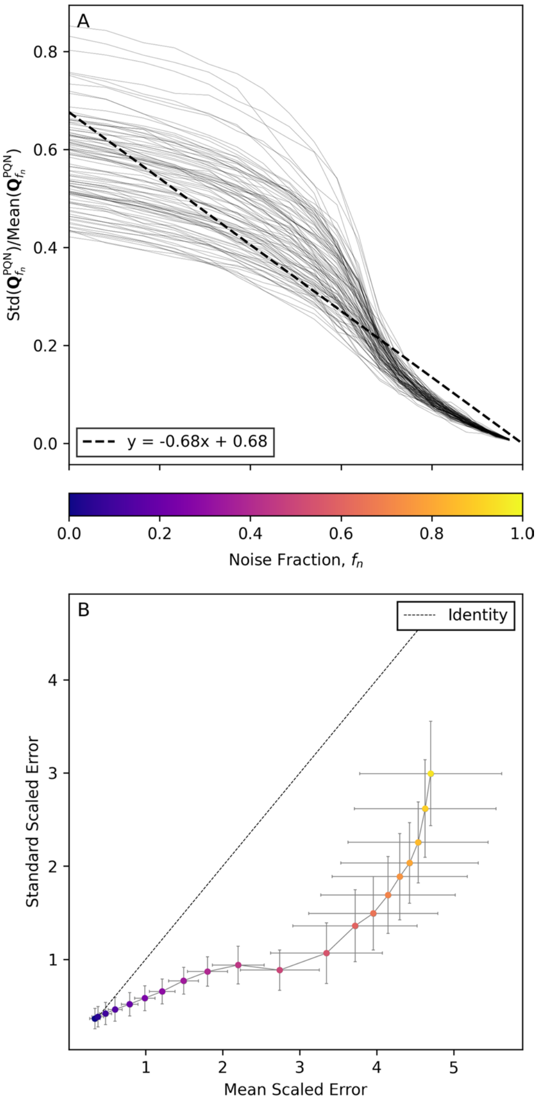


Fig. 9 Influence of the fraction of noisy data on the error of PQN calculation. Panel A illustrates the change of the coefficient of variance of $\mathbf{Q}^{\mathrm{PQN}}$ (Y-axis) as the noise fraction (f_n, Y-axis with the same tick labels as the color bar) increases. Panel B shows the error size of calculated Q^{PQN} to true V with mean scaling (X-axis) and standard scaling (Y-axis). The color of points relates to the noise fraction as depicted in the color bar

are many sources of contaminants in metabolome measurements. Noisy metabolites can be either introduced by biological means (e.g., metabolites that do not originate from sweat but from the surface of the skin in sweat measurements) [45] or by experimental handling [46]. As shown in Fig. 9, this noise in data negatively affects the performance of PQN. Thus, the goodness of PQN in the results of section "Comparison of PKM and MIX" is probably overestimated.

To get a more accurate view of the goodness of normalization of PQN and $\mathrm{MIX}_{\mathrm{minimal}}$, we tested their performance on synthetic data with different fractions of noise, f_n. In order to do so, we created 100 replicates of synthetic data sampled from real data (i.e., simulation v3) for 10 equidistant noise fractions ranging from $f_n = 0$ to $f_n = 0.9$ with $n_{\mathrm{metabolites}} = 60$. In all simulated data, only untargeted metabolites were affected by the introduction of noise, as we assumed that for targeted metabolites (i.e., $\ell = 4$) with known pharmacokinetic behaviour, one can be highly confident that the measurements are not originating from contaminants.

The rRMSE of PQN and $\text{MIX}_{\text{minimal}}$ is plotted in Fig. 10. Only when zero noise was present in the synthetic data set $\text{MIX}_{\text{minimal}}$ did not improve upon PQN. However, as the noise fraction increased, $\text{MIX}_{\text{minimal}}$ significantly outperformed PQN in terms of rRMSE. The *p*-values for all noise fractions are listed in the Additional file 1: Table S9.

The difference of rRMSE between PQN and $\text{MIX}_{\text{minimal}}$ in Fig. 10 is related to the difference of mean and standard scaled errors in Fig. 9B. PQN alone cannot utilize the improved performance of standard scaling as $\text{Std}(T(\mathbf{V}))$ has to be known for its calculation (Eq. 17b). However, when normalizing with $\text{MIX}_{\text{minimal}}$, $\text{Std}(T(\mathbf{V}))$ can be estimated from the pharmacokinetic part of the model (Eq. 9c) significantly improving its quality.

Application to real data

Caffeine network

Previously, we identified and quantified four metabolites (caffeine, paraxanthine, theobromine, and theophylline) in a time series after ingesting a single dose of caffeine [20]. To investigate the performance of normalization models on a real finger sweat data set, we split all measured $\widetilde{\mathbf{M}}$ time series into three parts that contained pairs of targeted metabolites each, only one shared by all, namely caffeine (compare Fig. 4 top and bottom network). Subsequently we fitted a $\text{PKM}_{\text{minimal}}$ and a $\text{MIX}_{\text{minimal}}$ model ($\ell = 2$) with adapted kinetics (Methods section "Real finger sweat metabolome data") through the three sub data sets. Due to the nature of the metabolite subnetworks (Fig. 4 bottom panel) it is possible to calculate two kinetic constants describing the absorption and elimination of caffeine ($k_a^{\text{caf}} = k_1'$ and $k_e^{\text{caf}} = k_2' + k_3'$) in all three cases. As the data for all three subnetworks was measured in the same experiment, we can assume that the

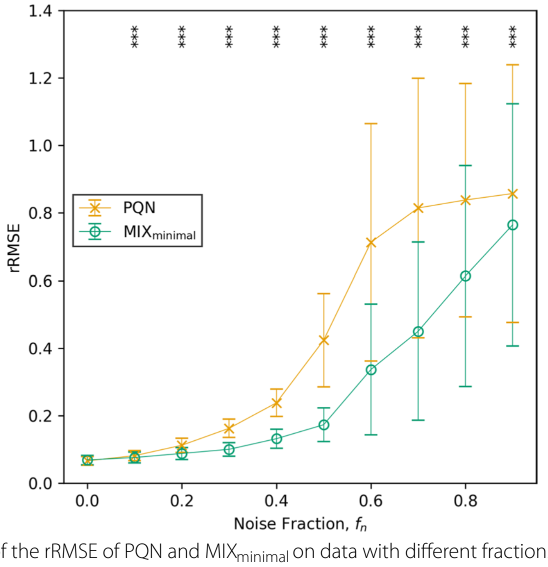


Fig. 10 Comparison of the rRMSE of PQN and $\text{MIX}_{\text{minimal}}$ on data with different fractions of noise. Significant differences in rRMSE between PQN and $\text{MIX}_{\text{minimal}}$ were tested with an one-sided pairwise Wilcoxon signed-rank test

underlying ground truth of these constants has to be the same. Therefore, by comparing the standard deviation of kinetic constants, it is possible to infer the performance of normalization methods.

In panels A and B of Fig. 11, the standard deviations of fitted kinetic constants within one measured $\widetilde{\mathbf{M}}$ time series are illustrated. Panel A shows that the standard deviations of the absorption constant of caffeine, k_a^{caf}, of $\mathrm{PKM}_{\mathrm{minimal}}$ are significantly larger than of the $\mathrm{MIX}_{\mathrm{minimal}}$ model ($p = 5.8 \times 10^{-4}, n = 37$, one-sided Wilcoxon signed-rank test). Likewise, a significant decrease in the size of standard deviations of $\mathrm{MIX}_{\mathrm{minimal}}$ was found compared to the previously published $\mathrm{PKM}_{\mathrm{minimal}}$ model ($p = 1.5\ 10^{-5}$) for the constant of caffeine elimination, k_e^{caf} (panel B, Fig. 11).

In panel E of Fig. 11, one exemplified normalized C time series of caffeine in sweat is depicted as fitted for all three subnetworks with $\mathrm{PKM}_{\mathrm{minimal}}$ and $\mathrm{MIX}_{\mathrm{minimal}}$,

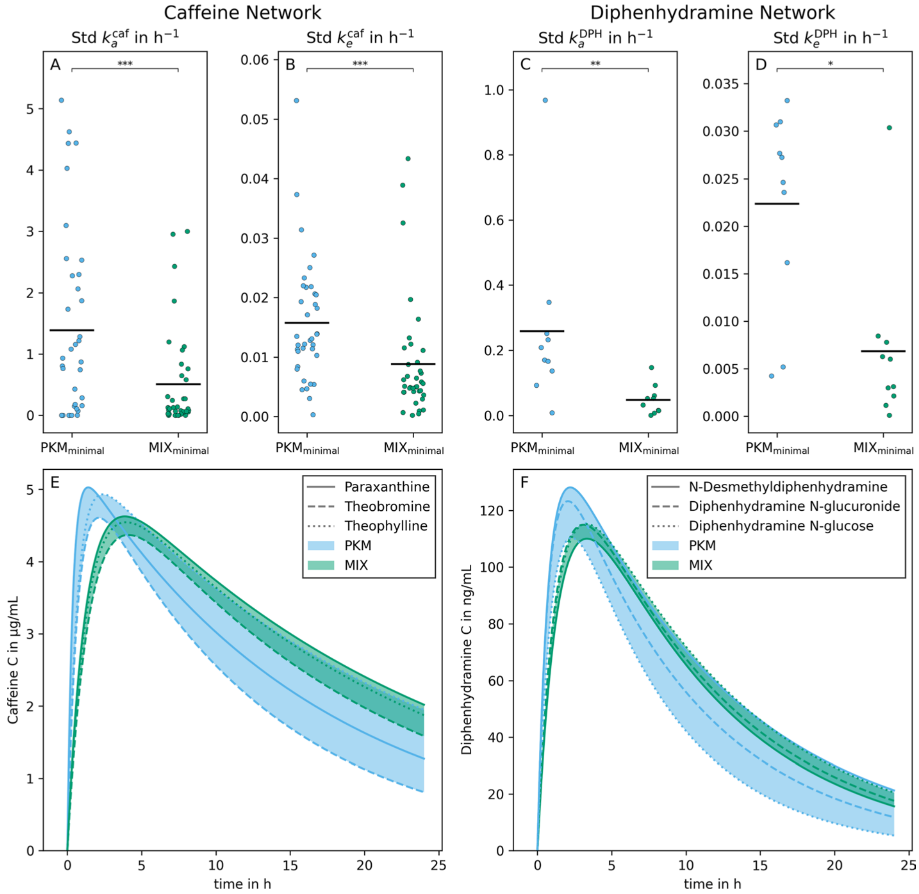


Fig. 11 Method validation with finger sweat (left column) and blood plasma (right column) data from Brunmair et al., 2021 [20] and Panitchpakdi et al., 2021 [41] respectively. On panels A to D, the standard deviations of constants of absorption and elimination of caffeine and diphenhydramine ($k_a^{\mathrm{caf}}, k_e^{\mathrm{caf}}, k_a^{\mathrm{DPH}}, k_e^{\mathrm{DPH}}$) between the three modeled subnetworks are plotted. The number of points per method corresponds to the number of concentrations time series present in both data sets (i.e., 37 and 10 for sweat and plasma, respectively). A one-sided Wilcoxon signed-rank test was used to test for significant differences. Panels E and F show the estimated concentration time series of caffeine and DPH plotted from the three different subnetworks. The lines are named after the second metabolite with a known kinetic present in the subnetwork; however, they all refer to *C* of caffeine and DPH. The colors of curves and the area between them indicate the results from normalization with $\mathrm{PKM}_{\mathrm{minimal}}$ or $\mathrm{MIX}_{\mathrm{minimal}}$, respectively

respectively. The selected time series illustrates the median of differences in standard deviations between $\mathrm{PKM}_{\mathrm{minimal}}$ and $\mathrm{MIX}_{\mathrm{minimal}}$ from panels A and B of Fig. 11. The area enclosed by the *C*s of $\mathrm{MIX}_{\mathrm{minimal}}$ models is smaller than from $\mathrm{PKM}_{\mathrm{minimal}}$.

We emphasize that in our original study, the caffeine degradation directly produces paraxanthine, theobromine, and theophylline; thus, pharmacokinetic parameters k_2, k_3, k_4 are explicitly linked [20]. Therefore, the kinetic network resembled specific kinetics of that metabolic pathway (Fig. 4 top panel). In contrast, in previous sections, we assumed that the underlying pathway structure is not known. Thus parameters are not linked, which implies that parameters are less constrained. Yet, in this section, we demonstrated that the fundamental improvement found by switching from PKM to a MIX model can also be translated back again to a more specific metabolic network (Fig. 4 bottom panel). In order to support this argument, we show the applicability of the $\mathrm{MIX}_{\mathrm{minimal}}$ normalization method on a real finger sweat data set. The results with real data emphasize the validity of the simulations done on synthetic data sets. They show that, especially when known metabolic networks are small, the $\mathrm{MIX}_{\mathrm{minimal}}$ model significantly improves the robustness of normalization and thus kinetic constants inferred from finger sweat time series measurements.

Diphenhydramine network

In the original study [41], the authors measured time series abundances in the blood plasma after the application of a single dose of diphenhydramine (DPH). $\widetilde{\mathbf{M}}$ from targeted DPH (known pharmacological constants, known kinetics) as well as untargeted metabolization products (N-desmethyl-DPH, DPH N-glucuronide, DPH N-glucose, known kinetics) and several other untargeted metabolites (unknown kinetics) were reported. Similar to sweat, although less pronounced, plasma also suffers from size effects (i.e., a systematic error in the measurements) introduced by biological means or preanalytical sample handling [47, 48]. Thus, we used the reported data as a second real data set for validation of the performance of $\mathrm{MIX}_{\mathrm{minimal}}$. The validation was performed in analogy to the caffeine study where a full network (Additional file 1: Fig. S4) is split into three subnetworks (Additional file 1: Fig. S5, for details see Methods section "Real blood plasma metabolome data").

In panels C and D of Fig. 11, the standard deviations of fitted kinetic constants within one measured $\widetilde{\mathbf{M}}$ and three fitted subnetworks are illustrated. Again, the standard deviations of k_a^{DPH} of $\mathrm{PKM}_{\mathrm{minimal}}$ are significantly larger than of $\mathrm{MIX}_{\mathrm{minimal}}$ ($p = 2{,}0 \times 10^{-3}, n = 10$, one-sided Wilcoxon signed-rank test, panel C). A similar significant decrease of the standard deviations are also found for k_e^{DPH} ($p = 3.2\ 10^{-2}$, panel D).

In panel F of Fig. 11, one exemplified normalized *C* time series of DPH in plasma is depicted as fitted for all three subnetworks with $\mathrm{PKM}_{\mathrm{minimal}}$ and $\mathrm{MIX}_{\mathrm{minimal}}$, respectively. The time series was selected as it is closest to the median of the differences in standard deviations between $\mathrm{PKM}_{\mathrm{minimal}}$ and $\mathrm{MIX}_{\mathrm{minimal}}$. It is visible that the area enclosed by the *C* resulting from the $\mathrm{MIX}_{\mathrm{minimal}}$ model is smaller than from $\mathrm{PKM}_{\mathrm{minimal}}$.

This validation illustrates the performance of the normalization models presented in this study on a data set that was measured independently from the development of said methods. The results of the plasma validation study are similar to the results observed for the finger sweat study; again, $\text{MIX}_{\text{minimal}}$ improves the robustness (i.e. reduces standard deviations) of size effect normalization.

Even though there is a significant decrease in the standard deviation of k_e^{DPH} with $\text{MIX}_{\text{minimal}}$ compared to $\text{PKM}_{\text{minimal}}$, $\text{MIX}_{\text{minimal}}$ also produced an outlier (Fig. 11D). The reason for this outlier is that on rare occasions, $\text{MIX}_{\text{minimal}}$ is not able to detect any size effects due to convergence issues (Additional file 1: Fig. S10A). To investigate these results, we performed synthetic data simulations (Additional file 1: Fig. S10B). There, we found that this behaviour of $\text{MIX}_{\text{minimal}}$ can be observed when two different $\mathbf{V}$ vectors are applied to ℓ and $\ell+$ metabolites. Therefore, we hypothesize that the clearly visible malfunction of $\text{MIX}_{\text{minimal}}$ to detect size effects (i.e. the variance of estimated $\mathbf{V}$ is close to 0) gives an indication to scientists that size effects might not be a major concern in such a data set. In this specific blood plasma time series measurement, for example, the size effects might have been too small compared to other error sources to be identified by $\text{MIX}_{\text{minimal}}$.

To summarize, with this validation, we show that the generalized normalization models, as implemented in this study, can directly be used for the normalization of real data as long as the modified Bateman function is able to describe the measured kinetics reasonably well and size effects are large enough to be detectable.

Discussion

In this study, we present a generalized framework for the PKM normalization model, first introduced in reference [20]. Moreover, we extend the existing model to incorporate untargeted metabolite information, dubbed as MIX model. Both models are implemented in Python and are available at GitHub https://github.com/Gotsmy/sweat_normalization.

The quality of normalization methods was tested on synthetic data sets. Synthetic data sets are necessary as it is impossible to obtain validation data without fundamentally changing the (finger) sweat sampling method as described above [20]. However, three different synthetic data generation methods (v1, v2, v3) were employed to ensure that synthetic data sets are as close to real data as possible. We found that when $n_{\text{metabolites}} \geq 60$, $\text{MIX}_{\text{minimal}}$ performs equally well or better than all other tested normalization methods.

Despite true V values remaining unknown, the real finger sweat data can be used as validation for relative robustness of normalization methods. There, $\text{MIX}_{\text{minimal}}$ significantly outperforms $\text{PKM}_{\text{minimal}}$. The decreased variance of kinetic constants estimated by $\text{MIX}_{\text{minimal}}$ likely originates from the fact that Q^{PQN} does not differ much for three subsets as long as sufficiently many $n_{\text{metabolites}} = 60$ are present in each subset. On the other hand, as only few data points are used for $\text{PKM}_{\text{minimal}}$ optimization, small errors in one of the two targeted metabolites' measured mass have a high potential to change the normalization result.

Additionally, the performance of $\text{PKM}_{\text{minimal}}$ and $\text{MIX}_{\text{minimal}}$ were compared on a blood plasma data set taken from a study independent of any measurements used for the development of the normalization models. There, we were able to demonstrate the same improvement from $\text{PKM}_{\text{minimal}}$ to $\text{MIX}_{\text{minimal}}$ in normalization robustness. Moreover, we show that the generalized normalization models as implemented as Python class in this study can be easily used for size effect normalization with little additional coding necessary.

To recapitulate, the proposed $\mathrm{MIX}_{\mathrm{minimal}}$ model has several crucial advantages over other tested methods.

- $\mathrm{MIX}_{\mathrm{minimal}}$ significantly outperforms $\mathrm{PKM}_{\mathrm{minimal}}$ in relative (rRMSE, $-43 \pm 12\,\%$) and absolute (RMSE, $-73 \pm 10\,\%$) errors with as little as 60 untargeted metabolites used as additional information (Fig. 7).
- $\mathrm{MIX}_{\mathrm{minimal}}$ is invariant to whether untargeted metabolites follow an easily describable kinetic concentration curve (Fig. 6).
- Without noise, $\mathrm{MIX}_{\mathrm{minimal}}$ performs equally well as PQN for relative abundances, but additionally, it estimates absolute values of V, similar to pharmacokinetic (PKM) models (Fig. 6).
- When noise is present $\mathrm{MIX}_{\mathrm{minimal}}$ also outperforms PQN for relative abundances (Fig. 10).
- $\mathrm{MIX}_{\mathrm{minimal}}$ performs well in this proof of principle study; moreover, it may be used as a basis for further improvements. Firstly, different, more sophisticated statistical normalization methods (e.g., EigenMS [27]) could be used as input for the PQN part of the model. Secondly, Bayesian priors describing uncertainties of different metabolites could be implemented over the λ parameter in a similar fashion as discussed in reference [49].
- Strikingly, the results showed that for all normalization methods tested, the RMSE and rRMSE values flattened once 60 metabolites were present in the original information. This suggested that the presented normalization models, especially $\mathrm{MIX}_{\mathrm{minimal}}$, can be applied even for biomatrices or analytical methods with as few as 60 compounds measured.
- Although $\mathrm{MIX}_{\mathrm{minimal}}$ was developed especially with sweat volume normalization in mind, it can be easily adapted for other biomatrices, e.g., plasma (Fig. 11).

Conclusion

In this study, we described and defined the MIX metabolomics time series normalization model and compared it to PKM. Subsequently, we elaborated several advantages of the $\mathrm{MIX}_{\mathrm{minimal}}$ model over PKM and previously published normalization methods. We are confident that this will further improve the reliability of metabolomic studies done on finger sweat and other conventional and non-conventional biofluids. However, we acknowledge that a more thorough investigation with data sets of several more quantified metabolites and determined sweat volumes needs to be carried out to assess the full potential of the proposed method.

Abbreviations

a_{sample}	Sampling skin area
b	Part of modified Bateman function
$C, \mathbf{C}$	Underlying concentration (vector)
c_0	Kinetic parameter
d	Kinetic parameter
F	Modified Bateman function
f_n	Noise fraction
i	Time point index
j	Metabolite index

k	Kinetic parameter
ℓ	Metabolites used for kinetic fitting
$\ell+$	Metabolites not used for kinetic fitting
$\mathcal{L}$	Loss
L	Loss function
lag	Kinetic parameter
$\widetilde{M}, \widetilde{\mathbf{M}}$	Measured mass (vector)
M^{ref}	Reference mass for PQN
m/z	Mass over charge ratio
$n_{\mathrm{metabolites}}$	Number of metabolites
$n_{\mathrm{time\ points}}$	Number of time points
p	*p*-value
q_{sweat}	Sweat rate
$Q^{\mathbf{C}}$	Median concentration fold change of two samples
$Q^{\mathbf{M}}$	Median mass fold change of two samples
$Q^{\mathrm{PQN}}, \mathbf{Q}^{\mathrm{PQN}}$	Normalization quotient (vector) calculated by PQN
R^2	Coefficient of determination
rRMSE	Relative measure of goodness of normalization
RMSE	Absolute measure of goodness of normalization
Std	Standard deviation
T	Transformation function
t	Time
$V, \mathbf{V}$	Collected (sweat) volume (vector)
V^{ref}	Reference volume for PQN
Var	Variance
v1, v2, v3	Synthetic data sets
Z	Scaling function
$\boldsymbol{\epsilon}$	Experimental error vector
$\boldsymbol{\theta}$	Kinetic parameter vector for fitting
λ	Loss weighting parameter
τ	Time to collect one sample

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-022-04918-1.

Additional file 1. Supplementary Figures, Tables, and Equations.

Acknowledgements

We thank Morgan Panitchpakdi and Shirley Tsunoda for their help in interpreting the plasma data set [41].

Author contributions

Conceptualization was done by MG and JZ. Funding was aquired by CG and JZ. Methodology was developed by MG and JB. Software was developed by MG. Formal analysis was done by MG and CB. Investigation was done by MG and JB. Resources were allocated by CG and JZ. Validation was done by MG, JB, and CB. The study was supervised by CG and JZ. Visualization was done by MG. Original draft was written by MG and JZ. Review and editing of the manuscript was done by all authors. All authors read and approved the final manuscript.

Funding

This research received no external funding. Open Access Funding by the University of Vienna.

Availability of data and materials

All analysis (except stated otherwise) was performed in Python 3.7 heavily relying on NumPy [50], Pandas [51], and SciPy [32]. Code for simulations, scripts for creation of figures and original and generated data is available on GitHub https://github.com/Gotsmy/sweat_normalization under the GNU GPL version 3 license.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 22 March 2022 Accepted: 2 September 2022
Published online: 16 September 2022

References

1. Jang M, Costa C, Bunch J, Gibson B, Ismail M, Palitsin V, Webb R, Hudson M, Bailey M. On the relevance of cocaine detection in a fingerprint. Sci Rep. 2020;10(1):1–7.
2. Delgado-Povedano M, Calderón-Santiago M, de Castro ML, Priego-Capote F. Metabolomics analysis of human sweat collected after moderate exercise. Talanta. 2018;177:47–65.
3. Brunmair J, Bileck A, Stimpfl T, Raible F, Del Favero G, Meier-Menches SM, Gerner C. Metabo-tip: a metabolomics platform for lifestyle monitoring supporting the development of novel strategies in predictive, preventive and personalised medicine. EPMA J. 2021:1–13
4. Czerwinska J, Jang M, Costa C, Parkin MC, George C, Kicman AT, Bailey MJ, Dargan PI, Abbate V. Detection of mephedrone and its metabolites in fingerprints from a controlled human administration study by liquid chromatography-tandem mass spectrometry and paper spray-mass spectrometry. Analyst. 2020;145(8):3038–48.
5. Calderón-Santiago M, Priego-Capote F, Turck N, Robin X, Jurado-Gámez B, Sanchez JC, De Castro MDL. Human sweat metabolomics for lung cancer screening. Anal Bioanal Chem. 2015;407(18):5381–92.
6. Cui X, Zhang L, Su G, Kijlstra A, Yang P. Specific sweat metabolite profile in ocular Behcet's disease. Int Immunopharmacol. 2021;97: 107812.
7. Harshman SW, Browder AB, Davidson CN, Pitsch RL, Strayer KE, Schaeublin NM, Phelps MS, O'Connor ML, Mackowski NS, Barrett KN, et al. The impact of nutritional supplementation on sweat metabolomic content: a proof-of-concept study. Front Chem. 2021;9:255.
8. Hussain JN, Mantri N, Cohen MM. Working up a good sweat-the challenges of standardising sweat collection for metabolomics analysis. Clin Biochemist Rev. 2017;38(1):13.
9. Harshman SW, Strayer KE, Davidson CN, Pitsch RL, Narayanan L, Scott AM, Schaeublin NM, Wiens TL, Phelps MS, O'Connor ML, et al. Rate normalization for sweat metabolomics biomarker discovery. Talanta. 2021;223: 121797.
10. Kuwayama K, Tsujikawa K, Miyaguchi H, Kanamori T, Iwata YT, Inoue H. Time-course measurements of caffeine and its metabolites extracted from fingertips after coffee intake: a preliminary study for the detection of drugs from fingerprints. Anal Bioanal Chem. 2013;405(12):3945–52.
11. Kuwayama K, Yamamuro T, Tsujikawa K, Miyaguchi H, Kanamori T, Iwata YT, Inoue H. Time-course measurements of drugs and metabolites transferred from fingertips after drug administration: usefulness of fingerprints for drug testing. Forensic Toxicol. 2014;32(2):235–42.
12. Baker LB. Physiology of sweat gland function: the roles of sweating and sweat composition in human health. Temperature. 2019;6(3):211–59.
13. Nyein HYY, Bariya M, Tran B, Ahn CH, Brown BJ, Ji W, Davis N, Javey A. A wearable patch for continuous analysis of thermoregulatory sweat at rest. Nat Commun. 2021;12(1):1–13.
14. Taylor NA, Machado-Moreira CA. Regional variations in transepidermal water loss, eccrine sweat gland density, sweat secretion rates and electrolyte composition in resting and exercising humans. Extreme Physiol Med. 2013;2(1):4.
15. Ando H, Noguchi R. Dependence of palmar sweating response and central nervous system activity on the frequency of whole-body vibration. Scand J Work Environ Health. 2003:216–219.
16. Zhong B, Jiang K, Wang L, Shen G. Wearable sweat loss measuring devices: from the role of sweat loss to advanced mechanisms and designs. Adv Sci. 2021:2103257.
17. Harshman SW, Pitsch RL, Smith ZK, O'Connor ML, Geier BA, Qualley AV, Schaeublin NM, Fischer MV, Eckerle JJ, Strang AJ, et al. The proteomic and metabolomic characterization of exercise-induced sweat for human performance monitoring: a pilot investigation. PLoS ONE. 2018;13(11):0203133.
18. Sonner Z, Wilder E, Heikenfeld J, Kasting G, Beyette F, Swaile D, Sherman F, Joyce J, Hagen J, Kelley-Loughnane N, et al. The microfluidics of the eccrine sweat gland, including biomarker partitioning, transport, and biosensing implications. Biomicrofluidics. 2015;9(3): 031301.
19. Du Q, Zhang Y, Wang J, Chang J, Wang A, Ren X, Liu B. Quantitative analysis of 17 hypoglycemic drugs in fingerprints using ultra-high-performance liquid chromatography/tandem hybrid triple quadrupole linear ion trap mass spectrometry. Rapid Commun Mass Spectrom. 2022;36(1):9199.
20. Brunmair J, Gotsmy M, Niederstaetter L, Neuditschko B, Bileck A, Slany A, Feuerstein ML, Langbauer C, Janker L, Zanghellini J, et al. Finger sweat analysis enables short interval metabolic biomonitoring in humans. Nat Commun. 2021;12(1):1–13.
21. Filzmoser P, Walczak B. What can go wrong at the data normalization step for identification of biomarkers? J Chromatogr A. 2014;1362:194–205.
22. Singh AS, Masuku MB. Sampling techniques and determination of sample size in applied statistics research: an overview. Int J Econ Commerce Manag. 2014;2(11):1–22.
23. Choi J, Bandodkar AJ, Reeder JT, Ray TR, Turnquist A, Kim SB, Nyberg N, Hourlier-Fargette A, Model JB, Aranyosi AJ, et al. Soft, skin-integrated multifunctional microfluidic systems for accurate colorimetric analysis of sweat biomarkers and temperature. ACS Sensors. 2019;4(2):379–88.

24. Kim SB, Koo J, Yoon J, Hourlier-Fargette A, Lee B, Chen S, Jo S, Choi J, Oh YS, Lee G, et al. Soft, skin-interfaced microfluidic systems with integrated enzymatic assays for measuring the concentration of ammonia and ethanol in sweat. Lab Chip. 2020;20(1):84–92.
25. Ragan TJ, Bailey AP, Gould AP, Driscoll PC. Volume determination with two standards allows absolute quantification and improved chemometric analysis of metabolites by nmr from submicroliter samples. Anal Chem. 2013;85(24):12046–54.
26. Warrack BM, Hnatyshyn S, Ott K-H, Reily MD, Sanders M, Zhang H, Drexler DM. Normalization strategies for metabonomic analysis of urine samples. J Chromatogr B. 2009;877(5–6):547–52.
27. Karpievitch YV, Nikolic SB, Wilson R, Sharman JE, Edwards LM. Metabolomics data normalization with eigenms. PLoS ONE. 2014;9(12): 116221.
28. Dieterle F, Ross A, Schlotterbeck G, Senn H. Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in 1h nmr metabonomics. Anal Chem. 2006;78(13):4281–90.
29. Li B, Tang J, Yang Q, Cui X, Li S, Chen S, Cao Q, Xue W, Chen N, Zhu F. Performance evaluation and online realization of data-driven normalization methods used in lc/ms based untargeted metabolomics analysis. Sci Rep. 2016;6(1):1–13.
30. Di Guida R, Engel J, Allwood JW, Weber RJ, Jones MR, Sommer U, Viant MR, Dunn WB. Non-targeted uhplc-ms metabolomic data processing methods: a comparative investigation of normalisation, missing value imputation, transformation and scaling. Metabolomics. 2016;12(5):93.
31. Macedo AN, Mathiaparanam S, Brick L, Keenan K, Gonska T, Pedder L, Hill S, Britz-McKibbin P. The sweat metabolome of screen-positive cystic fibrosis infants: Revealing mechanisms beyond impaired chloride transport. ACS Cent Sci. 2017;3(8):904–13.
32. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, Bright J, et al. Scipy 1.0: fundamental algorithms for scientific computing in python. Nat Methods. 2020;17(3):261–72.
33. Garrett ER. The bateman function revisited: a critical reevaluation of the quantitative expressions to characterize concentrations in the one compartment body model as a function of time with first-order invasion and first-order elimination. J Pharmacokinet Biopharm. 1994;22(2):103–28.
34. Brunius C, Shi L, Landberg R. Large-scale untargeted lc-ms metabolomics data correction using between-batch feature alignment and cluster-based within-batch signal intensity drift correction. Metabolomics. 2016;12(11):1–13.
35. Kvasnička A, Friedecký D, Tichá A, Hyšpler R, Janečková H, Brumarová R, Zadák Z. SLIDE—Novel Approach to Apocrine Sweat Sampling for Lipid Profiling in Healthy Individuals. Int J Molec Sci. 2021;22(15):8054.
36. Brunmair J, Gotsmy M, Niederstaetter L, Neuditschko B, Bileck A, Slany A, Feuerstein ML, Langbauer C, Janker L, Zanghellini J, et al. Finger sweat analysis enables short interval metabolic biomonitoring in humans. https://doi.org/10.5281/zenodo.5222967.
37. Chambers MC, Maclean B, Burke R, Amodei D, Ruderman DL, Neumann S, Gatto L, Fischer B, Pratt B, Egertson J, et al. A cross-platform toolkit for mass spectrometry and proteomics. Nat Biotechnol. 2012;30(10):918–20.
38. Tsugawa H, Ikeda K, Takahashi M, Satoh A, Mori Y, Uchino H, Okahashi N, Yamada Y, Tada I, Bonini P, et al. A lipidome atlas in ms-dial 4. Nat Biotechnol. 2020;38(10):1159–63.
39. Csajka C, Haller C, Benowitz N, Verotta D. Mechanistic pharmacokinetic modelling of ephedrine, norephedrine and caffeine in healthy subjects. Br J Clin Pharmacol. 2005;59(3):335–45.
40. Kamimori GH, Karyekar CS, Otterstetter R, Cox DS, Balkin TJ, Belenky GL, Eddington ND. The rate of absorption and relative bioavailability of caffeine administered in chewing gum versus capsules to normal healthy volunteers. Int J Pharm. 2002;234(1–2):159–67.
41. Panitchpakdi M, Weldon KC, Jarmusch AK, Gentry EC, Choi A, Sepulveda Y, Aguirre S, Sun K, Momper JD, Dorrestein PC, et al. Non-invasive skin sampling detects systemically administered drugs in humans. PloS one, 17(7), e0271794.
42. Panitchpakdi M, Weldon KC, Jarmusch AK, Gentry EC, Choi A, Sepulveda Y, Aguirre S, Sun K, Momper JD, Dorrestein PC, et al. Non-Invasive Skin Sampling Detects Systemically Administered Drugs in Humans. https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=deee382b163f4441afea5fda4b2a2bcf. Accessed 20 Dec 2021.
43. Wilcoxon F. Individual comparisons by ranking methods. Biometrics Bull. 1945;1(6):80–3.
44. van den Berg RA, Hoefsloot HC, Westerhuis JA, Smilde AK, van der Werf MJ. Centering, scaling, and transformations: improving the biological information content of metabolomics data. BMC Genomics. 2006;7(1):1–15.
45. Baker LB, Wolfe AS. Physiological mechanisms determining eccrine sweat composition. Eur J Appl Physiol. 2020;120(4):719–52.
46. da Silva RR, Vargas F, Ernst M, Nguyen NH, Bolleddu S, Del Rosario KK, Tsunoda SM, Dorrestein PC, Jarmusch AK. Computational removal of undesired mass spectral features possessing repeat units via a kendrick mass filter. J Am Soc Mass Spectrom. 2018;30(2):268–77.
47. Kamlage B, Maldonado SG, Bethan B, Peter E, Schmitz O, Liebenberg V, Schatz P. Quality markers addressing preanalytical variations of blood and plasma processing identified by broad and targeted metabolite profiling. Clin Chem. 2014;60(2):399–412.
48. Pinto J, Domingues MRM, Galhano E, Pita C, do Céu Almeida M, Carreira IM, Gil AM. Human plasma stability during handling and storage: impact on nmr metabolomics. Analyst 2014;139(5):1168–77.
49. Sheiner LB, Beal SL. Bayesian individualization of pharmacokinetics: simple implementation and comparison with non-Bayesian methods. J Pharm Sci. 1982;71(12):1344–8.
50. Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, Wieser E, Taylor J, Berg S, Smith NJ, Kern R, Picus M, Hoyer S, van Kerkwijk MH, Brett M, Haldane A, del Río JF, Wiebe M, Peterson P, Gérard-Marchant P, Sheppard K, Reddy T, Weckesser W, Abbasi H, Gohlke C, Oliphant TE. Array programming with NumPy. Nature. 2020;585(7825):357–62. https://doi.org/10.1038/s41586-020-2649-2.
51. Pandas Development Team, T.: Pandas-dev/pandas: Pandas. https://doi.org/10.5281/zenodo.3509134.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

A.3. Publication III: Sulfate limitation increases specific plasmid DNA yield and productivity in *E. coli* fed-batch processes.

Mathias Gotsmy, Florian Strobl, Florian Weiß, Petra Gruber, Barbara Kraus, Juergen Mairhofer, and Jürgen Zanghellini

My role was first author. I conceived the idea of sulfate limitation, implemented the code, drafted the manuscript, and compiled the figures.

Microbial Cell Factories

RESEARCH Open Access

Sulfate limitation increases specific plasmid DNA yield and productivity in *E. coli* fed-batch processes

Mathias Gotsmy[1,2], Florian Strobl[3], Florian Weiß[3], Petra Gruber[4], Barbara Kraus[4], Juergen Mairhofer[3] and Jürgen Zanghellini[1*]

Abstract

Plasmid DNA (pDNA) is a key biotechnological product whose importance became apparent in the last years due to its role as a raw material in the messenger ribonucleic acid (mRNA) vaccine manufacturing process. In pharmaceutical production processes, cells need to grow in the defined medium in order to guarantee the highest standards of quality and repeatability. However, often these requirements result in low product titer, productivity, and yield. In this study, we used constraint-based metabolic modeling to optimize the average volumetric productivity of pDNA production in a fed-batch process. We identified a set of 13 nutrients in the growth medium that are essential for cell growth but not for pDNA replication. When these nutrients are depleted in the medium, cell growth is stalled and pDNA production is increased, raising the specific and volumetric yield and productivity. To exploit this effect we designed a three-stage process (1. batch, 2. fed-batch with cell growth, 3. fed-batch without cell growth). The transition between stage 2 and 3 is induced by sulfate starvation. Its onset can be easily controlled via the initial concentration of sulfate in the medium. We validated the decoupling behavior of sulfate and assessed pDNA quality attributes (supercoiled pDNA content) in *E. coli* with lab-scale bioreactor cultivations. The results showed an increase in supercoiled pDNA to biomass yield by 33% and an increase of supercoiled pDNA volumetric productivity by 13 % upon limitation of sulfate. In conclusion, even for routinely manufactured biotechnological products such as pDNA, simple changes in the growth medium can significantly improve the yield and quality.

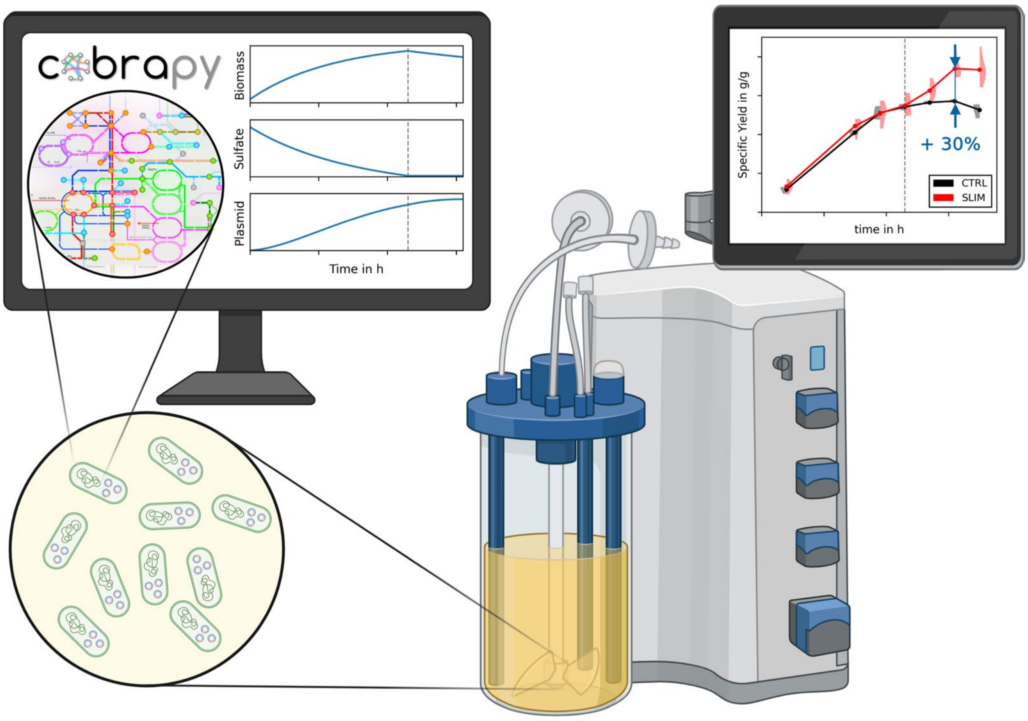
Highlights

- Genome-scale metabolic models predict growth decoupling strategies.
- Sulfate limitation decouples cell growth from pDNA production.
- Sulfate limitation increases the specific supercoiled pDNA yield by 33% and the volumetric productivity by 13%.
- We propose that sulfate limitation improves the biosynthesis of over 25% of naturally secreted products in *E. coli*.

Keywords Growth decoupling, Medium optimization, GSMM, FBA, dFBA, *E. coli*, Plasmid production

*Correspondence:
Jürgen Zanghellini
juergen.zanghellini@univie.ac.at
Full list of author information is available at the end of the article

Graphical Abstract



Introduction

Plasmid DNA (pDNA) is an important product of the pharmaceutical industry being primarily used as vectors for the transfection of mammalian cells. For example, pDNA can be directly injected in the form of a DNA vaccine [1]. Moreover, it is an important raw material for the production of mRNA vaccines, for example against SARS-CoV-2 [2, 3]. Additionally, pDNA can be used as a vector for gene therapy [4]. Regardless of the application, high amounts and high quality of pDNA are needed and the optimization of its production is of health-economic interest as pDNA is a relevant driver of manufacturing costs.

Apart from solely maximizing the yield of pDNA, three prerequisites are required for the design of a pDNA production process. Firstly, pDNA can be present in an open circular (oc), linear (l), or covalently closed circular (ccc), i.e., supercoiled form. Generally, the ccc form is considered more favorable for transfection of mammalian cells and, therefore, the fraction of supercoiling is of importance [5, 6]. Secondly, a loss of plasmid can be severely detrimental to the productivity during the fermentation process. Classically, this problem can be mitigated by the introduction of antibiotic resistance selection systems and the usage of antibiotic selection pressure during fermentation. However, these systems have several downsides. They, on the one hand, shift metabolic resources from the production of pDNA to the production of antibiotic resistance proteins [7]. On the other hand, special care has to be taken to remove residual antibiotics during pDNA purification and the absence thereof has to be validated. Therefore, several alternatives have been developed [8] although still not state-of-the-art yet for pDNA manufacturing. Thirdly, even though pDNA production in complex media generates higher yields, chemically defined media are preferred for the production of high-quality and safe pharmaceuticals [1].

Many strategies for the increase of pDNA production have been published with a large fraction using *E. coli* as production organism [9]. The methods range from the screening of favorable strains to metabolic engineering through knocking-in and -out of genes to antibiotic-free selection systems and other highly optimized production strains [9]. Most strategies for increasing pDNA production can be grouped into two conceptual approaches: (i) the reduction of cell growth; (ii) ensuring a constant supply of DNA precursor metabolites. The methods differ widely in the way one (or both) aims are achieved.

Early on researchers found that a low growth rate increases the specific productivity of pDNA [10]. To achieve this in batch fermentations Galindo et al. [11] designed a medium that releases glucose enzymatically and thus down-regulates glucose uptake and subsequently growth. Alternatively, Soto et al. [12] developed *E. coli* strain VH33 by knocking-out the main uptake pathway of glucose to achieve the same result. With this method, they could increase the production to 40 mg L^{-1} pDNA compared to 17 mg L^{-1} of the wild type strain [12]. Moreover, an optimization strategy for microaerobic environments was devised where *E. coli* strain W3110 improved pDNA production in presence of a recombinant expression of the *Vitreoscilla* hemoglobin protein [13]. In a subsequent study, this strain was tested in batch fermentations with different oxygen transfer rates and they concluded that as oxygen was depleted the growth rate decreased and the production of pDNA increased [14].

A strategy to ensure a constant supply of DNA precursor metabolites is, for example, the knocking-out of pyruvate kinase which forces metabolization of glucose over the pentose phosphate pathway [7, 15, 16]. Other methods utilize stoichiometric models to optimize the growth medium [17]. The authors concluded that the addition of the nucleosides adenosine, guanosine, cytidine, and thymidine as well as several amino acids can significantly improve pDNA production (60 mg L^{-1} in a batch fermentation). Martins et al. [18] optimized the growth medium for the high producer strain VH33 and concluded that the presence of aromatic amino acids (phenylalanine, tryptophan, tyrosine) is advantageous for redirecting molecules to the nucleotide synthesis pathways. Additionally, the effect of the amount and type of nitrogen source in the growth medium has been investigated on the production of pDNA to 213 mg L^{-1} [19]. Also, economical aspects of the medium design have been discussed [20].

Further potential for optimization is the pDNA itself. For example, reducing the size of the pDNA has been linked to higher volumetric yields [21]. However, a reduction might not always be possible, especially for therapeutic applications, where plasmid sizes are typically large (> 6 kb) [22]. Moreover, the pDNA yield of a process is highly dependent on the origin of replication. Currently, most plasmids carry a high copy number pUC origin of replication that allows up to 700 pDNA copies per cell [1]. Other approaches involved microaerobically induced [23] or heat-induced origins of replication that increase the plasmid copy number at higher temperatures than 37 °C [24, 25]. However, higher temperatures come with physiological trade-offs and, therefore, the amplitude and timing of heat induction are of importance [26].

Table 1 Lexicographic objectives for dFBA

Objective	
Max	Biomass production
Min	Sulfate uptake
Max	pDNA production

The order of optimization was top to bottom

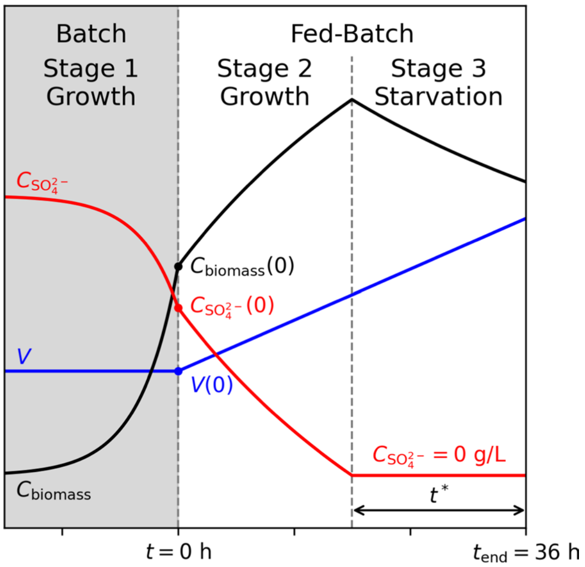


Fig. 1 Schematic of a three-stage growth-decoupled fed-batch pDNA production process (1. batch, 2. fed-batch with cell growth, 3. fed-batch without cell growth). Size of the variables and length of the stages are not to scale. As the batch process was kept unchanged, it was not included in the fed-batch optimization simulations. Instead, the simulation starts at the beginning of the feed ($t = 0$ h) with realistic values for the process variables at batch end (Additional file 1: Table S1)

Recently, also other production organisms were proposed, e.g. *Lactococcus lactis*. Contrary to *E. coli*, *L. lactis* is generally regarded as safe (GRAS) and thus simplifies the downstream processing [27, 28].

Here, we design a three-stage bioprocess, where cellular growth and pDNA production are decoupled. We use constraint-based modeling to (i) identify medium components that induce the switching and (ii) determine the optimal time point for switching between the phases such that the average volumetric productivity is maximized.

Methods

Metabolic modeling

All models and code used for the creation, simulation, and analysis of these models are available at https://github.com/Gotsmy/slim.

Model creation

For our analysis, we used *i*ML1515 [29], a genome-scale metabolic model of *Escherichia coli* strain K-12 substrain MG1655. All model modifications and simulations were performed in Python 3.10 using the CobraPy package [30].

To simulate plasmid production, an 8 base pair (bp) dummy pDNA metabolite with 50% GC content was added to the model. Subsequently, we added a pDNA production reaction, corresponding to the dummy plasmid's stoichiometry. pDNA polymerization cost was estimated to 1.36 mol adenosine triphosphate (ATP) per mol dNTP [7, 31]. Additionally, a pDNA sink reaction was introduced to make fluxes through the pDNA synthesis reaction feasible. An SBML version of the used model is available at https://github.com/Gotsmy/slim/tree/main/models.

Although 8 bp is an unrealistically small size for a real pDNA molecule, we emphasize that the actual length of the plasmid does not change the relative underlying stoichiometry. The reason we chose a small dummy plasmid is that it already has a molecular weight of approximately 4943.15 g mol^{-1}. If we included a multiple kbp sized plasmid into the model, we would have risked that its large molecular weight lead to numerical instabilities during simulation [32]. However, all results are shown in gram pDNA, therefore, the exact length of the dummy plasmid does not change the values.

Identification of decoupling compounds

Initially, we performed a parsimonious flux balance analysis (pFBA) [33] with biomass growth as objective. The maximal glucose uptake rate was set to 10 mmol g^{-1} h^{-1} and a non-growth associated maintenance requirement was set to 6.86 mmol ATP g^{-1} h^{-1} [29]. Exchange reactions with non-zero fluxes were used for the definition of the minimal medium. All exchange reactions that were not present in the minimal medium, except for H_2O and H^+, were turned off. To investigate the differences in uptake fluxes, an additional pFBA was performed with pDNA as the objective.

Next, we selected each of the minimal medium components and set the maximum exchange flux bound for this metabolite to 5, 25, 50, 75, and 100% of the flux during biomass growth. For each value, a production envelope (pDNA synthesis as a function of growth) was calculated. Decoupling medium components were identified as metabolites which, as their uptake flux decreased, the maximum pDNA synthesis potential increased at a maximum biomass growth.

Process simulation

We used dynamic flux balance analysis (dFBA) [34] to simulate the time evolution of the fed-batch processes. Only the feed phase was simulated (stage 2 and 3 in Fig. 1) as the batch remained unchanged. At every integration step, a lexicographic flux balance analysis (FBA), where all fluxes of interest were consecutively optimized, was performed [35]. The list of objectives is given in Table 1. The non-growth associated maintenance was kept at a constant value as before. We used SciPy's solve_ivp function for numerical integration [36].

Sulfate concentration, $C_{\text{SO}_4^{2-}}(t)$, in the medium was tracked as its depletion coincides with the metabolic switch from biomass growth to production. Due to a lack of knowledge, no uptake kinetics of sulfate were simulated. Practically, this meant that the sulfate uptake was calculated from the biomass stoichiometry and growth rate. The exchange reaction bounds were left unconstrained when $C_{\text{SO}_4^{2-}}(t) > 0$ and were blocked when $C_{\text{SO}_4^{2-}}(t) \leq 0$.

High copy number origin of replication plasmids typically replicate during the growth phase as they high-jack genomic DNA synthesis pathways. Therefore, we set a lower bound, $q^{\mu}_{\text{pDNA}} = 4.9$ mg g^{-1} h^{-1}, for the pDNA synthesis reaction. Moreover, it is unrealistic to assume that all available glucose is channeled towards pDNA production during the SO_4^{2-} starvation. Therefore, we set an upper bound to the synthesis reaction. Since its actual value was unknown, we tested several levels ranging from $q^{*}_{\text{pDNA}} = 4.9$ to 24.7 mg g^{-1} h^{-1}. Throughout this manuscript, that ratio of upper to lower bound of the pDNA synthesis reaction is referred to as

$$\kappa_{\text{pDNA}} = q^{*}_{\text{pDNA}} / q^{\mu}_{\text{pDNA}}. \tag{1}$$

We simulated 41 equidistant levels of $\kappa_{\text{pDNA}} \in [1, 5]$. Because of the implementation of the dFBA, the lower and upper bounds of the synthesis reaction can be interpreted as the pDNA synthesis fluxes during the growth and SO_4^{2-} starvation phase, respectively. We assumed negligible changes in overall biomass composition due to pDNA synthesis, which solely derives from external carbon sources.

To compare pDNA production processes, we calculated two performance indicators: the specific yield

$$Y_{\text{pDNA/biomass}}(t) := C_{\text{pDNA}}(t) / C_{\text{biomass}}(t), \tag{2}$$

and the average volumetric productivity

$$p_{\text{pDNA}}(t) := \frac{C_{\text{pDNA}}(t) - C_{\text{pDNA}}(0)}{t - t_0}. \tag{3}$$

Here, $t_0 = 0$ h indicates the start of feeding (see Fig. 1). With our parameters, see Additional file 1: Table S1, (3) simplifies to

$$p_{\text{pDNA}}(t) = C_{\text{pDNA}}(t)/t. \tag{4}$$

A schematic of a three-stage growth-decoupled fed-batch process is shown in Fig. 1. Initial conditions (at the start of the feed phase) resembled realistic values from the end of a batch process in a small bioreactor (Additional file 1: Table S1). Strategies with a linear (i.e. constant) and an exponential feeding rate,

$$r^{\text{feed}} = \begin{cases} r_{\text{lin}}^{\text{feed}}, \\ \mu C_{\text{biomass}}(0) V(0) Y_{\text{feed/biomass}} \exp(\mu t), \end{cases} \tag{5}$$

respectively, were tested. The specific glucose uptake rate was set to

$$q_{\text{glucose}}(t) = \frac{r^{\text{feed}} C_{\text{glucose}}^{\text{feed}}}{C_{\text{biomass}}(t) V(t)} \tag{6}$$

to ensure $C_{\text{glucose}}(t) = 0$ throughout the (non-starved) fed-batch phase (stage 2, Fig. 1). Here, $C_{\text{glucose}}^{\text{feed}}$ denotes the glucose concentration in the feed medium.

The dFBA simulation terminated once the maximal volume of 1L was reached (i.e. after 35.2 and 36.0 h in exponential and linear fed-batch, respectively). The initial sulfate concentration in the medium at the start of feed was varied in 301 equidistant steps to search for a productivity optimum. We assumed that sulfate was present in the medium at the start of the feed, while the feed medium was sulfate-free.

Identifying alternative targets

We screened 307 products with existing exchange reactions in *i*ML1515 [29] by performing lexicographic FBA (analogous to Table 1) with and without sulfate in the growth medium. Bioproducts for which the calculated synthesis rate improved during sulfate limitation were identified as potential products benefiting from a sulfate limited process design.

Validation experiments

Upstream process

All experiments were conducted with proline auxotroph *E. coli* K-12 strain JM108 [37], which previously had been used for pDNA production [8]. The cells were transformed with a plasmid of 12.0 kbp length and 53% GC content containing a pUC origin of replication and a kanamycin resistance gene. However, no kanamycin was added throughout the production process as U.S. Food and Drug Administration (FDA) and European Medicines Agency (EMA) recommend to avoid the use of antibiotics [1].

For fed-batch fermentations, *E. coli* JM108 were grown in a 1.8 L (1.0 L net volume, 0.5 L batch volume) computer-controlled bioreactor (DASGIP parallel bioreactor system, Eppendorf AG, Germany). The bioreactor was equipped with a pH probe and an optical dissolved oxygen probe (Hamilton Bonaduz AG, Switzerland). The pH was maintained at 7.0 ± 0.1 by addition of 12.5% ammonia solution; the temperature was maintained at 37 ± 0.5°C. The dissolved oxygen (O_2) level was stabilized above 30% saturation by controlling the stirrer speed, aeration rate, and gassing composition. Foaming was suppressed by the addition of 2 mL 1:10 diluted Struktol J673A antifoam suspension (Schill+Seilacher, Germany) to the batch medium and by the automatic addition of 1:10 diluted Struktol J673A controlled by a conductivity-operated level sensor. For the inoculation of the bioreactor, a seed culture was used (25 mL batch medium inoculated with 250 μL master cell bank in 250 mL baffled glass flasks at 37 °C with shaking at 180 rpm). The seed culture was incubated until a final OD_{600} of 2–4 was reached and a defined volume was transferred aseptically to the bioreactor to result in an initial OD_{600} of 0.015.

The fermentation process was designed for a final amount of 50 g cell dry mass (CDM) of which 1.51 g was obtained in a batch volume of 500 mL and 48.5 g during the feed phase via the addition of another 500 mL of feed medium. The amount of glucose for the specific medium was calculated based on a yield coefficient ($Y_{\text{biomass/glucose}}$) of 0.303 g g^{-1} and added as $C_6H_{12}O_6 \cdot H_2O$. For media preparation, all chemicals were purchased from Carl Roth GmbH (Germany) unless otherwise stated. Two different media compositions were compared regarding specific pDNA productivity: one with a limited sulfur source (SO_4^{2-} limitation) and one without a limited sulfur source (control). Feeding was initiated when the culture in the batch medium entered the stationary phase. A fed-batch regime with a linear substrate feed (0.26 g min^{-1} respectively 13.91 mL h^{-1})

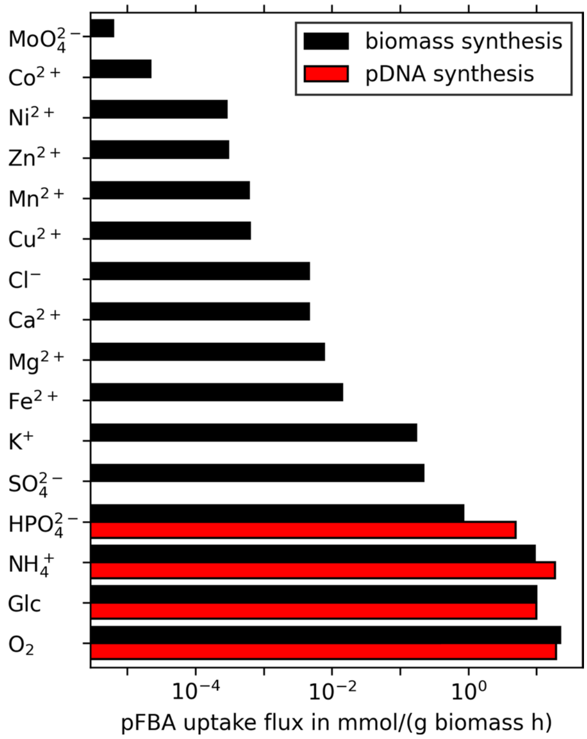


Fig. 2 pFBA uptake flux rates for all minimal medium components. Black bars represent fluxes for optimization of biomass synthesis, red bars represent fluxes of pDNA synthesis

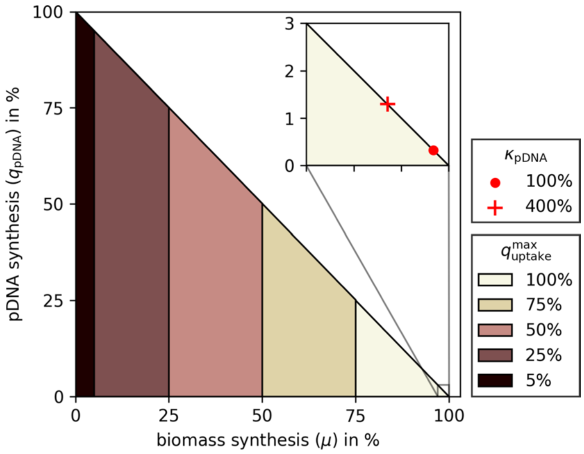


Fig. 3 Normalized pDNA production envelopes for different maximal uptake rates of (a single) decoupling nutrient. Ca^{2+}, Cl , Co^{2+}, Cu^{2+}, Fe^{2+}, K^+, Mg^{2+}, Mn^{2+}, MoO_4^{2-}, Ni^{2+}, SO_4^{2-}, and Zn^{2+} all result in identical sets of production envelopes. Note that all production envelopes include the line segment from (0|0) to (0|100). The inset shows the extreme points of a production envelope with a realistic pDNA production rate during biomass growth (red circle) and a potential 4-fold pDNA production rate increase (red cross)

was used for 35 h (approximately five generations). During SO_4^{2-} limitation fermentations, the simulations predicted that the provided sulfur was completely consumed at 23 h after feed start. The batch and fed-batch medium components are given in Additional file 1: Table S2. The cultivations of SO_4^{2-} limitation and control were conducted in six and three replicates, respectively.

Analysis

For off-line analysis (OD_{600}, CDM, pDNA product), the bioreactor was sampled during the fed-batch phase. The OD_{600} was measured using an Ultrospec 500 pro Spectrophotometer (Amersham Biosciences, UK), diluting the samples with phosphate-buffered saline to achieve the linear range of measurement. For the determination of CDM, 1 mL of cell suspension was transferred to pre-weighed 2.0 mL reaction tubes and centrifuged for 10 min at 16,100 rcf and 4°C with an Eppendorf 5415 R centrifuge. The supernatant was transferred to another reaction tube and stored at − 20°C for further analysis. As a washing step, the cell pellet was resuspended in 1.8 mL RO-H_2O, centrifuged, and the supernatant discarded. Afterwards, the pellet was resuspended in 1.8 mL RO-H_2O and finally dried at 105 °C for 24 h. The reaction tubes with the dried biomass were cooled to room temperature in a desiccator before re-weighing.

For pDNA product analysis, the sampling volume of the cell suspension, corresponding to 20 mg CDM, was estimated via direct measurement of the OD_{600}. The calculated amount was transferred to 2.0 mL reaction tubes and centrifuged at 16,100 rcf and 4 °C for 10 min. The supernatant was discarded, and the cell pellets were stored at − 20°C.

The content of pDNA in ccc-conformation was determined using AIEX-HPLC (CIMac pDNA-0.3 Analytical Column, 1.4 μL; BIA Separations d.o.o., Slovenia). The column separated open circular, linear, and supercoiled pDNA fractions into distinct peaks. Quantification was achieved using a calibration curve based on peak areas obtained from purified pDNA samples. For HPLC analysis cell disintegration was performed by an alkaline lysis method [38]. The obtained lysate was directly analyzed by HPLC (Agilent 1100 with a quaternary pump and diode-array detector (DAD)). Values derived from three biological replicates have a coefficient of variation lower than 10%.

The average volumetric productivity ($p_{\mathrm{pDNA}}(t)$) and the average specific yield ($Y_{\mathrm{pDNA/biomass}}(t)$) were calculated as given in Equations (3) and (2), respectively. The time-dependent pDNA synthesis rate ($q_{\mathrm{pDNA}}(t)$) was estimated via the finitediff Python package [39, 40].

Results

Key objective

We aim to design an efficient three-stage fed-batch process (Fig. 1) for pDNA production in *E. coli*, where cellular growth and pDNA production are separated.

In the following, we will

(i) use constraint-based modeling to identify medium components that enable switching from growth to production phase;

(ii) determine optimal switching time points to maximize the average productivity in a fed-batch fermentation; and

(iii) experimentally validate the computed strategies in a linear fed-batch process.

Identification of decoupling compounds

First, we used pFBA to compute a minimal set of uptake rates in the genome-scale metabolic model *i*ML1515 [29] supporting maximal aerobic growth of *E. coli* with glucose as carbon source. Similarly, we computed uptake rates for maximal pDNA production using the same constraints (Fig. 2). All calculated uptake rates are inflexible in the optima, except for Fe^{2+} and O_2. Their uptake can be further increased by conversion to and excretion of Fe^{3+} and H_2O. In contrast to biomass synthesis, pDNA production requires only glucose, O_2, NH_4^+, and HPO_4^{2-}, but no further nutrients. Therefore, we conclude that these remaining nutrients could potentially be used as decoupling agents separating pDNA synthesis from growth.

Next, for each decoupling nutrient, we restricted its uptake between zero and 100% of its rate at maximum growth and computed the corresponding pDNA production envelopes as a function of growth (Fig. 3). All twelve decoupling nutrients result in identical sets of production envelopes, which mirrors the fact that each decoupling nutrient is essential for growth but not required for pDNA production. Note that the one-to-one trade-off

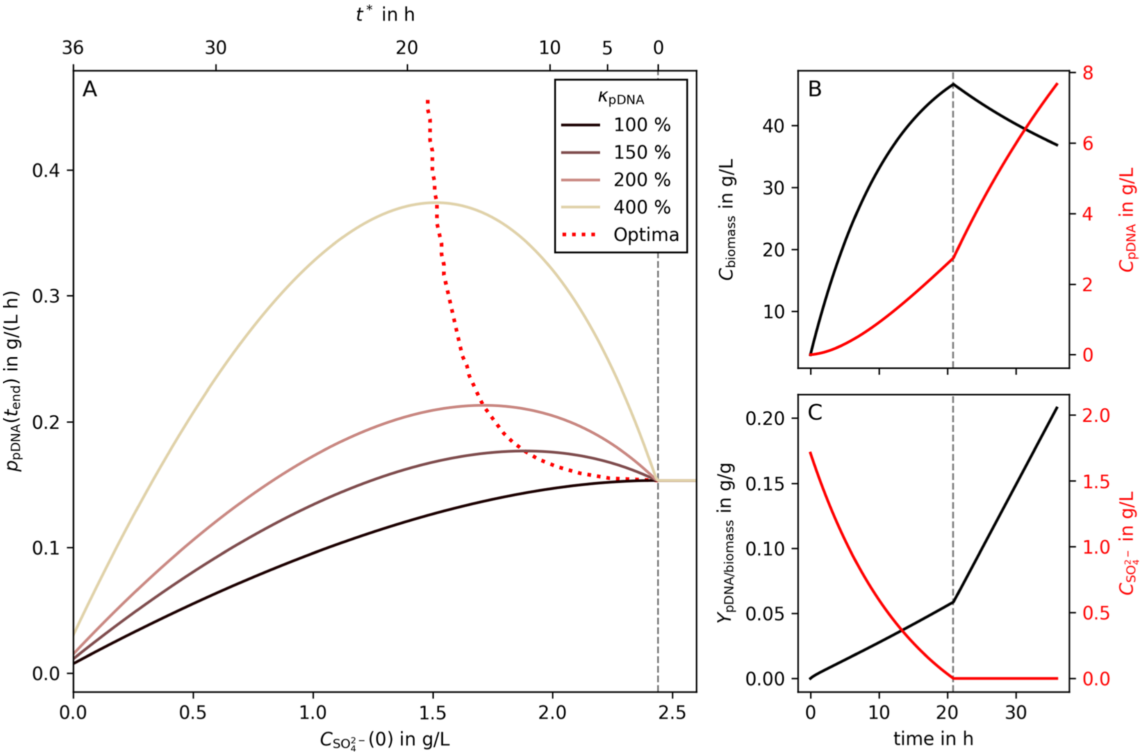


Fig. 4 Predicted optimal timing in a SO_4^{2-} limited linear fed-batch process. **A** shows average volumetric productivities of pDNA production (g $L^{-1}h^{-1}$) as a function of the initial SO_4^{2-} concentration in a linear fed-batch process. The full line represents different levels of increased pDNA production during starvation as percentages of pDNA production rate during biomass growth (κ_{pDNA}). The dotted line indicates the location of the optima for κ_{pDNA} between 100 and 500%. The second X-axis of **A** (top) illustrates the length of the sulfate starved process phase (t^*). For all modeled linear fed-batch processes, right of the gray dashed line, no SO_4^{2-} limitation occurred. **B** and **C** show the process curves of metabolites of interest in the optimal process for κ_{pDNA} = 200%. The gray dashed lines indicate the switching time point between the growth and production phases

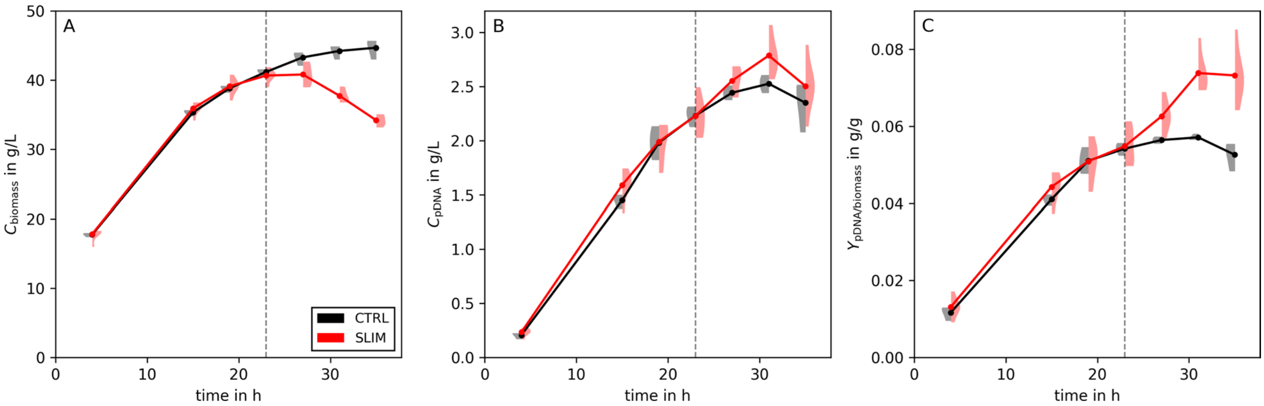


Fig. 5 Experimental results of sulfate limitation. **A** illustrates the biomass concentration, **B** the concentration of produced pDNA, and **C** the pDNA to biomass yield. The violins are calculated from triplicates of the control (i.e., no SO_4^{2-} limitation, CTRL, black) and from six replicates of the sulfate limited process (SLIM, red). The full lines and points are calculated from the mean of the replicates. The gray dashed line represents the estimated time of the switch from biomass growth to pDNA production (projected at 23 h)

between biomass production and pDNA synthesis, i.e., the upper limit of the production envelope, is a straight line between the points (0|100) and (100|0).

Realistically, pDNA production rates are significantly less than the theoretical value of 100% in Fig. 3. Therefore, in the inset of the same figure, the red markers illustrate the extreme points of more reasonable production envelopes. For example, the red circle (i.e., $\kappa_{\text{pDNA}} = 100\%$) illustrates average rates of pDNA production during cell growth. Even when the decoupling would lead to a boost by a factor of $\kappa_{\text{pDNA}} = 4$ (red cross), the resulting pDNA production flux would be only 1.3% of the theoretical maximum.

In the following, we focus on the impact of the six bulk non-metal elements (sulfur, phosphorus, oxygen, nitrogen, carbon, and hydrogen) that typically make up 97% (g g^{-1}) of the elemental biomass composition [41]. Moreover, except for potassium (and sulfate), all other predicted decoupling nutrients (iron, magnesium, calcium, chlorine, copper, manganese, zinc, nickel, cobalt, and molybdenum) are taken up at minute rates ($< 15\ \mu\text{mol g}^{-1}\text{h}^{-1}$, Fig. 2). Thus, exactly dosing their concentrations for limitation may be challenging in a bio-process. This leaves sulfate as the only predicted decoupling nutrient in a glucose-minimal medium.

Optimal sulfate limited processes

Decoupling production from growth during a bio-process raises the question of timing: when to best switch from growth to production phase to maximize performance.

In the following, we used dFBA [34] to track the time-dependent concentrations $C_i(t)$ of biomass, pDNA, glucose, and sulfate and determine the optimal initial sulfate concentration, $C_{SO_4^{2-}}(0)$, that maximize the average volumetric productivity

$$\max_{C_{SO_4^{2-}}(0)} p_{\text{pDNA}}\left(t_{\text{end}}, C_{SO_4^{2-}}(0)\right) \tag{7}$$

in a fed-batch process. Here t_{end} denotes the end of the bio-process, which terminates when the maximal volumetric capacity of the reactor is reached. Our simulations assume that pDNA production occurs (i) at a constant rate $q^*_{\text{pDNA}} = \kappa_{\text{pDNA}}\, q^{\mu}_{\text{pDNA}}$ during the sulfate starvation phase; (ii) at q^{μ}_{pDNA} during the growth phase.

We simulated sulfate limited fed-batch processes with a linear feed as sketched in Fig. 1 using the values listed in Additional file 1: Table S1. In all simulations the feed

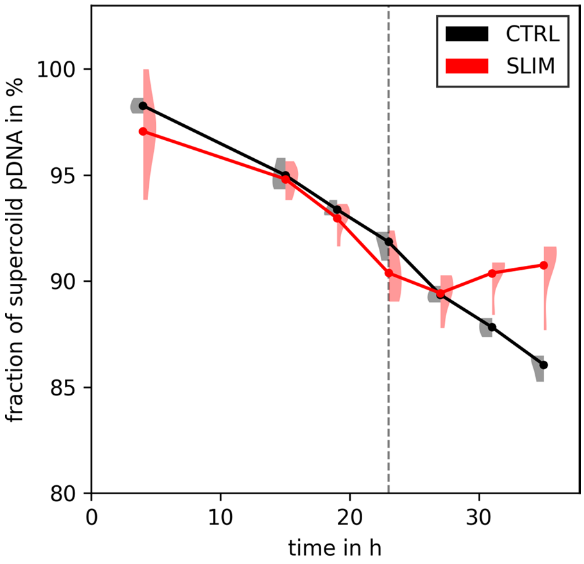


Fig. 6 Fraction of supercoiled (ccc) pDNA over time for SO_4^{2-} limited (SLIM, red) and control (CTRL, black) process. Experimental replicates are shown as violins, full lines and markers represent their means

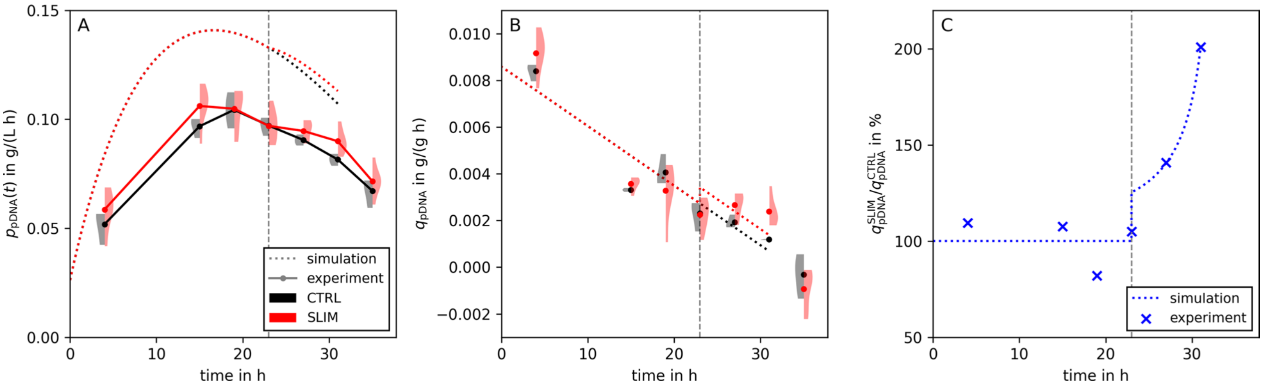


Fig. 7 Calculated experimental rates and comparison to the simulation. **A** illustrates the average volumetric productivity and **B** the pDNA production fluxes of control and SO_4^{2-} limited process (black CTRL and red SLIM, respectively). **C** shows the ratio of $q^{\text{SLIM}}_{\text{pDNA}}$ and $q^{\text{CTRL}}_{\text{pDNA}}$ (blue). Experimental replicates are shown as markers, dotted lines are calculated from simulations. To adjust the simulations to the rates obtained in the experiments, a linearly decreasing $q^{\text{CTRL}}_{\text{pDNA}}$ was fitted to experimental control data (black dotted line, **B**). Moreover, we fitted a parallel $q^{\text{SLIM}}_{\text{pDNA}}$ (red dotted line, **B**) to conform to the experimental flux ratio (panel **C**). The vertical gray dashed line represents the estimated time of switching from biomass growth to pDNA production (projected at 23 h)

rate was constant. Thus, the process length is always 36 h. Subsequently, we analyzed the impact of the length of the pDNA production phase (induced by sulfate starvation during the feed phase, i.e., stage 3 in Fig. 1) on the average volumetric productivity. An analogous analysis for a sulfate limited batch fermentation can be found in the Supplementary Notes A.3.1.

We observed distinct maxima in the average volumetric productivity of a linear fed-batch when $\kappa_{\text{pDNA}} > 1$ (red dotted line, Fig. 4A). Maxima occur at much longer starvation times compared to batch processes ($t^* = 19$ h versus 2.5 h at $\kappa_{\text{pDNA}} = 4$, Additional file 1: Fig. S1). Even compared to an equivalent exponential fed-batch, the optimal starvation is longer in a linear than in an exponential fed-batch process (Additional file 1: Fig. S2). For instance, at $\kappa_{\text{pDNA}} = 1.5$ a linear fed-batch process achieves an optimal $p_{\text{pDNA}} = 0.18$ g L^{-1} h^{-1} at $t^* = 12$ h, while an equivalent exponential fed-batch process reaches its optimum $p_{\text{pDNA}} = 0.096$ g L^{-1} h^{-1} at $t^* = 3.3$ h. Within our modelling assumptions, a linear fed-batch, even without starvation, outperforms an equivalent exponential fed-batch with (optimal) sulfate starvation as long as $\kappa_{\text{pDNA}} \leq 3.5$.

Typically, growth-decoupled processes suffer from a substantially decreasing (glucose) uptake rate during the production phase [42, 43]. Thus, our assumption of keeping an elevated q^*_{pDNA} constant over several hours may be unrealistic. Therefore, we investigated how the productivity optima change when the maximal feasible starvation length (t^*_{max}) is bounded. Yet, even in such cases, sulfate limited fed-batches perform better than standard processes without starvation (Additional file 1: Fig. S3).

Sulfate limitation experiments

The preceding analysis suggested that a three-stage fed-batch process with sulfate starvation will deliver superior pDNA production performances compared to a conventional, non-starved fed-batch process. To confirm this, we set up a linear fed-batch process with *E. coli* JM108 as host (see Sect. 2.2 for details). Based on small molecule production rates during sulfate starvation [43, 44], we assumed a $\kappa_{\text{pDNA}} = 2$ and consequently predicted $t^* = 13$ h. Thus, we computed the initial SO_4^{2-} concentration to be 3.8 g L^{-1} such that sulfate starvation occurs after 23 h in a 36 h bio-process.

Figure 5 highlights the feasibility of sulfate starvation (indicated as SLIM) to boost pDNA production in a (linear) fed-batch. Panel A illustrates the growth arrest due to sulfate starvation (compare the diverging lines to the right of the dashed line). Due to dilution, the biomass concentration (red) decreases if cells no longer grow. Yet, pDNA concentration keeps rising – even faster than in the unstarved control (compare red SLIM with black CTRL in panel B). Consequently, the specific pDNA yield rises too (panel C) reaching a maximum of 0.074 g g^{-1} after 31 h, which corresponds to an improvement of 29% compared to control ($p = 0.0005$, Additional file 1: Table S3).

Moreover, we compared the fraction of supercoiled pDNA between the SO_4^{2-} limited and control processes (Fig. 6). Up to 27 h after induction, there were no noticeable differences. After this point, the SO_4^{2-} limited process maintained a higher fraction of supercoiled pDNA ($+3\%, p = 0.0019$, Additional file 1: Table S3), resulting in a 33% ($p = 0.0001$) increase in the specific yield of supercoiled pDNA.

Beyond 31 h, pDNA concentration and specific yield decrease in both the SO_4^{2-} limited and the control process.

Next to the specific yield, also the average volumetric productivity increases by 10% at 31 h ($p = 0.0243$, Fig. 7A and Additional file 1: Fig. S3). If only the pharmacologically relevant supercoild fraction of pDNA is considered, the productivity increases by 13% ($p = 0.0052$, Additional file 1: Table S3).

To further investigate the experimental results, we computed the specific productivities in the sulfate limited and control fermentations (Fig. 7B). In both bio-processes q_{pDNA} decreases with time which is in contrast to our modeling assumptions.

Finally, we compared the specific and volumetric yield achieved by the experiments in this study (at $t = 31$ h) to published values. Table 2 shows that in terms of volumetric and specific yield the pDNA production strategy for our control values is already one of the best we could find, and with SO_4^{2-} limitation it performs better than all other published methods of our knowledge. Due to the extraordinary pDNA size in this study, the plasmid copy number is just above average. However, compared to the control, the plasmid copy number of the sulfate limited process increases by 29%.

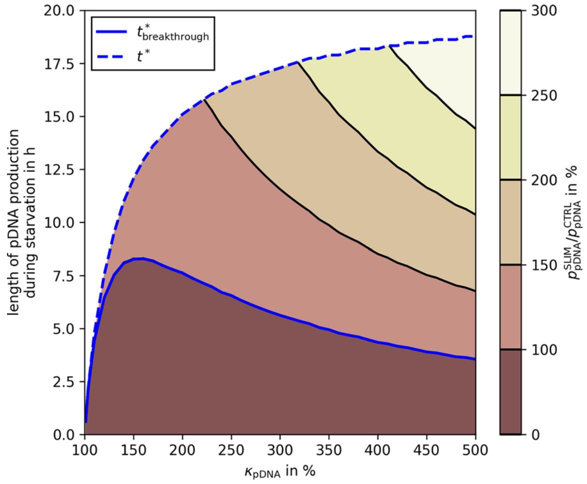


Fig. 8 Breakthrough production length during starvation in linear fed-batch processes. If the pDNA production during starvation can be held longer than the breakthrough production length ($t^*_{\text{breakthrough}}$, blue full line), the SO_4^{2-} limited process outperforms a control (i.e., not starved) process. A visual explanation of $t^*_{\text{breakthrough}}$ is given in Additional file 1: Fig. S4. The start of the starvation is defined by the optimal $c_{SO_4^{2-}}(0)$ calculated in Fig. 4A (red dotted line). The maximum pDNA production length during starvation is equal to the total starvation length t* (blue dashed line). The contour colors indicate the productivity of a SO_4^{2-} limited process compared to the control in % (color bar on the right)

Discussion

We aimed to improve pDNA productivity by designing a three-stage fed-batch process that separates cellular growth and production. Growth-decoupled processes are common design choices to enhance volumetric productivity in biochemical and biopharmaceutical production processes [43–46]. Especially with the advent of dynamic control in metabolic engineering that allows switching back and forth between metabolic growth and production phenotypes, interest in such (multi-stage) process designs has strongly grown [47]. While algorithms like MoVE [48] exist to identify intracellular metabolic switches, our focus here was on easily implementable medium modifications to induce these switches.

In this study, we identified twelve possible decoupling components (SO_4^{2-}, K^+, Ca^{2+}, Cl^-, and compounds of trace elements) for the production of pDNA (Fig. 2 and 3). All of them enable and regulate key functions in life [49]. Although trace elements act primarily as catalysts in enzyme systems, some of them, like copper and iron, play vital roles in energy metabolism [50]. However, we specifically selected SO_4^{2-} for further investigations because: (i) Sulfate is one of the six most prevalent elements in living organisms [41], which makes it comparably easy to measure and consequently determine the onset of starvation; (ii) Sulfate, in contrast to the other decoupling compounds, has a dedicated metabolic function that is well captured in the used genome-scale metabolic reconstruction *i*ML1515 of *E. coli*. [29]; (iii) Sulfate itself neither has a catalytic function nor a role in energy metabolism [51].

We simulated a three-stage fed-batch process where the transition from growth to production is triggered by the onset of sulfate starvation. Our computational model is based on two assumptions: (i) In each phase the specific pDNA production rate is constant; (ii) pDNA productivity increases upon starvation, i.e., $\kappa_{\text{pDNA}} = q^*_{\text{pDNA}}/q^{\mu}_{\text{pDNA}} > 1$. The latter aligns with experimental results of Masuda et al. [43], who reported a value of $\kappa_{\text{mevalonate}} = 1.16$ for mevalonate production during sulfate starvation.

In our experiments, we observed a decrease in the specific pDNA production for both the control and SO_4^{2-} limited process (Fig. 7B), challenging our assumption of a constant specific pDNA production rate. Investigating why q_{pDNA} decreases throughout the process will be the scope of further work. However, we implemented a

Table 2 Comparison of pDNA specific yields and volumetric yields in published studies

	Yield		Avg. vol.	Copy	Biomass	pDNA	Origin of	Selection	*E. coli*	Process
Reference	Specific	Volumetric	Productivity	Number	Conc.	Length	Replication	Pressure	Strain	Type
	[mg g^{-1}]	[mg L^{-1}]	[mg L^{-1}h^{-1}]	[plasmid cell^{-1}]	[gL^{-1}]	[kbp]				
[54]	10.8‡	58.3	3.9	NA	5.4	NA	NA	Ampicillin	BL21	Shake flask
[11]	15†*	57†	8.2†‡	298§	3.8‡	5.4	pUC	Ampicillin	DH5α	Shake flask
[13]	2.4†	8.0	0.8‡	86§	3.3‡	3.0	pUC	Kanamycin	BL21 *recA-*	Batch
[55]	2.4	19.9	0.3	42§	8.3‡	6.1	pUC	Ampicillin	VH34	Batch
[56]	2.4	30.0	2.7†‡	42§	12.5‡	6.1	pUC	Ampicillin	VH33	Batch
[19]	7.0	230	17.7	192§	34	3.9	pUC	Kanamycin	DH5α	Batch
[20]	7.9	35.9	2.3‡	529§	4.6†	1.6	pUC	Ampicillin	DH5α	Batch
[57]	11.3	39.4	1.1†‡	176§	3.5†	6.9	pUC	Ampicillin	DH5α	Batch
[17]	17.1‡	60	0.7‡	286§	3.5	6.4	ColE1	None	JM109	Batch
[16]	19.1	141	NA	553§	7.4‡	3.7	pUC	Kanamycin	GALG20	Batch
[18]	32.4	102.8	NA	526§	3.2	6.6	pBR322	Ampicillin	VH33	Batch
[10]	6.9‡	NA	NA	107§	NA	6.9	pUC	Ampicillin	W3110	Continuous
[58]	1.6	33	3.0	43§	20.6‡	4.0	pUC	None	DH5α	Exp. fed-perfusion
[59]	10.1	74.8	3.1‡	451§	8.3	2.4	pUC	Kanamycin	PFC	Fed-batch in shake flask
[12]	1.2	50	4.1	21§	41	6.1	pUC	Ampicillin	VH33	Exp. fed-batch
[60]	7.6	25.6	NA	118§	3.4‡	6.9	pUC	Ampicillin	DH5α	Exp. fed-batch
[57]	9.1	44	1.3‡	141§	4.8‡	6.9	pUC	Ampicillin	DH5α	Exp./lin. fed-batch
[61]	9.5*	250	11.3†‡	217§	26.4*	4.7	f1	Kanamycin	DH5α	Lin. fed-batch
[62]	9.8	140	3.9†‡	284§	14.3‡	3.7	ColE1	Kanamycin	GALG20	Exp. fed-batch
[63]	31.2*	1923	72.4‡	NA	62‡*	NA	pUC	Ampicillin	BL21 *recA*	Lin. fed-batch
[64]	15.5	759	31.2	426§	49‡	3.9	pUC	Kanamycin	DH5α	Exp. fed-batch
[8]	25.9	939	35.1	1142	36‡	3.5	ColE1	None	JM108*murselect*	Exp. fed-batch
[62]	55.8†‡*	2177	41†‡	965§	39†*	6.2	ColE1	Kanamycin	GALG20	Exp. fed-batch
[24]	51	2200	NA	841§	43‡	6.5	pUC	Kanamycin	DH5α	Exp. fed-batch
[65]	65*	1830	42†‡	1659§	28‡*	4.2	pUC	None	NTC4862	Exp. fed-batch
[25]	68*	2590	NA	1176§	38‡*	6.2	pUC	Kanamycin	DH5α	Exp. fed-batch
control	57.1±0.7	2528±68	81.5±2.2	510§	44.2±0.9	12.0	pUC	None	JM108	Lin. fed-batch
SO_4^{2-} limited	73.9±4.6	2788±161	89.9±5.2	660§	38.2±1.0	12.0	pUC	None	JM108	Lin. fed-batch

The last two rows show the results of the experiments of this study (all pDNA conformations). Values marked with † are estimated from published figures. Values marked with ‡ are calculated from reported values. Values marked with ∗ are converted from OD_{600} to cell dry mass (0.33g/L/OD_{600}, BNID109838 [53]). Values marked with §are estimated with 50% GC content and an *E. coli* cell mass of 110 fg (BNID100498 [53]). Values reported as NA were not accessible

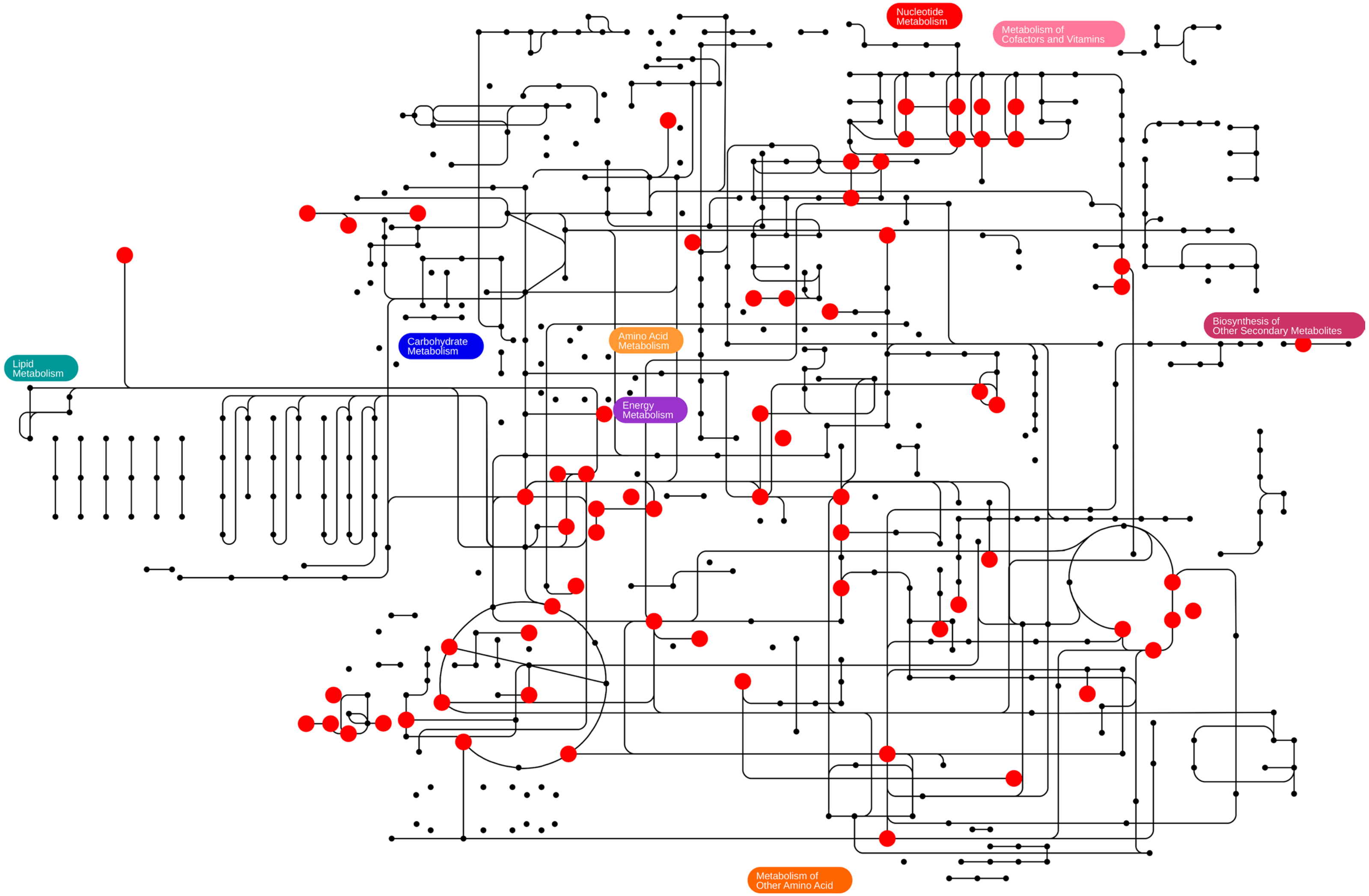

Nucleotide Metabolism
Metabolism of Cofactors and Vitamins
Biosynthesis of Other Secondary Metabolites
Lipid Metabolism
Carbohydrate Metabolism
Amino Acid Metabolism
Energy Metabolism
Metabolism of Other Amino Acid

Fig. 9 Applicability of sulfate limitation to other biotechnologcial products. Dots and lines indicate metabolites and reactions, respectively, annotated in *i*ML1515. Metabolites shown as red circles are targets for productivity improvement by sulfate limitation. The figure was created with iPath 3.0 [66]. An interactive version of the map is available at https://pathways.embl.de/selection/DfMG1nmoluPlvOcdQSt. A full list of identified targets is shown in Additional file 1: Table S4

time-dependent q_{pDNA} in additional simulations which demonstrate that the assumption of constant q_{pDNA} is not necessary for process improvements by SO_4^{2-} limitation (dotted lines in Fig. 7 and Additional file 1: Fig. S6) Even the optimal switching time changes by less than 1 h (Additional file 1: Fig. S7). Sulfate starvation always increases pDNA production, provided that $\kappa_{\text{pDNA}} > 1$, regardless of the process (batch, exponential or linear fed-batch). In fact, our data consistently shows higher specific pDNA production rates during starvation compared to control (Fig. 7C). This trend reinforces and validates our core assumption of $\kappa_{\text{pDNA}} > 1$ where the length of starvation required to maximize volumetric productivity strongly depends on its exact value.

Maintaining high κ_{pDNA} during sulfate starvation is a key requirement of our design. Our predictions are based on continuously elevated levels of pDNA productivity throughout starvation. For long starvation phases, this assumption may not be feasible [52]. However, Fig. 8 illustrates that this assumption is not particularly crucial. In the worst case (at $\kappa_{\text{pDNA}} = 160\%$), pDNA production needs to be maintained for 8.3 h to perform at least as well as a non-starved process. Our experimental data (Figs. 5 and 7) demonstrate that this is indeed feasible. Interestingly, if κ_{pDNA} is raised beyond 160%, the optimal starvation time increases too, but the minimally required length of pDNA production during starvation drops. This hints at a possible trade-off that may be explored in further process optimization steps.

In a growth-decoupled process, it is essential to, first, reach high biomass which can subsequently catalyze product formation. With our process settings (i.e., fixed final process volume), this is best achieved with a linear feeding regime (Additional file 1: Fig. S2), which quickly builds up biomass during the first few hours (compare Additional file 1: Fig. S2B and E).

We experimentally verified that a linear feeding strategy (which results in a continuously decreasing growth rate) outperforms the exponential feed even without sulfate limitation (data not shown). This is in agreement with literature which shows that a lower growth rate is preferential for pDNA production [10–12]. Interestingly a literature survey (Table 2) reveals that exponential feeding strategies are more frequently used which may explain why even our linear control process is able to outperform the majority of previously reported values. However, an ultimate comparison cannot be made as these studies used different plasmids which may significantly influence the evaluation metrics of the production process [21].

A key challenge in any growth-decoupled process is to maintain metabolic activity in non-growing metabolic states. Often a strong decrease in nutrient uptake is observed [43]. However, during sulfate starvation, glucose concentration in the reactor remained below the detection limit, indicating that cells consistently maintained glucose uptake equal to the glucose feed rate. This supports the validity of our assumption stated in Eq. 5. We speculate that this may be related to the fact that (i) due to the linear feed, the specific glucose uptake already dropped to 4% of its initial value at the onset of starvation – 81% lower than the (already) reduced specific glucose uptake rate during sulfate starvation reported by Masuda et al. [43]; (ii) sulfate starvation retains high ATP-levels compared to other nutrient limitations [45, 67].

Consistent with maintained metabolic activity, we detected acetate accumulation during starvation (Additional file 1: Fig. S5), which is a common sign of overflow metabolism in *E. coli* [68]. However, our theoretical predictions for the maximum acetate concentration exceeded the measured values, suggesting the existence of other byproducts. Identifying these will be the focus of future research.

In the future, additional refinement of the process model could be achieved by including a term for the metabolic burden of resistance protein synthesis. An appropriate computational framework has recently been published [69]. Several studies have shown that this may significantly influence the metabolism of a producing organism [7, 8].

In both control and SO_4^{2-} limited experiments, specific pDNA yields and concentrations dropped at the end of the bioprocess. This might be due to other limitations (e.g., the O_2 transfer rate [70]). Therefore, we suggest stopping the process at 31 h. At that point product concentration, average volumetric productivity, and specific yield are statistically significantly up by 10%, 10%, and 29%, respectively (compared to control). Considering the fraction of supercoiled pDNA, the sulfate limited process gains another 3% points to concentration and productivity, and 4% points to specific yield. A mechanistic interpretation of this interesting observation, however, is outside the scope of the current methodology and will be the focus of further work.

Despite using a defined, minimal medium and a comparatively large plasmid (12 kbp) our process outperforms previous reports [25] achieving an 8% increase in volumetric yield and a 9% boost in specific yield (Table 2). The latter is not only advantageous for higher pDNA quantities in sulfate limited experiments but, importantly, also aids in downstream processing [71, 72].

To showcase the broad applicability of sulfate limitation in biotechnological production, we explored the effect of sulfate limitation on 307 naturally secreted products of *E. coli* that are listed in the *i*ML1515 model [29]. Through lexicographic FBA with and without sulfate in the medium, we identified 83 compounds—more than 25% of those investigated—that would benefit from sulfate limitation in their production (Fig. 9). This underscores the significant potential of sulfate limited process design for a wide range of biotechnological products.

Conclusion

Based on genome-scale metabolic modeling, we have designed and successfully validated a three-stage, growth-decoupled fed-batch process for pDNA production in *E. coli*, cultivated in a minimal medium. We achieved the transition between the growth and production phases through sulfate starvation. This optimization led to statistically significant increases in key metrics: average supercoiled volumetric productivity (+13%), specific pDNA yield (+29%), and supercoiled specific pDNA yield (+33%). Overall, our process achieved a specific pDNA yield of 74 mg g^{-1} and a volumetric yield of 8 g L^{-1}, marking an increase of more than 8% compared to prior reports. Importantly, our process design may be benefitial to a wide range of bio-based products of industrial significance.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12934-023-02248-2.

Additional file 1: Table S1. Parameters for dFBA simulations. Cells marked with "–" are not applicable and cells marked with "var" are not constant in the respective process. **Table S2.** Growth media components of the experiments. Growth media were sterilized by filtration. The trace element solution was prepared in 5 M HCl and contained (g/L): 4.41 $CaCl_2 \cdot 2H_2O$, 3.34 $FeSO_4 \cdot 7H_2O$, 1.43 $CoCl_2 \cdot 6H_2O$, 1.03$MnSO_4 \cdot H_2O$, 0.15 $CuSO_4 \cdot 5H_2O$, 0.17 $ZnSO_4 \cdot 7H_2O$. † In total, 9.6 mL of a 200 g/L $MgSO_4 \cdot 7\ H_2O$ stock solution were pulsed into the bioreactor (3.2 mL at feed start, 3.2 mL at 7 h after feed start and 3.2 mL at 15 h after feed start). ‡ According to our stoichiometric model [29], the amount of $MgSO_4 \cdot 7\ H_2O$ in the batch medium alone can support more than 50 g of biomass. However, to ensure that Mg^{2+} is present in excess throughout the sulfate limited process we further added Mg^{2+} in the form of $MgCl_2 \cdot 6\ H_2O$. **Table S3.** Average values of several process variables of interest and their *p*-values compared to the control at 31 h (one-sided *t*-test). ccc pDNA corresponds to the supercoiled fraction of plasmid DNA measured.**Table S4.** List of potential targets for process optimization with sulfate limitation. We used lexicographic FBA (Table 1) to calculate maximal theoretical production rates during sulfate starvation (max. q^*) and sulfate excess (max. q^μ) in mmol g^{-1} h^{-1}. Compounds marked with † are displayed with their BiGG ID [75] to shorten the name. Previously described sulfate limitation target mevalonate is not naturally synthesized in E. coli [43] and, thus, not present in the list. **Figure S1.** Predicted optimal timing in a SO_4^{2-} limited linear fed-batch (**A, B, C**) and batch (**D, E, F**) process. **Figure S2.** Predicted optimal timing in a SO_4^{2-} limited linear (**A, B, C**) and exponential (**D, E, F**) fed-batch process. **Figure S3.** Predicted optimal timing in a SO_4^{2-} limited linear fed-batch process with different maximum starvation lengths. **Figure S4.** Visualization explanation of $t^*_{\text{breakthrough}}$. **Figure S5.** Acetate accumulation during control and SO_4^{2-} limited processes. **Figure S6.** Control and sulfate limited process with time-dependent q_{pDNA}. **Figure S7.** Predicted optimal timing in a SO_4^{2-} limited linear fed-batch process with variable q_{pDNA}.

Acknowledgements
Not applicable.

Author contributions
MG: conceptualization, methodology, soft- ware, formal analysis, investigation, visualization, writing—original draft, review and editing. FS: methodology, formal analysis, investigation, writing—review and editing. FW: methodology, formal analysis, investigation, writing - review and editing. PG: funding, writing—review and editing. BK: funding, writing—review and editing. JM: funding, writing - review and editing. JZ: conceptualization, funding, writing—original draft, review and editing.

Funding
Open access funding provided by University of Vienna. MG and JZ received funding from enGenes Biotech GmbH and Baxalta Innovations GmbH, a part of Takeda companies.

Availablility of data and materials
All scripts and data are available at https://github.com/Gotsmy/slim.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
FS and FW are employees of enGenes Biotech GmbH. JM is co-founder and Chief Executive Officer of enGenes Biotech GmbH. PG and BK are employees of Baxalta Innovation GmbH. Employees of Baxalta Innovations GmbH may be owners of stock and/or stock options. MG, JZ, FS, FW, and JM are authors of a patent application that has been filed on the basis of the reported results.

Author details
[1]Department of Analytical Chemistry, University of Vienna, Vienna 1090, Austria. [2]Doctorate School of Chemistry, University of Vienna, Vienna 1090, Austria. [3]enGenes Biotech GmbH, Vienna 1190, Austria. [4]Baxalta Innovations GmbH, A Part of Takeda Companies, Orth an der Donau 2304, Austria.

Received: 21 September 2023 Accepted: 11 November 2023
Published online: 28 November 2023

References

1. Mairhofer J, Lara AR. Advances in host and vector development for the production of plasmid DNA vaccines, cancer vaccines. Berlin: Springer; 2014. p. 505–41.
2. European Medicines Agency. 2020. Comirnaty assessment report, https://www.ema.europa.eu/en/documents/assessment-report/comirnaty-epar-public-assessment-report_en.pdf. Accessed 28 Jul 2022.
3. Schmidt A, Helgers H, Vetter FL, Juckers A, Strube J. Fast and flexible mRNA vaccine manufacturing as a solution to pandemic situations by adopting chemical engineering good practice-continuous autonomous operation in stainless steel equipment concepts. Processes. 2021;9:1874.
4. Ramamoorth M, Narvekar A. Non viral vectors in gene therapy-an overview. J Clin Diagnostic Res JCDR. 2015;9(1):GE01.
5. Cherng J-Y, Schuurmans-Nieuwenbroek N, Jiskoot W, Talsma H, Zuidam N, Hennink W, Crommelin D. Effect of DNA topology on the transfection

efficiency of poly ((2-dimethylamino) ethyl methacrylate)-plasmid complexes. J Controll Release. 1999;60:343–53.
6. Cupillard L, Juillard V, Latour S, Colombet G, Cachet N, Richard S, Blanchard S, Fischer L. Impact of plasmid supercoiling on the efficacy of a rabies DNA vaccine to protect cats. Vaccine. 2005;23:1910–6.
7. Cunningham DS, Koepsel RR, Ataai MM, Domach MM. Factors affecting plasmid production in Escherichia coli from a resource allocation standpoint. Microb Cell Factories. 2009;8:1–17.
8. Mairhofer J, Cserjan-Puschmann M, Striedner G, Nöbauer K, Razzazi-Fazeli E, Grabherr R. Marker-free plasmids for gene therapeutic applications-lack of antibiotic resistance gene substantially improves the manufacturing process. J Biotechnol. 2010;146:130–7.
9. Bower DM, Prather KL. Engineering of bacterial strains and vectors for the production of plasmid DNA. Appl Microbiol Biotechnol. 2009;82:805–13.
10. Wunderlich M, Taymaz-Nikerel H, Gosset G, Ramírez OT, Lara AR. Effect of growth rate on plasmid DNA production and metabolic performance of engineered Escherichia coli strains. J Biosci Bioeng. 2014;117:336–42.
11. Galindo J, Barrón BL, Lara AR. Improved production of large plasmid DNA by enzyme-controlled glucose release. Ann Microbiol. 2016;66:1337–42.
12. Soto R, Caspeta L, Barrón B, Gosset G, Ramírez OT, Lara AR. High cell-density cultivation in batch mode for plasmid DNA production by a metabolically engineered e, coli strain minimized overflow metabolism,. Biochem Eng J. 2011;56:165–71.
13. Jaén KE, Velazquez D, Delvigne F, Sigala JC, Lara AR. Engineering e. coli for improved microaerobic PDNA production. Bioprocess Biosyst Eng. 2019;42:1457–66.
14. Lara AR, Jaén KE, Folarin O, Keshavarz-Moore E, Büchs J. Effect of the oxygen transfer rate on oxygen-limited production of plasmid DNA by Escherichia coli. Biochem Eng J. 2019;150: 107303.
15. Cunningham DS, Liu Z, Domagalski N, Koepsel RR, Ataai MM, Domach MM. Pyruvate kinase-deficient Escherichia coli exhibits increased plasmid copy number and cyclic amp levels. J Bacteriol. 2009;191:3041–9.
16. Gonçalves GA, Prazeres DM, Monteiro GA, Prather KL. De novo creation of mg1655-derived e. coli strains specifically designed for plasmid DNA production. Appl Microbiol Biotechnol. 2013;97:611–20.
17. Wang Z, Le G, Shi Y, Wegrzyn G. Medium design for plasmid DNA production based on stoichiometric model. Process Biochem. 2001;36:1085–93.
18. Martins L, Pedro A, Oppolzer D, Sousa F, Queiroz J, Passarinha L. Enhanced biosynthesis of plasmid DNA from Escherichia coli vh33 using box-Behnken design associated to aromatic amino acids pathway. Biochem Eng J. 2015;98:117–26.
19. Islas-Lugo F, Vega-Estrada J, Alvis CA, Ortega-Lopez J, del Carmen Montes-Horcasitas M. Developing strategies to increase plasmid DNA production in Eescherichia coli dh5α using batch culture. J Biotechnol. 2016;233:66–73.
20. Danquah MK, Forde GM. Growth medium selection and its economic impact on plasmid DNA production. J Biosci Bioeng. 2007;104:490–7.
21. Kay A, O'Kennedy R, Ward J, Keshavarz-Moore E. Impact of plasmid size on cellular oxygen demand in Escherichia coli. Biotechnol Appl Biochem. 2003;38:1–7.
22. Folarin O, Nesbeth D, Ward JM, Keshavarz-Moore E. Application of plasmid engineering to enhance yield and quality of plasmid for vaccine and gene therapy. Bioengineering. 2019;6:54.
23. Jaén KE, Velázquez D, Sigala J-C, Lara AR. Design of a microaerobically inducible replicon for high-yield plasmid DNA production. Biotechnol Bioeng. 2019;116:2514–25.
24. Williams JA, Luke J, Langtry S, Anderson S, Hodgson CP, Carnes AE. Generic plasmid DNA production platform incorporating low metabolic burden seed-stock and fed-batch fermentation processes. Biotechnol Bioeng. 2009;103:1129–43.
25. Carnes AE, Luke JM, Vincent JM, Schukar A, Anderson S, Hodgson CP, Williams JA. Plasmid DNA fermentation strain and process-specific effects on vector yield, quality, and transgene expression. Biotechnol Bioeng. 2011;108:354–63.
26. Jaén KE, Lara AR, Ramírez OT. Effect of heating rate on PDNA production by e. coli. Biochem Eng J. 2013;79:230–8.
27. Monteiro G, Duarte S, Martins M, Andrade S, Prazeres D. High copy number plasmid engineering for biopharmaceutical-grade PDNA production in Lactococcus lactis. J Biotechnol. 2019;305:S23–4.
28. Duarte SO, Monteiro GA. Plasmid replicons for the production of pharmaceutical-grade PDNA, proteins and antigens by Lactococcus lactis cell factories. Int J Mol Sci. 2021;22:1379.
29. Monk JM, Lloyd CJ, Brunk E, Mih N, Sastry A, King Z, Takeuchi R, Nomura W, Zhang Z, Mori H, et al. I ml1515, a knowledgebase that computes Escherichia coli traits. Nat Biotechnol. 2017;35:904–8.
30. Ebrahim A, Lerman JA, Palsson BO, Hyduke DR. Cobrapy: constraints-based reconstruction and analysis for python. BMC Syst Biol. 2013;7:1–6.
31. Neidhardt FC, Ingraham JL, Schaechter M. Physiology of the bacterial cell; a molecular approach, 589.901 N397, Sinauer associates, 1990.
32. Sun Y, Fleming RM, Thiele I, Saunders MA. Robust flux balance analysis of multiscale biochemical reaction networks. BMC Bioinformatics. 2013;14:1–6.
33. Lewis NE, Hixson KK, Conrad TM, Lerman JA, Charusanti P, Polpitiya AD, Adkins JN, Schramm G, Purvine SO, Lopez-Ferrer D, et al. Omic data from evolved e. coli are consistent with computed optimal growth from genome-scale models. Mol Syst Biol. 2010;6:390.
34. Mahadevan R, Edwards JS, Doyle FJ III. Dynamic flux balance analysis of diauxic growth in Escherichia coli. Biophys J. 2002;83:1331–40.
35. Höffner K, Harwood SM, Barton PI. A reliable simulator for dynamic flux balance analysis. Biotechnol Bioeng. 2013;110:792–802.
36. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, Bright J, et al. Scipy 1.0: fundamental algorithms for scientific computing in python. Nat Methods. 2020;17:261–72.
37. Yanisch-Perron C, Vieira J, Messing J. Improved m13 phage cloning vectors and host strains: nucleotide sequences of the m13mpl8 and puc19 vectors. Gene. 1985;33:103–19.
38. Bimboim HC, Doly J. A rapid alkaline extraction procedure for screening recombinant plasmid DNA. Nucl Acids Res. 1979;7:1513–23.
39. Dahlgren B., Čertík O. finitediff, 2021. https://github.com/bjodah/finitediff. 10.5281/zenodo.5168369.
40. Fornberg B. Classroom note: calculation of weights in finite difference formulas. SIAM Review. 1998;40:685–91.
41. Lange HC, Heijnen JJ. Statistical reconciliation of the elemental and molecular biomass composition of Saccharomyces cerevisiae. Biotechnol Bioeng. 2001;75:334–44.
42. Klamt S, Mahadevan R, Hädicke O. When do two-stage processes outperform one-stage processes? Biotechnol J. 2018;13:1700539.
43. Masuda A, Toya Y, Shimizu H. Metabolic impact of nutrient starvation in mevalonate-producing Escherichia coli. Bioresour Technol. 2017;245:1634–40.
44. Willrodt C, Hoschek A, Bühler B, Schmid A, Julsing MK. Decoupling production from growth by magnesium sulfate limitation boosts de novo limonene production. Biotechnol Bioeng. 2016;113:1305–14.
45. Tokuyama K, Toya Y, Matsuda F, Cress BF, Koffas MA, Shimizu H. Magnesium starvation improves production of malonyl-coa-derived metabolites in Escherichia coli. Metabol Eng. 2019;52:215–23.
46. Stargardt P, Feuchtenhofer L, Cserjan-Puschmann M, Striedner G, Mairhofer J. Bacteriophage inspired growth-decoupled recombinant protein production in Escherichia coli. ACS Synth Biol. 2020;9:1336–48.
47. Hartline CJ, Schmitz AC, Han Y, Zhang F. Dynamic control in metabolic engineering: theories, tools, and applications. Metabol Eng. 2021;63:126–40.
48. Venayak N, von Kamp A, Klamt S, Mahadevan R. MoVE identifies metabolic valves to switch between phenotypic states. Nat Commun. 2018;9:5332.
49. Maret W, Blower P. Teaching the chemical elements in biochemistry: elemental biology and metallomics. Biochem Mol Biol Education. 2022;50:283–9.
50. N. R. C. U. C. O. D. A. Health. 1989. Trace Elements, in: Diet and Health: Implications for Reducing Chronic Disease Risk. Washington (DC): National Academies Press. 301.
51. Markovich D. Physiological roles and regulation of mammalian sulfate transporters. Physiol Rev. 2001;81:1499–533.
52. St John A, Goldberg A. Effects of starvation for potassium and other inorganic ions on protein degradation and ribonucleic acid synthesis in Escherichia coli. J Bacteriol. 1980;143:1223–33.

53. Milo R, Jorgensen P, Moran U, Weber G, Springer M. Bionumbers-the database of key numbers in molecular and cell biology. Nucl Acids Res. 2010;38:D750–3.
54. Xu Z-N, Shen W-H, Chen H, Cen P-L. Effects of medium composition on the production of plasmid DNA vector potentially for human gene therapy. J Zhejiang Univ Sci B. 2005;6:396.
55. Pablos TE, Soto R, Mora EM, Le Borgne S, Ramírez OT, Gosset G, Lara AR. Enhanced production of plasmid DNA by engineered Escherichia coli strains. J Biotechnol. 2012;158:211–4.
56. Cortés JT, Flores N, Bolívar F, Lara AR, Ramírez OT. Physiological effects of ph gradients on Escherichia coli during plasmid DNA production. Biotechnol Bioeng. 2016;113:598–611.
57. Dorward A, O'Kennedy RD, Folarin O, Ward JM, Keshavarz-Moore E. The role of amino acids in the amplification and quality of DNA vectors for industrial applications. Biotechnol Progress. 2019;35: e2883.
58. García-Rendón A, Munguía-Soto R, Montesinos-Cisneros RM, Guzman R, Tejeda-Mansir A. Performance analysis of exponential-fed perfusion cultures for PDNA vaccines production. J Chem Technol Biotechnol. 2017;92:342–9.
59. de la Cruz M, Ramírez EA, Sigala J-C, Utrilla J, Lara AR. Plasmid DNA production in proteome-reduced Escherichia coli. Microorganisms. 2020;8:1444.
60. O'Kennedy RD, Ward JM, Keshavarz-Moore E. Effects of fermentation strategy on the characteristics of plasmid DNA production. Biotechnol Appl Biochem. 2003;37:83–90.
61. O'Mahony K, Freitag R, Hilbrig F, Müller P, Schumacher I. Strategies for high titre plasmid DNA production in Escherichia coli dh5α. Process Biochem. 2007;42:1039–49.
62. Goncalves GA, Prather KL, Monteiro GA, Carnes AE, Prazeres DM. Plasmid DNA production with Escherichia coli galg20, a pgi-gene knockout strain: Fermentation strategies and impact on downstream processing. J Biotechnol. 2014;186:119–27.
63. Phue J-N, Lee SJ, Trinh L, Shiloach J. Modified Escherichia coli b (bl21), a superior producer of plasmid DNA compared with Escherichia coli k (dh5α). Biotechnol Bioeng. 2008;101:831–6.
64. Grijalva-Hernández F, Vega-Estrada J, Escobar-Rosales M, Ortega-López J, Aguilar-López R, Lara AR, Montes-Horcasitas MDC. High kanamycin concentration as another stress factor additional to temperature to increase PDNA production in e coli dh5α batch and fed-batch cultures. Microorganisms. 2019;7:711.
65. Nelson J, Rodriguez S, Finlayson N, Williams J, Carnes A. Antibiotic-free production of a herpes simplex virus 2 DNA vaccine in a high yield cgmp process. Human Vaccines Immunother. 2013;9:2211–5.
66. Darzi Y, Letunic I, Bork P, Yamada T. Ipath3. 0: interactive pathways explorer v3. Nucl Acids Res. 2018;46:W510–3.
67. Chubukov V, Sauer U. Environmental dependence of stationary-phase metabolism in bacillus subtilis and Escherichia coli. Appl Environ Microbiol. 2014;80:2901–9.
68. El-Mansi E, Holms W. Control of carbon flux to acetate excretion during growth of Escherichia coli in batch and continuous cultures. Microbiology. 1989;135:2875–83.
69. Oftadeh O, Hatzimanikatis V. Application of genome-scale models of metabolism and expression to the simulation and design of recombinant organisms. bioRxiv. 2023:2023–09.
70. Castan A, Enfors S-O. Characteristics of a do-controlled fed-batch culture of Escherichia coli. Bioprocess Eng. 2000;22:509–15.
71. Prather KJ, Sagar S, Murphy J, Chartrain M. Industrial scale production of plasmid DNA for vaccine and gene therapy: plasmid design, production, and purification. Enzyme Microbial Technol. 2003;33:865–83.
72. Williams JA, Carnes AE, Hodgson CP. Plasmid DNA vaccine vector design: impact on efficacy, safety and upstream production. Biotechnol Adv. 2009;27:353–70.
73. Wong P, Gladney S, Keasling JD. Mathematical model of the lac operon: inducer exclusion, catabolite repression, and diauxic growth on glucose and lactose. Biotechnol Progress. 1997;13:132–43.
74. Varma A, Palsson BO. Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type Escherichia coli w3110. Appl Environ Microbiol. 1994;60:3724–31.
75. King ZA, Lu J, Dräger A, Miller P, Federowicz S, Lerman JA, Ebrahim A, Palsson BO, Lewis NE. Bigg models: a platform for integrating, standardizing and sharing genome-scale models. Nucl Acids Res. 2016;44:D515–22.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

B. Appendix II: Supplementary Information

B.1. Supplementary Tables

Table B.1.: Parameters used for the calculation of the time series shown in Figure 1.9.

Parameter	Unit	Exponential Fed-Batch	Linear Fed-Batch
X_0	g	1.52	1.52
P_0	g	0	0
V_0	L	0.5	0.5
$Y_{X/G}$	$\mathrm{g\,g^{-1}}$	0.54	0.54
$Y_{P/G}$	$\mathrm{g\,g^{-1}}$	0.51	0.51
γ_P	$\mathrm{g\,h^{-1}\,g^{-1}}$	.0061	.0061
γ_M	$\mathrm{g\,h^{-1}\,g^{-1}}$	.16	.16
$\mu_{\max}$	$\mathrm{h^{-1}}$	–	2
μ	$\mathrm{h^{-1}}$	0.12	–
C_G	$\mathrm{g\,L^{-1}}$	330	330