UNIVERSITÄT
LEIPZIG

UFZ HELMHOLTZ
Zentrum für Umweltforschung

Master's thesis at the University of Leipzig

for the attainment of the degree: Master of Science

Faculty of Mathematics and Computer Science

Institute of Computer Science

Visualization and characterization of cross-feeding networks of microbial communities

Jens-Tilman Rau

Supervision:
Dr. Florian Centler (UFZ)
Prof. Dr. Peter Stadler (University of Leipzig)

Submitted by Jens-Tilman Rau, on the _____

Table of Contents

1. Introduction

Human biology, especially in school, is often taught in a way of explaining general functions of our body, eventually coming down to our genetics, which everything is based on. This deterministic idea of how the human body works as an autonomous programmed system is then often most challenged by the concept of epigenetics and the role of the human microbiome. The former is rather a new insight into how gene regulation is regulated by a complex system of modifiers, reacting to external environmental cues [1] but is still an integral part of the human body. However, the role of symbiotic microbiota really is a figurative add-on, in a way one could even think of it as "epi-human-genetics" (*epi* - greek prefix for "above"), because it goes beyond the abilities of the sole human body. If one defines the ability of an organism by the number of encoding genes, this is a major upgrade, multiplying the existing tool set by at least a factor of 150, just considering the human gut microbiota [2] [3].

In general microorganisms are the major contributor to Earth's biodiversity. But up to the beginning of the 21st century studies of microbial species, their constitution and functional properties, mostly involved *in vitro* cultivation. Findings therefore were culture-dependent, which posed a natural limit to further insights [4] [5]. The advent of bacterial ribosomal RNA sequences used to phylogenetically relate and enumerate constituents of mixed microbial communities then allowed in situ speciation of samples independent of cultivations [6]. Followed by the further evolution of the 16S rRNA marker gene technique and the upcoming revolution of DNA sequencing, from Sanger sequencing, over next generation sequencing to single molecule sequencing, studying the composition of microbial communities was raised to a whole new level. Metagenomic analysis now allows insights into the functional tool set of microorganisms or communities as an entirety, complemented by multiple omics methods like transcriptomics, proteomics and metabolomics, shedding light on the activity and interaction within these communities in their natural habitats [7]. This enabled ambitious projects like the Human Microbiome Project [8] to assess human associated microbiomes and their host interaction. It brings up the idea of a human 'supra-organism', which is the composite of both human cells and the cells of its microbiome. Omics data is now organized in databases like the famous KEGG database [9], the integrated microbial genomes database and comparative analysis system IMG [10], the metabolic pathways database MetaCyc [11] or, more recently and more specific, a

healthy human gut microbiome reference database [12]. This data can then be used to reconstruct artificial genome wide metabolic models of microbial communities. With these mathematical models, simulations can emulate a microbial community at different environmental conditions. These *in silico* studies can yield insights into the metabolic exchange of constituents of a community. Different scenarios can be simulated without much effort, once a model is created and fit for a simulator. Thus, e.g. different media compositions can be tried out, gene knockouts can be simulated, and model derived predictions might be transferred to *in vitro* studies for validation purposes. Yet in order to reach conclusions which are applicable to an experimental laboratory setting, the simulated data has to be analyzed, ideally implementing pipelines that can quickly asses the outcome of a simulation and provide an analytical framework. This is where data visualization comes into play, as the simulation of metabolic models yields a dynamic metabolic network of compounds and microorganisms, which can be represented by a graph with dynamically changing attributes. Its visual and graph theoretical properties can then be used to exploit interesting structural features to infer functional characteristics of the underlying metabolic model. As metabolic exchange networks can get very messy, it is crucial to implement techniques to lay out, simplify and highlight the network representation in a meaningful and directed way.

2. Theory

As huge volumes of data are continuously piling up, permeating into various aspects of life, data of biological origin does not make an exemption. Especially the advent of high throughput sequencing changed the pace at which data is obtained and it is still accelerating. And although computer power and data mining techniques evolved considerably during the same period, it seems that the bottleneck has shifted towards the curation and analysis of this data rather than how to obtain it [13]. Topic related data is organized in huge data bases, some even free to access. In system biology, this abundance of interconnected data offers immense prospects to mathematical models to describe complex interactive systems and their overall functions [14]. Genome-scale metabolic models are not new, but both quality and quantity of omics data allow reconstructing and refining models rather quickly today. Studying microbial communities, these mathematical formats of metabolic networks can then be used to simulate the evolution of a consortium, allowing computational studies examining cross-feeding interactions between the

species, testing media compositions or the effects of simulated knockouts on the system as a whole [15].The advantage of analyzing simulated data over experimental data is that in a time course experiment, temporal data is usually limited on sample frequency, whereas generated dynamic networks series are much more continuous regarding most measurements, thus allowing a higher resolution of the actual process. Furthermore, the actual metabolic exchange between the constituents of a microbial community is hard to measure *in situ*, as it requires tagging atoms in order to follow their journey through the metabolic web. This is possible in wet-lab experiments by incorporating stable isotope atoms into substrates [16] or using stable isotope amino acids [17]. By adding these to the reactor medium, the assimilation of the altered elements into proteins can be detected by mass spectrometry. The extent of incorporation is directly linked to the metabolic activity of the consortia and can identify key processes and constituents. However, these methods are suited for microbial communities of low complexity under laboratory conditions, not for *in situ* field experiments. In contrast, computational studies can examine the metabolic cross-feeding network of big microsoms like the human gut microbiome and examine exchange reactions between the single constituents, thus generating hypotheses which then might motivate target-oriented experiments.

2.1. Simulator *µbialSim*

The simulated data which is analyzed in this thesis was obtained by a batch and chemostat simulation of the SIHUMIx community in *µbialSim* [18] [19]. It is a dynamic Flux-Balance-Analysis-based numerical simulator, focused on evaluating microbial metabolic interactions. It takes a genome-scale-metabolic model and reactor parameters as an input, and predicts the evolution of medium composition and metabolic activity within the simulated reactor.

2.2. Metabolic interactions in microbial communities

Microbial communities are often very diverse biospheres, with many different species competing on a limited pool of substrates. One would think of a very competitive environment considering these circumstances, and there are many known antagonistic interactions between microorganisms, like producing toxins for instance [20]. And although those interactions might still prevail over beneficial synergistic interactions, auxotrophic bacteria are present, especially in nutrient rich environments [21]. The question of why organisms would abandon their autonomy by losing essential metabolic function trough gene loss and thus becoming dependent

on excreted products of other organisms is still not fully answered. Most arguments follow along the lines of the Black Queen hypothesis [22], which states that evolutionary the adaptive gene loss could be driven by selection, as omitting certain biosynthetic function could be favorable for an organism, if there are public goods produced by the community, which can be used by the organism at lower costs. For instance, M. Mee at al. find, that outsourcing the synthesis of costly amino-acids can be beneficial to an organism in a microbial community with metabolic cross-feeding activity [23]. The term "metabolic cross-feeding" was used inconsistently for various types of synergistic metabolic interactions among bacteria. N. Smith at al. therefore defined a new definition which says, that there is a metabolic cross-feeding interaction between bacteria strains, if a metabolite is produced by one strain and consumed by any other [24]. The introduction of artificially engineered interspecies dependencies could incorporate metabolic cross-feeding into communities and even yielded overall beneficial effects, like more uniform growth of the entire community [25]. As many auxotrophic bacteria are known today, this creates a void for key organisms as producer of public goods in microbial communities, which are essential to their community. In order to understand the dynamics within microbial consortia it is therefore crucial to identify metabolites different strains are cross-feeding on, and determine key producer of these intermediate compounds. This can help to better control microbial communities in biotechnological applications [26] and human health related issues regarding host microbiome interactions [27].

In order to study metabolic interactions within microsoms, the classical way would be culture-dependent studies. Ironically, if there are metabolic cross-feeding interactions in a microbial ecosystem, this approach might be hampered by the very existence of these interactions, as *in vitro* cultivation of a field sample might not provide an environment, where theses interactions are still in place and allow auxotrophic strains to grow [28]. Still, by growing environmental samples plated out on different media plates, supplemented with metabolites potentially involved in cross-feeding interaction, one can find auxotrophic bacteria. To identify actual cross-feeding behavior between organisms, a series of monocultures and cocultures on different growth media can identify the presence and type metabolic cross-feeding interactions [29]. Furthermore, as already mentioned above, the use of isotope substrates like amino acids and their measureable incorporation into proteins allows inference of active metabolic pathways and thus gains insights on key metabolic constituents. As these approaches are limited by the culturability of the sample

and the complexity of interactions, they are often complemented by computational culture-independent approaches, such as genome-scale metabolic models. These are then used to predict metabolic behavior at comparably low cost compared to complex experimental studies. However, eventually the verification of the generated hypotheses has to be carried out *in vitro* to really reach cause-effect conclusions.

2.3. Graph Theory

This paragraph describes the formal structure of a graph which is used in this thesis. It is based upon the definitions described in Diestel's book of graph theory [30].

A graph itself is basically a data structure, containing relational information. A graph G(V, E) consists of two distinct sets, the edge set E and the node set V, such that $E \subseteq [V]^2$. In a metabolic graph retrieved from the dynamic Flux-Balance-Analysis simulation of a genome-scale metabolic model of microbial communities in *µbialSim*, compounds and microorganisms form the node set V, and their relationship defined by coupled reactions form the edge set E. The temporal explicit direction of the flow of a reaction makes the corresponding graph a directed graph. As the node set consists of two distinct node types of compounds and microorganisms, and nodes of the same type cannot be connected to form an edge, the graph is called bipartite. The following section will therefore concentrate on the specific case of a bipartite directed graph.

Let G(V, E) be a bipartite graph, that contains the vertices V and edges E, then all vertices of G can be described as V(G) and all edges of G as E(G). The number of vertices V(G) is called the *order* of the graph G, the number of edges E(G) as the *size* of it. Within G, a particular edge e is described as an ordered pair of vertices. An edge with the starting vertex a and the end vertex b will therefore be denoted as $e = (a,b) \in E$. The order of a and b holds the information of the direction of the edge, so that $(a,b) \neq (b,a)$. The microbial networks used in this thesis consist out of two distinct vertex types, microorganisms M and compounds/metabolites C. M and C are subsets of V and their union represents V in its entirety. Therefore, G can also be written as G(M, C, E). Within any of these two sets, there is no edge in between its members, so that start vertex a and end vertex b of a random edge e cannot belong to the same vertex type. So if $e=(a,b) \in E(G)$ and e.g. $a \in M$, $b \in C$ has to be true, therefore G is *bipartite*. The number of incoming edges of a vertex v is called *indegree*, denoted as $deg^+(v)$, and the number of outgoing edges is

called *outdegree*, denoted as $deg^-(v)$. The *degree* of an vertex v, *deg(v)*, is the sum of outdegree and indegree.

A *path* within a graph G is a subgraph of G, described by a sequence of nodes and edges, which are connected in the graph, so that metaphorically speaking one could start at the first vertex and walk along the edge to the next vertex in the sequence until reaching the end vertex. Every vertex within this sequence is unique. The *length of a path* is the number of contained edges. If the length of a path is greater than two and the path is extended by an edge from its last vertex to its first one, then this graph is called a *cycle*. A *distance* d(a, b) of two nodes a and b of a graph G, is the shortest path in G with a as its starting vertex and b as its end vertex. The *diameter* of G is the longest distance between any pair of nodes within G, thus the longest of all shortest paths between any two vertices.

If there are additional attributes associated with the graph itself, or its elements E or V, it is also called a *weighted graph* or a *network*. Because the metabolic graph obtained by a simulation in *µbialSim* contains vertex attributes like compound concentration and biomass concentration, as well as reaction flux volumes and graph associated meta data, rather than simply carrying the structural information of the underlying metabolic graph, it is often referred to as a metabolic network. The term *dynamic network* is used, when referring to a sequence of networks, which can be exposed to structural and attribute related changes. Studies of metabolic interactions usually contain time series data, either obtained by sampling during experiments or through simulation at discrete time points within a certain time window. Each of these temporal data points then reflects a static metabolic network, and the entirety of all captured static networks form the dynamic metabolic network. In this text, the figurative term *slice* is often used to refer to one specific static network extracted from the time series data of the simulation at a specific time point.

Furthermore it should be noted, that in this thesis terms derived from *in vivo* biological studies and experiments like "*cultivation*", "*microbial growth*", "*media*" etc. are used to describe the artificially simulated counterparts of these terms. As this work solely relies on simulated data, paraphrasing these terms is often omitted for better readability. So eg. if "*cultivation*" is mentioned, it actually refers to the simulated cultivation, and "*media*" just describes the virtual

composition of compounds, etc. When referring to actual biological systems, this is explicitly mentioned.

2.4. Visualization

The visualization of graphs can serve as a complement to common graph analysis. The basis of visual graph analysis is the interplay of the visual graph representation itself, classical graph analysis using algorithms and exploiting graph theoretical properties, and user interaction [31]. The use of visualization can amplify cognition and help to understand complex data structures [32]. The process of visual data exploration is very interactive and integrated. Figure 1 shows a schematic workflow of an analytical process involving visual data analysis described by Keim et. al. [33],. The generation of "Data", in this case Multi-Omics experimental data available in various databases, led to the construction of sophisticated genome-scale metabolic models. By using these models in simulations, new data is created. The "Visual Data Exploration" of simulation results is then used to understand model dynamics or simulation parameters. This feedback cycle between "Visual Data Exploration" and "Model/Simulation" ultimately yields some "Knowledge", which in turn can be validated in experiments and thus closing the feedback cycle back to "Data".

To explore metabolic networks, the node-link diagram of a graph is the most used visual representation. It makes it easy to follow metabolic paths through the network and is rather intuitive [31]. However, node-link representations suffer from crossing edges as networks get really big. This makes size reduction an important step in creating representations. As metabolic networks are no bare graphs containing simply relational information, the visual representation of these attributes associated with nodes, edges, or the graph itself is very important in order to create meaningful visualizations.

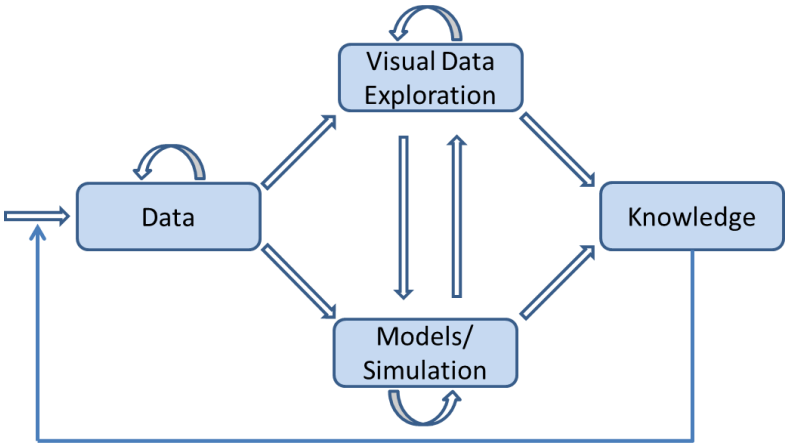

Figure 1: Process of data analysis
Visual analytics is integrated into an analytical process in order to gain knowledge, reproduced from Keim et al. [16].

Figure 2 provides an example of a stepwise approach to integrate these attributes into the visual representation in a way that supports the cognitive perception of their meaning. The depicted graph is a bipartite directed acyclic graph. Its nodes are therefore associated with a binary attribute, which indicates to which partition they belong two. Furthermore each edge holds a weight attribute, which is a numeric value. The first representation in Panel A simply shows a node-line diagram, representing the relationship between the entities of the graph. It does not include any further attribute information. In Panel B, the node attribute was mapped to the respective node color, whereas the edge weight value now correlates with the displayed edge width. The incorporation of the associated attributes into the visual representation increased the information content of the depiction, without cluttering the network with details. In Panel C, the layout of the nodes position was changed to a hierarchical order. Here the directed acyclic property of the graph is exploited, as directed acyclic graphs are directed graphs, which have at least one topological order. This makes for a quick and consistent hierarchic representation of the network by applying hierarchic layout algorithms. The Sugiyama approach is still the basis of many of these algorithms, creating hierarchical graph drawings of directed graphs, compromised of four main steps [34]. First, cycles are removed from the directed graph by deleting edges, creating a directed acyclic subgraph. If no additional constraint is applied at this step, the resulting subgraph is a random subgraph of the original directed graph. This step allows of a unidirectional drawing of the graph. In the second step, vertices are assigned to layers, which will later reflect their vertical positioning in the drawing. The subsequent step is finding an adequate order of the vertices within each layer. The main constraint of this step is usually to

avoid edge crossings in the final drawing. The final step then calculates horizontal and vertical coordinates of the vertices, which are the used for the final drawing. The result of this process in Panel C of Figure 2 highlights two interesting features of the underlying graph. One, the hierarchic representation emphasizes the flow direction within the network, and secondly, the bipartite nature of the graph. The latter is already captured by the node colors, but complemented with the chosen layout its meaning becomes more clear.

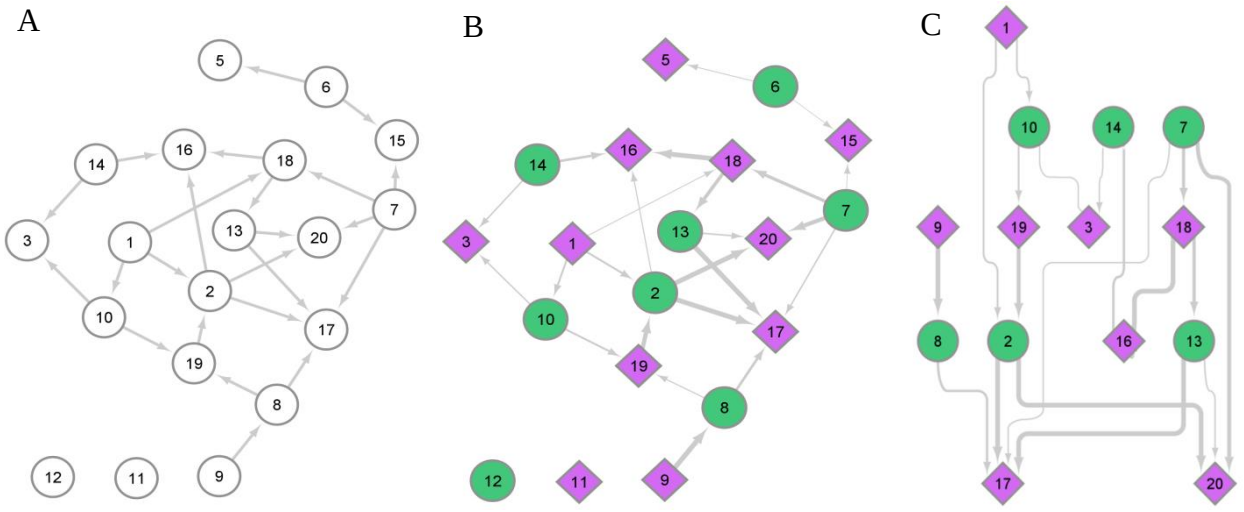

Figure 2: Visualization process of a random bipartite directed acyclic graph
A: Initial visualization, adapted organic layout, contains just relational information, no additional attributes
B: Incorporates edge weights and node type into the visualization of the graph
C: Edges with low edge weights and non-connected nodes are deleted; hierarchical layout applied

Metabolic networks have been visualized in many works and there are several tools available like MetDraw[35], ModelExplorer [36], FAME [37] or Escher [38]. Except for MetDraw, which creates static network views, these tools can create interactive metabolic maps of genome-scale metabolic models, and visualize a current metabolic state by including flux values of reactions, metabolite concentrations, etc. However these applications are tailored towards the use of single species models, and thus representations of the metabolic models are bipartite graphs of reactions and metabolites. Though some of these tools support compartmentalizing of the networks, to account for cell compartments and cell membrane, visualizations of microbial communities at a resolution down to the level of reactions would get very big and messy. VisANT 5.0 [39] implemented a specific layout in order to visualize multi-species networks, with expandable *metanodes*, each containing a metabolic model of one species, arranged around a center containing media compounds, which are transported in or out of the models. This approach

highlights inter species metabolic exchange and is better suited to visually explore cross-feeding of microbial networks. It furthermore incorporates dynamic flux visualization, allowing to the user to slide through time points of the time series while dynamic network properties are changing accordingly, to reflect the current temporal state.

The visualization of dynamic networks, with changing structure and attribute values of the underlying static temporal networks, poses an additional challenge to the visual analysis. Besides the obvious increase in data volume and computational cost, methods have to be implemented which can adequately reflect the changes of the dynamic attributes of the underlying graph to the viewer. The use of static network representation in animations is often used in this case, as for instance in this visualization of time series metabolomics data [40], but is often not sufficient as a stand-alone solution, as it rather conveys bigger and more structural changes due to the human perception [31]. For detailed analysis static representations are best suited, but therefore require condensing the temporal information of the dynamic network. This could be done by extracting temporal network slices at meaningful time points or the incorporation of time series data into the static representation, increasing the complexity of the visualization.

2.5. Scope

The main goal of this thesis is to provide a visual representation and characterization of the metabolic cross-feeding networks of the SIHUMIx microbial community [19] in a batch and chemostat process, simulated by *µbialSim* [18]. The visualization should be part of a visual data exploration of the dynamic changes within the metabolic network during the evolution of the simulated communities. By linearizing the metabolic networks in a way to preserve their cumulative flux, a hierarchical representation should emphasize the "trickle-down effect" of nutrients through the network, making it easy to follow along their metabolic pathways through the network. Complemented by an overall characterization of the simulated trajectories, metabolic cross-feeding interactions of microorganisms on intermediate compounds should be identified, shining light on important metabolic interdependencies and key organisms within the underlying metabolic model.

3. Methods

In order to assess the value of hierarchical visual representations of metabolic cross-feeding networks embedded into an overall visual graph analysis, two simulated datasets have been analyzed as described by the following methods.

3.1. Simulated Data

The analyzed simulated data is based on the SIHUMIx model [19]. The genome-scale metabolic data of this extended simplified human intestinal microbiota of eight species is based on the AGORA model, which holds a collection of 773 human gut species [41].

3.2. Data processing in R

As the data processing and visualization is programmed in R, simulation data from *µbialSim*, which is written in Matlab, has to be transferred to R. The data is split in three files, two containing static information on the future nodes and edges of the metabolic network, and one containing the dynamic attribute values (Table 1). These are read into three distinct data.frames in R. To specifically refer to these internal data structures in R, they are subsequently denoted as *nodes*, *edges*, and *trajectory*.

Table 1: Input files and corresponding R internal data structures for visual analysis in R and Cytoscape
Files exported from simulation data of *µbialSim*

Filename	Content and headers of files	R data.frame
nodes.txt	Attributes of microorganisms and compounds "id", "type", "name", "value"	*nodes*
edges.txt	Attributes of reaction (*relationship between nodes*) "from", "to", "reacID", "sense", flux"	*edges*
trajectory.txt	Dynamic attribute values (Compound and Biomass Concentration and reaction flux)	*trajectory*

Two main R packages were used in the analysis. The *igraph* package was used to calculate statistical properties of the graph, quick visualizations and exploratory analysis [42]. The *RCy3* package, which is also part of Bioconductor, was used to use *Cytoscape* functions from within R and conveniently also compatible to the graph class of *igraph* [43]. The network visualization shown in this thesis were generated with *Cytoscape* [44].

The information on reactions imported in the *edges* data.frame has to be altered, in order that both flux directions of a reaction are reflected as a distinct edge in the data structure, rather than one, with correct entries in *edges$from* and *edges$to*. This is due to the internal handling of reactions in *µbialSim*, where each reaction has a sense value s, indicating the directionality of the flux depending on the simulated flux value. If s = 1, a positive flux value indicates excretion of the metabolite and a negative flux value corresponds to an uptake of this metabolite into the cell of the respective organism. If s = -1, a positive flux value means metabolite uptake and a negative one metabolite excretion. By considering all reaction flux values stored in the *trajectory* data.frame, *edges* is adapted to contain all states of reactions (either excretion or uptake) which occur at least once during the entire time course. The new entries in *edges* due to the data transformation enforce creating new edge ids. To better identify the pair of edges describing two different directionalities of one reaction, the edge ids describing the uptake direction of a reaction are inferred from the edge ids of the excretion direction by appending a suffix "_c". Thus for instance, an arbitrary edge with the *edges$reacID* = "m_1_411" indicates an excretion of a metabolite, whereas the uptake of the same metabolite would be described by the edge with *edges$reacID* = "m_1_411_c". With every new entry in *edges*, the *trajectory* data.frame has to be adapted accordingly in order to still work as a lookup table for the dynamic flux values over the entire time course.

3.3. Obtaining static network slices of the dynamic network

In order to create static network representations, one has to extract temporal network slices from the time series data of the entire dynamic network. In theory, this could be done for any time points, thus creating a rather continuous series of networks, yet this comes at increased computational costs scaling with the size of the network and number of time points and does not necessarily yield benefits over selected network slices. To go into details of the metabolic interactions, one inevitably would have to choose time points anyway. This selection is customized, by considering the trajectories of biomass concentration and compound concentration, and defining time points at which assessment of the current metabolic interactions seems valuable. These can be phases of stable growth to investigate the characteristics of the metabolic networks during such periods, transition phases between two stable phases, or time windows, where there is a lot of dynamic change in the network and the examination of the underlying metabolic interactions could reveal mechanisms driving these changes. With the set

of all chosen time points, the dynamic attribute values at these time points are extracted from *trajectory*, yielding static graphs for each time point, referred to as network slices. These are then linearized, to obtain directed acyclic graphs.

3.4. Network Linearization

The idea behind the linearization of the network slices is to create directed acyclic graphs, which are convenient to display in hierarchical graph representation with common layout algorithms. These representations show a unidirectional flux direction through the network, making it easy to read. It also emphasizes the "trickle down" aspect of nutrients being poured into a metabolic network, which metabolizes the input further and further, profiting on its nutritional value, until the nutrients value is exploited and leaves the network .The layered layout consists of altering layers of microorganisms and compounds, which is due to the bipartite nature of the graph, therefore sometimes referred to as *microbial layer* or *compound layer*. This highly structured depiction of the graph is pleasing for the eye and allows quick orientation and intuitive understanding. Furthermore, by anticipating the linearization of the network slices, which is usually done by the hierarchical layout algorithms before visualizing by choosing a rather arbitrary directed acyclic subgraph of the input graph, one gains control to determine, which edges should be deleted in order to yield an directed acyclic subgraph of the original network slice.

Technically the objective goal of the linearization process is to preserve as much of the cumulative flux of the original network slice as possible (Table 2). This is obtained by ordering the edges of the original network slice by their flux value in decreasing order, and adding them to an empty graph. After each expansion, the acyclic property of the graph is reviewed. If the graph lost this property, the recently added edge is removed again, and the next edge in order expands the graph further. The algorithm returns the obtained directed acyclic graph, after looping through all ordered edges. It will also return edge ids of the removed edges, in order to identify them for analysis purposes.

3.5. Network Visualization

After loading the networks into Cytoscape via the *RCy3* API, their visual style is adapted according to its need, like mapping the flux values to the edge width, coloring nodes according to

their type etc. Depending on the desired representation, filters are applied to reduce the network in size. The hierarchical layouts used are the default one called “hierarchical layout”, and the “yfiles hierarchic layout” from the Cytoscape App “yFiles Layout Algorithms”. To map node positions of different networks, the implemented function “Copycat Layout” is used.

Table 2: Linearizing function for network slices

```
Linearize (slice S){
      orderedEdges = order(edges of S, by flux value, decreasing order)
      G = empty graph
      FOR (edge   in   orderedEdges):
            G = G + edge
            IF (isDirectedAcyclicGraph(G) != TRUE):
                  G = G - edge
      RETURN(G)
}
```

4. Results

In order to show the relevance of the implemented methods to visualize and analyze microbial networks, these methods were applied to two exemplary trajectories, simulated with *µbialSim* [18]. It is embedded in an overall characterization and evaluation of these microbial networks. The underlying microbial community model (meaning genome-scale metabolic model) for both simulations is identical and based on the SIHUMIx microbial community, consisting of eight microorganisms [19]. The difference is in the simulation parameters chosen within the simulator *µbialSim*, by setting different medium compositions and reactor process types.

4.1. Code

In order to enable further inspection of the created graphs and networks to the interested reader, https://github.com/Tille12/Masterarbeit holds a R-script and both datasets already preprocessed, so that the user can reconstruct part of the analysis this thesis was based on. For further instructions on how to use the script, please consider the "README.txt" file located in the repository behind the link. The repository further includes a PDF file of this manuscript and supplementary material including figures.

4.2. Initial Conditions and Simulation parameters

Table 3: Simulation parameters and initial conditions for both batch and chemostat simulation in *µbialSim*

	Batch – Dataset 1	**Chemostat- Dataset 2**
Specific biomass concentration* [gDW/L]	0.1	0.01
Compound concentration [mM]**	0.01	0.1
Biomass inflow concentration[gDW/L]	-	0
Compound inflow concentration [mM]**	-	0.1
Flow rate reactor [L/h]	-	0.01
Volume reactor [L]	1	0.25
Number of species	8	8
Number of compounds	139	139
Cultivation time [h]	1	400.1
Elapsed runtime µbialSim [h]	0.22	15.88

* biomass concentration was identical for all eight species
** compound concentration was identical for every compound of the 139 compounds in total

The main process related distinction between both trajectories is the reactor type, as Dataset 1 was obtained by simulating a batch process, whereas Dataset 2 by simulating a continuous fed reactor, with influent and effluent present at identical flow rates. This leads to different behavior, especially towards the end of the trajectories, as the metabolic activity in a batch system eventually and inevitably is going to face substrate depletion, while in a flow reactor the influent provides new substrate, in this case at a constant inflow rate of 1.04 $x10^{-3}$mmol/h for each compound (Table 3). Both communities consist of the same composition of microorganisms, starting at specific initial concentrations of 0.1gDW/L and 0.01gDW/L respectively. The influent of the chemostat system does not contain biomass, so there is no further external biomass source for both systems. The accumulated compound specific influent deposition into the *in silico* reactor over the entire time course is 0.42mmol. The simulated cultivation time is 1 h for the batch simulation and 400.1h for the chemostat simulation, with runtimes of µbialSim at 0.22h and 15.88h.

The next two sections 4.3 and 4.4 describe the basic characteristics of both trajectories, to provide a sound understanding of the simulated data. In some of the following figures, there are vertical lines displayed. These are always at the same time points on the timeline for each dataset, and refer to the slicing points of the respective networks, where temporal snapshots of the entire networks were taken and used for further analysis. In order to serve as visual anchor points over various graphs and to easily and consistently refer to these slices, these lines are included even in earlier plots where the entire temporal data is shown. In these cases, there might be no extensive reference or explanation to this graphical feature, as the intention there is to improve readability and informative value in combination with other plots.

4.3. Batch simulation

4.3.1. Characterization of trajectory

In Figure 3, the biomass concentration trajectories for all microorganisms of the SHIUMIx community are displayed. Starting at the same initial specific concentrations of 0.1gDW/L, the average end concentration after one hour of simulated cultivation time is 0.102gDW/L, with *Escherichia coli* as the most abundant species gaining a total of 0.06gDW/L biomass over the entire process. In comparison, *Bifidobacterium longum* as the least abundant organism built up 0.004gDW/L of biomass, which is roughly one fifteenth of the absolute growth of *E. coli*. In

general, *E. coli* sticks out and accounts for almost 37% of the total biomass accumulation, and among the rest, there is a clear but quite evenly spaced hierarchy. Looking at the growth rates over time (Figure 3), there is a sharp decline starting at around 0.05h into the cultivation, headed by *Clostridium butyricum* and shortly after followed by the rest of the species, except for *E. coli*. The stark difference in shape of the declining curve of the latter, going down slower and steadier, contributes significantly to the dominance in biomass gain, as after a similar head start of *C. butyricum* and *E. coli*, after approx. 0.2 hours, the latter is the only one still growing and comes close to a standstill only 0.55 hours later. After substrate depletion, no decline in biomass is observed for any species, which is due to the implementation of *µbialSim*, which does not account for starving populations. On the contrary, long periods of non-growth in a batch process can likely be interpreted as starving phases, translated to in vivo conditions, as substrate depletion ends further proliferation of the cells, so that the existing ones die off eventually, leading to a decay in cell count over time.

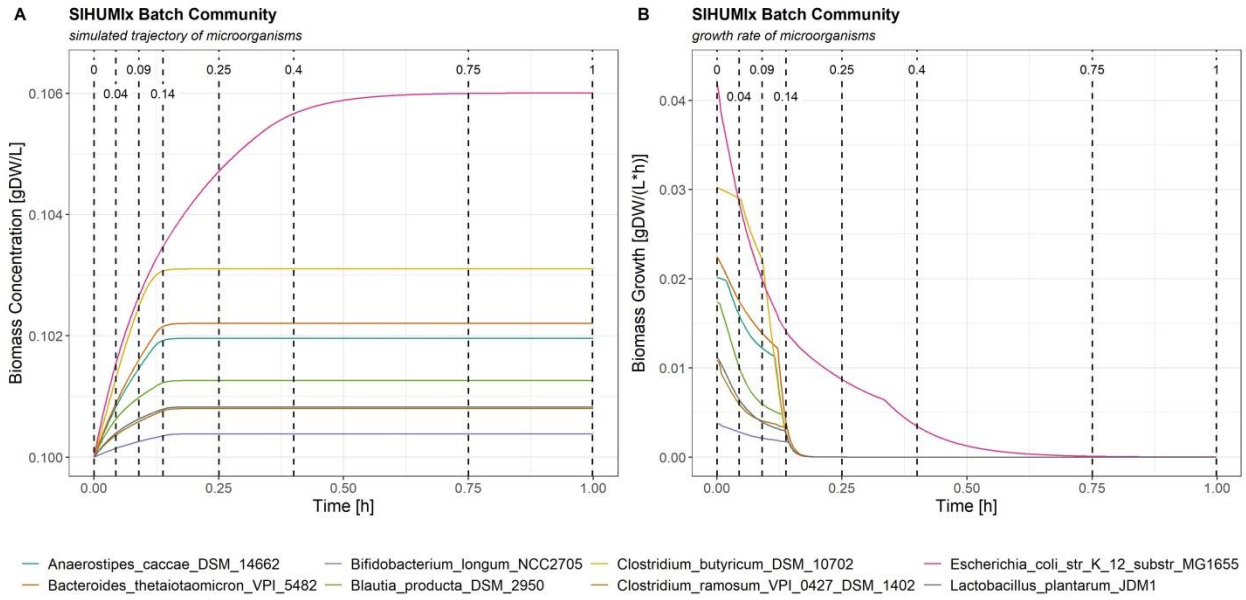

Figure 3: Trajectory of the SIHUMIx community in a simulated batch process
Biomass concentration (A) and biomass growth rate (B) over the entire simulation time. Vertical dashed lines refer to network slice time points on the batch dataset

Regarding the compounds of the virtual reactor Figure 4 gives an overview on the development of concentration (Panel A) change rate (Panel B) of a selection of 14/139 compounds. These compounds were among the most abundant and/or most fluctuating ones. Looking at the concentration curves in Panel A, proton stands out as the most abundant with an end

concentration of 0.4mmol/L, followed by acetate and formate at 0.26 and 0.24mmol/L respectively. These three constituents form 48 percent of the total composition of compounds at that time. The biggest amount of it is produced until 0.5 hours into the cultivation, with high production rates quickly falling within the first 0.14 hours, following a phase with very little change in production rates until approx. 0.35 hours and then finally falling down to insignificant low numbers after 0.5 hours. These three phases of high activity up to 0.14 hours, lower activity up to approx. 0.5 hours and the last phase with almost none, is also reflected in Figure 3, with all microorganisms growing in phase one, E. coli still striving in phase two and then also coming to a halt in phase 3.

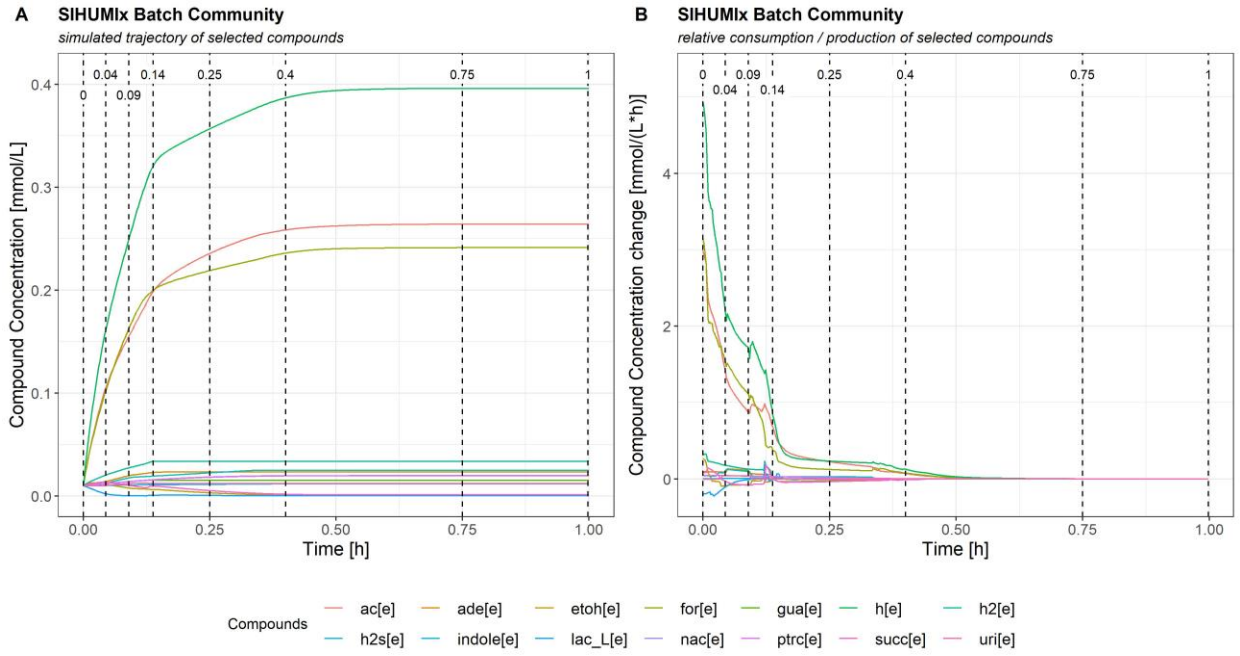

Figure 4: Trajectory of selected compounds in a simulated batch process
Concentration of 14 (out of 139) compounds (A) and corresponding temporal derivations (B). Compounds were selected by highest abundance and highest fluctuations summed up over the entire simulation time (excluding water for clarity), and are displayed with the following abbreviations: ac[e]: acetate, for[e]: formate, h2[e]: hydrogen, lac_L[e]: (S)-lactate, succ[e]: succinate, ade[e]:adenine, gua[e]: guanine, h2s[e]: hydrogen sulfide, nac[e]: nicotinate, ptrc[e]: putrescine, indole[e]: indole, uri[e]: uridine, h[e]: proton

4.3.2. Characterization of complete network

Entire networks of a model, including all possible cross-feeding interactions which can occur over the entire time course can become quite complex and hard to interpret. In

Figure 5, the network of the simulated cross-feeding SIHUMIx community in a batch reactor is displayed, showing all reactions, microorganisms and compounds, which at least once were present during the entire cultivation. The order of the network is 147, containing 8 microorganisms and 139 compounds. The underlying genome-scale metabolic model contains 684 exchange reactions, 575 of them at least once changing direction over the entire cultivation, reflected by additional reverse edges in the graph, which in total add up to 1259. The force directed layout algorithm behind the organic layout leads to dense hairball structures with highly connected nodes in the center. Arranged around this condensed structure, there are four groups of nodes, the top one containing *Bacteroides thetaiotaomicron*. The visual impression that this organism is kind of isolated from the others by looking at the depiction, is true in the sense that all eleven compounds in proximate distance to *B. thetaiotaomicron* are not connected to any other microorganism. The other three spatially separated groups of compounds right, left and below the center reveal that each is only connected to a specific pair of microorganisms. The depiction of the network already hints at these kinds of conclusions, but without further evaluation by manually zooming in and checking the edges and neighborhoods of nodes it is just vague and not implicitly clear from the chosen layout. The mere abundance of nodes and edges, accompanied by a lack of structure in the layout, make it really difficult for the viewer to directly extract information, without engaging, and shows the need for further refining the visual representation of the network.

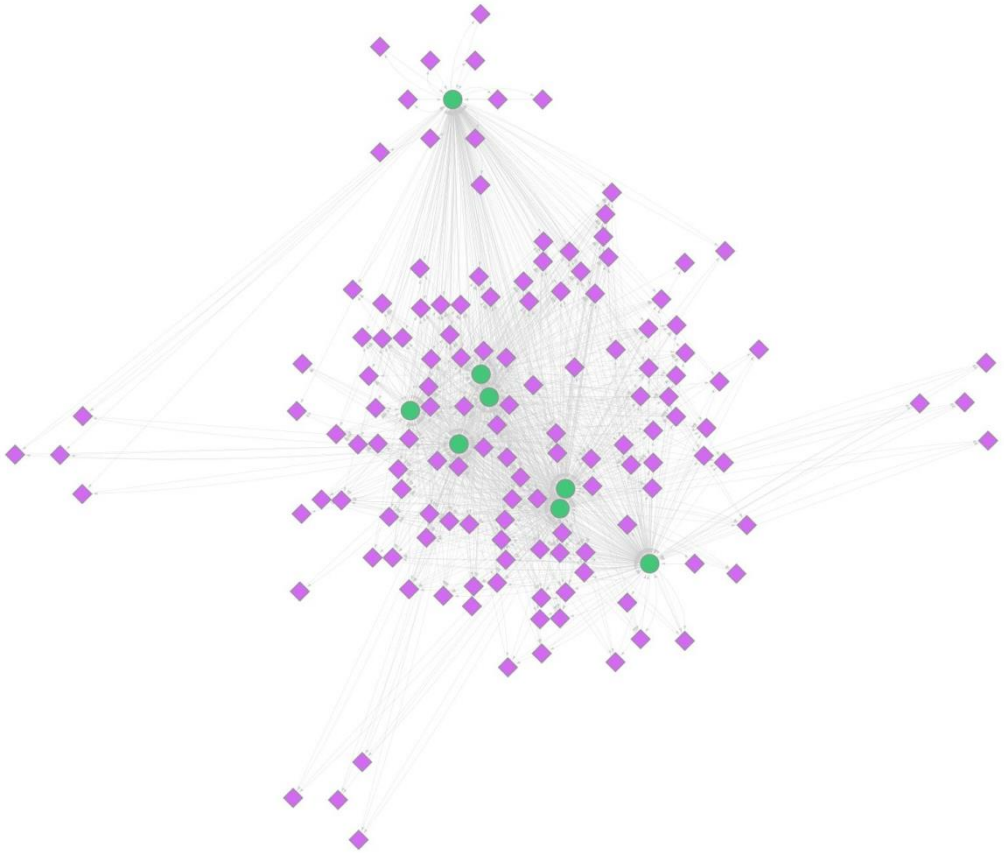

Figure 5: Entire network of the batch and chemostat simulation
Nodes represent microorganisms (green circles) and compounds (purple diamonds), and the connecting edges refer to their production and uptake reactions. This network integrates all edges and nodes, which are present in the simulated processes at any time point. The layout was generated in Cytoscape with the yFiles Organic Layout [44]

4.3.3. Pre-Linearization

Time points finding of network slices

As already described in the section above, the batch process can roughly be divided into three phases. Starting at t = 0h until approx. 0.14h, is a rapid growing phase. Continuing from there until roughly 0.5h into the cultivation, one can observe a moderate phase, with a subsequent phase of very low microbial growth until the end (t = 1h) of the process. In order to reflect typical network structures within these phases, time points were set in a way, so that every phase is covered at least ones. For instance, selected time points at 0h, 0.04h and 0.09h into the cultivation stand for phase one, time points at 0.14h right at the transition between two phases (see rapidly falling growth rates in Figure 3 and Figure 4). Time point 0.25h is chosen to reside

in a very stable part of the second phase, whereas at 0.4h compound production is finally declining towards zero and transitions into the last period, characterized by t = 0.75h and the end point of simulation at 1h.

Table 4: Graph parameters of the network slices of the simulated batch cultivation
Slices contain only connected nodes

Time [h]	**0**	**0.04**	**0.09**	**0.14**	**0.25**	**0.4**	**0.75**	**0.998**
Order	136	136	136	135	118	73	60	59
Size	546	557	561	500	426	72	59	58
Cumulative Flux [mM/h]	199.2	115.4	81.0	34.1	9.4	3.7	0.06	0.006
(thereof excretion)	*131.6*	*69.4*	*50.2*	*22.8*	*6.2*	*2.8*	*0.04*	*0.002*
Diameter	6	6	6	6	5	2	2	2

Characterization of pre-linearized network slices

Until the end of phase one, 0.14h into the cultivation, the basic constitution of the exchange network does not change considerably, the number of connected nodes stays almost the same (see Table 4). An decrease of 61 nodes from t = 0.09h to t = 0.14h hints to structural change in the network due to rapidly falling growth rates of the microorganisms, except for *E.coli*. At t = 0.25h this continuing trend leads to the first significant drop of the numbers of nodes in the network, although at this point all eight microorganisms are still present, which is not the case anymore at the next slice. Onwards from 0.4h into the simulation, the network enters a state, with E.coli as the single remaining growing organism, depleting the residues of the initial compounds still present in the reactor. Although flux rates continue to fall to very low levels, the structure of the network does not change much anymore until the end.

4.3.4. Linearization

Linearization of the extracted network slices of the entire dynamic network partly led to smaller linearized networks regarding network size, but no change regarding the order of a slice (see Table 5). As expected there was no structural change at the last three time points (for $t \geq 0.4$h), when growth in the microbial community is already reduced to a single organism, and thus no cycles should be present in the original slice. The set of all unique removed edges throughout the linearization process includes 59 edges, compared to a cumulative count of 115 removed edges.

28 of 59 are just removed once, whereas 5 of 59 edges are repeatedly removed at the first five time points. The diameter of the graph was reduced to four for the first three slices, and to five at t = 0.25h.

Table 5: Graph parameters of the linearized network slices of the simulated batch cultivation
Linearization of the network slices characterized in Table 4; non-connected nodes were omitted

Time [h]	**0**	**0.04**	**0.09**	**0.14**	**0.25**	**0.4**	**0.75**	**0.998**
Order	136	136	136	135	118	73	60	59
Size	522	528	534	477	414	72	59	58
Removed edges	24	29	27	23	12	0	0	0
Cum. Flux conservation[%] *	98.3	94.6	95.2	98.5	99.9	100	100	100
Diameter **	4 (6)	4 (6)	4 (6)	6 (6)	5 (6)	2 (2)	2 (2)	2 (2)

* cumulative flux in relation to pre-linearization cumulative flux
** value in brackets corresponds to diameter of the non-linearized slice at that time point

In line with the evaluation of characteristic graph parameters before and after linearization, the visual representation does not yield a big difference between full network slices and their linearized counterparts. At early stages of the simulation, the temporal network snapshots are very big in size and order, and therefore the relatively small number of omitted edges is not immediately obvious to the observer without further inspection. By applying a hierarchical layout on the linearized network and using the nodes' positions for the layout of the non-linearized network, the user can highlight the difference between the two. This is shown in Figure 6 for t = 0.04h into the cultivation. This representation still contains 93 nodes and 356 edges, and therefore is hard to inspect, especially presented in a static figure like in a manuscript, and therefore rather serves to point out bigger and more general structural features. To get into details, the network views of Cytoscape allow the user to zoom in, select, and edit the graph further as required. However, the topological sorted layers of the directed acyclic graph (Figure 6 without red edges) allow the user to still perceive a certain notion about the current network dynamics.

Figure 7 shows the linearized network slice at t = 0.09h. Edges with a flux smaller than 0.25mM/h are not displayed, as well as the compounds water and proton, and all non-connected nodes due to this simplifying process. The outcome still contains all eight microorganisms, 27 compounds and 53 edges. Most compounds are either consumed or produced in this reduced

depiction of the graph, just three compounds, ethanol (etoh[e]), succinate (succ[e]), and acetate (ac[e]) are neither just sources nor sinks. Ethanol is the only compound being visibly consumed by *B. producta* at this stage, which is produced by *B. longum* and *C. ramosum*. Looking at the data, the uptake of ethanol of *B. producta* at a rate of 0.43mM/h makes up 21% of all compounds consumed at that time, with fumarate following at 0.24mM/h and all other uptake rates smaller than 0.08mM/h. It then produces succinate, which is then consumed by two other species. In general, striving species at that point show a high connectivity, such as *E. coli* and *C. butyricum*, whereas *B. longum*, *C. ramosum* and *L. plantarum* show no uptake reaction with a flux higher than 0.25mM/h.

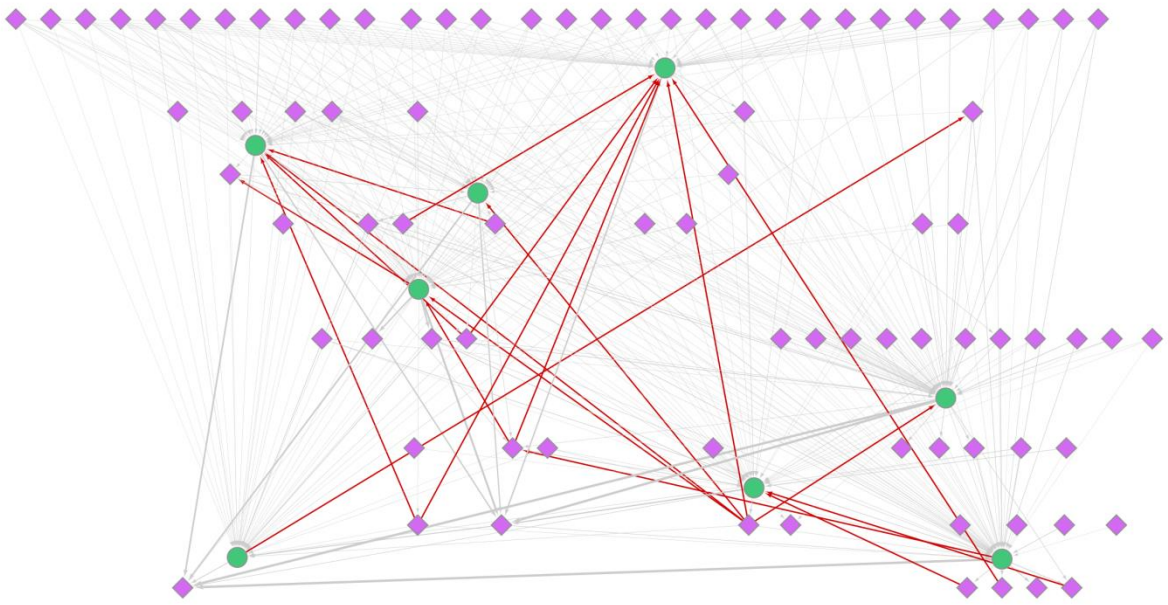

Figure 6: Network Slice at t=0.04h of batch cultivation with hierarchical layout
This graph shows a non-linearized representation of the batch process at t = 0.04h. Green circles represent organisms, purple diamonds compounds. Connecting edges represent the exchange reaction between the two. The layout was obtained, by first applying a hierarchical layout to the linearized version of this slice. This creates a layered topological order, with source compounds in the first layer on top, and sink compounds in the last layer at the bottom. By re-adding the edges which were deleted in the course of the linearizing step (marked in red), a depiction of the original graph is obtained, which clearly shows the difference. All of the red edges are back-edges, directed upwards, compared to the grey ones of the linearized graph, pointing downwards.

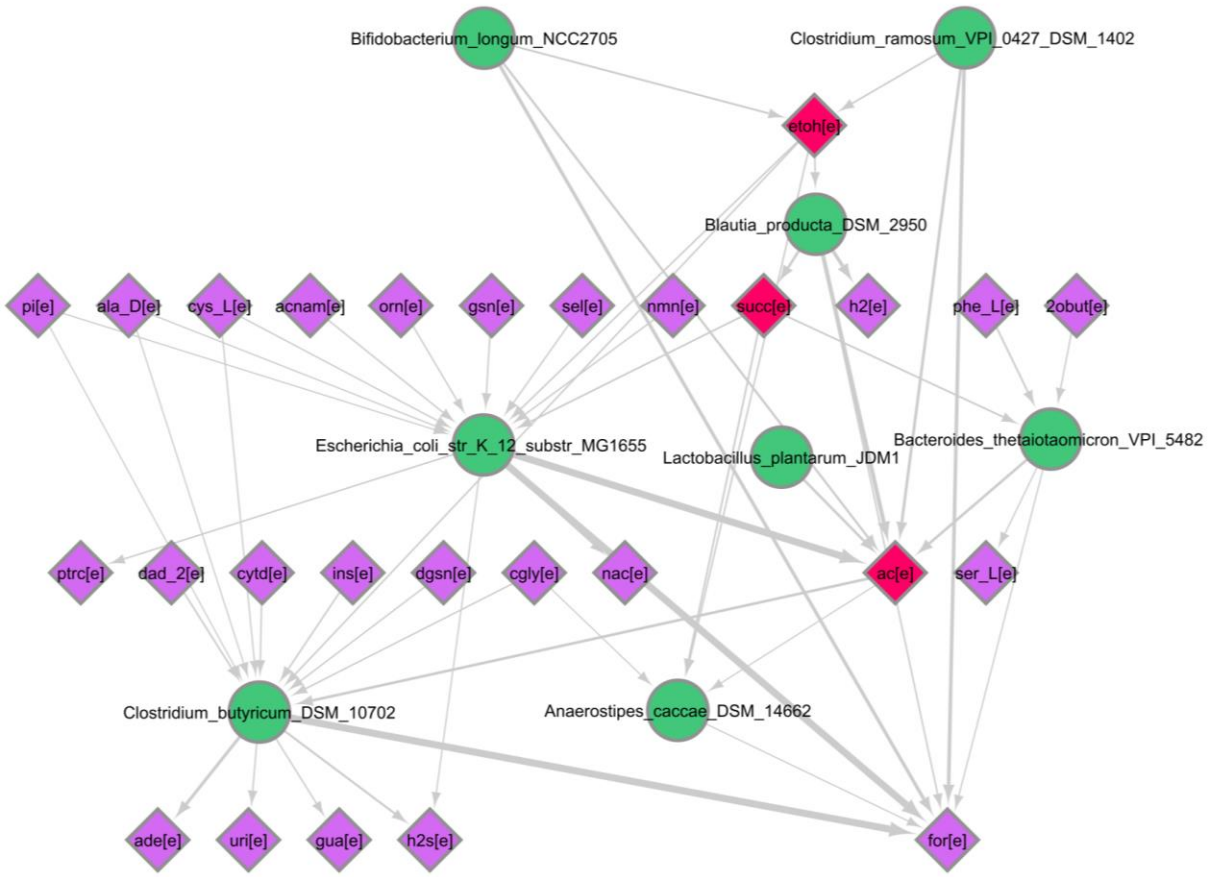

Figure 7: Network slice of batch simulation at t = 0.09 hours

This graph is a subgraph of the entire slice at t=0.09h, obtained by filtering out edges with a flux smaller than 0.25mM/h and omitting water and proton compound nodes for clarity. Afterwards the graph was stripped of all non-connected nodes. The green round nodes represent microorganisms, the purple diamond shaped ones compounds. Edges stand for excretion and uptake reaction of the organisms. The width of the edges is mapped to the flux through the related reaction at that moment, the direction indicates. Red marked nodes are compounds, with $deg^-(v) \neq 0$ and $deg^+(v) \neq 0$

4.4. Chemostat simulation

4.4.1. Characterization of trajectory

Microorganisms

Two graphs describe the microbial trajectory: Figure 8 depicts the overall time span of the cultivation and Figure 9 represents the microbial growth rates within the first 30h, to highlight the strong fluctuations during that time. Specific biomass concentration at the beginning of the chemostat cultivation is 0.01gDW/L for every species. Right from the start, all organisms are growing, until approx. 2 hours into the cultivation. The composition of the consortium at that time is dominated by *Clostridium butyricum* at 0.041gDW/L, while all other species

concentrations are between 0.012gDW/L and 0.027gDW/L. Except for *Escherichia coli* and *Clostridium ramosum*, these are maximum specific reactor concentrations regarding the entire simulation, as all these species are being continuously washed out afterwards. The total biomass gain within the reactor up to this point is 0.087gDW/L compared to 0.08gDW/L at the start, so overall biomass concentration doubles within the first two hours. Looking at Figure 9 it becomes apparent, that after the first two hours described above, growth rates go down rapidly, and microbial concentration of *E. coli* and *C. ramosum* separate from the rest of the species. Towards the end of the cultivation, biomass growth approaches a steady state with constant specific growth rates µ. For *Bifidobacterium longum* and *Lactobacillus plantarum* it is between 0.032/d and 0.035/d, the growth rate of *Anaerostipes caccae* and *Bacteroides thetaiotaomicron* lies between 0.026/d and 0.029/d, and *Blautia producta*, *C. butyricum*, *C. ramosum* and *E. coli* reside around 0.041/d. This is sufficient for *E.coli* and *C. ramosum* to still accumulate within the reactor at a very low rate, basically maintaining their concentration, but still leads to decaying concentration for the rest of the species, as they are steadily washed out with the effluent of the reactor. In general, it should be noted that jagged lines, especially at higher resolution of the time course (e.g. see Figure 10 Panel D), are the result of the implemented numeric method in µbialSim [18] which dynamically defines the size of the time steps, at which composition and activity of the simulated model are evaluated. One could reduce the size of a step to obtain smoother results, which then would prolong the runtime of the simulation.

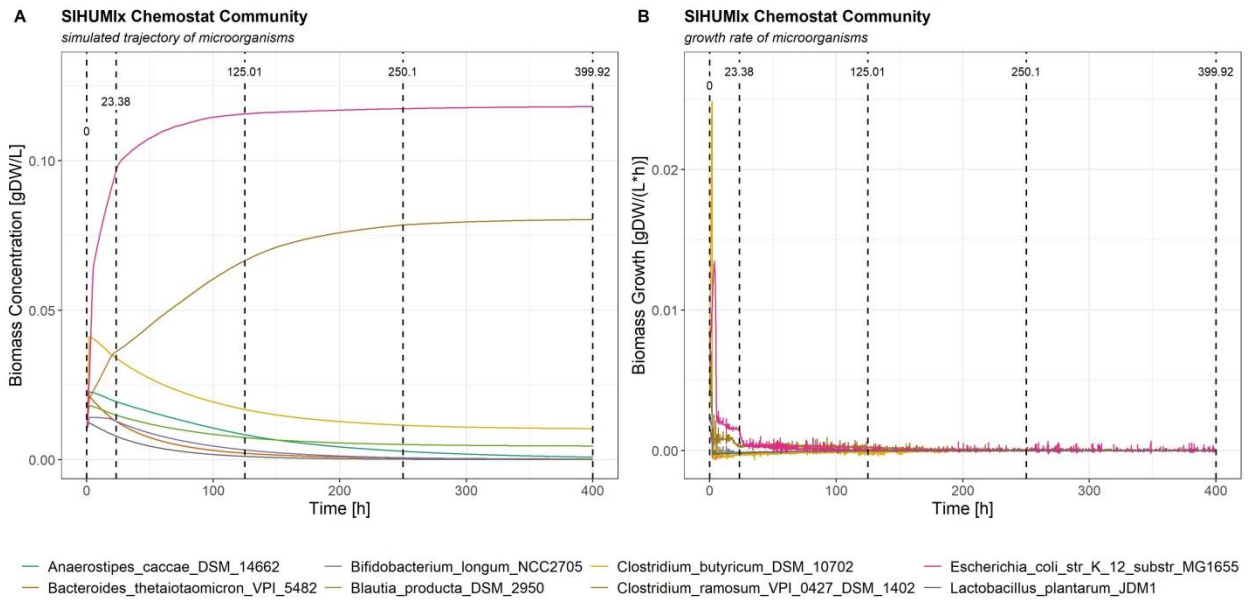

Figure 8: Trajectory of the SIHUMIx community in a simulated chemostat process
Biomass concentration (A) and biomass growth rate (B) over the entire simulation time. Vertical dashed lines refer to network slice time points on the chemostat dataset

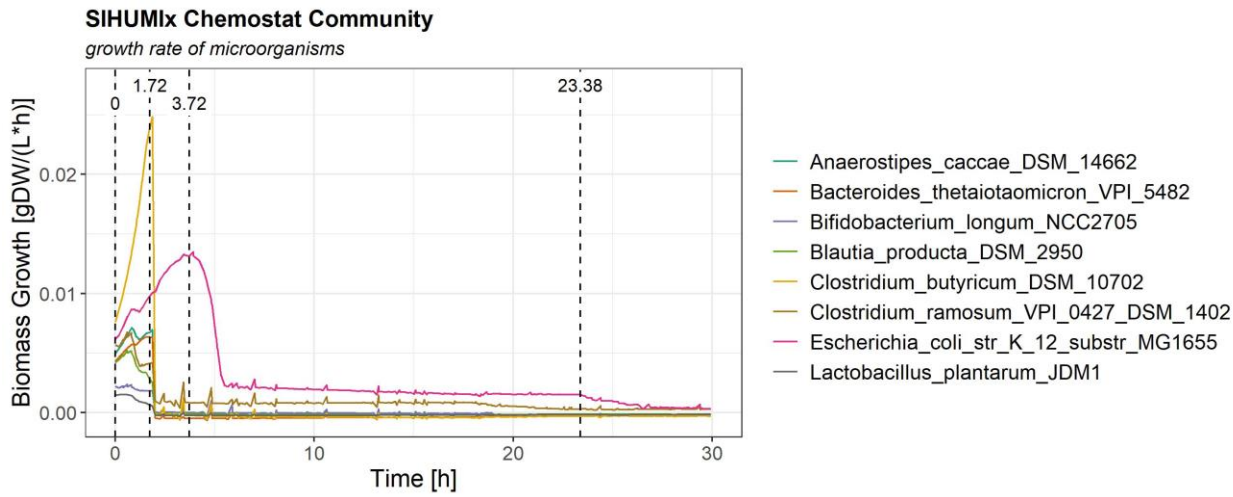

Figure 9: Detail of the SIHUMIx biomass growth rates in a simulated chemostat process
Biomass concentration (A) and biomass growth rate (B) over the first 30h out of 400h simulation in total. Vertical dashed lines refer to network slice time points on the chemostat dataset

Compounds

Due to the entire community growing from the start, compound consumption and production is particularly high at the beginning. Figure 10 Panel A reveals an obvious accumulation of both formate (for[e]) and acetate (ac[e]) over the entire time course, peaking at 5.22mM and 4.79mM at the end of the cultivation, representing almost 40% of the entire amount of compounds with

respect to the molecule numbers (for completeness: proton – h[e] would be at 6.54mM respectively approx. 26 %). Median compound concentration is at approx. 0.1mM at the end. Overall, Panel B shows especially high production rates right from the start with generally high

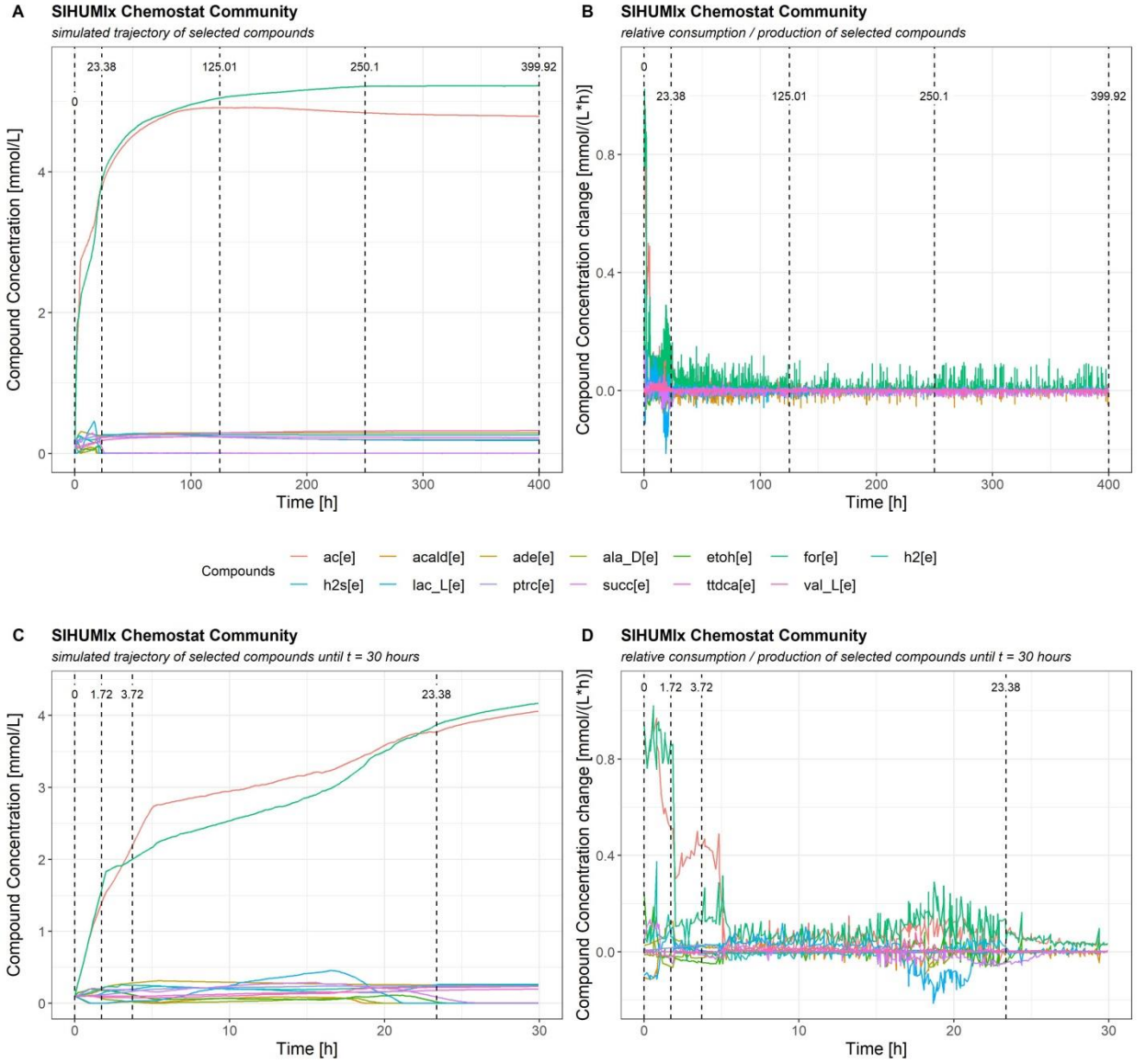

Figure 10: Trajectory of compounds in a simulated chemostat process

Concentration of 13 (out of 139) compounds (A, C) and corresponding relative time derivatives (B, D). Compounds were selected by highest abundance and/or highest fluctuations summed up over the entire simulation time (excluding water and proton for clarity). Panels A and B describe the entire simulation time course, whereas Panels C and D concentrate on the first 30h. Vertical dashed lines refer to network slice time points on the chemostat dataset. The compounds are displayed with the following abbreviations: ac[e]: acetate, acald[e]: acetaldehyde, ade[e]:adenine, ala_D[e] :D-alanine, etoh[e]: ethanol, for[e]: formate, h2[e]: hydrogen, h2s[e]: hydrogen sulfide, lac_L[e]: (S)-lactate, ptrc[e]: putrescine, succ[e]: succinate, ttdca[e]: tetradecanoate, val_L[e]: L-valine

amplitudes, which then decrease and peak again at around t = 18.7h. Afterwards amplitudes are going down again, with no visible change after around 125h into the cultivation. Looking into the first 30 hours in detail in Panel C and D, it becomes apparent that simultaneously to the collapse of microbial growth rates in Figure 9 at t = 2h, production of acetate and especially formate are dropping rapidly, then slightly rebound until falling quickly to low levels fluctuating around an accumulation rate of approx. 0.05mM/h. At this time a steady production of (S)-lactate jumpstarts, peaking in a concentration of approx. 0.45mM at t = 16.5 hours into the cultivation. With the transition of lactate production to lactate consumption, peaking at an uptake rate of -0.21mM/h of lactate at t = 18.7h, acetate and formate production rate rise again, their trend mirroring the one of lactate. In the following course, both compound consumption and production rates are going down again and level off at around t = 28h.

4.4.2. Characterization of complete network

The visualization of the entire metabolic network of the chemostat simulation yields the same network as the visualization of the entire batch network, shown in Figure 5. This states, that considering the set of all unique reactions being active at least once in the entire time course, both networks are identical. This is probably due to similar conditions at the beginning of the simulation, as the relative composition of the microbial community and the constituents of the medium are the same. However, it does not hold any other temporal information regarding when specific reactions were active, to which extent etc. Therefore the fact that the complete networks are identical does not suffice to argue on similar underlying dynamics during the simulation.

4.4.3. Pre-Linearization

Time points finding for network slices

In Figure 8, Figure 9 and Figure 10, vertical lines represent the time points, where network slices were extracted from the full dynamic network. To account for the high activity of the dynamic network within the first 30 hours, the first 4 time points are defined in this time window of the simulation. At t = 0h initial interactions are captured. The second slice resides at t = 1.7h, at the end of the growth phase considering almost all species of the consortia. After the sudden drop in microbial growth rates except for *E. coli.*, time point t = 3.7h sits at the point of maximum growth rate of E.coli, to reflect conditions at that point. Afterwards, at t = 23.4h, this slice should characterize the community interaction, when microbial growth rates are plateauing, before

finally leveling off at around t = 28h. The subsequent slices at t = 125h, t = 250h and t = 399.9h should assess the composition and microbial cross-feeding interaction at times, when the system seemingly enters a steady state.

Table 6: Graph parameters of the network slices of the simulated chemostat cultivation
Slices contain only connected nodes

Time [h]	**0**	**1.7**	**3.7**	**23.4**	**125**	**250**	**399.9**
Order	137	138	127	131	133	136	137
Size	549	571	432	522	513	513	521
Cumulative Flux [mM/h]	599.6	264.8	62.8	36.9	34.3	37.4	38.6
(thereof excretion)	*391.7*	*165*	*41.3*	*26.4*	*25*	27.3	28.3
Diameter	6	6	6	6	6	6	6

Characterization of pre-linearized network slices

The basic structure of the network stays within a range from 127 to 138 nodes, and 432 to 571 edges. The theoretical maximum would be 149 nodes and 684 edges, if all of their concentration or flux value would be greater than zero and form one connected component, see also section 4.3.2. The only obvious structural exception is at t = 3.7h into the cultivation, as it is the only slice with less than 130 nodes and less than 510 edges. The diameter of the network slices still stays the same at any time point. On the level of metabolic interactions, the cumulative flux of the slices reflects the high activity at the start, with conditions similar to a batch process, with quickly falling fluxes accounting for a steep decline of cumulative flux until at the fourth slice t = 23.4h, where it hits a plateau at around 37mM/h. From there on, cumulative flux values are fluctuating around this value. Overall, the excretion cumulative flux exceeds the uptake cumulative flux at every time point. However, within the first three slices until t = 3.7h, the proportion of excretion cumulative flux to entire cumulative flux is between 62% and 65%, rising to a range of 71% to 73% afterwards.

4.4.4. Linearization

The linearization of the network slices led to smaller networks in order and size for all time points (Table 6). The number of removed edges compared to the original slices is between 25 and 33 for each time point, except t = 3.7h with just 16 removed edges. In relation to the full

slices, this equates to a total removal of edges per time point between 5% and 6.3% of the original number. This reduction makes for a loss in cumulative flux of 0.1% up to 2.8% of the pre-linearization cumulative flux. Again t = 3.7h stands out with a minimal loss of 0.1%. In total, 188 are removed within the entire linearization process of all full network slices. Thereof 81 are unique, with 6 out of 81 specific edges repeatedly removed at every time point, yet most edges (57 out of 81) were removed twice at most. The diameter changed just once due to linearization, for the network slice at t = 0h from 6 to 4.

Table 7: Graph parameters of the linearized network slices of the simulated batch cultivation
Linearization of the network slices characterized in Table 4; non-connected nodes were omitted

Time [h]	**0**	**1.7**	**3.7**	**23.4**	**125**	**250**	**399.9**
Order	135	136	125	129	131	134	135
Size	510	526	402	478	471	469	477
Removed edges	25	33	16	29	27	29	29
Cum. Flux conservation[%] *	98.8	97.2	99.9	98.5	98.4	98.5	99
Diameter **	4 (6)	6 (6)	6 (6)	6 (6)	6 (6)	6 (6)	6 (6)

* cumulative flux in relation to pre-linearization cumulative flux
** value in brackets corresponds to diameter of the non-linearized slice at that time point

Looking at the linearized slice at t = 1.7h into the cultivation (Figure 11), *C. butyricum* sticks out in the last microbial layer of the network, with an in degree of 14 and an out degree of 5, acquiring most of the source compounds present. All visible edges hold a flux greater than 0.9mM/h due to a filtering step. Most connected compounds are formate and acetate, accumulating in the reactor. Looking at intermediate metabolites (red border paint), which are both consumed (red edges) and produced (green edges) at that time, there is a competition on ethanol (etoh[e]) and succinate (succ[e]) with a high metabolic rate as their concentration is decreasing. For instance, there are producer of ethanol with high production rates such as *C. ramosum* and *B. longum*, with three other species profiting on that. Succinate on the other hand is produced foremost by *B. producta* at high rates, which then too is distributed among three subsequent consumers. Acetate however, although an intermediate in the network representation, has *C. butyricum* as the only main consumer and is very abundant at that time, its function mere a sink, without a key cross-feeding role at that time. Overall, organisms which are mainly

producer of intermediates sit higher in the topological order, whereas the ones in the center tend to both feed and produce intermediates, with *C. butyricum* in the last row as the final consumer.

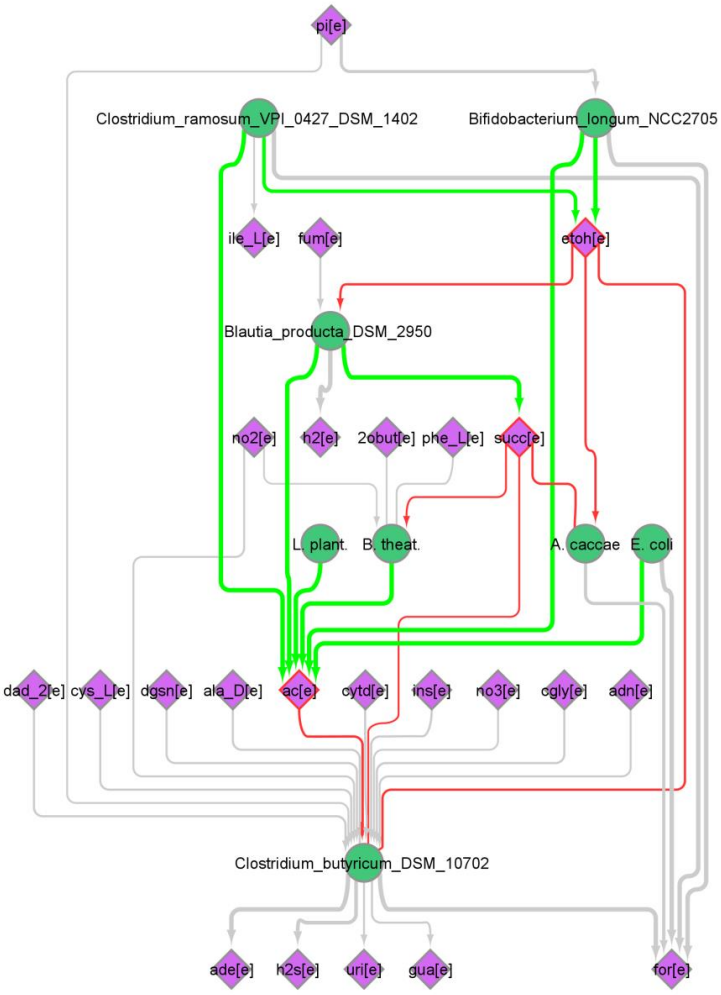

Figure 11: Network slice of the chemostat simulation at t = 1.7 h

This graph is a subgraph of the entire slice at t = 1.7h, obtained by filtering out edges with a flux smaller than 0.9mM/h and excluding water and proton compound nodes for clarity. Afterwards the graph was stripped of all non-connected nodes. Green round nodes represent microorganisms, purple diamonds represent compounds. Edges stand for excretion and uptake reaction of the organisms. The width of the edges is mapped to the flux through the related reaction at that moment, and the direction is indicated by arrows. Red marked nodes are intermediate compounds, with $deg^-(v)\neq 0$ and $deg^+(v)\neq 0$. Incoming edges to these nodes are marked in green, outgoing edges in red.

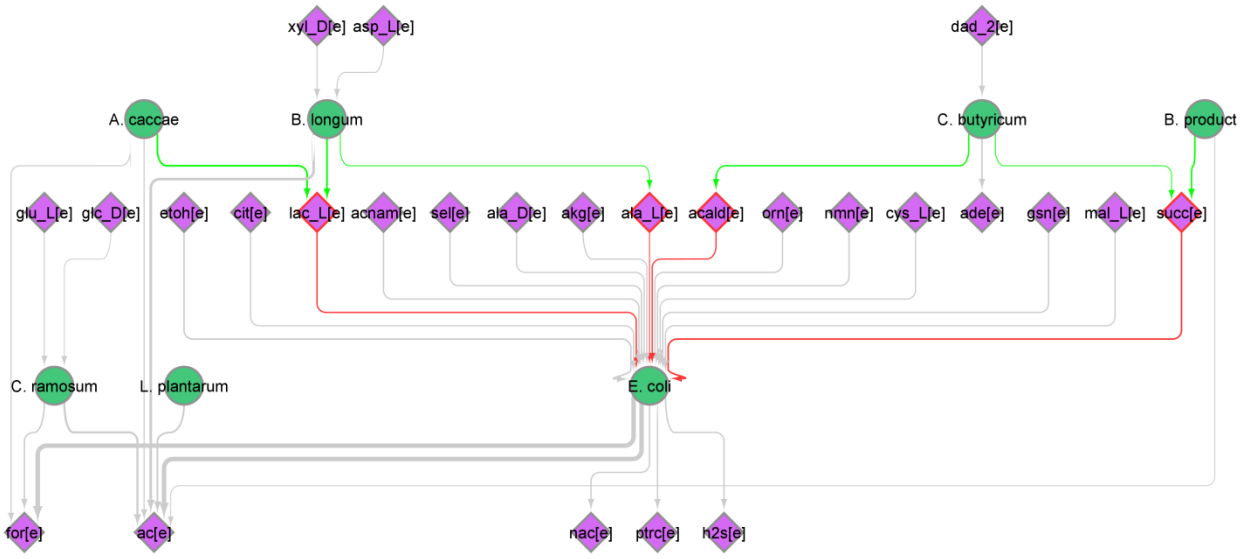

Figure 12: Network slice of the chemostat simulation at t = 3.7 h

This graph is a subgraph of the entire slice at t = 3.7h, obtained by filtering out edges with a flux smaller than 0.3mM/h and excluding water and proton compound nodes for clarity. Afterwards the graph was stripped of all non-connected nodes. Green round nodes represent microorganisms, purple diamonds represent compounds. Edges stand for excretion and uptake reaction of the organisms. The width of the edges is mapped to the flux through the related reaction at that moment, and the direction is indicated by arrows. Red marked nodes are intermediate compounds, with $deg^-(v) \neq 0$ and $deg^+(v) \neq 0$. Incoming edges to these nodes are marked in green, outgoing edges in red.

Continuing to the next slice at t = 3.7h, represented in Figure 11, the metabolic network interactions have changed quite a bit. The depiction is flatter and wider, consisting of just 2 layers of organisms. All edges visible account for a flux of greater than 0.3mM/h due to a filtering step. Regarding their role in the network, *E. coli* and *C. butyricum* account for the most striking difference compared to Figure 11. The latter changed position from the bottom row to the top and is visibly consuming just one compound and producing 2 intermediates, marked with red border paint. On the other hand, *E. coli* takes the prominent position of *C. butyricum* from the last slice as the most connected organism, consuming various compounds and all intermediate substrates. *B. producta* still serves as the biggest producer of succinate, which *C. butyricum* now produces instead of consuming it. As in the slice before, *L. plantarum* has no incoming edges indicating no uptake reactions of high volumes and resides as the least growing organisms at both time points. Succinate persists as a relevant intermediate compared to t = 1.7h, now joined by (S)-lactate (lac_L[e]), L-alanine (ala_l[e]) and acetaldehyde (acald[e]).

To assess metabolic interactions towards the final state of the chemostat reactor, Figure 13 shows slices at t = 125h and at the end of the cultivation at t = 399.9h. As cumulative flux in the network continuously decreases over time, a filter of 0.1mM/h was sufficient to achieve these two temporal networks. Although at first sight they can look quite different, these two networks are already quite alike. Structurally, the network in Panel A at t = 125h has two more layers, with *C. butyricum* moving into the top microbial layer. This is due to the fact that one product, acetaldehyde (acald[e]), is an intermediate compound at that time, being consumed by *B. longum* and *C. ramosum*, and C. butyricum being its only producer. As the latter does not consume any other intermediates, the topology sort of the hierarchal layout algorithm puts it on top of the network. At t = 399.9h acetaldehyde is merely a product of *C. butyricum* and *A. caccae*, not being metabolized any further, which means it does not restrict the topological rank of these organisms. Comparing nodes, there are three nodes (cgly[e], cys_L[e] and ala_L[e]) which are unique to the slice at t = 125h and two nodes (gua[e] and gdsn[e]) solely appearing in the second network at t = 399.9h. All of these are either sinks or sources in their respective network with one connected microorganism at most. Apart from acetaldehyde, both networks share two common intermediate compounds, ethanol and succinate, with equal connections to their respective consumers and producers. *B. producta* provides the biggest share of the production of succinate in both cases, consistent with the network slices analyzed at t = 1.7 h and t = 3.7 h. *B. longum* and *C. ramosum* produce ethanol, as already observed at time point t = 1.7h. Overall, both network slices at t = 125h and t = 399.9 show similar structural features in regard to important metabolite exchange reactions and share a large part of their nodes and edges. However, the linearization process at t = 125h removed two reactions from the network. One reaction is the uptake of lactate (lac_L[e]) of *C. ramosum*, the other one is the production of acetalhyde of *A. caccae*. These two edges would form cycles in the network, making lactate an intermediate compound in the network representation as well. Both edges would create relatively big cycles in regard to the size of the linearized network representation in Panel A. Similarly, the linearized network slice at t = 250h, which is structurally identical with the one at the end of the cultivation shown in Panel B, misses the consumption of acetaldehyde through *B. longum* and *C. ramosum* due to linearization, the latter also missing in the last network slice at t = 399.9h.

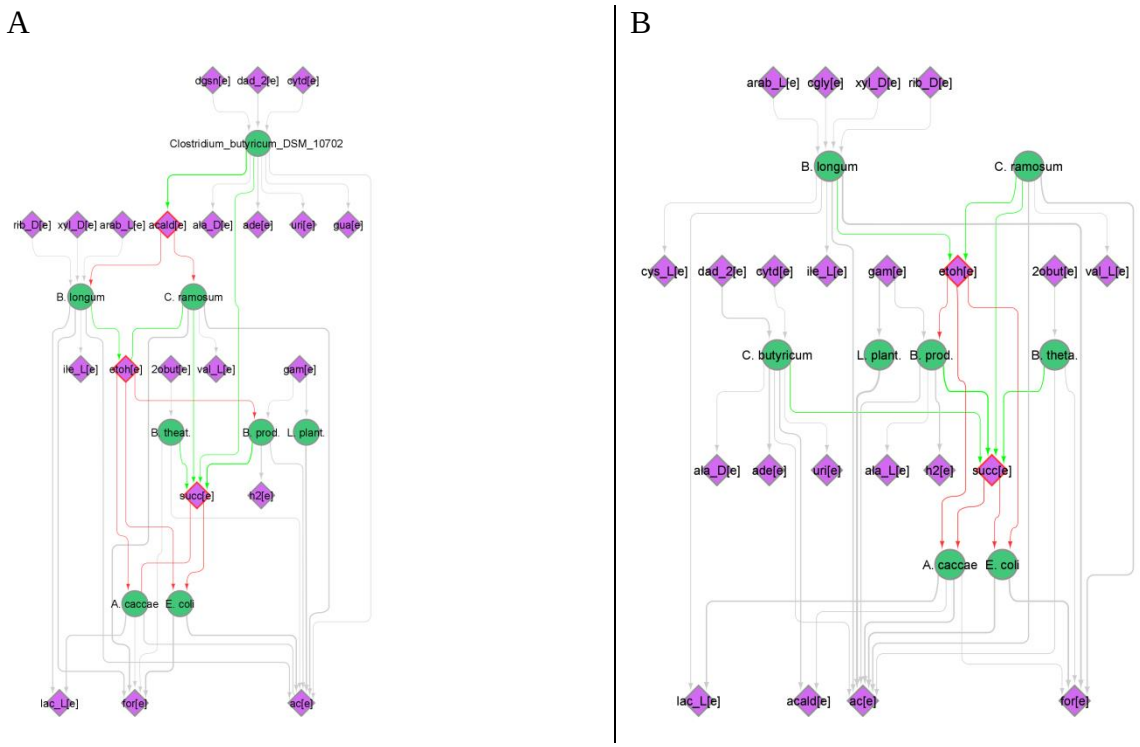

Figure 13: Linearized network slices of the chemostat simulation at t = 125h and t = 399.9h

This graph is a subgraph of the entire slice at t = 125h (Panel A) and t = 399.9h (Panel B), obtained by filtering out edges with a flux smaller than 0.1mM/h and excluding water and proton compound nodes for clarity. Afterwards the graph was stripped of all non-connected nodes. Green round nodes represent microorganisms, purple diamonds represent compounds. Edges stand for excretion and uptake reaction of the organisms. The width of the edges is mapped to the flux through the related reaction at that moment, and the direction is indicated by arrows. Red marked nodes are intermediate compounds, with $deg^-(v) \neq 0$ and $deg^+(v) \neq 0$. Incoming edges to these nodes are marked in green, outgoing edges in red..

5. Evaluation

Overall, the visual representation of the extracted and linearized network slices reflected the trajectories of the underlying simulation data well. Furthermore it enhanced understanding of the cause-effect interplay within the microbial network.

5.1. Visual representation and characterization

The visual analysis of the created metabolic networks can be carried out at different levels of detail. In Figure 6, the network slice of the batch simulation at t = 0.04h shows a rather complex and unfiltered version of the linearized slice at that time, which is then overlayed by the edges which were removed due to the linearization process. Although the static view limits the informative value of this representation, it still holds some important structural information. Knowing the flow direction of the underlying directed acyclic graph, the prominent organism in the top layer is mostly feeding on source compounds, which are not produced by other species in the consortium at that time. The overlayed red back edges show the exemption of this notion from the linearized (grey) version of the network. Considering eventual substrate depletion in a batch process, this seeming dependency on non-produced compounds is not the best precondition for further growth. Furthermore, even though distinct connections are hardly determinable, compounds of the first topological layer are more likely to be possibly consumed by other species, thus inferring a higher competition on those compounds. In general, if analyzing the network in Cytoscape, a rather complete representation of the metabolic network with this clear hierarchical structure, already integrating the removed edges of the linearization process, can be a good start into the analysis of the network slice.

A more detailed approach which is also more suitable for the static and spatially limited nature of a manuscript is Figure 7, representing the network slice of the batch simulation at t = 0.09h into the simulation. *E. coli* and *C. butyricum* are the driving species at that time according to growth rates. The representation confirms this, as their position within the hierarchy of the layout is in the ultimate and penultimate microbial layer, being both highly connected in comparison to the other species. Furthermore it is apparent, that both are feeding on several compounds, which are visibly solely consumed by them. Complementary they also feed on highly consumed intermediate compounds like ethanol and succinate.

Figure 11 is a network slice 1.7h after the chemostat cultivation started. *C. butyricum* holds a very prominent position in the last microbial layer of the hierarchical layout. The shape of the entire arrangement and the positioning of the organism kind of resembles a central flow valve. Looking at the metabolic interactions, it becomes clear that this impression actually reflects its function rather well: L-isoleucine (ile_L[e]), hydrogen (h2[e]) and formate (for[e]) are the only compounds not consumed by *C. butyricum* at that time. The only major flow figuratively "passing by" is the excretion of formate of the other species of the community. Although the layout states the obvious dominating role of *C. butyricum* correctly, the actual growing rates does not show that much of a difference then the visualization might infer without further data mining. This is especially true for *E. coli*, which is the second most growing species at that point at approx. half the rate than *C. butyricum*, but with no visible consuming reaction and only two outgoing reactions. This is due to the fact that the abundant incoming edges hold lower fluxes, which are hidden by the applied filter. Additional statistics like cumulative incoming flux could help in order to assess the state of interaction more accurately.

Proceeding further towards the end of the chemostat simulation, Figure 13 shows the comparison of the linearized network slice at t = 125h and t = 399.9h, the end of the cultivation. Within both networks, it is hard to point out a microorganism, which clearly stands out. This is even more difficult, when taking the eliminated edges into account, which are missing in the linearized representation of both networks. However, there is a prominent feature common in both networks. It is the intermediate compound succinate (succ[e]), which is produced by four organisms, among which *B. producta* is the biggest producer at both time points, and being consumed by two ones. The reoccurring of succinate as an intermediate compound within all analyzed network slices of the chemostat cultivation, hints at an important role of succinate as a cross-feeding compound. In this regard, *B. producta* has to be mentioned as a consistent producer of succinate over all selected time points. A recent paper supports this finding, stating that concentrations of succinate in human fecal samples are always very low, due to major genera in the human gut using it as a substrate and some producing it as an end product [45]. The genus of *Blautia* however is not among the end producers of succinate. Other noticeable compounds in this regard are ethanol and acetaldehyde, the latter showing up as an intermediate compound in the network slice at t = 399.9h too, if one would add the missing consumption of acetaldehyde by *E. coli* due to the linearization process.

In general it must be noted, that the interactive handling of the temporal network slices in Cytoscape allows the most benefit in analyzing and exploiting both visual properties and additional information stored in attributes, which can contain either meta data or hold customized calculated graph theoretical parameters in order for further analysis.

5.2. Linearization

The linearization of network slices of the dynamic networks of a simulated batch and chemostat process were helpful in complementing the overall analysis of these cross-feeding metabolic networks. The conservation of cumulative flux after linearization compared to the cumulative flux of the pre-linearized network slices was at least 95.2% for the batch related networks and 97.2% for the chemostat related ones. However, the process did not yield a reduction in network order greater than 5.7% either. The high conservation justifies the utilization of linearized networks in the analysis of metabolic interactions, as through linearization the hierarchical "trickle-down" nature of metabolic networks is emphasized and can be exploited for interesting structural features without losing too much information. The effect of simplification through linearization in the sense of pruning nodes and edges is rather negligible, as by favoring the persistence of high flux edges during the process, similar results can be obtained by filtering steps. In fact, while reducing the size of the linearized networks to a reasonable printable scale, the filter applied (flux value < x) often diminish the differences between the filtered linearized network slices and the filtered pre-linearized network slices, as excluded edges are by design rather small and thus filtered out first (see Figure 7, Figure 11, Figure 12). On the other hand, both linearized network slices of the chemostat simulation depicted in Figure 13 and the one mentioned in this context miss at least one edge compared to their pre-linearized counterpart, due to their relatively high flux value compared to all other metabolic reactions present at the respective time. By considering these backedges in further analysis, integrating them into the linearized representation of the graph, one could identify interesting metabolic cycles and currency metabolites which are continuously recycled within the metabolic network.

5.3. Conclusion

The linearization of temporal network slices guarantees a hierarchical layout representation of the graph. This representation facilitates intuitive understanding of the existing metabolic interaction network. The fixed flux direction allows for a fast but guarded assessment of the current roles of the microorganisms and compounds embedded in their metabolic network. Important visual features to take into account is the overall maximum topological order, the positioning of the organism or the compound within the hierarchical depiction of the graph and their connectedness. By highlighting intermediate compounds or compounds many species are competing on, it becomes easier to infer a deeper understanding of the function of certain constituents within the network. Visual properties like edge width representing flux volume can further contribute to a more integral notion of the metabolic interactions. Complemented with numbers from the underlying simulated data, a visually supported analysis can yield hints at interesting key compounds, reactions or organisms. Looking at temporal network slices distributed over the entire time course of a simulation, reoccurring patterns or strikingly distinct features could serve as starting points of further investigation or altering of simulation parameters for future test runs.

6. Outlook

In order to further extent the explanatory power of the visual analysis of cross-feeding metabolic networks, it needs to be complemented by extended graph analytical parameters. For instance centrality measures like betweenness could allow supporting findings from the visual analysis or hint to missed intermediate compounds due to filtering steps of the visualization process. Exploiting graph theoretical properties of the graph could also help to identify interesting time points. If this process is automatized, taking the entire time series data into account, it could help the user to select more meaningful network slices, which are interesting for further metabolic network analysis.

Another aspect would be the classification of the metabolites, for instance allowing the selection of carbon sources to concentrate on the carbon flow within the metabolic network. A desired selection of any sort could also already be incorporated into the linearizing algorithm, thus preferably keeping desired elements in the network and omitting others.

To thoroughly exploit the result of the linearization of the network slices, a systematic examination of the removed cycles could yield potential currency metabolites and identify further intermediate compounds, which are missed through linearization.

Additional visual representations could highlight interactions on species level. By disaggregating producing and consuming fluxes of metabolic compounds, one could display species dependent proportions of their total uptake amount of metabolites, which are produced by other species, and thus directly putting species into relation to each other. This could be especially relevant, if graphs become very big and therefore visualization requires reducing the displayed elements. Furthermore, a core set of relevant nodes and edges could be defined after overall analysis and put into an interactive animation, to visualize their dynamic behavior over the entire time course in a more condensed way.

Finally, an evaluation of hierarchic representation and linearization on big networks would be required in order to assess the applicability on networks with 100s of species.

Bibliography

[1] R. R. Kanherkar, N. Bhatia-Dey, and A. B. Csoka, "Epigenetics across the human lifespan," *Frontiers in Cell and Developmental Biology*, vol. 2, no. SEP. Frontiers Media S.A., p. 49, 09-Sep-2014.

[2] S. R. Gill *et al.*, "Metagenomic analysis of the human distal gut microbiome," *Science (80-.).*, vol. 312, no. 5778, pp. 1355–1359, Jun. 2006.

[3] Q. Zhang *et al.*, "Circulating mitochondrial DAMPs cause inflammatory responses to injury," *Nature*, vol. 464, no. 7285, pp. 104–107, 2010.

[4] P. Hugenholtz, B. M. Goebel, and N. R. Pace, "Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity," *Journal of Bacteriology*, vol. 180, no. 18. American Society for Microbiology, pp. 4765–4774, 1998.

[5] R. I. Amann, W. Ludwig, and K. H. Schleifer, "Phylogenetic identification and in situ detection of individual microbial cells without cultivation," *Microbiological Reviews*, vol. 59, no. 1. American Society for Microbiology, pp. 143–169, 1995.

[6] N. R. Pace, D. A. Stahl, D. J. Lane, and G. J. Olsen, "The Analysis of Natural Microbial Populations by Ribosomal RNA Sequences," Springer, Boston, MA, 1986, pp. 1–55.

[7] N. J. Loman and M. J. Pallen, "Twenty years of bacterial genome sequencing," *Nature Reviews Microbiology*, vol. 13, no. 12. Nature Publishing Group, pp. 787–794, 01-Dec-2015.

[8] P. J. Turnbaugh, R. E. Ley, M. Hamady, C. M. Fraser-Liggett, R. Knight, and J. I. Gordon, "The Human Microbiome Project," *Nature*, vol. 449, no. 7164. Nature Publishing Group, pp. 804–810, 18-Oct-2007.

[9] M. Kanehisa *et al.*, "From genomics to chemical genomics: new developments in KEGG.," *Nucleic Acids Res.*, vol. 34, no. Database issue, pp. D354–D357, Jan. 2006.

[10] V. M. Markowitz *et al.*, "IMG: The integrated microbial genomes database and comparative analysis system," *Nucleic Acids Res.*, vol. 40, no. D1, pp. D115–D122, Jan.

2012.

[11] P. D. Karp, M. Riley, S. M. Paley, and A. Pellegrini-Toole, “The MetaCyc database,” *Nucleic Acids Res.*, vol. 30, no. 1, pp. 59–61, Jan. 2002.

[12] C. H. King *et al.*, “Baseline human gut microbiota profile in healthy people and standard reporting template,” *PLoS One*, vol. 14, no. 9, p. e0206484, Sep. 2019.

[13] V. Marx, “The big challenges of big data,” *Nature*, vol. 498, no. 7453, pp. 255–260, Jun. 2013.

[14] P. Kohl, E. J. Crampin, T. A. Quinn, and D. Noble, “Systems Biology: An Approach,” *Clin. Pharmacol. Ther.*, vol. 88, no. 1, pp. 25–33, Jul. 2010.

[15] M. Kumar, B. Ji, K. Zengler, and J. Nielsen, “Modelling approaches for studying the microbiome,” *Nat. Microbiol.*, vol. 4, no. 8, pp. 1253–1267, 2019.

[16] N. Jehmlich, C. Vogt, V. Lünsmann, H. H. Richnow, and M. von Bergen, “Protein-SIP in environmental studies,” *Curr. Opin. Biotechnol.*, vol. 41, pp. 26–33, 2016.

[17] X. Chen, S. Wei, Y. Ji, X. Guo, and F. Yang, “Quantitative proteomics using SILAC: Principles, applications, and developments,” *Proteomics*, vol. 15, no. 18, pp. 3175–3192, Sep. 2015.

[18] D. Popp and F. Centler, “μbialSim: constraint-based dynamic simulation of complex microbiomes,” *bioRxiv*, p. 716126, 2019.

[19] N. Becker, J. Kunath, G. Loh, and M. Blaut, “Human intestinal microbiota: Characterization of a simplified and stable gnotobiotic rat model,” *Gut Microbes*, vol. 2, no. 1, Jan. 2011.

[20] B. C. Kirkup and M. A. Riley, “Antibiotic-mediated antagonism leads to a bacterial game of rock-paper-scissors in vivo,” *Nature*, vol. 428, no. 6981, pp. 412–414, Mar. 2004.

[21] G. D’Souza, S. Waschina, S. Pande, K. Bohl, C. Kaleta, and C. Kost, “Less is more: Selective advantages can explain the prevalent loss of biosynthetic genes in bacteria,” *Evolution (N. Y).*, vol. 68, no. 9, pp. 2559–2570, Sep. 2014.

[22] J. J. Morris, R. E. Lenski, and E. R. Zinser, "The black queen hypothesis: Evolution of dependencies through adaptive gene loss," *MBio*, vol. 3, no. 2, May 2012.

[23] M. T. Mee, J. J. Collins, G. M. Church, and H. H. Wang, "Syntrophic exchange in synthetic microbial communities," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 111, no. 20, pp. E2149–E2156, May 2014.

[24] N. W. Smith, P. R. Shorten, E. Altermann, N. C. Roy, and W. C. McNabb, "The Classification and Evolution of Bacterial Cross-Feeding," *Front. Ecol. Evol.*, vol. 7, p. 153, May 2019.

[25] M. Ziesack *et al.*, "Engineered Interspecies Amino Acid Cross-Feeding Increases Population Evenness in a Synthetic Bacterial Consortium," *mSystems*, vol. 4, no. 4, Aug. 2019.

[26] M. Cavaliere, S. Feng, O. S. Soyer, and J. I. Jiménez, "Cooperation in microbial communities and their biotechnological applications," *Environ. Microbiol.*, vol. 19, no. 8, pp. 2949–2963, Aug. 2017.

[27] C. L. Maynard, C. O. Elson, R. D. Hatton, and C. T. Weaver, "Reciprocal interactions of the intestinal microbiota and immune system," *Nature*, vol. 489, no. 7415. Nature Publishing Group, pp. 231–241, 13-Sep-2012.

[28] A. D'Onofrio *et al.*, "Siderophores from Neighboring Organisms Promote the Growth of Uncultured Bacteria," *Chem. Biol.*, vol. 17, no. 3, pp. 254–264, Mar. 2010.

[29] G. D'Souza, S. Shitut, D. Preussger, G. Yousif, S. Waschina, and C. Kost, "Ecology and evolution of metabolic cross-feeding interactions in bacteria," *Natural Product Reports*, vol. 35, no. 5. Royal Society of Chemistry, pp. 455–488, 01-May-2018.

[30] R. Diestel, *Graph theory*, 3rd ed. Springer-Verlag Heidelberg, New York, 2005.

[31] T. von Landesberger *et al.*, "Visual analysis of large graphs: State-of-the-art and future research challenges," *Eurographics Symp. Geom. Process.*, vol. 30, no. 6, pp. 1719–1749, 2011.

[32] A. Kerren, J. Stasko, J.-D. Fekete, and C. North, *Information Visualization*, Lecutre No. Springer Berlin Heidelberg, 2008.

[33] D. Keim, G. Andrienko, J.-D. Fekete, C. Görg, J. Kohlhammer, and G. Melançon, "Visual Analytics: Definition, Process, and Challenges," in *Information Visualization: Human-Centered Issues and Perspectives*, A. Kerren, J. T. Stasko, J.-D. Fekete, and C. North, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 154–175.

[34] K. Sugiyama, S. Tagawa, and M. Toda, "Methods for Visual Understanding of Hierarchical System Structures," *IEEE Trans. Syst. Man Cybern.*, vol. 11, no. 2, pp. 109–125, 1981.

[35] P. A. Jensen and J. A. Papin, "MetDraw: automated visualization of genome-scale metabolic network reconstructions and high-throughput data," *Bioinformatics*, vol. 30, no. 9, pp. 1327–1328, May 2014.

[36] N. Martyushenko and E. Almaas, "ModelExplorer - software for visual inspection and inconsistency correction of genome-scale metabolic reconstructions," *BMC Bioinformatics*, vol. 20, no. 1, p. 56, Jan. 2019.

[37] J. Boele, B. G. Olivier, and B. Teusink, "FAME, the Flux Analysis and Modeling Environment," *BMC Syst. Biol.*, vol. 6, no. 1, p. 8, Jan. 2012.

[38] Z. A. King, A. Dräger, A. Ebrahim, N. Sonnenschein, N. E. Lewis, and B. O. Palsson, "Escher: A Web Application for Building, Sharing, and Embedding Data-Rich Visualizations of Biological Pathways," *PLOS Comput. Biol.*, vol. 11, no. 8, p. e1004321, Aug. 2015.

[39] B. R. Granger, Y.-C. Chang, Y. Wang, C. DeLisi, D. Segrè, and Z. Hu, "Visualization of Metabolic Interaction Networks in Microbial Communities Using VisANT 5.0," *PLOS Comput. Biol.*, vol. 12, no. 4, p. e1004875, Apr. 2016.

[40] L. F. Buchweitz *et al.*, "Visualizing metabolic network dynamics through time-series metabolomic data," *BMC Bioinformatics*, vol. 21, no. 1, p. 130, Dec. 2020.

[41] S. Magnúsdóttir *et al.*, "Generation of genome-scale metabolic reconstructions for 773

members of the human gut microbiota," *Nat. Biotechnol.*, vol. 35, no. 1, pp. 81–89, Jan. 2017.

[42] T. N. Gábor Csárdi, "The igraph software package for complex network research," *nterJournal Complex Syst. 1695*, p. 78, 2006.

[43] J. A. Gustavsen, S. Pai, R. Isserlin, B. Demchak, and A. R. Pico, "RCy3: Network biology using Cytoscape from within R," *F1000Research*, vol. 8, p. 1774, Dec. 2019.

[44] P. Shannon *et al.*, "Cytoscape: A software Environment for integrated models of biomolecular interaction networks," *Genome Res.*, vol. 13, no. 11, pp. 2498–2504, Nov. 2003.

[45] K. Oliphant and E. Allen-Vercoe, "Macronutrient metabolism by the human gut microbiome: Major fermentation by-products and their impact on host health," *Microbiome*, vol. 7, no. 1, pp. 1–15, Jun. 2019.

Statutory Declaration

I hereby declare this thesis as my own work and that I have not called upon the help of a third party. In addition, I affirm that neither I nor anybody else has submitted this thesis or parts of it to obtain credits elsewhere before. I have clearly marked and acknowledged all quotations or references that have been taken from the works of other. All secondary literature and other sources are marked and listed in the bibliography. The same applies to all charts, diagrams and illustrations as well as to all Internet sources.

Leipzig, Date	Name, Signature