

Computational strategies employing stratification and network analysis to advance lipidomics-driven research

Tim Daniel Rose

Vollständiger Abdruck der von der TUM School of Life Sciences der Technischen Universität München zur Erlangung eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitz: Prof. Dr. Mathias Wilhelm

Prüfer*innen der Dissertation:

1. TUM Junior Fellow Dr. Josch Konstantin Pauling
2. Prof. Dr. Corinna Dawid
3. Jun. Prof. Dr. Anne-Christin Hauschild

Die Dissertation wurde am 01.09.2022 bei der Technischen Universität München eingereicht und durch die TUM School of Life Sciences am 02.02.2023 angenommen.

Acknowledgments

First of all, I would like to thank my supervisor Josch Pauling. You believed in me very early and convinced me to move with you to your new lab in Freising. We spent countless hours discussing projects and introducing me to the fascinating world of lipidomics. Thank you for always being supportive and providing me with great opportunities to extend my horizon.

Another big thanks go to all current and past members of the LipiTUM group. Especially to Vivian and Nikolai for the exciting discussions and all the fun times. Thanks Nikolai, for the time we spent together developing and debugging LINEX. It was great working with you. Big thanks also go to the whole exbio Team for providing me with a great working environment in the lab and all the social activities. Thank you Martina, for helping with all the administrative issues and doing everything to make our work life as easy as possible.

A special thanks also go to the collaborators in the LiSyM consortium, especially Andrej and Olga. Working with you on this project for a long time and collaborating at the eye level was an honor.

Thank you Olya, for being the best partner I could imagine. For supporting me, for motivating me, for criticizing me, building me up, and just being there.

Finally, I want to thank my family. Thank you for always supporting me in all my decisions and giving me the best advice. You taught me always to reach further and make the best out of every situation. I will always be grateful for your support and appreciation.

And thanks to everyone who was not mentioned here but helped to make my journey a truly unique experience. Off to new adventures.

Abstract

Understanding diseases on the molecular level can enable precise diagnostic methods and more cost-effective treatments with fewer side effects. Systems medicine research tries to achieve this by acquiring large quantities of molecular data and analyzing them with computational methods to understand diseases' molecular heterogeneity and mechanisms. Lipids are biomolecules that are gaining more importance in systems medicine. They are involved in energy storage, signal transduction, and the composition of membranes. Mass spectrometry-based lipidomics can identify and quantify large numbers of molecular lipid species from biological samples. However, computational methods for integrating lipidomics in systems medicine workflows are still missing. This publication-based dissertation presents computational methods and applications to advance lipidomics-driven research.

First, the disease subtyping Molecular Signature identification using Biclustering (MoSBi) is presented. It utilizes the predictions of multiple biclustering algorithms to predict robust sample groups and characteristic molecular signatures. The method was developed to be highly interpretable using scalable network visualizations of the predictions. Further, the performance was evaluated on multiple synthetic and experimental datasets to show the advantages over other methods. The following publication utilized the method in a liver lipidomics study on non-alcoholic fatty liver disease (NAFLD). A common disease in modern societies, strongly associated with obesity. The analysis revealed molecular subtypes describing the progression of the disease that could classify patients into subgroups using lipid markers. This shows the potential of lipids as potential biomarkers for systems medicine.

Furthermore, the Lipid Network Explorer (LINEX) is presented. A method to create and analyze lipid species networks for lipidomics data. So far, systematic lipid metabolic networks are not available. They are necessary to functionally analyze and interpret lipidomics data for systems medicine. In a combined visualization of lipid networks with statistical measures, LINEX can show systematic alterations of the lipidome between experimental conditions. To make LINEX networks more comprehensive, LINEX² is presented as a preprint. It utilizes curated lipid class reactions from public databases and includes a network enrichment algorithm to predict enzymatic dysregulations.

Finally, two co-author contributions on computational drug repurposing strategies are discussed. Predicting drug repurposing candidates for new diseases can be the last computational step in a systems medicine workflow.

In summary, the work presented in this dissertation presents novel computational algorithms for subtyping and functional lipidomics interpretation and drug repurposing strategies. This can make it possible to add lipidomics as another molecular dimension to systems medicine research. All methods were developed with high accessibility to promote the interpretation of complex molecular data in clinical research.

Kurzfassung

Ein Verständnis von Krankheiten auf molekularer Ebene kann neue Diagnose- und Behandlungsverfahren ermöglichen, die präziser sind als symptom-basierte. In der Systemmedizin werden große Mengen molekularer Daten akquiriert und analysiert, um Krankheitsmechanismen und deren Heterogenität zu verstehen. Lipide sind Biomoleküle, die an Relevanz für die Systemmedizin gewinnen. Sie spielen eine wichtige Rolle für die zelluläre Energiespeicherung, Signalverarbeitung und sind Hauptbestandteil von Zellmembranen. Mit Massenspektrometrie kann eine Vielzahl von Lipiden quantifiziert werden aus biologischen Proben. Computergestützte Methoden, um Lipid Daten systematisch im systemmedizinischen Kontext zu analysieren fehlen bisher. In dieser kumulativen Dissertation werden Methoden und Anwendungen von computergestützten Methoden zur Integration von Lipidomik Daten in die Systemmedizin vorgestellt.

Zuerst wird die Subtypisierungsmethode Molecular Signature identification using Biclustering (MoSbi) präsentiert. Sie nutzt die Vorhersagen von mehreren Biclustering Algorithmen, um robuste Patientengruppen und molekulare Charakteristika vorherzusagen. Die Methode legt den Fokus auf Interpretierbarkeit der Resultate durch Netzwerkvisualisierungen. Zusätzlich wurde die Methode auf einer Vielzahl von synthetischen und experimentellen Daten evaluiert, um die Vorteile gegenüber anderen Methoden zu zeigen. In der darauffolgenden Publikation wurde Methodik auf Leber Lipidomik Daten von Patienten mit nichtalkoholischer Fettleber angewandt. Das ist eine weit verbreitete Krankheit in der modernen Zivilisation, die oft mit Übergewichtigkeit einhergeht. Die Analyse zeigte molekulare Subtypen auf, die den Verlauf der Krankheit beschreiben und anhand derer Patienten klassifiziert werden konnten, durch Identifizierung von Lipid Markern. Dadurch konnte das Potential von Lipiden als Biomarker gezeigt werden.

Des Weiteren wird der Lipid Network Explorer (LINEX) präsentiert, eine Methode zur Erstellung und Analyse von Lipid Species Netzwerken von Lipidomik Daten. Bis zu diesem Zeitpunkt waren systematische Lipid Metabolische Netzwerke nicht verfügbar. Diese sind wichtig, um Lipidomik Daten in der Systemmedizin funktional zu interpretieren. In einer kombinierten Visualisierung mit statistischen Kenngrößen können systematische Lipidomveränderungen aufgezeigt werden. Um Lipid Netzwerke zu vervollständigen, wird LINEX² als Preprint vorgestellt. Dafür nutzt die Software kuratierte Lipid Reaktionen aus öffentlichen Datenbanken. Auch wird ein Algorithmus präsentiert, um enzymatische Dysregulierung vorherzusagen.

Schlussendlich werden zwei Beiträge zur computergestützten medikamentösen Neu Indikation diskutiert. Vorhersage von Kandidaten zur medikamentösen Neu Indikation einer Krankheit ist der letzte Schritt einer systemmedizinischen Analyse.

Zusammengefasst werden in dieser Dissertation neue Algorithmen für die Subtypisie-

rung, der funktionellen Analyse und der medikamentösen Neu Indikation vorgestellt. Dies kann es ermöglichen, die Lipidomik als zusätzliche molekulare Dimension zur Systemmedizin hinzuzufügen. Alle Methoden wurden entwickelt mit Hinblick auf Nutzbarkeit und Interpretierbarkeit von komplexen molekularen Daten in der klinischen Forschung.

Publication Record

- TD Rose et al. "MoS_{Bi}: Automated signature mining for molecular stratification and subtyping" **Proceedings of the National Academy of Sciences** 2022, 119 (16), e2118210119; doi: 10.1073/pnas.2118210119
- N Köhler, M Höring, B Czepukojc, TD Rose et al., "Kupffer cells are protective in alcoholic steatosis" **Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease** 2022, 166398; doi: 10.1016/j.bbadis.2022.166398
- O Vvedenskaya* and TD Rose* et al. "Non-alcoholic fatty liver disease Stratification by Liver Lipidomics" **Journal of Lipid Research** 2021, 62, 100104; doi: 10.1016/j.jlcr.2021.100104
- N Köhler* and TD Rose* et al. "Investigating Global Lipidome Alterations with the Lipid Network Explorer" **Metabolites** 2021, 11 (8), 488; doi: 10.3390/metabo11080488
- G Galindez* and J Matschinske* and S Sadegh* and TD Rose* and M Salgado-Albarrán* and J Späth* et al. "Lessons from the COVID-19 pandemic for advancing computational drug repurposing strategies" **Nature Computational Science** 2021, 1, 33-41; doi: 10.1038/s43588-020-00007-6
- S Sadegh, J Matschinske, DB Blumenthal, G Galindez, T Kacprowski, M List, R Nasirigerdeh, M Oubounyt, A Pichlmair, TD Rose et al. "Exploring the SARS-CoV-2 virus-host-drug interactome for drug repurposing" **Nature Communications** 2020, 11, 3518; doi: 10.1038/s41467-020-17189-2

*These authors contributed equally

Contents

Acknowledgments	ii
Abstract	iii
Kurzfassung	iv
Publication Record	vi
1. Motivation	1
2. Introduction	4
2.1. Metabolism	5
2.1.1. Example: metabolism in cancer	10
2.2. Lipids	11
2.2.1. Lipid metabolism	13
2.2.2. Functions of lipids in health and disease	17
2.3. Lipidomics	18
2.4. Other omics technologies	25
2.5. Computational Biology	27
2.5.1. Systems biology & systems medicine	27
2.5.2. Data-driven approaches	28
2.5.3. Prior-knowledge integrating approaches	31
2.5.4. Computational lipidomics	33
2.5.5. Drug repurposing	34
2.5.6. Challenges in computational biology	34
2.6. Objective	35
3. Methods	37
3.1. An Ensemble biclustering method for disease subtyping	37
3.2. Creation of rule-based lipid networks for a functional lipidome interpretation	38
3.3. Deriving hypothesis for enzymatic dysregulation	39
4. Publication summaries	43
4.1. MoSBI: Automated signature mining for molecular stratification and subtyping	43
4.2. Nonalcoholic fatty liver disease stratification by liver lipidomics	45
4.3. Investigating Global Lipidome Alterations with the Lipid Network Explorer .	47

5. Unpublished Results	49
5.1. Lipid network and moiety analysis for revealing enzymatic dysregulation and mechanistic alterations from lipidomics data	49
6. Discussion	53
7. Conclusion	63
A. Appendix	66
A.1. MoSBI: Automated signature mining for molecular stratification and subtyping	66
A.2. Nonalcoholic fatty liver disease stratification by liver lipidomics	77
A.3. Investigating Global Lipidome Alterations with the Lipid Network Explorer .	92
A.4. Preprint: Lipid network and moiety analysis for revealing enzymatic dysregulation and mechanistic alterations from lipidomics data	112
A.5. Co-author contributions	132
A.5.1. Exploring the SARS-CoV-2 virus-host-drug interactome for drug repurposing	132
A.5.2. Lessons from the COVID-19 pandemic for advancing computational drug repurposing strategies	133
List of Figures	134
Acronyms	136
Bibliography	139

1. Motivation

Metabolic diseases have become highly prevalent in modern society. For example, around 25% of the world's population is affected by non-alcoholic fatty liver disease (NAFLD). An even higher prevalence is among obese individuals [1].

In a healthy state, cells maintain a molecular balance of reactions and interacting biomolecules to adapt to changing conditions. This is known as homeostasis [2]. In metabolic diseases, the metabolism is especially disturbed, and therefore the focus for understanding such diseases. The most common metabolic disease is obesity. In Germany, approximately 12.9% of the population is obese, which puts affected people at risk of coronary heart disease or diabetes mellitus [3]. Germany's estimated population with diabetes mellitus is between 7.2 and 9.8% [4]. Obesity is also a risk factor for cardiovascular diseases (CVD), which is the leading cause of death worldwide [5]. non-alcoholic fatty liver disease (NAFLD) is strongly associated with diabetes and obesity, and around 28% of patients develop liver damage through fibrosis and steatohepatitis [6]. These diseases are also a high economic burden. For CVD these costs in the US are estimated at around \$320 billion [7]. Diseases do not only result in deaths and hospital stays but cause a reduced quality of life for affected individuals [8]. NAFLD results in estimated costs of €4.33 billion in Germany and \$103 billion in the United States in the year 2016 [9].

NAFLD is characterized by an excessive accumulation of lipids in the liver. Lipids are a highly diverse class of metabolites. They are essential for long-term energy storage and are the main constituents of cellular membranes. This enables cells to separate from the environment and create compartments with distinct functions. Lipids are also crucial for a variety of metabolic and signaling processes. The increasing prevalence of metabolic diseases requires a detailed understanding of the lipidome on the structural level to understand their cellular functions and regulation. This can be enabled by mass spectrometry-based lipidomics protocols. Lipids are also becoming more critical for precise drug delivery [10]. Just recently, they received attention for COVID-19 mRNA vaccines. The mRNA is transported in a lipid vesicle that can be taken up by cells.

Effectively treating diseases requires understanding the altered molecular processes affected by the disease. The availability of high-resolution molecular data is crucial to understanding diseases, but not sufficient. Generating clinically relevant knowledge from heterogeneous molecular data is challenging and requires computational algorithms [11]. These efforts fall under the umbrella of systems medicine, where the functions of complex biological systems are studied in health and disease [12]. Systems medicine can provide a better disease understanding and offers new opportunities for medicine to be predictive and personalized [12]. With computational methods, diseases can be subtyped [13], molecular disease mechanisms identified [14], or biomarkers predicted [15]. In particular, subtyping

aims at unraveling molecular differences within a disease cohort. This can promote more individualized treatments of patients [16]. A disease is divided by stages, risk factors, or a molecular mechanism during subtyping. Deciphering such differences within a patient cohort can enable precise treatments for specific disease subtypes. Disease mechanisms can promote the selection and development of therapeutics tailored towards a specific disease manifestation [17]. For example, lipid enzymes and their regulating pathways can be potential drug targets if a mechanistic connection to a disease has been identified. Biomarkers can provide early-stage information for the prognosis of a disease or reveal risk factors. Personalized medicine diagnoses can also be more cost-effective [18], potentially reducing the health system's costs. Due to the importance of lipids in biological processes, they are essential for understanding and treating diseases [19]. Developing computational approaches to efficiently subtype diseases on a molecular basis and adding lipids as another dimension for biomedical research is the goal of this publication-based dissertation.

In the first publication, I present the computational subtyping method Molecular Signature identification using Biclustering (MoSBi) [20]. This work aims at making subtyping more accessible and interpretable. MoSBi combines the results from multiple algorithms to make predictions more robust. To offer value for clinical research, computational methods for disease subtyping need to promote the interpretation of subtype predictions. The visualization of such results was also an unsolved problem for this class of algorithms. In a systems medicine context, this can be beneficial to visualize results because the method allows additional health confounders to be incorporated into the visualization. By overcoming hurdles of accessibility, the approach makes computational disease subtyping more usable for clinical research and molecular data, where stratification is required.

In the second publication, I utilized liver lipidomics data of a large patient cohort with different stages of NAFLD to reveal disease subtypes, and potential biomarkers [21] (in the following, referred to as NAFLD stratification publication). This was not only a perfect opportunity to apply MoSBi to find novel disease subtypes but is of great importance for gaining more insights into the most common liver disorder. Finding markers for NAFLD and predicting the progression is of enormous interest to public health. Markers that resemble the disease progression could enable early detection of the disease and provide more insights into the molecular status of the disease for treatment decisions. From a biological perspective, lipidome alterations in the liver can be investigated to quantify changes in the abundance and composition of lipids during disease progression.

Extracting disease mechanisms requires a functional understanding of molecular biological interactions. Mass spectrometry has enabled the detection of the molecular composition of lipids. However, large-scale interaction networks for lipids are not available. Therefore, the third publication presents the Lipid Network Explorer (LINEX) method [22]. The approach computes lipid networks and visualizes them with statistical measures to interpret lipidome alterations quantitatively. This approach is complemented by a network analysis method that can propose mechanistic hypotheses for enzymatic dysregulation from lipidomics data. This work is attached to the dissertation as a preprint [23]. These methods make it possible to interpret lipid data functionally and give it more relevance for clinical research. A functional

understanding of systematic lipid alterations has great potential for systems medicine.

The final step of assessing diseases from a systems medicine standpoint is the development of treatments. The development of new drugs can cost several billion dollars. Drug discovery is an especially difficult step during the development [24]. In contrast, identifying new targets for existing drugs can be a more cost-effective alternative [25]. This is called drug repurposing, and computational methods play a crucial role in predicting drug repurposing candidates. I co-authored two publications focusing on drug repurposing of COVID-19 [26, 27]. First, the development of an interaction network-based method to predict drug targets and candidates [27]. Second, in a review on drug repurposing strategies and the proposal of a standardized development cycle [26]. Drug repurposing is the last step in a systems medicine workflow that from disease subtyping, biomarker identification, and mechanism detection to find drug targets. It further shows the importance of computational methods to advance drug repurposing research.

Outline

In the following introduction, I will explain the biological and computational basics of the publications. In sections 2.1 & 2.2, the principles of the cellular metabolism and lipids are covered with examples of dysregulation in diseases. These are crucial to understanding the variety of existing lipids and the structure of the lipid metabolism, which serves as a basis for the work on creating lipid networks. This is followed by an introduction to the experimental methodologies for measuring lipids in biological samples. Knowing the experimental details provides the reader with the necessary knowledge for discussing future computational lipidome research opportunities and limitations. Next, I give an overview of other omics techniques. Finally, computational biology is introduced in section 2.5. The section gives an overview of types of computational research and shows how the previously introduced aspects of molecular biology are utilized to get a better understanding of the cellular machinery. It also gives an overview of the different types of computational biological research for systems biology and where the developed methods of this dissertation fit in.

Followed by the introduction is a summary of the developed computational methods and contributions to the presented publications. In the discussion, the general applicability and limitations of the results are described, and an integrated workflow is presented. I also describe the limitations of current lipidomics research and the implications for computational method development. In the end, the opportunities for computational methods for personalized medicine are discussed.

2. Introduction

Before introducing lipids and their metabolism, which are central to this dissertation, a general background in molecular biology is required. It will provide an overview of how lipids are embedded in the molecular biology of the cell and how they relate to other biomolecules.

The fundamental building blocks of a living cell are deoxyribonucleic acid (DNA), ribonucleic Acid (RNA), proteins, metabolites, and lipids. As famously formulated in the central dogma of molecular biology by Francis Crick [28], information flows from DNA over RNA to proteins. This view on molecular biology is partially outdated. Nowadays, it is known that epigenetic modifications such as methylations also influence the activity of genes [29] as well as metabolites and lipids can alter protein activity. Information is stored as a sequence of nucleic acids, known as genes on the DNA. The RNA polymerase enzyme can transcribe genes to RNA. RNAs that encode proteins are also known as messenger RNA (mRNA). The mRNA sequence can be translated into an amino acid chain using ribosomes, a complex molecule composed of ribosomal RNA and proteins. The amino acid chain then folds into a chemically and thermodynamically optimal state [30], known as a protein. Proteins fulfill a variety of functions. They can serve as transcription factors, where they activate or inhibit gene transcription. Other proteins act as signaling molecules, forwarding sensory information that can activate or inhibit transcription factors. Some proteins build the cytoskeleton by creating polymer fibers, providing structural integrity for cells and highways for intracellular transport proteins. Enzymes, another subset of proteins, catalyze chemical reactions of metabolites or lipids. Metabolites are diverse organic molecules that provide energy for cellular processes, and they are the building blocks for all biomolecules, such as proteins or DNA. Lipids, also known as fats or oils, can store energy efficiently. Furthermore, lipids are the basis of cellular membranes that allow a cell to have compartments and separate itself from the environment.

An essential aspect of a living cell is that biomolecules dynamically interact. Genes regulate each other, metabolites are catalyzed by enzymes and regulate their activity, and proteins interact are used to forward information [31]. Figure 2.1 shows an overview of the information flow and interactions. Hence, changes or defects of only one gene can influence the whole system. Examples of such changes are defects in the synthesis of species sugar chains, so-called glycans, that increase mortality in the first years of life [32], somatic mutations resulting in cancer development [33], or obesity causing type 2 diabetes [34].

Crucial for cell functions is that biomolecules can often interact with multiple other molecules. An example of this is the phosphorylation of proteins. Over 40% of proteins have phosphorylations during their lifetime, created by kinases [35]. Kinases are also proteins that phosphorylate other proteins or metabolites. They are usually able to phosphorylate proteins at various phosphorylation sites. The same holds for transcription factors, which regulate the

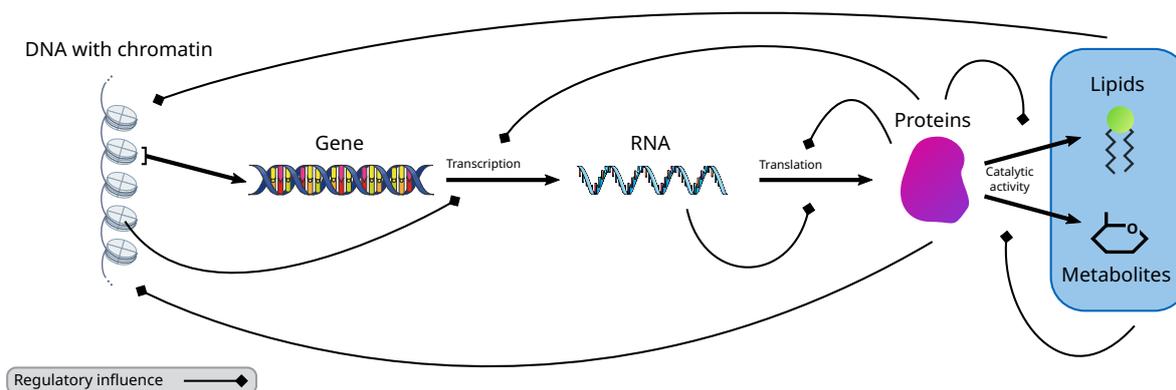


Figure 2.1.: Information flow and regulation between biomolecules. Genes are transcribed to RNA, which are then translated to proteins. Specific proteins are then able to catalyze metabolic reactions. All layers are interconnected and have regulatory influences by activating, inhibiting, or providing building blocks for other biomolecules.

License information: dna-nucleotides-ribbon and rna icon by Servier <https://smart.servier.com/> are licensed under CC-BY 3.0 Unported <https://creativecommons.org/licenses/by/3.0/>

expression of genes. They can bind to multiple sequence motifs and, therefore, influence the expression of many genes. Such Multi-specificity can also be observed in metabolic reactions. Many lipid enzymes catalyze metabolic reactions for lipids with different fatty acyls. While this can have drastic implications for diseases caused by a single gene or protein failure, it enables the molecular machinery to adapt to the state of the whole system by altering the activity of central biomolecules. Understanding the importance of multi-specificity is also crucial for my publications on lipid networks. It will be more explicitly elaborated for the lipid metabolism in section 2.2.1.

Before introducing lipids and their functions, I will introduce the basics of metabolism. Energy metabolism is crucial for synthesizing lipids, and specific metabolites are the precursors of lipids. As an example of systematic alterations in diseases, the well-studied metabolism in cancer is explained.

2.1. Metabolism

Metabolism is commonly associated with providing energy for all processes in a cell. But metabolism is more than just the production of energy from food. Energy storage is done with metabolites, separation of cells from the environment via membranes, and in signaling processes, metabolites are involved. Metabolites are also the building blocks for all complex biomolecules.

Metabolites are small molecules compared to polymers such as proteins with hundreds

of amino acids. In a cell, they are converted to other metabolites by metabolic reactions. Most reactions do not occur spontaneously under the conditions in a cell. This is due to the activation energy of every reaction that is required for every reaction. The same reason we need sparks to kindle a fire. Enzymes fulfill this purpose for metabolic reactions. They bind metabolites and stabilize reaction transitions through their active pockets [36]. This biochemical principle was already described by Linus Pauling in 1946 [37]. While many enzymes are multispecific, meaning they can catalyze multiple reactions, various enzymes are necessary to catalyze all metabolic reactions in a cell. The model "Recon 2.2", a manually curated collection of known metabolic reactions, contains 5324 metabolites, 7785 reactions, and 1675 associated genes [38] (the genes mainly encode for enzymes). And this does not include thousands of lipid species with their corresponding reactions. This is due to the complexity of the lipid metabolism (explained in section 2.2.1), but also because of a general lack of integration of lipidomics into bioinformatics workflows, which is one of the topics that are addressed in this thesis.

One of the most critical energy sources is glucose. It comes mainly from plant-based food in the form of starch, a polymer of glucose molecules. Starch is broken down in the saliva of the mouth by amylases, a class of enzymes that are also excreted by the pancreas [39].

Before the enzymatic machinery in a cell can start working, metabolites need to enter the cell. Metabolite transporters are membrane-based transporters that can channel metabolites through membranes. One essential class of transporters is the GLUT family. They transport glucose through the membrane without needing external energy [40]. Therefore, they can only transport glucose along the osmotic gradient. Another class of transporters, the SLC5 family, couples glucose transport to a Na^+ gradient and can actively import glucose into a cell [41]. Transporters for many different metabolites exist (e.g. [42, 43, 44, 45]).

Once glucose is imported into cells, it can be broken down to release energy. This happens in the central carbon metabolism. In the following, I will introduce the main pathways of the central carbon metabolism to give an overview of metabolic principles for energy generation and biosynthesis.

Glycolysis

Glycolysis is one of the main pathways to supply energy and metabolites for biosynthesis [46]. It is evolutionary conserved and can be found in all domains of life. By binding a phosphate group, energy is generated from glucose by converting Adenosine diphosphate (ADP) to Adenosine triphosphate (ATP). The bound is highly energetic, and releasing or transferring the phosphate groups is an exergonic reaction. Exergonic reactions have a strongly negative free energy change, which is irreversible under normal conditions. Irreversibility is defined by an equilibrium of a reaction that lies firmly on the side of products. The glycolysis starts with the glucose sugar (Figure 2.2), consisting of 6 carbon atoms (C6). Kinase enzymes activate glucose twice by phosphorylation, which requires energy in the form of ATP. It is essential to consider that the pathway requires an initial energy investment before energy can be released. After this, the C6 sugar with two phosphorylations (Fructose 6-bisphosphate) is broken down into two C3 molecules Glyceraldehyde 3-phosphate (G3P)

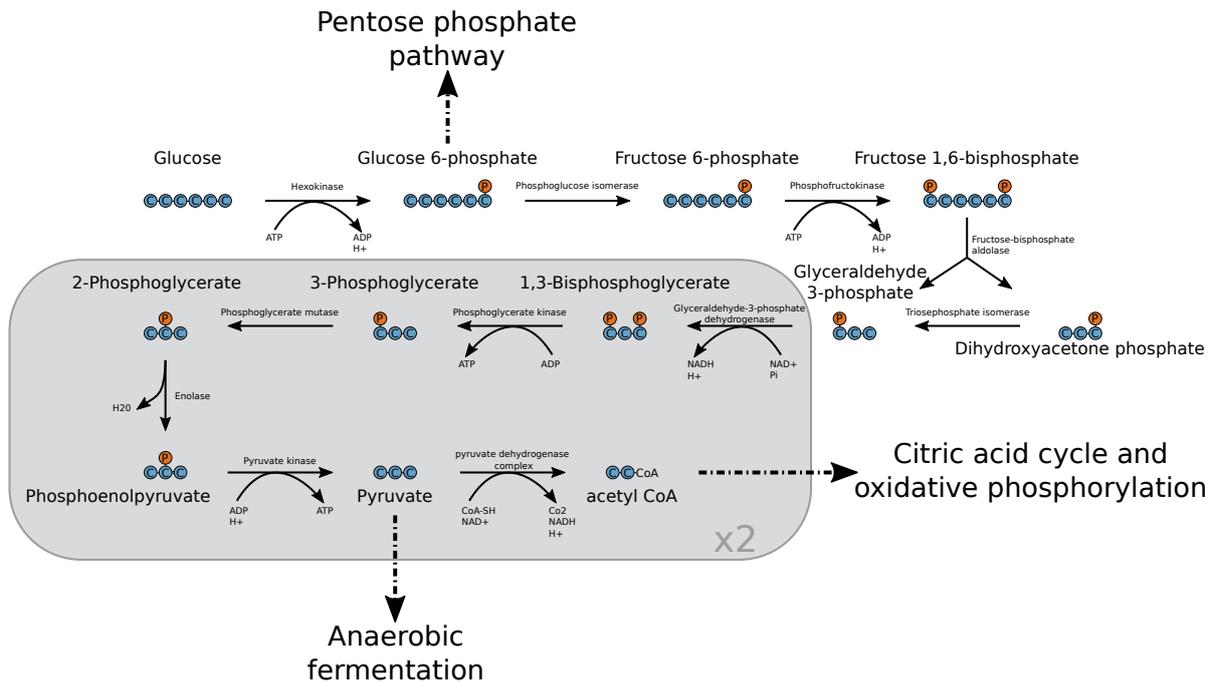


Figure 2.2.: Glycolysis pathway. C6 glucose is broken down into two C3 pyruvate molecules. In this process, ATP and NADH are generated. Abbreviations: Adenosine diphosphate (ADP), Adenosine triphosphate (ATP), oxidized Nicotinamide adenine dinucleotide (NAD^+), reduced Nicotinamide adenine dinucleotide (NADH), Coenzyme A (CoA), Proton (H^+).

and Dihydroxyacetone phosphate (Figure 2.2). An isomerase converts the latter to G3P. This is followed by two crucial reactions for energy generation and shows the principle of metabolic energy generation. G3P is oxidized at the first carbon atom, and a free phosphate is binding, producing 1,3-Bisphosphoglycerate. Electrons from the oxidized G3P are transferred to oxidized Nicotinamide adenine dinucleotide (NAD^+) resulting in reduced Nicotinamide adenine dinucleotide (NADH). The phosphoglycerate kinase then phosphorylates ADP using the previously bound phosphate from 1,3-Bisphosphoglycerate. Finally, one more ATP per molecule is produced, yielding Pyruvate. In total, 2 ATP have to be invested for one glucose molecule, and 4 ATP and one NADH are produced.

In an aerobic environment, the additional electrons of NADH can be forwarded to oxygen to produce more energy. However, pyruvate is reduced when no oxygen is available, or cells cannot perform oxidative phosphorylation. This can result in the production of either ethanol and CO_2 or lactate. The recovered NAD^+ can then be reused for oxidation of G3P.

Citric acid cycle

With the availability of oxidizing agents, more energy can be generated from the complete oxidation of pyruvate. Furthermore, it is crucial for the efficient energy extraction from

proteins and lipids [47]. Before pyruvate can enter the citric acid cycle, also known as the tricarboxylic acid (TCA) cycle, it is further oxidized and bound to Coenzyme A (CoA). In this reaction, CO_2 is released and NAD^+ reduced and the product of the reaction is acetyl-CoA (Figure 2.2). The pathway is called a cycle because, in the first step, the acetyl group is binding to oxaloacetate, which is also the end product and can be reused to start another cycle. Oxaloacetate can also be synthesized from pyruvate by a reaction catalyzed by the pyruvate carboxylase. It creates a bond between pyruvate and carbonic acid under ATP hydrolysis. The TCA cycle consists of 9 reactions with the following summed-up reaction equation:



The synthesized Guanosine triphosphate (GTP) can also be used to phosphorylate ADP. Since the main currency for energy in cells is ATP, the energy saved in the reduced NAD^+ and flavin adenine dinucleotide (FAD) is then used to synthesize ATP. This is happening in the oxidative phosphorylation.

Oxidative Phosphorylation

ATP is generated in the oxidative phosphorylation by a proton translocation through a membrane, driven by a proton gradient [48]. An excellent summary of the history of the findings and debate on the pathway can be found in the publication of Nath et al. [49]. Creating and maintaining a proton gradient requires a particular environment. Therefore, the TCA cycle and oxidative phosphorylation occur in the matrix of mitochondria. Mitochondria have an outer and inner membrane. The inner membrane space is called the matrix, and the space in between the two membranes is the intermembrane space. The pathway consists of a series of protein complexes at the inner mitochondrial membrane. They are electron acceptors from NADH and FADH and forward them from one complex to the next. In this process, they pump protons from the matrix to the intermembrane space and forward them to O_2 synthesizing H_2O . As a result, a proton gradient across the inner membrane is generated, with a high concentration of protons in the intermembrane space and a low concentration in the matrix. The ATP Synthase protein complex utilizes this gradient by allowing protons to pass through to the matrix and using the motion to synthesize ATP from ADP and free phosphate.

The pathways described so far are all catabolic, meaning metabolites are broken down to produce energy. However, anabolism, the synthesis of new metabolites as building blocks for biomolecules, is another important aspect of metabolism. Three significant pathways that contribute to anabolism are discussed below.

Pentose phosphate pathway

The Pentose phosphate pathway (PPP) fulfills several important functions. It provides reduced nicotinamide adenine dinucleotide phosphate (NADPH) for the fatty acid synthesis (more on that in Section 2.2.1), C5 sugars for nucleic acid synthesis, and the C4 sugar Erythrose 4-phosphate as an amino acid precursor [50].

The PPP can be divided into oxidative and non-oxidative parts. In the oxidative part, glucose 6-phosphate is oxidized to Ribulose 5-phosphate in three reactions under release of CO₂ and two NADPH. This is followed by the non-oxidative part, where the reactions are reversible. There are two versions of this non-oxidative part. The canonical (of F-type) pathway and the L-type pathway. In the canonical version, two enzymes catalyze all reactions. First is the transketolase that transfers a C2 carbon fragment from one phospho-sugar to another, and the second is the transaldolase that transfers a C3 fragment from one phospho-sugar to another. They can synthesize C3, C4, C6, and C7 phospho-sugars from two initial C5 phospho-sugars. The L-type pathway contains not only the transketolase but an additional aldolase that combines a C3 and C5 molecule to Octulose 1,8-bisphosphate. With a phosphotransferase, C7 bisphosphate molecules can be synthesized. The same end products can be synthesized as in the canonical pathway. Since C7 and C8 mono- and bisphosphate sugars could be observed [51], and the transaldolase can be found in humans [52] a combination of proposed versions of the pathway is likely.

Because of the reversibility of the reactions in the PPP, it can, on the one hand, supply the glycolysis with intermediates for the energy metabolism, and on the other hand, glycolysis can provide the precursors for the pathway [53]. This makes it a powerful buffer for stressful situations.

Gluconeogenesis

The pathway that synthesizes glucose is called gluconeogenesis. It plays a vital role in mammals during starvation [54]. Gluconeogenesis occurs predominantly in the liver when glycogen stores are depleted. As precursors for glucose metabolites such as lactate, pyruvate, glycerol, and certain amino acids can be used [55]. The precursors are converted to Phosphoenolpyruvate (see Figure 2.2). From there, reverse glycolysis is possible. The pathway shares many enzymes with the glycolysis that are reversed based on substrate and product concentrations [56]. For example, there are differences in the enzyme glucose-6-phosphatase, converting glucose 6-phosphate into glucose, instead of the hexokinase that catalyzes the opposite direction in the glycolysis.

Gluconeogenesis is an excellent example of cooperation in multicellular organisms, where liver cells invest energy to synthesize the more easily transportable glucose to provide energy for other organs. Since it is an important pathway, it is also regulated by the hormone insulin [54].

2.1.1. Example: metabolism in cancer

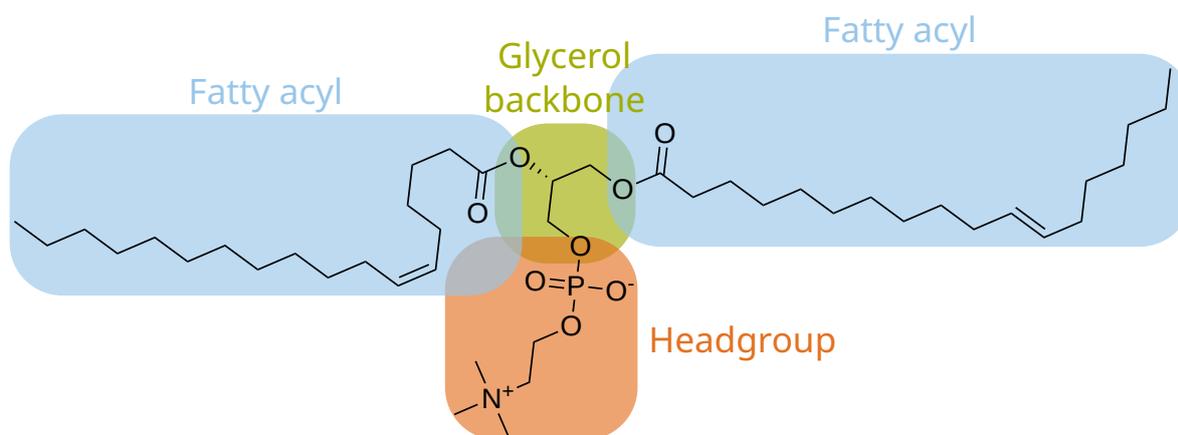
Because of metabolism's central role in providing energy and biosynthesis of all cellular parts, metabolic reprogramming is often observed in diseases. This phenomenon is incredibly well studied in cancer and known as a hallmark of cancer [57]. Oncogenes do not only alter the metabolism in general but also remodel the lipid metabolism, for example, via the EFGR [58] or KRAS [59] signaling proteins. Understanding the molecular basis of cancer development is of immense interest to systems medicine research. Cancer is a very heterogeneous disease, and many subtypes can be identified. In the MoSbi publication, I worked with several cancer-related datasets to evaluate the method for subtyping complex diseases. Here, I will explain the metabolism in cancer as an example of the adaption of metabolic pathways in diseases.

Cancer cells are growing and dividing very fast compared to normal tissue. This is also reflected in an altered metabolism. The effect is famously known as the Warburg effect, named after the scientist Otto Warburg, who discovered it in the 1920s [60]. It describes the principal activity of aerobic glycolysis in cancer cells. Oxidative phosphorylation and the TCA cycle, which provide the most energy, are down-regulated compared to the glycolytic activity. There are several theories for this behavior. The two leading theories are oxidative stress [61] and precursor demands [62]. First, fast-growing cells have not only a high demand for energy but also for precursors of biomolecules [62]. They are required to synthesize other biomolecules required for cell division and growth. For example, the cancer cells facilitate metabolic flux from the glycolysis into the PPP [63]. This provides ribose sugars that are necessary for nucleotide synthesis and NADPH for fatty acid synthesis [64] that are building blocks for the membranes. Cancer cells also excrete specific metabolites, such as lactate or glutamate. Since this is also observed in oxygen-rich environments, lactate excretion is not fully understood [62]. Glutamate excretion is believed to stimulate nucleotide production [65].

The second theory for increased glycolysis is an often low oxygen environment in cancer tissue. The growth of blood vessels cannot always keep up with the fast growth of cancer. Limited oxygen supply paired with oxidative phosphorylation can result in oxidative stress [61, 66]. Oxidative stress is defined as the occurrence of reactive oxygen species that result in tissue damage [67]. They can occur during the oxidative phosphorylation from electron leakage [68]. At high concentrations, the cell's antioxidant capacity cannot remove them sufficiently anymore. Using just glycolysis as the primary energy supply can reduce this stress.

Regulated is this metabolic reprogramming by the differential activity of central signaling pathways including mTOR, PI3K, and AKT [66], but also circular non-coding RNAs have been found as regulators of glycolytic enzymes in cancer [69]. These differences in metabolic pathway activity between healthy and cancer tissue make it a potential target for therapeutics inhibiting such anabolic activity [70].

In this section, I presented different metabolic pathways and showed that pathway definitions are often blurry and that reactions can be part of multiple pathways. Furthermore, the



Phosphatidylcholine (18:1(11E)/18:1(6Z))

Figure 2.3.: Structure of a Phosphatidylcholine (PC) lipid. The lipid consists of a glycerol backbone (green), a phospho-choline headgroup (orange), and two fatty acyls (blue). Lipid structure received from the SwissLipids database (identifier: SLM:000008040).

cancer example showed that there are often complex relationships between metabolic functions and regulations that are not always well understood. This is already a first indication of why working with networks covering all known metabolism parts can be advantageous.

In the next section, I will introduce lipids, a versatile set of molecules essential for long-term energy storage, membrane integrity, and signaling cascades. Their metabolism is highly entangled with the previously introduced pathways.

2.2. Lipids

Lipids are molecules that are characterized by high hydrophobicity. They are commonly known as fats or oils. Due to their heterogeneity, lipids can be divided into several groups and classes. The most common group comprises glycerolipids consisting of a glycerol backbone, one to three fatty acyls, and sometimes a headgroup. Figure 2.3 shows an example of such a lipid. Fatty acyls (or fatty acids, when not bound via an ester bond) are aliphatic chains that can have one or several double bonds. In the example, two fatty acyls of carbon length 18 with one double bond can be seen. The right fatty acyl has the double bond at the 11th position in a *trans* configuration (E), counting from the first carbon of the chain. In contrast, the left fatty acid has the double bond at the 6th position in a *cis* (Z) configuration. The lipid headgroup in Figure 2.3 is a phosphocholine and therefore belongs to the Phosphatidylcholine (PC) lipid class. Two other important groups are sterol-based lipids and sphingolipids. Both can also have bonds with fatty acyls and headgroups.

Several nomenclatures for lipid names are available. Lipid-specific short-nomenclatures are

more commonly used than the official naming of molecules by the International Union of Pure and Applied Chemistry (IUPAC). Most common are the LIPID MAPS [71] or the Liebisch shorthand nomenclature [72]. For lipid fragments, a dedicated lipid nomenclature is available, as used by the ALEX123 software [73]. In this dissertation the LIPID MAPS nomenclature is used, which defines the lipid in Figure 2.3 as a PC(18:1(11E)/18:1(6Z)). Different resolutions for lipid names are possible, depending on the identification level of the lipid (more on that in section 2.3). The naming of PC(18:1(11E)/18:1(6Z)) follows the highest resolution. It defines the stereospecifically numbered (sn) position of each fatty acid at the glycerol (18:1(11E) at *sn-1* and 18:1(6Z) at *sn-2*), as well as the number of carbon atoms and double bond positions. PC(18:1/18:1) describes the lipid with fatty acyl positions but without double bond positions. PC(18:1_18:1) is the lipid without knowing the exact fatty acyl positions. This and all previous descriptions are called "molecular species" since they resolve the molecular compositions of the fatty acyls. Finally, the sum formula PC(36:2) only describes the sum of fatty acyl lengths and the combined number of double bonds.

Lipids are very amphiphilic molecules, meaning they have hydrophobic (fatty acyls) and hydrophilic (head group) properties (Figure 2.3). Therefore, they are not or only partially soluble in water. Lipids with polar headgroups, such as PCs can form bilayers or micelles in hydrophilic environments such as cells [74] (examples in Figure 2.4). Micelles with an outer layer of polar lipids can be used to store or transport non-polar lipids, e.g., Triacylglycerol (TG) species that consist of a glycerol backbone with three fatty acyls (Figure 2.5) and therefore do not have a polar headgroup.

Lipid bilayers are the central part of cellular membranes, enabling a separation between the intracellular space and the environment or between cellular compartments. It also prevents polar molecules from passing through the membrane. Dedicated membrane transporters are available, as explained for glucose transporters in Section 2.1. The composition of biological membranes is commonly described by the fluid mosaic model [75]. It defines the membrane as a mixture of proteins and lipids forming a bilayer, where biomolecules are diffusing freely [75]. This can be compared to a two-dimensional fluid solution of diffusing lipids and proteins along the membrane. The membrane's fluidity has to be maintained through temperature changes, especially for organisms that don't maintain their temperature. Lipid class and fatty acid composition of membrane lipids can be adapted to achieve this. Highly unsaturated fatty acyls (many double bonds) induce bends in the carbon chain (as can be seen in Figure 2.3), which results in higher distances between lipids in the membrane (higher fluidity). The fluid mosaic model has been challenged by the lipid raft model [76]. It describes the membrane as a self-organized structure that emerges from associations between different lipids and proteins [77]. Based on this model, the bilayer is not a passive solvent but provides local accumulations of distinct lipids and proteins [78]. These can facilitate environments for specific protein or lipid functions, such as signaling processes.

Lipids are very effective for storing energy. With lipid synthesis, the cell can store energy long-term and adapt its function. During degradation, energy can be released. Lipids can also be converted into each other to fulfill different functions. In the following section, I

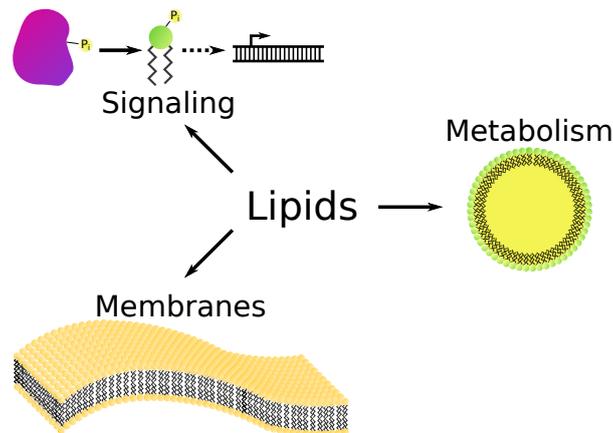


Figure 2.4.: Important functions of lipids in the cell: As signaling molecules, long-term energy storage, and main membrane component.

will explain these crucial parts of lipid metabolism. This is crucial to understanding lipid functions and roles in different parts of the metabolism and signaling pathways.

2.2.1. Lipid metabolism

Lipid metabolism can be divided into the metabolism of complex lipids and fatty acids. However, complex lipids require fatty acids for their assembly. Fatty acids also carry the majority of energy that can be stored in lipids. The energy can also be released in a pathway known as β -oxidation.

Fatty acid synthesis

The synthesis of saturated fatty acids is an iterative process that involves acetyl-CoA, the end product of glycolysis (Figure 2.2) and NADPH as primary substrates. NADPH as a reducing equivalent is primarily synthesized in the PPP. In each iteration, an acyl chain is extended by two carbon atoms [79]. The iteration starts with the carboxylation of an acetyl-CoA to malonyl-CoA. Then, under the release of CO_2 , a ketoacyl is created. In three further reactions, the ketoacyl is reduced, dehydrated, and again reduced by oxidizing two NADPHs. The result of the iteration is an extended acyl-CoA chain. Odd-chain fatty acids, such as 15:0 or 17:0 can be synthesized by starting with Propionyl-CoA (a C3 chain) instead of acetyl-CoA [80].

Elongation of saturated fatty acids is only one part of the fatty acid synthesis. As already mentioned, fatty acids can have double bonds in the acyl chain, affecting their shape and functional properties. Dehydrogenases catalyze the desaturation of fatty acids (i.e., introducing double bonds). These enzymes oxidize the acyl chain by reducing FAD. However, not all desaturations can be catalyzed by human enzymes. Desaturation from oleic acid (C18:1- Δ 9) to linoleic acid (C18:2- ω 6), and from there to α -linoelic acid (C18:3- ω 3) is only synthesized in plants [81]. Therefore, they are essential fatty acids that come from the diet. These can

be utilized as substrates for further elongations and desaturations in humans to synthesize e.g. arachidonic acid (C_{20:4- ω 6}) or docosahexaenoic (C_{22:6- ω 3}). Since the focus of this thesis lies on the metabolism of complex lipids, I will not discuss all details of fatty acid synthesis, but interested readers can get an excellent overview of the different enzymes involved in the pathway and their regulation in the publication of Guillou et al. [81].

β -oxidation

Another essential part of fatty acid metabolism is catabolism. Fatty acids store energy at a high density. They are oxidized in the mitochondria to make the energy available for other cellular pathways. The pathway is called β -oxidation since the third carbon of the acyl chain (the β -position carbon) is oxidized in the process. Just like fatty acid synthesis, β -oxidation is an iterative pathway, where in each iteration acetyl-CoA is cleaved-off the acyl-chain while reducing one NAD⁺, one FAD, and consuming one H₂O molecule [82]. The resulting acetyl-CoA can then enter the TCA cycle and used to synthesize ATP. Odd chain fatty acids are oxidized similarly, with the final product being Propionyl-CoA. This metabolite can be transformed to Succinyl-CoA and also enter the TCA cycle

Oxidation of unsaturated fatty acids is also possible via this pathway. This requires two additional enzymes. An enoyl-isomerase that moves a double bond from the Δ -3 to the Δ -2 position, and a reductase that reduces the double bond [82].

Complex lipid metabolism

Fatty acids are an integral part of complex lipids (Example in Figure 2.3). Therefore, they are required for the first step of assembling complex lipids. The synthesis of glycerolipids begins with a reaction catalyzed by the glycerolphosphate acyltransferase, which creates an ester bond between a G3P and a fatty acyl-CoA [83]. The resulting lipid is a Lyso phosphatidic acid (LPA) (Figure 2.5). An LPA acyltransferase can then attach another fatty acid to the lipid, thus synthesizing a Phosphatidic acid (PA). After hydrolyzing the phosphate headgroup, the class Diacylglycerol (DG) lipid is formed. Another ester bond can be created between a DG and a fatty acid, which is catalyzed by a DG acyltransferase, converting the DG to a TG [84]. TGs are typical storage lipids in adipose tissue. Enzymes to catalyze the reverse reaction also exist. Lipases hydrolyse TG to DG, Monoacylglycerol (MG) and fatty acids [85]. Also a DG can be converted back into a PA species, by phosphorylation at the glycerol backbone (catalyzed by the DG kinase) [86].

PA also serves as a precursor for other lipid classes, such as Phosphatidylinositol (PI). The Cytidine diphosphate (CDP)-DG synthase uses PA and Cytidine triphosphate as substrates to synthesize CDP-DG. In the next step, the PI synthase binds inositol to the phosphate group of CDP-DG, while hydrolysing Cytidine monophosphate, creating a PI lipid [87]. PI can be further phosphorylated at various positions of the inositol ring [86]. This is especially important in signaling cascades.

Other classes of glycerophospholipids, such as Phosphatidylethanolamine (PE) and PC require DG as a metabolic precursor. PE is synthesized by DG reacting with CDP-ethanolamine,

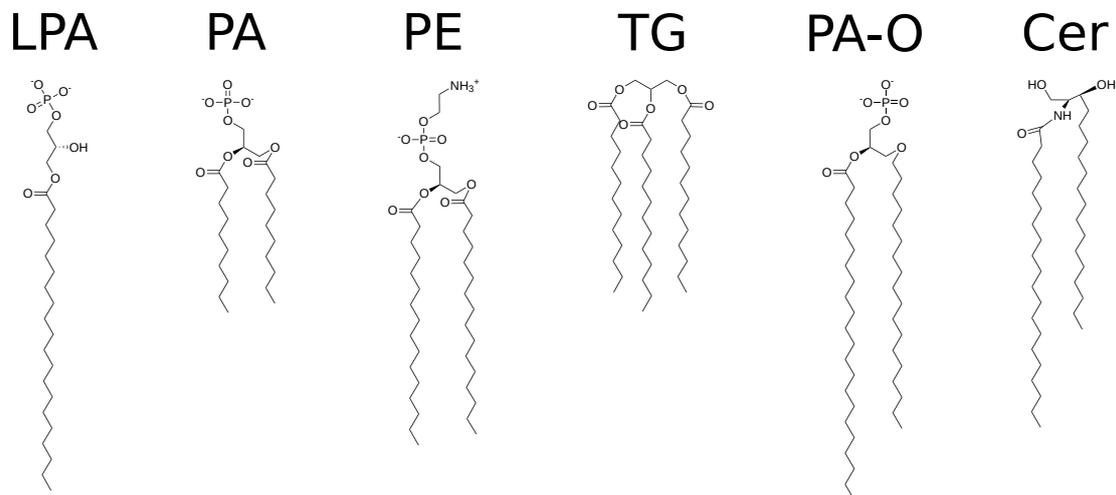


Figure 2.5.: Chemical structures of common lipid classes. Abbreviations: Lyso phosphatidic acid (LPA), Phosphatidic acid (PA), Phosphatidylethanolamine (PE), Diacylglycerol (DG), Triacylglycerol (TG), and Ether Phosphatidate (PA-O).

catalyzed by the choline/ethanolamine phosphotransferase (CEPT) [88]. Alternatively, it can be created by decarboxylation of Phosphatidylserine (PS). PC also can be synthesized by the CEPT, which catalyzes a reaction of DG and CDP-choline. This de-novo synthesis of PE & PC is also known as the Kennedy pathway [89, 90]. Another way is by three iterative methylations of a PE (phosphatidylethanolamine N-methyltransferase) [88].

Not all glycerophospholipids have fatty acids attached by ester bonds. Ether lipids can also be commonly found in cellular membranes. An Ether Phosphatidate (PA-O) can be seen in Figure 2.5. The ether lipid biosynthesis pathway starts with introducing an ester bond between a Glycerone-phosphate and a fatty acid [91]. This bond is then oxidized by the alkylglycerone-phosphate synthase, resulting in a Ether Lyso Phosphatidate (LPA-O). Several pathways can bind more fatty acids or create headgroup modifications [91]. These pathways are often the same as for the non-ether counterparts.

Sphingolipids are another important class of lipids that are not based on a glycerol backbone. They consist of a sphingoid backbone, where an additional fatty acid and headgroup can be attached. An example is Ceramide (Cer), a subclass of sphingolipids (Figure 2.5). They are synthesized by a reaction of a fatty acid with serine (Serine-palmitoyl transferase), creating a ketosphinganine, which is then reduced to a sphinganine [92]. The sphinganine can bind a fatty acid, producing a Cer. By adding or modifying headgroups of the Cer, other lipid classes can be synthesized (e.g., Sphingomyelin (SM)). It was recently discovered that the pathway of synthesizing sphingolipids in bacteria varies in the order or reactions compared to eukaryotes [93]. In this pathway, ketosphinganine first binds a fatty acid before reducing it to Cer.

So far, I have introduced the metabolism of complex lipids only on the lipid class level and not for specific molecular species. This is because lipid enzymes are highly multi-specific, a crucial aspect that was the basis for the lipid network methods presented in this dissertation [22, 23]. I will explain this using two lipid enzyme families as examples. The Phospholipase A₂ (PLA₂) and Membrane-bound O-acyltransferase (MBOAT) families. MBOAT enzymes attach a activated fatty acid (bound to CoA) to lysophospholipids. Gijón et al. [94] investigated the preference for different MBOAT enzymes towards phospholipid classes and fatty acid substrates. MBOAT1 showed a strong preference toward C18 fatty acids, especially for PS. Another family member, MBOAT7 preferred arachidonic acid and PI lipids. MBOAT2 & 5 were less specific with substrates and catalyzed the reaction over a broader range of phospholipids. Hayashi et al. [95] performed a similar study for PLA₂ enzymes. This class of enzymes hydrolyses fatty acyls from glycerophospholipids. They also showed that different PLA₂ members can catalyze the reaction over various fatty acyl and lipid class substrates. With e.g. cPLA₂ being more specific and sPLA₂ less specific. The two studies provide great insight into how lipid enzymes act on the lipidome. Instead of changing the abundances of specific lipid molecular species, they act on multiple lipids and alter the lipidome's lipid classes and fatty acids distribution. PLA₂ and MBOAT are opposing parts of the Land's cycle that is remodeling the lipidome [96].

This multispecificity shows how difficult it is to find all associations between individual lipid species and enzymes. As I will explain in section 2.3 about the experimental measurements of lipids, many lipids have only been identified recently, and more will be in the future with technological advances. Filling these gaps computationally and creating comprehensive reaction networks for lipids was a central part of this dissertation.

Lipid transport

Lipid metabolism happens not only within a cell; lipid transport across organs is also crucial for organisms. The transport of lipids is not as easy as for other water-soluble metabolites. Their hydrophobic properties lead to the formation of micelles or double layers. Therefore, dedicated transport mechanisms have been developed. We can differentiate between vesicular and non-vesicular. In non-vesicular pathways, lipid-transfer proteins carry individual lipids through the cell [97]. However, in this way, they can move only small amounts of lipids. Especially in multicellular organisms, where stored fats in the form of TG must be transported through the bloodstream from adipose tissue to the liver and vice versa, vesicular lipid transport is utilized. A significant class of lipid vesicles is lipoproteins. They consist of a core of TGs, and cholesteryl esters, covered by a monolayer of phospholipids and proteins [98]. Lipoproteins are characterized by size and weight, and are commonly grouped into five classes: Chylomicrons, very-low-density lipoprotein (VLDL), low-density lipoprotein (LDL), intermediate-density lipoprotein, (IDL) high-density lipoprotein (HDL). HDLs carry the lowest percentage of TG and cholesterol [98]. Their function is to transport lipids from peripheral (adipose) tissue to the liver for hydrolysis [99].

Different lipoproteins have been associated with cardiovascular diseases [100]. They are commonly investigated in blood diagnostics as health indicators. Cholesterol and lipoprotein

research was awarded the Nobel prize in medicine in 1964 and 1985 [101, 102]. However, current technologies make it possible to investigate lipids not only on the lipoprotein level but to identify and quantify different lipid species. Such fine-grained insights can bring more detailed insights into disease-related alterations in lipid metabolism.

2.2.2. Functions of lipids in health and disease

When thinking about the function of lipids, an important question is why such a variety of lipids exists. Cell membranes comprise predominantly of glycerophospholipids, in particular PC [103]. Long-term energy storage is done through TGs and cholesterol. But why are there so many glycerophospholipids and other lipid classes?

PI plays an important part in cellular signaling cascades [104]. It can be phosphorylated at the inositol ring. In this way, it participates, e.g., in the PI3K/Akt pathway [105] that is important for cell growth and proliferation. In this pathway, PI3K phosphorylates PI-bisphosphate, which enables Akt to bind with it, which then, in turn, can directly activate mTOR [106], a regulator of many processes, such as metabolism and mRNA translation.

Cardiolipins (CLs) are lipids consisting of two PAs connected via glycerol, bound to their phosphate headgroups. This gives them unique physicochemical properties [107]. For example, they can promote curvature of the membrane. For that reason, they can be found in the inner membrane of mitochondria, which is strongly curved.

Sphingolipids participate in apoptotic signaling pathways, and they mix poorly with Cholesterol and therefore tend to form microdomains [108]. As Globosides, which are sphingolipids with polysaccharide headgroups, they contribute to the extracellular matrix. By aggregating certain lipid classes in the membrane, cells can regulate membrane dynamics, such as expansion or vesicle transport [109]. Lipids can also support certain folding states of membrane-bound proteins and hence their function by stabilizing or destabilizing them [110]. In a recent publication, researchers found that a higher variability of lipids can confer robustness to environmental changes [111]. Also, fatty acids can fulfill essential functions. Arachidonic acid and its derivatives are involved in inflammation, pain, and fever processes [112]. They can also be found in the stratum corneum, the outer protective layer of human skin [113].

These are not all functions and pathways where lipids are involved. However, it shows the diversity and requirements for all the different lipids found in organisms. Because of their importance, it is also clear that lipids are involved in many diseases.

Lipids in diseases

Obesity is the most commonly known lipid-associated disease. Together with other factors, obesity can lead to metabolic syndrome. It is defined by an excessive flux of free fatty acids (FFA), a pro-inflammatory state, and induces a high risk for type-2 diabetes, as well as CVD [114]. A strong correlation between Cholesterol and CVD has been observed [115]. High concentrations of FFA can activate receptors that promote lipoprotein lipase activity [116]. This leads to TG release from VLDLs and results in Cholesterol rich HDLs. Furthermore,

FFAs are the main drivers of lipotoxicity [117]. Lipotoxicity describes the accumulation of lipids in non-adipose tissue that can lead to cellular dysfunction and cell death [118]. In the heart, FFA accumulation results in TG synthesis, as seen in CVD [118]. Also, insulin resistance of type-2 diabetes patients can be related to lipotoxicity [117].

In cancer, an altered lipid metabolism can also be observed. Oncogenic activation of the PI3K/Akt pathway results in lipid synthesis [19]. De-novo lipid synthesis is required for growth and proliferation [119]. Especially synthesis of saturated lipids has been observed [120]. But also, the breakdown of storage lipids to meet energy demands is accelerated. Therefore, liver cancer, predominantly hepatocellular carcinoma, drastically impacts the concentration of lipoproteins in the blood because of its importance in the lipoprotein metabolism [121]. Other examples are thyroid diseases, a hormone gland that also influences lipoprotein metabolism that can have an impact on HDL & LDL levels [122].

Other diseases where an altered lipid metabolism was shown are neurodegenerative diseases, such as Parkinson's and Alzheimer's [123, 124], and viral infections. Lipids play an essential role in virus-host interactions in the form of receptors, fusion cofactors, or modification of membrane curvature [125]. Their membrane receptors could also interfere with the lipoprotein metabolism [121].

Due to the relevance of lipids in many biological processes and especially the dysregulation of lipid metabolism in many diseases, studying lipids can lead to a better understanding of diseases. Investigating the role of lipids in the non-alcoholic fatty liver disease (NAFLD) was the goal of one publication in this dissertation [21]. Working with lipid data requires understanding the experimental procedure and computational processing of the data. This is also necessary for discussing limitations and the potential of lipidomics for future research. Therefore, mass spectrometry-based lipidomics, the research field that measures lipid abundances of biological systems, is introduced in the next Section.

2.3. Lipidomics

Deciphering the complex lipidome of organisms is the goal of lipidomics. Nowadays, lipidomics experiments are commonly performed using mass spectrometry (MS) [126, 127]. Before the wide availability of mass spectrometers, lipids were usually measured by thin-layer chromatography (TLC). This method applied lipids in an organic mobile phase to a stationary silica gel [128]. Lipid classes eluted at different positions after a particular time based on their interactions with the stationary phase. They could then be identified from standards and labeling with a dye [128]. Also, a two-dimensional TLC can be performed by applying mobile phases with different elution powers after each other [129]. TLC made it possible to identify the abundances of different lipid classes. However, the identification of individual lipid species was not possible.

Using mass spectrometry coupled with Liquid chromatography (LC), over a thousand lipids species can be identified [130]. This allows more profound insights into changes in the lipidome. The workflow of a typical lipidomics experiment can be seen in Figure 2.6. After

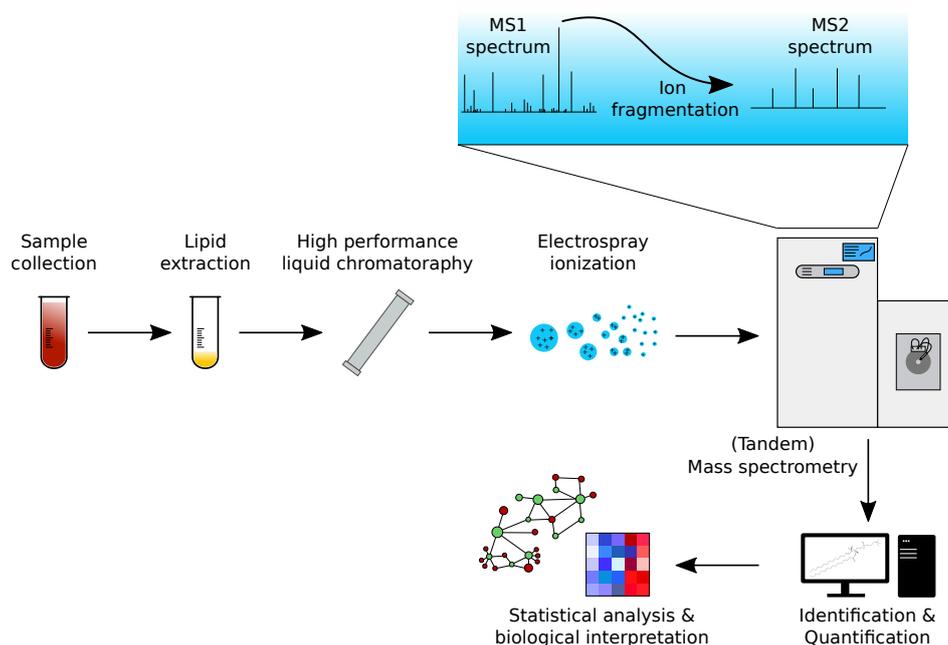


Figure 2.6.: Common workflow of mass spectrometry-based lipidomics experiments.

sample collection, lipids are extracted and optionally separated in a liquid chromatography step. After that, lipids are ionized and enter the mass spectrometer. It detects Mass-to-charge (m/z) ratios of ions and, when performing a tandem-MS experiment, can break apart the ions and measure their fragment m/z ratios. From the mass spectra, dedicated software tools identify and quantify lipids. The resulting lipidomics data can then be used for statistical analysis and biological interpretation. In many ways, lipidomics and metabolomics workflows are very similar. The most significant differences are in the extraction due to the non-polarity of lipids and computational identification. Each step of the experimental lipidomics pipeline influences the quality of the data and the resulting level of lipid identification, which in turn influences the computational analysis and interpretation. The variability of the workflow can yield different results. In a clinical context, standardization of such workflows is crucial. All necessary steps are explained below to give an overview of the complexity of a lipidomics experiment.

Sample collection

The first step in lipidomics is the collection of samples. In clinical lipidomics, blood [131] or tissue [132] samples are commonly used. Because of potential oxidization or enzyme activity that might alter the lipid composition even after sample collection, samples should be stored at -80°C [133]. Before starting an experiment, it should also be decided how many samples are necessary to perform the desired statistical analysis. Adding additional samples later might induce batch effects, which must be computationally removed. However, this requires

knowledge of batch processing of each experimental step.

Lipid extraction

After sample collection, lipid extraction is performed. It is done to remove all non-lipid molecules from the solution. Unwanted and less volatile compounds such as proteins or specific metabolites can lead to ion suppression in the MS [134], thus reducing the sensitivity of identifications in the experiment. The extraction should be performed in a cold environment, e.g. in a cold room or on ice. This ensures equal conditions for each sample and reduces the chance of undesired enzymatic or chemical reactions inside the sample after collection. Before the extraction starts, internal standards are added to the sample material. These resemble typical lipid classes that are expected in the sample but unlikely to exist in the organism, e.g., synthetically modified lipids [135]. They serve the purpose of monitoring the consistency of sample preparation and absolute quantification.

The most common protocols for lipid extraction are the Folch [136] and the Bligh & Dyer method [137]. Both methods are based on Chloroform, methanol, and water and aim to split the polar phase, which includes metabolites and proteins, from the non-polar phase containing lipids. The Bligh & Dyer method is faster and requires fewer volumes [137]. Also, other extractions that use, e.g., methyl tert-butyl ether (MTBE), have been developed [133]. Additionally, fractionation might be used to separate polar and non-polar lipids [130].

As I observed in the NAFLD stratification study [21] that the extraction day can also result in batch effects. This can be due to temperature alterations or newly mixed chemical solutions for the extraction. Therefore, reporting lipidomics data should also include extraction batch information for large cohorts.

Separation

Before ionization and injection into the MS, a separation layer is often applied. This step is not performed in direct-injection (also called shotgun lipidomics) experiments, where the extract is ionized directly. Due to the complexity of the lipidome, separation is required to detect the complete structural identity of lipids from complex biological samples [138].

The most common separation technique for lipidomics is LC. It separates lipids by chemical properties such as polarity. An LC consists of a high-pressure pump and a chromatography column. The column contains particles that interact with the lipids, also known as the stationary phase. Under high pressure, the column is loaded with the lipid solution, and a mobile phase is added that is pumped through the column. Through interactions of lipids with the mobile and stationary phase, lipids pass the columns at different times, which is captured as the retention time. Different LC techniques are available for lipidomics that differ in stationary and mobile phase [138]. In reverse-phase chromatography, a non-polar stationary phase is used with a polar mobile phase. This results in polar lipids eluting first, whereas more non-polar lipids eluting later. Normal phase chromatography is the opposite, with a stationary and non-polar mobile phase. Hence, lipids are eluting, starting with highly non-polar lipids. Finally, hydrophilic interaction chromatography uses a polar stationary

phase and an aqueous mobile phase. In this chromatography method, polar lipids elute earlier than non-polar lipids. It increases the sensitivity of the MS identifications since a less diverse set of lipids is measured at each retention time [139]. LC columns with different lengths, particle sizes, and chemical properties have been shown to influence lipid identification performance [140]. Therefore, the choice of LC technology and columns should be carefully evaluated.

Gas chromatography (GC) is mainly used for FFAs. The principle is that the solvent is inserted into the GC and heated up. Compounds vaporize at different temperatures based on the interactions with the stationary phase. It has provided a good separation for fatty acid (and derivative) characterization [141]. For complex lipids, GC requires derivatization to increase stability and volatility [142] and therefore is typically not used.

Another, more recently used separation technique for lipidomics is Ion mobility spectrometry (IMS). This separation is done post-ionization and often combined with LC. It adds another dimension to the measurement, the Collision cross-section (CCS). The principle relies on interactions of ions in the gas phase with a neutral drift gas that yields separation of ions based on size and chemical properties [143]. Multiple versions of IMS are available. For example, in drift tube IMS, ions travel through an electric field and are slowed down by the drift gas. In trapped IMS, ions are trapped, and the trapping voltage is gradually reduced while a drift gas flows through the trap. Based on their collisions with the drift gas, ions leave the trap at different time points. IMS happens on the scale of milliseconds and therefore is a high-speed separation technique, compared to LC, which is on the scale of minutes [143]. In lipidomics, IMS can help to minimize sample loss and preserve the biomolecular context (e.g., glycolipid conjugates) [144]. It has been shown to provide a better sensitivity with low sample amounts [145].

Ionization

A MS measures the m/z ratios of ions. Not all lipids are ions in their natural state. Ionization methods can charge molecules such as lipids such that the MS can detect them. The two main ionization methods used for lipidomics are Electrospray ionization (ESI) and Matrix-assisted laser desorption/ionization (MALDI).

For developing the ESI method, John Fenn received the Nobel prize in chemistry in 2002 [146]. ESI ionizes lipids that are in a liquid solvent. The solvent with lipids is ejected from a metal capillary with electric potential and forms a cone at the front [147]. Charged droplets are emitted at the tip of the cone, which is supported by a gas flow. The solvent is rapidly evaporating from the droplet. While the droplets shrink, the charge density builds up, due to a maximum number of elementary charges on a droplet (Rayleigh limit) [147]. Droplets exceeding this limit form even smaller droplets [147]. This way, droplets become smaller and smaller until single molecules and potential adducts are left. Adducts can come from the sample matrix or mobile phase solvents [148]. These nano droplets then enter the mass spectrometer (Figure 2.6). ESI is known as a soft ionization method. However, fragmentation of ions can occur and lead to losses of, e.g., hydroxyl, ammonia, or formate groups [148]. Buffer additives in the solvent are also commonly added to modify the pH and improve

ionization efficiency [138]. The exact mechanism of ionization has not been understood, but two physical models have been proposed [149]. The process results in lipids that have one or multiple charges. ESI is commonly combined with LC and IMS.

The other commonly used ionization method is MALDI. For MALDI, the sample is embedded in a matrix of often aromatic molecules, such as 2,5-dihydroxybenzoic acid. The solvent is added in a liquid phase, evaporates, and matrix molecules co-crystallize with the analytes [150]. Then a laser pulse is directed to the sample. This results in evaporation of analytes and ion formation with the matrix molecules. Different laser wavelengths, pulse widths, and matrix elements influence the ion formation [151]. MALDI is especially suited for spatial lipidomics analysis since the laser beam can be directed at different positions, such as a histological tissue sample [152]. It can also be combined with IMS for lipidomics [153].

Finally, it is essential to mention that some lipids ionize better with negative adducts, whereas others are mainly positively ionized.

Mass Spectrometry

m/z values of ionized lipids are measured in the mass spectrometer. If ions are additionally fragmented and product ions measured, it is called tandem MS or MS². The layout of most instruments is very similar. Since most machines nowadays are tandem MS, I will describe their basic setup. Ions enter the instrument and are accelerated by an electric field. After that, they enter a quadrupole mass filter. It consists of four parallel electrodes with oscillating voltages. By setting the voltages for a specific frequency and amplitude (described by Mathieu's differential equation), only ions of a specific m/z value pass the device [154]. Others collide with the electrodes or walls. Selected ions then proceed to fragmentation in the collision cell. In there, ions collide at high speeds with neutral molecules, inducing breaks in covalent bonds. These fragments can be lipid specific and, therefore, helpful for identifying molecules (More on that in the next paragraph). After this, ions enter the mass analyzer. The most common mass analyzers are the time-of-flight (TOF), quadrupole, and orbitrap.

As the name implies, the TOF mass analyzers measure the flight time of accelerated ions. It is based on the inertia of m/z ratios of molecules, which is higher for heavier molecules. Often, ions are accelerated and redirected by an electromagnetic repulsion. Their drift can be measured by an ion detector and the m/z ratio calculated. The quadrupole measures ions similarly to the quadrupole mass filter. It is set for specific m/z values or cycles through ranges of m/z values. The amount of passing ions is detected for each m/z value. In the orbitrap mass analyzer, ions are orbiting through an elliptic spindle, and the induced current is measured [155]. Through Fourier transformation, m/z values can be identified. A good overview of mass spectrometer types and their potential for lipidomics was published by Köfeler et al. [139].

Many mass spectrometers have additional parts, such as ion traps, to accumulate ions in specific steps, but that does not alter the basic principle of filtering, fragmentation, and mass detection. Due to positive or negative charges of ions, an MS measurement is done either in negative or positive mode, where it accelerates only one type of charge. When a simple MS experiment is performed without fragmentation, the initial mass filtering and fragmentation

steps are not performed, and ions proceed directly to the mass analyzer.

In lipidomics, the analysis can be differentiated between targeted and untargeted [156]. Targeted lipidomics is performed when known lipid species should be confirmed or quantified in a sample. It is performed, by filtering for specific precursors (for simple MS) and fragment ions (for MS2). This requires knowledge about the mass and fragmentation patterns of the target molecule. The most common instrument of choice for this analysis is the triple-quadrupole mass spectrometer [156], first developed by Yost et al. [157]. The untargeted analysis aims at identifying all (or as many as possible) lipids from a sample. For MS1, as many ions as possible within a mass range are detected. Untargeted MS2 can be generally divided into Data-dependent acquisition (DDA) and Data-independent acquisition (DIA). DDA selects and fragments ions based on the measured intensities of the precursor ions. Ions exceeding a certain intensity are fragmented, and the fragmentation results are detected. This is the most common MS2 mode in lipidomics [140]. A restriction is that low-abundant lipids might not be fragmented and that within a time interval, only a limited number of ions can be fragmented. The problem can be partially solved by multiple DDA MS2 runs with the iterative exclusion of the previously fragmented ions [158]. In DIA, ions are fragmented independently of their measured abundance. One DIA method is All ion fragmentation (AIF), which fragments all ions without a prior mass filter [159]. This mode loses information about the precursor ions of the fragments, and fragments of all ions are mixed. An alternative is Sequential window acquisition of all theoretical fragment ion spectra (SWATH) MS [160, 161]. In SWATH small windows of mass ranges are selected for fragmentation and cycled through. This method was initially developed for proteomics, but it can also be applied for metabolomics and lipidomics, where it could benefit absolute quantification [162].

Lipid Identification & Quantification

Identifying lipids from MS1 spectra is based on reference databases of lipids. Masses of lipids combined with possible adducts can be matched to observed peaks. Additionally, isotopic peaks can be incorporated to improve the identification [163]. However, such identification is still prone to errors since multiple molecules can have the same mass. It also does not allow the identification of lipids on the molecular species level. For example, the mass of PC(18:2_18:2) is identical to the mass of PC(16:0_20:4) (both have the sum formula PC(36:4)).

MS2 can provide the information to identify molecular species. Since fragmentation of lipids commonly occurs at the ester bonds of fatty acyls and the headgroup, these can be predicted or are available in libraries, such as the ALEX¹²³ lipid calculator [73]. Figure 2.7 shows the MS2 spectrum of a PC(18:3_18:3) in negative scan mode. For example, a fatty acid fragment (m/z 277.22) can be observed. Another example is the neutral loss of fatty acid and parts of the headgroup (m/z 502.30). All fragment information combined can increase confidence in the identification of a lipid.

Software tools can automate the identification of hundreds of lipids. Non-commercial software for shotgun lipidomics is e.g. ALEX [164], ALEX123 [165], or LipidXplorer [166]. LC-MS(2) identification can be performed with MS-DIAL [167], Lipid Data Analyzer [168], LipidMatch [169] or LipidHunter [170]. Because quantifying lipids on fragments is chal-

2. Introduction

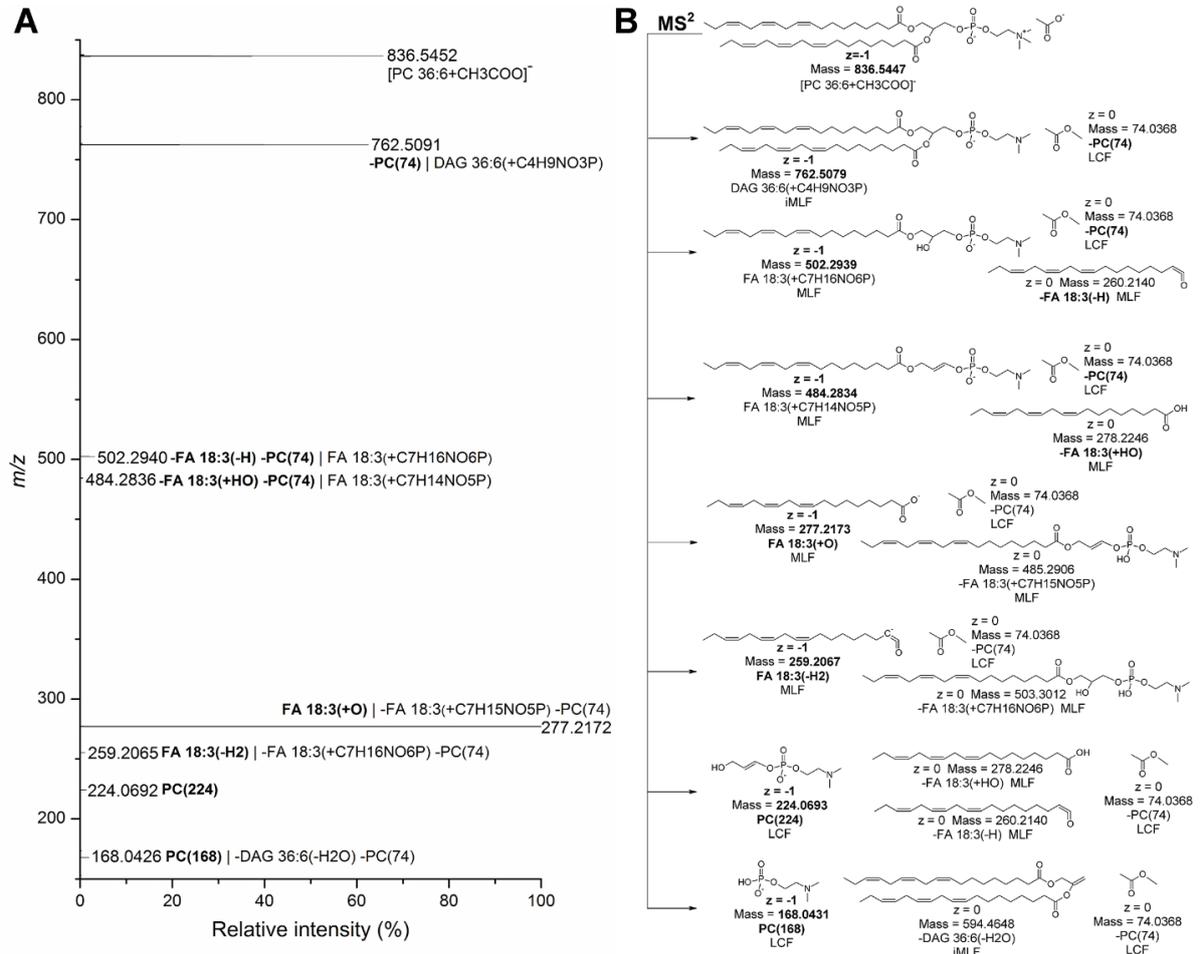


Figure 2.7.: Fragmentations of a PC(18:3_18:3) with a CH₃COO⁻ adduct in negative scan mode. (A) MS² spectra of the m/z 836.5 precursor. (B) Potential annotated fragments of the peaks in A.

The figure is taken from the publication of J. K. Pauling et al. [73]. It has not been modified and is originally provided under the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>).

lenging, it is commonly done on MS1 peaks, using identified lipids from MS2 spectra, if available. The Peak area or intensity of identified lipid peaks is used for quantification. Baseline correction from blanks can be done correctly for noise in the measurement [169]. If internal standards of known concentrations were added at the beginning of the extraction, they could be used for (semi-)absolute quantification. In this case, the intensity of lipids of the same class (or lipids with a similar retention time) is used to normalize peaks and predict concentrations from a standardization curve [171].

2.4. Other omics technologies

Lipidomics is not the only "omics" discipline. Compared to other techniques, it is a relatively new field. Omics techniques have enabled a systematic molecular analysis of organisms. While the main focus of this dissertation lies in lipidomics, other omics techniques are commonly used for molecular biological research and can give valuable insights into biological functions. The disease subtyping method I present in this dissertation [20] can be used with any omics data. In the systematic evaluation of the method, transcriptomics, proteomics, and metabolomics data were used. I will introduce some of the most widely used omics methods in the following.

Capturing more precise and complete data is dependent on technological advancement. The first human genome was sequenced in 2001 [172]. However, the first gap-free sequence of a whole human genome has only been published in 2022 [173]. Genomics can reveal mutations that might be risk factors for diseases [174]. In contrast to the first sequencing method published in 1977 by Sanger et al. [175], next-generation sequencing (NGS) is massively parallelized and high-throughput [176].

Observing and understanding the changes in gene expression in disease development can be achieved by transcriptomics. Transcriptomics first was traditionally measured with microarrays, a targeted method, where transcripts bind to the provided oligomers. Currently, transcriptomics is mainly done by mRNA sequencing. With this method, transcripts are individually sequenced after breaking them into smaller reads and mapped to a genome [177]. In contrast to the genome, the transcriptome is more dynamic, since cells can alter their gene expression.

As already mentioned, mRNA is not the final product of converting a gene into a functional unit. After transcription comes translation, where a nucleotide sequence is translated into an amino acid sequence, resulting in a protein. Therefore, proteomics can be beneficial in investigating the functional state of a sample. Similarly to lipidomics, proteomics is performed with MS. Proteins are split into smaller peptides prior to ionization. In the MS, the m/z ratios of peptides are measured and additionally fragmented. The peptide chain commonly breaks at the peptide bond, leaving peptide fragments broken at different places of the chain. Since the masses of amino acids are well known, the fragments can be computationally put together and quantified. Thousands of proteins can be identified in a sample using MS. Mass spectrometry-based proteomics has the advantage that it can also be used to identify posttranslational modifications, such as phosphorylations that, e.g., activate the catalytic

activity of a protein. MS-based proteomics can also be combined with protocols such as affinity purification [178]. This way, proteins can bind and build complexes to a specific target before measuring. Only proteins bound to the target proceed to the MS. This makes it possible to identify interactions between proteins.

Metabolomics is performed using MS or nuclear magnetic resonance spectroscopy (NMR). With MS, molecules are identified based on their m/z ratio, whereas in NMR intra-molecular and inter-molecular resonances between protons (or other atomic nuclei) are measured for identification [179]. MS yields higher sensitivity and reliable identification, whereas NMR provides highly quantitative and reproducible measurements [180]. Identifying and quantifying metabolites or lipids only gives a snapshot of the current state of the metabolism. Even if several time points are measured, it cannot be tracked which metabolic pathways are active or into which end products metabolites are converted. The idea of ^{13}C fluxomics is to achieve this by labeling metabolites. In nature, two stable isotopes of the carbon atom exist, ^{12}C with 6 protons and 6 neutrons in the atomic core, and ^{13}C with 6 protons and 7 neutrons. The relative abundance of ^{13}C carbon in nature is only around 1% [181]. Molecules with ^{13}C carbon have a higher mass than the molecule without the isotope, which can be detected in an MS. By introducing artificially ^{13}C enriched metabolites to a sample, e.g., glucose, and performing metabolomics over a series of time intervals, it can be observed that other molecules with the ^{13}C carbon become higher abundant. Calculating the flux through metabolic pathways from this data can be done computationally but requires knowledge about active enzymatic reactions in a sample [182].

This is not a complete list of omics techniques, but it gives an overview of some of the most popular ones. For omics data, it is also important at what resolution and dimension it is measured. Typically, omics data is measured from bulk samples. This means that cells from cell culture, tissue, or a blood sample are processed and the biomolecules of interest isolated. This is usually done to get a minimum required amount of sample to detect the molecules. However, such a sample can contain many different types of cells. Experimental comparisons between samples from cancerous and healthy tissue on the bulk level might not reflect the high heterogeneity of cancer cells within a tumor [183]. A complete understanding of the tumor would require investigating its molecular biology on the single-cell level.

Therefore, single-cell omics techniques have become more prominent. Single-cell RNA sequencing was first published in 2009, and the technology has made big progress since then [184]. For instance, subclasses of cells in the brain cortex could be revealed using single-cell RNA sequencing [185]. In cancer research, it can be utilized to understand metastasis development [186]. Single-cell analysis is possible not only for RNA data. Also, metabolomics/lipidomics can be performed on the single-cell level using MALDI MS coupled with microscopy [187]. Another important aspect of understanding the function of cells within a tissue is their spatial location. A functional tissue contains multiple cell types that are specialized and often require a specific cellular environment. Spatial omics techniques can be used to achieve this. They aim to give a two-dimensional representation of, e.g., metabolite abundances. For lipidomics, this is achieved by MS imaging [188].

Gaining knowledge from large amounts of biological omics data can be facilitated by computational methods, such as the ones developed in this dissertation. The following section introduces essential concepts of computational biology.

2.5. Computational Biology

The Computational biology (CB) research field aims to produce knowledge about biological processes and systems by analyzing and modeling biomedical data. This can range from developing novel computational methods to applying analysis and modeling approaches. It is a highly interdisciplinary research field since it requires understanding experimental potentials/limitations, statistics and mathematical modeling, and the biological system. The work of a computational biologist often starts after data has been acquired and is complementary to experimental work by helping to generate knowledge from the data.

Biology has become a highly quantitative science [189] because research questions usually involve many genes or connections. Large quantities of "omics" data contain various information about the state and change of a biological system, be it from case-control experiments or big clinical cohorts with different disease manifestations and origins. Molecular biology consists of complex interactions between thousands of molecules that have dynamic effects on each other. CB approaches can be used to identify disease mechanisms, infer regulatory interactions, provide low-dimensional representations, or enable multimodal data integration [190]. Therefore, computational biology utilizes methods and concepts from mathematics, statistics, machine learning, and physics to model biological data.

A quote from an article about computational biology summarises it very well: "Computational biology turns ideas to hypotheses" [191]. This highlights that the outcome of a CB analysis are hypotheses about phenomena, which require external confirmation or validation by additional experiments that, e.g., test the proposed mechanism on other patient data that was not part of the original analysis.

2.5.1. Systems biology & systems medicine

The computational methods explained in this section focus on studying systems' biological phenomena. In systems biology, cellular function is studied as a whole; instead of investigating single entities, such as single genes or lipids [192]. Two main paradigms exist to study systems biology [193]. In bottom-up analysis, a subset of previously known connections is used to predict changes in the system. This relies on precise (quantitative) knowledge of interactions. The top-down method tries to find new interactions and regulations from vast molecular biological data. Top-down can also profit from prior knowledge in the form of molecular interactions. Bottom-up research requires more targeted data, with many perturbations to estimate the systemic response, whereas top-down studies work with big datasets and fewer perturbations [193]. Experimental data and computational models must be integrated to understand complex biological behavior [194]. Especially for top-down research, omics data is necessary, capturing the whole system's properties. The results of a top-down analysis

can serve as a starting point for targeted bottom-up research. The computational methods presented in the publications of this dissertation follow the top-down paradigm. They were designed to identify disease subgroups, hypothesize mechanisms, and visualize trends in complex omics data.

Applying systems biological research methods to investigate diseases is known as systems medicine. Systems Medicine wants to overcome current limitations of disease complexity [12]. The aim is to make medicine predictive, preventive, personalized, and participatory, referred to as P4 medicine [195]. To reach this goal, molecular environment, socioeconomic interactions, and co-morbidities of diseases are analyzed [196]. A particular challenge is integrating all types of (molecular) biological information [197]. To get a systematic understanding of a disease on the molecular scale, all levels of biology need to be considered. This is where CB comes in, to convert data into knowledge that reveals disease mechanisms, diagnostic markers, subtyping, and drug targets [197]. Molecular interaction networks and regulatory information offer additional benefits to the computational analysis in systems medicine [198]. In the long term, systems medicine can personalize medicine by treating molecular causes rather than symptoms [195].

This dissertation's computational methods and applications fit into the systems medicine workflow for subtyping, biomarker detection, and mechanism investigation. While they work independently in their sub-domain, I will present an integrated workflow in the discussion. Systems medicine is very similar to systems biology and shares many principles. Therefore, they can also be applied in a general systems biological context to study the behavior of cellular systems in a top-down fashion.

In the following sections, I will explain and summarize concepts focusing on analyzing clinical omics data. I divided the section into data-driven approaches, which work entirely with the quantitative information provided by the omics experiment, and prior knowledge-based approaches, which additionally take prior knowledge into account.

2.5.2. Data-driven approaches

While simple statistical methods, such as a t-test that compare two population means [199] can help manifest differences between two experimental groups, more sophisticated methodologies are necessary to unravel the complex relationships and interactions in molecular omics data.

Machine learning algorithms are promising to infer biological knowledge from such data [190]. The algorithms are designed to learn and find patterns in data. Combined with the right research question, their findings can be interpreted as molecular biological insights. Machine learning algorithms used in CB can mainly be divided into supervised and unsupervised learning methods.

Unsupervised machine learning

Unsupervised machine learning works with unlabeled data, meaning that data is used as an input without providing additional information, such as the experimental conditions. Dimensionality reduction algorithms aim at representing multi-dimensional data at a low dimensionality, which humans can capture. Prominent representatives are Principal Component Analysis (PCA), Uniform Manifold Approximation and Projection (UMAP) [200], and t-Distributed Stochastic Neighbor Embedding (t-SNE) [201]. They make it possible to visualize data, e.g., hundreds of lipid features in a two-dimensional plot. PCA shows linear combinations of features that exhibit the most variance in the data. t-SNE and UMAP are non-linear methods that aim to preserve similarities between samples and embed them in a low-dimensional space. Such methods are often the first step in the analysis to get an overview of the data. In the NAFLD stratification study, we used dimensionality reduction to visualize global differences between experimental groups and potential batch effects. They can also be used to investigate multiple omics layers jointly [202].

Unsupervised approaches based on neural networks are becoming more popular with better data availability. For example, to learn low-dimensional representations for integrating multiple datasets [203] or multiple omics layers [204].

Another group of unsupervised learning algorithms is clustering. Clustering algorithms group samples into subsets (clusters), where samples within one cluster are similar, whereas non-similar samples belong to different clusters [205]. Typically, clustering algorithms require the number of clusters as a parameter. Therefore, they require an initial guess or hypothesis on the number of expected groups in the data. However, methods to estimate the best number of clusters from the data have been developed [206]. Probably the most simple clustering algorithm is k-means. The algorithm assigns samples to clusters based on their distance to centroids. Their positions are updated to the mean positions of all the assigned samples until the algorithm converges. Other clustering algorithms are based on densities (e.g., DBSCAN), distributions (Gaussian mixture models), or hierarchical trees. Different clustering approaches can yield different results on different datasets and should therefore be evaluated to select the most appropriate tool for an analysis [207]. The application of clustering for clinical omics data is, e.g., the de-novo grouping of disease patients based on molecular data resembling traits such as treatment response, survival, or recurrence [208]. This can provide an alternative view of the disease based on the molecular phenotype and be less biased than commonly used disease definitions. Molecular subtyping can enable more efficient treatments [209], since patients show similar molecular alterations. Especially in the clinical context, high interpretability of stratification is important [210].

Biclustering is a special case of clustering, where samples and features (e.g., lipids) are clustered simultaneously. They not only find sample subgroups but also subsets of features that are characteristic of them [20]. This can yield additional insight for the identification and characterization of patient subgroups. In a systems medicine context, it can be particularly interesting for disease subtyping since the disease does not affect all biomolecules but manifests in a subset of genes, proteins, lipids, or pathways. Many algorithms have been developed, with different objectives and heuristics [211]. Biclustering is the primary technology for the

MoSBI method [20]. The publication also provides a more detailed overview of biclustering algorithms. As shown in this work, biclustering and unsupervised approaches generally have a vast potential for disease subtyping since they can find novel groups in the data.

Supervised machine learning

In contrast to unsupervised learning, labeled data is used for supervised machine learning. Typical tasks for supervised machine learning methods are classification or regression. The simplest form of regression is linear regression. A linear model fits a linear relationship between one or multiple independent variables and a dependent variable. Another type is Cox proportional hazards regression, which quantifies the survival between two patient groups [212]. Such a model can be used to, e.g., predict cancer development [213].

Classification models learn to differentiate samples based on categorical variables. In a clinical context, disease manifestations can be classified based on omics data [214]. This way, future patients can be diagnosed based on their molecular data. Examples of classification algorithms are decision trees, support vector machines, or naive Bayes. I utilized classification models to investigate how well NAFLD subgroups can be differentiated from each other [21].

Also, for supervised machine learning, neural network-based methods have been developed and deliver promising results for complex CB problems. For example, they are used to predict gene expression from regulatory interactions [215] or three-dimensional structures of proteins from their amino acid sequence [216].

The concept of feature importance plays an essential role in applying supervised machine learning methods in omics data. In many supervised machine learning algorithms, the feature importance can be quantified. This describes how much a feature or combination of features contributes to predicting a particular label. These can serve as biomarkers for a disease [217]. Therefore, machine learning applications for biomedical omics data often are applied to identify potential biomarkers [218] explicitly. As shown in the NAFLD stratification study, also a combination of supervised and unsupervised methods can be used to predict potential biomarkers [21] (Appendix A.2).

A common problem for supervised machine learning methods on omics data is overfitting. It can occur when a high number of features are available for a comparably low number of samples. In this case, models use many parameters to incorporate each feature, allowing them to fit the data perfectly. However, a perfect fit means that all training samples can be perfectly predicted, but labels for new samples that were not part of the training are predicted poorly. So the algorithm learns the training data's properties and does not generalize beyond the training data. This is not desirable since computational biologists are interested in finding genuine relationships in the data that provide insights into the molecular machinery instead of data-specific artifacts [219]. Multiple solutions have been proposed to cope with the problem. Examples are regularization, where the number of predictive variables in a model is reduced [220], splitting the data into a training and test set, or applying more complex strategies such as cross-validation.

Machine learning algorithms are not the only algorithms applied to omics data in CB.

Also, other heuristic algorithms or statistical methods are commonly used. Here, I want to highlight specific methods that infer networks from omics data. Networks are not only visually appealing for data analysis but can resemble so far unknown interactions of biomolecules in a cell. Many studies employ correlation networks, where a connection between two features indicates a correlation above a certain threshold. They are used across all omics disciplines and can indicate co-expression or co-regulation [221]. Another example is partial correlation which regularizes the number of edges [222]. A typical application is to infer gene regulatory networks, for which various computational methods have been published [223].

2.5.3. Prior-knowledge integrating approaches

An advantage for the computational analysis of omics data can be the integration of prior knowledge. Research in molecular biology in the last century has led to extensive knowledge about molecular interactions, functions, and pathways. This knowledge can be utilized to enhance computational methods and increase interpretability.

Central to the integration of prior knowledge is its accessibility. A variety of biological databases are available. Examples of databases (that were also partially used for the work in this dissertation) are as follows. The Kyoto Encyclopedia of Genes and Genomes (KEGG) stores molecular pathways, structural information, disease perturbations, and other molecular biological information [224]. Information about protein-protein interactions, which can be utilized to create corresponding networks, is available in the STRING database [225]. Metabolic reactions can be found, e.g., in the Rhea database [226].

Pathway analysis methods

A typical research question is which known pathways are significantly dysregulated between two experimental or clinical conditions. Two approaches are mainly used for such analysis: Gene Set Enrichment Analysis (GSEA) and Over-representation Analysis (ORA). Both methods utilize information about gene participation in pathways from databases such as Gene Ontology (GO) or KEGG.

GSEA uses a list of genes (or other features) ranked by correlation with the phenotype of interest [227]. These can be p-values of t-tests or feature importance of a machine learning model. The algorithm then scores if the genes corresponding to a pathway accumulate predominantly at the top of the list. A p-value can be calculated by perturbing the feature ranking. The rationale is that all genes should show a substantial alteration in an active or dysregulated pathway. ORA utilizes a similar principle but works with a non-ranked set of features and tests if more features than expected, belonging to one pathway, are in the set of features [228] (using a hypergeometric test). Such a feature list can result from t-tests after setting a significance threshold. Pathway enrichment is not only possible for genes, but also methods for metabolites [229] and lipids [230] have been developed. For the evaluation and systematic comparison of computational methods, pathway enrichment algorithms are often used [231, 232, 233] since they provide simple and interpretable insights into the results of an analysis. The resulting pathways can provide a hypothesis for dysregulated pathways

in diseases [234, 235]. As with every computationally generated hypothesis, it requires an additional experimental validation by individually investigating a pathway's activity.

GSEA and ORA are both top-down systems biological methods. However, bottom-up pathways analysis is also typical, especially in metabolic research. This analysis of individual pathways (or reaction networks) is known as kinetic modeling [236]. Kinetic models use ordinary differential equations or stochastic numerical methods [237] to simulate metabolic pathways. They can show how functional behavior emerges from dynamic concentration changes [238]. Rate equations model enzymatic reactions, e.g., the Michaelis-Menten kinetics [239]. These can incorporate regulation by metabolites or signaling cascades [238]. Parameterization of kinetic models can be very challenging due to many parameters. Therefore, whole-cell models, including many pathways, have been developed [240], but are still rare. As an example application, kinetic modeling was used to understand the dynamics and vulnerabilities of fatty acid metabolism [241]. In the clinical context, kinetic models can potentially recapitulate variations in the dynamics of disease development in patients [242]. Other applications are in modeling the metabolism on a larger scale. This has been done for the glucose metabolism of the liver and transport to other organs [243].

Network analysis methods

Pathways are human-defined molecular subsets that were grouped by one particular molecular function. As discussed with the Pentose phosphate pathway (PPP), borders between pathways are not always clear. This means that they can be biased toward certain well-studied functions. Therefore, a de-novo network analysis that works on a whole interaction network, such as protein-protein interactions from the STRING database [225] or a genome-scale metabolic network, e.g., the previously mentioned Recon network [38]. By combining complete interaction networks with molecular omics data in a top-down systems biological approach, the activity or dysregulation of specific network modules can be inferred and provide a de-novo pathway analysis.

Strategies to incorporate biological interactions for data analysis vary and are very diverse. For example, a method by Benedetti et al. [244] builds correlation networks of features entirely based on omics data. Interaction networks are utilized in the last step to find a correlation cut-off. The cut-off is selected such that the resulting network has the highest overlap with the prior knowledge network. This gives the method the freedom to find correlations that are not previously known. On the other hand, it makes interpretation of the network more challenging. Other approaches, such as active module identification or network enrichment, work with a different principle.

Network enrichment methods start from a complete interaction network and combine it with omics data, where each feature corresponds to an edge or node in the network. A heuristic is then used to identify a subnetwork that shows the highest activity, variation in samples or can explain differences between disease groups. Unsupervised methods for network enrichment have been developed [245], as well as supervised methods [246]. It was also utilized to combine network enrichment with feature selection of machine learning algorithms [208]. With network enrichment, pathways associated with liver fibrosis [247] or

hubs for mutations in cancer [248] have been identified. Alterations in interactions can also be investigated in this way [249]. In this dissertation, I co-developed an enrichment algorithm specifically for lipid networks [23]. Enrichment on lipid networks was not possible so far because of missing comprehensive lipid networks combined and unique requirements to include the enzyme multispecificity into the modeling.

An important network-based method specific to metabolic networks is constrained or stoichiometric modeling. It utilizes the stoichiometry of metabolic reactions and constraints about maximum and minimum reaction activity [250]. The most famous method for constrained-based modeling is Flux Balance Analysis (FBA). It studies the behavior of genome-scale metabolic networks under steady-state conditions. A steady-state occurs when the metabolite concentrations do not change; therefore, the same amount of nutrition is consumed as an organism grows/generates a product. The objective of an FBA is usually to maximize the growth of an organism. But also other objectives, such as ATP production, energy reduction, or product synthesis, can be investigated [251]. A great potential of the method is that it can simulate a metabolic steady-state under several conditions, e.g., gene knockouts or limited nutrition availability [252]. FBA analysis has been successfully applied to, e.g., find drug targets for Tuberculosis by simulating knockouts [253] or investigate the metabolic states of breast cancer [254]. But FBA is not the only method that analyses metabolic networks under steady-state conditions, other strategies that sample flux distributions [255] or investigate flux variability [250] have been developed. Stoichiometric network analysis approaches can be integrated with metabolic flux data [256] or mRNA sequencing data [257] to resemble experimental conditions.

2.5.4. Computational lipidomics

Only a few computational methods are available for systematically mining and interpreting lipidomics data. Examples are the lipidr [258], BioPAN [259], LION/web [230], or the LUX Score [260]. The lipidr software focuses on statistics on fatty acyl chains and a lipid set enrichment [258]. LION/web implements an ontology enrichment for lipid functions, chemical, and physical properties [230]. The LUX Score computes an embedding of lipidomes into a chemical space [260]. Finally, BioPAN is a network-based method that can predict active linear pathways on the class, fatty acid, or lipid species level [259]. These methods and more algorithms are further discussed in the attached manuscripts on lipid networks (Appendix A.3 & A.4).

Such a low number of algorithms and computational methods compared to other omics disciplines makes it challenging to analyze lipidomics data in every experimental scenario and gain functional knowledge. Because of big differences between lipids and other metabolites in structural and metabolic aspects, methods designed for metabolomics analysis cannot be used out of the box for lipidomics analysis. Further, differences in lipid identification, either on the sum- or molecular species level, complicate the development of general methods.

Networks are crucial for integrating lipidomics data with other omics data. They can include metabolic reactions between lipid species and links to the corresponding lipid enzymes. Also, signaling processes and lipid transport can be encoded in networks. This is also important for

systems medicine research. Interaction networks can point towards molecular mechanisms of diseases in combination with quantitative omics data. For this dissertation, two lipid network-based approaches have been developed [22, 23]. They aim to functionally understand systematic lipidomic alterations and compute hypotheses for enzymatic dysregulation from lipidomics data.

2.5.5. Drug repurposing

Two co-author contributions in this dissertation focus on drug repurposing [26, 27]. Drug repurposing is the application of drugs for other diseases. New drug development and certification are associated with high costs and time-consuming processes. Therefore, drug repurposing is especially interesting for rare diseases [261]. But also for cancer, promising repurposing candidates have been identified [262]. In the United States, around 30% of newly marketed drugs are repurposed [25]. Computation approaches are crucial for drug repurposing. They are used to predict novel drug targets on the molecular level through pathway/network mapping, docking simulations, signature matching, or genetic associations [261]. During the Sars-CoV-2 pandemic, drug repurposing gained strong interest to respond to the disease [26]. In this context, I contributed to a project on repurposing using interaction networks [27].

Computational drug repurposing is also facing many challenges. In particular, the transition from computational predictions to experimental validation and clinical trials. This is the main focus of a review I co-authored about lessons from drug repurposing research during the COVID-19 pandemic [26].

2.5.6. Challenges in computational biology

Computational biology (CB) comes with many challenges. Overfitting as a potential problem of machine learning algorithms was already mentioned. But this is not the only challenge. Problems for computational biologists can already occur before the computational work starts. Biases in clinical trials can happen at any trial stage, for example, during patient selection [263]. Cohorts can be subject to gender, ethnic, or social biases. These might affect the generalizability of a study outcome. After cohort acquisition, batch effects can be introduced in the sample processing. They describe differences between samples that are introduced because of processing of samples by different people, changed solvents, or environmental differences [264]. Standardized protocols and randomization can help reduce and make it possible to cope with them. One strategy to tackle batch effects in lipidomics data is described for the NAFLD stratification publication [21] (Appendix A.2). To detect batch effects, computational biologists need information about the exact experimental procedure. When comparing different cohorts from different studies, standardized procedures and protocols are necessary to achieve comparability [133]. Standardization is critical in lipidomics, where different nomenclatures and levels of identification are used.

More challenges occur when applying computational methods to quantitative omics data. Depending on the analytical method for measuring, the resulting data can have different

distributions or sparsity. For example, single-cell data tends to be more sparse, i.e., more missing values per sample [265]. Computational methods are often developed with certain assumptions about the distribution. Therefore, researchers need to check whether the data is suitable for analysis using the selected method. This is even more challenging when integrating multiple omics layers with different distributions and modalities. Various methods are usually available that address the same (or similar) problem. Evaluations of computational methods and comparisons can often be found in the literature, e.g., the DREAM challenges [266]. Another potential issue of utilizing CB approaches is parameterization. Parameter choices should be carefully evaluated for each data set. Evaluating the influence of parameters and data modalities for CB methods is crucial. Therefore, in the MoSBi publication a systematic influence of these properties on biclustering algorithms was evaluated [20] (Appendix A.1).

Mapping omics data to networks and pathways is another challenge in CB [267]. A big problem lies in the bias of biological networks towards over-studied areas. Parts of networks can be over-represented. In protein-protein interaction networks, systematically higher connectivity towards commonly studied proteins has been observed [233]. On the contrary, e.g., large parts of the metabolome remain unstudied and cannot be mapped to pathways [268]. This is a particular challenge for lipids, where systematic lipid species networks are unavailable. With the Lipid Network Explorer (LINEX) framework, this problem was addressed [22, 23] (Appendix A.3 & A.4). Such biases can strongly affect the analysis, and the prior knowledge quality needs to be evaluated. One needs to pay attention in the analysis that specific biomolecules are highlighted just because they have higher connectivity instead of exhibiting exciting behavior.

Because of these challenges, computational biologists need to process each dataset carefully and check if assumptions for specific computational methods hold and what potential limitations or biases are. I will return to this topic about specific challenges during my research in the discussion.

2.6. Objective

The introduction gave an overview of lipid metabolism, the pipeline of lipidomics experiments, and computational biology. While there have been many achievements over the last decades, there are still critical challenges in integrating lipidomics into systems medicine research. This publication-based dissertation aims to address some of these challenges. In the following chapters, the developed and applied methods are presented and summarized in the context of the publications. The dissertation includes three first-author publications, two co-author contributions, and a method available as a preprint.

The three first-author publications are: First, the biclustering ensemble method MoSBi for the automated stratification of omics data [20]. Integrating multiple algorithms and profiting from them overcomes the specificities of single algorithms. The publication presents the method, an extensive evaluation and comparison to other methods, and possible applications. Using networks makes it possible to visualize biclustering results on a large scale. Second, a

NAFLD liver lipidomics study. Using the previously developed biclustering method, subtypes that resemble the progression of the disease are discovered [21]. This shows the applicability of the biclustering method to discover novel disease subtypes and potential molecular markers for disease progression. Third, the LINEX method generates data-specific lipid networks and combines them with statistics to provide a functional analysis of lipidomics data [22]. It addresses the problem of mapping lipidomics data to metabolic networks.

Next, unpublished work is discussed, extending LINEX [23]. It provides a method to create more comprehensive lipid-metabolic networks and a network enrichment algorithm to derive hypotheses for enzymatic dysregulation from lipidomics data. The work addresses the problem of inferring the alterations of transcripts/proteins from lipidomics data.

Finally, two co-author contributions are mentioned that arose during the Sars-CoV-2 pandemic and focused on drug repurposing. I participated in a study that developed an interaction network-based method to find drug repurposing candidates [27] (summary is available in Appendix A.5.1). In the other publication, I co-authored a review that evaluated methods utilized for finding drug repurposing candidates for the disease [26] (summary is available in Appendix A.5.2).

3. Methods

The following chapter gives an overview of the developed computational methods of this dissertation: The biclustering ensemble method Molecular Signature identification using Biclustering (MoSBi) for sample subtyping, the Lipid Network Explorer (LINEX) for functionally analyzing lipidome alterations, and based on this, the Lipid Network Explorer 2 (LINEX²) that computes more comprehensive lipid networks and includes a network enrichment algorithm. The summaries provide an intuition about the principles of the algorithms. The entire explanations and definitions can be found in the full publications (section A.1, A.3, A.4).

3.1. An Ensemble biclustering method for disease subtyping

The first publication [20] (Section 4.1) presents the ensemble biclustering algorithm MoSBi. Similar to clustering, biclustering is a class of unsupervised machine learning algorithms. The difference between clustering and biclustering algorithms is visualized in Figure 3.1. Clustering algorithms use all features of omics data to identify sample groups, whereas biclustering groups samples and features simultaneously. This makes it possible to analyze molecular signatures for each subgroup to guide the biological interpretation and downstream analysis.

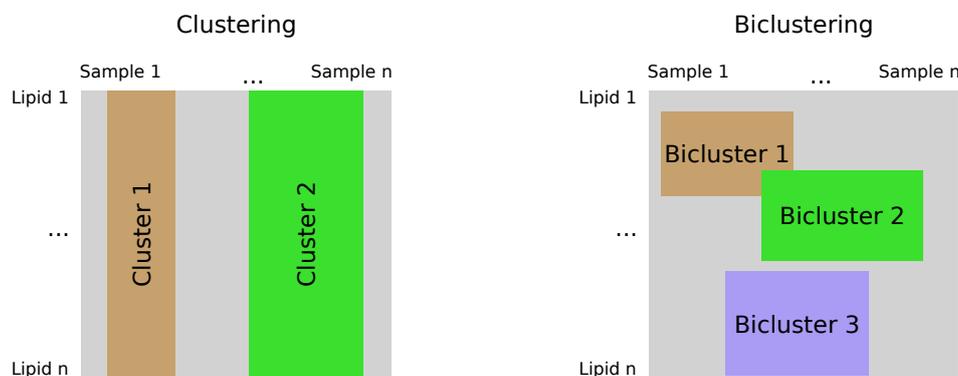


Figure 3.1.: Difference between clustering (left) and biclustering (right) on an omics data matrix.

Unaltered figure taken from Vvedenskaya and Rose et al. [21] (Supplementary Figure 3SA). Original figure distributed under the CC BY-NC-ND license (<https://creativecommons.org/licenses/by-nc-nd/4.0/>).

As an ensemble approach, MoSBi integrates the results of previously published biclustering

algorithms. This can be beneficial because each biclustering algorithm has different objectives and definitions of biclusters. By combining multiple algorithms into a joint ensemble output, we can profit from each algorithm and gain more confidence in their predictions.

The MoSbi algorithm consists of several steps. First, Each biclustering algorithm is executed independently. Second, an overlap between each pair of predicted biclusters is computed, resulting in a similarity matrix for all biclusters. For instance, the Jaccard index can be used as an overlap or similarity metric. The overlap is calculated independently if two biclusters were predicted by the same algorithm or not. Third, an error model is utilized to filter for randomly overlapping biclusters. The method aims at finding strongly overlapping biclusters that point towards the same underlying data structure. Therefore, slight overlaps between biclusters that can occur randomly are filtered out. In the fourth step, MoSbi builds a network of biclusters from the filtered similarity matrix. Connections between biclusters indicate higher than random overlaps. Finally, network communities are extracted using the Louvain method [269]. The network communities can then be interpreted as ensemble biclusters.

A formal description of the algorithm can be found in the Methods Section of the publication [20] (attached to Section 4.1). The second publication used the algorithm to stratify NAFLD patients (Vvedenskaya et al. [21], Section 4.2).

The network visualization is central to the MoSbi workflow. It gives the user an overview of the biclustering predictions and their overlaps. Colorings of biclusters offer additional benefits. Biclusters colored by algorithms can offer insight into the similarity between the results of biclustering algorithms or within one algorithm. Coloring by sample groups can reveal the enrichment of certain conditions within bicluster communities. This serves as a starting to investigate those communities further in downstream analysis.

Match scores were used to evaluate the algorithm performance in the publication. They quantify the match between predicted biclusters and ground truth. We defined the ground truth by simulating data with implanted biclusters or known disease subtypes from clinical omics data. The gene match score has been widely used to evaluate biclustering algorithms. To apply it on simulated data, we extended the score to a two-dimensional version. The match score can define two metrics: relevance and recovery. Bicluster relevance describes how well predicted biclusters match the ground truth and can be calculated. The bicluster recovery quantifies how well each bicluster matched the ground truth on average.

A detailed description of the method and evaluation can be found in the original publication [20] (Appendix A.1. The MoSbi algorithm was implemented as an R package and web service. Both can be used to execute all included biclustering algorithms, execute the ensemble workflow and visualize the results. The source code and web service links can be found in the publication [20] (attached to Section 4.1).

3.2. Creation of rule-based lipid networks for a functional lipidome interpretation

The LINEX is a framework to create, visualize and analyze lipid networks [22]. Due to a large number of theoretically possible lipid species in an organism, LINEX builds data-dependent

lipid species networks. Hence, it requires lipidomics data with quantified lipid species as an input. As already mentioned in the introduction, multiple lipid nomenclatures exist. We use the LipidLynxX software to convert them into one format [270].

The network is based on metabolic rules. They describe lipid class reactions that LINEX then extrapolates to lipid species. Fatty acid reactions can also be defined to visualize the fatty acid metabolism in the network. Class reactions are divided into headgroup and fatty acid-related reactions. A rule describing the reaction between PE and PS only affects the headgroup. Therefore, it requires two lipid species with the same fatty acid composition (or sum formula, if only sumspecies are provided). Fatty acid-related class reactions (e.g., LPA \rightarrow PA) require fatty acid composition for all but one fatty acid in the lipid with more fatty acids. For example, the rule can be extrapolated to the species reaction LPA(18:1) \rightarrow PA(16:0_18:1).

LINEX also offers the possibility of adding fatty acid reactions to the network. The fatty acid metabolism is usually independent of the lipid class metabolism, and elongations or desaturations do not occur for fatty acyls bound to a lipid. However, it can help to visualize fatty acid effects. For example, a desaturation in the network results in the following connection: LPA(18:0) \rightarrow LPA(18:1).

Creating a lipid network is only part of the LINEX method. Another important aspect is the combination of statistical measures shows quantitative changes in the lipidome. P-values of hypothesis testing can be shown as node color or size on the network. Correlations between lipids in one experimental condition of changes in correlation between conditions can be highlighted on the network edges. Systematic changes in parts of the lipidome can intuitively be visualized with this. Figure 3.2 shows examples for different LINEX visualizations.

A detailed description of the method and evaluation can be found in the original publication [22] (Appendix A.3). LINEX was implemented in the python programming language and comes with a web service, where users can upload lipidomics data and analyze the created networks interactively in the browser. Links to the source code and web service can be found in the publication [22] (attached to Section 4.3).

3.3. Deriving hypothesis for enzymatic dysregulation

Finally, Lipid Network Explorer 2 (LINEX²) was developed to create more comprehensive lipid networks and a network-based algorithm to interpret lipidomic changes [23]. In contrast to version 1, LINEX² changes the algorithm and knowledge base for the computation of data-driven lipid metabolic networks. Instead of relying on user-defined metabolic rules, curated lipid-metabolic reactions from public databases are used. Lipid-related reactions from the Rhea [271] and Reactome [272] database were parsed and curated. The databases were selected because Rhea is crosslinked with the SwissLipids database and includes specific molecular species reactions and Reactome because of its general overview of the lipid metabolism (but only on the class level). After curation, over 3000 reactions were annotated from both databases for various lipid classes, and organisms [23]. The annotation allows the construction of comprehensive lipid networks without advanced knowledge from users about lipid reactions.

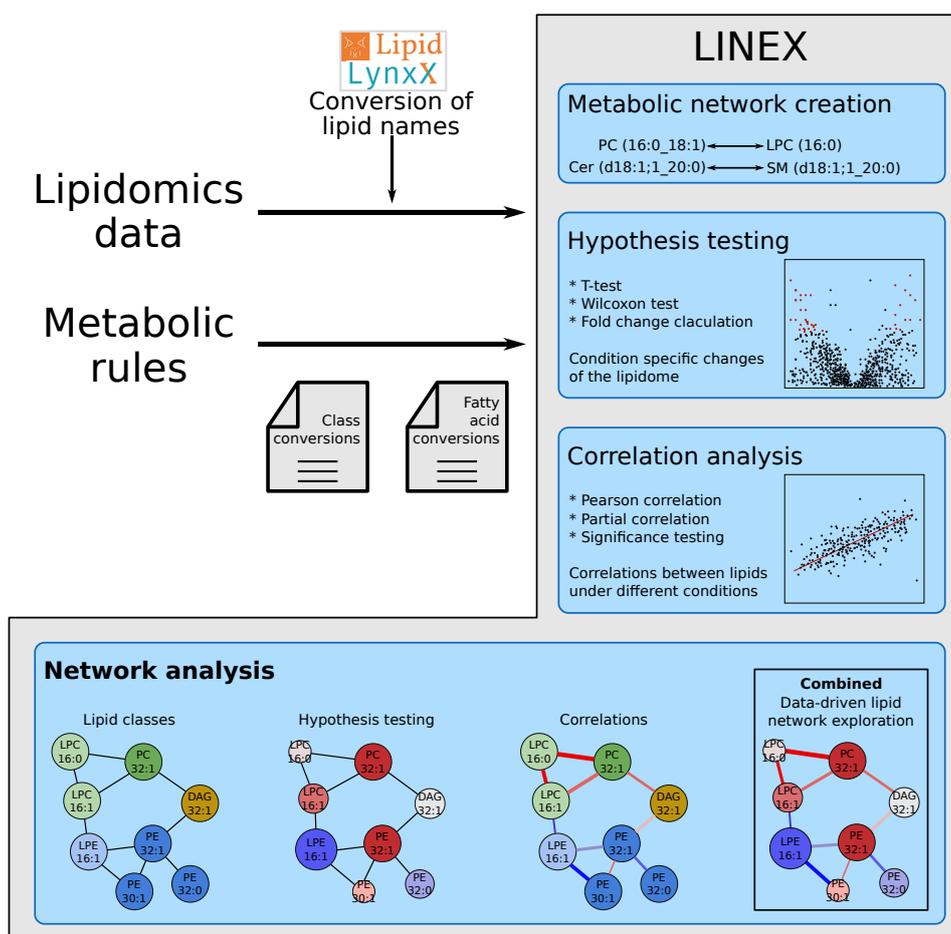


Figure 3.2.: Workflow of LINEX. Lipidomics data and metabolic rules are the input for the algorithm. This is used to create a lipid species network and combine it with statistical measures, such as hypothesis tests or correlation analysis. Unaltered figure taken from Köhler and Rose et al. [22]. Original figure distributed under the CC BY license (<https://creativecommons.org/licenses/by/4.0/>).

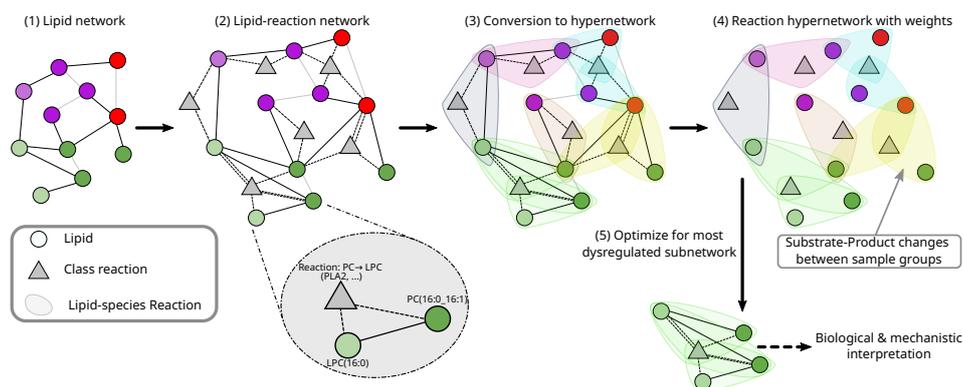


Figure 3.3.: Schema of the LINEX² network enrichment algorithm. A LINEX² network is converted into a hypergraph. Each hyperedge represents one lipid species reaction and connects substrates, products, and class reaction nodes. Hyperedges are weighted by changes in the substrate-to-product ratio between experimental conditions as an approximation for enzymatic dysregulation. A local network search is performed to find the maximally dysregulated subnetwork. Figure adapted from Rose et al. [23]. Permission granted by the authors.

Another novelty of LINEX² is a network enrichment algorithm for inferring enzymatic dysregulation from lipidomics data. The idea of the network enrichment is to utilize the multispecificity of the lipid enzymes in the network to infer systematic changes in substrates and products. The workflow of the algorithm is depicted in Figure 3.3.

First, lipid class reaction nodes are introduced to the network. They connect to lipids that participate in at least one lipid species reaction derived from the corresponding class reaction. This network is converted into a hypergraph, where each hyperedge represents a lipid species reaction with lipid-substrates, -products, and reaction nodes [23] (Figure 3.3 step 3). The hypergraph is transformed into a new representation, where hyperedges are nodes, and edges connect them if they share lipids or lipid class reactions [23]. In the following steps, each hyperedge is weighted by changes in the substrate-to-product ratio between experimental conditions. Substrate-product changes are calculated using the lipidomics data [23]. Finally, combined with simulated annealing, a local network search is employed to find the maximally weighted (dysregulated) subnetwork. A full description is available in the preprint (Appendix A.4).

For the interpretability of the solution, the workflow also includes the computation of a p-value, where (unconnected) lipid species are samples and their score computed [23]. The distribution of random solutions is then utilized to calculate an empirical p-value. It provides a measure of whether the computed subnetwork (mechanistic) has a significantly higher score to explain changes in the network than unconnected solutions [23].

A detailed description of the method and evaluation can be found in the preprint [23] (Appendix A.4). The LINEX² algorithm is available as a python package (<https://pypi.org/project/linex2/>) and web service (<https://exbio.wzw.tum.de/linex2>, source code:

<https://gitlab.lrz.de/lipitum-projects/linex>).

4. Publication summaries

4.1. MoSBI: Automated signature mining for molecular stratification and subtyping

Citation

"MoSBI: Automated signature mining for molecular stratification and subtyping" Tim D. Rose, Thibault Bechtler, Octavia-Andreea Ciora, Kim A. L. Le, Florian Molnar, Nikolai Köhler, Jan Baumbach, Richard Röttger, and Josch K. Pauling; In: *Proceedings of the National Academy of Sciences* 119.16 (2022): e2118210119; doi: <https://doi.org/10.1073/pnas.2118210119>

Summary

A variety of biclustering algorithms have been published in the last twenty years. However, selecting a specific algorithm for new data is challenging, since there is no universally fitting algorithms. Each algorithm has different heuristics and objectives, such that they can yield different insights into experimental omics data.

To address this challenge, we developed the biclustering ensemble method Molecular Signature identification using Biclustering (MoSBI). The method integrates the results of different biclustering algorithms into a network of bicluster similarities, from which highly overlapping bicluster communities can be extracted.

We evaluated the potential of the method for molecular disease stratification on clinical cancer data sets with known subtypes. This included transcriptomics, proteomics, and metabolomics data. In comparison to individual biclustering approaches, MoSBI showed the most consistent high performance in terms of recovering the known subtypes, and pathway enrichment of the resulting (ensemble) biclusters. We also evaluated the performance on simulated data with known properties. Our analysis did not show a clearly outstanding individual biclustering algorithm there, and again MoSBI was consistently high performing.

Furthermore, we investigated the applicability of MoSBI on multi-omics data. For this, we used a public breast cancer cohort, transcriptomics, micro RNA, and protein data for each patient. The method can be applied to each dataset independently and on the combined multi-omics data to consistently recover disease subtypes.

Overall, we showed that the ensemble method has advantages over individual biclustering algorithms and removes the need to parameterize them independently. Biclustering has great potential to improve the analysis of clinical omics data for exploratory research and

patient stratification. To make this analysis accessible for researchers of all fields, we made the method is available as an R package and web service.

Contribution

I developed the methodology and implemented it as an R package. I conceptualized, supervised, and co-implemented the evaluation of the algorithm on synthetic and experimental omics data. Further, I generated all Figures for the main manuscript, drafted the original version of the manuscript, and contributed to the revised manuscript.

As stated in the publication: "T.D.R., J.B., R.R., and J.K.P. designed research; T.D.R., T.B., O.-A.C., K.A.L.L., F.M., N.K., and J.K.P. performed research; T.D.R. and J.K.P. contributed new reagents/analytic tools; T.D.R., T.B., O.-A.C., K.A.L.L., F.M., N.K., and J.K.P. analyzed data; and T.D.R., N.K., J.B., R.R., and J.K.P. wrote the paper." [20].

Availability

The publication is available in Appendix A.1. "Copyright © 2022 the Author(s). Published by PNAS. This open access article is distributed under Creative Commons Attribution-NonCommercial-NoDerivatives License 4.0 (CCBY-NC-ND)." [20] (<https://creativecommons.org/licenses/by-nc-nd/4.0/>).

4.2. Nonalcoholic fatty liver disease stratification by liver lipidomics

Citation

"Nonalcoholic fatty liver disease stratification by liver lipidomics" Olga Vvedenskaya[†], Tim D. Rose[†], Oskar Knittelfelder, Alessandra Palladini, Judith A. H. Wodke, Kai Schuhmann, Jacobo M. Ackerman, Yuting Wang, Canan Has, Mario Brosch, Veera Raghavan Thangapandi, Stephan Buch, Thomas Züllig, Jürgen Hartler, Harald C. Köfeler, Christoph Röcken, Ünal Coskun, Edda Klipp, Witigo Von Schoenfels, Justus Gross, Clemens Schafmayer, Jochen Hampe, Josch K. Pauling, and Andrej Shevchenko; In: *Journal of Lipid Research* 62 (2021): 100104; doi: <https://doi.org/10.1016/j.jlcr.2021.100104>

[†] These authors contributed equally.

Summary

The non-alcoholic fatty liver disease (NAFLD) is common in western societies. It is characterized by a accumulation of neutral lipids in the liver and correlates with obesity. In some patients, the disease can transform into non-alcoholic steatohepatitis (NASH), which yields to fibrosis and inflammation of the liver. The disease is diagnosed by histological inspection of liver biopsies.

In this study, we aimed to investigate the lipidomic changes of the liver during disease progression. For this, shotgun lipidomics of 365 patients with healthy liver, obese patients, NAFLD, and non-alcoholic steatohepatitis (NASH) was performed. A strong increase of neutral lipids with disease progression was confirmed from this data.

Next, we utilized the ensemble biclustering approach to access changes of other parts of the lipidome. We identified bi-directional changes of Sphingomyelin (SM) species. Some increase with disease progression, whereas others decreased. Using these SMs as features for classification, we could differentiate for instance healthy from NASH patients, but failed to differentiate obese from NAFLD patients. However, this is crucial for a robust disease diagnosis. Therefore, we identified subgroups from the biclusters, after observing that many NAFLD patients clustered together with either healthy or NASH patients.

The subgroups, in combination with the SMs showed systematic changes of the disease progression with in patients diagnosed with NAFLD. Some NAFLD patients were from their lipidomic profile very close to obese patients, whereas others exhibited a profile similarly to NASH patients. The subgroups, identified from the lipidomics data were confirmed with clinical information about the patients.

Contribution

I conceptualized and performed the computational analysis (biclustering, classification, data normalization/correction) of the lipidomics data generated for this study. I contributed to the

interpretation of the results, made Figure 2-4, and participated in writing the manuscript.

As stated in the publication: "O.V., T.D.R., J.Hampe, J.K.P., and A.S. conceptualization; O.V., O.K., A.P., J.A.H.W., J.M.A., C.H., M.B., V.R.T., S.B., and J.K.P. data curation; T.D.R., A.P., J.A.H.W., J.M.A., and C.H. formal analysis; H.C.K., Ü.C., E.K., J.Hampe, J.K.P., and A.S. funding acquisition; O.V., T, D.R., O.K., Y.W., T.Z., and J.Hartler investigation; T.D.R., O.K., J.A.H.W., K.S., Y.W., J.Hartler, H.C.K., J.K.P., and A.S. methodology; O.V., M.B., J.Hampe, and A.S. project administration; M.B., S.B., C.R., W.v.S., J.G., and C.S. resources; T.D.R., A.P., J.M.A., C.H., and J.K.P. software; H.C.K., C.R., Ü.C., E.K., W.v.S., J.G., C.S., J.Hampe, J.K.P., and A.S. supervision; O.V., O.K., K.S., and Y.W. validation; O.V., T.D.R., and A.P. visualization; O.V., T.D.R., J.K.P., and A.S. writing–original draft; M.B., V.R.T., S.B., J.Hartler, H.C.K., Ü.C., E.K., and C.S. writing–review and editing." [21]

Availability

The publication is available in Appendix A.2. "© 2021 THE AUTHORS. Published by Elsevier Inc on behalf of American Society for Biochemistry and Molecular Biology. This is an open access article under the CC BY-NC-ND license (<https://creativecommons.org/licenses/by-nc-nd/4.0/>)" [21].

4.3. Investigating Global Lipidome Alterations with the Lipid Network Explorer

Citation

"Investigating Global Lipidome Alterations with the Lipid Network Explorer" Nikolai Köhler[†], Tim D. Rose[†], Lisa Falk, and Josch K. Pauling; In: *Metabolites* 11(8) (2021): 488; doi: <https://doi.org/10.3390/metabo11080488>

[†] These authors contributed equally.

Summary

Mapping of complex lipids onto metabolic reaction networks is challenging, because of the multispecificity of lipid enzymes (as explained in Section 2.2.1). Therefore, data-specific lipid networks are necessary, that generate a reaction network based on the identified lipids in an experiment.

In this publication, we presented the Lipid Network Explorer (LINEX), as software to generate lipid networks and visualize them in combination with statistical characteristics of the altered lipidome between two experimental conditions. Lipid species reactions are created from metabolic rules that serve as templates of lipid class of fatty acid reactions. This results in context-specific lipid networks that unite the class and fatty acid metabolism. The networks can then be used to visualize lipidomic changes between conditions. For example, with fold-changes, significance levels, or changes of correlation.

We applied the method to three public lipidomics datasets, to show its applicability and how novel insights can be gained from lipidomics data. On colorectal cancer lipidomics, we observed that individual lipids might not show significant changes, whereas in the network, similar patterns between metabolically closely related lipids can be found. Next, we discussed the importance of lipidomic coverage for computing data-specific lipid networks. And finally, on lipidomics data from the serum of aging humans, we showed how global lipidome alterations can be visualized with the methodology.

Another focus of the publication was, to provide web-based software that enables the easy use of LINEX to analyze lipidomics data. This is especially important because there is only a small number of lipidomics-specific software available to systematically investigate such data. Therefore, LINEX is an addition to existing software solutions but offers novel insights into global and local changes of the lipidome from a functional standpoint.

Contribution

I conceptualized the methodology, co-developed the software, and performed the application of the method on public lipidomics data (all equally with the co-first author N.K.). I also contributed to the writing, creation of Figures, and revision of the manuscript.

As stated in the publication: "Conceptualization: N.K., T.D.R. and J.K.P.; Software: N.K., T.D.R. and L.F.; Validation: N.K. and T.D.R.; Writing—original draft: N.K., T.D.R. and J.K.P.; Writing—reviewing & editing: N.K., T.D.R. and J.K.P.; Supervision: J.K.P." [22].

Availability

The publication is available in Appendix A.3. "© 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>)" [22].

5. Unpublished Results

5.1. Lipid network and moiety analysis for revealing enzymatic dysregulation and mechanistic alterations from lipidomics data

Availability

The following work was not peer-reviewed at the time of submission but is available as a preprint on bioRxiv [23]. Since this work was a significant part of my research, it is presented in this dissertation. In this chapter, a summary of the results in the preprint and my contribution are shown. The full preprint manuscript is available in appendix A.4.

Results

LINEX² builds on the previously developed Lipid Network Explorer (LINEX) [22]. This method was limited in several aspects. It required user-defined lipid class reactions, which require detailed knowledge about lipid metabolic pathways if reactions beyond the default ones are needed [23]. While LINEX provided a network visualization of the lipidome, no algorithm was available to mine it systematically.

LINEX² incorporates lipid class reactions from manually curated reaction database entries. With this, lipid networks incorporating over 90 lipid classes can be built. Network extension, the algorithm that creates data-specific lipid networks, can also evaluate biochemical reactions with more than one lipid substrate and product. Furthermore, a network enrichment algorithm for the lipid networks was developed. It uses the multispecificity of lipid enzymes to infer enzymatic dysregulation from lipidomics data (Figure 5.1A, the method is explained in 3.3).

As a proof of principle, the network enrichment was applied to liver lipidomics of mice under non-alcoholic fatty liver conditions [273]. The algorithm was used to find the maximum difference between WT and MBOAT7 knock-out samples. MBOAT7 catalyzes the acyltransferation of a fatty acyl CoA on Lyso-PI, producing a PI. It is known to have a strong preference for arachidonic acid as a substrate. The score optimization process can be seen in Figure 5.1B. It highlights the advantage of simulated annealing to overcome local maxima. The enrichment algorithm recovered the correct lipid class reaction from the lipidomics data (Figure 5.1C) and the lipid species reactions pointed towards long-chain highly unsaturated substrated. This strongly indicated that MBOAT7 was at the center of dysregulation. The LINEX² enrichment was the first algorithm to achieve this on lipidomics data. "*LINEX² cannot differentiate between the exact enzyme for this reaction. However, in contrast to e.g., PLA2, MBOAT7 only catalyzes LPI \rightarrow PI class reactions. Additionally, MBOAT7 is known for a higher affinity*

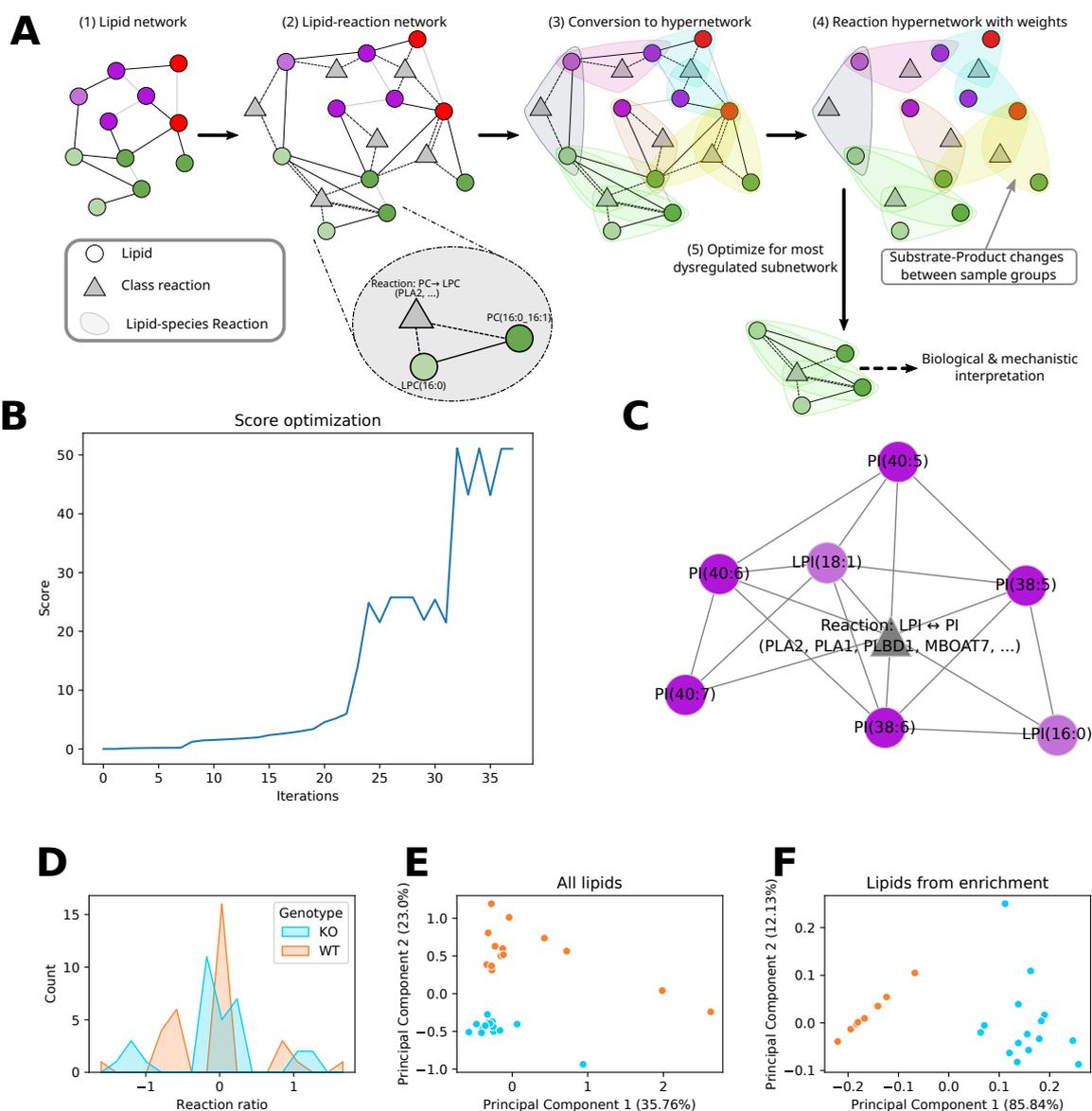


Figure 5.1.: (A) Workflow of the LINEX² network enrichment algorithm. (B) Score optimization of the network enrichment comparing wild type (WT) and MBOAT7 knock-out liver lipidomics. Data taken from Thangapandi et al. [273]. The same data is used for the plots (C-F). (C) Optimized subnetwork. (D) Ratio distribution of the LPI ↔ PI reaction for all molecular species reactions. (E) Principal component analysis of the lipidomics data with all lipids and (D) only for the lipids in the subnetwork shown in (C)

The figure is taken from the preprint of Rose et al. [23]. Permission granted by the authors.

for AA [...]. This preference can also be observed in the solution [...] for the edge between LPI(18:1) and PI(38:5), under the assumption that this reaction can only occur if the molecular composition of PI(38:5) is PI(18:1_20:4). Furthermore, all other reactions between LPis and PIs are only possible for the addition/removal of fatty acyls with at least 20 carbon atoms and 4 double bonds. These results are not surprising, because of the structural similarity of AA to other (very)-long-chain polyunsaturated fatty acids [...]" [23]. The results indicate that the network enrichment approach and knowledge about lipid metabolism can provide strong hypotheses for enzymatic dysregulation from lipidomics data.

In the next step, the overall distribution of substrate-to-product ratios for the LPI → PI class reaction was evaluated (Figure 5.1D). "The distributions show a peak around zero, indicating that many reaction ratios are not influenced by the MBOAT7 knock-out (KO). However, two more peaks around 1 and -1 can be observed for both conditions, where the peaks of the KO are shifted slightly more towards absolutely higher values. Despite these subtle differences, it is not possible to draw a hypothesis towards a mechanistic explanation including fatty acid-specific effects" [23]. This highlights the importance of the enrichment algorithm that can pick up on such fatty acid-specific effects. Principal component analysis of all lipids (Figure 5.1E) and lipids predicted by the enrichment (Figure 5.1F) was performed. It can be seen that both plots show a clear distinction between the WT and knock-out samples. However, in Figure 5.1F, this difference describes the primary source of variance (85.84%), in contrast to the analysis with all lipids, where the second principal component describes this [23].

In the preprint, LINEX² was also applied to lipidomics data of lean and obese adipocytes. Together with a lipid moiety analysis, we showed systematic differences in the neutral lipid metabolism and membrane lipids. The enrichment provided us with a hypothesis about changing PC to PE ratios that indicate membrane expansion for membrane lipids.

Overall, it shows that LINEX² can support the computational analysis of lipidomics data and automate analysis workflows to create hypotheses about mechanisms for lipidomic changes.

Citation

"Lipid network and moiety analysis for revealing enzymatic dysregulation and mechanistic alterations from lipidomics data" [Tim D. Rose[†]](#), Nikolai Köhler[†], Lisa Falk, Lucie Klischat, Olga E. Lazareva, Josch K. Pauling; In: *bioRxiv* (2022) 2022.02.04.479101 version 2; doi: <https://doi.org/10.1101/2022.02.04.479101>

[†] These authors contributed equally.

Contribution

As stated in the manuscript: "JKP supervised the project and secured the funding. NK, [TDR](#), and JKP planned and conceptualized the work. NK and [TDR](#) developed the web service. NK, OEL, and [TDR](#) designed and implemented the network enrichment procedure. LF, LK, and [TDR](#) parsed and curated the reaction databases, and implemented the network extension. NK and [TDR](#) applied,

5. Unpublished Results

validated, and interpreted the approach on lipidomics data. NK, OEL, TDR, and JKP wrote the manuscript. All authors read, reviewed, and accepted the manuscript in its final form."

6. Discussion

Computational methods are essential for identifying complex relationships from large-scale omics data. They can provide mechanistic insights into disease development and possible treatment targets in the clinical context. In this dissertation, two methods for the computational analysis of omics data were presented, the biclustering ensemble approach MoSBI and the LINEX/LINEX² framework for lipid network analysis. The goal was to enable systems medicine approaches that include lipidomics to get a new perspective of the molecular characteristics of diseases. Investigating alterations of the lipidome can be of great importance for understanding metabolic diseases. In the following discussion, I will go through all the presented work and discuss them in the context of contributing to systems medicine and computational biology. Furthermore, I will propose an integrated computational workflow, combining all approaches to show that they are not individual contributions to the field but can be applied together to tackle disease complexity. Finally, the challenges and prospects for integrating lipidomics into systems medicine research are discussed.

Interpretable subtyping with MoSBI

In the publication of MoSBI, I aimed to develop a computational method for patient stratification that works universally. The need to select algorithms, parameterize them, and make custom visualizations of the data should be reduced. Algorithms that require extensive parameterization and are challenging to execute can be used by the computational biology community but are less likely to be utilized by a broader scientific userbase [274]. To address these challenges, MoSBI was developed as a biclustering ensemble approach. In this way, MoSBI can profit from predictions of multiple biclustering algorithms. This is important since it allows MoSBI to build consensus biclusters from predictions of algorithms that identify different types of biclusters. This distinguishes MoSBI from previous biclustering ensemble approaches, which aimed at finding consensus biclusters for the same algorithms, which are applied multiple times to the same data [275, 276, 277]. In the publication, I showed that MoSBI could achieve a robust high performance without optimization of parameters for individual biclustering algorithms. However, I also found that individual algorithms can achieve significantly better results with the correct parameters [20], which are difficult to obtain since biclustering is an unsupervised method. Future work on the systematic parameterization of biclustering algorithms has the potential to make this class of algorithms even more useful for application on omics data. The applicability of MoSBI goes beyond just the performance of algorithms. I showed that it also drastically reduces the number of predicted biclusters by creating consensus biclusters of patterns in the data that were predicted multiple times and providing the frequencies of how often specific samples or

features were predicted. This can estimate the confidence for each data point as part of the predicted bicluster. An essential aspect is the network visualization of MoSBI. Interpretation of biclustering predicting requires a scalable visualization. Biclustering can be applied to disease subtyping in a systems medicine context, but interpreting many predicted biclusters is not trivial. The MoSBI visualization aims at aiding the interpretation by visually grouping similar biclusters and the possibility to color the network with other confounders. This increases the applicability of MoSBI in a clinical context since complex relationships to other health confounders can be visualized alongside the biclustering results. The network visualization can also be a basis for future biclustering algorithms and the consolidation of their results since it is a highly scalable method. Novel biclustering algorithms can also be easily added to the MoSBI framework. It can be beneficial to MoSBI predictions since they might find patterns in the data which were not found by any other biclustering algorithm. While the results showed that the MoSBI performance could be better than each algorithm individually, MoSBI predictions strongly rely on their predictions. Another contribution of the MoSBI publication for the biclustering/subtyping field is the development of a workflow to create simulated data to evaluate algorithms. The work showed that different data properties strongly influence performances. Future research in this field can reuse the pipeline to evaluate performances on various properties. This workflow allows biclustering algorithms to be compared on a diverse and standardized set of data characteristics.

Revealing heterogeneity in the NAFLD liver lipidome

Computational method publications usually focus on performance metrics and potential applications to showcase the developed methodology's usability. These controlled testing environments do not always represent the challenges of novel data analysis. In the NAFLD study [21], I was able to apply MoSBI to identify previously unknown disease subgroups based on lipidomics data. Starting from the biclustering results, we found characteristic Sphingomyelin (SM) alterations showing differences between the disease groups. Further, NAFLD patients were subgrouped based on their co-occurrence in bicluster communities with either healthy or NASH patients. Classifiers were trained on SM markers, to distinguish NAFLD subgroups. The analysis and results would not have been possible without the MoSBI workflow. Canonical clustering analysis might have identified similar subgroups, but a simultaneous detection of characteristic features would not have been possible. Using a single biclustering algorithm would have resulted in an incomplete set of predicted biclusters for which no appropriate visualization would have been possible that was crucial for identifying the subgroups. Network visualization of biclustering results, provided by MoSBI, made it possible to see a systematic co-occurrence of sample groups. The computational workflow of this publication can also serve as a blueprint for other clinical omics studies because biclustering analysis is not very commonly used. Starting with unsupervised (ensemble) biclustering analysis for feature selection and subtyping, followed by supervised classification for performance evaluation, and finally, validation of subgroups on additional clinical patient information. The results generated with MoSBI show the importance of lipidomics in NAFLD research. A potential limitation of the study is that only sum species for lipids were

systematically quantified. This level of identification was enough to provide many insights into the NAFLD liver lipidome. However, molecular lipid species can potentially reveal in even greater detail the functional associations of lipids. In the next section, I will focus more on the importance of structural lipid identification. In another publication on animal models of alcoholic fatty liver disease, MoS_{Bi} revealed characteristic signatures of Ceramide lipids. The researchers hypothesized a protective role of Kupffer liver cells in alcoholic steatosis [278]. This is another example that showed the usability of MoS_{Bi} for gaining insight from omics data.

MoS_{Bi} can identify characteristic features from any numeric data. This is not restricted to molecular data. In a clinical context, for example, electronic health records can be used to stratify patients based on their medical history. Features are not necessarily functionally related when working with omics data and predicting biclusters. Highly correlating features can be co-regulated but do not have to be. Biclustering methods incorporating interaction networks have been developed [245] for a more functional interpretation of the results. However, this also restricts the feature selection to potential biases in interaction networks [233]. For the functional interpretation of lipidomics data, we then developed LINEX.

Functional analysis of lipids using networks

LINEX aims to overcome the gap between identified lipid species and functionally interpret quantitative lipidome alterations. Systems medicine research integrating lipidomics data for drug target prediction and disease mechanism identification requires a functional interpretation of lipidomics data. By computing data-specific lipid networks, LINEX addresses a common discrepancy between identified lipids and database information about lipid metabolic reactions. For instance, the human metabolic network Recon3D contains 427 phospholipid reactions. However, it is still inaccurate because "the exact position of the double bond is not specified or the exact lipid composition of a phospholipid or triglyceride or sphingomyelin species is missing" [279]. This is also the case in other reaction databases, such as Rhea [271]. Reactions are often undefined for fatty acyls (marked as "R", e.g. in RHEA:10604) or specified for certain fatty acyls but are incomplete (e.g. RHEA:46360) and lack the observed broad multispecificity of lipid enzymes [94, 95]. LINEX solves this problem by constructing data-specific lipid networks from lipid class reactions, assuming that a lipid species reaction can occur for all identified possible substrate-product pairs. LINEX² provides a significant update to the previous version since it works out of the box with various lipid classes and is supported by the knowledge from reaction databases. The idea of creating data-specific lipid networks is not new and has been previously implemented in the BioPAN software [259]. A significant difference is that BioPAN operates only on the sum species level. With improvements in lipidomics technology, molecular species identification becomes more common and therefore requires computational methods that can work with this type of data.

While data-specific lipid networks are advantageous for accurately depicting likely reactions in the measured samples, it is also a clear limitation. Especially low abundant lipids might not always be identified in experiments. This creates gaps in the networks that have technical reasons. As technology evolves with better separation techniques and more sensitive

mass spectrometers, this will become less of a problem. Until a complete lipidome can be reliably measured, inferring nodes to improve network connectivity might be a solution. In metabolism research, inference is often used to construct metabolic networks from genomic data [280] or predict networks from identified metabolites [281]. Inference in lipid networks could profit from prior knowledge about reference lipidomes from similar organisms/tissues or measured fatty acid compositions. Prediction of likely lipid species that improve the network connectivity could also benefit experimentalists, which can try to confirm those predictions in targeted lipidomics and improve their lipidomics coverage.

After computation of a lipid species network, LINEX combined the network visualization with statistical measures to facilitate interpretation of the data. We showed that such a network visualization could reveal systematic changes in the lipidome and put them into a pathway context [22]. The LINEX approach to put the interactive network visualization in the center and project additional characteristics onto the network for an exploratory functional analysis is novel in the computational lipidomics field. Usability for lipidomics researchers was also a high priority in the development to make the workflow accessible. Recently, Pérez-Martí et al. [282] visualized compositional fatty acid saturation changes on a lipid class network. Borgmeyer et al. [283] mapped lipid species to a Reactome and KEGG lipid class network to show associations to enzymes. This shows that other researchers also utilize networks' potential to assess the lipidome. LINEX can profit from these ideas and could be updated in the future to include innovative approaches for lipidome analysis on one platform. This was already done by adapting a glycan substructure methodology [284] for lipid moieties and integrating it with the LINEX framework.

One of the main goals of LINEX² was the development of an enrichment methodology that can infer enzymatic dysregulation from lipid networks. We showed how a proof of principle of MBOAT7 knock-out data and created hypotheses for lipid alterations in adipocytes [23]. LINEX² assumes that a few changing multispecific enzymes should impact many lipid species reactions. In future work, probability distributions for reaction changes can be used to more accurately depict the candidates for enzymatic dysregulation systematically over the entire network. The LINEX² enrichment methodology might be improved with more quantitative data about lipid enzyme specificity. So far, LINEX² predicts candidates from dysregulated enzymes but cannot automatically rank those. With more knowledge about, e.g., fatty acyl preferences of lipid enzymes, enzyme candidates could be ranked based on the fatty acid composition of the lipids in the predicted subnetwork. Preferably, quantitative enzyme dynamics should be measured experimentally, but docking or molecular dynamics simulations also could provide approximations for these values. In contrast to MoSBI, LINEX/LINEX² was only previously validated on previously published data were used. As already explained, such data does not always reflect the challenges of new experimental data. Future studies will have the chance to prove the approach's potential in new data.

In a systems medicine context, inferring mechanisms can help uncover potential drug targets. The possibility of generating hypotheses of potential mechanisms is crucial to understanding diseases. Therefore, LINEX² is a critical contribution to integrating lipidomics into the field of systems medicine. It can also be a basis for integrating lipidomics with other

omics disciplines, such as proteomics data.

Computational drug repurposing

Revealing disease mechanisms and subgroups is essential for understanding diseases, but the final goal is treatments. As already presented in the introduction, drug repurposing can be a possible solution that is cheaper than new development and certification of drugs. I contributed to a publication about drug repurposing in SARS-Cov-2 [27] (contribution shown in appendix A.5.1). In this work, we presented a network-based approach to finding drug targets inhibiting human interactions with viral proteins. A web interface increased the accessibility for every researcher to create hypotheses for follow-up experiments or simulations. However, despite many efforts in computational drug repurposing, the predictions did not make it into clinical trials. We address this discrepancy in a review paper, to which I contributed [26] (contribution shown in appendix A.5.2). In this work, we identified the lack of standardized data and computational results as the primary source for this problem. As a solution, we proposed a unified drug repurposing strategy. It includes standardized data, accessible workflows for computational repurposing methods, the proposal of combinatorial treatments, expert guides analysis, and experimental candidate validation. Of the many studies we reviewed, most included only parts of this strategy and had no follow-up research that experimentally investigated the proposed candidates. Using the unified repurposing strategy, I believe drug repurposing has a vast potential for systems medicine.

An integrated computational workflow

The computational methods presented in this dissertation work independently and address different research questions. However, a systems medicine workflow is more complex and cannot be addressed by individual methods that solve only specific problems. The previously mentioned steps of subgrouping, biomarker detection, and disease mechanism identification require multiple methods that work hand in hand. The approaches MoS_{Bi}, LINEX/LINEX², and drug repurposing are parts of combined workflow to tackle disease complexity. This combined workflow is shown in Figure 6.1.

For gaining insight from (multi-)omics data of a disease cohort, MoS_{Bi} is an initial exploratory analysis step. Different omics datasets can be analyzed individually or combined, as we showed in the publication [20]. This results in patient subgroups with characteristic signatures. In combination with supervised learning, predicted signatures can serve as biomarkers for stages or subtypes of a disease. Such an application was already depicted in the NAFLD publication [21]. In the next step, LINEX² can be utilized to interpret lipidomic changes between subgroups. Of course, this is only possible if lipidomics has been measured. Especially for metabolic diseases, functional lipidomic differences between subgroups within a disease are essential to understand. Lipid networks can reveal systematic alterations in the lipidome composition between subgroups. Trends towards certain lipid classes or fatty acyl desaturation can indicate changes in cellular metabolism. With the enrichment analysis, hypotheses for altered enzyme activity can be generated. This can provide potential

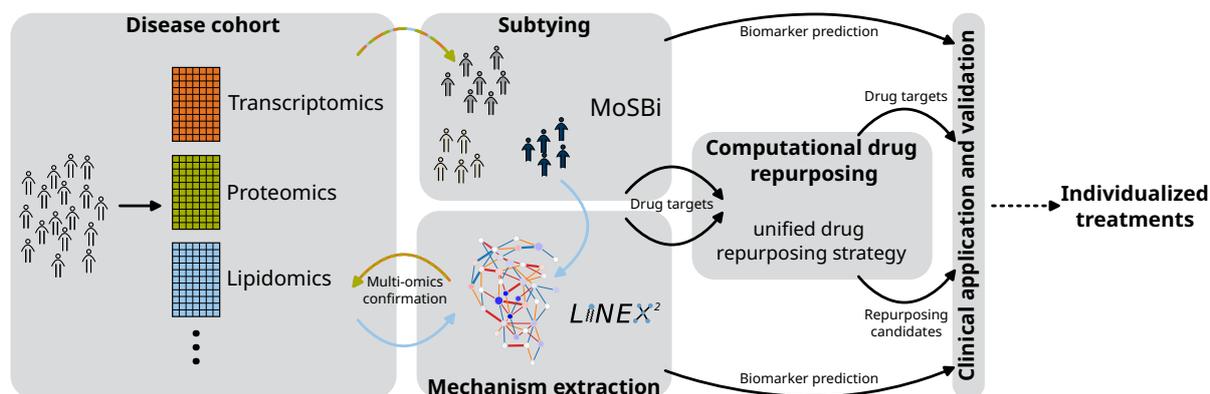


Figure 6.1.: Integration of MoSBI, LINEX², and computational drug repurposing in a systems medicine context for a disease cohort with multi-omics data.

mechanistic explanations for some observed differences in lipid abundances. The functional analysis of lipidomics data can be used to refine biomarker predictions by putting them into an interpretable and mechanistic context.

Hypotheses from LINEX² can then be supported or falsified with transcriptomics or proteomics data. LINEX² predictions also do not include a directionality, meaning if the predicted enzymes increase or decrease activity. This can be further investigated with enzyme protein or transcript abundances. Lipid-enzyme dysregulation can also be combined with transcriptional regulatory networks or signaling pathways. Such analysis can be based on interaction networks in combination with characteristic subgroup features predicted by MoSBI.

Predictions of MoSBI and LINEX² together with further downstream analysis can be used as potential biomarkers and serve as drug targets. This is where computational drug repurposing starts. As presented in the review [26], such repurposing could be done based on interaction networks, where drugs are predicted that target proteins directly or indirectly interact with the protein of interest. We presented several options for utilizing such an interaction network-based analysis workflow [27]. Another option is based on docking between potential drugs and proteins. Computational drug repurposing requires extensive experimental validation but can significantly decrease the time to bring potential treatments to the clinic [26].

An integrated workflow of all computational methods, part of this dissertation, for accessing complex diseases can benefit future individualized treatments. In every molecular biological research, where samples can be subgrouped and lipidomic alterations are investigated, a combined application of MoSBI and LINEX² can offer additional insights into lipidomic regulation.

Challenges of lipidomics for clinical research

Lipidomics is playing a growing role in clinical research. Cancer research has shown its potential to facilitate understanding, prognosis, and treatment [285, 286, 287]. Similarly,

it is gaining importance in metabolically-related disease research such as diabetes and cardiovascular disease [288]. For non-alcoholic fatty liver disease (NAFLD), we showed that lipidomics could also be used for disease stratification and identifying potential biomarkers [21]. The work can serve as a starting point for adding another dimension to the diagnosis of the disease, typically based on histology. Lipidomics can provide more quantitative support for the different stages of the disease.

However, the systematic use of lipidomics in clinics has a few obstacles. This is summarized very well by a publication title of Liebisch et al. [133]: "*Reporting of lipidomics data should be standardized*". The work was published in 2017, but problems in lipidomics data standardization continue to exist. The authors discuss many aspects of lipidomics, but here I want to focus on the lacking standardization for computational research. Different nomenclatures for lipids are usually the first challenges that occur in the development of lipidomics software. The introduction has already mentioned nomenclatures (Section 2.2). Methods such as LINEX require a standardized format of lipid names to work. Software packages that address this problem have been developed and can convert between nomenclatures and decompose complex lipid structures [270, 289]. However, this requires constant software updates when new nomenclatures are added or updates in naming conventions must be implemented. For new researchers entering the field, these differences can cause challenges for the usage of computational methods.

Another challenge that impacted this dissertation's work was the lack of standardization in lipid databases. For example, LIPID MAPS [290] and SwissLipids [291] are two comprehensive lipid databases that list various information for lipid species. They are cross-linked, meaning if a lipid is available in both databases, they contain the link to the other resource. But there are also inconsistencies. For instance, LIPID MAPS contains only molecular species with a defined sn and double bond positions. In current lipidomics experiments, lipids are typically not identified on such a level. Developing software that can consistently work with all databases is not straightforward.

The situation is even more challenging for lipids in reaction databases. Earlier in the discussion, I mentioned inconsistencies in reaction databases, where some reactions are confirmed for certain molecular species. In the Reactome database [272], reactions are mainly provided on the lipid class level but have varying identifiers for the same lipid classes. For LINEX², this required an extensive manual curation of the resources to make them compatible for usage in the algorithmic framework.

For the processing of lipidomics data, standardization is another crucial aspect. In the NAFLD study [21], I observed strong batch effects related to the lipid extraction date of the samples. For such a significant cohort, sample processing usually has to be spread over several days, processed by several people, or chemical solutions have to be renewed in between. More sources for batch effects can be drift in MS sensitivity, LC column performance, or ionization efficiency, which are likely to occur in big cohorts where the instrument is running for several days [292, 293]. All these factors influence the peak size and position during mass spectrometry. Our study showed that internal lipid standards could help reduce batch effects but not entirely. Standardized reporting of all experimental parameters, randomization

during sample collection, and regular quality control runs will make lipidomics experiments more reproducible. As we did in the NAFLD study, batch effects can be computationally removed but require knowledge of all potential sources.

With the foundation of the Lipidomics Standards Initiative [294], the field moves towards more standards and naming conventions. This will also facilitate the development of new computational approaches for lipidomics data analysis. As discussed in the drug repurposing review [26], standardization is not only a problem of lipidomics. We identified this lacking standardization also as the main obstacle to the success of drug repurposing. Standardization has advantages in identifying lipids and connecting them on a systems biological level. However, these standards always need to be adapted to allow CB approaches to incorporate more detailed information. The latest developments in lipid fragmentation using electron-activated dissociation make it possible to resolve lipid structures fully [295]. New efforts must standardize the terminology for these fragmentation patterns such that identification software can utilize them. Also, downstream lipidomics analysis approaches must be adapted to work with such data.

A general challenge of computational biology that is not only limited to lipidomics is non-standardized software interfaces and data types. Such incompatibilities often require individual solutions for each computational biology research project, where methods must be combined into one pipeline. There are several approaches to make bioinformatics methods compatible, such as the scverse for single-cell data [296] or the systems biology markup language. Of course, there are always new data types and computational goals which cannot always be compatible with other methods. However, using standard data formats and providing interfaces for specific downstream analysis can make the tools more usable for researchers, prevent reimplementations, and reduce mistakes because of a limited understanding of the software. A central part of MoSbi is the compatibility layer for all the included biclustering methods. This makes it possible to analyze all biclustering predictions using a standardized bicluster format. LINEX uses compatibility software for lipid nomenclatures to work out of the box with various lipid formats.

Prospects for computational lipidomics

Computational analysis for the functional interpretation of lipidomics data is still in its infancy. Integration of lipidomics with other data such as transcriptomics or proteomics is typically done individually [297]. LINEX² is a step toward a multi-omics integration of lipidomics data. While the approach itself cannot be used directly for multi-omics data, networks with annotated enzymes have the potential for network integration. Network models, such as Recon 3D, have already started integrating lipid species reactions in genome-scale metabolic models [279]. However, data-specific lipid networks are more flexible and can account for newly identified lipids in experiments. Lipid network enrichment in combination with multi-omics data, e.g., supported by enzyme abundances, can improve the quality of hypothesis generation.

New perspectives for computational lipidomics come with the more widespread use of MS2. Only with fragmentation can lipid species be resolved structurally. This usually means

identifying the fatty acid composition. With current technology, lipids can be fully structurally identified [298, 295]. Computational methods can utilize this information to provide an in-depth analysis of the lipidome. This is especially important for understanding changes in the fatty acid metabolism, where double bond positions have an impact on the function [299]. Complex lipids result from fatty acid metabolism, lipid class metabolism, and all their metabolic precursors. Therefore, diseases can have an impact on all these moieties of lipid species. A complete understanding of the dynamics of the lipidome can only be achieved if a significant proportion of the lipidome can be quantified on a structural level. MS² is crucial to achieving this. Fully characterized lipid species are also a basis for dynamic modeling of the fatty acid and complex lipid metabolism because they reflect the structural differences that affect enzymatic constants at the lowest level. Of course, this only works with detailed functional information about lipid reactions in databases, which are currently many on the lipid class level. This probably has historical reasons. Before using MS for lipidomics, TLC was typically used for measuring lipids. With this technique, lipids were only identified on the lipid class level. Therefore, lipid enzymes were also only studied at the lipid class level. In the future, quantitative enzyme assays are necessary that provide quantitative information about the association constants for molecular lipid species. This will be of great value for the computational analysis of lipid metabolism.

Another area for novel computational methods is spatial lipidomics, also known as MS imaging. Spatial metabolomics/lipidomics can reveal, for example, heterogeneous metabolic profiles within cancers [300]. Omics independent workflows and methods for spatial data have been developed (e.g. [301, 302, 303]), which can also be applied to lipidomics. Functional analysis for spatial lipidomics is more challenging because MS imaging is usually only performed on the MS1 level without fragmentation. This does not allow the identification of the fatty acid composition of complex lipids and limits the possibilities for functional analysis and interpretation. Spatial MS2 lipidomics could enable a detailed functional understanding of lipidomic processes on the sub-tissue level. For transcriptomics, a combination of single-cell data has been used to decompose cell types in spatial transcriptomics [304]. For lipidomics, high-resolution lipid identification on bulk data could be used to infer fatty acid compositions of spatial lipidomics. Computational analysis of spatial and single-cell lipidomics, typically based on MS imaging, [187, 305] has an enormous potential to model the heterogeneity within tissues or between cells but has to overcome many challenges, including sparsity and lower levels of identification.

Computational approaches for personalized medicine

Incorporating molecular information into therapy and diagnosis is the primary goal of personalized or systems medicine. As explained in the introduction, computational methods are integral to developing systems medicine approaches. They can create hypotheses for molecular disease mechanisms, patient stratification, or prognosis prediction. A big challenge of bringing patient stratification into the clinics is interpretability [210]. Computational decision-making must be understandable if physicians base their decisions on them. Therefore, many approaches focus on the explainability of their results [306, 210]. Consensus methods,

such as MoS_{Bi} can also be a solution since they base predictions on the results of multiple algorithms.

Artificial intelligence methods based on neural networks are also gaining relevance for biomedical research. They are commonly used to interpret imaging data, predict biochemical structures, or discover drugs [307]. They can yield more accurate predictions with enough data available but often lack explainability. Computational methods are not the only limiting factors for personalized medicine. Data bias, socio-environmental factors, and data privacy can impact research [308].

Another area where metabolism research is becoming more critical for personalized medicine is drug research [309]. The effect of a drug depends not only on the performance of the active substance but also on how the patient metabolizes it. This can be modeled with pharmacokinetic models that try to predict drug safety and efficacy [310]. Also, the gut microbiome plays a crucial role in drug metabolism [311]. All these factors have to be considered for a truly personalized drug treatment. Computational models can potentially integrate various information to predict the optimal dosage or substance for each patient if enough data is available.

The importance of data availability for systems medicine research cannot be overstated. For instance, where large amounts of samples are stored, biobanks can provide cohorts for medical research [312]. Of course, privacy concerns regarding sensitive patient data have to be considered. Federated machine learning approaches have been developed, which spread the training of computational models over several institutions and generate consensus models [313]. Only model parameters and not patient data are then shared. In this way, sensitive data does not leave, e.g., a hospital, but can be used to improve computational predictions in a privacy-conserving way by multiple research institutions.

7. Conclusion

In this dissertation, computational strategies for the functional analysis of lipidomics data and disease subtyping have been developed. The work was driven by the lack of approaches to integrating lipidomics into systems medicine workflows. Lipids and their metabolism are highly relevant, especially for metabolic diseases, and can yield new insights into disease processes and serve as markers. Approaches to analyzing vast amounts of lipids must be interpretable to be of value for clinical research.

To address these issues, two computational frameworks were presented. The MoSBi method is a biclustering ensemble method that reduces the need for the selection of biclustering algorithms. With its network visualization, results from large datasets can be efficiently visualized and interpreted. Its ability to stratify patients *de novo* was proven on lipidomics data from a NAFLD cohort. This analysis also revealed potential biomarkers for the disease progression. The LINEX framework was presented, which can functionally interpret lipidomics data by creating data-specific lipid metabolic networks. With a novel enrichment algorithm, hypotheses for enzymatic dysregulation can be predicted from lipidomics data. Further, I presented contributions to computational drug repurposing, which is an essential step from disease mechanisms to precision treatments.

Both MoSBi and LINEX/LINEX² are crucial advances for disease subtyping and functional lipidomics analysis. In an extensive evaluation of synthetic and experimental omics data, MoSBi showed robust and scalable results. With the work on NAFLD stratification, I showed that MoSBi is suitable for *de-novo* patient stratification and explorative analysis of large clinical cohorts. While the lack of alternative methods prevented the systematic comparison of LINEX, the power of the approach was shown on knock-out data and various clinically relevant lipidomics datasets. All methods were designed, keeping accessibility and interpretability in mind. By utilizing networks in both methods, complex relations can be visualized to reveal global and local relationships. In LINEX, this is achieved by adding statistical measures to networks, and in MoSBi through visualizing clinical or experimental confounders. LINEX also showed that networks offer a benefit for lipidomics analysis, which can reveal previously unknown trends. Furthermore, LINEX² serves as a proof of principle that it is possible to derive hypotheses of enzymatic dysregulation from lipidomics data through an automated computational algorithm. Networks are also crucial for the presented drug repurposing method. They make it possible to visualize drug targets, drug candidates, and related proteins. With web services, these advanced algorithms are accessible to all researchers. Furthermore, MoSBi can be directly used for a multi-omics analysis, and functional connections between lipids and enzymes, provided by LINEX², can be a starting point for downstream analysis with proteomics or transcriptomics data.

The approaches can systematically provide new insights but are, of course, also limited

in specific aspects. MoSBi overcomes algorithm specificities by integrating their results but still relies on the performance of each algorithm. As shown, perfect algorithm parameters, which are difficult to estimate, can significantly improve the predictions. Patterns captured by none of the included algorithms will also not be a part of MoSBi predictions. Therefore, novel biclustering algorithms can still be valuable for MoSBi. A central limitation of LINEX and LINEX² is the creation of lipid networks. The lipid metabolism is of high complexity, which can only be partially evaluated in the process of computing networks. This makes creating networks for less studied parts of the lipidome challenging. The problem also holds for the enrichment, which is based on these networks and can potentially propagate this bias. Large-scale quantitative lipid enzyme assays can help to overcome this. Computational drug repurposing also comes with many limitations, which are the main focus of the review I contributed to [26]. These mainly lack standardized data, reporting, and experimental evaluation of predicted repurposing candidates.

Despite the mentioned limitations, my work showed new workflows to mine clinical (lipid-)omics data. Automatically creating hypotheses for enzymatic dysregulation from lipid data can provide opportunities for applying lipidomics in clinical research. This includes biomarker detection and identifying disease mechanisms related to lipid metabolism. With MoSBi biclustering is more accessible for disease subtyping. As shown on liver lipidomics of NAFLD patients with MoSBi, the approaches can promote a molecular understanding of highly relevant diseases in our modern society. Potential applications of MoSBi go even beyond molecular biology. It can work with any numerical data to stratify samples, e.g., for disease subtyping with electronic health records. MoSBi can be utilized for multi-dimensional clinical research to advance disease understanding. With LINEX² high-resolution lipidomics data can be mined to study perturbations of the lipid metabolism. This gives lipidomics research an even higher potential to identify disease mechanisms directly related to lipid metabolism. Analytical progress will make it possible to identify all structural features of lipids systematically. From this LINEX² can profit by providing more accurate lipid network representations that will make it possible to discover more about the regulation of the lipidome. Applications for the methods are also outside systems medicine. In all areas of systems biology research, molecular mechanisms can be identified with MoSBi and LINEX. With all the presented methods, a systems medicine workflow shows how clinical research benefits from computational methods to systematically gain knowledge about the molecular state of diseases.

Computational biology has the potential to have a significant impact on the future of medicine. Data availability and algorithmic advances facilitate new applications, but many challenges in standardization, explainability, and robustness must be overcome. Lipidomics still faces many challenges before it can be routinely employed for precision medicine [314]. However, it has vast potential to provide biomarkers or prognostic indicators [315]. With an increasing prevalence of metabolic diseases, the need to understand lipid-based disease mechanisms and markers will also increase. Patients will profit from this by receiving more personalized diagnoses and precise treatments with fewer side effects. With the work of this dissertation, I hope to contribute to making lipidomics analysis more interpretable and

7. Conclusion

facilitating patient stratification. I look forward to seeing computational biology thrive in the future to provide more possibilities for understanding the nature of molecular biology and diseases.

A. Appendix

A.1. MoSBI: Automated signature mining for molecular stratification and subtyping



MoSBI: Automated signature mining for molecular stratification and subtyping

Tim Daniel Rose^a, Thibault Bechtler^a, Octavia-Andreea Ciora^a, Kim Anh Lilian Le^a, Florian Molnar^a, Nikolai Köhler^a, Jan Baumbach^b, Richard Röttger^c, and Josch Konstantin Pauling^a

Edited by David Donoho, Stanford University, Stanford, CA; received October 4, 2021; accepted February 28, 2022

The improving access to increasing amounts of biomedical data provides completely new chances for advanced patient stratification and disease subtyping strategies. This requires computational tools that produce uniformly robust results across highly heterogeneous molecular data. Unsupervised machine learning methodologies are able to discover *de novo* patterns in such data. Biclustering is especially suited by simultaneously identifying sample groups and corresponding feature sets across heterogeneous omics data. The performance of available biclustering algorithms heavily depends on individual parameterization and varies with their application. Here, we developed MoSBI (molecular signature identification using biclustering), an automated multialgorithm ensemble approach that integrates results utilizing an error model-supported similarity network. We systematically evaluated the performance of 11 available and established biclustering algorithms together with MoSBI. For this, we used transcriptomics, proteomics, and metabolomics data, as well as synthetic datasets covering various data properties. Profiting from multialgorithm integration, MoSBI identified robust group and disease-specific signatures across all scenarios, overcoming single algorithm specificities. Furthermore, we developed a scalable network-based visualization of bicluster communities that supports biological hypothesis generation. MoSBI is available as an R package and web service to make automated biclustering analysis accessible for application in molecular sample stratification.

stratification | biclustering | subtyping | multiomics | pathomechanism

Optimizing treatments and improving patients' health is the goal of precision medicine. In contrast to canonical medicine, where treatments are prescribed empirically (1), precision medicine aims to identify individually adapted treatments. Nowadays, diseases are commonly diagnosed based on the International Classification of Diseases. This assumes that diseases show similar symptoms in every individual; hence treatments are meant to act on the majority of symptoms. Patient stratification for precision medicine builds on the idea that a cohort of patients with varying or similar symptoms might have different molecular causes. They can then be stratified on the molecular level and divided into subgroups (2). Therefore, precision medicine wants to move away from classical disease definitions to characteristic signatures of molecular alterations which enable individualized treatments.

Achieving this requires an understanding of molecular disease mechanisms. Unsupervised machine learning methods are best suited since they uncover the inherent structure of the given data and do not require labeled data, which might be biased toward classical disease understandings (3). Unsupervised clustering methods seek to identify distinct subgroups over the entire features set, but it is unrealistic to assume that diseases manifest in all features. Instead, they are limited to a subtype-specific subset. Biclustering algorithms can meet this requirement.

Molecular data is usually available in data matrices with patient samples as columns and biomolecular features as rows. Biclustering algorithms cluster samples and biomolecules of a data matrix simultaneously. This results in sample groups with a molecular subset that characterizes the group. Numerous algorithms have been published, which try to tackle the problem from different angles. An overview of important concepts was published by Madeira and Oliveira (4).

Similar to clustering (5), evaluations of biclustering algorithms have shown differences in performance under various real-life and synthetic conditions (6, 7). A common way to improve the results of machine learning techniques is ensemble approaches, for example, for biomarker discovery (8). The goal is to improve robustness, consistency, novelty, and stability over what single algorithms could achieve (9). Also, for biclustering problems, ensemble algorithms have been proposed (10–16). Most of these ideas are adaptations of approaches for ensemble clustering. Some of the proposed methods have not been

Significance

Molecular patient stratification and disease subtyping are ongoing and high-impact problems that rely on the identification of characteristic molecular signatures. Current computational methods show high sensitivity to custom parameterization, which leads to inconsistent performance on different molecular data. Our new method, MoSBI (molecular signature identification using biclustering), 1) enables so far unmatched high performance for stratification and subtyping across datasets of various different biomolecules, 2) provides a scalable solution for visualizing the results and their correspondence to clinical factors, and 3) has immediate practical relevance through its automatic workflow where individual selection, parameterization, screening, and visualization of biclustering algorithms is not required. MoSBI is a major step forward with a high impact for clinical and wet-lab researchers.

Author contributions: T.D.R., J.B., R.R., and J.K.P. designed research; T.D.R., T.B., O.-A.C., K.A.L.L., F.M., N.K., and J.K.P. performed research; T.D.R. and J.K.P. contributed new reagents/analytic tools; T.D.R., T.B., O.-A.C., K.A.L.L., F.M., N.K., and J.K.P. analyzed data; and T.D.R., N.K., J.B., R.R., and J.K.P. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

Copyright © 2022 the Author(s). Published by PNAS. This open access article is distributed under Creative Commons Attribution-NonCommercial-NoDerivatives License 4.0 (CC BY-NC-ND).

¹To whom correspondence may be addressed. Email: josch.pauling@tum.de.

This article contains supporting information online at <https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2118210119/-DCSupplemental>.

Published April 11, 2022.

implemented (10, 13) and therefore not easily accessible, while others are single-algorithm ensemble approaches that cannot overcome the limitations of one algorithm.

The analysis and interpretation of biclustering results can profit from visualizations, which show the content or relations between biclusters. Many approaches have been developed (17–23), which are often bound to specific algorithms or do not scale well for many biclusters (18).

Here we propose a multialgorithm biclustering ensemble approach for the stratification of molecular samples. In the manuscript, we 1) introduce the methodology and network visualization; 2) evaluate the performance on multiple experimental metabolomics, proteomics, and transcriptomics datasets; 3) with a framework for synthetic data generation, evaluate the approach on synthetic data; 4) apply our approach in a multiomics context; and 5) present open-source software to make biclustering more accessible for research.

Results

A Multialgorithm Ensemble Biclustering Approach. The steps of our ensemble approach (MoSbi—molecular signature identification using biclustering) are described in Fig. 1A; for full details, please refer to *Materials and Methods*. At first, we selected a set of established or recently developed biclustering algorithms (Table 1), which are executed independently. Next, similarities between all biclusters are calculated. The similarity is described by the degree of overlap, meaning the more samples and features shared between biclusters, the higher their similarity. Highly similar biclusters point toward the same pattern in the data. Similarities are filtered for random overlaps, and a bicluster network is generated with biclusters as nodes and connections between them if they exceed a higher than random similarity (for details, see *Materials and Methods*). This removes overlaps of biclusters that are likely to occur randomly and do not carry meaningful overlaps. The same network without the filtered random overlaps is shown in *SI Appendix, Fig. S1*. While biclusters with similar disease subtypes are still close together, the overall connectivity in the network is significantly higher. The example network shown in Fig. 1A reveals several highly connected communities in the network, which are not as strongly connected with each other. By using the Louvain modularity, such communities can be extracted and converted into ensemble biclusters. Two thresholds control the size of the resulting ensemble biclusters. We previously successfully utilized the principle of MoSbi to identify *de novo* subtypes of nonalcoholic liver disease based on clinical lipidomics data (34).

Before evaluating the performance of MoSbi on multiple omics datasets, we selected a public thymic epithelial tumor dataset (35) to show the application and potential of our approach. Ku et al. (35) measured the proteome of 134 tumor, tumor-adjacent, and normal thymus samples and revealed significant differences in the proteome signatures of thymoma subtypes. In Fig. 1A, the similarity network of predicted biclusters colored by sample groups can be seen. Node sizes were scaled according to the number of samples. It provides an overview of the match of predicted biclusters with known information about samples, in this case, cancer subtypes/tissues. While being a central part of the workflow, to compute ensemble biclusters, networks also serve as a visualization of biclustering predictions.

It is immediately obvious that clusters of nodes can be found in the network, indicating biclusters with a similar set of samples and features. This can be observed by similar color distributions of biclusters clustering together. The clusters show high

intraconnectivity, but also connections to other clusters. This means that some signatures are shared between network communities. After applying the Louvain modularity, these clusters result in network communities.

Some communities, in particular, communities 2, 4, and 8, predominantly consist of type A, B, and AB thymoma. We performed Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment of protein sets from the ensemble biclusters (Fig. 1B). All selected communities showed significant repair mechanism pathways, which is well known for tumors to influence those pathways. Additionally, community 2, which includes samples of all thymoma subtypes, indicating a common signature on the proteomic level, showed two cancer-related terms.

We investigated the occurrence of proteins and samples in biclusters belonging to one community (Fig. 1C). A difference in the distributions of samples and features can be observed. The distribution of features is strongly positively skewed with a very low mode. In contrast, the sample distribution, for example, for communities 5 and 8, has a mode close to one. This shows that biclusters inside the same community (after filtering edges for random overlaps) can carry very different features and samples. By setting thresholds, ensemble biclusters can be restricted to point to consistent patterns in the data or to allow for variability.

In Fig. 1D, we visualize the affiliation of biclusters to algorithms, which predicted them. This reveals that biclustering algorithms tend to identify overlapping regions in the data, resulting in highly connected communities consisting only of one algorithm. This shows the necessity of taking the results of multiple biclustering algorithms into account and relying on not one but many different algorithms to capture patterns in the data beyond the specificities of a single algorithm. While we observe a good overlap of some network communities with the tumor subtypes, some individual (unconnected) biclusters also show high overlaps, for instance, with the type B thymoma. The strength of the MoSbi algorithm lies in the aggregation of biclusters. The visualization also helps to identify and analyze these individual biclusters, if they exhibit a high consensus with relevant information such as biological factors.

The results above demonstrate the power and utility of the workflow to establish a sophisticated biclustering analysis, to generate biological hypotheses.

Individual Biclustering Algorithms vs. MoSbi. Next, we compared the individual performances of available biclustering algorithms and contrasted them with the performance of MoSbi. For that, we selected six published and publicly available datasets from the metabolomics, transcriptomics, and proteomics disciplines (details in *SI Appendix, Table S1*). All datasets were analyzing cancer tissues or investigated cancer subtypes. As a gold standard, we used the condition match score to quantify the overlap between predicted biclusters and sample labels (see *Materials and Methods*), where the relevance describes how well predicted biclusters correspond to known labels, and recovery describes how well the labels were recovered by predictions. Additionally, Gene Ontology (GO) and KEGG pathway enrichment was performed to evaluate the gene sets in predicted biclusters.

The match between predictions and sample groups can be seen in Fig. 2A. It reveals a heterogeneous performance of the individual biclustering algorithms. Spectral only predicted biclusters in two out of the six scenarios. The iterative signature algorithm (Isa) has the highest recovery on the Tang et al. (36) metabolomics and Ku et al. (35) proteomics data and both transcriptomics datasets but has a poor performance on Yang et al. (37) metabolomics and Wiśniewski et al. (38) proteomics data. While having a good

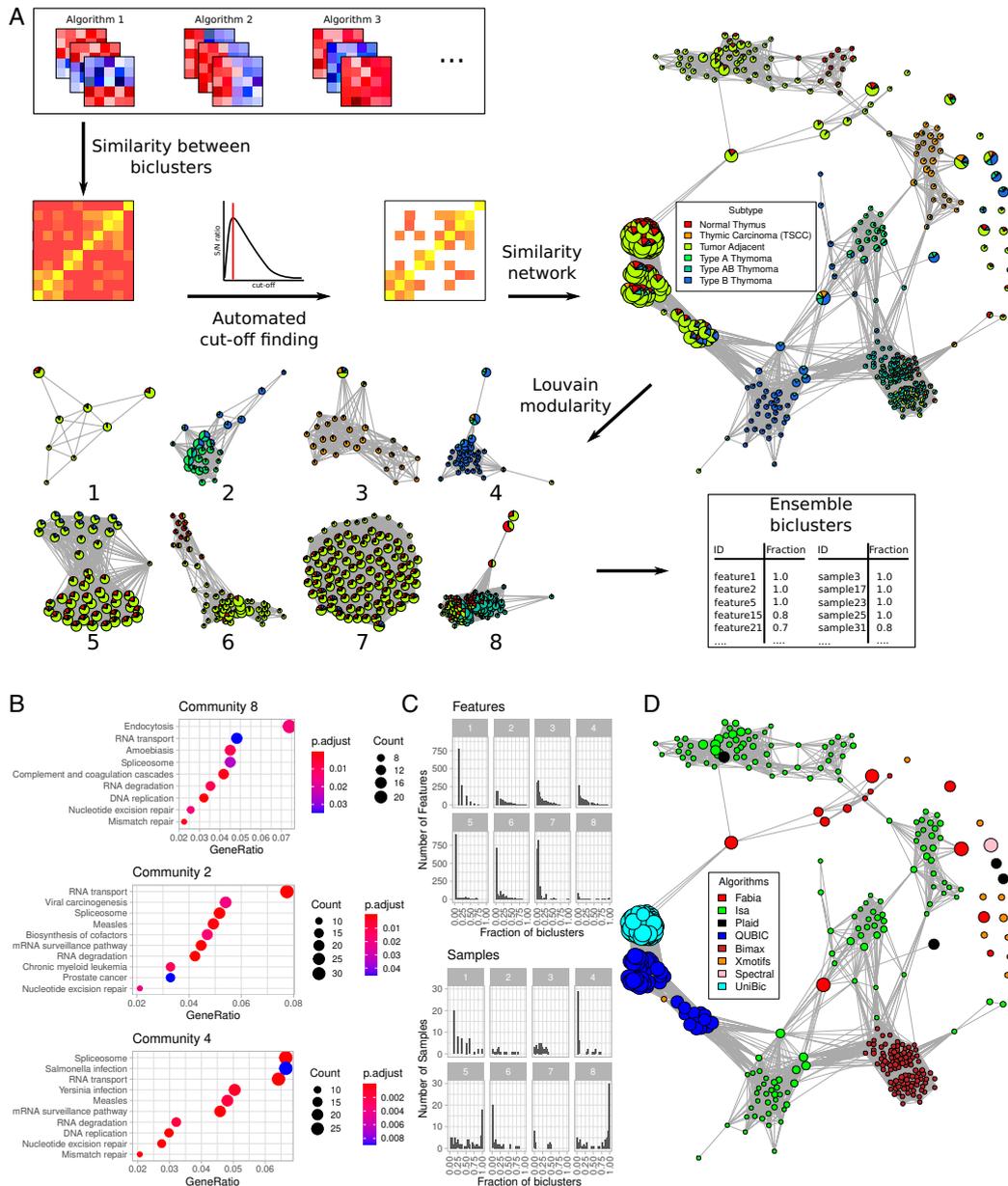


Fig. 1. Workflow of MoSbi with exemplary network visualizations. (A) Steps of the MoSbi approach. First, biclusters are predicted by multiple algorithms, and a similarity matrix is computed, which is then filtered for larger than random overlaps, using an error model. The matrix is then converted to a network that can be visualized with meta-information about samples or features. Louvain communities are then extracted and converted into ensemble biclusters. As an example, the bicluster network of proteomics data from Ku et al. (35) is shown. Nodes represent biclusters, with edges between them if their overlap exceeds the error threshold. (B) KEGG pathway enrichment for features of selected communities 2, 4, and 8. (C) Frequency of features (*Upper*) and samples (*Lower*) in biclusters that belong to one community. (D) Bicluster network of proteomics data from Ku et al. (35). Node colors represent algorithms, by which they were predicted.

Downloaded from https://www.pnas.org by University of Florida on April 18, 2022 from IP address 128.227.1.167.

Table 1. List of evaluated biclustering algorithms in alphabetical order

Algorithm	Publication
BicARE	Gestraud et al. (24)
Bimax	Prelić et al. (25)
CC	Cheng and Church (26)
Fabia	Hochreiter et al. (27)
Isa	Bergmann et al. (28)
Plaid	Lazzeroni and Owen (29)
QUBIC	Zhang et al. (30)
Quest	Murali and Kasif (31)
Spectral	Kluger et al. (32)
UniBic	Wang et al. (33)
Xmotifs	Murali and Kasif (31)

The results of algorithms can be imported and accessed with our MoSBI R package or executed using the webtool.

recovery, Isa never scores best on relevance. Similar behavior can be observed for Plaid, which, on average, performs very well for relevance, but shows low recoveries. It can also be observed that Plaid is the only algorithm that reached a relevance and recovery higher than 0.5, and achieved this in one proteomics dataset. We then applied our ensemble approach to the predictions of all algorithms per dataset (Fig. 2A, black marker). The ensemble approach is one of the two best performing tools in either recovery or relevance in all other datasets, except for the Tang et al. (36) metabolomics data, where we could observe high overlaps with other clinical confounders (SI Appendix, Fig. S2). On metabolomics data, with fewer features compared to sequencing data, the communities can additionally be visualized as cooccurrence networks (SI Appendix, Fig. S3). Over all six datasets, MoSBI performed second best, on average, by relevance and second best by recovery after Plaid and Isa, which both have poorer performances on the other scale.

To investigate the performance of the algorithms on the gene level, we performed KEGG pathway and GO biological process enrichment for the proteomics and transcriptomics datasets. In KEGG enrichment (Fig. 2B), The Biclustering Analysis and Results Exploration (BicARE) algorithm predicted the most biclusters with at least one significantly enriched term in three datasets. Interestingly, it did not stand out when investigating sample group labels. The ensemble method again showed a better performance than the average of biclustering algorithms. The same holds for the enrichment of biological processes with GO terms (Fig. 2C). Since all investigated proteomic and transcriptomic datasets were cancer related, we searched specifically for enriched KEGG pathways including the word “cancer,” “carcinoma,” or “tumor” (Fig. 2D). On the Wiśniewski et al. (38) proteomics data, only Isa and BicARE found significant terms for biclusters, but only at very low frequencies. In the Ku et al. (35) proteomics data, MoSBI found the most significant terms and, on the two transcriptomics datasets, the second most after Fabia (factor analysis for bicluster acquisition) and BicARE.

This reveals that individual biclustering algorithms peak in one or another measure or dataset, but in an unpredictable manner. However, the MoSBI ensemble approach is more consistent and therefore more reliable for biclustering analysis.

Performance on Synthetic Data. Evaluation on experimental data is preferable since it accurately resembles the real-life application of biclustering and stratification. Unfortunately, two-dimensional (2D) gold standards are usually not available, since many factors are influencing the molecular state of samples. Synthetic data can overcome this problem. This is frequently done to evaluate biclustering algorithms (6, 25, 39).

Based on the synthetic data generation of Prelić et al. (25), we developed a workflow to create synthetic scenarios, where one or multiple properties can be investigated (SI Appendix, *Synthetic Evaluation Scenarios*). We repeated previous scenarios from Prelić et al. (25) and added scenarios, covering sparsity, overlaps, and mixed sizes (SI Appendix, Table S2), and evaluated them on biclustering algorithms (*Materials and Methods* and SI Appendix, Figs. S6–S10). Since molecular omics data can include missing values, we investigated the effect of sparsity on the performance of biclustering algorithms (Fig. 3A). While the overall performance of all algorithms decreases with increased sparsity, Fabia and Isa showed a higher resilience until a sparsity of 20% (percentage of missing values in the matrix; SI Appendix), after which the results deteriorated. The relevance was more robust against sparsity and did not decrease as strongly as the recovery.

So far, synthetic evaluation has focused on the assessment of individual characteristics of the data (e.g., noise or size). Using our workflow and knowledge from previous synthetic scenarios, we defined a complex scenario, incorporating all previously mentioned manipulations to the data (Fig. 3B). We evaluated all approaches in this scenario and added a negative binomial background to simulate unique molecular identifier RNA sequencing (RNAseq) data (Fig. 3B, *Left*). Performance analysis separated the tools into two groups: clearly higher performing tools consisting of Fabia, Isa, and MoSBI, and the rest performing significantly inferiorly. Fabia shows the best recovery, and the ensemble approach shows the best relevance, but only marginally above Fabia and Isa. Even with the poor performance of many algorithms, MoSBI can still achieve high recovery and relevance. Algorithm selection has an influence on every ensemble approach; therefore, excluding the worst-performing algorithms from the ensemble approach yields a high increase of the relevance of the ensemble approach, while the recovery remains similar (SI Appendix, Fig. S11A).

Being an average, the relevance does not characterize every distribution correctly, but is widely used in biclustering evaluation studies. We investigated the relevance distribution of all algorithms independently (Fig. 3C) and combined (Fig. 3D). Some distributions are skewed. The combined distribution is positively skewed, showing that the majority of biclusters have a very low overlap with the gold standard. Predictions by the ensemble approach show a different distribution (Fig. 3E), where the majority of biclusters have a score above 0.5. Since an ensemble approach is sensitive to the performance of the underlying biclustering algorithms, we selected the best-performing algorithms and repeated the analysis (SI Appendix, Fig. S11 C and D). As can be seen, the performance of MoSBI is even more evident, showing the importance of the utilized algorithms. On the other hand, it shows that the approach can achieve a good performance, even with some poorly performing algorithms included. By combining highly overlapping biclusters, MoSBI can reduce the number of mismatched biclusters. This also shows that the relevance distribution can give more detailed insights into algorithm performance. MoSBI additionally reduces the number of biclusters drastically, making an investigation of all predictions more manageable. Analysis of the MoSBI parameters (SI Appendix, Fig. S12) showed that the row and column thresholds should be in the range of 0.02 and 0.2. The relevance increases with higher minimum community size thresholds, whereas the recovery decreases. An application-specific trade-off has to be decided by users. The number of randomizations for the similarity cutoff estimation does not affect the performance of MoSBI.

To investigate the performance of all algorithms under the best conditions, we optimized their parameters to achieve the best possible performance (SI Appendix, Fig. S13). This showed that algorithms can produce markedly better results given correct

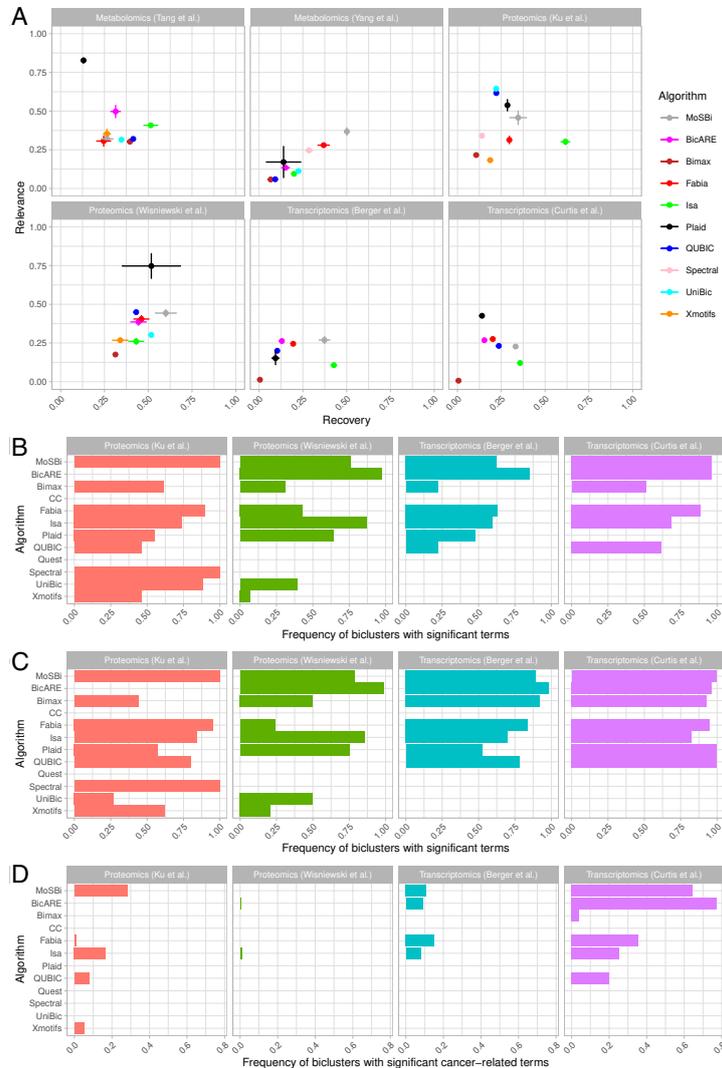


Fig. 2. Performance of MoSBI and individual biclustering algorithms on cancer-related omics data. Data was used from Ku et al. (35), Wisniewski et al. (38), Berger et al. (41), Curtis et al. (44), Tang et al. (36), and Yang et al. (37). For further information about the data, see *SI Appendix, Table S1*. (A) Recovery and relevance for the condition match score of biclustering tools based on samples for cancer (subtypes). (B) Frequency of predicted biclusters per algorithm, with one or more significant KEGG terms (adjusted *P* value cutoff < 0.05). (C) Frequency of biclusters with one or more significant GO terms from the “biological process” category. (D) Frequency of predicted biclusters per algorithm, with one or more cancer-related KEGG terms.

parameters compared to their standard parameters, in Fig. 3B. However, this is time consuming and only possible for data with an existing gold standard. The differences between the two complex synthetic scenarios showed that parameters and performances vary widely between datasets. Therefore, an ensemble method offers an easier method to achieve good performance independently of parameter optimization.

Biclustering in a Multiomics Context. Since biclustering requires a data matrix as input, it can naturally be applied to multiomics data, when merged into one data matrix. To investigate the performance of MoSBI in a multiomics context, we used the

TCGA breast cancer cohort from the Xena Platform (40), which provides omics data for multiple breast cancer subtypes. RNAseq, microRNA (miRNA), and protein data were run independently and combined for all biclustering algorithms. All resulting bicluster networks (Fig. 4A) appear similar, with big basal communities and multiple communities consisting mainly of the LumA or LumB subtype, often highly interconnected. The protein data network shows a less distinct basal community, whereas the miRNA data network shows Her2 samples mixed with LumB samples.

In the next step, we evaluated the performance of the biclustering algorithms on the different data types. A consistent perform-

A. Appendix

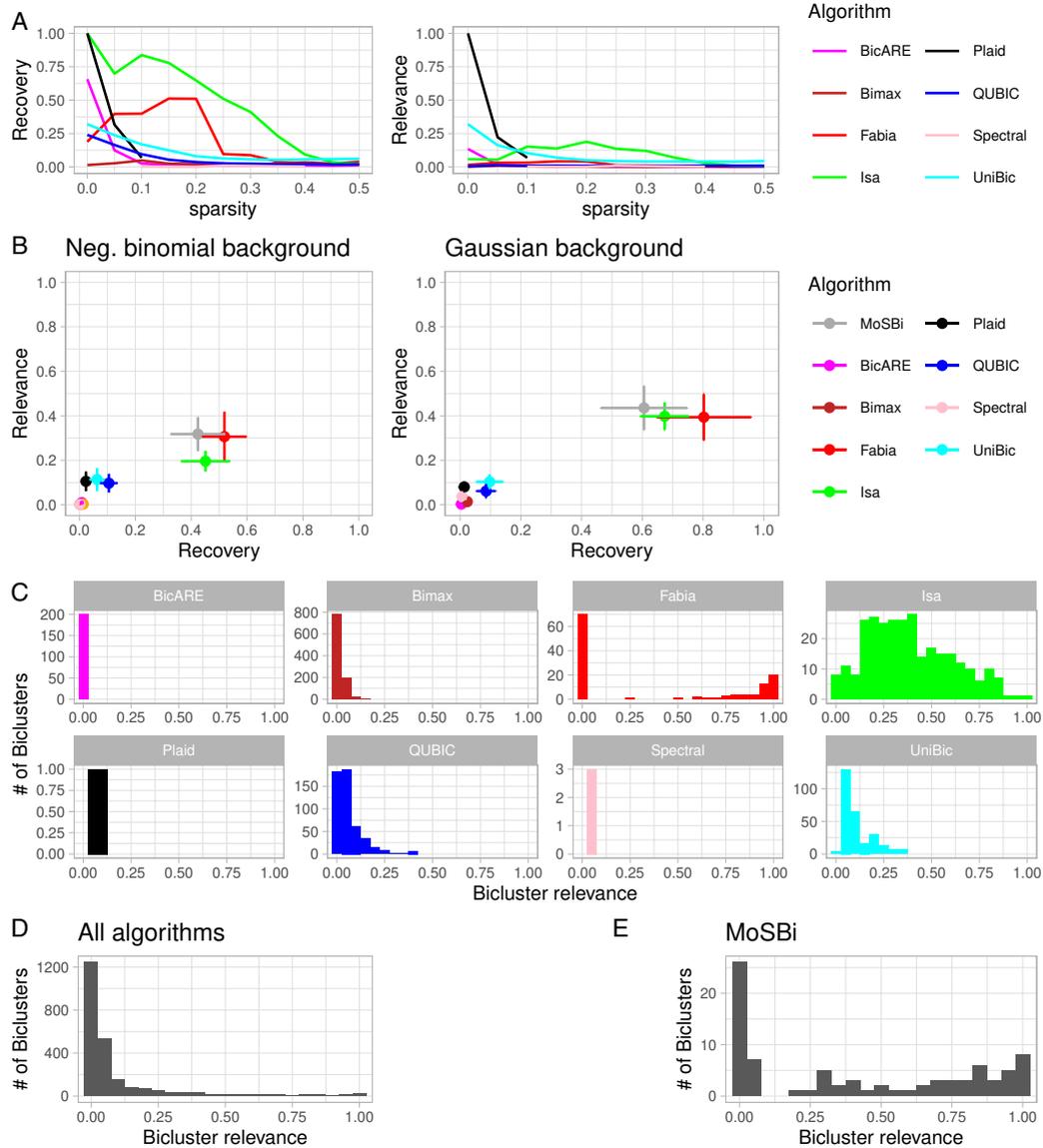


Fig. 3. Evaluation of biclustering algorithms on synthetic data. (A) Recovery and relevance of biclustering algorithms with increasing sparsity, for one hidden shift bicluster. (B) Performance of biclustering algorithms and ensemble approach on a synthetic scenario including different bicluster types, sizes, sparsity, and noise with a negative binomial distributed background (Left) and normally distributed background (Right). (C) Relevance distribution of biclustering algorithms for the scenario shown in B, Right. (D) Relevance distribution of all algorithms summed up from C. (E) Relevance distribution of the predictions of the ensemble approach using the biclusters from D.

mance of most algorithms can be observed (Fig. 4B), with only Plaid showing a high increase in relevance on the multiomics data compared to the other datasets, and not identifying any biclusters on the protein data. Only with MoSBI, a relevance and recovery higher than 0.25 could be observed in all four datasets. This shows that the multiomics data did not yield a big performance increase for most algorithms, but rather that all data types carry

the information to identify subtypes, with the ensemble approach being the most robust throughout all data types.

While we did not find big differences in the overall performance, we next looked at the recovery of the subtypes individually (Fig. 4C). Most algorithms did not recover all subtypes equally well. Isa has the highest recovery for basal (above 0.75 for RNAseq and multiomics) and worst for normal (all below 0.25). Fabia

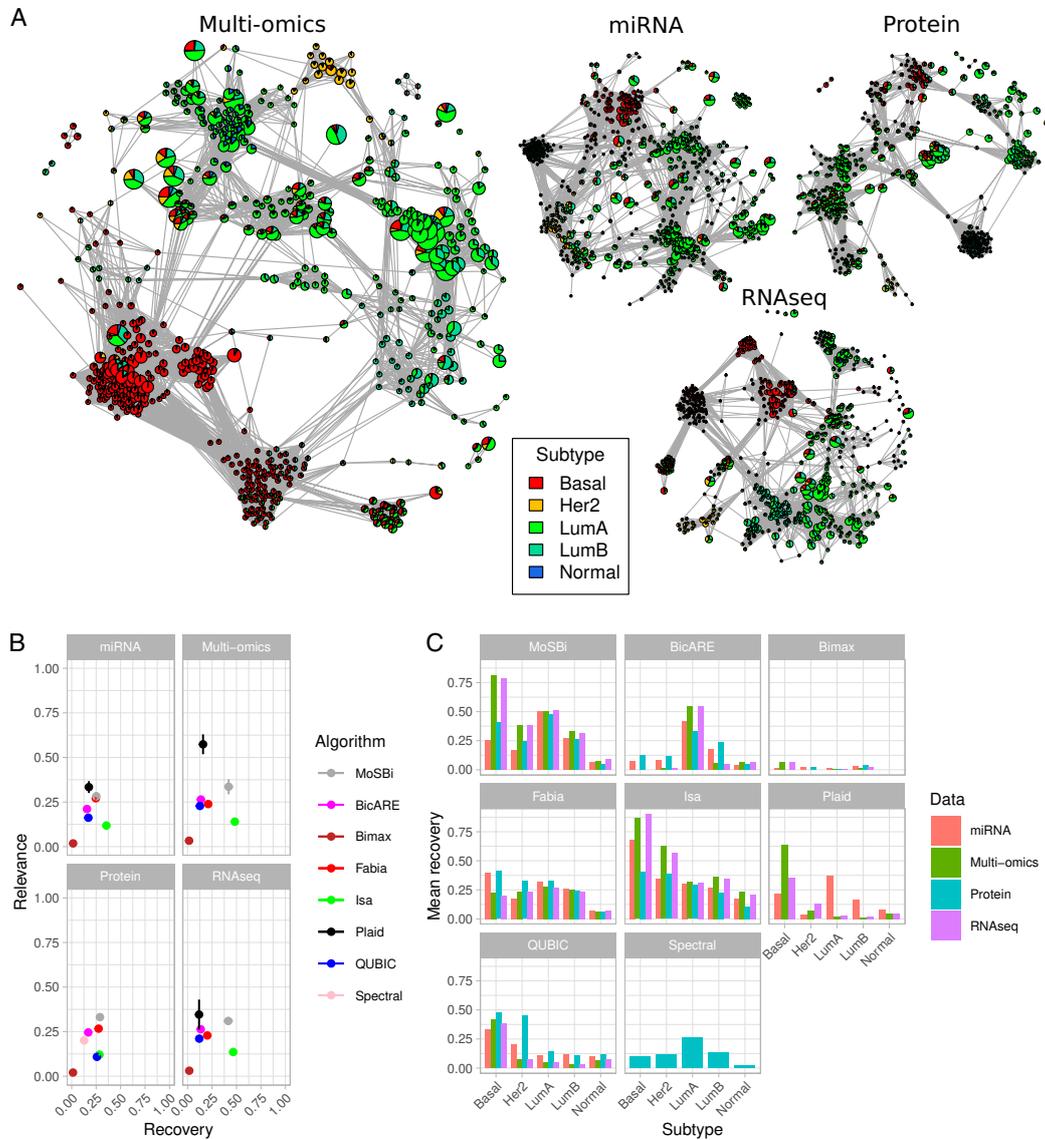


Fig. 4. Biclustering on breast cancer multiomics data. (A) Bicluster similarity networks on TCGA breast cancer miRNA, Protein expression, RNAseq, and combined data. Biclusters are colored by subtype, and node size is proportional to sample size. (B) Relevance and recovery for the condition match score on the datasets from A for each algorithm individually and combined with MoSbi. All algorithms were executed 10 times. (C) Recovery for subtypes on the datasets from A for each algorithm individually and combined with MoSbi. All algorithms were executed 10 times.

exhibits a more equal distribution, except for normal, which has a low recovery throughout all algorithms. It can again be observed that all data types are similarly able to identify subtypes. In MoSbi, the basal subtype has a better recovery in RNAseq and multiomics data. Another interesting observation is that BicARE consistently recovers the LumA subtype through all data types.

In this analysis, we can show that multiomics biclustering is possible and can add value to the results. However, an individual biclustering analysis on all data types is also possible and yields

similar performance. However, a combined analysis might be beneficial for a biological interpretation of biclusters, which consists of features from different omics types.

The MoSbi Software Suite. To make our ensemble approach and biclustering algorithms, in general, accessible for scientists and provide an easy-to-use interface, we developed the MoSbi suite for the identification of molecular signatures using biclustering. MoSbi is available as an R package on biocon-

Downloaded from https://www.pnas.org by University of Florida on April 18, 2022 from IP address 128.227.1.167.

ductor (<https://bioconductor.org/packages/mosbi/>) and web-app (<https://exbio.wzw.tum.de/mosbi>).

Many biclustering algorithms, such as Isa (41) and Fabia (27) or the biclust package, use different result formats for returning biclusters. Therefore, we developed a unified framework, which is able to import predictions from various biclustering algorithms to simplify the analysis of biclustering algorithms and apply our ensemble approach. Our network-based visualizations are also available in MoSBI, which can be used with our ensemble approach or single biclustering algorithms. The framework can be extended to offer support for new biclustering algorithms and integrate them into the workflow. Networks can be exported as graphML for compatibility with tools such as Cytoscape (42).

The web app allows users without programming knowledge to stratify samples with our ensemble approach and profit from visualizations. Additionally, all biclustering algorithms can be accessed and executed with all parameters independently, if users are interested in specific algorithms. We also provide a docker image of the web tool, which allows it to be deployed locally.

Discussion

Stratification of patients based on molecular omics data is a challenging task and requires modern computational tools. Unsupervised approaches are suited to identify novel subgroups in the data. Biclustering is able to find meaningful patterns in modern omics data. In contrast to traditional clustering, algorithms not only output sample subgroups but, additionally, feature subsets that characterize this similarity and can be further analyzed, for example, for functional associations to find disease mechanisms. We developed a biclustering ensemble approach, which takes the results of multiple biclustering algorithms and computes ensemble biclusters using a network-based approach. This is based on the assumption that biclustering algorithms predict highly overlapping biclusters, which we could validate in our work. Various biclusters pointing to the same underlying data structures can indicate robust biclusters, which are then identified with MoSBI. We showed this on thymic epithelial tumor data (35), where we were able to retrieve known cancer subtypes.

We demonstrated the application of MoSBI on cancer-related datasets and showed the possibility of performing a multiomics analysis using biclustering. On various synthetic and experimental datasets, we assessed the performance of different biclustering algorithms and compared them to our ensemble approach. While Fabia and Isa, on average, performed best of all considered biclustering algorithms, no algorithm performed best in all scenarios and can be universally recommended. MoSBI did not always stand out, but it achieved a robust good performance in most scenarios. While the optimization of algorithm parameters on synthetic data could significantly improve the results, it leads to extensive run times and requires gold standard annotations, which are usually not available in real data, indicating that MoSBI is a preferable choice for biclustering. Additionally, it markedly reduces the number of biclusters. The network visualization gives an overview of the results and, compared to other methods (18), scales well with an increased number of biclusters.

The advantage of our ensemble approach over other biclustering ensemble approaches is that it is not algorithm specific and, via the MoSBI suite, is accessible as an application programming interface (API) and graphical user interface. Unfortunately, some proposed approaches lack implementation (10, 13). An ensemble method based on the calculation of similarities between biclusters was proposed by Hanczar and Nadif (12), where the

authors calculated overlaps based on sums of overlaps of rows and columns, which can result in nonzero similarities for biclusters that share rows but no columns and are, in fact, not overlapping (SI Appendix, Fig. S14). They proposed the method as a single-algorithm ensemble approach that applied hierarchical clustering on the similarity matrix. This introduces another parameter for the number of consensus biclusters and assigns each bicluster to an ensemble bicluster, even with low overlap. Our approach avoids this by using the Louvain modularity to find the optimal split of the network into communities. We also introduce an error model for ensemble biclustering that removes random, and therefore misleading, overlaps from the similarity network. Additionally, MoSBI makes further analysis easier, since it reduces the number of predictions while maintaining similar performance. With MoSBI, we provide a tool to make the application of multialgorithm ensemble biclustering with scalable visualizations applicable for all kinds of noninformatics users possible. However, as an ensemble approach, MoSBI relies on the performance of multiple biclustering algorithms. We showed how the selection of biclustering algorithms can influence the results of MoSBI (SI Appendix, Fig. S11 C and D). While MoSBI is robust against a few badly performing algorithms, the majority of algorithms need to identify reasonable biclusters for MoSBI in order to work correctly. With new developments and available algorithms, MoSBI can be extended to improve performance in the future.

Similar to other unsupervised methods such as clustering, biclustering is often only the first step in data analysis. This comes with the challenge to inspect and interpret the results before further deciding about follow-up analysis steps. A particular challenge can be the difference in sizes of (ensemble) biclusters. It is important to consider the number of samples included for a molecular signature that corresponds to a phenotype, to evaluate its robustness. A direct comparison of biclusters with big differences in size should therefore be handled with care. The MoSBI framework allows for simple visualization of this but still requires manual supervision.

Our methodology offers an advanced perspective on biclustering and can visualize detailed properties of predictions. We demonstrated how a bicluster network analysis provides additional biological and structural insights into data. Clinical or experimental conditions can be associated with biological features. Using our approach, biclustering has the potential to play a significant role in disease subtyping and understanding.

Materials and Methods

The biclustering ensemble algorithm consists of four major steps. These are the execution of multiple biclustering algorithms, followed by a similarity computation for all returned biclusters, filtering of the similarity matrix for random overlaps, and community detection on the similarity network. In the following, all steps are described in detail.

Algorithms. Given an input matrix $M \in \mathbb{R}^{R \times C}$, we utilize different biclustering algorithms (Table 1) and collect their results in one combined list of biclusters $B = [B_1, B_2, \dots, B_n]$, where $B_i = (B_i^r, B_i^c)$ and $B_i^r \subseteq [1, \dots, R]$, $B_i^c \subseteq [1, \dots, C]$ is a set of row and column indices of the matrix M that belongs to a bicluster B_i . We implemented interfaces for all algorithms in our R package to generate this list using one unified API.

Similarity Metrics. In the next step, pairwise similarities between all biclusters in B are computed. This is done using common similarity metrics, where the similarity is expressed as a 2D overlap between biclusters. To do so, we treated a bicluster matrix as a 2D area and computed their similarity in terms of overlapping areas. This is different than the additive similarity as proposed by Hanczar

and Nadif (12). One implemented metric is the Jaccard index. Our adaption resulted in the following formula:

$$J(B_1, B_2) = \frac{|B_1 \cap B_2|}{|B_1 \cup B_2|} = \frac{|B_1' \cap B_2'| \times |B_1' \cap B_2'|}{(|B_1'| \times |B_2'|) + (|B_2'| \times |B_2'|) - (|B_1' \cap B_2'| \times |B_1' \cap B_2'|)}$$

Besides the widely used Jaccard index, also the Bray-Curtis similarity, overlap coefficient, and Fowlkes-Mallows index were implemented in a similar 2D fashion. This results in a similarity matrix S with $S_{ij} = J(B_i, B_j)$. Note that MoSbi can use any other definition of similarity as well. Biclusters fully contained in other ones are evaluated with the same metric. Hence, they exhibit a similarity based on their overlap as described above.

Error Model. Since biclusters can have random overlaps that do not represent meaningful interactions, we estimate a cutoff to filter for such overlaps in the similarity matrix. This is done by randomly generating a list of biclusters B' such that $B' = [B_1', B_2', \dots, B_n']$, $|B'| = |B|$, and $B_i' = (B_i^r, B_i^c)$, where B_i^r and B_i^c are randomly drawn without replacement from $[1, \dots, R]$ and $[1, \dots, C]$ correspondingly such that $|B_i^r| = |B_i^c|$ and $|B_i'| = |B_i|$. To estimate the best cutoff c^* for the values in the similarity matrix S , we treat the S as an adjacency matrix and optimize c^* for the biggest ratio between remaining edges in S and S' , where $S'_{ij} = J(B_i', B_j')$ (to increase robustness, multiple randomizations K of B are used),

$$c^* = \operatorname{argmax}_c \frac{\sum_{ij} \Theta_c(S_{ij})}{\sum_k (\sum_{ij} \Theta_c(S_{ij}^k)) / K}$$

with

$$\Theta_c: \mathbb{R} \rightarrow \{0, 1\} \\ x \mapsto \begin{cases} 0: & x < c \\ 1: & x \geq c \end{cases}$$

This results in the final and filtered similarity matrix S^{c^*} where

$$S_{ij}^{c^*} \mapsto \begin{cases} 0: & S_{ij} < c^* \\ S_{ij}: & S_{ij} \geq c^* \end{cases}$$

Community Detection. Finally, S^{c^*} is used as an adjacency matrix with biclusters as nodes, and edges representing similarities. We compute the weighted Louvain modularity (43), with similarities as weights, to find bicluster communities in the network. These highly similar bicluster communities can then be converted into ensemble biclusters using three parameters: `min_size` (default = 2) which defines the minimum number of biclusters in a community to convert a community into a bicluster, where smaller communities are not considered; and `row_threshold` and `col_threshold` (default = 0.1), the minimum frequency of occurrence of a row/column element in a bicluster community to be taken over into an ensemble bicluster: For example, with values of 0.5, only genes and samples will be part of the new ensemble bicluster if they occur in at least 50% of all biclusters in the corresponding community.

Implementation. MoSbi is free software. The workflow was implemented in the R programming language (version ≥ 3.6) and C++17. The web interface was realized with the Shiny web framework for R (version 1.4.0.2). The workflow can be executed from our web app on our servers or on a local machine using a public Docker image. For higher throughput or for the integration of our approach into a bioinformatics pipeline, the R package can be used directly.

Visualizations. Network visualizations of the MoSbi package are implemented in R using the "igraph" package. Interactive plots in the web tool use the "visNetwork" library. All other visualizations use the "ggplot2" library in R.

Cooccurrence Networks. For cooccurrence networks, biclusters from one community were selected. From this, a new network is computed with samples and features as nodes. Edges can occur between samples and samples, samples and features, and features and features. An edge is drawn between two nodes if they occur together in at least one bicluster of the community. Edges are

weighted by the number of biclusters, where two nodes cooccur. For the visualization, a network layout is computed, which takes the edge weights into account.

Match Score. The performance of biclustering algorithms and MoSbi was evaluated by comparing their overlap to labeled gold standard data. We used the commonly applied gene match score,

$$MS_G(M_1, M_2) = \frac{1}{|M_1|} \sum_{(G_1, C_1) \in M_1} \max_{(G_2, C_2) \in M_2} \frac{|G_1 \cap G_2|}{|G_1 \cup G_2|}$$

where M_1 and M_2 are two sets of biclusters, with each bicluster consisting of a set of genes G_i and conditions C_i (rows and columns) (25). To investigate sample/condition overlaps, we define the according condition match score,

$$MS_C(M_1, M_2) = \frac{1}{|M_1|} \sum_{(G_1, C_1) \in M_1} \max_{(G_2, C_2) \in M_2} \frac{|C_1 \cap C_2|}{|C_1 \cup C_2|}$$

On synthetic data, where a 2D gold standard is available, we define the 2D match score as the multiplicative score of both dimensions,

$$MS_{2D}(M_1, M_2) = \frac{1}{|M_1|} \sum_{(G_1, C_1) \in M_1} \max_{(G_2, C_2) \in M_2} \frac{|C_1 \cap C_2|}{|C_1 \cup C_2|} \times \frac{|G_1 \cap G_2|}{|G_1 \cup G_2|}$$

The scores can be used to compute relevance and recovery. Let M_{opt} be a set of implanted biclusters or a gold standard, and let M be the output of a biclustering algorithm. Then, the average bicluster relevance is defined as $MS(M, M_{opt})$ and describes to what extent the biclusters found by the algorithm correspond to the true hidden biclusters in the gene, condition, or both dimensions. Similarly, the average bicluster recovery is defined as $MS(M_{opt}, M)$ and describes how well each of the true biclusters is recovered by the algorithm. The recovery and relevance score both have an optimal value of one, indicating a perfect overlap, and zero, indicating no overlap.

The match scores describe a normalized sum of values. To investigate how well all individual biclusters predicted by one algorithm match the gold standard, we investigated the relevance distribution $RD = [rd_1, rd_2, \dots, rd_n]$ with n as the number of biclusters in set of biclusters M and

$$rd_i = \max_{(G_{opt}, C_{opt}) \in M_{opt}} \frac{|C_i \cap C_{opt}|}{|C_i \cup C_{opt}|} \times \frac{|G_i \cap G_{opt}|}{|G_i \cup G_{opt}|}$$

where C_i and G_i are the columns and rows of bicluster M_i .

Experimental Omics Data. We evaluated the biclustering algorithms and MoSbi on six publicly available metabolomics (36, 37), proteomics (35, 38), and transcriptomics (41, 44) datasets (SI Appendix, Table S1). Feature-wise z scores were computed for all datasets, and, prior to that, log2 transformed [except for Ku et al. (35) and Curtis et al. (44), which already showed a normal distribution]. Transcriptomics data were filtered for genes with 80% coverage in all samples and filtered the 5,000 most variant genes, to reduce algorithm runtime. Gene set/pathway enrichment was performed using the "clusterProfiler" R package using the "enrichGO" (biological process enrichment) and "enrichKEGG" functions.

TCGA breast cancer data were downloaded from the Xena Platform (40, 45). RNAseq transcriptomics data were processed as described above, and miRNA and protein data were filtered for 80% coverage in all samples and z-score transformed. Only samples occurring in all three datasets were considered for the individual and multiomics analysis, which resulted in 484 samples with measurements for all three data types.

Synthetic Data Generation. To investigate the performance of tools in a controlled environment with a fully known gold standard, we developed a pipeline to generate synthetic datasets with implanted biclusters and additional properties such as noise and sparsity. The pipeline is shown in SI Appendix, Fig. S1. A detailed description of all synthetic scenarios is available in SI Appendix.

Data Availability. The source code is available for the R package (<https://github.com/tDrose/mosbi>) and for the web application (<https://gitlab.lrz.de/lipitum-projects/mosbi-webapp>). Both are published under the aGPLv3 license. The code and all used data for the evaluation that was performed for this work is available on figshare: <https://doi.org/10.6084/m9.figshare.19096070.v1> (46).

Previously published data were used for this work (35–38, 40, 41, 44).

All other study data are included in the article and/or *SI Appendix*.

ACKNOWLEDGMENTS. T.D.R., N.K., and J.K.P. are funded by the Bavarian State Ministry of Science and the Arts in the framework of the Bavarian Research

Institute for Digital Transformation (Grant LipiTUM). J.B. was partially funded by his VILLUM Young Investigator Grant 13154. The work by J.B. was also supported by the German Federal Ministry of Education and Research within the framework of the e:Med research and funding concept (Grant 01ZX1910D).

Author affiliations: ^aLipiTUM, TUM School of Life Sciences, Technical University of Munich (TUM), 65354 Freising, Germany; ^bDepartment for Mathematics and Computer Science, University of Southern Denmark, 5230 Odense, Denmark; and ^cInstitute for Computational Systems Biology, University of Hamburg, 22607 Hamburg, Germany

- M. R. Trusheim, E. R. Berndt, F. L. Douglas, Stratified medicine: Strategic and economic implications of combining drugs and clinical biomarkers. *Nat. Rev. Drug Discov.* **6**, 287–293 (2007).
- S. Khakabimamaghani, M. Ester, "Bayesian biclustering for patient stratification" in *Pacific Symposium on Biocomputing 2016*, R. A. Altman et al., Eds. (World Scientific, 2016), pp. 345–356.
- O. Lazareva et al., BiCoN: Network-constrained biclustering of patients and omics data. *Bioinformatics* **37**, 2398–2404 (2020).
- S. C. Madeira, A. L. Oliveira, Biclustering algorithms for biological data analysis: A survey. *IEEE ACM Trans. Comput. Biol. Bioinformatics* **1**, 24–45 (2004).
- C. Winiew, J. Baumbach, R. Röttger, Comparing the performance of biomedical clustering methods. *Nat. Methods* **12**, 1033–1038 (2015).
- V. A. Padilha, J. G. Ricardo, A systematic comparative evaluation of biclustering techniques. *BMC Bioinformatics* **18**, 55 (2017).
- B. Pontes, R. Giraldez, J. S. Aguilár-Ruiz, Biclustering on expression data: A review. *J. Biomed. Inform.* **57**, 163–180 (2015).
- U. Neumann et al., Compensation of feature selection biases accompanied with improved predictive performance for binary classification by using a novel ensemble feature selection approach. *BioData Min.* **9**, 36 (2016).
- S. Vega-Pons, J. Ruiz-Shulcloper, A survey of clustering ensemble algorithms. *Int. J. Pattern Recognit. Artif. Intell.* **25**, 337–372 (2011).
- B. Hanczar, M. Nadif, Ensemble methods for biclustering tasks. *Pattern Recognit.* **45**, 3938–3949 (2012).
- G. Aggarwal, N. Gupta, "BiEtopt: Biclustering ensemble using optimization techniques" in *IEEE International Conference on Data Mining*, H. Xiong, G. Karypis, B. M. Thuraisingham, D. J. Cook, X. Wu, Eds. (Institute of Electrical and Electronics Engineers, 2013), pp. 181–192.
- B. Hanczar, M. Nadif, Using the bagging approach for biclustering of gene expression data. *Neurocomputing* **74**, 1595–1605 (2011).
- B. Hanczar, M. Nadif, "Unsupervised consensus functions applied to ensemble biclustering" in *Proceedings of the 3rd International Conference on Pattern Recognition Applications and Methods*, M. De Marsico, A. Tabbone, A. Fred, Eds. (SciTePress, 2014), pp. 30–39.
- G. Aggarwal, N. Gupta, "BEM: bicluster ensemble using mutual information" in *2013 12th International Conference on Machine Learning and Applications*, M. Arif Wani et al., Eds. (IEEE Computer Society, 2013), pp. 321–324.
- L. Yin, Y. Liu, Ensemble biclustering gene expression data based on the spectral clustering. *Neural Comput. Appl. Ensemble biclustering gene expression data based on the spectral clustering*, 2403–2416 (2018).
- A. Kasim, *Applied Biclustering Methods for Big and High-Dimensional Data using R* (Chapman and Hall, 2016).
- J. Heinrich, R. Seifert, M. Burch, D. Weiskopf, *BiCluster Viewer: A Visualization Tool for Analyzing Gene Expression Data* (Springer, 2011).
- H. Aouabed, R. Santamaría, M. Eloumi, VisBicCluster: A Matrix-Based bicluster visualization of expression data. *J. Comput. Biol.* **27**, 1384–1396 (2020).
- G. A. Grothaus, A. Mufti, T. M. Murali, Automatic layout and visualization of biclusters. *Algorithms Mol. Biol.* **1**, 15 (2006).
- R. Santamaría, R. Theron, L. Quintales, BicOverlapper 2.0: Visual analysis for gene expression. *Bioinformatics* **30**, 1785–1786 (2014).
- S. Barkow, S. Bleuler, A. Prelic, P. Zimmermann, E. Zitzler, BiCAT: A biclustering analysis toolbox. *Bioinformatics* **22**, 1282–1283 (2006).
- M. Streit, et al., Fuzzy force-directed bicluster visualization. *BMC Bioinformatics* **15**, S4 (2014).
- R. Santamaría, R. Theron, L. Quintales, BicOverlapper: A tool for bicluster visualization. *Bioinformatics* **24**, 1212–1213 (2008).
- P. Gestraud, I. Brito, E. Barillot, *BicARE: Biclustering Analysis and Results Exploration*, R package version 1.52.0, <https://doi.org/doi:10.18129/B9.bioc.BicARE> (2020).
- A. Prelic et al., A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics* **22**, 1122–1129 (2006).
- Y. Cheng, G. M. Church, Biclustering of expression data. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **8**, 93–103 (2000).
- S. Hochreiter, et al., FABIA: Factor analysis for bicluster acquisition. *Bioinformatics* **26**, 1520–1527 (2010).
- S. Bergmann, J. Ihmels, N. Barkai, Iterative signature algorithm for the analysis of large-scale gene expression data. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* **67**, 031902 (2003).
- L. Lazerzeri, A. Owen, Plaid models for gene expression data. *Stat. Sin.* **12**, 61–86 (2002).
- Y. Zhang, et al., QUBIC: A bioconductor package for qualitative biclustering analysis of gene co-expression data. *Bioinformatics* **33**, 450–452 (2016).
- T. M. Murali, S. Kasif, Extracting conserved gene expression motifs from gene expression data. *Pac. Symp. Biocomput.* **2003**, 77–88 (2002).
- Y. Kluger, R. Basri, J. T. Chang, M. Gerstein, Spectral biclustering of microarray data: Co-clustering genes and conditions. *Genome Res.* **13**, 703–716 (2003).
- Z. Wang, G. Li, R. W. Robinson, X. Huang, UniBic: Sequential row-based biclustering algorithm for analysis of gene expression data. *Sci. Rep.* **6**, 23466 (2016).
- O. Vvedenskaya et al., Nonalcoholic fatty liver disease stratification by liver lipidomics. *J. Lipid Res.* **62**, 100104 (2021).
- X. Ku et al., Deciphering tissue-based proteome signatures revealed novel subtyping and prognostic markers for thymic epithelial tumors. *Mol. Oncol.* **14**, 721–741 (2020).
- X. Tang et al., A joint analysis of metabolomics and genetics of breast cancer. *Breast Cancer Res.* **16**, 415 (2014).
- Y. Yang et al., Integrated microbiome and metabolome analysis reveals a novel interplay between commensal bacteria and metabolites in colorectal cancer. *Theranostics* **9**, 4101–4114 (2019).
- J. R. Wiśniewski et al., Absolute proteome analysis of colorectal mucosa, adenoma, and cancer reveals drastic changes in fatty acid metabolism and plasma membrane transporters. *J. Proteome Res.* **14**, 4005–4018 (2015).
- K. Eren, M. Deveci, O. Küçükçunç, Ü. V. Çatalyürek, A comparative analysis of biclustering algorithms for gene expression data. *Brief Bioinform.* **14**, 279–292 (2013).
- M. J. Goldman et al., Visualizing and interpreting cancer genomics data via the Xena platform. *Nat. Biotechnol.* **38**, 675–678 (2020).
- A. C. Berger et al., Cancer Genome Atlas Research Network, A comprehensive pan-cancer molecular study of gynecologic and breast cancers. *Cancer Cell* **33**, 690–705.e9 (2018).
- P. Shannon et al., Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).
- V. D. Blondel, J. L. Guillaume, R. Lambiotte, E. Lefebvre, Fast unfolding of communities in large networks. *J. Stat. Mechanics Theory Exper.* **2008**, P10008 (2008).
- C. Curtis et al.; METABRIC Group, The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486**, 346–352 (2012).
- M. J. Goldman, TCGA Breast Cancer (BRCA). Xena Browser. [https://xenabrowser.net/datapages/?cohort=TCGA%20Breast%20Cancer%20\(BRCA\)&removeHub=https%3A%2F%2Fxcena.treehouse.gi.ucsc.edu%3A443](https://xenabrowser.net/datapages/?cohort=TCGA%20Breast%20Cancer%20(BRCA)&removeHub=https%3A%2F%2Fxcena.treehouse.gi.ucsc.edu%3A443). Accessed 14 December 2020.
- T. D. Rose et al., MoSbi - Data & scripts for biclustering algorithm evaluation. Figshare. <https://doi.org/10.6084/m9.figshare.19096070.v1>. Deposited 31 January 2022.

A.2. Nonalcoholic fatty liver disease stratification by liver lipidomics



Nonalcoholic fatty liver disease stratification by liver lipidomics

Olga Vvedenskaya^{1,*}, Tim Daniel Rose^{2,3}, Oskar Knittelfelder¹, Alessandra Palladini^{3,4}, Judith Andrea Heidrun Wodke⁵, Kai Schuhmann¹, Jacobo Miranda Ackerman¹, Yuting Wang¹, Canan Has¹, Mario Brosch^{6,7}, Veera Raghavan Thangapandi^{6,7}, Stephan Buch^{6,7}, Thomas Züllig⁸, Jürgen Hartler^{9,10}, Harald C. Köfeler⁸, Christoph Röcken¹¹, Ünal Coskun^{3,4,12}, Edda Klipp⁵, Witigo von Schoenfels^{13,14}, Justus Gross¹⁵, Clemens Schafmayer¹⁵, Jochen Hampe⁶, Josch Konstantin Pauling^{2,*}, and Andrej Shevchenko^{1,3*}

¹Max Planck Institute of Molecular Cell Biology and Genetics, Dresden, Germany; ²LipiTUM, Chair of Experimental Bioinformatics, TUM School of Life Sciences, Technical University of Munich, Munich, Germany; ³Paul Langerhans Institute Dresden of the Helmholtz Zentrum Munich at the University Hospital Carl Gustav Carus, Technische Universität (TU) Dresden, Dresden, Germany; ⁴German Center for Diabetes Research (DZD e.V.), Neuherberg, Germany; ⁵Theoretical Biophysics, Humboldt-Universität zu Berlin, Berlin, Germany; ⁶Department of Medicine I, University Hospital Dresden and ⁷Center for Regenerative Therapies Dresden (CRTD), Technische Universität (TU) Dresden, Dresden, Germany; ⁸Core Facility Mass Spectrometry, Medical University of Graz, Graz, Austria; ⁹Institute of Pharmaceutical Sciences, and ¹⁰Field of Excellence BioHealth, University of Graz, Graz, Austria; ¹¹Department of Pathology, University Hospital Schleswig-Holstein, Kiel, Schleswig-Holstein, Germany; ¹²Department of Membrane Biochemistry and Lipid Research, University Hospital Carl Gustav Carus of Technische Universität Dresden, Dresden, Germany; ¹³Department of Visceral and Thoracic Surgery, University Hospital Schleswig-Holstein, Kiel Campus, Christian-Albrechts-University Kiel, Kiel, Germany; ¹⁴Christian Albrechts University in Kiel Center of Clinical Anatomy Kiel, Schleswig-Holstein, Germany; and ¹⁵Department of General, Visceral, Vascular and Transplant Surgery, Rostock University Medical Center, Rostock, Germany

Abstract Nonalcoholic fatty liver disease (NAFLD) is a common metabolic dysfunction leading to hepatic steatosis. However, NAFLD's global impact on the liver lipidome is poorly understood. Using high-resolution shotgun mass spectrometry, we quantified the molar abundance of 316 species from 22 major lipid classes in liver biopsies of 365 patients, including nonsteatotic patients with normal or excessive weight, patients diagnosed with NAFL (nonalcoholic fatty liver) or NASH (nonalcoholic steatohepatitis), and patients bearing common mutations of NAFLD-related protein factors. We confirmed the progressive accumulation of di- and triacylglycerols and cholesteryl esters in the liver of NAFL and NASH patients, while the bulk composition of glycerophospho- and sphingolipids remained unchanged. Further stratification by biclustering analysis identified sphingomyelin species comprising n24:2 fatty acid moieties as membrane lipid markers of NAFLD. Normalized relative abundance of sphingomyelins SM 43:3;2 and SM 43:1;2 containing n24:2 and n24:0 fatty acid moieties, respectively, showed opposite trends during NAFLD progression and distinguished NAFL and NASH lipidomes from the lipidome of nonsteatotic livers. Together with several glycerophospholipids containing a C22:6 fatty

acid moiety, these lipids serve as markers of early and advanced stages of NAFL.

Supplementary key words NAFLD • shotgun lipidomics • lipid biomarkers • sphingomyelins • liver biopsies • NAFL; • NASH • biclustering analysis • steatosis • liver lipidome

Nonalcoholic fatty liver disease (NAFLD) is a metabolic dysfunction histologically characterized by hepatic fat accumulation (hepatic steatosis) in the absence of heavy alcohol consumption in past medical history (1). NAFLD affects up to 30% of adults and up to 80% of obese and diabetic individuals worldwide (2). The prevalence and severity of NAFLD are higher in men although in postmenopausal women, the NAFLD rate increases (3). NAFLD is subdivided into nonalcoholic fatty liver (NAFL), nonalcoholic steatohepatitis (NASH), cirrhosis, and hepatocellular carcinoma (4, 5). In contrast to NASH, hepatic steatosis in NAFL may occur with no significant inflammation.

The molecular background and pathophysiology of NAFL and why and how it progresses to NASH are poorly understood (6–9). While the intracellular accumulation of triacylglycerols (TG), diacylglycerols (DG) and free fatty acids (FFA) is a metabolic hallmark of NAFLD, it is unclear whether the disease also alters a broader scope of lipids (10–16). Mutations in NAFLD risk factors, e.g., *PNPLA3* or *MBOAT7*, affect the liver

[‡]These authors contributed equally to this work.

*For correspondence: Josch Konstantin Pauling, josch.pauling@wzw.tum.de; Andrej Shevchenko, shevchenko@mpi-cbg.de.



lipidome (17–31). It is also conceivable that lipidome remodeling may contribute to or be associated with the onset and propagation of NAFLD (15). However, lipidomic evidence is mostly semiquantitative and based on limited lipid class coverage (32). Previous studies were often focused on the composition of energy storage lipids (12, 13, 33), while membrane and signaling lipids received less attention (28, 34, 35). Chiappini *et al.* (36) used TOF-SIMS imaging to analyze 104 lipid species in biopsies of 61 NAFLD patients and suggested a NASH lipidomic signature comprising 32 lipid species, although the analysis was biased to most abundant and best ionized classes, e.g., phosphatidylcholines (PC) and TG. The study by Gorden *et al.* (13) covered 186 lipids in liver biopsies and plasma of 91 patients and revealed a combination of plasma biomarkers (including polyunsaturated glycerol- and glycerophospholipids together with long-chain ceramides (Cer)) that distinguished NASH from NAFL. However, lipid markers and their fold changes between disease states reported by independent studies were not concordant. Also, if NAFLD globally alters the liver lipidome and if changes in the abundance of glycerophospho- and sphingolipids (for convenience, here we termed them as membrane lipids) corroborate major clinical indices and correlate with disease severity is an open question, particularly because no reference values for individual liver lipids and their physiological variation are available.

To better understand how NAFLD is transforming the human liver lipidome, we assembled a cohort of 365 histologically characterized biopsies reflecting its progression from nonsteatotic obesity to overt NASH, together with appropriate nonsteatotic and nonobese controls. We then used shotgun mass spectrometry (37, 38) to systematically quantify the molar abundance of 316 species from 22 major lipid classes that encompassed membrane and energy storage lipids including cholesterol. Biclustering analysis of the curated lipidomics dataset recognized several signatures comprising specific membrane lipids that enabled patient's stratification at different stages of NAFLD independently of progressive accumulation of DG and TG.

MATERIALS AND METHODS

Cohort recruitment, study design, and ethic approval

The study protocol accords the ethical guidelines of the 1975 Declaration of Helsinki and was approved by the authority of Universität Kiel (D425/07, A111/99) before the study commenced. All patients had given their written informed consent. In total, 365 individuals (124 males; 241 females) from 17 to 85 years of age and whose BMI was in the range of 14.8–83.6 (kg/m²) were recruited within the time period of 2007–2016. NASH and NAFL were defined by the NAFLD activity score (NAS) as described (39). Phenotyping of the entire cohort was performed using standardized histology protocol (17) in a blinded fashion by a board-certified surgical

pathologist (C. R.) having the specialization in hepatopathology (details are in [supplemental data](#), Histology).

Alcohol consumption was assessed by self-reporting; subjects with average alcohol consumption of more than 30 g/day in men or 20 g/day in women (an equivalent of three and two standard alcoholic drinks per day, respectively) were not enrolled (40). The collected metadata included age, sex, BMI, blood test, including gamma-glutamyl transferase (GGT), medication taken by each patient, and mutation status of the following genes: *PNPLA3*, *TM6SF2*, *MBOAT7*, *HSD17B13*, *SERPINA1* (*SERPINA 5* and *SERPINA 2*). Evidence for insulin resistance was not available.

Common chemicals and lipid standards

Synthetic lipid standards (see [supplemental data](#), Common chemicals and lipid standards for complete list) were purchased from Avanti Polar Lipids (Alabaster, AL, USA). Individual standards were mixed and diluted with methyl-*tert*-butyl ether (MTBE)/methanol (MeOH) 10:3 (v/v) (see [supplemental data](#), Common chemicals and lipid standards for details).

Annotation of lipid classes and species

Glycerolipids are referred to TG and DG; glycerophospholipids and lyso-glycerophospholipids to phosphatidic acids (PA), phosphatidylinositols (PI), phosphatidylserines (PS), phosphatidylglycerols (PG), phosphatidylethanolamines (PE), phosphatidylcholines (PC), ether phosphatidylethanolamines (PE O-), ether phosphatidylcholines (PC O-), lyso-phosphatidic acids (LPA), lyso-phosphatidylinositols (LPI), lyso-phosphatidylcholines (LPC), and lysophosphatidylethanolamines (LPE); sphingolipids to ceramides (Cer) and sphingomyelins (SM); sterols to cholesterol (Chol) and cholesteryl esters (CE). Species of glycerol- and glycerophospholipids and cholesteryl esters are annotated as <lipid class> <total number of carbon atoms> : <total number of double bonds> in both (or, for lyso-lipids and cholesteryl esters, in one) fatty acid or fatty alcohol moieties (moiety). Sphingolipids are annotated as <lipid class> <total number of carbon atoms> : <total number of double bonds>; <total number of hydroxyl groups> at the ceramide backbone.

Sample preparation for shotgun lipidomics

Biopsies (wet weight of 4.2–21.9 mg) were shock-frozen in liquid nitrogen ensuring an ex vivo time of less than 40 s and stored at –80°C freezer. Prior to lipid extraction, the tissues were homogenized in 300 µl of isopropanol using zirconium beads; the total protein content was determined by Pierce 660 assay (Thermo Fisher Scientific, USA). Lipids were extracted from aliquots containing an equivalent of 50 µg of total protein by adding 700 µl of MTBE/MeOH 10:3 (v/v) containing the internal standard mix (41, 42) (see details in [supplemental data](#), Sample preparation). After evaporation of the organic phase, lipid extracts were reconstituted in 600 µl of 2:1 (v/v) MeOH / CHCl₃ and stored at –20°C. Ten microliters of a lipid extract were diluted with 90 µl of the spray solution (4:2:1 isopropanol/MeOH/CHCl₃ (v/v/v) containing 7.5 mM ammonium formate) for mass spectrometric analysis. Samples were analyzed in technical duplicates.

Lipid identification and quantification by shotgun mass spectrometry

The mass spectrometric analysis was performed on a Q Exactive instrument (Thermo Fisher Scientific, Bremen,

Germany) equipped with a robotic nanoflow ion source TriVersa NanoMate (Advion BioSciences, Ithaca, NY) using nanoelectrospray chips with spraying nozzle diameter of 4.1 μm . The ion source was controlled by the Chipsoft 8.3.1 software (Advion BioSciences). Ionization voltage was +0.96 kV in positive and -0.96 kV in negative mode; backpressure was 1.25 psi in both modes (43). Temperature of the ion transfer capillary was 200 °C; S-lens RF level was 50%. FT MS spectra were acquired within the range of m/z 400–1,000 in positive and m/z 350–1,000 in negative ion mode at the target mass resolution of $R_{m/z\ 200}=140,000$; automated gain control (AGC) of 3×10^6 and maximal injection time of 3 s. In both modes t -SIM spectra were acquired at the same mass resolution and m/z range as above; AGC of 5×10^4 ; maximum injection time of 650 ms; width of isolation window of 20 Th. The inclusion list of masses targeted by t -SIM started at m/z 355 in negative and m/z 405 in positive ion mode and other masses were computed by adding 10 Da increment (i.e., m/z 355, 365, 375) up to m/z 1,005. Free cholesterol was quantified by parallel reaction monitoring (PRM) FT MS/MS (41, 44, 45) during the same analysis. The number of micro-scans was set to 1; width of precursor isolation window of 0.8 Th; normalized collision energy (nCE): 12.5%; AGC: 5×10^4 and maximum injection time of 3 s.

Raw t -SIM spectra were subjected to repetition rate filtering by PeakStrainer software (46) and then stitched together by SIMStitcher software (47). Lipids were identified by LipidXplorer software (48) by accurately determined m/z (mass accuracy better than 5 ppm) and quantified by comparing the isotopically corrected abundances of their molecular ions with the abundances of internal standards of the same lipid class. All internal standards were detected in all samples. MS² validation of selected species by HCD FT MS/MS was applied as described in supplemental data, MS² validation.

Raw data processing

The LipidXplorer output was processed using several steps. Technical replicates were averaged and spectra of QC samples (details in supplemental data, Pilot sub-cohort and Quality Control) set aside. In each sample lipid abundances for which the standard deviation (SD) exceeded 40% were set to Not-a-Number (NaN). Lipids whose abundance exceed the minimal value determined for this species by less than 2-fold were exempted from SD filtering (supplemental data, Raw data processing and supplemental Fig. S1). Also, cholesterol values were not SD-filtered because they were determined by PRM. Next, we grouped the patient samples according to their disease status (normal control, healthy obese, NAFL, NASH, and none). The abundance of lipid species that were detected in less than 15% of all patient samples of the same group was set to zero. Finally, we applied plate bias correction using the ComBat approach (49) as detailed in supplemental data, Raw data processing. The final dataset comprised the molar abundance of 316 lipid species from 22 lipid classes in 365 patients (supplemental dataset S1).

Bioinformatics and data mining

Biclustering analysis was performed using the isa (50) and qbic (51) algorithms. To eliminate redundancy of biclusters, we screened them using a consensus approach (see supplemental data, Bioinformatics and data mining).

Classifications were performed using random forest classifiers from the “randomForest” R package. Hundred

classifiers were trained per scenario on bootstrapped lipidomics data. The models were evaluated using precision recall (PR) and receiver operator characteristics (ROC) area under the curve (AUC) values on the test set. Feature importance was computed using gini index averaged over all bootstrapped models (For further details, see supplemental data, Bioinformatics and data mining).

Lipidomics impact of mutations

The impact of mutations was probed within each group (supplemental data, Genotyping), in which the lipidomes of homozygotes, heterozygotes, and normal genotypes were compared pairwise (details are in supplemental data, Statistical analysis of mutation-correlated lipidome changes).

RESULTS

Study cohort

Percutaneous or surgical liver biopsies were taken from in total of 365 Caucasian patients (124 males and 241 females) admitted to the University Hospital Schleswig-Holstein, Kiel, Germany during 2007–2016. Control biopsies were collected during gastrointestinal surgery for pathologies having no direct association with NAFLD. Each biopsy was examined by the same surgical pathologist (C. R.) (17) and assessed according to the NAFLD activity score (NAS) (39).

Patients were further sorted into four basal and one additional group (52) as outlined below. Biopsies were classified according to the presence of steatosis and inflammation as the main criteria to differentiate simple steatosis (NAFL) from steatohepatitis (NASH) (52). Therefore, individuals whose biopsies were having histological fat content $\leq 5\%$; BMI $< 30 \text{ kg/m}^2$; no histologically proven liver pathology; no lobular inflammation and absence of significant (stage 2 or higher) fibrosis formed the normal control (NC) group. Healthy obese (HO) group comprised patients having no NAFLD per clinical and histological evidence, yet their BMI exceeded 30 kg/m^2 . Patients were assigned to NASH and NAFL groups if the histological examination revealed steatosis with or without inflammation, respectively. In total, 337 patients were sorted into NC ($n = 49$), HO ($n = 51$), NAFL ($n = 143$), and NASH ($n = 94$) groups (Table 1; supplemental Table S1, age and BMI distribution is in supplemental Fig. S4).

Within the cohort, 28 patients who did not receive NAS because of serious liver conditions, such as histologically confirmed cancer with abnormal and higher than control levels of GGT, alkaline phosphatase (AP), and bilirubin, were assigned to the additional group “None.” Their lipidomes helped to delineate the impact of a broader spectrum of liver diseases different from NAFLD. The impact of obesity (53) was evaluated by comparing lipidomes of NC and HO groups. Since mean BMI of HO, NAFL, and NASH groups was similar, their lipidomes could be compared with no further adjustment. The mean age of NC group was

TABLE 1. Clinicopathological characteristics of the study cohort

Group	Number of patients			NAS Parameters ^a						Mean age, years	Mean BMI kg/m ²
	Total	Female	Male	NAS Fat	NAS Ballooning	NAS Inflammation	NAS Average	Fibrosis ^a			
NC	49	25	24	0	0	0	0	0	67.7	24.1	
HO	51	47	4	0	0	0	0	0	42.3	46.7	
NAFL	143	94	49	1	0	0	0	0.35	45.6	47.1	
NASH	94	62	32	2	0	1	4	1	45.1	49.3	
None	28	13	15	n.a.	n.a.	n.a.	n.a.	n.a.	57.1	33.7	

^aMedian values, in arbitrary units assigned according to (39).

12 years higher than of NAFL/NASH/None groups because of fewer younger people undergoing gastrointestinal surgery.

Shotgun lipidomics of liver biopsies

Although surgical biopsies were similar in size, they varied in weight by as much as 5-fold and also differed in fat and fibrotic content (Table 1). Since they were unique and histologically inhomogeneous, each specimen was analyzed as a single sample with no independently processed replicates. In an aliquot of each biopsy lysate, we first determined the total protein content and used it for subsequent normalization of molar lipid abundances.

To quantify the lipidomes, we employed high-resolution shotgun Fourier transform mass spectrometry (FT MS) that relies on direct nanoflow infusion of total lipid extracts (37, 38). During shotgun analysis, internal standards spiked into biopsy lysates prior lipid extraction are ionized together with endogenous lipids and enable their absolute (molar) quantification (17, 42, 54–56). Because of high and variable content of TG together with abundant chemical background, we analyzed liver extracts by the method of *t*-SIM (for targeted single ion monitoring) (47). In each analysis FT MS spectra were successively acquired from partially overlapping *m/z* windows of 20 Th and then these spectra were stitched together into a master spectrum by SIM-Stitcher software. Although typical *t*-SIM acquisition (supplemental Fig. S5) required ca. 11 min per sample (compared with less than 1 min required for conventional FT MS analysis), it increased the number of quantifiable lipid species by ca. 45% (57).

To ensure spectra acquisition consistency and enable batch correction of lipid abundances, we created a QC standard by pooling equal aliquots of 36 extracts that, according to a preliminary experiment, reflected extreme values of the total lipid content and represented both genders, stages of inflammation and fibrosis. In each aliquot of the QC sample, we quantified 255 lipid species from 19 lipid classes. The molar abundance of more than 98% of lipid species differed by less than 5% and only cholesterol (quantified by the method of PRM (57–59)) and a few TG species were detected with higher SD (supplemental Fig. S2).

The final dataset covering 365 biopsies comprised normalized molar abundances (in pmol per μ g of total

protein) of 316 lipid species from 22 lipid classes: Out of the total of 316 lipid species, only 114 were detected in more than 98% of biopsies (supplemental dataset S2).

Liver lipidome of the normal control group

The NC group ($n = 49$) combined patients with no apparent liver pathology and obesity, whose lipidome we assumed as basal (Fig. 1). The mol% of lipid classes generally corroborates previous reports (13) (supplemental Table S2), although we detected considerably more Cer, PS, SM, and lyso-lipids.

Since the liver is a hub of lipoprotein biosynthesis, we further examined if the NC lipidome differs from plasma lipidome (42, 56, 60). While both lipidomes have similar mol% proportions of PC, SM, and Chol, plasma contained 10-fold more CE (supplemental Fig. S6). Liver lipidome contains a larger variety of lyso-species (liver-specific are LPG, LPI, LPS, LPA), while LPC is more abundant in mol%. TG and DG are also more abundant in the liver, as well as PE, PG, PI, PS, and Cer. Lipid class profiles in the liver showed no pronounced gender bias (supplemental Fig. S6) that was apparent in healthy plasma (42).

Mutation-associated changes of the liver lipidome

We further evaluated how common mutations in NAFLD risk genes: *TM6SF2* variant rs58542926 (hcv89463510), *PNPLA3* variant rs738409 (hcv7241), *MBOAT7* variant rs641738 (hcv8716820), *SERPINA1* variant PiZ (Glu342Lys) rs28929474 (hcv34508510), *SERPINA1* variant PiS affected the liver lipidome. Within NC and NASH groups liver lipidomes of homozygotes, heterozygotes and normal genotype carriers were compared pairwise (Table 2). Note that, for consistency, we compared absolute (in moles per μ g of total protein) rather than relative (in mol% within each class) lipid abundances, although the latter could reveal more affected species (17) (supplemental Dataset S1).

MBOAT7 mutation-dependent changes were only detected in the NC group and confined to PI 36:1 (Table 2), consistent with its phosphatidylinositol *O*-acyltransferase activity and the role in NAFLD pathogenesis (17). In the NASH group *PNPLA3* rs738409 mutation elevated the levels of CE 22:2 and TG 56:7, while the total abundance of CE and TG classes did not change (supplemental Figs. S7, S8A,B) (19, 20, 29, 31, 61–63).

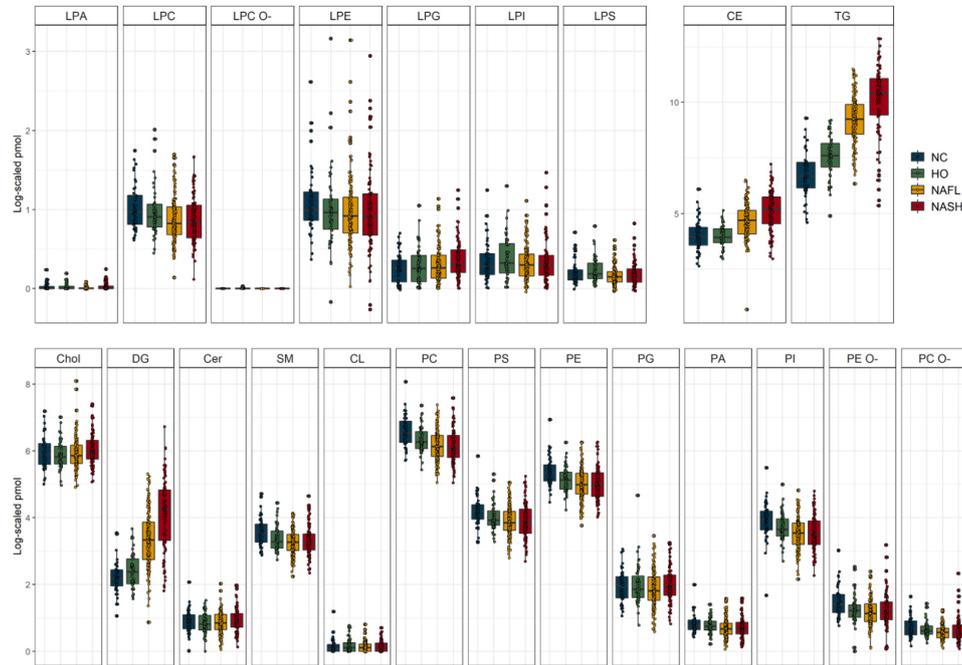


Fig. 1. Lipid class composition (in pmol per μg of total protein; \log_2 scaled) of liver biopsies in the four main groups of patients. The total abundance of each lipid class was calculated by adding up molar abundances of lipid species. Boxing highlights values between 25% and 75% quartiles; vertical lines connect minimum and maximum values excluding outliers. Filled circles stand for the lipid class abundance in individual biopsies (average of three technical replicates); black lines within boxes indicate median values. Color coding is shown in the inset at the right-hand side.

Condition-associated changes in lipid classes

There was no marked difference between the total abundance of lipid classes in disease groups compared with NC (Fig. 1), except significantly higher levels of TG and, to lesser extent, of DG and CE in NAFL and NASH patients (supplemental Figs. S9–S11). Although the mean BMI of HO patients was almost 2-fold higher compared with NC and was as high as in NAFL and NASH patients, the abundance of TG in HO and NC was similar.

In principal component analysis (PCA) plots of patients' lipidomes, the gradient across PC1 (principal

component 1) from NC (at the left) to NASH (at the right) (Fig. 2A) reflected the increased abundance of glycerolipids and CE (Fig. 2B). Interestingly, the profiles of hydrocarbon chain length and unsaturation of fatty acid moieties in TG and DG species were the same in all patient groups (Fig. 2C).

We further looked if TG and DG accumulated in the liver of NAFL and NASH patients were compositionally different from the adipose tissue. The ten most abundant TG species in liver (current dataset) and white adipose tissue (WAT) (64, 65) (supplemental

TABLE 2. *PNPLA3* and *MBOAT7* mutation-dependent changes of lipidome in NASH and NC groups

Lipid ^a	Mutation	Group	Mutation Status	Abundance Ratio
CE 22:2	<i>PNPLA3</i> rs738409	NASH	Heterozygote/No mutation	4.0 ^b
TG 56:7	<i>PNPLA3</i> rs738409	NASH	Heterozygote/No mutation	6.8 ^c
TG 56:7	<i>PNPLA3</i> rs738409	NASH	Homozygote/No mutation	15.1 ^c
TG 56:7	<i>PNPLA3</i> rs738409	NASH	Homozygote/Heterozygote	2.2 ^b
PI 36:1	<i>MBOAT7</i> rs641738	NC	Heterozygote/No mutation	2.2 ^b
PI 36:1	<i>MBOAT7</i> rs641738	NC	Homozygote/No mutation	4.5 ^d

^aSignificantly changed lipid in the two compared groups.

^bSignificance $P < 0.05$.

^cSignificance $P < 0.005$.

^dSignificance $P < 0.01$.

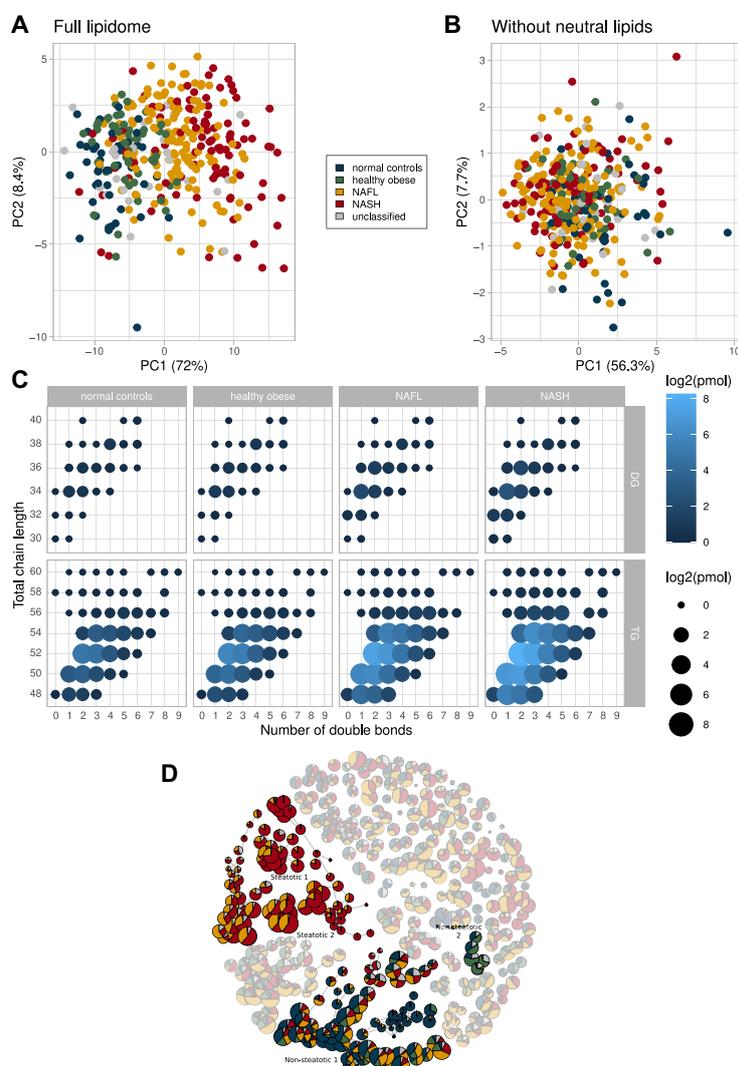


Fig. 2. Segregation of patient groups by the composition of lipidomes. A: PCA plot for full lipidomes of liver biopsies of 365 patients, whose disease group is indicated by color (coding scheme is in the inset); (B) PCA plot of lipidomes from which glycerolipids and CE were omitted. C: Length and unsaturation of fatty acid moieties in TG and DG species. Circle size and color reflect lipid abundances. D: Similarity network of biclusters. Node size is proportional to the number of patients in the bicluster. Highlighted are four network components (annotated) comprising the largest number of connected biclusters attributed to steatotic (in red and amber for NASH and NAFL) or nonsteatotic patients (in blue and green for NC and HO).

Table S3) were almost the same suggesting that TG metabolism in tissues is not organ-specific. However, they differed from TG in plasma of obese patients (66), presumably because in tissues TG accumulation sequesters lipotoxic FFA, while plasma TG are packed into lipoproteins for transporting via bloodstream (65).

Molecular stratification of patient groups by biclustering

Lipid class composition of liver biopsies (Figs. 1, 2) appears to be conserved and, apart from major increase in glycerolipids and CE, offers limited molecular insight or diagnostic perspective. We reasoned that

better molecular stratification of patients could rely on clusters of individual species spanning multiple lipid classes whose molar abundance coherently changes with the disease progression. Specifically, we employed a biclustering ensemble approach (Bioinformatics and data mining), in which lipidome compositions are clustered and the results consolidated into networks of connected components comprising lipid compositions specific for a selection of patients (Fig. 2D; supplemental Figs. S13, S14). Lipid compositions and meta-data could be compared across biclusters and reveal lipid signatures specific for a disease condition or patient group. Within all patients' lipidomes, we recognized two components comprising higher proportion of steatotic (NASH and NAFL) and two components with mostly nonsteatotic (NC and HO) patients (Fig. 2D, supplemental Table S4). Their lipidomes were clearly separated by PCA (supplemental Fig. S12A).

We first tested if lipids identified by biclustering could detail the trend toward steatosis and identify markers distinguishing nonsteatotic from steatotic groups by training random forest-based classifiers.

These groups were not separated by common clinical indices, e.g., total bilirubin, GGT, AP, alanine aminotransferase (ALT), and aspartate transaminase (AST) (supplemental Fig. S15B), but were readily distinguished by lipidome compositions (Fig. 3A). Classification performance was evaluated using the precision recall and ROCs (supplemental Fig. S16).

Lipids most significant for the classification were mono- and diunsaturated TG and DG (e.g., TG 50:1, TG 52:2, or DG 34:1) (supplemental Fig. S15A). To further test if unsaturation of fatty acid moieties in TG plays a role, we trained two classifiers with TG having in total 1–4 and 5–8 double bonds, respectively. Both models performed significantly better than random. However, the median Precision Recall Area Under the Curve (PR AUC) of the model with more saturated TG showed better performance (supplemental Fig. S15C, D).

Taken together, our analysis revealed several mono- and diunsaturated DG and TG species as lipid markers of steatosis while common membrane lipids (including glycerophospholipids (GPL), Chol, or Cer) were not in the markers list with one notable exception of two SM

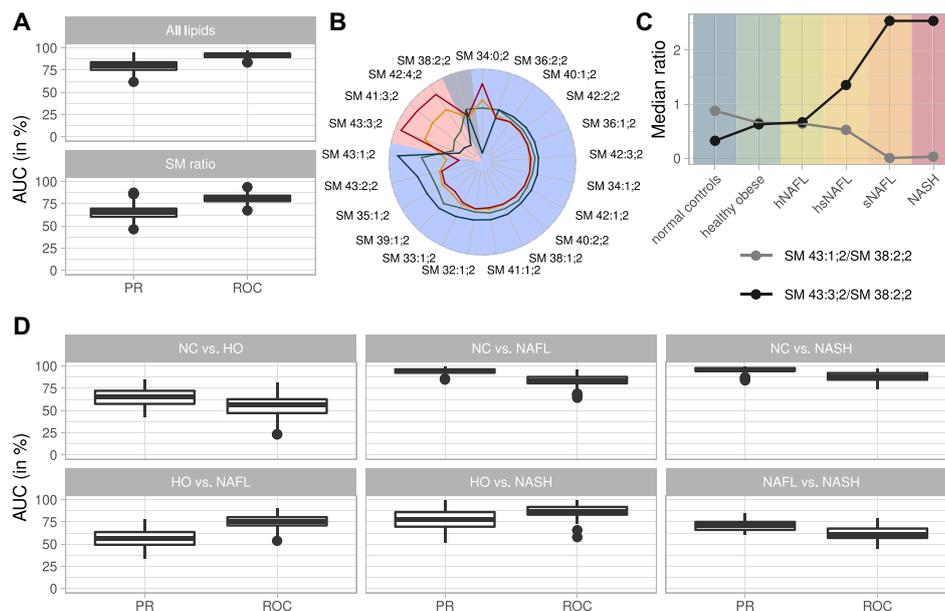


Fig. 3. Classification of disease groups by lipid markers. A: Classification of steatotic and nonsteatotic patients reported as an area under the curve (AUC) for precision recall (PR) and receiver operator characteristics (ROC). Classification based on all lipids (upper panel) and only by the ratio of SM 43:1:2 and SM 43:3:2 normalized to SM 38:2:2 (lower panel). B: Changes in the abundance of SM species in steatotic (red background) and nonsteatotic (blue background) components. SM 38:2:2 belongs to none of the two components (gray background). Lines indicate disease progression: NC (blue), HO (green), NAFL (yellow), and NASH (red) patient groups. Axes indicate relative abundances of SM species; data points per species sum up to 100%. C: Medians of the ratios of the abundance of SM 43:1:2 to SM 38:2:2 and SM 43:3:2 to SM 38:2:2 for the patient groups, including subgroups of NAFL. D: Classifications of all patient groups against each other using the two SM ratios.

species (supplemental Fig. S15A) that also showed clear NAFLD-dependent profile (Fig. 3B).

sphingomyelin species sharing n24:2 fatty acid moiety are same-class membrane lipid markers of NAFLD

We plotted the relative abundances of all SM species found in steatotic and nonsteatotic groups (Fig. 3B). The abundance of species from the nonsteatotic group (such as SM 42:2;2, SM 40:1;2, SM 42:1;2) either changed marginally or, as for SM 43:1;2, markedly decreased with NAFLD progression with the exception of SM 34:0;2 (Fig. 3B & supplemental Fig. S17).

In contrast, the abundance of SM 41:3;2, SM 43:3;2, and SM 42:4;2 from the steatotic group followed the opposite trend: it consistently increased with NAFLD progression. To elaborate on this finding, we first validated their identification by high-resolution HCD FT MS/MS (supplemental Materials and Methods, MS² validation). We note that these lipids were previously identified in human plasma by the method of LC-MSⁿ. Major species of SM 41:3;2 and SM 43:3;2 were recognized as SM d17:1/n24:2 and SM d19:1/n24:2, respectively (67); and SM 42:4;2 was SM d18:2/n24:2 (60), suggesting that these three SM may be sharing n24:2 fatty acid moiety. In contrast to SM 41:3;2, another compositionally related yet more saturated sphingomyelin SM 43:1;2 was attributed to nonsteatotic group and its abundance dropped with NAFLD progression.

To validate the molecular composition of liver sphingomyelins, we subjected a few combined extracts of steatotic and nonsteatotic biopsies to targeted LC-MSⁿ (67) (method details are in supplemental data, Identification of SM molecular species by LC-MSⁿ). The analysis (supplemental Table S5) confirmed that SM 34:0;2 belongs to dihydro sphingomyelins, the lipid class associated with fat deposition in organs, including the liver and pancreas (68). SM 42:4;2 and SM 43:1;2 comprised unsaturated (n24:2) and saturated (n24:0) fatty acid moieties, respectively (supplemental Fig. S19; supplemental Table S5). Because of their low abundance, LC-MSⁿ analysis of SM 43:3;2 and SM 41:3;2 was inconclusive.

We noticed that the abundance of SM 41:3;2 and SM 43:3;2 (both comprising n24:2) as compared with SM 41:1;2 (comprising n24:0) followed opposite trends (Fig. 3B), and we examined if their ratio to some unchanged SM species could distinguish different stages of NAFLD. We computed the ratios of SM 43:1;2 (decreasing in NAFLD) and of SM 43:3;2 (increasing in NAFLD) to SM 38:2;2 (SM d18:2/n20:0; see supplemental Table S5) that did not associate with steatotic and nonsteatotic groups and whose abundance was not affected by NAFLD (Fig. 3B). These ratios could be determined directly from a shotgun spectrum of a total lipid extract without prior adjustment to the abundance of internal standards or biopsy size (Fig. 3C). Strikingly, they distinguished nonsteatotic and steatotic subcohorts

with only slightly lower specificity than all lipids (Fig. 3A). SM-based classification was unaffected by obesity: it marginally distinguished NC and HO, but readily delineated NC from NAFL and NASH (Fig. 3D). To corroborate these findings, we further examined mol% profiles of species of SM and Cer classes in the four patient groups (supplemental Figs. S17, S18). They confirmed the notable decrease in the marker SM 41:1;2 as well as of the most abundant SM 34:1;2 together with the concomitant increase of SM 41:3;2 and SM 43:3;2 in NAFL and NASH subgroups, as compared with NC and HO. This, however, did not affect the abundance of matching Cer species: the mean value of Cer 34:1;2 was unchanged and none of Cer with odd number of carbon atoms was detectable.

We noted that SM ratios were less specific when we compared HO and NAFL (Fig. 3D). At the same time, HO and NASH were distinguished notably better albeit there was no global difference between their lipidomes except higher abundance of neutral lipids. Therefore, we hypothesized that the cohort of NAFL patients may be compositionally heterogeneous and consist of smaller, yet compositionally distinct subcohorts reflecting some intermediate stages of NAFLD pathogenesis. In this way, SM ratios and likely other lipid markers could reflect the transition from initial and, likely, reversible stage(s) of NAFL towards NASH.

Heterogeneity of NAFL lipidome and its molecular markers

We noticed that a component “Non-steatotic 1” (Fig. 2C) covers the largest number of nonsteatotic (NC+HO) patients, but also some patients with NAFL and NASH. Similarly, “Steatotic 2” mainly covers steatotic (NAFL and NASH) patients and almost no patients of NC and HO groups. We hypothesized that lipidomes of some NAFL patients may have higher similarity to the lipidomes of HO or of NASH. Based on the similarity of lipidome compositions, we used biclustering to divide NAFL patients into four subgroups that clustered together with nonsteatotic (h- for healthy) individuals, with NASH (s- for sick) individuals or neither h- nor s- (hs-) (Fig. 4). To test if this lipidome-based clustering reflects the disease progression also within the NAFL group, we compared the median values of individual histopathological indices (e.g., liver fat mass, ballooning, fibrosis, and inflammation) that were used for calculating NAS. We observed clear disease-related trends for each index and also for the total NAS (supplemental Table S6). Within NAFL subgroups, fat and ballooning expectantly increased from h-NAFL to s-NAFL, whereas fibrosis and inflammation did not, indicating that developing steatosis has not yet led to NASH. Expectantly, BMI did not change progressively between the subgroups.

Next, we subjected the patients subgroups to random-forest classification according to the following scenarios: (a) h-NAFL versus HO, (b) hs-NAFL versus HO, (c)

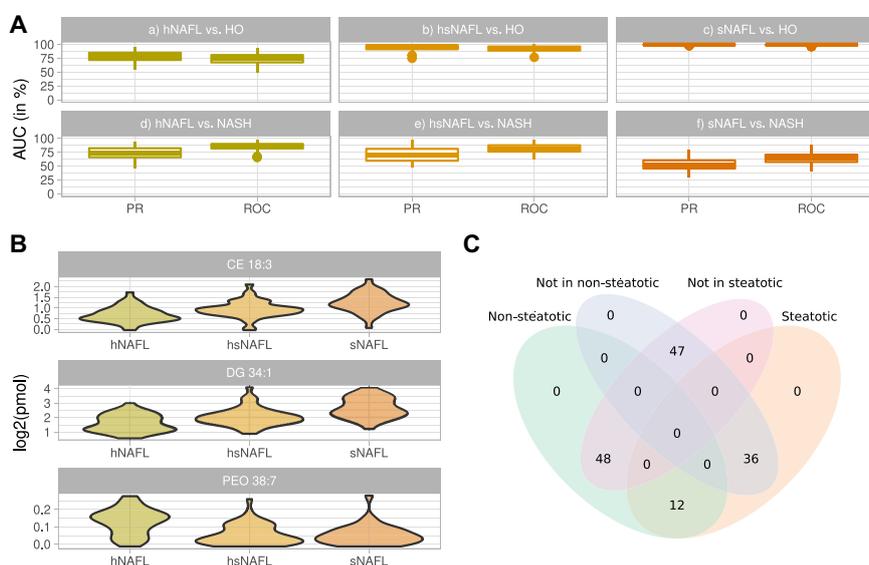


Fig. 4. Classification of NAFL subgroups. A: Classification of NAFL subgroups against HO and NASH groups by ratios of SM 43:1:2 to SM 38:2:2 and SM41:3:2 to SM 38:2:2 whose trend lines are shown in Fig. 3C. B: Violin plots of \log_2 -scaled abundances of CE 18:3, DG 34:1, and PE O-38:7 in the three NAFL subgroups. Color coding of in panels A and B is the same as in Fig. 3C. C: Distribution of NAFL patients in bicluster components indicating the representative number of patients in each subgroup.

s-NAFL versus HO, (d) s-NAFL versus NASH and (e) h-NAFL versus NASH, (f) h-NAFL versus NASH on the full lipidome (supplemental Fig. S20). As anticipated, we observed that the classifications performance increased from (a) to (c) and from (d) to (f) indicating that the lipidome of NAFL subgroups consistently changed with the disease progression. The change of SM ratios in subgroups of NAFL and also in four major groups of patients visualized this trend (Fig. 3C).

In these classifications DG 34:1, but also CE 18:3 and CE 18:2 were the most discriminating markers (supplemental Fig. S21). We next tested, if h-NAFL, hs-NAFL, and s-NAFL patients, selected solely by comparing their lipidomes, are also distinguishable histologically? If so, what marker lipids could reflect the transition from h-NAFL to s-NAFL and further toward NASH? The examination of histological indices indicated gradual increase in both inflammation and steatosis (supplemental Table S6). However, none of them reached typical NASH values, apart from % of accumulated fat. In their lipidomes, the abundance of CE 18:3 and DG 34:1 gradually increased from h-NAFL to s-NAFL (Fig. 4B).

We also classified lipidomes of NAFL patients from the steatotic bicluster component against all remaining NAFL patients and identified PE O-38:7 as the most discriminating marker (supplemental Fig. S22). Examination by HCD FT MS/MS identified it as plasmalogen

PE O-16:1/22:6 (supplemental Fig. S23 and supplemental Materials and Methods, MS² validation) whose abundance dropped already in hs-NAFL subgroup down to s-NAFL level. Another three polyunsaturated glycerophospholipids: LPE 22:6, PC 38:6, and PC 40:7 showed similar classification power (supplemental Fig. S20). Their sum formula suggests that, similar to PE O-38:7 and LPE 22:6, they also comprise C22:6 fatty acid moiety. We argue that stepwise decrease in the abundance of C22:6 - containing glycerophospholipids might indicate the onset of transition of h-NAFL toward s-NAFL, despite that there was only a marginal difference between histological indices of h-NAFL and hs-NAFL subgroups (supplemental Table S6). Interestingly, the classifications based on the SM ratios showed similar specificity (supplemental Fig. S17).

Taken together, our analysis distinguished non-steatotic (NC and HO) and steatotic (NAFL and NASH) cohorts by the accumulation of storage lipid markers TG 50:1, DG 34:1, CE 18:3 and, independently, by the ratios of sphingomyelins SM 43:3:2 and SM 41:3:2 SM to the reference SM 38:2:2. Furthermore, the SM ratios together with coherently changing C22:6 - containing glycerophospholipids (most specifically PE O-38:7, but also LPE 22:6, PC 38:6 and PC 40:7) revealed the compositional heterogeneity of the NAFL lipidome and how it changes, once the progressing disease gradually remodels HO-like lipidome toward NASH (Fig. 5).

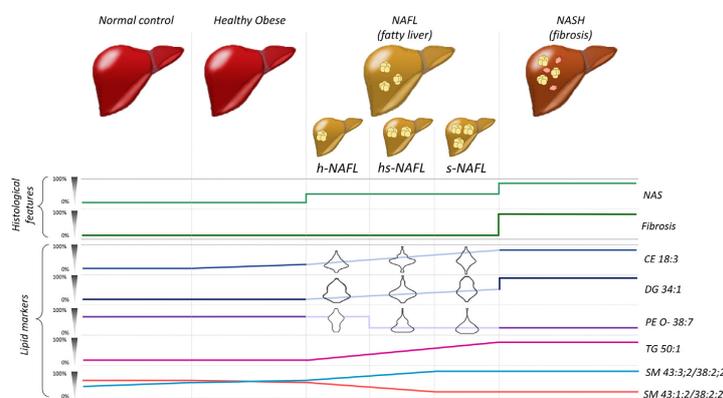


Fig. 5. Schematic trends of abundances of lipidomics and histological markers during NAFLD. The magnitude of change (in %) of lipid (Figs. 3 and 4) and histological (supplemental Table S4) markers scaled to the difference between NC and NASH. Cartoon images of the liver indicate progressive accumulation of fat eventually leading to inflammation and fibrosis.

DISCUSSION

Excessive accumulation of TG is the histological hallmark of NAFLD. While this alone links NAFLD to altered lipid metabolism, little was known if and how NAFLD globally alters the liver lipidome, particularly its membrane and signaling complements.

Our study confirmed the accumulation of neutral lipids along with NAFLD progression. Interestingly, the molecular composition of TG and DG species did not change and, altogether, was similar to white adipose tissue (WAT). Accordingly, biclustering analysis identified monounsaturated TAG 50:1 and DAG 34:1, but also unsaturated CE 18:3 as most specific markers of NAFLD progression whose abundance steadily increases from HO to NASH (Figs. 3–5). The impact of five risk genes implicated in NAFLD development corroborated the molecular specificity of *PNPLA3* and *MBOAT7*, but did not alter the liver lipidome globally.

Interestingly, the levels of glycerophospholipids and of sphingolipids, e.g., Cer and (with a few notable exceptions) SM were not perturbed significantly (Fig. 1). Also, no apparent association with altered SM/Cer metabolism emerged from the transcriptomics analysis (69). Although liver inflammation and apoptosis in NASH are supposed to increase the levels of Cer also by cleaving SM (9, 13, 32, 70, 71), we did not observe it in the biopsies. However, if SM directly contributed to or are associated with steatosis and its transformation to NASH was an open question (72). Specific bidirectional changes of the four SM species supported the robust same lipid class disease state classification. Conveniently, ratios of SM markers could be computed from raw intensities of corresponding molecular peaks: they did not depend on the biopsy size and did not require the full lipidome quantification. To the best of our knowledge, the prospective marker lipids SM 43:3:2, SM

41:3:2, and SM 42:4:2 were not previously spotted in the pathophysiological context of metabolic syndrome. Therefore, it seems promising to use targeted quantification to follow their levels in plasma as it might yield a convenient marker for independent and noninvasive stratification of NAFLD.

Sphingolipids with n24:2 moieties are produced from linoleic (C18:2) fatty acid (73). CE 18:2 is the most abundant cholesteryl ester in both the liver and plasma, and it is also enriched in NAFLD patients. It is therefore conceivable that enhanced biosynthesis of SM comprising n24:2 reflects increasing availability of free linoleic acid, while the abundance of SM with saturated or monounsaturated fatty acid moieties tends to decrease (Fig. 3B). However, relative changes of the abundance of SM species between steatotic and non-steatotic conditions are not proportional to their absolute (molar) content and are also sphingosine-backbone dependent. We noticed that two out of three marker SM comprise sphingosine backbone having odd number of carbon atoms. Therefore, it would be interesting to assess the contribution of microbiome lipids (reviewed in (74)) or *de novo* synthesis from branched amino acids (75).

One of the most intriguing findings of this work was the lipidomics evidence of the NAFL cohort heterogeneity (Figs. 3–5) that was not apparent from the histological examination (Table 1). In the NAFL subgroups that we conveniently termed as hs- and s-NAFL the levels of plasmalogen PE O-38:7, but also PC 38:6, PC 40:7 and LPI 22:6 (all comprising C22:6 moiety) dropped down to the level typical for NASH patients. In the same NAFL subgroups the dynamics of histopathological indices (supplemental Table S6) was concordant with progressive steatosis with no hallmarks of fibrosis and inflammation. We therefore speculate that altered SM ratios together with decreased levels of C22:6-

containing glycerophospholipids might indicate a turning point in NAFLD pathogenesis, where the transition from NAFL to NASH becomes irreversible.

While at this stage we cannot offer a suitable mechanistic explanation, we hypothesize that depletion of C22:6 containing GPL might affect the levels of PPAR α (76) and SREBP1c (77) and, eventually, promote steatosis and inflammation. Also, we cannot rule out that the decreased abundance of these lipids hints at enhanced oxidative stress, despite that shotgun profiling revealed no notable accumulation of oxidized TG and glycerophospholipids.

Irrespective of NAFLD stage, the organism strived to maintain the compositional identity of the hepatocyte membrane lipidome and changes were limited in both molecular scope and magnitude. The unexpected link between species-specific SM metabolism and NAFLD progression, together with NAFL-specific change in the levels of C22:6 - containing glycerophospholipids could contribute to both patients stratification and mechanistic understanding of the lipidome regulation during the disease.

Last but not least, this work created an open and transparent lipidomic resource that could be expanded with or cross-validated by further independent studies reporting molar abundances of liver lipids, irrespective of their research objectives and employed analytical methods. Studies design and reported evidence should enable direct comparison of lipidomics data, rather than their context-dependent interpretations and trends. Eventually, this may help to establish reference values of molar lipid abundances and use them for molecular diagnostics of a broad spectrum of liver pathologies. It will also complement ongoing efforts to create harmonized lipidomic resources for liquid and solid biopsies sharing similar principles of collecting and organizing the data (41, 44, 55, 60, 78).

Data availability

All data concerned with this study are presented within the manuscript, [supplemental data](#), and [supplemental datasets](#). 

Supplemental data

This article contains [supplemental data](#) (13, 17, 39, 42, 49, 50, 51, 64, 66, 67, 79–86).

Acknowledgments

The data that support the findings of this study are available from the corresponding author upon reasonable request. We thank Prof Juergen Graessler for his advice on the statistical analysis of the data. We are grateful to Prof Kai Simons, Dr Christian Klose, Dr Sider Penkov, and members of Shevchenko lab for fruitful discussions and experimental support.

Author contributions

O. V., T. D. R., J. Hampe, J. K. P., and A. S. conceptualization; O. V., O. K., A. P., J. A. H. W., J. M. A., C. H., M. B., V. R. T.,

S. B., and J. K. P. data curation; T. D. R., A. P., J. A. H. W., J. M. A., and C. H. formal analysis; H. C. K., Ü. C., E. K., J. Hampe, J. K. P., and A. S. funding acquisition; O. V., T. D. R., O. K., Y. W., T. Z., and J. Hartler investigation; T. D. R., O. K., J. A. H. W., K. S., Y. W., J. Hartler, H. C. K., J. K. P., and A. S. methodology; O. V., M. B., J. Hampe, and A. S. project administration; M. B., S. B., C. R., W. v. S., J. G., and C. S. resources; T. D. R., A. P., J. M. A., C. H., and J. K. P. software; H. C. K., C. R., Ü. C., E. K., W. v. S., J. G., C. S., J. Hampe, J. K. P., and A. S. supervision; O. V., O. K., K. S., and Y. W. validation; O. V., T. D. R., and A. P. visualization; O. V., T. D. R., J. K. P., and A. S. writing—original draft; M. B., V. R. T., S. B., J. Hartler, H. C. K., Ü. C., E. K., and C. S. writing—review and editing.

Author ORCIDs

Oskar Knittelfelder  <https://orcid.org/0000-0002-1565-7238>

Thomas Züllig  <https://orcid.org/0000-0002-8483-0962>

Jürgen Hartler  <https://orcid.org/0000-0002-1095-6458>

Harald C. Köfeler  <https://orcid.org/0000-0002-2725-9616>

Christoph Röcken  <https://orcid.org/0000-0002-6989-8002>

Jochen Hampe  <https://orcid.org/0000-0002-2421-6127>

Funding and additional information

This study was supported by the German Ministry of Research and Education (BMBF) through the Liver Systems Medicine (LiSyM) network grant (to J. H., E. K., and A. S.); Lipidomics and Informatics for Life Sciences (LIFS) unit of de.nBi consortium (to A. S.); TRR83 grant from Deutsche Forschungsgemeinschaft (DFG) to A. S. and U. C.; German Federal Ministry of Education and Research (BMBF) grant to the German Center for Diabetes Research (DZD e.V.) for U. C. T. D. R. and J. K. P. were supported by the Bavarian State Ministry of Science and the Arts coordinated by the Bavarian Research Institute for Digital Transformation (bidt). H. C. K. is supported by the Austrian Federal Ministry of Education, Science and Research (grant BMWFW-10.420/0005-WF/V/3c/2017) and by HRSM project “Explorative Lipidomics seltener und chronischer Krankheiten.” J. H. gratefully acknowledges financial support by the University of Graz.

Conflict of interest

The authors declare that they have no conflicts of interest with the contents of this article.

Abbreviations

ALT, alanine aminotransferase; AP, alkaline phosphatase; AST, aspartate transaminase; AUC, area under the curve; CE, cholesteryl ester; Cer, ceramide; Chol, cholesterol; DG, diacylglycerol; FFA, free fatty acid; GGT, gamma-glutamyl transferase; GPL, glycerophospholipid; HO, healthy obese; LPA, lyso-phosphatidic acid; LPC, lyso-phosphatidylcholine; LPE, lysophosphatidylethanolamine; LPI, lyso-phosphatidylinositol; NAFLD, nonalcoholic fatty liver disease; NASH, nonalcoholic steatohepatitis; NC, normal control; PA, phosphatidic acid; PC, phosphatidylcholine; PC O-, ether phosphatidylcholine; PE, phosphatidylethanolamine; PE O-, ether phosphatidylethanolamine; PG, phosphatidylglycerol; PI, phosphatidylinositol; PR, precision recall; PS, phosphatidylserine; ROC, receiver operator characteristic; SM, sphingomyelin; TG, triacylglycerol.

Manuscript received May 11, 2021, and in revised from July 20, 2021. Published, JLR Papers in Press, August 10, 2021, <https://doi.org/10.1016/j.jlr.2021.100104>

REFERENCES

- Marjot, T., Moolla, A., Cobbold, J. F., Hodson, L., and Tomlinson, J. W. (2020) Nonalcoholic fatty liver disease in adults: current concepts in etiology, outcomes, and management. *Endocr. Rev.* **41**, 66–117
- Musso, G., Gambino, R., De Michieli, F., Cassader, M., Rizzetto, M., Durazzo, M., Fagà, E., Silli, B., and Pagano, G. (2003) Dietary habits and their relations to insulin resistance and postprandial lipemia in nonalcoholic steatohepatitis. *Hepatology*. **37**, 909–916
- Lonardo, A., Nascimbeni, F., Ballestri, S., Fairweather, D., Win, S., Than, T. A., Abdelmalek, M. F., and Suzuki, A. (2019) Sex differences in nonalcoholic fatty liver disease: state of the art and identification of research gaps. *Hepatology*. **70**, 1457–1469
- Friedman, S. L., Neuschwander-Tetri, B. A., Rinella, M., and Sanyal, A. J. (2018) Mechanisms of NAFLD development and therapeutic strategies. *Nat. Med.* **24**, 908–922
- Hardy, T., Oakley, F., Anstee, Q. M., and Day, C. P. (2016) Nonalcoholic fatty liver disease: pathogenesis and disease spectrum. *Annu. Rev. Pathol. Mech. Dis.* **11**, 451–496
- Paschos, P., and Paletas, K. (2009) Non alcoholic fatty liver disease two-hit process: multifactorial character of the second hit. *Hypokratia*. **13**, 128
- Kleiner, D. E., Brunt, E. M., Wilson, L. A., Behling, C., Guy, C., Contos, M., Cummings, O., Yeh, M., Gill, R., Chalasani, N., Neuschwander-Tetri, B. A., Diehl, A. M., Dasarthy, S., Terrault, N., Kowdley, K., et al (2019) Association of histologic disease activity with progression of nonalcoholic fatty liver disease. *JAMA Netw. Open*. **2**, e1912565
- Lamotte, A., Leclercq, I., and Lanthier, N. (2020) The mechanisms of steatosis pathogenesis during NASH development. *Acta Gastroenterol. Belg.* **83**, 1575
- Simon, J., Ouro, A., Ala-Ibanibo, L., Presa, N., Delgado, T. C., and Martínez-Chantar, M. L. (2020) Sphingolipids in non-alcoholic fatty liver disease and hepatocellular carcinoma: ceramide turnover. *Int. J. Mol. Sci.* **21**, 40
- Tamura, S., and Shimomura, I. (2005) Contribution of adipose tissue and de novo lipogenesis to nonalcoholic fatty liver disease. *J. Clin. Invest.* **115**, 1139–1142
- Listenberger, L. L., Han, X., Lewis, S. E., Cases, S., Farese, R. V., Ory, D. S., and Schaffer, J. E. (2003) Triglyceride accumulation protects against fatty acid-induced lipotoxicity. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 3077–3082
- Alkhoury, N., Dixon, L. J., and Feldstein, A. E. (2009) Lipotoxicity in nonalcoholic fatty liver disease: not all lipids are created equal. *Expert Rev. Gastroenterol. Hepatol.* **3**, 445–451
- Gorden, D. L., Myers, D. S., Ivanova, P. T., Fahy, E., Maurya, M. R., Gupta, S., Min, J., Spann, N. J., McDonald, J. G., Kelly, S. L., Duan, J., Sullards, M. C., Leiker, T. J., Barkley, R. M., Quehenberger, O., et al (2015) Biomarkers of NAFLD progression: a lipidomics approach to an epidemic. *J. Lipid Res.* **56**, 722–736
- Sanders, F. W. B., and Griffin, J. L. (2016) De novo lipogenesis in the liver in health and disease: more than just a shunting yard for glucose. *Biol. Rev.* **91**, 452–468
- Knebel, B., Fahlbusch, P., Dille, M., Wahlers, N., Hartwig, S., Jacob, S., Kettel, U., Schiller, M., Herebian, D., Koellmer, C., Lehr, S., Müller-Wieland, D., and Kotzka, J. (2019) Fatty liver due to increased de novo lipogenesis: alterations in the hepatic peroxisomal proteome. *Front. Cell Dev. Biol.* **7**, 248
- Cohen, J. C., Horton, J. D., and Hobbs, H. H. (2011) Human fatty liver disease: old questions and new insights. *Science*. **332**, 1519–1523
- Thangapandi, V. R., Knittelfelder, O., Brosch, M., Patsenker, E., Vvedenskaya, O., Buch, S., Hinz, S., Hendricks, A., Nati, M., Herrmann, A., Rekhade, D. R., Berg, T., Matz-Soja, M., Huse, K., Klipp, E., et al (2021) Loss of hepatic Mboat7 leads to liver fibrosis. *Gut*. **70**, 940–950
- Krawczyk, M., Rau, M., Schattenberg, J. M., Bantel, H., Pathil, A., Demir, M., Kluwe, J., Boettler, T., Lammert, F., and Geier, A. (2017) Combined effects of the PNPLA3 rs738409, TM6SF2 rs58542926, and MBOAT7 rs641738 variants on NAFLD severity: a multicenter biopsy-based study. *J. Lipid Res.* **58**, 247–255
- Stickel, F., and Hampe, J. (2012) Genetic determinants of alcoholic liver disease. *Gut*. **61**, 150–159
- Stickel, F., Moreno, C., Hampe, J., and Morgan, M. Y. (2017) The genetics of alcohol dependence and alcohol-related liver disease. *J. Hepatol.* **66**, 195–211
- Jonas, W., and Schürmann, A. (2020) Genetic and epigenetic factors determining NAFLD risk. *Mol. Metab.* **50**, 101111
- Trépo, E., and Valenti, L. (2020) Update on NAFLD genetics: from new variants to the clinic. *J. Hepatol.* **72**, 1196–1209
- Strnad, P., Buch, S., Hamesch, K., Fischer, J., Rosendahl, J., Schmelz, R., Brueckner, S., Brosch, M., Heimes, C. V., and Woditsch, V. (2019) Heterozygous carriage of the alpha1-antitrypsin Pi* Z variant increases the risk to develop liver cirrhosis. *Gut*. **68**, 1099–1107
- Mancina, R. M., Dongiovanni, P., Petta, S., Pingitore, P., Meroni, M., Rametta, R., Borén, J., Montalcini, T., Pujia, A., and Wiklund, O. (2016) The MBOAT7-TMC4 variant rs641738 increases risk of nonalcoholic fatty liver disease in individuals of European descent. *Gastroenterology*. **150**, 1219–1230
- Di Sessa, A., Umamo, G. R., Cirillo, G., Del Prete, A., Iacomino, R., Marzuillo, P., and Del Giudice, E. M. (2018) The membrane-bound O-Acyltransferase7 rs641738 variant in pediatric non-alcoholic fatty liver disease. *J. Pediatr. Gastroenterol. Nutr.* **67**, 69–74
- Thabet, K., Asimakopoulos, A., Shojaei, M., Romero-Gomez, M., Mangia, A., Irving, W. L., Berg, T., Dore, G. J., Grønbaek, H., and Sheridan, D. (2016) MBOAT7 rs641738 increases risk of liver inflammation and transition to fibrosis in chronic hepatitis C. *Nat. Commun.* **7**, 1–8
- Thabet, K., Chan, H. L. Y., Petta, S., Mangia, A., Berg, T., Boonstra, A., Brouwer, W. P., Abate, M. L., Wong, V. W., and Nazmy, M. (2017) The membrane-bound O-acyltransferase domain-containing 7 variant rs641738 increases inflammation and fibrosis in chronic hepatitis B. *Hepatology*. **65**, 1840–1850
- Luukkonen, P. K., Zhou, Y., Hyötyläinen, T., Leivonen, M., Arola, J., Orho-Melander, M., Orešič, M., and Yki-Järvinen, H. (2016) The MBOAT7 variant rs641738 alters hepatic phosphatidylinositols and increases severity of non-alcoholic fatty liver disease in humans. *J. Hepatol.* **65**, 1263–1265
- Kawano, Y., and Cohen, D. E. (2013) Mechanisms of hepatic triglyceride accumulation in non-alcoholic fatty liver disease. *J. Gastroenterol.* **48**, 434–441
- Buch, S., Stickel, F., Trépo, E., Way, M., Herrmann, A., Nischalke, H. D., Brosch, M., Rosendahl, J., Berg, T., Ridinger, M., Rietschel, M., McQuillin, A., Frank, J., Kiefer, F., Schreiber, S., et al (2015) A genome-wide association study confirms PNPLA3 and identifies TM6SF2 and MBOAT7 as risk loci for alcohol-related cirrhosis. *Nat. Genet.* **47**, 1443–1448
- Sliz, E., Sebert, S., Würtz, P., Kangas, A. J., Soinen, P., Lehtimäki, T., Kähönen, M., Viikari, J., Männikkö, M., Ala-Korpela, M., Raitakari, O. T., and Kettunen, J. (2018) NAFLD risk alleles in PNPLA3, TM6SF2, GCKR and LYPLAL1 show divergent metabolic effects. *Hum. Mol. Genet.* **27**, 2214–2223
- Montefusco, D. J., Allegood, J. C., Spiegel, S., and Cowart, L. A. (2018) Non-alcoholic fatty liver disease: insights from sphingolipidomics. *Biochem. Biophys. Res. Commun.* **504**, 608–616
- Wang, H., Quiroga, A. D., and Lehner, R. (2013) Analysis of lipid droplets in hepatocytes. *Methods Cell Biol.* **116**, 107–127
- Lovric, A., Granér, M., Björnson, E., Arif, M., Benfeitas, R., Nymän, K., Ståhlman, M., Pentikäinen, M. O., Lundbom, J., Hakkarainen, A., Sirén, R., Nieminen, M. S., Lundbom, N., Lauerma, K., Taskinen, M. R., et al (2018) Characterization of different fat depots in NAFLD using inflammation-associated proteome, lipidome and metabolome. *Sci. Rep.* **8**, 1–14
- Alonso, C., Noureddin, M., Lu, S. C., and Mato, J. M. (2019) Biomarkers and subtypes of deranged lipid metabolism in non-alcoholic fatty liver disease. *World J. Gastroenterol.* **25**, 3009–3020
- Chiappini, F., Coilly, A., Kadar, H., Gual, P., Tran, A., Desterke, C., Samuel, D., Duclos-Vallée, J. C., Touboul, D., Bertrand-Michel, J., Brunelle, A., Guettier, C., and Le Naour, F. (2017) Metabolism dysregulation induces a specific lipid signature of nonalcoholic steatohepatitis in patients. *Sci. Rep.* **7**, 1–17
- Ryan, E., and Reid, G. E. (2016) Chemical derivatization and ultrahigh resolution and accurate mass spectrometry strategies for “shotgun” lipidome analysis. *Acc. Chem. Res.* **49**, 1596–1604

38. Schwudke, D., Schuhmann, K., Herzog, R., Bornstein, S. R., and Shevchenko, A. (2011) Shotgun lipidomics on high resolution mass spectrometers. *Cold Spring Harb. Perspect. Biol.* **3**, a004614
39. Kleiner, D. E., Brunt, E. M., Van Natta, M., Behling, C., Contos, M. J., Cummings, O. W., Ferrell, L. D., Liu, Y., Torbenson, M. S., Unalp-Arida, A., Yeh, M., McCullough, A. J., Sanyal, A. J., and Nonalcoholic Steatohepatitis Clinical Research Network (2005) Design and validation of a histological scoring system for nonalcoholic fatty liver disease. *Hepatology* **41**, 1313–1321
40. Danielsson, A. K., Lundin, A., and Andréasson, S. (2018) Using mobile phone technology to treat alcohol use disorder: study protocol for a randomized controlled trial. *Trials* **19**, 1–8
41. Sales, S., Knittelfelder, O., and Shevchenko, A. (2017) Lipidomics of human blood plasma by high-resolution shotgun mass spectrometry. *Methods Mol. Biol.* **1619**, 203–212
42. Sales, S., Graessler, J., Ciucci, S., Al-Atrih, R., Vihervaara, T., Schuhmann, K., Kauhainen, D., Sysi-Aho, M., Bornstein, S. R., Bickel, M., Cannistraci, C. V., Ekroos, K., and Shevchenko, A. (2016) Gender, contraceptives and individual metabolic predisposition shape a healthy plasma lipidome. *Sci. Rep.* **6**, 1–14
43. Schuhmann, K., Almeida, R., Baumert, M., Herzog, R., Bornstein, S. R., and Shevchenko, A. (2012) Shotgun lipidomics on a LTQ Orbitrap mass spectrometer by successive switching between acquisition polarity modes. *J. Mass Spectrom.* **47**, 96–104
44. Vvedenskaya, O., Wang, Y., Ackerman, J. M., Knittelfelder, O., and Shevchenko, A. (2019) Analytical challenges in human plasma lipidomics: a winding path towards the truth. *Trends Anal. Chem.* **120**, 115277
45. Knittelfelder, O., Prince, E., Sales, S., Fritzsche, E., Wöhner, T., Brankatschk, M., and Shevchenko, A. (2020) Sterols as dietary markers for *Drosophila melanogaster*. *Biochim. Biophys. Acta Mol. Cell Biol. Lipids* **1865**, 158683
46. Schuhmann, K., Thomas, H., Ackerman, J. M., Nagornov, K. O., Tsybin, Y. O., and Shevchenko, A. (2017) Intensity-independent noise filtering in FT MS and FT MS/MS spectra for shotgun lipidomics. *Anal. Chem.* **89**, 7046–7052
47. Schuhmann, K., Srzentić, K., Nagornov, K. O., Thomas, H., Gutmann, T., Coskun, Ü., Tsybin, Y. O., and Shevchenko, A. (2017) Monitoring membrane lipidome turnover by metabolic 15N labeling and Fourier transform mass spectrometry. *Anal. Chem.* **89**, 12857–12865
48. Herzog, R., Schwudke, D., Schuhmann, K., Sampaio, J. L., Bornstein, S. R., Schroeder, M., and Shevchenko, A. (2011) A novel informatics concept for high-throughput shotgun lipidomics based on the molecular fragmentation query language. *Genome Biol.* **12**, 1–25
49. Johnson, W. E., Li, C., and Rabinovic, A. (2007) Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118–127
50. Bergmann, S., Ihmels, J., and Barkai, N. (2003) Iterative signature algorithm for the analysis of large-scale gene expression data. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* **67**, 031902
51. Li, G., Ma, Q., Tang, H., Paterson, A. H., and Xu, Y. (2009) QUBIC: a qualitative biclustering algorithm for analyses of gene expression data. *Nucleic Acids Res.* **37**, e101
52. Brosch, M., Kattler, K., Herrmann, A., von Schönfels, W., Nordström, K., Seehofer, D., Damm, G., Becker, T., Zeissig, S., Nehring, S., Reichel, F., Moser, V., Thangapandi, R. V., Stichel, F., Baretton, G., et al. (2018) Epigenomic map of human liver reveals principles of zoned morphogenic and metabolic control. *Nat. Commun.* **9**, 1–11
53. Saywar, R., Pierce, N., and Koppe, S. (2018) Obesity and nonalcoholic fatty liver disease: current perspectives. *Diabetes Metab. Syndr. Obes. Targets Ther.* **11**, 533–542
54. Surma, M. A., Herzog, R., Vasilj, A., Klose, C., Christinat, N., Morin-Rivron, D., Simons, K., Masoodi, M., and Sampaio, J. L. (2015) An automated shotgun lipidomics platform for high throughput, comprehensive, and quantitative analysis of blood plasma intact lipids. *Eur. J. Lipid Sci. Technol.* **117**, 1540–1549
55. Wang, Y., Hinz, S., Uckeremann, O., Hönscheid, P., von Schönfels, W., Burmeister, G., Hendricks, A., Ackerman, J. M., Baretton, G. B., Hampe, J., Brosch, M., Schafmayer, C., Shevchenko, A., and Zeissig, S. (2020) Shotgun lipidomics-based characterization of the landscape of lipid metabolism in colorectal cancer. *Biochim. Biophys. Acta Mol. Cell Biol. Lipids* **1865**, 158579
56. Bowden, J. A., Heckert, A., Ulmer, C. Z., Jones, C. M., Koelmel, J. P., Abdullah, L., Ahonen, L., Alnouti, Y., Armando, A. M., Asara, J. M., Bamba, T., Barr, J. R., Bergquist, J., Borchers, C. H., Brandsma, J., et al. (2017) Harmonizing lipidomics: NIST interlaboratory comparison exercise for lipidomics using SRM 1950-Metabolites in frozen human plasma. *J. Lipid Res.* **58**, 2275–2288
57. Knittelfelder, O., Traikov, S., Vvedenskaya, O., Schuhmann, A., Segeletz, S., Shevchenko, A., and Shevchenko, A. (2018) Shotgun lipidomics combined with laser capture microdissection: a tool to analyze histological zones in cryosections of tissues. *Anal. Chem.* **90**, 9868–9878
58. Higashi, T., and Shimada, K. (2004) Derivatization of neutral steroids to enhance their detection characteristics in liquid chromatography-mass spectrometry. *Anal. Bioanal. Chem.* **378**, 875–882
59. Casanovas, A., Hannibal-Bach, H. K., Jensen, O. N., and Ejsing, C. S. (2014) Shotgun lipidomic analysis of chemically sulfated sterols compromises analytical sensitivity: recommendation for large-scale global lipidome analysis. *Eur. J. Lipid Sci. Technol.* **116**, 1618–1620
60. Quehenberger, O., and Dennis, E. A. (2011) The human plasma lipidome. *N. Engl. J. Med.* **365**, 1812–1823
61. Bruschi, F. V., Claudel, T., Tardelli, M., Caligiuri, A., Stulnig, T. M., Marra, F., and Trauner, M. (2017) The PNPLA3 I148M variant modulates the fibrogenic phenotype of human hepatic stellate cells. *Hepatology* **65**, 1875–1890
62. BasuRay, S., Wang, Y., Smagris, E., Cohen, J. C., and Hobbs, H. H. (2019) Accumulation of PNPLA3 on lipid droplets is the basis of associated hepatic steatosis. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 9521–9526
63. Luukkonen, P. K., Nick, A., Hölttä-Vuori, M., Thiele, C., Iso-Kuorti, E., Lallukka-Brück, S., Zhou, Y., Hakkarainen, A., Lundbom, N., Peltonen, M., Orho-Melander, M., Orešič, M., Hyötyläinen, T., Hodson, L., Ikonen, E., et al. (2019) Human PNPLA3-I148M variant increases hepatic retention of polyunsaturated fatty acids. *JCI Insight* **4**, e127902
64. Al-Sari, N., Suvitaival, T., Mattila, I., Ali, A., Ahonen, L., Trost, K., Henriksen, T. F., Pociot, F., Dragsted, L. O., and Legido-Quigley, C. (2020) Lipidomics of human adipose tissue reveals diversity between body areas. *PLoS One* **15**, e0228521
65. Alves-Bezerra, M., and Cohen, D. E. (2017) Triglyceride metabolism in the liver. *Compr. Physiol.* **8**, 1–8
66. Graessler, J., Schwudke, D., Schwarz, P. E. H., Herzog, R., Shevchenko, A., and Bornstein, S. R. (2009) Top-down lipidomics reveals ether lipid deficiency in blood plasma of hypertensive patients. *PLoS One* **4**, e6261
67. Hardter, J., Armando, A. M., Trötz Müller, M., Dennis, E. A., Köfeler, H. C., and Quehenberger, O. (2020) Automated annotation of sphingolipids including accurate identification of hydroxylation sites using MSⁿ data. *Anal. Chem.* **92**, 14054–14062
68. Wu, Z. E., Fraser, K., Kruger, M. C., Sequeira, I. R., Yip, W., Lu, L. W., Plank, L. D., Murphy, R., Cooper, G. J. S., Martin, J.-C., Hollingsworth, K. G., and Poppitt, S. D. (2021) Untargeted metabolomics reveals plasma metabolites predictive of ectopic fat in pancreas and liver as assessed by magnetic resonance imaging: the TOPI-Asia study. *Int. J. Obes.* **45**, 1844–1854
69. Sen, P., Govaere, O., Sinojta, T., McGlinchey, A., Geng, D., Ratziu, V., Bugianesi, E., Schattenberg, J. M., Vidal-Puig, A., and Allison, M. (2021) Quantitative genome-scale analysis of human liver reveals dysregulation of glycosphingolipid pathways in progressive nonalcoholic fatty liver disease. *medRxiv*. <https://doi.org/10.1101/2021.02.09.21251354>
70. Iqbal, J., Walsh, M. T., Hammad, S. M., and Hussain, M. M. (2017) Sphingolipids and lipoproteins in health and metabolic disorders. *Trends Endocrinol. Metab.* **28**, 506–518
71. Aguilera-Romero, A., Gehin, C., and Riezman, H. (2014) Sphingolipid homeostasis in the web of metabolic routes. *Biochim. Biophys. Acta* **1841**, 647–656
72. Bikman, B. T., and Summers, S. A. (2011) Sphingolipids and hepatic steatosis. *Adv. Exp. Med. Biol.* **721**, 87–97
73. Edagawa, M., Sawai, M., Ohno, Y., and Kihara, A. (2018) Widespread tissue distribution and synthetic pathway of polyunsaturated C24:2 sphingolipids in mammals. *Biochim. Biophys. Acta Mol. Cell Biol. Lipids* **1863**, 1441–1448
74. Adolph, T. E., Grander, C., Moschen, A. R., and Tilg, H. (2018) Liver–microbiome axis in health and disease. *Trends Immunol.* **39**, 712–723
75. Wallace, M., Green, C. R., Roberts, L. S., Lee, Y. M., McCarville, J. L., Sanchez-Gurmaches, J., Meurs, N., Gengatharan, J. M., Hoyer,

- J. D. and Phillips, S. A. (2018) Enzyme promiscuity drives branched-chain fatty acid synthesis in adipose tissues. *Nat. Chem. Biol.* **14**, 1021–1031
76. Jump, D. B. (2008) N-3 polyunsaturated fatty acid regulation of hepatic gene transcription. *Curr. Opin. Lipidol.* **19**, 242–247
77. Botolin, D., Wang, Y., Christian, B., and Jump, D. B. (2006) Docosahexaenoic acid (22:6,n-3) regulates rat hepatocyte SREBP-1 nuclear abundance by Erk- and 26S proteasome-dependent pathways. *J. Lipid Res.* **47**, 181–192
78. Burla, B., Arita, M., Arita, M., Bendt, A. K., Cazenave-Gassiot, A., Dennis, E. A., Ekroos, K., Han, X., Ikeda, K., and Liebisch, G. (2018) MS-based lipidomics of human blood plasma: a community-initiated position paper to develop accepted guidelines. *J. Lipid Res.* **59**, 2001–2017
79. Li, P., Wu, Q., and Burges, C. J. (2008) McRank: learning to rank using multiple classification and gradient boosting. *Adv. Neural Inf. Process. Syst.*, 897–904
80. Triebel, A., Trötzmüller, M., Hartler, J., Stojakovic, T., and Köfeler, H. C. (2017) Lipidomics by ultrahigh performance liquid chromatography - high resolution mass spectrometry and its application to complex biological samples. *J. Chromatogr. B.* **1053**, 72–80
81. Fauland, A., Köfeler, H. C., Trötzmüller, M., Knopf, A., Hartler, J. J., Eberl, A., Chitraju, C., Lankmayr, E., and Spener, F. (2011) A comprehensive method for lipid profiling by liquid chromatography-ion cyclotron resonance mass spectrometry. *J. Lipid Res.* **52**, 2314–2322
82. Hartler, J., Triebel, A., Ziegl, A., Trötzmüller, M., Rechberger, G. N., Zeleznik, O. A., Zierler, K. A., Torta, F., Cazenave-Gassiot, A., and Wenk, M. R. (2017) Deciphering lipid structures based on platform-independent decision rules. *Nat. Methods* **14**, 1171–1174
83. Hartler, J., Trötzmüller, M., Chitraju, C., Spener, F., Köfeler, H. C., and Thallinger, G. G. (2011) Lipid Data Analyzer: unattended identification and quantitation of lipids in LC-MS data. *Bioinformatics* **27**, 572–577
84. Hanczar, B., and Nadif, M. (2011) Using the bagging approach for biclustering of gene expression data. *Neurocomputing* **74**, 1595–1605
85. Srnad, P., Buch, S., Hamesch, K., Fischer, J., Rosendahl, J., Schmelz, R., Brueckner, S., Brosch, M., Heimes, C. V., Woditsch, V., Scholten, D., Nischalke, H. D., Janciauskiene, S., Mandorfer, M., Trauner, M., *et al.* (2019) Heterozygous carriage of the alpha1-antitrypsin Pi* Z variant increases the risk to develop liver cirrhosis. *Gut* **68**, 1099–1107
86. Stöckel, F., Buch, S., Nischalke, H. D., Weiss, K. H., Gotthardt, D., Fischer, J., Rosendahl, J., Marot, A., Elamly, M., and Casper, M. (2018) Genetic variants in *PNPLA3* and *TM6SF2* predispose to the development of hepatocellular carcinoma in individuals with alcohol-related cirrhosis. *Am. J. Gastroenterol.* **113**, 1475–1483

A.3. Investigating Global Lipidome Alterations with the Lipid Network Explorer



Article

Investigating Global Lipidome Alterations with the Lipid Network Explorer

Nikolai Köhler [†], Tim Daniel Rose [†], Lisa Falk and Josch Konstantin Pauling ^{*}

LipiTUM, Chair of Experimental Bioinformatics, TUM School of Life Sciences, Technical University of Munich, 85354 Freising, Germany; nikolai.koehler@tum.de (N.K.); tim.rose@wzw.tum.de (T.D.R.); lisa.falk@tum.de (L.F.)

^{*} Correspondence: josch.pauling@wzw.tum.de

[†] These authors contributed equally to this work.

Abstract: Lipids play an important role in biological systems and have the potential to serve as biomarkers in medical applications. Advances in lipidomics allow identification of hundreds of lipid species from biological samples. However, a systems biological analysis of the lipidome, by incorporating pathway information remains challenging, leaving lipidomics behind compared to other omics disciplines. An especially uncharted territory is the integration of statistical and network-based approaches for studying global lipidome changes. Here we developed the Lipid Network Explorer (LINEX), a web-tool addressing this gap by providing a way to visualize and analyze functional lipid metabolic networks. It utilizes metabolic rules to match biochemically connected lipids on a species level and combine it with a statistical correlation and testing analysis. Researchers can customize the biochemical rules considered, to their tissue or organism specific analysis and easily share them. We demonstrate the benefits of combining network-based analyses with statistics using publicly available lipidomics data sets. LINEX facilitates a biochemical knowledge-based data analysis for lipidomics. It is available as a web-application and as a publicly available docker container.

Keywords: computational lipidomics; computational systems biology; network biology; bioinformatics; lipidomics; lipids; metabolic networks



Citation: Köhler, N.; Rose, T.D.; Falk, L.; Pauling, J.K. Investigating Global Lipidome Alterations with the Lipid Network Explorer. *Metabolites* **2021**, *11*, 488. <https://doi.org/10.3390/metabo11080488>

Academic Editor: Hunter N. B. Moseley

Received: 6 July 2021

Accepted: 27 July 2021

Published: 28 July 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Lipids play a central role in biology for membranes, energy metabolism and signaling processes. Lipidomics is gaining impact in systems biology and medicine as lipids are an important molecular dimension for the investigation of biological mechanisms, stratification of patients, and disease subtyping. Recent advances in extraction protocols, high resolution Mass Spectrometry (MS) and methods for the identification and quantification of lipids allow for more comprehensive and complex lipidomes to be measured. However, the analysis of lipidomics data does not end with quantification. To interpret changes of the lipidome and embed them into a systems biological context, dedicated computational approaches are necessary. The software tools lipidr [1] and LipidSuite [2] provide statistical methods to mine and perform differential analysis of lipidomics data. They implement a “Lipid Set Enrichment Analysis” and “Lipid chain analysis” to investigate the regulation of lipid classes, carbon chains or saturations. These approaches incorporate lipid-specific characteristics into the statistical analysis. However, the possibility to investigate associations between lipids is missing.

Association networks from molecular omics data can offer benefits for data analysis, as biological networks carry information about functional interactions of biomolecules. Examples are Protein-Protein Interaction (PPI) networks, Gene Regulatory (GR) networks, or metabolic networks. In the case of lipid metabolic networks, these characterize transformations of lipids catalyzed by enzymes. Dedicated bioinformatics tools such as Key-PathwayMiner [3,4], DOMINO [5] or HotNet2 [6] have been developed, which extract

functionally associated network modules enriched with deregulated genes/proteins from PPI networks in a case/control setting. Such network modules can hint towards biochemical mechanisms, which connect a phenotype to its underlying molecular machinery. Applying network-based computational methods on lipidomics data remains challenging. One reason is that reaction databases carry information mainly on a lipid class level but not on a molecular species level [7,8]. Since modern lipidomics experiments provide measurements on the sum or molecular species level, more fine-grained reaction information can be utilized. Therefore, (partial) correlation networks of lipids species can be used to investigate data-driven interactions between lipids.

Correlation networks are a common method for the analysis of metabolomics/lipidomics data [9–11]. They show relationships between lipids entirely based on pairwise correlations over all measured samples. While they can reveal novel relationships between lipids, they do not describe functional associations between them. Recently it was shown that correlation networks can profit from incorporating prior knowledge into cut-off selection [12], providing an alternative to purely data-driven or purely knowledge-driven metabolic networks. An interplay between functional and data-driven associations could therefore be beneficial for the analysis of lipidomics experiments.

Functional analysis of lipid data is already possible with tools such as LION/web [13] or BioPAN [14], which enrich lipids based on an ontology or pathways. LION/web identifies lipid-associated terms in lipidomes [13] and associates biological functions to lipidomics data. BioPAN visualizes biochemical pathways of lipids, which can be investigated on the lipid class, species or fatty acid (FA) metabolism level. Additionally, BioPAN provides quantitative scores for the activity of pathways. However, they focus on the enrichment of pathways or reaction chains rather than on a global analysis of the lipidome.

Another approach for the global qualitative analysis of the lipidome is the LUX Score [15]. The methodology embeds the lipidome in a chemical space, such that lipids are close to each other if they exhibit a high chemical similarity (based on SMILES notation of chemical structures). The LUX Score also operates on the lipid species level. It provides an overview of chemical properties and a qualitative comparison of lipidomes.

Here we present the Lipid Network Explorer (LINEX), a flexible web-application (app) to create, visualize and analyze functional lipidomics networks. It combines enzymatic transformations between lipids with correlations and statistical properties that can be superimposed onto the network. This enables a global and a local view on the lipidome. The tool thereby provides a basis for introducing graph-theoretical and network-topological approaches into the analysis of lipidomics data. We further present applications of LINEX on available lipidomics data sets and show the benefits of a network-based analysis.

2. Results

We developed LINEX to visualize and analyze functional associations of lipids on networks (Figure 1), enabling the investigation of lipidomics data in the context of metabolic reactions. In such networks, lipids are represented as nodes, while edges indicate a connection via enzymatic reactions of lipid classes or FAs (Figure A2a in Appendix A). These reactions are encoded as rules customizable by the user. This way, condition-, tissue-, or organism-specific lipid metabolic properties can be incorporated into an analysis with LINEX. As default settings, common reactions of glycerol-, glycerophospho- and sphingolipids as well as typical FA modifications are included. LINEX then combines reactions of lipid class and FA metabolism into one network to give a comprehensive overview of lipid species metabolism.

On the basis of experimental lipidomics data, and optional sample group annotation, data specific metabolic networks are computed. Supported by a data driven lipid network exploration, correlation analysis and hypothesis testing can be added to the network representation (Figure 1) for a combined analysis.

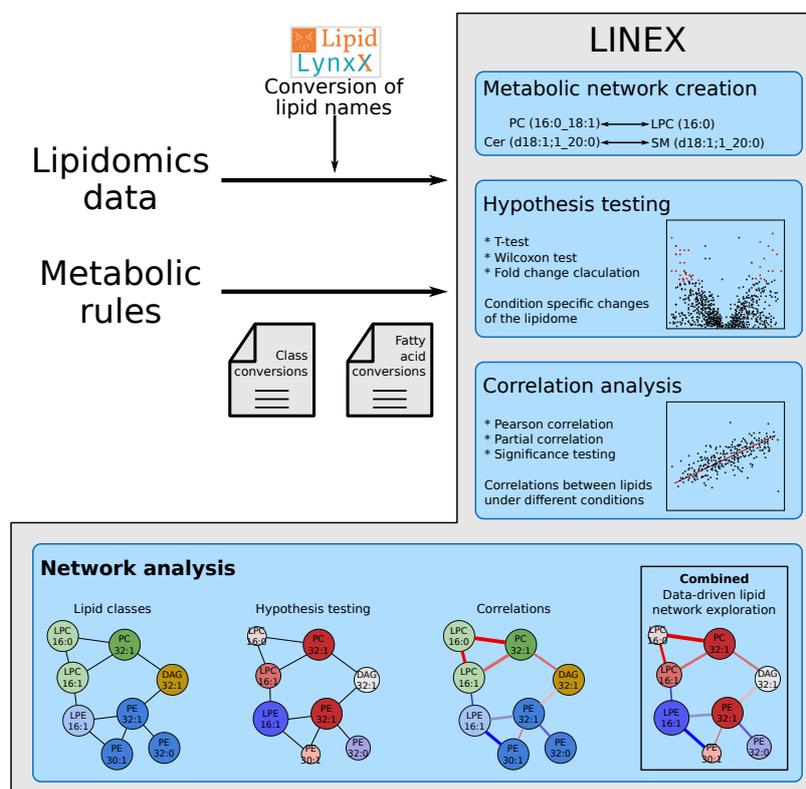


Figure 1. Workflow of the LINEX approach. Lipidomics data and optionally customized metabolic rules are uploaded by the user. The data are used to generate an experiment-specific lipid network, which can be visualized together with statistical measures such as correlation and fold change.

LINEX is available as a web-app (<https://exbio.wzw.tum.de/linex/> (accessed on 27 July 2021)), where lipidomics data can be uploaded (Figure A2a), networks computed and interactively visualized (Figure A2b). The lipidomics data have to be uploaded as one table (data from two ion modes have to be processed and combined by the users to one table prior to the analysis with LINEX). Additionally, the networks and all computed statistical measures can be downloaded (Figure A2c). In the following, we apply LINEX to three publicly available lipidomics datasets. They were selected to cover technical aspects such as MS1, MS2 and lipidome coverage and experimental designs such as case-control, time series and multi-group conditions. On those, we present our workflow to analyze combined metabolic and data driven lipid networks.

All networks shown in the results section are available as interactive HTML files (Supplementary Data 1–3).

2.1. Lipidomics of Colorectal Cancer

We investigated lipidomics data from Wang et al. [16] about a lipidomics characterization of colorectal cancer patients. The authors identified and quantified 342 lipid species

from 20 different lipid classes. According to the authors, no global changes of the lipidome were detected, but alterations in individual lipids were observed.

The network computed by LINEX (Figure 2a, interactive network: Supplementary Data 1) shows a global view on the changes of the lipidome between colorectal tumor and normal mucosa. In the network, each node represents a lipid species, and each edge between a pair of lipids indicates a biochemical reaction capable of transforming the lipids into each other on the class or FA level. Edges are colored by changes of correlation from healthy to cancer condition. Node colors represent the log fold change between healthy and cancer samples, with red indicating increased and blue indicating decreased lipid levels under healthy conditions. Node sizes indicate the negative log₁₀ FDR-values of a lipid between the two conditions, where more strongly altered lipids are displayed as larger nodes.

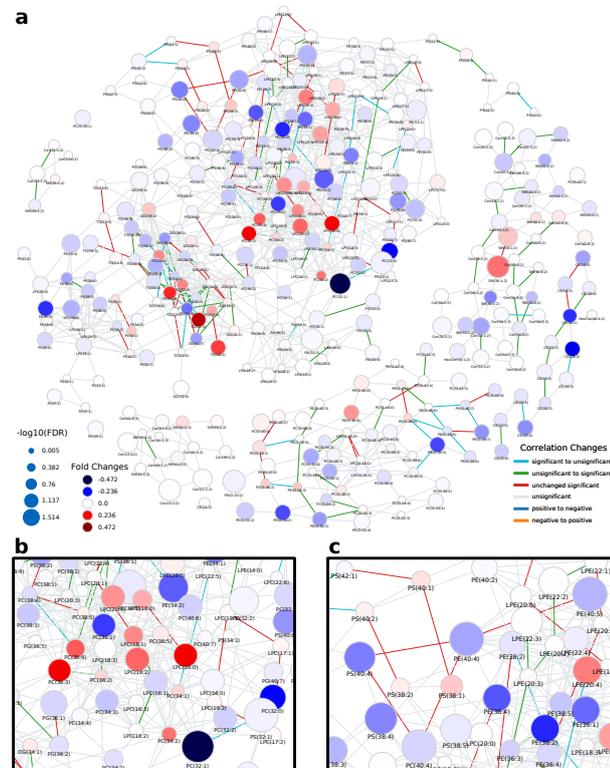


Figure 2. Lipid network of colorectal cancer lipidomics data from Wang et al. [16]. (a) Full lipid network with node size scaled by negative log₁₀ of *p*-values for comparison between healthy and cancer tissue. Lipids are colored by log fold change between healthy and cancer tissue. Blue colors indicate lower levels of lipids in the healthy condition compared to the tumor and red higher levels in healthy samples. Edges are colored by changes of correlation for lipids from the healthy to cancer condition. For example, green indicates a non-statistically significant correlation in the healthy condition and a statistically significant correlation in the tumor, where the correlation has the same sign. (b) Subnetwork showing PC and LPC nodes. (c) Subnetwork showing mainly unsaturated glycerophospholipids.

At first glance, it can be observed that the majority of reactions (edges) between lipid species do not represent significant correlations in either of the two conditions (FDR < 0.05, used throughout the manuscript as the significance cut-off). However, highly intra-connected parts of the network (local communities) can be observed, which exhibit significant correlations, indicated by colored edges. Some examples are triacylglycerol (TG) and diacylglycerol (DG) species (Figure A3a). While the fold changes of individual species are heterogeneous, a trend of higher unsaturated TG species increasing in tumor tissue and higher saturated TG species decreasing is observable. In particular, correlations between highly unsaturated TGs (52:5, 54:5, 54:6, 54:7) remain significant over both conditions, while others occur (green) or disappear (cyan) when comparing normal mucosa to tumor mucosa. This indicates changes in the regulation of the FA metabolism related to neutral lipids.

A big part of the network comprises the metabolism of GPLs. The network shows a set of phosphatidylcholine (PC) and lyso-phosphatidylcholine (LPC) species, which decrease in tumor samples and are metabolically closely related via reactions catalyzed by the MBOAT7 and PLA2 enzymes (Figure 2b). MBOAT7 expression has previously been associated with gastrointestinal cancer risk [17] as well as lipid-linked liver diseases [18], which we were able to link to lipidome alterations by only considering the LINEX network. The respective set of lipids is surrounded by PC, phosphatidylethanolamine (PE) and LPC species, which show the opposite behavior. We could also observe an interesting pattern of correlation of poly-unsaturated GPLs (Figure 2c). Here, PC, phosphatidylserine (PS) and PE species which have a sum composition of 40:4, and were all found to be significantly upregulated in the original publication additionally show functional correlations between each other, independent of the condition. This is a strong indication of a common mechanism regulating these lipid species.

In the metabolism of phosphatidylinositol (PI), high fold changes could be observed in poly-unsaturated PI species, while some highly connected lyso-phosphatidylinositol (LPI) species 18:2 and 16:0 did not seem to be influenced by the tumor (Figure 2a, left). The authors argued that ether lipids might play a role in tumor progression, especially lower levels of phosphatidylethanolamine ether (PEO) indicating higher oxidative stress. Our analysis shows a close biochemical connection between downregulated PEO species (Figure A3c). Other PEO species (e.g., PE(O-38:5) to PE(O-36:5), or PE(O-40:6) to PE(O-40:7)), which increase in the tumor condition only show significant correlation in healthy samples, revealing a diverging pattern in ether-PE. A reaction chain of ceramides with significant correlations could be observed in the sphingolipid metabolism component of the network (Figure A3b). While the Cers themselves are not significant, their correlations show a clear co-regulation. This shows that changes of individual lipids might not always be significant, but a combined network analysis with functional interactions and correlations can nevertheless reveal interesting relations between lipids as well as indicate putative common regulatory mechanisms.

2.2. Lipidome Alterations in Aging Brain of Mice

Next, we investigated a lipidomics experiment from Tu et al. [19] about lipidome changes in the aging brain of mice between the age of 4 weeks to 52 weeks. Although not compatible with the LipidLynxX [20] converter, we manually added Sulfatide and Hex2Cer to the metabolic rules. In contrast to the previous data set, we could observe very few correlations between lipids (Figure A4). To standardize the coloring of lipids in networks, we developed a unified color scheme on the lipid class level (see Section 4). The types of reactions forming edges between lipids are mainly chain length modifications and desaturations. Lipid headgroup modifications can be observed primarily between GPLs (Figure 3, interactive network: Supplementary Data 2). FA additions/removals are only found between DG(18:1_22:0) and three TG species. Figure 3 shows a subnetwork of highly saturated TG species, which are only connected via FA reactions. We first observed a decrease of TG species from 4 to 12 weeks, followed by a strong increase of TG levels starting

from the age of 32 weeks. This may be an indication for increased de novo lipogenesis, which might be explained with FAS (fatty acid synthase, preferentially synthesizes palmitic and stearic acid), SCD-1 (stearoyl-CoA desaturase, synthesizes palmitoleic and oleic acid), and GPAT-1 (glycerol-3-phosphate acyltransferase, preference for saturated FAs) enzyme activity [21]. This is an advantage of LINEX, which can depict relations of lipids also based on FA metabolism. The example also shows the importance of coverage of the lipidome. The more species available, the better connections between lipids can be inferred, ultimately helping to understand lipid metabolic alterations. The particular example lacks lyso-glycerophospholipids, which play a central role in the metabolism. Many lipids remain unconnected in this example or form components of less than four lipids, which makes the biological interpretation of the lipidome in the network context challenging (Figure A4).

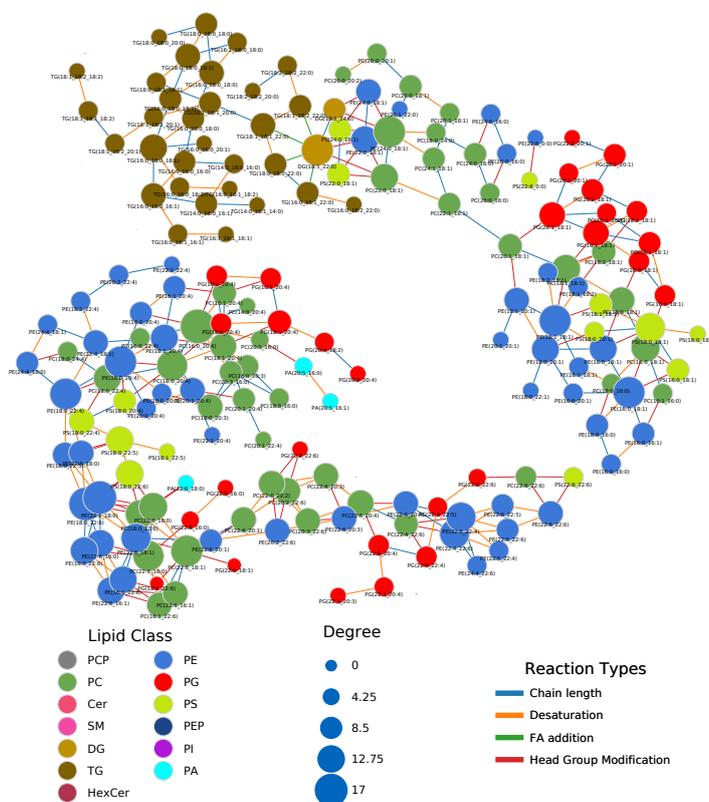


Figure 3. Part of the lipid network of the lipidomics data from Tu et al. [19]. Shown are the two main components of the GPL metabolism. Nodes are colored by lipid class, and edges are colored by reaction type. Node sizes represent the degree.

In the previous example on the lipidome of colorectal cancer patients, one GPL component could be observed. Based on the data of Tu et al. [19], multiple such components can be found. The two biggest components can be seen in Figure 3. Both share a similar set of FAs from C16 to C22. The topological structures of both components also show similarities. Many triangles of PC, PE and PS species can be found, which share the

same FA signature and are converted into each other by headgroup modifications (e.g., PC(18:0_18:1), PE(18:0_18:1), PS(18:0_18:1) or PC(18:0_20:4), PE(18:0_20:4), PS(18:0_20:4)). In some cases, additional connections to phosphatic acid (PA) or phosphatidylglycerol (PG) can be found. Other GPLs are connected purely via FA reactions (e.g., PE(22:5_22:6)). This pattern shows that certain FA combinations for GPLs seem favorable for enzymatic reactions, because they do not only occur in pairs but directly for up to five different lipid classes, which can be converted into each other.

Tu et al. [19] reported an overall decrease of GPLs and increase of sphingolipids and neutral lipids. With LINEX, we could visualize this trend on the whole lipidome (Figure A5). The global changes from the 4 week to the 12 week measurements were specific on the molecular species level, with small fold changes from 12 to 24 week old mice. The next change from 24 to 32 week probes showed the previously mentioned effect clearly with the GPL components being mainly decreased (red) and the rest mainly increased (blue). Interestingly, the ether lipids increased and therefore behaved opposite to the other GPLs. Finally, the comparison of 32 to 52 week old mice showed a similar pattern as the previous comparison, but with increased fold changes, especially in highly connected GPL such as PE(18:1_18:1), PC(16:0_20:4) or PE(22:4_22:6).

2.3. Healthy Human Reference Plasma Lipidome in Aging

As a third example, we are showcasing plasma lipidome data from a human reference population presented in Kyle et al. [22], which comprises 136 samples and 302 lipids, mostly identified as molecular species. All patients do not suffer from any diagnosed disease and represent the United States population in terms of age and sex distribution. To enable statistical comparisons, we grouped the patients by age (see Section 4 for details) and investigated the changes of the lipidome from young to old.

Many edges in the network (Figure 4, interactive network: Supplementary Data 3) show non-statistically significant correlations in any of the age groups, as indicated by the large fraction of gray edges, especially in the area rich in PCs and PEs in the upper right part of the lipid network (compare Figure A6). Those areas, which show statistically significant correlations, do so in half of the groups, namely at the 'Toddler', 'Child' and 'Elder' stage. While these reactions affect PCs and PEs with a variety of molecular compositions, most of these reactions are FA related, which becomes especially clear for PC species with odd-chain FA on the lower right side of the subnetwork. Interestingly, many lipids in this subnetwork show differential abundances between toddlers and children (Figure A6a), which is accompanied by a higher density of strong correlations. For comparisons including young adults (Figure A6c,d), both the number of lipid species with a higher probability of being different between sample groups and the number of edges with changes in correlation are much lower in this area of odd-chain PCs. Considering the general structure of the subnetworks shown in Figure A6, these two groups show an interesting behavior with respect to the position of lipid species with high absolute fold changes, which are located mostly on the outside of the network, corresponding to lower node degree and betweenness centrality. Most changes in correlation, however, are happening in the inner part around higher connected nodes, especially lyso-species. A possible explanation for this phenomenon is that changes in the center of the network are propagated to more peripheral parts, while intermediate nodes stay nearly unaffected in their abundances, as reactions producing and transforming them are changing their activities to the same degree.

In contrast to the part of the network shown in Figure A6, the subnetwork depicted in Figure A7 mainly comprises TG species and is only lightly connected. This is possibly due to few reported DG species, which would be connected to multiple TGs similar to LPC species connecting PCs. Considering all four age comparisons (i) Toddler vs. Child (ii) Child vs. Teenager (iii) Teenager vs. Adult (iv) Adult vs. Elder), most edges are either statistically significantly correlated in multiple comparisons or in none. This indicates constant metabolic activities shared across different age stages. Generally patients

grouped as children, teenagers and young adults (see Section 4 for details) only show minor differences in TG levels (Figure A7b,c).

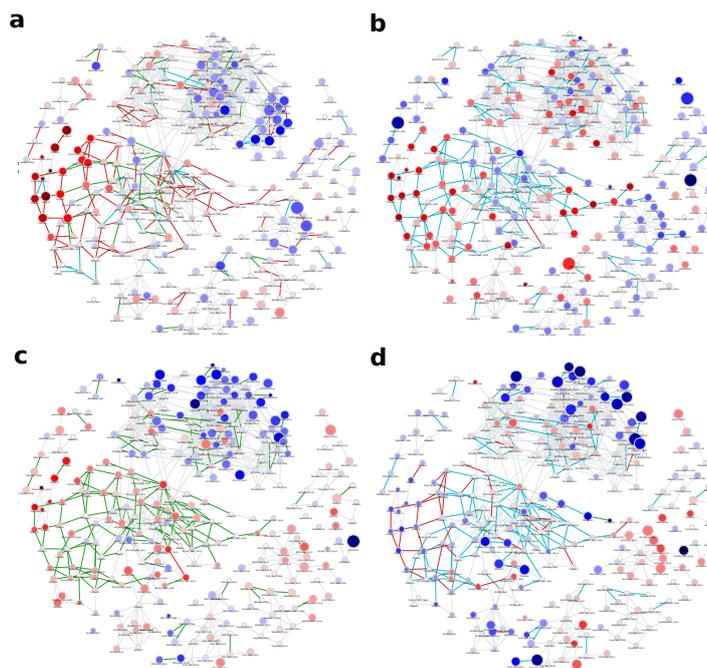


Figure 4. Global age-related plasma lipidome changes in a healthy human reference population from Kyle et al. [22]. Node colors represent log fold-changes with blue being negative, i.e., lower in the first condition, and red being positive. Node sizes are proportional to $-\log_{10}(\text{FDR})$ values. Edge colors indicate changes in correlation values between the respective conditions. For edge color groups see legend in Figure 2b. (a) Toddler vs. Child (b) Child vs. Teenager (c) Teenager vs. Adult (d) Adult vs. Elder.

Investigating the changes from toddler to child in Figure A7a, shows that most TGs, which are differentially abundant, exhibit a chain length of 44 to 48 and 0 to 3 double bonds. On the one hand, most of these lipids are connected by edges representing strong correlations in both age groups. On the other hand, connections to unchanged lipids are mostly connected via edges that are only significant in the children group and represent FA elongations. As most of the species are only identified as sum species, potential FA-specific patterns cannot be observed. However, because the described changes apply to a very limited set of total chain lengths, FA-specific elongation patterns may play a major role in changing TG levels between toddlers and children.

For the comparison of adults to elder (Figure A7d), the previously described TG species are not differentially abundant, even though they are strongly correlated with each other. However, the few species with low p -values in the subnetwork comprise longer fatty acyls (a sum of 54 to 58 hydrocarbons), are more unsaturated (6 to 11 double bonds), and are located in two separate areas of the subnetwork. These lipids are sequentially connected via edges of the same type of correlation change ("significant to insignificant"),

referring to a statistically significant correlation in younger adults between two lipids, which is not statistically significant in older adults), e.g., TG(58:9), TG(58:10) and TG(58:11).

3. Discussion

Existing bioinformatics tools for lipidomics data analysis are mainly based on the lipid class metabolism, ontologies, the chemical space or correlations. With LINEX, a new type of analysis for lipidomics is available. We combined established statistical measures as already used in other lipidomics analysis approaches such as *lipidr* [1] and functional associations between lipids. The tool BioPAN [14] offers an analysis of lipid networks and aims to find active reaction chains. LINEX takes a different approach and focuses on visualizing statistics on networks and hence revealing global trends of the lipidome and local shifts of lipids through metabolic reactions. The LUX Score [15] also visualizes global alterations of the lipidome but does not show functional associations between lipids as LINEX does.

We applied LINEX to publicly available lipidomics data and were able to reveal new insights into the regulation of lipid metabolism in addition to the originally reported ones showing the advantages of a combined lipid network analysis for the biological interpretation of lipidomics experiments. Going beyond statistical comparisons of individual lipids, but considering functional associations between lipids together with correlations and a differential analysis of sample groups, we move towards a systems biological approach for the analysis of complex lipidomes.

With its versatile visualization options, LINEX offers lipid researchers the possibility to investigate lipidome changes on a global scale while also revealing specific local associations of lipids. Furthermore, the possibility to visualize changes in (partial) correlations between lipid pairs along with reaction types allows for a more holistic view on enzymatic changes affecting lipid metabolism to develop hypotheses about biological mechanisms. The visualized networks can be downloaded and shared as fully interactive standalone files.

As with all correlation analyses, LINEX can suffer from induced spurious correlation through indirect effects. Especially in the case of unmeasured reaction partners, both correlations and partial correlations are subject to possible false-positives. Therefore, results based on these metrics should always be interpreted with caution. Beyond the issue of spurious, undetectable lipids and low coverage can limit the interpretability of LINEX results, as important connections between different parts of the network may be missing. Future work on lipid metabolic networks has to aim at reducing the impact of these effects on data interpretation and the selection of putatively interesting subnetworks.

A particular challenge is the multi-specificity of many enzymes catalyzing lipid metabolic reactions, meaning they can catalyze conversions of multiple molecular lipid species belonging to the same lipid class. Hence, lipid metabolic networks have to be generated specifically for each dataset. This makes the workflow for lipid-metabolic networks fundamentally different to working with PPI or GR networks. Dedicated algorithms such as KeyPathwayMiner [3,4], DOMINO [5] or HotNet2 [6] perform an enrichment of deregulated genes on the whole network of possible interactions. However, with lipid species networks, the networks themselves carry information about the composition of the lipidome and its associations. Therefore, a direct application of common network enrichment tools for other biological networks is not possible. With the availability of molecular reaction networks by LINEX, we enable a combined analysis of lipidomics data and provide a basis to develop algorithms specifically for lipid networks, which integrate network (topological) approaches with statistical techniques. They hold the potential to associate changes in individual lipid species with global patterns in the lipid reaction network, thereby allowing them to go beyond pathway enrichment algorithms. This lays the foundation for further improvements in the analysis of lipid metabolic networks, integrating biochemical and statistical measures. With such approaches, the discovery of

condition-specific network motifs will be possible. These motifs can then be used to define disease (sub-)types and to link conditions similar in their molecular lipid network patterns.

LINEX can be used to compare multiple conditions and switch between different network views to investigate systemic trends of lipidome changes. The versatility of LINEX allows users to create dataset-specific lipid-reaction networks, visualize and analyze the networks utilizing topological and statistical properties, as well as a standardized lipid class color scheme, and adapt the analysis to specific organisms, compartments or conditions, without requiring any programming knowledge, making it accessible not only to bioinformaticians but all lipidomics researchers. LINEX provides a novel view on the lipidome and can help to mechanistically understand remodeling of the lipidome. It can assist the community in mechanistic interpretation of lipid alterations and hypothesis generation.

4. Materials and Methods

4.1. Webtool

The LINEX web tool was implemented in python using the Django web framework. It is publicly available at <https://exbio.wzw.tum.de/linex/> (accessed on 27 July 2021). The code is available at <https://gitlab.lrz.de/lipitum-projects/linex> (accessed on 27 July 2021). Interactive network visualizations were generated using the visjs-network library along with utilities from the pyvis [23] package. To achieve simple portability to other platforms with all dependencies, LINEX is running in a Docker environment and can be deployed locally.

4.2. Lipid Name Conversion

Lipidomics data often uses different lipid naming conventions. LINEX uses Lipid Lynx [20] to convert and standardize lipid names in order to recognize them. All lipids recognized by Lipid Lynx can be used by LINEX, if lipid class information and lipid class conversions are available. If they are not available by default, they can be extended by the user.

4.3. Dynamic Network Creation

The inference of lipid metabolic networks in LINEX is implemented in a modular way by splitting transforming reactions into two broad categories: class or headgroup-related transformations and fatty acid-related (FA-related) transformations. Two given lipid species are connected in the network if they either share all their FA(s) and their headgroups are connected by a reaction, or if both lipids have the same headgroup and exactly one FA pair is transformable, according to a set of input rules. If two lipids from different classes only differ in the number of FAs, e.g., a PC and a LPC, a connection is drawn if the “larger” (PC) lipid species contains all FAs present in the “smaller” (LPC) lipid and the missing FA is in a user-defined pool of possible FAs. The decision process with pre-defined FA rules is depicted in Figure A1a. Additionally, FA reactions are evaluated (elongation, desaturation and oxidation), connecting lipids of the same class if they differ in a chain length of two, a desaturation or oxidation (on the molecular species level this is considered for individual FAs). While this type of inferred connection is based on biochemical reactions, it only represents a heuristic. All edges of this type can interactively be hidden with one click. Further details for matching between lipids of different structural resolutions with examples can be found in Appendix B.

Due to the nature of the matching procedures, it is not possible to cover many-to-many reactions such as the modification of a ceramide with a phosphocholine group from a phosphatidylcholine to a sphingomyelin and a diacylglycerol.

Default rules for both lipid class reactions and FA reactions are available. The default lipid classes and their connections are shown in Figure A1b. Because of the versatility of the implementation, user-defined customization to any desired condition and organism are possible for both sets of rules. Furthermore, it is possible to manually customize enzyme annotation for all headgroup modifying reactions.

LINEX can handle three levels of FA resolution, sum composition, molecular species and sn-specific lipid annotations, but profits from identification of all FAs, due to higher specificity of the assigned edges. In order to utilize the maximum amount of information, mixed identification levels within a dataset are allowed. When matching species on sum composition level to species of higher structural resolution, the list of allowed FAs (Table A1) is used to determine whether a FA addition is possible under the given conditions. The only requirement for using LINEX is a lipid nomenclature compatible with Lipid LynxX [20], as internal lipid mapping depends on a unified nomenclature.

4.4. Lipid Class Color Scheme

We developed a color scheme to color lipids based on their class. This scheme is available in Supplementary Data 4 and on the linex website: <https://exbio.wzw.tum.de/linex/download> (accessed on 27 July 2021). It supports colors for 46 common lipid classes. Groups of lipids have similar colors, with lyso-species being brighter and other classes darker. Colors are available as hex codes.

4.5. Statistical Methods

For analyzing changes between sample groups, multiple statistical measures are included, which can be separated into lipid species, i.e., nodes, specific and reaction, i.e., edge, specific metrics.

To compare lipid abundances, (log) fold-changes and binary statistical tests are available. End-users can choose between parametric (*t*-test) and non-parametric (Wilcoxon signed-rank test [24]) depending on their data distributions. All *p*-values are automatically reported as Benjamini-Hochberg corrected False Discovery Rates (FDR) [25]. These can be visualized as node color or size.

Additionally, three theoretical graph measures are computed for each node, namely degree, betweenness centrality [26] and closeness centrality [27]. These are, in contrast to the above metrics, independent of sample groups and visualized as node size or color.

Edge-related measures are based on correlations and partial-correlations. In order to compare two groups, (partial) correlation changes are sorted into five discrete groups, which represent whether the correlation between two lipids stayed (in-)significant, turned (in-)significant or changed its sign. In the network visualization, they are represented by the coloring of edges.

All statistical measures were computed using *scipy* [28] and *scikit-learn* [29]. For graph-related measures, the *NetworkX* [30] package was used.

LINEX does not provide data pre-processing options. Therefore, input data has to be readily processed (sample normalization, batch correction, normalization to internal standards or log-transformation). Future updates will be announced on the website: <https://exbio.wzw.tum.de/linex/> (accessed on 27 July 2021).

4.6. Experimental Data Processing

For the evaluation, publicly available lipidomics datasets were used. The data from Wang et al. [16] was reformatted and lipid names converted with Lipid LynxX [20]. No further modifications were done to the quantified measurements. Lipidomics data from Tu et al. [19] was downloaded from the MetaboLights database [31] (Study ID: MT-BLS562 and MTBLS495). Prior to uploading the data, reported as peak areas, it was quotient-normalized [32] and generalized log₂ transformed. Healthy human reference population data of the plasma lipidome was taken from Kyle et al. [22]. Unsupported lipid classes, namely Sulfatide and Carnitine, two Endocannabinoids and Co-Enzyme Q10 were removed, and LPE-P was manually added to the lipid class settings file. Three ceramide species were measured in positive and negative mode. For these, only the negative mode information was used. Lipidomics data were downloaded from the MassIVE repository at <https://doi.org/10.25345/C5P11F> (MSV000085508; accessed on 27 July 2021). Patient metadata used can be found on figshare [33]. In order to compare age-related changes,

patients were grouped into 4 groups. Toddler: 0 to 36 months; Child: 4–12 years; Teenager: 13–19 years; Adult: 20–49 years; Elderly: 50–81 (old patient).

Supplementary Materials: The following are available online at <https://www.mdpi.com/article/10.3390/metabo11080488/s1>, Supplementary Data 1: Interactive HTML of the network shown in Figure 2, Supplementary Data 2: Interactive HTML of the network shown in Figure 3, Supplementary Data 3: Interactive HTML of the network shown in Figure 4, Supplementary Data 4: Lipid Class Color Scheme.

Author Contributions: Conceptualization: N.K., T.D.R. and J.K.P.; Software: N.K., T.D.R. and L.F.; Validation: N.K. and T.D.R.; Writing—original draft: N.K., T.D.R. and J.K.P.; Writing—reviewing & editing: N.K., T.D.R. and J.K.P.; Supervision: J.K.P. All authors have read and agreed to the published version of the manuscript.

Funding: This project was funded by the Bavarian State Ministry of Science and the Arts in the framework of the Bavarian Research Institute for Digital Transformation (bidt, grant LipiTUM).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The developed software is open source. The source code is available at: <https://gitlab.lrz.de/lipitum-projects/linex> (accessed on 27 July 2021). For the analysis, publicly available lipidomics data was used (See methods section).

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

FA	fatty Acid
GPL	glycerophospholipid
GR	Gene Regulatory
LGPL	lyso-glycerophospholipid
LPC	lyso-phosphatidylcholine
LPE	lyso-phosphatidylethanolamine
LPI	lyso-phosphatidylinositol
MS	Mass Spectrometry
PA	phosphatic acid
PC	phosphatidylcholine
PE	phosphatidylethanolamine
PEO	phosphatidylethanolamine Ether
PG	phosphatidylglycerol
PI	phosphatidylinositol
PPI	Protein-Protein Interaction
PS	phosphatidylserine

Appendix A

Table A1. LINEX Default Fatty Acids. This list is used when lipids from different classes that only differ in the number of FAs are matched. Users can customize this list for their specific experimental conditions.

Saturated FAs	Monounsaturated FAs	Polyunsaturated FAs
14:0	16:1	18:2
15:0	18:1	20:2
16:0	20:1	20:3
17:0		20:4
15:0		20:5
20:0		22:4

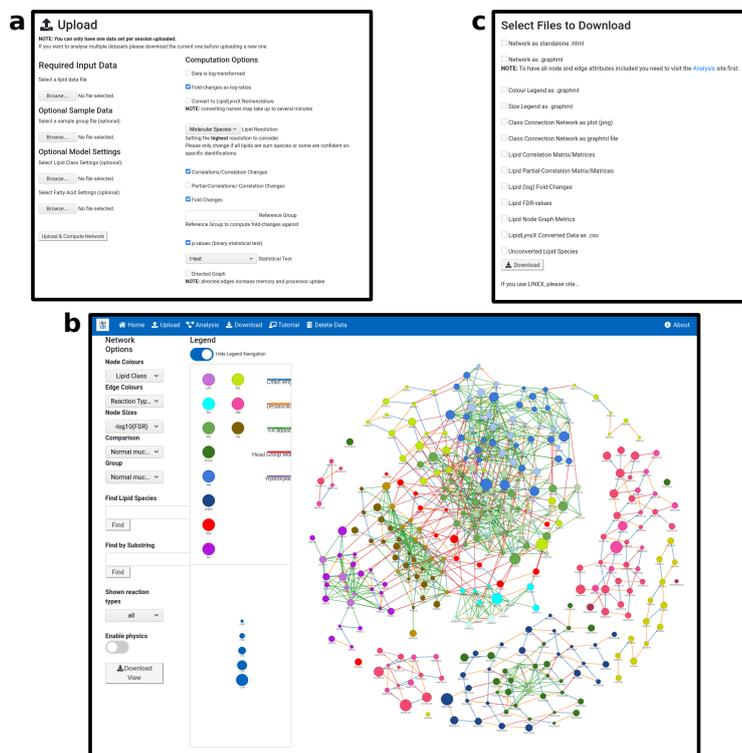


Figure A2. Main interfaces of the LINEX web-app. (a) Upload of lipidomics data with optional group labels for samples. Statistical methods can be selected for the visualization on the resulting network. Additionally, information about metabolic reactions and lipid classes can be uploaded to extend the network. (b) Analysis page. Here, the lipid networks can be interactively investigated and statistical or biochemical properties can be shown. (c) Download page. The network can be downloaded including all computed statistical measures.

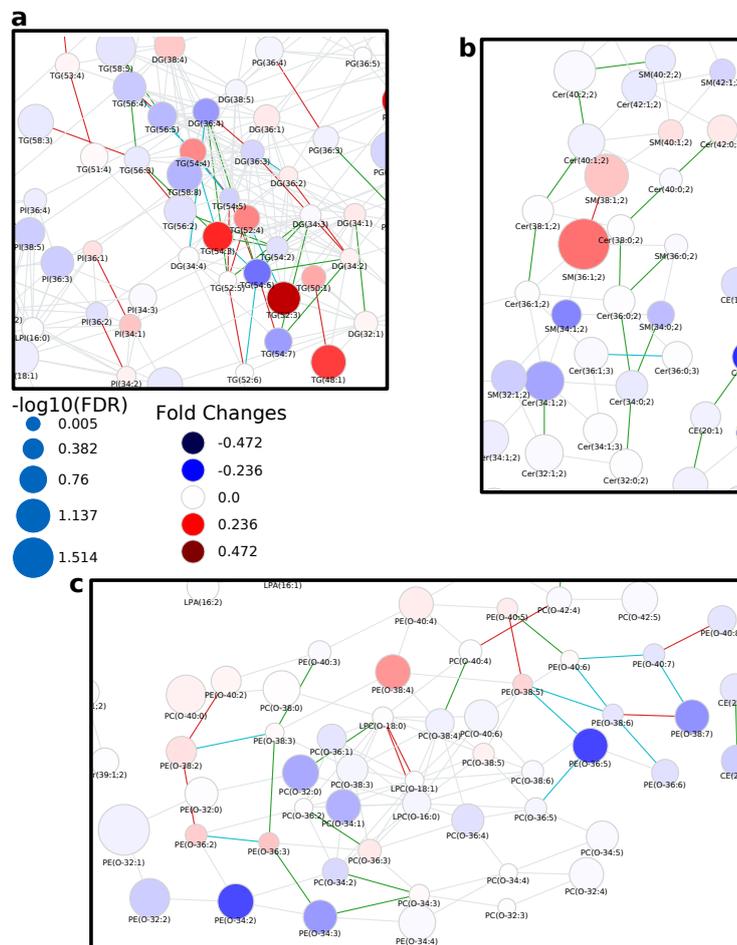


Figure A3. Detailed views on subnetworks of the lipidomics data of Wang et al. [16] showing the metabolism of (a) TG and DG, (b) ether lipids, and (c) sphingolipids. The full network can be seen in Figure 2. Nodes are colored by fold change and node size is scaled by $-\log_{10}$ of multiple testing corrected p -value. Edges are colored by correlation changes (see Figure 2).

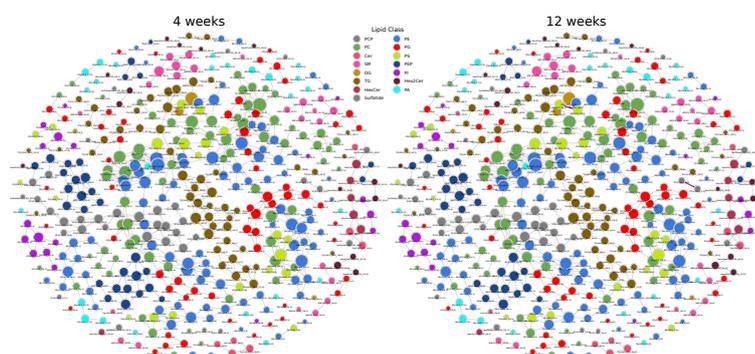


Figure A4. Lipid networks of the lipidomics data from Tu et al. [19]. Nodes are colored by lipid class and edges show correlations between lipids for each mouse age group. Significant and negative correlations are blue, significant and positive correlations red, and insignificant correlations gray. Other time points show similarly less significant correlations (not shown here).

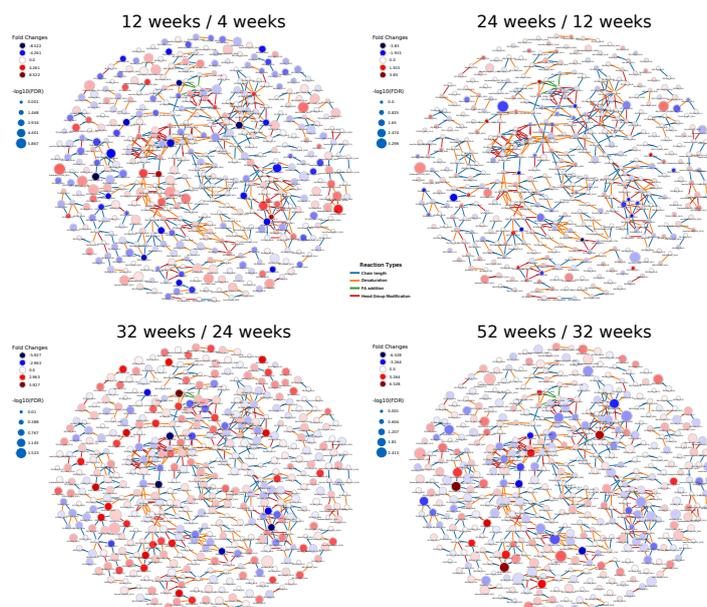


Figure A5. Fold changes of lipids visualized on lipid networks of the lipidomics data from Tu et al. [19]. Node size scaled by negative log₁₀ of the *p*-values for comparison between healthy and cancer tissue. Lipids are colored by log fold change between mouse brain age groups. Blue indicates negative fold changes and red positive fold changes (e.g., higher levels in 12 weeks compared to 4 weeks are red). Edges are colored by reaction type. Chain length modification (blue), desaturation (orange), fatty acid addition (green) and head group modification (red).

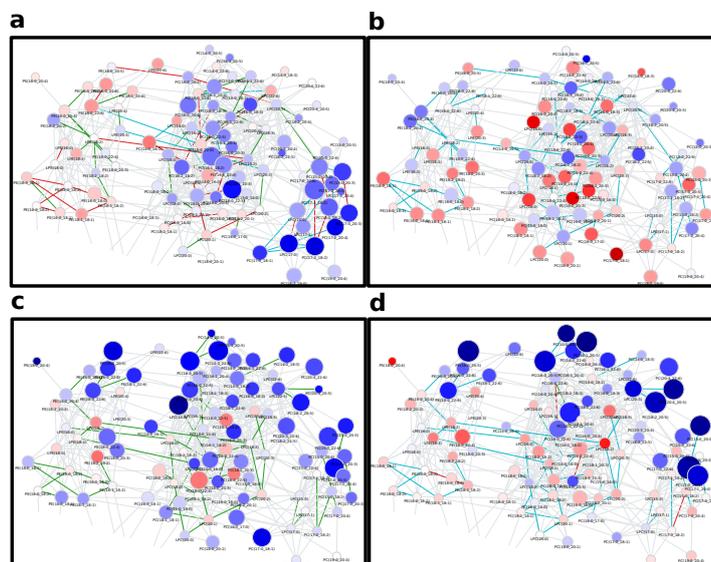


Figure A6. Detailed view on the PC/PE subnetwork from Kyle et al. [22] comparing (a) Toddler to Children, (b) Children to Teenager, (c) Teenager to Young Adults and (d) Young Adults to Older Adults.

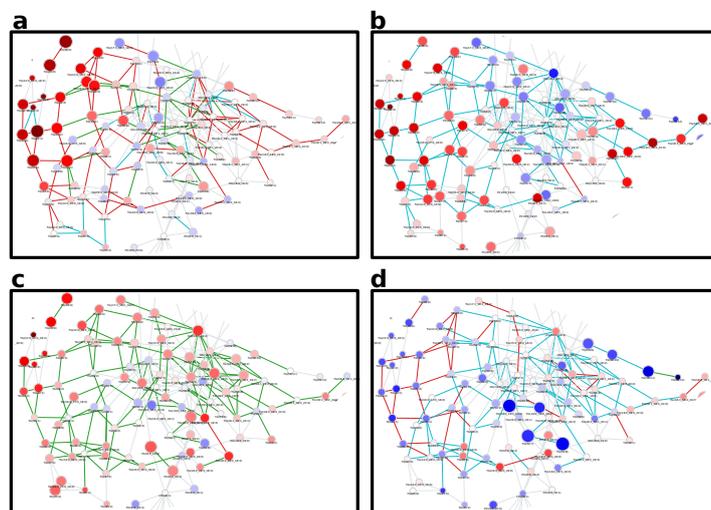


Figure A7. Neutral lipid subnetwork based on Kyle et al. [22] comparing (a) Toddler to Children, (b) Children to Teenager, (c) Teenager to Young Adults and (d) Young Adults to Older Adults.

Appendix B

In order to give a better intuition on how the rules work, we want to give three examples representing the basic types of reactions possible based on molecular species annotations.

PC(16:0_18:0)—PC(18:1_16:0): Both lipids share the same headgroup and have the same number of FAs. Therefore, the only possible reaction can be on FA level. Since 16:0 is shared in both, the remaining FAs need to be transformable. According to the default rules (Figure A1a), 18:0 → 18:1 fulfills the criteria for a desaturation, because the number of carbon atoms as well as the number of hydroxy groups stay the same, while the number of double bonds is changed by exactly one. As such fatty acid modifications are not known for esterified fatty acids, this edge represents a heuristic rather than a direct biochemical reaction. Users can remove all edges of this type in the interactive network visualization.

DG(16:0_18:0)—TG(18:1_18:0_16:0): While these lipids share the same headgroup they differ in the number of FAs. The first step in the further workflow is now to check whether the FAs in the DG, the species with fewer FAs, are both present in the putative reaction partner. As this is the case, we know that DG(16:0_18:0) and TG(18:1_18:0_16:0) are connected via the addition of an 18:1 FA. If these lipids were given as sum species, the difference between their sum compositions—34:0 and 52:1, respectively—would have been used to find the missing FA and a subsequent check of whether the resulting FA 18:1 is in the list of allowed FAs (see Table A1 for the default values) would have decided over whether the reaction is considered possible or not.

PE(16:0_18:0)—PC(16:0_18:0): The two species are composed of different headgroups; hence, the only possible reaction is a headgroup modification. For such a reaction, the lipids need to have the exact same FA composition. On sum species level, this requirement is loosened to both lipids having to have the same number of FAs and the same sum composition. Subsequently, the lipid class connection table (Figure A1b) is queried to validate whether a reaction transforming one headgroup into the other exists. Because this is the case, based on default settings, PE(16:0_18:0) and PC(16:0_18:0) are connected in the network.

References

- Mohamed, A.; Molendijk, J.; Hill, M.M. Lipidr: A Software Tool for Data Mining and Analysis of Lipidomics Datasets. *J. Proteome Res.* **2020**, *19*, 2890–2897. [[CrossRef](#)] [[PubMed](#)]
- Mohamed, A.; Hill, M.M. LipidSuite: Interactive web server for lipidomics differential and enrichment analysis. *Nucleic Acids Res.* **2021**, *49*, W346–W351. [[CrossRef](#)]
- Alcaraz, N.; Pauling, J.; Batra, R.; Barbosa, E.; Junge, A.; Christensen, A.G.L.; Azevedo, V.; Ditzel, H.J.; Baumbach, J. KeyPathwayMiner 4.0: condition-specific pathway analysis by combining multiple omics studies and networks with Cytoscape. *BMC Syst. Biol.* **2014**, *8*, 99. [[CrossRef](#)]
- Dhakar, K.; Zarecki, R.; van Bommel, D.; Knossow, N.; Medina, S.; Öztürk, B.; Aly, R.; Eizenberg, H.; Ronen, Z.; Freilich, S. Strategies for Enhancing Degradation of Linuron by sp. Strain SRS 16 Under the Guidance of Metabolic Modeling. *Front. Bioeng. Biotechnol.* **2021**, *9*, 602464. [[CrossRef](#)]
- Levi, H.; Elkon, R.; Shamir, R. DOMINO: A network-based active module identification algorithm with reduced rate of false calls. *Mol. Syst. Biol.* **2021**, *17*, e9593. [[CrossRef](#)]
- Leiserson, M.D.M.; Vandin, F.; Wu, H.T.; Dobson, J.R.; Eldridge, J.V.; Thomas, J.L.; Papoutsaki, A.; Kim, Y.; Niu, B.; McLellan, M.; et al. Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nat. Genet.* **2015**, *47*, 106–114. [[CrossRef](#)]
- Kopczynski, D.; Coman, C.; Zahedi, R.P.; Lorenz, K.; Sickmann, A.; Ahrends, R. Multi-OMICS: A critical technical perspective on integrative lipidomics approaches. *Biochim. Biophys. Acta Mol. Cell Biol. Lipids* **2017**, *1862*, 808–811. [[CrossRef](#)]
- Poupin, N.; Vinson, F.; Moreau, A.; Batut, A.; Chazalviel, M.; Colsch, B.; Fouillen, L.; Guez, S.; Khoury, S.; Dalloux-Chioccioli, J.; et al. Improving lipid mapping in Genome Scale Metabolic Networks using ontologies. *Metabolomics* **2020**, *16*, 44. [[CrossRef](#)]
- Köberlin, M.S.; Snijder, B.; Heinz, L.X.; Baumann, C.L.; Fauster, A.; Vladimer, G.I.; Gavin, A.C.; Superti-Furga, G. A Conserved Circular Network of Coregulated Lipids Modulates Innate Immune Responses. *Cell* **2015**, *162*, 170–183. [[CrossRef](#)] [[PubMed](#)]
- Yetukuri, L.; Katajamaa, M.; Medina-Gomez, G.; Seppänen-Laakso, T.; Vidal-Puig, A.; Oresic, M. Bioinformatics strategies for lipidomics analysis: characterization of obesity related hepatic steatosis. *BMC Syst. Biol.* **2007**, *1*, 12. [[CrossRef](#)] [[PubMed](#)]
- Wong, G.; Chan, J.; Kingwell, B.A.; Leckie, C.; Meikle, P.J. LICRE : unsupervised feature correlation reduction for lipidomics. *Bioinformatics* **2014**, *30*, 2832–2833. [[CrossRef](#)] [[PubMed](#)]
- Benedetti, E.; Pučić-Baković, M.; Keser, T.; Gerstner, N.; Büyükközkcan, M.; Štambuk, T.; Selman, M.H.J.; Rudan, I.; Polašek, O.; Hayward, C.; et al. A strategy to incorporate prior knowledge into correlation network cutoff selection. *Nat. Commun.* **2020**, *11*, 5153. [[CrossRef](#)]
- Molenaar, M.R.; Jeucken, A.; Wassenaar, T.A.; van de Lest, C.H.A.; Brouwers, J.F.; Helms, J.B. LION/web: A web-based ontology enrichment tool for lipidomic data analysis. *Gigascience* **2019**, *8*, giz061. [[CrossRef](#)]

14. Gaud, C.; Sousa, B.C.; Nguyen, A.; Fedorova, M.; Ni, Z.; O'Donnell, V.B.; Wakelam, M.J.O.; Andrews, S.; Lopez-Clavijo, A.F. BioPAN: A web-based tool to explore mammalian lipidome metabolic pathways on LIPID MAPS. *F1000Res* **2021**, *10*, 4. [[CrossRef](#)]
15. Marella, C.; Torda, A.E.; Schwudke, D. The LUX Score: A Metric for Lipidome Homology. *PLoS Comput. Biol.* **2015**, *11*, e1004511. [[CrossRef](#)]
16. Wang, Y.; Hinz, S.; Uckermann, O.; Hönscheid, P.; von Schönfels, W.; Burmeister, G.; Hendricks, A.; Ackerman, J.M.; Baretton, G.B.; Hampe, J.; et al. Shotgun lipidomics-based characterization of the landscape of lipid metabolism in colorectal cancer. *Biochim. Biophys. Acta Mol. Cell Biol. Lipids* **2020**, *1865*, 158579. [[CrossRef](#)] [[PubMed](#)]
17. Heinrichs, S.K.M.; Hess, T.; Becker, J.; Hamann, L.; Vashist, Y.K.; Butterbach, K.; Schmidt, T.; Alakus, H.; Krasniuk, I.; Höblinger, A.; et al. Evidence for PTGER4, PSCA, and MBOAT7 as risk genes for gastric cancer on the genome and transcriptome level. *Cancer Med.* **2018**, *7*, 5057–5065. [[CrossRef](#)]
18. Thangapandi, V.R.; Knittelfelder, O.; Brosch, M.; Patsenker, E.; Vvedenskaya, O.; Buch, S.; Hinz, S.; Hendricks, A.; Nati, M.; Herrmann, A.; et al. Loss of hepatic Mboat7 leads to liver fibrosis. *Gut* **2021**, *70*, 940–950. [[CrossRef](#)] [[PubMed](#)]
19. Tu, J.; Yin, Y.; Xu, M.; Wang, R.; Zhu, Z.J. Absolute quantitative lipidomics reveals lipidome-wide alterations in aging brain. *Metabolomics* **2017**, *14*, 5. [[CrossRef](#)]
20. Ni, Z.; Fedorova, M. LipidLynxX: lipid annotations converter for large scale lipidomics and epilipidomics datasets. *bioRxiv* **2020**. [[CrossRef](#)]
21. Balgoma, D.; Pettersson, C.; Hedeland, M. Common Fatty Markers in Diseases with Dysregulated Lipogenesis. *Trends Endocrinol. Metab.* **2019**, *30*, 283–285. [[CrossRef](#)]
22. Kyle, J.E.; Stratton, K.G.; Zink, E.M.; Kim, Y.M.; Bloodsworth, K.J.; Monroe, M.E.; Waters, K.M.; Webb-Robertson, B.J.M.; Koeller, D.M.; Metz, T.O. A resource of lipidomics and metabolomics data from individuals with undiagnosed diseases. *Sci. Data* **2021**, *8*, 114. [[CrossRef](#)]
23. Perrone, G.; Unpingco, J.; Lu, H.M. Network visualizations with Pyvis and VisJS. *arXiv* **2020**, arXiv:2006.04951.
24. Wilcoxon, F. Individual Comparisons by Ranking Methods. *Biom. Bull.* **1945**, *1*, 80. [[CrossRef](#)]
25. Benjamini, Y.; Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B (Methodol.)* **1995**, *57*, 289–300. [[CrossRef](#)]
26. Freeman, L.C. A Set of Measures of Centrality Based on Betweenness. *Sociometry* **1977**, *40*, 35. [[CrossRef](#)]
27. Bavelas, A. Communication Patterns in Task-Oriented Groups. *J. Acoust. Soc. Am.* **1950**, *22*, 725. [[CrossRef](#)]
28. Virtanen, P.; Gommers, R.; Oliphant, T.E.; Haberland, M.; Reddy, T.; Cournapeau, D.; Burovski, E.; Peterson, P.; Weckesser, W.; Bright, J.; et al. SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nat. Methods* **2020**, *17*, 261–272. [[CrossRef](#)]
29. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
30. Hagberg, A.; Schult, D.; Swart, P. Exploring Network Structure, Dynamics, and Function Using Networkx. In Proceedings of the 7th Python in Science Conference (SciPy 2008), Pasadena, CA, USA, 19–24 August 2008; pp. 11–15.
31. Haug, K.; Cochrane, K.; Nainala, V.C.; Williams, M.; Chang, J.; Jayaseelan, K.V.; O'Donovan, C. MetaboLights: A resource evolving in response to the needs of its scientific community. *Nucleic Acids Res.* **2020**, *48*, D440–D444. [[CrossRef](#)]
32. Dieterle, F.; Ross, A.; Schlotterbeck, G.; Senn, H. Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in 1H NMR metabolomics. *Anal. Chem.* **2006**, *78*, 4281–4290. [[CrossRef](#)]
33. Demographic Information for Reference Population. Available online: <https://doi.org/10.6084/m9.figshare.12440342> (accessed on 11 May 2021).

A.4. Preprint: Lipid network and moiety analysis for revealing enzymatic dysregulation and mechanistic alterations from lipidomics data

Preprint by Rose et al. [23]. Rights for publication in this dissertation have been granted by the authors.

Lipid network and moiety analysis for revealing enzymatic dysregulation and mechanistic alterations from lipidomics data

Tim D. Rose^{1,†} Nikolai Köhler^{1,†} Lisa Falk¹ Lucie Klischat¹
Olga E. Lazareva^{2,3,4,5} Josch K. Pauling^{1,*}

¹ LipiTUM, Chair of Experimental Bioinformatics, TUM School of Life Sciences, Technical University of Munich, 85354 Freising, Germany

² Chair of Experimental Bioinformatics, TUM School of Life Sciences, Technical University of Munich, 85354 Freising, Germany

³ Division of Computational Genomics and Systems Genetics, German Cancer Research Center (DKFZ), 69120 Heidelberg, Germany

⁴ Junior Clinical Cooperation Unit Multiparametric methods for early detection of prostate cancer, German Cancer Research Center (DKFZ), Heidelberg, Germany

⁵ European Molecular Biology Laboratory, Genome Biology Unit, 69117 Heidelberg, Germany

† These authors contributed equally

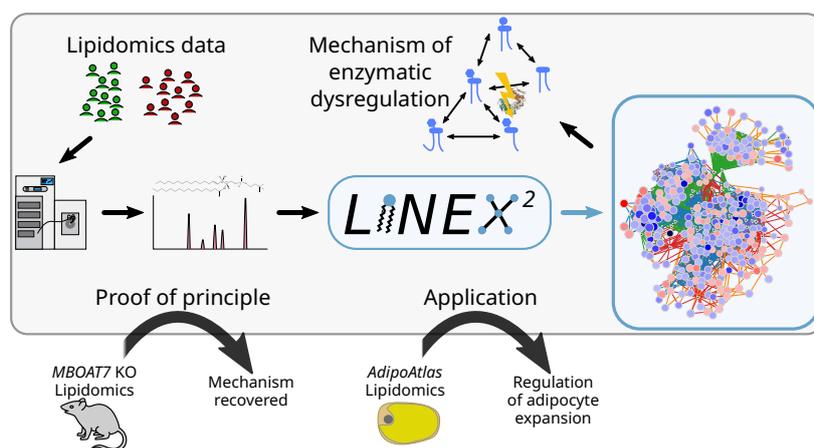
* Correspondence to josch.pauling@tum.de

Abstract

Lipidomics is of growing importance for clinical and biomedical research due to many associations between lipid metabolism and diseases. The discovery of these associations is facilitated by improved lipid identification and quantification. Sophisticated computational methods are advantageous for interpreting such large-scale data for understanding metabolic processes and their underlying (patho)mechanisms. To generate hypothesis about these mechanisms, the combination of metabolic networks and graph algorithms is a powerful option to pinpoint molecular disease drivers and their interactions. Here we present LINEX² (Lipid Network Explorer), a lipid network analysis framework that fuels biological interpretation of alterations in lipid compositions. By integrating lipid-metabolic reactions from public databases we generate dataset-specific lipid interaction networks. To aid interpretation of these networks we present an enrichment graph algorithm that infers changes in enzymatic activity in the context of their multi-specificity from lipidomics data. Our inference method successfully recovered the MBOAT7 enzyme from knock-out data. Furthermore, we mechanistically interpret lipidomic alterations of adipocytes in obesity by leveraging network enrichment and lipid moieties. We address the general lack of lipidomics data mining options to elucidate potential disease mechanisms and make lipidomics more clinically relevant.

Keywords Network Enrichment · Lipid metabolic networks · Lipidomics · Disease mechanisms

Graphical Abstract



LINEX² (Lipid Network Explorer) is a framework to visualize and analyze quantitative lipidomics data. The included algorithms offer new perspectives on the lipidome and can propose potential mechanisms of dysregulation.

- Using the Reactome and Rhea databases, a comprehensive set of lipid class reactions is included and utilized to map the lipidome on custom data-specific networks.
- With a novel network enrichment method, enzymatic dysregulation can be recovered from lipidomics data.
- We validate its usability on data with a central lipid enzymatic deficiency.
- LINEX² is the first capable of such analysis and includes complimentary analysis options for structural lipid analysis. It is freely available as a web service (<https://exbio.wzw.tum.de/linex2>).

1 Introduction

Lipids play a fundamental role in cells across all domains of life. They are not only crucial for the long-term storage of energy but can also influence the activity and occurrence of membrane proteins [1], as well as signalling and inflammatory processes [2, 3]. Therefore, diseases are also influenced by lipids. This is known not only for liver and metabolic diseases [4, 5] but also e.g. various cancers [6, 7, 8, 9]. Despite their essential role in many biological processes, excessive accumulation of lipids, especially in non-adipose tissues can lead to lipotoxicity [10, 11]. Hence, to fully understand diseases on the molecular level, changes in the lipidome have to be characterized and their regulation understood.

Nowadays, an increasing part of the lipidome can be identified and quantified using mass spectrometry (MS). The field, also known as lipidomics, is becoming more relevant for clinical applications and biomarker research [12]. While MS-based lipidomics is not yet used for diagnoses, potential biomarkers have been discussed [13, 14, 15] and disease stratifications based on lipidomics proposed [16, 17]. To gain more insights into disease mechanisms, it is necessary to go beyond classification and prediction by proposing functional interpretations of lipid changes and links to other omics layers. Due to the complexity of both acquired lipidomics data as well as the regulatory mechanisms behind lipid metabolism, dedicated computational tools are of great importance for unraveling these associations.

Such interactions can be studied through biological networks. On the metabolic level, these networks describe reactions between metabolites that are catalyzed by enzymes. When considering lipid

bioRxiv preprint doi: <https://doi.org/10.1101/2022.02.04.479101>; this version posted May 23, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Rose and Köhler et al.

Preprint version 2 - May 23, 2022

metabolic networks an additional constraint is the inherent complexity of the lipidome and its chemical reactions. Lipid enzymes commonly catalyze more than one reaction, this is referred to as multispecificity [18]. This usually means that one enzyme catalyzes a reaction for a group of lipids that e.g. belong to one lipid class but differ in their fatty acyl composition. The combinatorial complexity makes generating lipidome scale metabolic networks for an organism inefficient but instead requires data-specific networks [19, 20, 21].

Metabolic networks are commonly studied with dynamic modeling or constraint based modeling. These techniques allow predictions of the system dynamics, for example the distribution of energy resources. Parameterization of such models requires large amounts of data covering the entire molecular state [22]. Especially metabolic fluxes and well-characterized enzyme kinetics are important, which are often not available in a clinical setting.

Another way to analyze biological networks is through network enrichment. By comparing two experimental conditions, the goal is to find highly connected molecular subnetworks that are enriched with significant genes, proteins, or metabolites. The rationale behind this approach is to propose a mechanistic hypothesis for observed dysregulations. Many algorithms have been developed over the years [23, 24, 25, 26, 27], mainly with a focus on protein-protein interaction (PPI) or gene-regulatory networks. A dedicated method for metabolomics data is included in the MetExplore analysis and visualization software [28]. Their MetaboRank [29] algorithm is an enrichment and network-based fingerprint recommendation method. For lipid networks, an algorithm implemented in the BioPAN software is available, which is capable of creating lipid networks and running de-novo pathway enrichment on them [20, 21]. However, it does not consider reactions involving (de-)esterification. LINEX (Lipid Network Explorer) is a network-based method, which we previously developed [19], addressing this. It combines lipid class and fatty acid metabolism to provide comprehensive networks for computational analysis and lipidomics data interpretation. Using the LINEX framework we previously showed in several studies [19] that new insights into lipidome-wide data can be generated using lipid networks and that central alterations are often metabolically highly related. A limitation of this method is that lipid class reactions have to be entered by users. This requires detailed knowledge about lipid metabolism if reactions beyond the default are required. In contrast to de-novo enrichment on large-scale biological networks, pathway enrichment identifies significantly altered categorized pathways. For metabolites, this can be performed with the KEGG [30] or Reactome database [31]. A recent lipid-specific method is the Lipid Ontology web service (LION/web), which performs an ontology-based enrichment incorporating biological and chemical properties of lipids [32]. So far, no method is available, that puts the multispecificity of lipid enzymes into the center of interpreting lipidomic changes.

Here we present LINEX², a redesigned and extended framework, which addresses the shortcomings of lipid-network based methods. Lipid reactions are based on database information. This provides links to other omics disciplines. Furthermore, we developed a lipid-network enrichment algorithm, that incorporates multispecific enzyme links. The method enables the generation of mechanistic hypothesis from lipidomics data. We successfully applied our method to lipidomics data of a knock-out study and reveal potential dysregulations of the lipid metabolism in the adipose tissue of obese humans. This can help to better translate lipidomics into clinical application [33, 34] and improve our understanding of the role of lipid metabolism in disease mechanisms.

2 Results

2.1 A framework for lipid network analysis

The workflow of a lipidomics experiment can be divided into five steps: sampling, sample preparation, data acquisition, data processing, and data interpretation [35]. LINEX² is aiming at the biological interpretation of lipidomics data (Figure 1). The LINEX² builds data-specific lipid metabolic networks. To obtain these networks, we developed a network extension algorithm (Figure 1, purple box), where metabolic reactions on the lipid class level and fatty acid reactions are extended to the lipid species level. Network extension is possible with molecular species (e.g. DG(16:0_18:1)) or sum species data (e.g. DG(34:1)). Sum species are internally converted to molecular species, to incorporate modifications or additions/removals of fatty acids. This is achieved by finding sets of fatty acyls matching the sum composition using common fatty acids defined per lipid class (e.g. DG(16:0_18:1), DG(16:1_18:0), or DG(14:0_20:1) for DG(34:1)). If molecular species are identified

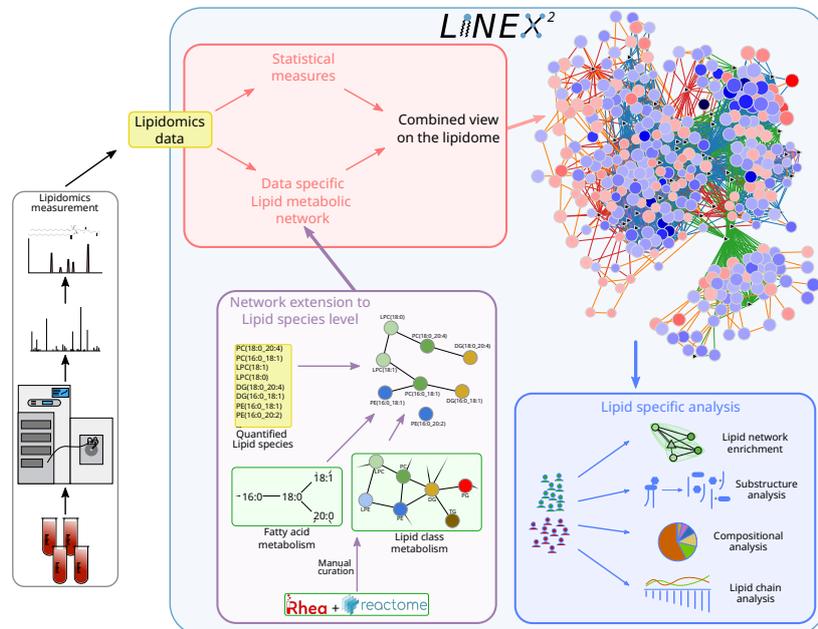


Figure 1: Lipidomics data is used as an input to LINEX². The lipids are then utilized to perform network extension that converts lipid class and fatty acid metabolic networks to lipid species, which are then visualized together with statistical measures such as t-tests or correlations. The network is also used as a basis for lipid substructure, compositional, and lipid chain analysis. A lipid network enrichment algorithm, that takes enzymatic multi specificity into account, can be used to generate hypotheses for enzymatic dysregulation.

but not quantified they can be used instead of inferring fatty acyl sets, as reported in some studies [16]. An example for the network extension is the lipid class reaction between a Phosphatidylcholine (PC) and a Diacylglycerol (DG) ($PC \rightarrow DG$), where the phosphocholine headgroup is cleaved off, is applied to the molecular lipid species $PC(16:0_18:1) \rightarrow DG(16:0_18:1)$ (for a detailed description see Materials & Methods section Network extension). Also, fatty acid reactions, such as elongation or desaturation can optionally be added to the network as heuristics, e.g. for Lyso-PC(18:0) (LPC(18:0)) \rightarrow LPC(18:1). Since such reactions usually do not occur on complex lipids directly, but rather as activated fatty acids, they help to visualize fatty acid-specific effects on the network, as previously shown [19], and facilitate computational network analysis.

2.2 Comprehensive curation of lipid-metabolic reactions

The basis for our network extension are publicly available metabolic reaction databases. To provide a comprehensive overview of lipid metabolism, we curated lipid class reactions from the Rhea [36] and Reactome [31] databases (Figure 2A). As a reference for lipid classes, we updated the lipid classes from the ALEX123 lipid database [37] (see Data Availability section). During curation, we removed all transport reactions and specialized modifications such as oxidations or fatty acid branching, which cannot be annotated to standardized lipid classes or are not generalizable for automated network extension. Curation resulted in over 3000 annotated reactions from both databases combined (Figure 2A) across organisms, including organism-specific reactions from Reactome. The top three organisms including the most reactions from Reactome are *Homo sapiens* (HSA), *Rattus norvegicus* (RNO), and *Mus musculus* (MMU) (Figure 2B).

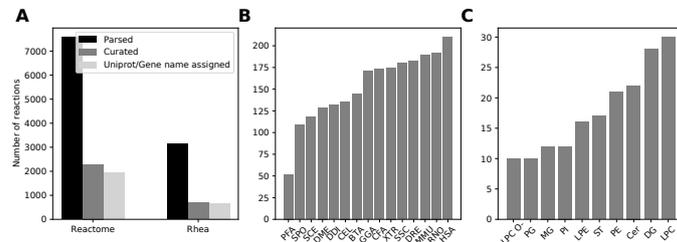


Figure 2: **A** Number of lipid-reactions parsed from Reactome and Rhea databases (black), after curation for available lipid classes and number of curated reactions (dark-grey), which Uniprot or gene name annotations were available (light-grey). **B** Curated reactions per organism from the Reactome database (Rhea does not list details about organisms). **C** Top ten lipid classes with the most curated class reactions.

In cellular lipid metabolism multiple enzymes may catalyze the same lipid class reactions but exhibit different substrate affinities based on the molecular fatty acyl composition. We made all annotated enzymes per class reaction available. After database processing, LPC is the lipid class participating in most reactions (Figure 2C), followed by DG. All reaction identifiers are individually linked, providing a reference to the original database entries in the network.

To keep the freely available LINEX² software up-to-date, user contributions for new lipid classes and lipid-metabolic reactions can be made using an online form (<https://exbio.wzw.tum.de/linex2>). This way LINEX² can be updated in a community effort to enhance support for less studied parts of the lipidome.

2.3 An approach to analyzing lipid networks

For interpreting quantitative changes in molecular networks, network enrichment can be a powerful approach. In the context of metabolic or lipid networks, such methods can reveal underlying changes in enzymatic activity. However, it is more challenging than network enrichment of e.g. proteomics data on PPI networks where changes in protein amounts correspond directly to functional changes of the nodes in the network. In PPI networks changes in protein abundances correspond directly to functional changes of the nodes, representing proteins, in the network. However, when analyzing (lipid-)metabolic networks enzymatic changes can only be approximated from changes in metabolite abundances between experimental conditions. In lipid-metabolic networks, an additional challenge comes from the multispecificity of involved enzymes. In LINEX²-networks (as implemented in the network extension) every edge between two lipid species corresponds to an enzymatic reaction, therefore enzymes can correspond to multiple edges.

Our method is designed to explicitly take multispecificity into account. Therefore, a hypernetwork, establishing connections not only between lipids but also reactions, is required. In the hypernetwork more than two nodes can be connected with one (hyper)edge. Based on this representation, the enrichment algorithm can easily connect solutions from the same class reaction, promoting solutions explainable by a few metabolic reactions. Figure 3A shows the workflow of the enrichment analysis (for details see the Materials & Methods section Network enrichment). We start with a LINEX²-network, where reactions are represented as edges (1). In the next step, we add lipid class reactions as a second type of nodes to the network (2). Edges between a class reaction node and all lipid species participating in this reaction are introduced, in addition to lipid-lipid edges, that represent conversions. This network is converted to a hypernetwork, where each hyperedge represents a lipid species reaction with lipid-substrates, -products, and reaction nodes (3). For each hyperedge (lipid species reaction), the dysregulation is quantified by the relative change of the lipid substrate-product ratio or difference between two experimental conditions (4). Considering both substrates and products is especially important for reversible reactions [39]. The reaction network is then used to find a maximally dysregulated subnetwork by employing a simulated annealing-supported local search (5). Heuristic reactions are penalized in the objective function of the network enrichment and serve only

A. Appendix

bioRxiv preprint doi: <https://doi.org/10.1101/2022.02.04.479101>; this version posted May 23, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Rose and Köhler et al.

Preprint version 2 - May 23, 2022

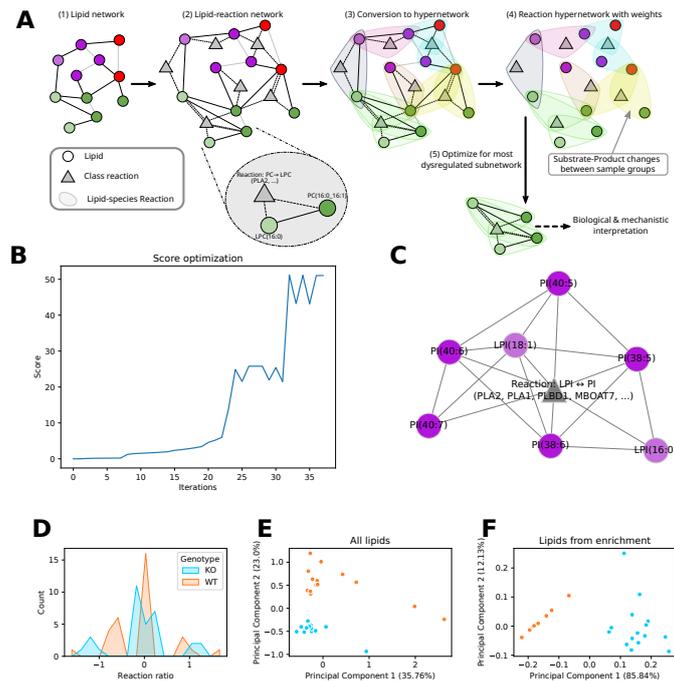


Figure 3: A Description of network enrichment workflow. In brief, the lipid network is converted into a hypernetwork, in which hyperedges correspond to lipid species reactions. Based on the computed dysregulation per hyperedge, an optimization algorithm finds the subnetwork with the maximum dysregulation **B** Optimal subnetwork predicted by the enrichment algorithm for mice liver lipidomics data by Thangapandi et al. [38]. The comparison is between wild-type and MBOAT7 knock-out samples. The resulting subnetwork shows the LPI ↔ PI reaction at the center, surrounded by polyunsaturated PI species and two LPI species. **C** Progression of the objective function score during optimization that yielded the subnetwork in **B**. **D** Substrate-product ratio distribution for the LPI ↔ PI class reaction for all lipid species reactions per genotype (MBOAT7 deficient (KO) and wild type (WT)). **E** Principal component analysis of full lipidomics data and **F** of a subset of the lipidomics data containing only the lipids from the enriched subnetwork from **B**. The color code is the same as in **D** for both plots.

to increase connectivity. Additionally, the number of class reactions in the network can be penalized to favor parsimonious solutions with a simple mechanistic explanation.

2.4 Inferring known enzymatic dysregulation from a knock-out study

As a proof of principle for the enrichment, we selected data from Thangapandi et al. [38]. In this study, the authors compared liver lipidomics of mice with a hepatospecific deficiency of MBOAT7 (KO) to wild-type (WT) mice under non-alcoholic fatty liver disease (NAFLD) condition. MBOAT7 catalyzes the class reaction fatty acyl-CoA + LPI \rightarrow PI + CoA with a specific preference for Arachidonic acid (20:4(ω -6), AA) [40]. The data from Thangapandi et al. [38] is well suited for testing our enrichment algorithm because the enzymatic origin of lipidomic changes in liver tissue is known and the lipidome is affected by the disease.

Figure 3B shows the score progression during the optimization of the algorithm. The temporary plateau at a score of 25 shows the need for global approximation methods such as simulated annealing. In Figure 3C the optimal subnetwork is shown (full network available in the supplement). It consists only of PI, LPI species, and one class reaction. This class reaction represents the transformation between LPI and PI. LINEX² cannot differentiate between the exact enzyme for this reaction. However, in contrast to e.g. PLA2, MBOAT7 only catalyzes LPI \rightarrow PI class reactions. Additionally, MBOAT7 is known for a higher affinity for AA [40]. This preference can also be observed in the solution in Figure 3C for the edge between LPI(18:1) and PI(38:5), under the assumption that this reaction can only occur if the molecular composition of PI(38:5) is PI(18:1_20:4). Furthermore, all other reactions between LPIs and PIs are only possible for the addition/removal of fatty acyls with at least 20 carbon atoms and 4 double bonds. These results are not surprising, because of the structural similarity of AA to other (very)-long-chain polyunsaturated fatty acids (Supplementary Figure S1). While LINEX² is not able to directly pinpoint MBOAT7, the results demonstrate its capability to find strong hypotheses for enzymatic dysregulation from lipidomics data.

To evaluate the enrichment results, we implemented an empirical p-value estimation procedure (detailed description in Materials & Methods section Network enrichment). This is computed by comparing the score of the final solution to a distribution of sets of unconnected reactions of the same size, to evaluate whether the final connected subnetwork has a significantly higher score. The MBOAT7 enrichment result (Figure 3C) has a p-value of 0.0018, indicating the likeliness of the mechanistic solution.

When investigating the distributions of the LPI \leftrightarrow PI class reaction (i.e. over all respective lipid species reactions) per genotype (Figure 3D), no strong distribution shift in one direction can be observed. The distributions show a peak around zero, indicating that many reaction ratios are not influenced by the MBOAT7 knock-out (KO). However, two more peaks around 1 and -1 can be observed for both conditions, where the peaks of the KO are shifted slightly more towards absolutely higher values. Despite these subtle differences, it is not possible to draw a hypothesis towards a mechanistic explanation including fatty acid-specific effects. In Figure 3E we plotted the principal component analysis (PCA) of the full lipidomics data. In contrast, Figure 3F shows the PCA plot based only on the lipidomics data for the lipid species present in the enrichment subnetwork (Figure 3C). In the PCA of all lipids, PC2 reflects the variance corresponding to the genotype, explaining 23% of the total variance. However, after selecting the LPI and PI species from the enrichment solution, the genotypic difference makes up for the majority of the variance with almost 86%. This means that the lipids in the subnetwork (Figure 3C) represent the effect of the MBOAT7 knock-out almost entirely.

These results demonstrate the ability of the enrichment analysis to develop reasonable hypotheses on enzymatic dysregulation based on lipidomics data. The result not only shows an increased variance corresponding to the genotype but also allows mechanistic lipid species-specific explanations.

2.5 A mechanistic hypothesis for adipocyte expansion in obesity

We further aimed at improving our understanding of the changes in lipid-metabolism of lipid-related diseases. For this purpose, we selected the AdipoAtlas [41], a reference lipidome of adipose tissue in lean and obese humans. The authors identified 1636 molecular lipid species, out of which 737 were quantified. Out of all semi-absolutely quantified lipid species, only 26 - solely carnitines - could not be mapped out of the box, showing that a full reference lipidome can be analyzed with the LINEX² database integration.

bioRxiv preprint doi: <https://doi.org/10.1101/2022.02.04.479101>; this version posted May 23, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Rose and Köhler et al.

Preprint version 2 - May 23, 2022

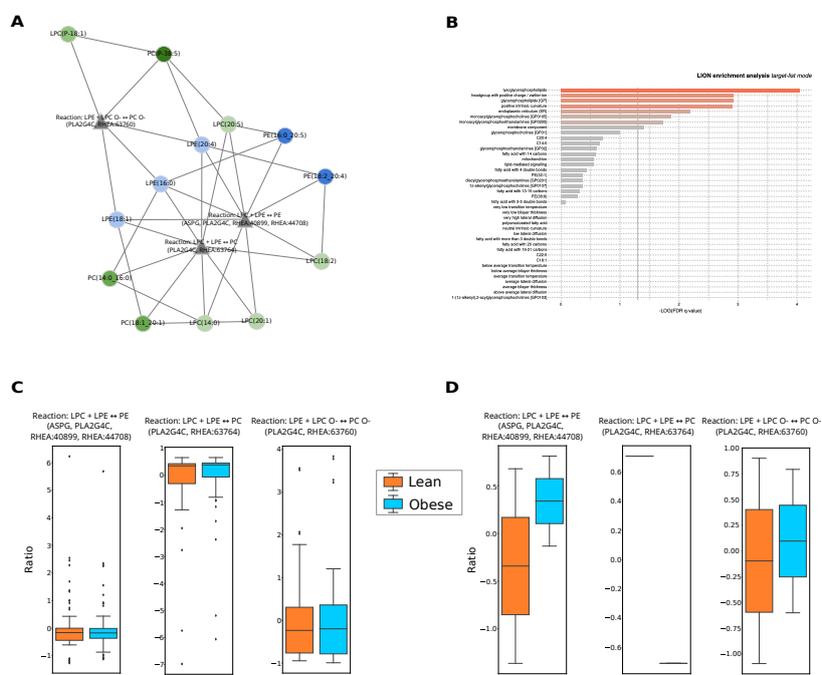


Figure 4: LINEX² application on the AdipoAtlas data. **A** Subnetwork returned by the introduced enrichment algorithm. The enriched subnetwork contains three reaction nodes, all representing fatty acid transfer between lysophospho- and phospholipids. Furthermore, the network shows a preference for long-chain polyunsaturated fatty acids. **B** LION enrichment using the lipids in the subnetwork (A) as targets in the 'target list mode'. **C** Distribution of the substrate to product changes (see Methods - Substrate-product change calculation) for the three reactions present in A over all possible lipid species combinations from the AdipoAtlas data. **D** Distribution of the substrate to product changes using only the lipid species combinations identified in A. Both C and D ratios are shown as per-reaction z-scores.

2.5.1 Network analysis indicates a mechanism for adipocyte expansion

We used our network enrichment algorithm, which resulted in the subnetwork shown in Figure 4A. The subnetwork contains three reactions, which all represent an acyl-transferase reaction between Lyso-Phospholipids. Investigating the reaction ratios of these three class reactions over all possible species reactions shows equal distributions between obese and lean (Figure 4C). However, considering the species reactions present in the subnetwork reveals differences between the groups with respect to the reaction ratios (Figure 4D). These reactions are catalyzed by the Phospholipase A2 Group IVC (PLA2G4C) and the asparaginase (ASPG), which both have lipase and acyl-transferase activity. It has been shown that PLA2 Group IV members preferably act on the sn-2 position and that polyunsaturated fatty-acyls are commonly transferred by them [42]. This preference is reflected in the subnetwork. Literature research shows that PLA2G4C has been reported to be differentially expressed in obese individuals [43, 44] and products of (c)PLA2 activity are known mediators of adipose tissue metabolism [45].

The prevalence of acyl-transferase reactions in the subnetwork suggests a transfer of FAs between lipids with a Phosphocholine and a Phosphoethanolamine headgroup and their respective Lyso-Phospholipid species. The ratio of LPC/LPE to PC/PE as well as the ratio of lipids with a Phos-

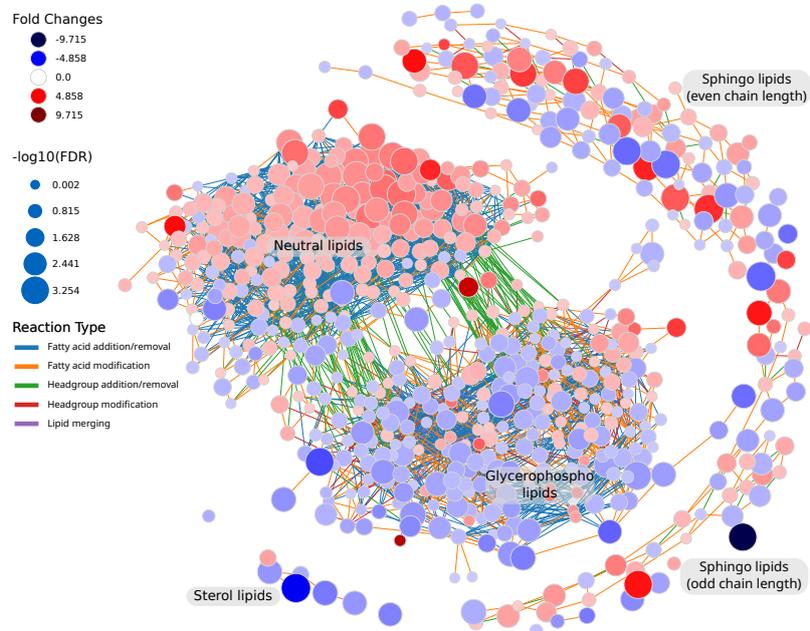


Figure 5: Lipidomics data from the AdipoAtlas visualized with LINEX. In the network lipids are represented as circular nodes. The red color of lipid nodes represents a positive fold change from lean to obese condition, and blue a negative fold change. Edge color indicates the type of reaction connecting two nodes. An interactive version of the network as well as all other analyses conducted with LINEX are available in an HTML file in the supplement.

phocholine headgroup to lipids with a Phosphoethanolamine headgroup influences the membrane curvature [46, 47]. This property is important because adipocytes expand in obesity [48]. A change in this ratio has also been associated with altered membrane integrity and fluidity [49, 50]. We confirmed this with a Lipid Ontology (LION) enrichment analysis [32], where we used the lipids of the enriched subnetworks as a target list (Figure 4B). The analysis resulted in membrane curvature and other membrane-related terms. Additionally, we observed similar behavior in the development of mesenchymal stem cells to adipogenic cells based on data from Levental et al. [51] (Supplementary Figure S2B). These insights further support the practical feasibility of our reaction enrichment approach.

2.5.2 Lipid moieties show alterations in neutral lipid composition

Despite changes in the Glycerophospholipid composition that are an indication for adipocyte expansion, accumulation of neutral storage lipids is a major hallmark for obesity. This is also reflected in the network representation of the AdipoAtlas lipidome (Figure 5). It shows increased TG and DG levels in obese samples, and an overall decrease in Glycerophospholipids. Neutral lipid species containing poly-unsaturated FAs have especially high fold changes (Figure 5, Supplementary Figure S3). Concerning chain length, we observe that TG species with a sum length >30 and <57 are accumulated in obese samples (Supplementary Figure S4A). Since this pathway of the lipid metabolism was not picked up by the network enrichment as the strongest dysregulated part, we wanted to further investigate the compositional changes of neutral lipids. For this, we developed a lipid moiety analysis. It quantifies common substructures of lipids across the lipidome to show trends in changes of the lipidome composition (Supplementary Figure S5). As moieties, we define sum length, number of

bioRxiv preprint doi: <https://doi.org/10.1101/2022.02.04.479101>; this version posted May 23, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Rose and Köhler et al.

Preprint version 2 - May 23, 2022

double bonds, head groups, and their combinations. The results go in hand with the observations on the lipid network. Especially lipid species with a sum length >45 and 2 to 3 double bonds show a sharp increase in obesity, predominantly TG species with a length of 49 and 53. Also Sterol esters show significant changes in disease progression. The observed changes in the TG composition are in accordance with previously published results [52]. This analysis can provide additional insights into the lipid metabolism and complement the network analysis.

2.6 LINEX² Software

The LINEX² software framework for analysis and visualization of lipid networks is available as a web service at <https://exbio.wzw.tum.de/linex2>. Lipidomics data and sample annotations can be uploaded as .csv files, to perform not only network enrichment and visualization, but also summarizing statistics, lipid chain analysis [53], and moiety analysis. Results can be viewed and downloaded in an interactive format. For high-throughput analysis, a python package is also available (<https://pypi.org/project/linex2/>).

3 Discussion

We present a method to generate and analyze lipid-metabolic networks. Using curated lipid class reactions from common metabolic databases our method computes data-specific lipid networks. Furthermore, we developed a network enrichment algorithm, to propose hypotheses for enzymatic dysregulation from lipidomics data. As a proof of principle, we applied the approach to liver lipidomics data, where the deficient MBOAT7 enzyme was successfully identified from the data.

Network enrichment for molecular biological data analysis has first been applied in 2002 [54] and a variety of methods have been developed since then. The challenge in generating mechanistic hypothesis from metabolomics or lipidomics data lies in the fact that dysregulation on the enzymatic level is not measured directly. Instead it can only be inferred based on changes in the metabolome, unless full-scale proteomics experiments are run in addition. For lipid networks, only one tool, BioPAN, is available so far [20]. In contrast to our proposed network enrichment algorithm, this method is searching for activated reaction chains between lipids of the same sum composition. The scope of the LINEX² enrichment differs from BioPAN, by searching for dysregulation of multispecific enzymes that likely affect lipids of the same class with different sets of fatty acyls. Another difference is in the network computation. LINEX² includes fatty acyl addition/removal, enabling insights such as the MBOAT7 example we show in this work. To illustrate how LINEX² compares to BioPAN [20], we computed the BioPAN network (Supplementary Figure S6A) as well as the predicted list of active reactions. The results do not include LPI species and only one reaction chain with a PI species (Supplementary Figure S6B). Therefore a hypothesis on MBOAT7 dysregulation cannot be drawn from this method. Nguyen et al. [21] performed a network optimization based on changes in lipid abundances and literature mining of lipid-enzyme interactions. However, they do not infer quantitative values for reactions and no implementation is available.

Hence, LINEX² lipid network enrichment is the only available method that aims at inferring enzymatic dysregulation from lipidomics data. An important aspect of the method is the usage of hypernetworks, to take the multispecificity of lipid enzymes into account, which increases confidence in the retrieved mechanism. Beyond its role in lipid metabolism multispecificity also plays a role in other biological processes, for instance in the glycan metabolism, where enzymes extend various branched glycan structures [55] or DNA methylation [56]. Our principle of network analysis and enrichment could be extended into these fields and help to discover underlying dysregulation.

A limitation of our enrichment algorithm is that it computes substrate-product ratios independent from each other. In reality, however, reactions are linked through shared substrates or products and metabolic changes are propagated through the network. These effects can be due to, e.g. metabolic self-regulation [57] and structural or signaling functions. Since each lipid species takes part in a plethora of reactions, results of altered enzymatic activity might not be observed directly for the substrates and products of that reaction. This is also the case for multiple reactions, which form a consecutive transformation sequence that change at the same time. However, assuming the principle of maximum parsimony, disordered conditions are most likely caused by alterations in only a few enzymatic steps, making the settings for such inaccurate approximations rare cases. Our network extension method depends on generalizable reaction rules. Therefore, manual curation of reaction

databases was necessary. Due to a better coverage of commonly measured lipid classes, metabolic databases may be susceptible to research bias. We address this bias by using lipid class reactions instead of enzymes, to prevent well-studied enzymes participating in many reactions from being favourably selected. Additionally, the network enrichment is avoiding bias by correcting for the number of lipid participants in the reaction.

A limitation for the generated hypothesis is the missing knowledge about enzymatic specificity. Therefore, our method is constrained to returning a set of candidate enzymes, which are attributed to the same type of reaction, without pinpointing individual enzymes. With more data available, such as the work from Hayashi et al. [42], better estimates for fatty acid-specific subnetworks can be made.

Lipids exhibit a plethora of structural and signaling functions, beyond energy metabolism. Therefore it is important to not only consider the biosynthesis of lipids, but also the change of biophysical properties. Consequently, a comprehensive computational analysis of the lipidome should include both network-based as well as lipid property-related methods. We also showed this by generating additional insights through the use of LION [32], lipid chain analysis [53], and lipid moiety analysis, which quantifies common lipid substructure features.

With the ability to connect enzymatic activity to lipidomics data, LINEX² provides the basis for a knowledge-driven integration of lipidomics with proteomics data. The inclusion of quantitative proteome information could further improve the performance of the enrichment algorithm presented in this paper and open up the possibility of directly identifying causal proteins. This could be of great value for the causal interpretation of lipidome changes, which would directly translate into relevance for clinical applications, due to the many associations of lipids with various disorders [44, 13, 8, 7, 16].

With our LINEX² web service, we offer new analysis methods for lipidomic data, ranging from network visualization to generating hypotheses for dysregulation. Freely available through a user-friendly interface, lipidomics researchers do not need to be experts in bioinformatics to perform sophisticated analyses of the lipidome in a metabolic context. Moreover, LINEX² networks can be the basis for further methodological developments that help to enhance the biological interpretability of lipidomics experiments by enabling inference of metabolic regulation from lipid data.

4 Materials & Methods

4.1 Database parsing & curation

We obtained lipid-related reactions from the Rhea [36] and Reactome [31] databases. From Rhea, all reactions involving lipids were parsed (based on ChEBI ontology, a subclass of CHEBI:18059). All reactions included in the category “Metabolism of Lipids” for all available organisms (e.g. R-HSA-556833 for *Homo sapiens*) were parsed from Reactome.

After parsing, all lipids and reactions were manually curated. Lipids were annotated and assigned to classes according to an updated version of lipid nomenclature from Pauling et al. [37] (Supplementary Table 1). Lipids that could not be annotated were not considered. Lipids are commonly composed of a headgroup, a backbone, and a set of attached fatty acids. From the databases, we extracted reactions showing conversions between common lipid classes, which are usually based on changes in one of these three attributes of lipids. We classified these lipid class reactions with at least one annotated lipid available into different categories: headgroup modification (e.g. PS ↔ PE), headgroup addition/removal (e.g. DG ↔ PA), fatty acid addition/removal (e.g. LPC ↔ PC), lipid merging (e.g. PA + PG ↔ CL) (see next section and Supplementary Figure S7 for more detailed descriptions). Fatty acid reactions on complex lipids are heuristics and can be manually added or banned by the user. Default available reactions are fatty acid elongation (increasing the chain length by 2), fatty acid desaturation (adding one double bond), and hydroxylation/oxidation (adding one hydroxylation/oxidation to a fatty acid).

4.2 Network extension to species level

Curated class reactions from databases are used to infer lipid species networks. All steps of this network extension are explained below. To properly evaluate the reactions, molecular lipid species are required. This means that for each lipid the attached fatty acid must be available. Therefore, all

bioRxiv preprint doi: <https://doi.org/10.1101/2022.02.04.479101>; this version posted May 23, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Rose and Köhler et al.

Preprint version 2 - May 23, 2022

lipid species, which are only available as sum species, are converted into a set of possible molecular species. As an example, a PC(40:2) has to be converted into possible molecular species such as PC(20:0_20:2) or PC(22:2_18:0). For this, possible common (class-specific) fatty acids can be added by the user. Only if at least one molecular species can be generated that has the same sum formula as the original sum species, it is considered for the network extension.

Extension of lipid class metabolic networks to lipid species networks can be divided into two steps: extension of the class metabolism and fatty acid metabolism.

Extension of class metabolism Lipid class reactions are evaluated using the defined reaction categories (headgroup removal/addition, headgroup modification, fatty acid addition/removal, and lipid merging) plus ether heuristic. For each reaction, all lipids from the user data, which match the lipid classes that participate in a reaction are selected. Reactions with more than one lipid class as substrate and product are only possible or available for certain reaction categories. If possible, these are explicitly mentioned. The reaction evaluations are under the condition that a lipid class reaction for the substrate-product set exists. For a “Headgroup modification” reaction the substrate and product lipids require the same set of fatty acids, e.g. PS(18:0_16:0) \leftrightarrow PE(18:0_16:0), (Supplementary Figure S7A). A “Headgroup addition/removal” also requires the substrate and product lipids to have the same set of fatty acids, e.g. DG(18:0_18:1) \leftrightarrow PA(18:0_18:1) (Supplementary Figure S7B). The reaction is also possible for two lipids as substrates and two lipids as products, e.g. PC + Cer \leftrightarrow DG + SM (Supplementary Figure S7C). In this case, the headgroup is shifted from one lipid to another. For this evaluation, two substrate product pairs are matched for the lipid donating the headgroup (PC \leftrightarrow DG) and the lipid accepting it (Cer \leftrightarrow SM). These are then evaluated independently if at least one reaction per pair can be found. The “Fatty acid addition/removal” reactions in the case of one lipid as substrate and product require one lipid with one less fatty acid and the fatty acids of the lipid with fewer fatty acids to be contained in the other lipid, e.g. DG(18:0_18:1) \leftrightarrow TG(18:0_18:1_16:0) (Supplementary Figure S7D). For two substrates and two products, a fatty acid is shifted from one lipid to another, e.g. PE + MLCL \leftrightarrow CL + LPE. Again, two substrate product pairs are matched for the lipid donating the fatty acid (PC \leftrightarrow LPC) and the lipid accepting it (MLCL \leftrightarrow CL). They are evaluated independently and the edges are added to the network if two pairs can be found which donate/accept the same fatty acid. Another case exists for reactions with two substrates and one product (e.g. LPC + LPC \leftrightarrow PC). Also here, a fatty acid is shifted from one lipid to another, however, the donor is not considered a lipid, after the fatty acid is removed. Similarly, a pair of lipids accepting the fatty acid is formed (LPC \leftrightarrow PC). Edges are then added to the network if the accepting and donating lipids have combined the same fatty acids as the resulting lipid. The reaction type “Lipid merging” describes two lipids that are bound together by a reaction, e.g. PG + PA \leftrightarrow CL (Supplementary Figure S7E). The molecular species of the substrates require the same combined fatty acids as the resulting lipid for this reaction to occur and be added to the network. We additionally consider fatty acid ether exchange as heuristics. These optional connections are edges between lipid classes and their corresponding ether classes if they share the same set of fatty acids, e.g. LPA(18:1) \leftrightarrow LPA(O-18:1), to improve network connectivity and stress fatty acid-specific effects in the network. Edges in the network are undirected since we cannot conclude the net flux of a reaction from the lipidomics data, especially since for most reactions, counterparts in the opposite direction exist.

Extension of fatty acid metabolism Fatty acid synthesis and modification occurs commonly on activated fatty acids and they are not bound to complex lipids. However, to increase network connectivity, fatty acid reactions on complex lipids can be added to the network. As described earlier, this is done through user-defined reactions. A fatty acid reaction, e.g. PC(18:0_16:0) \rightarrow PC(18:1_16:0), here fatty acid desaturation for the fatty acid 18:0 to 18:1, requires two lipids of the same lipid class and all but one identical fatty acid. Only one fatty acid modification is considered per reaction. In the case of elongation, the non-identical fatty acids require the same amount of double bonds, and other modifications have to differ by the length of two carbon atoms, e.g. 18:0 - 20:0. A desaturation requires fatty acids, which differ by a double bond, with all other attributes being the same. Custom fatty acid metabolism rules can be added, by providing the numeric changes of fatty acid attributes, such as length, double bonds, or modifications. Additionally, reactions between two specific fatty acids can be excluded. For example, the desaturation of fatty acyl 18:2 to 18:3 is not possible in humans.

In the network representation lipids are shown in the provided resolution. In the case of sum species, lipid nodes can also be shown as molecular species (based on the possible molecular species, as

explained earlier). Sum species including their statistical properties are then projected onto multiple potential molecular species.

4.3 Network enrichment

We developed a novel network enrichment algorithm for lipid networks. It aims to find the most dysregulated lipid subnetwork between two experimental conditions providing a hypothesis for enzymatic alteration/dysregulation. The methodology involves 1. building a reaction network from a standard LINEX network and calculation of substrate-product changes per reaction. 2. Utilization of a local search algorithm to find the heaviest connected subgraph (i.e. the subgraph with the largest average substrate-product change) and 3. an empirical p-value estimation. All steps are described below.

Reaction network building To convert the lipid network to a reaction network, we generate a unique reaction identifier for each reaction (edge) in the network extension. This is especially important for reactions with more than one substrate and product, with multiple edges corresponding to one lipid species reaction. In the next step, all lipid species reactions are converted to a new network representation with reactions as nodes. Edges between two reaction nodes are drawn, if the reaction belongs to the same lipid class reaction or at least one lipid species can be found in both reactions.

Substrate-product change calculation The Substrate-product change is calculated independently for each reaction r_i of the set of all reaction nodes R in the network. It describes the relative substrate to product change between two experimental conditions. It can be calculated using absolute or relative substrate-product change. The data for the calculation consists of a set of measured samples N . With C denoting the subset of samples belonging to the control condition and D the samples belonging to the disease condition. The absolute substrate product difference Γ^a for reaction r_i for of the disease samples D is calculated as

$$\Gamma_{r_i}^{a,D} = \frac{\sum_{n \in D} \left(\frac{1}{|p|} \sum_{p \in r_i} p_n - \frac{1}{|s|} \sum_{s \in r_i} s_n \right)}{|D|}$$

with $\frac{1}{|p|} \sum_{p \in r_i} p_n$ as the mean of all lipid products concentrations from reaction r_i in sample n and $\frac{1}{|s|} \sum_{s \in r_i} s_n$ corresponding for substrates. We choose the mean over the sum here to avoid a bias towards reactions with unequal numbers of lipid products or substrates. The relative substrate-product difference Γ^r is calculated by

$$\Gamma_{r_i}^{r,D} = \frac{\sum_{n \in D} \left(\left(\prod_{p \in r_i} p_n \right)^{\frac{1}{|p|}} / \left(\prod_{s \in r_i} s_n \right)^{\frac{1}{|s|}} \right)}{|D|}$$

The $|p|$ th- and $|s|$ th-root are used as bias correction factors in an analogous fashion to using the mean over the sum in the absolute difference calculation. From the user, the absolute or relative score can be used to compute the final reaction score that compares both experimental sets C and D . It is calculated as follows:

$$\text{Score}(r_i) = \frac{|\Gamma_{r_i}^D - \Gamma_{r_i}^C|}{\Gamma_{r_i}^C}$$

This can be done with the relative or absolute substrate-product change. As previously explained, reactions of the fatty acid metabolism or other lipid conversions are heuristic, to improve network connectivity. They do not occur directly on the lipid level. For that reason, They are also considered in the network enrichment but penalized (default = -1) to favor the selection of the non-heuristic reactions.

Local search and simulated annealing Local search is a heuristic approach that is usually applied to hard optimization problems [58]. Local search investigates the search space by applying local changes to candidate solutions, such that the objective function value is increasing. The changes are applied until no more local improvements can be made. To avoid stagnation in a local maximum, the simulated annealing procedure [59] allows non-optimal solutions and thus increases the exploration

space. The probability of accepting a suboptimal solution depends on the temperature parameter T , which decreases over time at rate α :

$$T = T_0 \cdot \alpha^n$$

where T_0 is the initial temperature, α is the rate of decrease and n is the iteration number. If no more local improvements are possible, a random solution is accepted under the following condition:

$$e^{\frac{o_{n-1} - o_n}{-T}} > \text{uniform}(0, 1)$$

where o_{n-1} and o_n are objective function scores at iterations $n-1$ and n correspondingly.

We employ local search on the reaction network $G = (V, E)$. Starting from a (random) set of connected starting nodes, also called seed, the local search can perform three actions for improvement in the objective function scores: node addition, node deletion, and node substitution. A minimum and maximum size for the subnetwork have to be entered as parameters, preventing the algorithm from selecting too small or big solutions. The action that allows improving the current value of the objective function is accepted, and thus a candidate solution is modified at each iteration. The algorithm terminates when a) no further improvements are possible, b) the simulated annealing condition is not satisfied, or c) the number of maximum iterations is reached. The best-identified subnetwork is returned. The objective function score of a reaction subnetwork $G^* = (V^*, E^*)$ is computed as follows:

$$o = \frac{\sum_{v_i \in V^*} \text{Score}(v_i)}{|V^*| \times (p \times |\text{CR}(V^*)|)}$$

with a user defined penalty p for the number of different lipid class reactions in the subnetwork and $\text{CR}(V^*)$, the set of different lipid class reactions in the set nodes V^* . If the reaction network consists of unconnected components, the local search is run for each component independently and a subgraph for each component is returned.

Subnetwork p-value The network enrichment algorithm results in a subnetwork with a score for each run. To indicate if this subnetwork/score provides a significant insight compared to an equally sized random set of reactions, we compute an empirical p-value. For that, we sample reactions in the range of the minimum and maximum subnetwork size. These reactions are not connected, as in the subnetwork of the enrichment. This creates a distribution of scores. The distribution is then used to estimate a p-value for the solution found by the enrichment. The number of samples can be decided by the user, with more samples giving a better estimate of the distribution at increased runtime. The rationale behind sampling unconnected solutions is to estimate how much the connected (mechanistic) subnetwork scores compared to unconnected (non-mechanistic) solutions.

In the implementation for the LINEX² web service, the local search is run multiple times (default=5), each with random seeds. The best result can then be optionally used as a seed for another local search run, to improve this result. The best score achieved in all local search runs is then returned to the user.

4.4 Reaction ratio plots

Visualizations of reaction ratios were performed for each lipid class reaction individually. Reaction ratios per sample are computed in the same as for the substrate-product change calculation, without averaging over all individuals:

$$\frac{\left(\prod_{p \in r_i} p_n\right)^{\frac{1}{|p|}}}{\left(\prod_{s \in r_i} s_n\right)^{\frac{1}{|s|}}}$$

All ratios per experimental condition are compiled into a list and the density for each considered experimental condition is plotted.

4.5 Lipid moiety analysis

The (combined) abundance of lipid features was implemented inspired by the glycan substructure method by Bao et al. [55]. We used the same vectorization and weighting as the authors, but with lipid substructures as features. These were: headgroup, backbone, independent fatty acyls, sum length of fatty acyls, sum double bonds of fatty acyls, and fatty acyl hydroxylations. The features were

bioRxiv preprint doi: <https://doi.org/10.1101/2022.02.04.479101>; this version posted May 23, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Rose and Köhler et al.

Preprint version 2 - May 23, 2022

weighted independently or in combination of pairs by occurrence in each lipid per sample. To find the most discriminative feature combinations, we train a regression model with sample groups as target variables and extract its coefficients. A summary of the workflow can be found in Supplementary Figure S5.

4.6 Lipid chain analysis

We implemented lipid chain analysis in python according to the proposed method by Mohamed, Molendijk, and Hill [53]. For each lipid class, lipid species with the same sum length of fatty acids are summed up per sample and a mean over all samples of one experimental condition is calculated. After that, the fold change between a selected control and e.g. a disease condition is calculated for each sum length per lipid class. The result is then plotted with an ascending fatty acid length on the x-axis, showing class-specific fatty acid length fold changes between conditions.

4.7 Webtool and data upload

The web service is built with the Django web framework (<https://www.djangoproject.com/>) in the python programming language (version 3.8, <https://www.python.org/>). PostgreSQL (<https://www.postgresql.org/>) is used as a back-end database for Django, to store data, networks, and all computed attributes. Cookies are used to connect a browser session to uploaded data, their corresponding computed networks, and analyses. For interactive network visualizations, vis-network [60] is used and other interactive plots are done with Plotly [61]. All other user-site functionalities are implemented in plain JavaScript. PDF versions of networks are generated with the NetworkX package [62] in conjunction with the matplotlib library [63]. The backend was implemented in python. To achieve compatibility across operating systems, LINEX² can be built in a Docker environment. In the public LINEX² version, uploaded user data is temporarily stored on our server for a certain time or can be deleted manually by the user (for further information see <https://exbio.wzw.tum.de/linex/request-data-delete>). However, using the provided Dockerfiles LINEX² can also be easily run locally on any computer (for instructions check the source code repository). LINEX² is free software, published under the aGPLv3 license. The source code is available at <https://gitlab.lrz.de/lipitum-projects/linex>. While we adapted the procedure to generate lipid species networks, the original LINEX version can still be accessed through the website (marked as version 1).

Identified and quantified lipidomics data with optional sample labels can be uploaded to LINEX². Lipidomics data must be uploaded as a table with samples, lipids, and their corresponding concentrations/amounts. To convert lipids into our internal programming model, we recommend the LIPID MAPS nomenclature [64]. However, we integrated the LipidLynx [65] software, which can convert multiple lipid nomenclatures, increasing the compatibility of LINEX² with multiple formats. A tutorial is available on the website (<https://exbio.wzw.tum.de/linex/tutorial>).

4.8 Statistical measures implemented in LINEX²

To enable a combined visualization of the biochemical connections between lipid species and quantitative lipidomics measurements, LINEX² offers the possibility to project different statistical metrics onto the species networks.

To characterize the changes in lipid levels between different experimental conditions, we provide unpaired parametric (t-test), non-parametric (Wilcoxon rank-sum test) test options, and a paired parametric test (Wilcoxon signed-rank test). All resulting p-values are automatically false-discovery rate (FDR) corrected using the Benjamini-Hochberg procedure [66]. Furthermore, fold changes are computed to showcase effect size. For the computation of these metrics, we used the scipy package [66, 67] in conjunction with the statsmodels package [68]. All measures can be visualized either as node sizes or node colors.

Measures for lipid connections (i.e. edges in the network) are correlation-based. Specifically, the options provided are spearman's correlation and partial correlation. All correlations above a user-specified significance threshold (default = 0.05) are set to 0 automatically. Correlation values can be visualized in the network representation through edge colors.

bioRxiv preprint doi: <https://doi.org/10.1101/2022.02.04.479101>; this version posted May 23, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Rose and Köhler et al.

Preprint version 2 - May 23, 2022

4.9 Analyzed data sets

The lipidomics data for MOBAT7 WT and knockout mice were taken from Thangapandi et al. [38]. No further processing was done and the data was analyzed as provided by the authors. Data for the Adipo Atlas was used as provided in the supplement of Lange et al. [41]. The comparison of MCS to adipogenic cells is coming from the supplement of Levental et al. [51]. Lipid species measured in less than 50% of all samples were removed before analysis with LINEX². For all data sets analyzed with LINEX², HTML files with the LINEX² output are available in the supplement.

Data Availability Statement

LINEX is free software. Source code: GitLab (aGPLv3 License): <https://gitlab.lrz.de/lipitum-projects/linex>

Figure reproducibility: <https://gitlab.lrz.de/lipitum-projects/LINEX2-paper-code>

ALEX123 lipid classes and curated database reaction: https://gitlab.lrz.de/lipitum-projects/LINEX2_package/-/tree/master/LINEX2/data

Author Contributions

JKP supervised the project and secured the funding. NK, TDR, and JKP planned and conceptualized the work. NK and TDR developed the web service. NK, OEL, and TDR designed and implemented the network enrichment procedure. LF, LK, and TDR parsed and curated the reaction databases, and implemented the network extension. NK and TDR applied, validated, and interpreted the approach on lipidomics data. NK, OEL, TDR, and JKP wrote the manuscript. All authors read, reviewed, and accepted the manuscript in its final form.

Acknowledgments

This project was funded by the Bavarian State Ministry of Science and the Arts in the framework of the Bavarian Research Institute for Digital Transformation (bidit; JKP, NK, TDR: Junior Research Group LipiTUM; O.L.: Doctoral Fellow).

Conflicts of interest

The authors declare no conflicts of interest.

References

- [1] John A Allen, Robyn A Halverson-Tamboli, and Mark M Rasenick. "Lipid raft microdomains and neurotransmitter signalling". en. In: *Nat. Rev. Neurosci.* 8.2 (Feb. 2007), pp. 128–140.
- [2] Charles N Serhan, Nan Chiang, and Thomas E Van Dyke. "Resolving inflammation: dual anti-inflammatory and pro-resolution lipid mediators". en. In: *Nat. Rev. Immunol.* 8.5 (May 2008), pp. 349–361.
- [3] Valerio Chiurchiù, Alessandro Leuti, and Mauro Maccarrone. "Bioactive Lipids and Chronic Inflammation: Managing the Fire Within". en. In: *Front. Immunol.* 9 (Jan. 2018), p. 38.
- [4] Stella Bernardi et al. "The Complex Interplay between Lipids, Immune System and Interleukins in Cardio-Metabolic Diseases". en. In: *Int. J. Mol. Sci.* 19.12 (Dec. 2018).
- [5] Chih-Hao Lee, Peter Olson, and Ronald M Evans. "Minireview: lipid metabolism, metabolic diseases, and peroxisome proliferator-activated receptors". en. In: *Endocrinology* 144.6 (June 2003), pp. 2201–2207.
- [6] Claudio R Santos and Almut Schulze. "Lipid metabolism in cancer". en. In: *FEBS J.* 279.15 (Aug. 2012), pp. 2610–2623.
- [7] Janel Suburu and Vong Q Chen. "Lipids and prostate cancer". In: *Prostaglandins Other Lipid Mediat.* 98.1-2 (May 2012), pp. 1–10.

bioRxiv preprint doi: <https://doi.org/10.1101/2022.02.04.479101>; this version posted May 23, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Rose and Köhler et al.

Preprint version 2 - May 23, 2022

- [8] Jingting Jiang, Peter Nilsson-Ehle, and Ning Xu. "Influence of liver cancer on lipid and lipoprotein metabolism". en. In: *Lipids Health Dis.* 5.1 (Mar. 2006), pp. 1–7.
- [9] Alicja Pakiet et al. "Changes in lipids composition and metabolism in colorectal cancer: a review". en. In: *Lipids Health Dis.* 18.1 (Jan. 2019), pp. 1–21.
- [10] Jean E Schaffer. "Lipotoxicity: when tissues overeat". In: *Curr. Opin. Lipidol.* 14.3 (June 2003), p. 281.
- [11] J M Weinberg. "Lipotoxicity". In: *Kidney Int.* 70.9 (Nov. 2006), pp. 1560–1566.
- [12] Markus R Wenk. "The emerging field of lipidomics". en. In: *Nat. Rev. Drug Discov.* 4.7 (July 2005), pp. 594–610.
- [13] Ding Liu et al. "Identification of lipid biomarker from serum in patients with chronic obstructive pulmonary disease". en. In: *Respir. Res.* 21.1 (Sept. 2020), p. 242.
- [14] Furong Yan, Hong Zhao, and Yiming Zeng. "Lipidomics: a promising cancer biomarker". en. In: *Clin. Transl. Med.* 7.1 (July 2018), p. 21.
- [15] Francesca Perrotti et al. "Advances in Lipidomics for Cancer Biomarkers Discovery". en. In: *Int. J. Mol. Sci.* 17.12 (Nov. 2016).
- [16] Olga Vvedenskaya et al. "Nonalcoholic fatty liver disease stratification by liver lipidomics". en. In: *J. Lipid Res.* 62 (Aug. 2021), p. 100104.
- [17] Adam Stefanko et al. "Lipidomic approach for stratification of acute myeloid leukemia patients". en. In: *PLoS One* 12.2 (Feb. 2017), e0168781.
- [18] S Gatt and Y Barenholz. "Enzymes of Complex Lipid Metabolism". en. In: *Annual Review of Biochemistry* 42.1 (1973), pp. 61–90.
- [19] Nikolai Köhler et al. "Investigating Global Lipidome Alterations with the Lipid Network Explorer". en. In: *Metabolites* 11.8 (July 2021).
- [20] Caroline Gaud et al. "BioPAN: a web-based tool to explore mammalian lipidome metabolic pathways on LIPID MAPS". en. In: *F1000Res.* 10 (Jan. 2021), p. 4.
- [21] An Nguyen et al. "Host lipidome analysis during rhinovirus replication in HBECs identifies potential therapeutic targets". In: *Journal of Lipid Research* 59.9 (2018), pp. 1671–1684. ISSN: 0022-2275. DOI: <https://doi.org/10.1194/jlr.M085910>. URL: <https://www.sciencedirect.com/science/article/pii/S0022227520335367>.
- [22] Amit Rai and Kazuki Saito. *Omics data input for metabolic modeling*. 2016.
- [23] Nicolas Alcaraz et al. "Efficient key pathway mining: combining networks and OMICS data". en. In: *Integr. Biol.* 4.7 (July 2012), pp. 756–764.
- [24] Hagai Levi, Ran Elkon, and Ron Shamir. *DOMINO: a network-based active module identification algorithm with reduced rate of false calls*. 2021.
- [25] Zijian Ding, Wenbo Guo, and Jin Gu. *ClustEx2: Gene Module Identification using Density-Based Network Hierarchical Clustering*. 2018.
- [26] Haisu Ma et al. "COSINE: COndition-SpecIfic sub-NEtwork identification using a global optimization method". en. In: *Bioinformatics* 27.9 (May 2011), pp. 1290–1298.
- [27] Susan Dina Ghiassian, Jörg Menche, and Albert-László Barabási. "A DIseAse MOdule De-tection (DIAMOnD) algorithm derived from a systematic analysis of connectivity patterns of disease proteins in the human interactome". en. In: *PLoS Comput. Biol.* 11.4 (Apr. 2015), e1004120.
- [28] Ludovic Cottret et al. "MetExplore: collaborative edition and exploration of metabolic networks". en. In: *Nucleic Acids Res.* 46.W1 (July 2018), W495–W502.
- [29] Clément Frainay et al. "MetaboRank: network-based recommendation system to interpret and enrich metabolomics results". en. In: *Bioinformatics* 35.2 (Jan. 2019), pp. 274–283.
- [30] Minoru Kanehisa and Susumu Goto. "KEGG: Kyoto Encyclopedia of Genes and Genomes". en. In: *Nucleic Acids Res.* 28.1 (Jan. 2000), pp. 27–30.
- [31] Bijay Jassal et al. "The reactome pathway knowledgebase". en. In: *Nucleic Acids Res.* 48.D1 (Jan. 2020), pp. D498–D503.
- [32] Martijn R Molenaar et al. "LION/web: a web-based ontology enrichment tool for lipidomic data analysis". en. In: *Gigascience* 8.6 (June 2019).
- [33] Jiawei Lv et al. "Clinical lipidomics: a new way to diagnose human diseases". en. In: *Clin. Transl. Med.* 7.1 (Apr. 2018), p. 12.

A. Appendix

bioRxiv preprint doi: <https://doi.org/10.1101/2022.02.04.479101>; this version posted May 23, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Rose and Köhler et al.

Preprint version 2 - May 23, 2022

- [34] Linlin Zhang, Xianlin Han, and Xiangdong Wang. “Is the clinical lipidomics a potential goldmine?” en. In: *Cell Biol. Toxicol.* 34.6 (Dec. 2018), pp. 421–423.
- [35] Thomas Züllig, Martin Trötz Müller, and Harald C Köfeler. “Lipidomics from sample preparation to data analysis: a primer”. en. In: *Anal. Bioanal. Chem.* 412.10 (Apr. 2020), pp. 2191–2209.
- [36] Thierry Lombardot et al. “Updates in Rhea: SPARQLing biochemical reaction data”. en. In: *Nucleic Acids Res.* 47.D1 (Jan. 2019), pp. D596–D600.
- [37] Josch K Pauling et al. “Proposal for a common nomenclature for fragment ions in mass spectra of lipids”. In: *PLoS One* 12.11 (Nov. 2017), e0188394.
- [38] Veera Raghavan Thangapandi et al. “Loss of hepatic Mboat7 leads to liver fibrosis”. en. In: *Gut* 70.5 (May 2021), pp. 940–950.
- [39] Wolfram Liebermeister and Edda Klipp. “Bringing metabolic networks to life: convenience rate law and thermodynamic constraints”. en. In: *Theor. Biol. Med. Model.* 3 (Dec. 2006), p. 41.
- [40] Miguel A Gijón et al. “Lysophospholipid acyltransferases and arachidonate recycling in human neutrophils”. en. In: *J. Biol. Chem.* 283.44 (Oct. 2008), pp. 30235–30245.
- [41] Mike Lange et al. “AdipoAtlas: A reference lipidome for human white adipose tissue”. en. In: *Cell Rep Med* 2.10 (Oct. 2021), p. 100407.
- [42] Daiki Hayashi, Varnavas D Mouchlis, and Edward A Dennis. “Omega-3 Versus Omega-6 Fatty Acid Availability is Controlled by Hydrophobic Site Geometries of Phospholipase A₂”. en. In: *J. Lipid Res.* (Aug. 2021), p. 100113.
- [43] Nicholas J Carruthers et al. *The human type 2 diabetes-specific visceral adipose tissue proteome and transcriptome in obesity*. 2021.
- [44] Laura Jackisch et al. “Differential expression of Lp-PLA₂ in obesity and type 2 diabetes and the influence of lipids”. en. In: *Diabetologia* 61.5 (May 2018), pp. 1155–1166.
- [45] Marcia J Abbott, Tianyi Tang, and Hei Sook Sul. *The role of phospholipase A₂-derived mediators in obesity*. 2010.
- [46] Leonid Chernomordik. “Non-bilayer lipids and biological fusion intermediates”. In: *Chemistry and physics of lipids* 81.2 (1996), pp. 203–213.
- [47] N Fuller and RP Rand. “The influence of lysolipids on the spontaneous curvature and bending elasticity of phospholipid membranes”. In: *Biophysical journal* 81.1 (2001), pp. 243–254.
- [48] Kirsty L Spalding et al. “Dynamics of fat cell turnover in humans”. en. In: *Nature* 453.7196 (June 2008), pp. 783–787.
- [49] Zhaoyu Li et al. “The ratio of phosphatidylcholine to phosphatidylethanolamine influences membrane integrity and steatohepatitis”. In: *Cell metabolism* 3.5 (2006), pp. 321–331.
- [50] Rosie Dawaliby et al. “Phosphatidylethanolamine is a key regulator of membrane fluidity in eukaryotic cells”. In: *Journal of Biological Chemistry* 291.7 (2016), pp. 3658–3667.
- [51] Kandice R Levental et al. “ ω -3 polyunsaturated fatty acids direct differentiation of the membrane phenotype in mesenchymal stem cells to potentiate osteogenesis”. en. In: *Sci Adv* 3.11 (Nov. 2017), eaao1193.
- [52] Chong Yew Tan et al. “Adipose tissue fatty acid chain length and mono-unsaturation increases with obesity and insulin resistance”. en. In: *Sci. Rep.* 5 (Dec. 2015), p. 18366.
- [53] Ahmed Mohamed, Jeffrey Molendijk, and Michelle M Hill. “lipidr: A Software Tool for Data Mining and Analysis of Lipidomics Datasets”. en. In: *J. Proteome Res.* 19.7 (July 2020), pp. 2890–2897.
- [54] Trey Ideker et al. “Discovering regulatory and signalling circuits in molecular interaction networks”. en. In: *Bioinformatics* 18 Suppl 1 (2002), S233–40.
- [55] Bokan Bao et al. “Correcting for sparsity and interdependence in glycomics by accounting for glycan biosynthesis”. en. In: *Nat. Commun.* 12.1 (Aug. 2021), p. 4988.
- [56] J Walter, T A Trautner, and M Noyer-Weidner. “High plasticity of multispecific DNA methyltransferases in the region carrying DNA target recognizing enzyme modules”. en. In: *EMBO J.* 11.12 (Dec. 1992), pp. 4445–4450.
- [57] SR Hackett et al. “Systems-level analysis of mechanisms regulating yeast metabolic flux”. In: *Science* 354.6311 (2016), aaf2786. DOI: 10.1126/science.aaf2786. eprint: <https://www.science.org/doi/pdf/10.1126/science.aaf2786>. URL: <https://www.science.org/doi/abs/10.1126/science.aaf2786>.

A. Appendix

bioRxiv preprint doi: <https://doi.org/10.1101/2022.02.04.479101>; this version posted May 23, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Rose and Köhler et al.

Preprint version 2 - May 23, 2022

- [58] Emile Aarts and Jan Karel Lenstra. *Local Search in Combinatorial Optimization*. en. Princeton University Press, June 2018.
- [59] P J van Laarhoven and E H Aarts. *Simulated Annealing: Theory and Applications*. en. Springer Science & Business Media, Mar. 2013.
- [60] Giancarlo Perrone, Jose Unpingco, and Haw-Minn Lu. "Network visualizations with Pyvis and VisJS". In: *arXiv preprint arXiv:1802.03426* (June 2020). arXiv: 2006.04951 [cs.SI].
- [61] Plotly Technologies Inc. "Collaborative data science". In: *Montréal, QC* (2015).
- [62] Aric A Hagberg, Daniel A Schult, and Pieter J Swart. "Exploring Network Structure, Dynamics, and Function Using Networkx". en. In: *Proceedings of the 7th Python in Science Conference (SciPy2008)*. Pasadena, CA USA, 2008, pp. 11–15.
- [63] John D Hunter. *Matplotlib: A 2D Graphics Environment*. 2007.
- [64] Eoin Fahy et al. "LIPID MAPS online tools for lipid research". en. In: *Nucleic Acids Res.* 35.Web Server issue (July 2007), W606–12.
- [65] Zhixu Ni and Maria Fedorova. "LipidLynxX: a data transfer hub to support integration of large scale lipidomics datasets". In: *bioRxiv* 2020.04.09.033894 (2020).
- [66] Yoav Benjamini and Yosef Hochberg. "Controlling the false discovery rate: A practical and powerful approach to multiple testing". en. In: *J. R. Stat. Soc.* 57.1 (Jan. 1995), pp. 289–300.
- [67] Pauli Virtanen et al. "SciPy 1.0: fundamental algorithms for scientific computing in Python". en. In: *Nat. Methods* 17.3 (Mar. 2020), pp. 261–272.
- [68] Skipper Seabold and Josef Perktold. *Statsmodels: Econometric and Statistical Modeling with Python*. 2010.

A.5. Co-author contributions

A.5.1. Exploring the SARS-CoV-2 virus-host-drug interactome for drug repurposing

Citation

"Exploring the SARS-CoV-2 virus-host-drug interactome for drug repurposing" Sepideh Sadegh, Julian Matschinske, David B. Blumenthal, Gihanna Galindez, Tim Kacprowski, Markus List, Reza Nasirigerdeh, Mhaned Oubounyt, Andreas Pichlmair, Tim D. Rose, Marisol Salgado-Albarrán, Julian Späth, Alexey Stukalov, Nina K. Wenke, Kevin Yuan, Josch K. Pauling, and Jan Baumbach; In: *Nature Communications* 11, 3518 (2020); doi: <https://doi.org/10.1038/s41467-020-17189-2>

Summary

The global pandemic of the Sars-CoV-2 virus required the development of new treatments and vaccinations in a short time. A way to achieve this is drug repurposing, where drugs, previously developed for other diseases, are used for COVID-19 patients.

In this publication, the interaction network-based method CoronaVirus Explorer (CoVex) was developed. It visualizes the virus-host-drug interactome and can suggest drug repurposing candidates. The interaction network was built, by combining, identified virus-host protein interactions, protein-protein interaction networks, and drug-target interactions. Researchers can explore the interactome and utilize network algorithms that prioritize drug repurposing candidates from seeds. Such seeds can emerge from previous research results or hypotheses. They can be drugs, human proteins, or viral proteins. The network algorithms are able to find the shortest paths or network communities that connect seeds, with drugs, or viral proteins. This can identify potential pathways that potentially inhibit viral replication. Different scenarios are showcased in the manuscript that shows applications of the web-based software. The results of a CoronaVirus Explorer (CoVex) analysis can then be used to further investigate the mechanisms and potential of key interactions between viral-, human proteins, and drugs.

Contribution

I contributed to the integration of the interactome and data analysis. Further, I contributed to the writing and Figures.

As stated in the publication: "S.S., J.M., J.B., M.L., T.K., J.K.P., A.P., and A.S. conceived and designed the study. S.S. and J.M. were in charge of overall direction, planning, and supervision. S.S., G.G., T.D.R., M.S.-A., and N.K.W. performed the acquisition, integration, and interpretation of data. S.S., D.B.B., M.L., and K.Y. developed and adapted the algorithms for network-based drug repurposing. J.M., R.N., M.O., and J.S. implemented the web platform. All authors provided critical feedback and helped in the interpretation of data, manuscript writing, and the improvement of the platform." [27].

A.5.2. Lessons from the COVID-19 pandemic for advancing computational drug repurposing strategies

Citation

"Lessons from the COVID-19 pandemic for advancing computational drug repurposing strategies" Gihanna Galindez, Julian Matschinske, Tim D. Rose, Sepideh Sadegh, Marisol Salgado-Albarrán, Julian Späth, Jan Baumbach, and Josch K. Pauling; In: *Nature Computational Science* 1, 33–41 (2021); doi: <https://doi.org/10.1038/s43588-020-00007-6>

Summary

During the COVID-19 pandemic, many scientific efforts focused on drug repurposing. In this review, we summarized and discussed, how drug repurposing research was performed and what challenges occurred.

First, common data resources for drug repurposing were assessed. Then, computational repurposing work was divided into virus-targeting and host-targeting approaches. Virus-targeting approaches commonly utilized docking simulations or neural networks, to find potential inhibitors for viral proteins. Host-targeting approaches used quantitative omics data to match signatures of drug application or interaction networks to find closely connected to drug targets. Furthermore, we compared their predictions to drugs in a trial at the point of publication.

Finally, a unified drug repurposing strategy was proposed to overcome the problems we observed in the literature. This included standardized molecular data and result reporting, a combination of host- and virus targeting approaches, combinatorial treatment suggestions, expert-guided analysis, and candidate validation.

Contribution

I focused with M.S.A. on the evaluation of virus-targeting approaches and comparison of their predictions to clinical trials. Together with all other co-authors, I equally contributed to the unified repurposing strategy.

As stated in the publication "G.G., J.M., T.D.R., S.S., M.S.A., J.S., J.B. and J.K.P. contributed equally to the manuscript writing. J.B. and J.K.P. were in charge of overall direction, planning and supervision. All authors provided critical feedback and helped to improve the manuscript." [26].

List of Figures

2.1. Information flow and regulation between biomolecules. Genes are transcribed to RNA, which are then translated to proteins. Specific proteins are then able to catalyze metabolic reactions. All layers are interconnected and have regulatory influences by activating, inhibiting, or providing building blocks for other biomolecules. License information: dna-nucleotides-ribbon and rna icon by Servier https://smart.servier.com/ are licensed under CC-BY 3.0 Unported https://creativecommons.org/licenses/by/3.0/	5
2.2. Glycolysis pathway. C6 glucose is broken down into two C3 pyruvate molecules. In this process, ATP and NADH are generated. Abbreviations: Adenosine diphosphate (ADP), Adenosine triphosphate (ATP), oxidized Nicotinamide adenine dinucleotide (NAD ⁺), reduced Nicotinamide adenine dinucleotide (NADH), Coenzyme A (CoA), Proton (H ⁺).	7
2.3. Structure of a Phosphatidylcholine (PC) lipid. The lipid consists of a glycerol backbone (green), a phospho-choline headgroup (orange), and two fatty acyls (blue). Lipid structure received from the SwissLipids database (identifier: SLM:000008040).	11
2.4. Important functions of lipids in the cell: As signaling molecules, long-term energy storage, and main membrane component.	13
2.5. Chemical structures of common lipid classes. Abbreviations: Lyso phosphatidic acid (LPA), Phosphatidic acid (PA), Phosphatidylethanolamine (PE), Diacylglycerol (DG), Triacylglycerol (TG), and Ether Phosphatidate (PA-O).	15
2.6. Common workflow of mass spectrometry-based lipidomics experiments.	19
2.7. Fragmentations of a PC(18:3_18:3) with a CH ₃ COO ⁻ adduct in negative scan mode. (A) MS ₂ spectra of the m/z 836.5 precursor. (B) Potential annotated fragments of the peaks in A. The figure is taken from the publication of J. K. Pauling et al. [73]. It has not been modified and is originally provided under the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/).	24
3.1. Difference between clustering (left) and biclustering (right) on an omics data matrix. Unaltered figure taken from Vvedenskaya and Rose et al. [21] (Supplementary Figure 3SA). Original figure distributed under the CC BY-NC-ND license (https://creativecommons.org/licenses/by-nc-nd/4.0/).	37

3.2.	Workflow of LINEX. Lipidomics data and metabolic rules are the input for the algorithm. This is used to create a lipid species network and combine it with statistical measures, such as hypothesis tests or correlation analysis. Unaltered figure taken from Köhler and Rose et al. [22]. Original figure distributed under the CC BY license (https://creativecommons.org/licenses/by/4.0/).	40
3.3.	Schema of the LINEX ² network enrichment algorithm. A LINEX ² network is converted into a hypergraph. Each hyperedge represents one lipid species reaction and connects substrates, products, and class reaction nodes. Hyperedges are weighted by changes in the substrate-to-product ratio between experimental conditions as an approximation for enzymatic dysregulation. A local network search is performed to find the maximally dysregulated subnetwork. Figure adapted from Rose et al. [23]. Permission granted by the authors.	41
5.1.	(A) Workflow of the LINEX ² network enrichment algorithm. (B) Score optimization of the network enrichment comparing WT and MBOAT7 knock-out liver lipidomics. Data taken from Thangapandi et al. [273]. The same data is used for the plots (C-F). (C) Optimized subnetwork. (D) Ratio distribution of the LPI ↔ PI reaction for all molecular species reactions. (E) Principal component analysis of the lipidomics data with all lipids and (D) only for the lipids in the subnetwork shown in (C) The figure is taken from the preprint of Rose et al. [23]. Permission granted by the authors.	50
6.1.	Integration of MoSbi, LINEX ² , and computational drug repurposing in a systems medicine context for a disease cohort with multi-omics data.	58

Acronyms

- ADP** Adenosine diphosphate. 6–8, 134
- AIF** All ion fragmentation. 23
- ATP** Adenosine triphosphate. 6–8, 14, 33, 134
- CB** Computational biology. 27, 28, 30, 34, 35, 60
- CCS** Collision cross-section. 21
- CDP** Cytidine diphosphate. 14, 15
- CEPT** choline/ethanolamine phosphotransferase. 15
- Cer** Ceramide. 15
- CL** Cardiolipin. 17
- CoA** Coenzyme A. 7, 8, 13, 14, 16, 49, 134
- CoVex** CoronaVirus Explorer. 132
- CVD** cardiovascular diseases. 1, 17, 18
- DDA** Data-dependent acquisition. 23
- DG** Diacylglycerol. 14, 15, 134
- DIA** Data-independent acquisition. 23
- DNA** deoxyribonucleic acid. 4
- ESI** Electrospray ionization. 21, 22
- FAD** flavin adenine dinucleotide. 8, 13, 14
- FBA** Flux Balance Analysis. 33
- FFA** free fatty acids. 17, 18, 21
- G3P** Glyceraldehyde 3-phosphate. 6, 7, 14

- GC** Gas chromatography. 21
- GO** Gene Ontology. 31
- GSEA** Gene Set Enrichment Analysis. 31, 32
- GTP** Guanosine triphosphate. 8
- H⁺** Proton. 7, 134
- HDL** high-density lipoprotein. 16–18
- IMS** Ion mobility spectrometry. 21, 22
- KEGG** Kyoto Encyclopedia of Genes and Genomes. 31
- LC** Liquid chromatography. 18, 20–22, 59
- LDL** low-density lipoprotein. 16, 18
- LINEX** Lipid Network Explorer. iii, iv, 2, 35–40, 47, 49, 53, 55–57, 59, 60, 63, 64, 135
- LINEX²** Lipid Network Explorer 2. iii, iv, 37, 39, 41, 49–51, 53, 55–60, 63, 64, 135
- LPA** Lyso phosphatidic acid. 14, 15, 39, 134
- LPA-O** Ether Lyso Phosphatidate. 15
- m/z** Mass-to-charge. 19, 21–26, 134
- MALDI** Matrix-assisted laser desorption/ionization. 21, 22, 26
- MBOAT** Membrane-bound O-acyltransferase. 16, 49, 50, 56, 135
- MG** Monoacylglycerol. 14
- MoSBi** Molecular Signature identification using Biclustering. iii, iv, 2, 10, 30, 35, 37, 38, 43, 53–58, 60, 62–64, 135
- mRNA** messenger RNA. 4, 25, 33
- MS** mass spectrometry. 18–23, 25, 26, 59–61
- NAD⁺** oxidized Nicotinamide adenine dinucleotide. 7, 8, 14, 134
- NADH** reduced Nicotinamide adenine dinucleotide. 7, 8, 134
- NADPH** reduced nicotinamide adenine dinucleotide phosphate. 9, 10, 13

- NAFLD** non-alcoholic fatty liver disease. iii, 1, 2, 18, 20, 29, 30, 34, 36, 38, 45, 54, 55, 57, 59, 60, 63, 64
- NASH** non-alcoholic steatohepatitis. 45, 54
- NGS** next-generation sequencing. 25
- NMR** nuclear magnetic resonance spectroscopy. 26
- ORA** Over-representation Analysis. 31, 32
- PA** Phosphatidic acid. 14, 15, 17, 39, 134
- PA-O** Ether Phosphatidate. 15, 134
- PC** Phosphatidylcholine. 11, 12, 14, 15, 17, 51, 134
- PCA** Principal Component Analysis. 29
- PE** Phosphatidylethanolamine. 14, 15, 39, 51, 134
- PI** Phosphatidylinositol. 14, 16, 17, 49
- PLA₂** Phospholipase A₂. 16
- PPP** Pentose phosphate pathway. 9, 10, 13, 32
- PS** Phosphatidylserine. 15, 16, 39
- RNA** ribonucleic Acid. 4, 10, 26, 43
- SM** Sphingomyelin. 15, 45, 54
- sn** stereospecifically numbered. 12, 59
- SWATH** Sequential window acquisition of all theoretical fragment ion spectra. 23
- TCA** tricarboxylic acid. 8, 10, 14
- TG** Triacylglycerol. 12, 14–18, 134
- TLC** thin-layer chromatography. 18, 61
- TOF** time-of-flight. 22
- t-SNE** t-Distributed Stochastic Neighbor Embedding. 29
- UMAP** Uniform Manifold Approximation and Projection. 29
- VLDL** very-low-density lipoprotein. 16, 17
- WT** wild type. 49–51, 135

Bibliography

- [1] Z. Younossi et al. "Global Perspectives on Nonalcoholic Fatty Liver Disease and Nonalcoholic Steatohepatitis". In: *Hepatology* 69 (6 2019), pp. 2672–2682. ISSN: 1527-3350. DOI: 10.1002/HEP.30251.
- [2] W. B. Cannon. "ORGANIZATION FOR PHYSIOLOGICAL HOMEOSTASIS". In: *Physiological Reviews* 9.3 (1929), pp. 399–431. DOI: 10.1152/physrev.1929.9.3.399. eprint: <https://doi.org/10.1152/physrev.1929.9.3.399>.
- [3] A. Konnopka, M. Bödemann, and H. H. König. "Health burden and costs of obesity and overweight in Germany". In: *European Journal of Health Economics* 12 (4 2011), pp. 345–352. ISSN: 16187598. DOI: 10.1007/S10198-010-0242-6.
- [4] C. Heidemann and C. Scheidt-Nave. "Prevalence, incidence and mortality of diabetes mellitus in adults in Germany – A review in the framework of the Diabetes Surveillanc". In: *Journal of Health Monitoring*. Vol. 2. 3. Robert Koch-Institut, Epidemiologie und Gesundheitsberichterstattung, 2017. DOI: 10.17886/RKI-GBE-2017-062.
- [5] L. J. Laslett et al. "The Worldwide Environment of Cardiovascular Disease: Prevalence, Diagnosis, Therapy, and Policy Issues: A Report From the American College of Cardiology". In: *Journal of the American College of Cardiology* 60 (25 SUPPL. 2012), S1–S49. ISSN: 07351097. DOI: 10.1016/J.JACC.2012.11.002.
- [6] P. Angulo. "Nonalcoholic Fatty Liver Disease". In: *New England Journal of Medicine* 346 (16 2002), pp. 1221–1231. ISSN: 0028-4793. DOI: 10.1056/NEJMRA011775.
- [7] D. Giedrimiene and R. King. "Abstract 207: Burden of Cardiovascular Disease (CVD) on Economic Cost. Comparison of Outcomes in US and Europe". In: *Circulation: Cardiovascular Quality and Outcomes* 10 (suppl_3 2017). ISSN: 1941-7713. DOI: 10.1161/CIRCOUTCOMES.10.SUPPL_3.207.
- [8] Z. M. Younossi and L. Henry. "Economic and Quality-of-Life Implications of Non-Alcoholic Fatty Liver Disease". In: *PharmacoEconomics* 33 (12 2015), pp. 1245–1253. ISSN: 11792027. DOI: 10.1007/S40273-015-0316-5.
- [9] Z. M. Younossi et al. "The economic and clinical burden of nonalcoholic fatty liver disease in the United States and Europe". In: *Hepatology* 64 (5 2016), pp. 1577–1586. ISSN: 15273350. DOI: 10.1002/HEP.28785.
- [10] A. G. Kohli et al. "Designer lipids for drug delivery: From heads to tails". In: *Journal of Controlled Release* 190 (2014), pp. 274–287. ISSN: 0168-3659. DOI: 10.1016/J.JCONREL.2014.04.047.

- [11] B. Berger, J. Peng, and M. Singh. "Computational solutions for omics data". In: *Nature Reviews Genetics* 14 (5 2013), pp. 333–346. ISSN: 1471-0064. DOI: 10.1038/nrg3433.
- [12] C. Auffray, Z. Chen, and L. Hood. "Systems medicine: The future of medical genomics and healthcare". In: *Genome Medicine* 1 (1 2009), pp. 1–11. ISSN: 1756994X. DOI: 10.1186/GM2/METRICS.
- [13] T. Nguyen et al. "A novel approach for data integration and disease subtyping". In: *Genome Research* 27 (12 2017), pp. 2025–2039. ISSN: 1088-9051. DOI: 10.1101/GR.215129.116.
- [14] Y. Hasin, M. Seldin, and A. Lusk. "Multi-omics approaches to disease". In: *Genome Biology* 18 (1 2017), pp. 1–15. DOI: 10.1186/S13059-017-1215-1.
- [15] D. C. Collins et al. "Towards Precision Medicine in the Clinic: From Biomarker Discovery to Novel Therapeutics". In: *Trends in Pharmacological Sciences* 38 (1 2017), pp. 25–40. ISSN: 0165-6147. DOI: 10.1016/J.TIPS.2016.10.012.
- [16] S. Saria and A. Goldenberg. "Subtyping: What It is and Its Role in Precision Medicine". In: *IEEE Intelligent Systems* 30 (4 2015), pp. 70–75. ISSN: 15411672. DOI: 10.1109/MIS.2015.60.
- [17] Y. Yamamoto et al. "Current Status, Issues and Future Prospects of Personalized Medicine for Each Disease". In: *Journal of Personalized Medicine* 12 (3 2022), p. 444. ISSN: 2075-4426. DOI: 10.3390/JPM12030444.
- [18] K. A. Phillips et al. "The economic value of personalized medicine tests: what we know and what we need to know". In: *Genetics in Medicine* 16 (3 2013), pp. 251–257. ISSN: 1530-0366. DOI: 10.1038/gim.2013.122.
- [19] C. R. Santos and A. Schulze. "Lipid metabolism in cancer". In: *The FEBS Journal* 279 (15 2012), pp. 2610–2623. ISSN: 1742-4658. DOI: 10.1111/J.1742-4658.2012.08644.X.
- [20] T. D. Rose et al. "MoSBi: Automated signature mining for molecular stratification and subtyping". In: *Proceedings of the National Academy of Sciences* 119 (16 2022). ISSN: 0027-8424. DOI: 10.1073/PNAS.2118210119.
- [21] O. Vvedenskaya et al. "Nonalcoholic fatty liver disease stratification by liver lipidomics". In: *Journal of Lipid Research* 62 (2021), p. 100104. ISSN: 15397262. DOI: 10.1016/J.JLR.2021.100104.
- [22] N. Köhler et al. "Investigating Global Lipidome Alterations with the Lipid Network Explorer". In: *Metabolites* 11 (8 2021), p. 488. ISSN: 2218-1989. DOI: 10.3390/METABO11080488.
- [23] T. D. Rose et al. "Lipid network and moiety analysis for revealing enzymatic dysregulation and mechanistic alterations from lipidomics data". In: *bioRxiv* (2022), p. 2022.02.04.479101. DOI: 10.1101/2022.02.04.479101.
- [24] M. D. Rawlins. "Cutting the cost of drug development?" In: *Nature Reviews Drug Discovery* 3 (4 2004), pp. 360–364. ISSN: 1474-1784. DOI: 10.1038/nrd1347.

- [25] K. Park. "A review of computational drug repurposing". In: *Translational and Clinical Pharmacology* 27 (2 2019), pp. 59–63. ISSN: 22890882. DOI: 10.12793/TCP.2019.27.2.59.
- [26] G. Galindez et al. "Lessons from the COVID-19 pandemic for advancing computational drug repurposing strategies". In: *Nature Computational Science* 1 (1 2021), pp. 33–41. ISSN: 2662-8457. DOI: 10.1038/s43588-020-00007-6.
- [27] S. Sadegh et al. "Exploring the SARS-CoV-2 virus-host-drug interactome for drug repurposing". In: *Nature Communications* 11 (1 2020), pp. 1–9. ISSN: 2041-1723. DOI: 10.1038/s41467-020-17189-2.
- [28] F. H. Crick. "On protein synthesis". In: *Symp Soc Exp Biol*. Vol. 12. 138-63. 1958, p. 8.
- [29] R. Singal and G. D. Ginder. "DNA Methylation". In: *Blood* 93 (12 1999), pp. 4059–4070. ISSN: 0006-4971. DOI: 10.1182/BLOOD.V93.12.4059.
- [30] T. E. Creighton. "Protein folding." In: *Biochemical Journal* 270 (1 1990), p. 1. ISSN: 02646021. DOI: 10.1042/BJ2700001.
- [31] P. Aloy and R. B. Russell. "Structural systems biology: modelling protein interactions". In: *Nature Reviews Molecular Cell Biology* 7 (3 2006), pp. 188–197. ISSN: 1471-0080. DOI: 10.1038/nrm1859.
- [32] H. H. Freeze. "Genetic defects in the human glycome". In: *Nature Reviews Genetics* 7 (7 2006), pp. 537–551. ISSN: 1471-0064. DOI: 10.1038/nrg1894.
- [33] I. Martincorena and P. J. Campbell. "Somatic mutation in cancer and normal cells". In: *Science* 349 (6255 2015), pp. 1483–1489. ISSN: 10959203. DOI: 10.1126/SCIENCE.AAB4082.
- [34] M. I. McCarthy, P. F. ; 2, and . J. Tuomilehto. "How Obesity Causes Diabetes: Not a Tall Tale". In: *Science* 307 (5708 2005), pp. 373–375. ISSN: 00368075. DOI: 10.1126/SCIENCE.1104342.
- [35] P. V. Hornbeck et al. "PhosphoSite: A bioinformatics resource dedicated to physiological protein phosphorylation". In: *PROTEOMICS* 4 (6 2004), pp. 1551–1561. ISSN: 1615-9861. DOI: 10.1002/PMIC.200300772.
- [36] R. L. Schowen. "How an enzyme surmounts the activation energy barrier". In: *Proceedings of the National Academy of Sciences of the United States of America* 100 (21 2003), p. 11931. ISSN: 00278424. DOI: 10.1073/PNAS.2235806100.
- [37] L. Pauling. "Molecular architecture and biological reactions". In: *Chemical and engineering news* 24.10 (1946), pp. 1375–1377.
- [38] N. Swainston et al. "Recon 2.2: from reconstruction to model of human metabolism". In: *Metabolomics* 12 (7 2016), pp. 1–7. ISSN: 15733890. DOI: 10.1007/S11306-016-1051-4/TABLES/2.
- [39] P. J. Butterworth, F. J. Warren, and P. R. Ellis. "Human alpha-amylase and starch digestion: An interesting marriage". In: *Starch - Stärke* 63 (7 2011), pp. 395–405. ISSN: 1521-379X. DOI: 10.1002/STAR.201000150.

- [40] G. W. Gould and G. D. Holman. "The glucose transporter family: structure, function and tissue-specific expression." In: *Biochemical Journal* 295 (Pt 2 1993), p. 329. ISSN: 02646021. DOI: 10.1042/BJ2950329.
- [41] E. M. Wright. "Glucose transport families SLC5 and SLC50". In: *Molecular Aspects of Medicine* 34 (2-3 2013), pp. 183–196. ISSN: 0098-2997. DOI: 10.1016/J.MAM.2012.11.002.
- [42] D. L. Jack, N. M. Yang, and M. H. Saier. "The drug/metabolite transporter superfamily". In: *European Journal of Biochemistry* 268 (13 2001), pp. 3620–3639. ISSN: 1432-1033. DOI: 10.1046/J.1432-1327.2001.02265.X.
- [43] Y. Kanai and M. A. Hediger. "The glutamate/neutral amino acid transporter family SLC1: Molecular, physiological and pharmacological aspects". In: *Pflugers Archiv European Journal of Physiology* 447 (5 2004), pp. 469–479. ISSN: 00316768. DOI: 10.1007/S00424-003-1146-4.
- [44] F. Palmieri et al. "Mitochondrial metabolite transporters". In: *Biochimica et Biophysica Acta (BBA)-Bioenergetics* 1275 (1-2 1996), pp. 127–132.
- [45] Z. T. Schug, J. V. Voorde, and E. Gottlieb. "The metabolic fate of acetate in cancer". In: *Nature Reviews Cancer* 16 (11 2016), pp. 708–717. ISSN: 1474-1768. DOI: 10.1038/nrc.2016.87.
- [46] A. Boiteux and B. Hess. "Design of glycolysis". In: *Philosophical Transactions of the Royal Society of London. B, Biological Sciences* 293 (1063 1981), pp. 5–22. ISSN: 09628436. DOI: 10.1098/RSTB.1981.0056.
- [47] M. Akram. "Citric Acid Cycle and Role of its Intermediates in Metabolism". In: *Cell Biochemistry and Biophysics* 68 (3 2013), pp. 475–478. ISSN: 1559-0283. DOI: 10.1007/S12013-013-9750-1.
- [48] Y. Hatefi. "The mitochondrial electron transport and oxidative phosphorylation system". In: *Annual review of biochemistry* 54 (1 1985), pp. 1015–1069.
- [49] S. Nath and J. Villadsen. "Oxidative phosphorylation revisited". In: *Biotechnology and Bioengineering* 112 (3 2015), pp. 429–437. ISSN: 1097-0290. DOI: 10.1002/BIT.25492.
- [50] T. Wood. *The pentose phosphate pathway*. Elsevier, 2012.
- [51] J. F. Williams, P. F. Blackmore, and M. G. Clark. "New reaction sequences for the non-oxidative pentose phosphate pathway". In: *Biochemical Journal* 176 (1 1978), pp. 257–282. ISSN: 0264-6021. DOI: 10.1042/BJ1760257.
- [52] S. Thorell et al. "The three-dimensional structure of human transaldolase". In: *FEBS Letters* 475 (3 2000), pp. 205–208. ISSN: 0014-5793. DOI: 10.1016/S0014-5793(00)01658-6.
- [53] "The return of metabolism: biochemistry and physiology of the pentose phosphate pathway". In: *Biological Reviews* 90 (3 2015), pp. 927–963. ISSN: 1469-185X. DOI: 10.1111/BRV.12140.
- [54] M. Hatting et al. "Insulin regulation of gluconeogenesis". In: *Annals of the New York Academy of Sciences* 1411 (1 2018), pp. 21–35. ISSN: 1749-6632. DOI: 10.1111/NYAS.13435.

- [55] J. H. Exton. "Gluconeogenesis". In: *Metabolism* 21 (10 1972), pp. 945–990. ISSN: 0026-0495. DOI: 10.1016/0026-0495(72)90028-5.
- [56] H. G. Hers and L. Hue. "Gluconeogenesis and related aspects of glycolysis". In: *Annual review of biochemistry* 52 (1 1983), pp. 617–653.
- [57] D. Hanahan and R. A. Weinberg. "Hallmarks of cancer: The next generation". In: *Cell* 144 (5 2011), pp. 646–674. ISSN: 00928674. DOI: 10.1016/J.CELL.2011.02.013.
- [58] J. Bi et al. "Oncogene Amplification in Growth Factor Signaling Pathways Renders Cancers Dependent on Membrane Lipid Remodeling". In: *Cell Metabolism* 30 (3 2019), 525–538.e8. ISSN: 1550-4131. DOI: 10.1016/J.CMET.2019.06.014.
- [59] A. M. Gouw et al. "Oncogene KRAS activates fatty acid synthase, resulting in specific ERK and lipid signatures associated with lung adenocarcinoma". In: *Proceedings of the National Academy of Sciences of the United States of America* 114 (17 2017), pp. 4300–4305. ISSN: 10916490. DOI: 10.1073/PNAS.1617709114.
- [60] O. Warburg, F. Wind, and E. Negelein. "The Metabolism of Tumors in the Body". In: *The Journal of General Physiology* 8 (6 1927), p. 519. ISSN: 15407748. DOI: 10.1085/JGP.8.6.519.
- [61] M. Potter, E. Newport, and K. J. Morten. "The Warburg effect: 80 years on". In: *Biochemical Society Transactions* 44 (5 2016), pp. 1499–1505. ISSN: 0300-5127. DOI: 10.1042/BST20160094.
- [62] R. J. DeBerardinis and N. S. Chandel. "We need to talk about the Warburg effect". In: *Nature Metabolism* 2 (2 2020), pp. 127–129. ISSN: 2522-5812. DOI: 10.1038/s42255-020-0172-2.
- [63] K. C. Patra and N. Hay. "The pentose phosphate pathway and cancer". In: *Trends in Biochemical Sciences* 39 (8 2014), pp. 347–354. ISSN: 09680004. DOI: 10.1016/j.tibs.2014.06.005.
- [64] M. V. Liberti and J. W. Locasale. "The Warburg Effect: How Does it Benefit Cancer Cells?" In: *Trends in Biochemical Sciences* 41 (3 2016), pp. 211–218. ISSN: 0968-0004. DOI: 10.1016/J.TIBS.2015.12.001.
- [65] A. Nilsson et al. "Quantitative analysis of amino acid metabolism in liver cancer links glutamate excretion to nucleotide synthesis". In: *Proceedings of the National Academy of Sciences* 117 (19 2020), pp. 10294–10304. ISSN: 0027-8424. DOI: 10.1073/PNAS.1919250117.
- [66] R. J. D. Berardinis and N. S. Chandel. "Fundamentals of cancer metabolism". In: *Science Advances* 2 (5 2016). ISSN: 23752548. DOI: 10.1126/SCIADV.1600200.
- [67] D. J. Betteridge. "What is oxidative stress?" In: *Metabolism* 49 (2 2000), pp. 3–8. ISSN: 0026-0495. DOI: 10.1016/S0026-0495(00)80077-3.
- [68] M. Frisard and E. Ravussin. "Energy metabolism and oxidative stress". In: *Endocrine* 29 (1 2006), pp. 27–32. ISSN: 1559-0100. DOI: 10.1385/ENDO:29:1:27.

- [69] T. Yu et al. "CircRNAs in cancer metabolism: A review". In: *Journal of Hematology and Oncology* 12 (1 2019), pp. 1–10. ISSN: 17568722. DOI: 10.1186/S13045-019-0776-8.
- [70] U. E. Martinez-Outschoorn et al. "Cancer metabolism: a therapeutic perspective". In: *Nature Reviews Clinical Oncology* 14 (1 2016), pp. 11–31. ISSN: 1759-4782. DOI: 10.1038/nrclinonc.2016.60.
- [71] E. Fahy et al. "Update of the LIPID MAPS comprehensive classification system for lipids¹". In: *Journal of Lipid Research* 50 (SUPPL. 2009), S9–S14. ISSN: 0022-2275. DOI: 10.1194/JLR.R800095-JLR200.
- [72] G. Liebisch et al. "Update on LIPID MAPS classification, nomenclature, and shorthand notation for MS-derived lipid structures". In: *Journal of Lipid Research* 61 (12 2020), pp. 1539–1555. ISSN: 0022-2275. DOI: 10.1194/JLR.S120001025.
- [73] J. K. Pauling et al. "Proposal for a common nomenclature for fragment ions in mass spectra of lipids". In: *PLOS ONE* 12 (11 2017), e0188394. ISSN: 1932-6203. DOI: 10.1371/JOURNAL.PONE.0188394.
- [74] F. Callegarin et al. "Lipids and biopackaging". In: *Journal of the American Oil Chemists' Society* 74 (10 1997), pp. 1183–1192. ISSN: 1558-9331. DOI: 10.1007/S11746-997-0044-X.
- [75] S. J. Singer and G. L. Nicolson. "The Fluid Mosaic Model of the Structure of Cell Membranes". In: *Science* 175 (4023 1972), pp. 720–731. ISSN: 00368075. DOI: 10.1126/SCIENCE.175.4023.720.
- [76] K. Simons and E. Ikonen. "Functional rafts in cell membranes". In: *Nature* 387 (6633 1997), pp. 569–572. ISSN: 1476-4687. DOI: 10.1038/42408.
- [77] D. Lingwood and K. Simons. "Lipid rafts as a membrane-organizing principle". In: *Science* 327 (5961 2010), pp. 46–50. ISSN: 00368075. DOI: 10.1126/SCIENCE.1174621.
- [78] K. Simons and D. Toomre. "Lipid rafts and signal transduction". In: *Nature Reviews Molecular Cell Biology* 1 (1 2000), pp. 31–39. ISSN: 1471-0080. DOI: 10.1038/35036052.
- [79] S. J. Wakil. "Mechanism of fatty acid synthesis". In: *Journal of Lipid Research* 2 (1 1961), pp. 1–24. ISSN: 0022-2275. DOI: 10.1016/S0022-2275(20)39034-9.
- [80] M. Pfeuffer and A. Jaudszus. "Pentadecanoic and Heptadecanoic Acids: Multifaceted Odd-Chain Fatty Acids". In: *Advances in Nutrition* 7 (4 2016), pp. 730–734. ISSN: 2161-8313. DOI: 10.3945/AN.115.011387.
- [81] H. Guillou, P. G. P. Martin, and T. Pineau. "Transcriptional Regulation of Hepatic Fatty Acid Metabolism". In: *Lipids in Health and Disease, Subcellular Biochemistry*. Vol. 49. Springer, Dordrecht, 2008, pp. 3–47. DOI: 10.1007/978-1-4020-8831-5_1.
- [82] H. Schulz. "Beta oxidation of fatty acids". In: *Biochimica et Biophysica Acta (BBA) - Lipids and Lipid Metabolism* 1081 (2 1991), pp. 109–120. ISSN: 0005-2760. DOI: 10.1016/0005-2760(91)90015-A.
- [83] R. G. H. Downer. "Lipid metabolism". In: *Comprehensive insect physiology, biochemistry and pharmacology* 10 (1985), pp. 77–113.

- [84] W. E. Lands. "Lipid Metabolism". In: *Annual Review of Biochemistry* 34 (1965), pp. 313–346. ISSN: 00664154. DOI: 10.1146/ANNUREV.BI.34.070165.001525.
- [85] F. K. Winkler, A. D'Arcy, and W. Hunziker. "Structure of human pancreatic lipase". In: *Nature* 343 (6260 1990), pp. 771–774. ISSN: 1476-4687. DOI: 10.1038/343771a0.
- [86] D. Poccia and B. Larijani. "Phosphatidylinositol metabolism and membrane fusion". In: *Biochemical Journal* 418 (2 2009), pp. 233–246. ISSN: 0264-6021. DOI: 10.1042/BJ20082105.
- [87] N. J. Blunsom and S. Cockcroft. "Phosphatidylinositol synthesis at the endoplasmic reticulum". In: *Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids* 1865 (1 2020), p. 158471. ISSN: 1388-1981. DOI: 10.1016/J.BBALIP.2019.05.015.
- [88] J. N. van der Veen et al. "The critical role of phosphatidylcholine and phosphatidylethanolamine metabolism in health and disease". In: *Biochimica et Biophysica Acta (BBA) - Biomembranes* 1859 (9 2017), pp. 1558–1572. ISSN: 0005-2736. DOI: 10.1016/J.BBAMEM.2017.04.006.
- [89] E. P. Kennedy and S. B. Weiss. "The Function of Cytidine Coenzymes in the Biosynthesis of Phospholipides". In: *Journal of Biological Chemistry* 222 (1 1956), pp. 193–214. ISSN: 0021-9258. DOI: 10.1016/S0021-9258(19)50785-2.
- [90] F. Gibellini and T. K. Smith. "The Kennedy pathway—De novo synthesis of phosphatidylethanolamine and phosphatidylcholine". In: *IUBMB Life* 62 (6 2010), pp. 414–428. ISSN: 1521-6551. DOI: 10.1002/IUB.337.
- [91] K. Watschinger and E. R. Werner. "Orphan enzymes in ether lipid metabolism". In: *Biochimie* 95 (1 2013), pp. 59–65. ISSN: 0300-9084. DOI: 10.1016/J.BIOCHI.2012.06.027.
- [92] A. H. Merrill. "De Novo Sphingolipid Biosynthesis: A Necessary, but Dangerous, Pathway *". In: *Journal of Biological Chemistry* 277 (29 2002), pp. 25843–25846. ISSN: 0021-9258. DOI: 10.1074/JBC.R200009200.
- [93] G. Stankeviciute et al. "Convergent evolution of bacterial ceramide synthesis". In: *Nature Chemical Biology* (2021), pp. 1–8. ISSN: 1552-4469. DOI: 10.1038/s41589-021-00948-7.
- [94] M. A. Gijón et al. "Lysophospholipid Acyltransferases and Arachidonate Recycling in Human Neutrophils". In: *Journal of Biological Chemistry* 283 (44 2008), pp. 30235–30245. ISSN: 0021-9258. DOI: 10.1074/JBC.M806194200.
- [95] D. Hayashi, V. D. Mouchlis, and E. A. Dennis. "Omega-3 versus Omega-6 fatty acid availability is controlled by hydrophobic site geometries of phospholipase A2s". In: *Journal of Lipid Research* 63 (2021), p. 100113. ISSN: 15397262. DOI: 10.1016/J.JLR.2021.100113.
- [96] V. Varadharajan, W. J. Massey, and J. M. Brown. "Membrane-Bound O-Acyltransferase 7 (MBOAT7) Driven Phosphatidylinositol Remodeling in Advanced Liver Disease." In: *Journal of Lipid Research* (2022), p. 100234. ISSN: 0022-2275. DOI: 10.1016/J.JLR.2022.100234.

- [97] S. Lev. "Non-vesicular lipid transport by lipid-transfer proteins and beyond". In: *Nature Reviews Molecular Cell Biology* 11 (10 2010), pp. 739–750. ISSN: 1471-0080. DOI: 10.1038/nrm2971.
- [98] H. N. Ginsberg. "LIPOPROTEIN PHYSIOLOGY". In: *Endocrinology and Metabolism Clinics* 27 (3 1998), pp. 503–519. ISSN: 0889-8529. DOI: 10.1016/S0889-8529(05)70023-2.
- [99] A. Jomard and E. Osto. "High Density Lipoproteins: Metabolism, Function, and Therapeutic Potential". In: *Frontiers in Cardiovascular Medicine* 7 (2020), p. 39. ISSN: 2297055X. DOI: 10.3389/FCVM.2020.00039/BIBTEX.
- [100] M. Banach et al. "Intensive LDL-cholesterol lowering therapy and neurocognitive function". In: *Pharmacology & Therapeutics* 170 (2017), pp. 181–191. ISSN: 0163-7258. DOI: 10.1016/J.PHARMTHERA.2016.11.001.
- [101] *The Nobel Prize in Physiology or Medicine 1964 - Summary*. URL: <https://www.nobelprize.org/prizes/medicine/1964/summary/> (visited on 08/22/2022).
- [102] *The Nobel Prize in Physiology or Medicine 1985 - Press release*. URL: <https://www.nobelprize.org/prizes/medicine/1985/press-release/> (visited on 08/22/2022).
- [103] A. A. Noga and D. E. Vance. "Insights into the requirement of phosphatidylcholine synthesis for liver function in mice". In: *Journal of Lipid Research* 44 (10 2003), pp. 1998–2005. ISSN: 0022-2275. DOI: 10.1194/JLR.M300226-JLR200.
- [104] W. F. Boss and Y. J. Im. "Phosphoinositide Signaling". In: *Annual Review of Plant Biology* 63 (2012), pp. 409–429. ISSN: 15435008. DOI: 10.1146/ANNUREV-ARPLANT-042110-103840.
- [105] R. Liu et al. "PI3K/AKT pathway as a key link modulates the multidrug resistance of cancers". In: *Cell Death & Disease* 11 (9 2020), pp. 1–12. ISSN: 2041-4889. DOI: 10.1038/s41419-020-02998-6.
- [106] B. A. Hemmings and D. F. Restuccia. "PI3K-PKB/Akt Pathway". In: *Cold Spring Harbor Perspectives in Biology* 4 (9 2012). ISSN: 19430264. DOI: 10.1101/CSHPERSPECT.A011189.
- [107] M. Ren, C. K. Phoon, and M. Schlame. "Metabolism and function of mitochondrial cardiolipin". In: *Progress in Lipid Research* 55 (1 2014), pp. 1–16. ISSN: 0163-7827. DOI: 10.1016/J.PLIPRES.2014.04.001.
- [108] C. Mencarelli and P. Martinez-Martinez. "Ceramide function in the brain: when a slight tilt is enough". In: *Cellular and Molecular Life Sciences* 70 (2 2012), pp. 181–203. ISSN: 1420-9071. DOI: 10.1007/S00018-012-1038-X.
- [109] A. L. Santos and G. Preta. "Lipids in the cell: organisation regulates function". In: *Cellular and Molecular Life Sciences* 75 (11 2018), pp. 1909–1927. ISSN: 1420-9071. DOI: 10.1007/S00018-018-2765-4.
- [110] A. Laganowsky et al. "Membrane proteins bind lipids selectively to modulate their structure and function". In: *Nature* 510 (7503 2014), pp. 172–175. ISSN: 1476-4687. DOI: 10.1038/nature13419.

- [111] E. Tookmanian et al. "Hopanoids Confer Robustness to Physicochemical Variability in the Niche of the Plant Symbiont *Bradyrhizobium diazoefficiens*". In: *Journal of Bacteriology* (2022). Ed. by A. Becker. ISSN: 0021-9193. DOI: 10.1128/JB.00442-21.
- [112] P. Davies et al. "The Role of Arachidonic Acid Oxygenation Products in Pain and Inflammation". In: *Annual Review of Immunology* 2 (1984), pp. 335–357. ISSN: 07320582. DOI: 10.1146/ANNUREV.IY.02.040184.002003.
- [113] J. M. Jungersted et al. "Lipids and skin barrier function – a clinical perspective". In: *Contact Dermatitis* 58 (5 2008), pp. 255–262. ISSN: 1600-0536. DOI: 10.1111/J.1600-0536.2008.01320.X.
- [114] R. H. Eckel, S. M. Grundy, and P. Z. Zimmet. "The metabolic syndrome". In: *The Lancet* 365 (9468 2005), pp. 1415–1428. ISSN: 0140-6736. DOI: 10.1016/S0140-6736(05)66378-7.
- [115] S. Bernardi et al. "The Complex Interplay between Lipids, Immune System and Interleukins in Cardio-Metabolic Diseases". In: *International Journal of Molecular Sciences* 19 (12 2018), p. 4058. ISSN: 1422-0067. DOI: 10.3390/IJMS19124058.
- [116] C. H. Lee, P. Olson, and R. M. Evans. "Minireview: Lipid Metabolism, Metabolic Diseases, and Peroxisome Proliferator-Activated Receptors". In: *Endocrinology* 144 (6 2003), pp. 2201–2207. ISSN: 0013-7227. DOI: 10.1210/EN.2003-0288.
- [117] J. M. Weinberg. "Lipotoxicity". In: *Kidney International* 70 (9 2006), pp. 1560–1566. ISSN: 0085-2538. DOI: 10.1038/SJ.KI.5001834.
- [118] J. E. Schaffer. "Lipotoxicity: when tissues overeat". In: *Current opinion in lipidology* 14.3 (2003), pp. 281–287.
- [119] J. Suburu and Y. Q. Chen. "Lipids and prostate cancer". In: *Prostaglandins & Other Lipid Mediators* 98 (1-2 2012), pp. 1–10. ISSN: 1098-8823. DOI: 10.1016/J.PROSTAGLANDINS.2012.03.003.
- [120] A. Pakiet et al. "Changes in lipids composition and metabolism in colorectal cancer: a review". In: *Lipids in Health and Disease* 18 (1 2019), pp. 1–21. ISSN: 1476-511X. DOI: 10.1186/S12944-019-0977-8.
- [121] J. Jiang, P. Nilsson-Ehle, and N. Xu. "Influence of liver cancer on lipid and lipoprotein metabolism". In: *Lipids in Health and Disease* 5 (1 2006), pp. 1–7. ISSN: 1476511X. DOI: 10.1186/1476-511X-5-4.
- [122] L. H. Duntas. "Thyroid Disease and Lipids". In: *Thyroid* 12 (4 2004), pp. 287–293. ISSN: 10507256. DOI: 10.1089/10507250252949405.
- [123] H. Xicoy, B. Wieringa, and G. J. M. Martens. "The Role of Lipids in Parkinson's Disease". In: *Cells* 8 (1 2019), p. 27. ISSN: 2073-4409. DOI: 10.3390/CELLS8010027.
- [124] P. Foley. "Lipids in Alzheimer's disease: A century-old story". In: *Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids* 1801 (8 2010), pp. 750–753. ISSN: 1388-1981. DOI: 10.1016/J.BBALIP.2010.05.004.

- [125] N. S. Heaton and G. Randall. "Multifaceted roles for lipids in viral infection". In: *Trends in Microbiology* 19 (7 2011), pp. 368–375. ISSN: 0966-842X. DOI: 10.1016/J.TIM.2011.03.007.
- [126] M. Wenk. "The emerging field of lipidomics". In: *Nature Reviews Drug Discovery* 4 (7 2005), pp. 594–610. ISSN: 1474-1784. DOI: 10.1038/nrd1776.
- [127] M. R. Wenk. "Lipidomics: New Tools and Applications". In: *Cell* 143 (6 2010), pp. 888–895. ISSN: 0092-8674. DOI: 10.1016/J.CELL.2010.11.033.
- [128] R. M. Deranieh, A. S. Joshi, and M. L. Greenberg. "Thin-Layer Chromatography of Phospholipids". In: *Methods in molecular biology (Clifton, N.J.)* 1033 (2013), p. 21. ISSN: 10643745. DOI: 10.1007/978-1-62703-487-6_2.
- [129] B. Fuchs et al. "Lipid analysis by thin-layer chromatography—A review of the current state". In: *Journal of Chromatography A* 1218 (19 2011), pp. 2754–2774. ISSN: 0021-9673. DOI: 10.1016/J.CHROMA.2010.11.066.
- [130] M. Lange et al. "AdipoAtlas: A reference lipidome for human white adipose tissue". In: *Cell Reports Medicine* 2 (10 2021), p. 100407. ISSN: 2666-3791. DOI: 10.1016/J.XCRM.2021.100407.
- [131] D. Schwudke et al. "Shotgun Lipidomics on High Resolution Mass Spectrometers". In: *Cold Spring Harbor Perspectives in Biology* 3 (9 2011), a004614. ISSN: 19430264. DOI: 10.1101/CSHPERSPECT.A004614.
- [132] A. S. Woods and S. N. Jackson. "Brain tissue lipidomics: Direct probing using matrix-assisted laser desorption/ionization mass spectrometry". In: *The AAPS Journal* 8 (2 2006), E391–E395. ISSN: 1550-7416. DOI: 10.1007/BF02854910.
- [133] G. Liebisch et al. "Reporting of lipidomics data should be standardized". In: *Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids* 1862 (8 2017), pp. 747–751. ISSN: 1388-1981. DOI: 10.1016/J.BBALIP.2017.02.013.
- [134] T. M. Annesley. "Ion Suppression in Mass Spectrometry". In: *Clinical Chemistry* 49 (7 2003), pp. 1041–1044. ISSN: 0009-9147. DOI: 10.1373/49.7.1041.
- [135] X. Jiang, K. Yang, and X. Han. "Direct quantitation of psychosine from alkaline-treated lipid extracts with a semi-synthetic internal standard". In: *Journal of Lipid Research* 50 (1 2009), pp. 162–172. ISSN: 0022-2275. DOI: 10.1194/JLR.D800036-JLR200.
- [136] J. Folch, M. Lees, and G. H. S. Stanley. "A simple method for the isolation and purification of total lipids from animal tissues". In: *J Biol Chem* 226 (1 1957), pp. 497–509.
- [137] E. G. Bligh and W. J. Dyer. "A rapid method of total lipid extraction and purification". In: *Canadian journal of biochemistry and physiology* 37 (8 1959), pp. 911–917.
- [138] M. Lange et al. "Liquid Chromatography Techniques in Lipidomics Research". In: *Chromatographia* 82 (1 2019), pp. 77–100. ISSN: 1612-1112. DOI: 10.1007/S10337-018-3656-4.

- [139] H. C. Köfeler et al. "Mass Spectrometry Based Lipidomics: An Overview of Technological Platforms". In: *Metabolites* 2 (1 2012), pp. 19–38. ISSN: 2218-1989. DOI: 10.3390/METAB02010019.
- [140] A. Criscuolo et al. "Rational selection of reverse phase columns for high throughput LC–MS lipidomics". In: *Chemistry and Physics of Lipids* 221 (2019), pp. 120–127. ISSN: 0009-3084. DOI: 10.1016/J.CHEMPHYSLIP.2019.03.006.
- [141] K. Sandra and P. Sandra. "Lipidomics from an analytical perspective". In: *Current Opinion in Chemical Biology* 17 (5 2013), pp. 847–853. ISSN: 1367-5931. DOI: 10.1016/J.CBPA.2013.06.010.
- [142] W. J. Griffiths et al. "CHAPTER 1 Lipidomics Basics". In: *New Developments in Mass Spectrometry* (2020), pp. 1–24. ISSN: 20457553. DOI: 10.1039/9781788013109-00001.
- [143] G. Paglia et al. "Applications of ion-mobility mass spectrometry for lipid analysis". In: *Analytical and Bioanalytical Chemistry* 407 (17 2015), pp. 4995–5007. ISSN: 16182650. DOI: 10.1007/S00216-015-8664-8/FIGURES/8.
- [144] M. Kliman, J. C. May, and J. A. McLean. "Lipid analysis and lipidomics by structurally selective ion mobility-mass spectrometry". In: *Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids* 1811 (11 2011), pp. 935–945. ISSN: 1388-1981. DOI: 10.1016/J.BBALIP.2011.05.016.
- [145] C. G. Vasilopoulou et al. "Trapped ion mobility spectrometry and PASEF enable in-depth lipidomics from minimal sample amounts". In: *Nature Communications* 11 (1 2020), pp. 1–11. ISSN: 2041-1723. DOI: 10.1038/s41467-019-14044-x.
- [146] *The Nobel Prize in Chemistry 1964 - Summary*. URL: <https://www.nobelprize.org/prizes/chemistry/2002/summary/> (visited on 08/22/2022).
- [147] L. Konermann et al. "Unraveling the mechanism of electrospray ionization". In: *Analytical Chemistry* 85 (1 2013), pp. 2–9. ISSN: 00032700. DOI: 10.1021/AC302789C.
- [148] J. Draper et al. "Metabolite signal identification in accurate mass metabolomics data with MZedDB, an interactive m/z annotation tool utilising predicted ionisation behaviour 'rules'". In: *BMC Bioinformatics* 10 (1 2009), pp. 1–16. ISSN: 14712105. DOI: 10.1186/1471-2105-10-227/FIGURES/4.
- [149] M. Wilm. "Principles of Electrospray Ionization". In: *Molecular & Cellular Proteomics* 10 (7 2011), p. M111.009407. ISSN: 1535-9476. DOI: 10.1074/MCP.M111.009407.
- [150] K. Dreisewerd. "The desorption process in MALDI". In: *Chemical Reviews* 103 (2 2003), pp. 395–425. ISSN: 00092665. DOI: 10.1021/CR010375I.
- [151] R. Zenobi and R. Knochenmuss. "Ion formation in MALDI mass spectrometry". In: *Mass Spectrometry Reviews* 17 (5 1998), pp. 337–366. DOI: 10.1002/(SICI)1098-2787(1998)17:5%3C337::AID-MAS2%3E3.0.CO;2-S.
- [152] K. Ščupáková et al. "Spatial Systems Lipidomics Reveals Nonalcoholic Fatty Liver Disease Heterogeneity". In: *Analytical Chemistry* 90 (8 2018), pp. 5130–5138. ISSN: 15206882. DOI: 10.1021/ACS.ANALCHEM.7B05215.

- [153] K. V. Djambazova et al. "Resolving the Complexity of Spatial Lipidomics Using MALDI TIMS Imaging Mass Spectrometry". In: *Analytical Chemistry* 92 (19 2020), pp. 13290–13297. ISSN: 15206882. DOI: 10.1021/ACS.ANALCHEM.0C02520.
- [154] P. E. Miller and M. B. Denton. "The quadrupole mass filter: basic operating concepts". In: *Journal of chemical education* 63 (7 1986), p. 617.
- [155] R. A. Zubarev and A. Makarov. "Orbitrap mass spectrometry". In: *Analytical Chemistry* 85 (11 2013), pp. 5288–5296. ISSN: 00032700. DOI: 10.1021/AC4001223.
- [156] H. C. Lee and T. Yokomizo. "Applications of mass spectrometry-based targeted and non-targeted lipidomics". In: *Biochemical and Biophysical Research Communications* 504 (3 2018), pp. 576–581. ISSN: 0006-291X. DOI: 10.1016/J.BBRC.2018.03.081.
- [157] R. A. Yost and C. G. Enke. "Selected Ion Fragmentation with a Tandem Quadrupole Mass Spectrometer". In: *Journal of the American Chemical Society* 100 (7 1978), pp. 2274–2275. ISSN: 15205126. DOI: 10.1021/JA00475A072.
- [158] J. P. Koelmel et al. "Expanding Lipidome Coverage Using LC-MS/MS Data-Dependent Acquisition with Automated Exclusion List Generation". In: *Journal of the American Society for Mass Spectrometry* 28 (5 2017), pp. 908–917. ISSN: 18791123. DOI: 10.1007/S13361-017-1608-0.
- [159] S. Naz et al. "Development of a Liquid Chromatography-High Resolution Mass Spectrometry Metabolomics Method with High Specificity for Metabolite Identification Using All Ion Fragmentation Acquisition". In: *Analytical Chemistry* 89 (15 2017), pp. 7933–7942. ISSN: 15206882. DOI: 10.1021/ACS.ANALCHEM.7B00925.
- [160] L. Krasny et al. "SWATH mass spectrometry as a tool for quantitative profiling of the matrisome". In: *Journal of Proteomics* 189 (2018), pp. 11–22. ISSN: 1874-3919. DOI: 10.1016/J.JPROT.2018.02.026.
- [161] L. C. Gillet et al. "Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: A new concept for consistent and accurate proteome analysis". In: *Molecular and Cellular Proteomics* 11 (6 2012). ISSN: 15359484. DOI: 10.1074/MCP.0111.016717.
- [162] M. Raetz, R. Bonner, and G. Hopfgartner. "SWATH-MS for metabolomics and lipidomics: critical aspects of qualitative and quantitative analysis". In: *Metabolomics* 16 (6 2020), pp. 1–14. ISSN: 15733890. DOI: 10.1007/S11306-020-01692-0.
- [163] J. Hartler et al. "Lipid Data Analyzer: unattended identification and quantitation of lipids in LC-MS data". In: *Bioinformatics* 27 (4 2011), pp. 572–577. ISSN: 1367-4803. DOI: 10.1093/BIOINFORMATICS/BTQ699.
- [164] P. Husen et al. "Analysis of Lipid Experiments (ALEX): A Software Framework for Analysis of High-Resolution Shotgun Lipidomics Data". In: *PLOS ONE* 8 (11 2013), e79736. ISSN: 1932-6203. DOI: 10.1371/JOURNAL.PONE.0079736.

- [165] S. R. Ellis et al. "Automated, parallel mass spectrometry imaging and structural identification of lipids". In: *Nature Methods* 15 (7 2018), pp. 515–518. ISSN: 1548-7105. DOI: 10.1038/s41592-018-0010-6.
- [166] R. Herzog et al. "LipidXplorer: A Software for Consensual Cross-Platform Lipidomics". In: *PLOS ONE* 7 (1 2012), e29851. ISSN: 1932-6203. DOI: 10.1371/JOURNAL.PONE.0029851.
- [167] H. Tsugawa et al. "A lipidome atlas in MS-DIAL 4". In: *Nature Biotechnology* 38 (10 2020), pp. 1159–1163. ISSN: 1546-1696. DOI: 10.1038/s41587-020-0531-2.
- [168] J. Hartler et al. "Deciphering lipid structures based on platform-independent decision rules". In: *Nature Methods* 14 (12 2017), pp. 1171–1174. ISSN: 1548-7105. DOI: 10.1038/nmeth.4470.
- [169] J. P. Koelmel et al. "LipidMatch: An automated workflow for rule-based lipid identification using untargeted high-resolution tandem mass spectrometry data". In: *BMC Bioinformatics* 18 (1 2017), pp. 1–11. ISSN: 14712105. DOI: 10.1186/S12859-017-1744-3.
- [170] Z. Ni et al. "LipidHunter Identifies Phospholipids by High-Throughput Processing of LC-MS and Shotgun Lipidomics Datasets". In: *Analytical Chemistry* 89 (17 2017), pp. 8800–8807. ISSN: 15206882. DOI: 10.1021/ACS.ANALCHEM.7B01126.
- [171] J. P. Koelmel et al. "Software tool for internal standard based normalization of lipids, and effect of data-processing strategies on resulting values". In: *BMC Bioinformatics* 20 (1 2019), pp. 1–13. ISSN: 14712105. DOI: 10.1186/S12859-019-2803-8.
- [172] J. C. Venter et al. "The sequence of the human genome". In: *Science* 291 (5507 2001), pp. 1304–1351. ISSN: 0036-8075. DOI: 10.1126/SCIENCE.1058040.
- [173] S. Nurk et al. "The complete sequence of a human genome". In: *Science* 376 (6588 2022), pp. 44–53. ISSN: 0036-8075. DOI: 10.1126/SCIENCE.ABJ6987.
- [174] G. Abraham and M. Inouye. "Genomic risk prediction of complex human disease and its clinical application". In: *Current Opinion in Genetics & Development* 33 (2015), pp. 10–16. ISSN: 0959-437X. DOI: 10.1016/J.GDE.2015.06.005.
- [175] F. Sanger, S. Nicklen, and A. R. Coulson. "DNA sequencing with chain-terminating inhibitors." In: *Proceedings of the National Academy of Sciences of the United States of America* 74 (12 1977), pp. 5463–5467. ISSN: 00278424. DOI: 10.1073/PNAS.74.12.5463.
- [176] L. Liu et al. "Comparison of next-generation sequencing systems". In: *Journal of Biomedicine and Biotechnology* 2012 (2012). ISSN: 11107243. DOI: 10.1155/2012/251364.
- [177] Z. Wang, M. Gerstein, and M. Snyder. "RNA-Seq: a revolutionary tool for transcriptomics". In: *Nature reviews Genetics* 10 (1 2009), p. 57. ISSN: 14710056. DOI: 10.1038/NRG2484.
- [178] T. Bürckstümmer et al. "An efficient tandem affinity purification procedure for interaction proteomics in mammalian cells". In: *Nature Methods* 3 (12 2006), pp. 1013–1019. ISSN: 1548-7105. DOI: 10.1038/nmeth968.

- [179] S. K. Bharti and R. Roy. "Quantitative ^1H NMR spectroscopy". In: *TrAC Trends in Analytical Chemistry* 35 (2012), pp. 5–26. ISSN: 0165-9936. DOI: 10.1016/J.TRAC.2012.02.007.
- [180] Z. Pan and D. Raftery. "Comparing and combining NMR spectroscopy and mass spectrometry in metabolomics". In: *Analytical and Bioanalytical Chemistry* 387 (2 2007), pp. 525–527. ISSN: 16182642. DOI: 10.1007/S00216-006-0687-8/FIGURES/1.
- [181] *Isotopic Abundance of Carbon Atoms : SHIMADZU (Shimadzu Corporation)*. URL: https://www.shimadzu.com/an/service-support/technical-support/analysis-basics/gcms/fundamentals/whatcarbon_atoms.html (visited on 04/18/2022).
- [182] S. Niedenführ, W. Wiechert, and K. Nöh. "How to measure metabolic fluxes: a taxonomic guide for ^{13}C fluxomics". In: *Current Opinion in Biotechnology* 34 (2015), pp. 82–90. ISSN: 0958-1669. DOI: 10.1016/J.COPBIO.2014.12.003.
- [183] C. E. Meacham and S. J. Morrison. "Tumour heterogeneity and cancer cell plasticity". In: *Nature* 501 (7467 2013), pp. 328–337. ISSN: 1476-4687. DOI: 10.1038/nature12624.
- [184] A. A. Kolodziejczyk et al. "The Technology and Biology of Single-Cell RNA Sequencing". In: *Molecular Cell* 58 (4 2015), pp. 610–620. ISSN: 1097-2765. DOI: 10.1016/J.MOLCEL.2015.04.005.
- [185] A. Zeisel et al. "Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq". In: *Science* 347 (6226 2015), pp. 1138–1142. ISSN: 10959203. DOI: 10.1126/SCIENCE.AAA1934.
- [186] D. A. Lawson et al. "Single-cell analysis reveals a stem-cell program in human metastatic breast cancer cells". In: *Nature* 526 (7571 2015), pp. 131–135. ISSN: 1476-4687. DOI: 10.1038/nature15260.
- [187] L. Rappez et al. "SpaceM reveals metabolic states of single cells". In: *Nature Methods* 18 (7 2021), pp. 799–805. ISSN: 1548-7105. DOI: 10.1038/s41592-021-01198-0.
- [188] K. Ščupáková et al. "Spatial Systems Lipidomics Reveals Nonalcoholic Fatty Liver Disease Heterogeneity". In: *Analytical Chemistry* 90 (8 2018), pp. 5130–5138. ISSN: 15206882. DOI: 10.1021/ACS.ANALCHEM.7B05215.
- [189] D. Noble. "The rise of computational biology". In: *Nature Reviews Molecular Cell Biology* 3 (6 2002), pp. 459–463. ISSN: 1471-0080. DOI: 10.1038/nrm810.
- [190] F. Raimundo et al. "Machine learning for single-cell genomics data analysis". In: *Current Opinion in Systems Biology* 26 (2021), pp. 64–71. ISSN: 2452-3100. DOI: 10.1016/J.COISB.2021.04.006.
- [191] F. Markowetz. "All biology is computational biology". In: *PLOS Biology* 15 (3 2017), e2002050. ISSN: 1545-7885. DOI: 10.1371/JOURNAL.PBIO.2002050.
- [192] H. Kitano. "Systems biology: A brief overview". In: *Science* 295 (5560 2002), pp. 1662–1664. ISSN: 00368075. DOI: 10.1126/SCIENCE.1069492.
- [193] F. J. Bruggeman and H. V. Westerhoff. "The nature of systems biology". In: *Trends in Microbiology* 15 (1 2007), pp. 45–50. ISSN: 0966-842X. DOI: 10.1016/J.TIM.2006.11.003.

- [194] H. Kitano. "Computational systems biology". In: *Nature* 420 (6912 2002), pp. 206–210. ISSN: 1476-4687. DOI: 10.1038/nature01254.
- [195] M. Flores et al. "P4 medicine: how systems medicine will transform the healthcare sector and society". In: *Personalized Medicine* 10 (6 2013), pp. 565–576. ISSN: 17410541. DOI: 10.2217/PME.13.57.
- [196] J. Bousquet et al. "Systems medicine and integrated care to combat chronic noncommunicable diseases". In: *Genome Medicine* 3 (7 2011), pp. 1–12. ISSN: 1756994X. DOI: 10.1186/GM259/FIGURES/4.
- [197] L. Hood and M. Flores. "A personal view on systems medicine and the emergence of proactive P4 medicine: predictive, preventive, personalized and participatory". In: *New Biotechnology* 29 (6 2012), pp. 613–624. ISSN: 1871-6784. DOI: 10.1016/J.NBT.2012.03.004.
- [198] A. W. Goldman et al. "Bioregulatory systems medicine: An innovative approach to integrating the science of molecular networks, inflammation, and systems biology with the patient's autoregulatory capacity?" In: *Frontiers in Physiology* 6 (Aug 2015), p. 225. ISSN: 1664042X. DOI: 10.3389/FPHYS.2015.00225/BIBTEX.
- [199] Student. "The probable error of a mean". In: *Biometrika* (1908), pp. 1–25.
- [200] L. McInnes, J. Healy, and J. Melville. "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction". In: *arXiv* (2018). DOI: 10.48550/arxiv.1802.03426.
- [201] L. Van der Maaten and G. Hinton. "Visualizing data using t-SNE." In: *Journal of machine learning research* 9.11 (2008).
- [202] L. Cantini et al. "Benchmarking joint multi-omics dimensionality reduction approaches for the study of cancer". In: *Nature Communications* 12 (1 2021), pp. 1–12. ISSN: 2041-1723. DOI: 10.1038/s41467-020-20430-7.
- [203] M. Lotfollahi et al. "Mapping single-cell data to reference atlases by transfer learning". In: *Nature Biotechnology* 40 (1 2021), pp. 121–130. ISSN: 1546-1696. DOI: 10.1038/s41587-021-01001-7.
- [204] M. Kang, E. Ko, and T. B. Mersha. "A roadmap for multi-omics data integration using deep learning". In: *Briefings in Bioinformatics* 23 (1 2022). ISSN: 14774054. DOI: 10.1093/BIB/BBAB454.
- [205] L. Rokach and O. Maimon. "Clustering Methods". In: *Data Mining and Knowledge Discovery Handbook* (2005), pp. 321–352. DOI: 10.1007/0-387-25465-X_15.
- [206] R. Kothari and D. Pitts. "On finding the number of clusters". In: *Pattern Recognition Letters* 20 (4 1999), pp. 405–416. ISSN: 0167-8655. DOI: 10.1016/S0167-8655(99)00008-2.
- [207] C. Wiwie, J. Baumbach, and R. Röttger. "Comparing the performance of biomedical clustering methods". In: *Nature Methods* 2015 12:11 12 (11 2015), pp. 1033–1038. ISSN: 1548-7105. DOI: 10.1038/nmeth.3583.

- [208] S. J. Larsen, H. H. Schmidt, and J. Baumbach. “De Novo and Supervised Endophenotyping Using Network-Guided Ensemble Learning”. In: *Network and Systems Medicine* 3 (1 2020), pp. 8–21. DOI: 10.1089/SYSM.2019.0008.
- [209] C. Wang, R. Machiraju, and K. Huang. “Breast cancer patient stratification using a molecular regularized consensus clustering method”. In: *Methods* 67 (3 2014), pp. 304–312. ISSN: 1046-2023. DOI: 10.1016/J.YMETH.2014.03.005.
- [210] G. Valdes et al. “MediBoost: a Patient Stratification Tool for Interpretable Decision Making in the Era of Precision Medicine”. In: *Scientific Reports* 6 (1 2016), pp. 1–8. ISSN: 2045-2322. DOI: 10.1038/srep37854.
- [211] S. C. Madeira and A. L. Oliveira. “Biclustering algorithms for biological data analysis: A survey”. In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 1 (1 2004), pp. 24–45. ISSN: 15455963. DOI: 10.1109/TCBB.2004.2.
- [212] V. S. Stel et al. “Survival Analysis II: Cox Regression”. In: *Nephron Clinical Practice* 119 (3 2011), pp. c255–c260. ISSN: 16602110. DOI: 10.1159/000328916.
- [213] A. G. Singal et al. “Machine Learning Algorithms Outperform Conventional Regression Models in Predicting Development of Hepatocellular Carcinoma”. In: *The American journal of gastroenterology* 108 (11 2013), p. 1723. ISSN: 00029270. DOI: 10.1038/AJG.2013.332.
- [214] C. X. Li et al. “Integration of multi-omics datasets enables molecular classification of COPD”. In: *European Respiratory Journal* 51 (5 2018). ISSN: 0903-1936. DOI: 10.1183/13993003.01930-2017.
- [215] Z. Zhang et al. “Deep learning in omics: a survey and guideline”. In: *Briefings in Functional Genomics* 18 (1 2019), pp. 41–57. ISSN: 20412657. DOI: 10.1093/BFGP/ELY030.
- [216] J. Jumper et al. “Highly accurate protein structure prediction with AlphaFold”. In: *Nature* 596 (7873 2021), pp. 583–589. ISSN: 1476-4687. DOI: 10.1038/s41586-021-03819-2.
- [217] M. Leclercq et al. “Large-scale automatic feature selection for biomarker discovery in high-dimensional omics data”. In: *Frontiers in Genetics* 10 (MAY 2019), p. 452. ISSN: 16648021. DOI: 10.3389/FGENE.2019.00452.
- [218] F. Rohart et al. “mixOmics: An R package for ‘omics feature selection and multiple data integration”. In: *PLOS Computational Biology* 13 (11 2017), e1005752. ISSN: 1553-7358. DOI: 10.1371/JOURNAL.PCBI.1005752.
- [219] J. G. Greener et al. “A guide to machine learning for biologists”. In: *Nature Reviews Molecular Cell Biology* 23 (1 2021), pp. 40–55. ISSN: 1471-0080. DOI: 10.1038/s41580-021-00407-0.
- [220] A. Acharjee et al. “Comparison of regularized regression methods for omics data”. In: *Metabolomics* 3 (3 2013), p. 1.

- [221] A. Fukushima et al. "Integrated omics approaches in plant systems biology". In: *Current Opinion in Chemical Biology* 13 (5-6 2009), pp. 532–538. ISSN: 1367-5931. DOI: 10.1016/J.CBPA.2009.09.022.
- [222] S. Epskamp and E. I. Fried. "A tutorial on regularized partial correlation networks." In: *Psychological Methods* 23 (4 2018), p. 617. ISSN: 1939-1463. DOI: 10.1037/MET0000167.
- [223] X. Hu et al. "Integration of single-cell multi-omics for gene regulatory network inference". In: *Computational and Structural Biotechnology Journal* 18 (2020), pp. 1925–1938. ISSN: 2001-0370. DOI: 10.1016/J.CSBJ.2020.06.033.
- [224] M. Kanehisa et al. "KEGG: integrating viruses and cellular organisms". In: *Nucleic Acids Research* 49 (D1 2021), pp. D545–D551. ISSN: 0305-1048. DOI: 10.1093/NAR/GKAA970.
- [225] B. Snel et al. "STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene". In: *Nucleic acids research* 28 (18 2000), pp. 3442–3444. ISSN: 1362-4962. DOI: 10.1093/NAR/28.18.3442.
- [226] R. Alcántara et al. "Rhea—a manually curated resource of biochemical reactions". In: *Nucleic Acids Research* 40 (D1 2012), pp. D754–D760. ISSN: 0305-1048. DOI: 10.1093/NAR/GKR1126.
- [227] A. Subramanian et al. "Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles". In: *Proceedings of the National Academy of Sciences* 102 (43 2005), pp. 15545–15550. ISSN: 00278424. DOI: 10.1073/PNAS.0506580102.
- [228] C. Backes et al. "GeneTrail—advanced gene set enrichment analysis". In: *Nucleic Acids Research* 35 (suppl_2 2007), W186–W192. ISSN: 0305-1048. DOI: 10.1093/NAR/GKM323.
- [229] M. Kankainen et al. "MPEA—metabolite pathway enrichment analysis". In: *Bioinformatics* 27 (13 2011), pp. 1878–1879. ISSN: 1367-4803. DOI: 10.1093/BIOINFORMATICS/BTR278.
- [230] M. R. Molenaar et al. "LION/web: a web-based ontology enrichment tool for lipidomic data analysis". In: *GigaScience* 8 (6 2019), pp. 1–10. ISSN: 2047217X. DOI: 10.1093/GIGASCIENCE/GIZ061.
- [231] V. A. Padilha and R. J. Campello. "A systematic comparative evaluation of biclustering techniques". In: *BMC Bioinformatics* 18 (1 2017), pp. 1–25. ISSN: 14712105. DOI: 10.1186/S12859-017-1487-1/FIGURES/15.
- [232] A. Prelić et al. "A systematic comparison and evaluation of biclustering methods for gene expression data". In: *Bioinformatics* 22 (9 2006), pp. 1122–1129. ISSN: 1367-4803. DOI: 10.1093/BIOINFORMATICS/BTL060.
- [233] O. Lazareva et al. "On the limits of active module identification". In: *Briefings in Bioinformatics* 22 (5 2021). ISSN: 14774054. DOI: 10.1093/BIB/BBAB066.
- [234] X. Fang et al. "Genetic network and gene set enrichment analysis to identify biomarkers related to cigarette smoking and lung cancer". In: *Cancer Treatment Reviews* 39 (1 2013), pp. 77–88. ISSN: 0305-7372. DOI: 10.1016/J.CTRV.2012.06.001.

- [235] M. Murohashi et al. "Gene set enrichment analysis provides insight into novel signalling pathways in breast cancer stem cells". In: *British Journal of Cancer* 102 (1 2009), pp. 206–212. ISSN: 1532-1827. DOI: 10.1038/sj.bjc.6605468.
- [236] R. Heinrich and S. Schuster. *The regulation of cellular systems*. Springer Science & Business Media, 1996.
- [237] D. T. Gillespie. "A general method for numerically simulating the stochastic time evolution of coupled chemical reactions". In: *Journal of Computational Physics* 22 (4 1976), pp. 403–434. ISSN: 0021-9991. DOI: 10.1016/0021-9991(76)90041-3.
- [238] H. Link, D. Christodoulou, and U. Sauer. "Advancing metabolic models with kinetic information". In: *Current Opinion in Biotechnology* 29 (1 2014), pp. 8–14. ISSN: 0958-1669. DOI: 10.1016/J.COPBIO.2014.01.015.
- [239] K. A. Johnson and R. S. Goody. "The original Michaelis constant: Translation of the 1913 Michaelis-Menten Paper". In: *Biochemistry* 50 (39 2011), pp. 8264–8269. ISSN: 00062960. DOI: 10.1021/BI201284U/SUPPL_FILE/BI201284U_SI_001.PDF.
- [240] A. Bordbar et al. "Personalized Whole-Cell Kinetic Models of Metabolism for Discovery in Genomics and Pharmacodynamics". In: *Cell Systems* 1 (4 2015), pp. 283–292. ISSN: 2405-4712. DOI: 10.1016/J.CELS.2015.10.003.
- [241] K. van Eunen et al. "Biochemical Competition Makes Fatty-Acid beta-Oxidation Vulnerable to Substrate Overload". In: *PLOS Computational Biology* 9 (8 2013), e1003186. ISSN: 1553-7358. DOI: 10.1371/JOURNAL.PCBI.1003186.
- [242] M. M. Islam, W. L. Schroeder, and R. Saha. "Kinetic modeling of metabolism: Present and future". In: *Current Opinion in Systems Biology* 26 (2021), pp. 72–78. ISSN: 2452-3100. DOI: 10.1016/J.COISB.2021.04.003.
- [243] M. König and H. G. Holzhütter. "Kinetic modeling of human hepatic glucose metabolism in type 2 diabetes mellitus predicts higher risk of hypoglycemic events in rigorous insulin therapy". In: *Journal of Biological Chemistry* 287 (44 2012), pp. 36978–36989. ISSN: 00219258. DOI: 10.1074/JBC.M112.382069.
- [244] E. Benedetti et al. "A strategy to incorporate prior knowledge into correlation network cutoff selection". In: *Nature Communications* 11 (1 2020), pp. 1–12. ISSN: 2041-1723. DOI: 10.1038/s41467-020-18675-3.
- [245] O. Lazareva et al. "BiCoN: network-constrained biclustering of patients and omics data". In: *Bioinformatics* 37 (16 2021), pp. 2398–2404. ISSN: 1367-4803. DOI: 10.1093/BIOINFORMATICS/BTAA1076.
- [246] N. Alcaraz et al. "KeyPathwayMiner: Detecting Case-Specific Biological Pathways Using Expression Data". In: *Internet Mathematics* 7 (4 2011), pp. 299–313. ISSN: 15427951. DOI: 10.1080/15427951.2011.604548.
- [247] M. D. M. AbdulHameed et al. "Systems Level Analysis and Identification of Pathways and Networks Associated with Liver Fibrosis". In: *PLOS ONE* 9 (11 2014), e112193. ISSN: 1932-6203. DOI: 10.1371/JOURNAL.PONE.0112193.

- [248] M. D. Leiserson et al. "Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes". In: *Nature Genetics* 47 (2 2014), pp. 106–114. ISSN: 1546-1718. DOI: 10.1038/ng.3168.
- [249] F. Cheng et al. "Comprehensive characterization of protein–protein interactions perturbed by disease mutations". In: *Nature Genetics* 53 (3 2021), pp. 342–353. ISSN: 1546-1718. DOI: 10.1038/s41588-020-00774-y.
- [250] S. Klamt, O. Hädicke, and A. von Kamp. "Stoichiometric and constraint-based analysis of biochemical reaction networks". In: *Large-Scale Networks in Engineering and Life Sciences* 65 (2014), pp. 263–316. ISSN: 21643725. DOI: 10.1007/978-3-319-08437-4_5/TABLES/2.
- [251] K. J. Kauffman, P. Prakash, and J. S. Edwards. "Advances in flux balance analysis". In: *Current Opinion in Biotechnology* 14 (5 2003), pp. 491–496. ISSN: 0958-1669. DOI: 10.1016/J.COPBIO.2003.08.001.
- [252] J. D. Orth, I. Thiele, and B. O. Palsson. "What is flux balance analysis?" In: *Nature Biotechnology* 28 (3 2010), pp. 245–248. ISSN: 1546-1696. DOI: 10.1038/nbt.1614.
- [253] K. Raman, P. Rajagopalan, and N. Chandra. "Flux Balance Analysis of Mycolic Acid Pathway: Targets for Anti-Tubercular Drugs". In: *PLOS Computational Biology* 1 (5 2005), e46. ISSN: 1553-7358. DOI: 10.1371/JOURNAL.PCBI.0010046.
- [254] K. R. Shechter et al. "Metabolic memory underlying minimal residual disease in breast cancer". In: *Molecular Systems Biology* 17 (10 2021), e10141. ISSN: 1744-4292. DOI: 10.15252/MSB.202010141.
- [255] E. Murabito et al. "Monte-Carlo Modeling of the Central Carbon Metabolism of *Lactococcus lactis*: Insights into Metabolic Regulation". In: *PLOS ONE* 9 (9 2014), e106453. ISSN: 1932-6203. DOI: 10.1371/JOURNAL.PONE.0106453.
- [256] A. Wagner et al. "Metabolic modeling of single Th17 cells reveals regulators of autoimmunity". In: *Cell* 184 (16 2021), 4168–4185.e21. ISSN: 10974172. DOI: 10.1016/J.CELL.2021.05.045.
- [257] J. M. Lee, E. P. Gianchandani, and J. A. Papin. "Flux balance analysis in the era of metabolomics". In: *Briefings in Bioinformatics* 7 (2 2006), pp. 140–150. ISSN: 1467-5463. DOI: 10.1093/BIB/BBL007.
- [258] A. Mohamed, J. Molendijk, and M. M. Hill. "lipidr: a software tool for data mining and analysis of lipidomics datasets". In: *Journal of Proteome Research* (2020). ISSN: 1535-3893. DOI: 10.1021/acs.jproteome.0c00082.
- [259] C. Gaud et al. "BioPAN: a web-based tool to explore mammalian lipidome metabolic pathways on LIPID MAPS". In: *F1000Research* 10 (2021), p. 4. ISSN: 2046-1402. DOI: 10.12688/f1000research.28022.1.
- [260] C. Marella, A. E. Torda, and D. Schwudke. "The LUX Score: A Metric for Lipidome Homology". In: *PLOS Computational Biology* 11 (9 2015). Ed. by J. L. Reed, e1004511. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1004511.

- [261] S. Pushpakom et al. "Drug repurposing: progress, challenges and recommendations". In: *Nature Reviews Drug Discovery* 18 (1 2019), pp. 41–58. ISSN: 1474-1784. DOI: 10.1038/nrd.2018.168.
- [262] L. Sleire et al. "Drug repurposing in cancer". In: *Pharmacological Research* 124 (2017), pp. 74–91. ISSN: 1043-6618. DOI: 10.1016/J.PHRS.2017.07.013.
- [263] G. S. May et al. "The randomized clinical trial: bias in analysis." In: *Circulation* 64 (4 1981), pp. 669–673.
- [264] W. Han and L. Li. "Evaluating and minimizing batch effects in metabolomics". In: *Mass Spectrometry Reviews* 41 (3 2022), pp. 421–442. ISSN: 1098-2787. DOI: 10.1002/MAS.21672.
- [265] D. Lähnemann et al. "Eleven grand challenges in single-cell data science". In: *Genome Biology* 21 (1 2020), pp. 1–35. ISSN: 1474-760X. DOI: 10.1186/S13059-020-1926-6.
- [266] G. Stolovitzky, D. Monroe, and A. Califano. "Dialogue on Reverse-Engineering Assessment and Methods". In: *Annals of the New York Academy of Sciences* 1115 (1 2007), pp. 1–22. ISSN: 1749-6632. DOI: 10.1196/ANNALS.1407.021.
- [267] R. Nussinov. "Advancements and Challenges in Computational Biology". In: *PLOS Computational Biology* 11 (1 2015), e1004053. ISSN: 1553-7358. DOI: 10.1371/JOURNAL.PCBI.1004053.
- [268] R. Breitling, A. R. Pitt, and M. P. Barrett. "Precision mapping of the metabolome". In: *Trends in Biotechnology* 24 (12 2006), pp. 543–548. ISSN: 0167-7799. DOI: 10.1016/J.TIBTECH.2006.10.006.
- [269] V. D. Blondel et al. "Fast unfolding of communities in large networks". In: *Journal of Statistical Mechanics: Theory and Experiment* 2008 (10 2008). ISSN: 1742-5468. DOI: 10.1088/1742-5468/2008/10/P10008.
- [270] Z. Ni and M. Fedorova. "LipidLynxX: lipid annotations converter for large scale lipidomics and epilipidomics datasets". In: *bioRxiv* (2020), p. 2020.04.09.033894. DOI: 10.1101/2020.04.09.033894.
- [271] P. Bansal et al. "Rhea, the reaction knowledgebase in 2022". In: *Nucleic Acids Research* 50 (D1 2022), pp. D693–D700. ISSN: 0305-1048. DOI: 10.1093/NAR/GKAB1016.
- [272] B. Jassal et al. "The reactome pathway knowledgebase". In: *Nucleic Acids Research* 48 (D1 2020), pp. D498–D503. ISSN: 0305-1048. DOI: 10.1093/NAR/GKZ1031.
- [273] V. R. Thangapandi et al. "Loss of hepatic Mboat7 leads to liver fibrosis". In: *Gut* 70 (5 2021), pp. 940–950. ISSN: 0017-5749. DOI: 10.1136/GUTJNL-2020-320853.
- [274] M. List, P. Ebert, and F. Albrecht. "Ten Simple Rules for Developing Usable Software in Computational Biology". In: *PLOS Computational Biology* 13 (1 2017), e1005265. ISSN: 1553-7358. DOI: 10.1371/JOURNAL.PCBI.1005265.
- [275] B. Hanczar and M. Nadif. "Ensemble methods for biclustering tasks". In: *Pattern Recognition* 45 (11 2012), pp. 3938–3949. ISSN: 0031-3203. DOI: 10.1016/J.PATCOG.2012.04.010.

- [276] G. Aggarwal and N. Gupta. “BEMI bicluster ensemble using mutual information”. In: *Proceedings - 2013 12th International Conference on Machine Learning and Applications, ICMLA 2013* 1 (2013), pp. 321–324. DOI: 10.1109/ICMLA.2013.65.
- [277] L. Yin and Y. Liu. “Ensemble biclustering gene expression data based on the spectral clustering”. In: *Neural Computing and Applications* 30 (8 2018), pp. 2403–2416. ISSN: 09410643. DOI: 10.1007/S00521-016-2819-1/TABLES/7.
- [278] N. Köhler et al. “Kupffer cells are protective in alcoholic steatosis”. In: *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease* 1868 (6 2022), p. 166398. ISSN: 0925-4439. DOI: 10.1016/J.BBADIS.2022.166398.
- [279] E. Brunk et al. “Recon3D enables a three-dimensional view of gene variation in human metabolism”. In: *Nature Biotechnology* 36 (3 2018), pp. 272–281. ISSN: 1546-1696. DOI: 10.1038/nbt.4072.
- [280] Y. Yamanishi, J. P. Vert, and M. Kanehisa. “Supervised enzyme network inference from the integration of genomic data and chemical information”. In: *Bioinformatics* 21 (suppl_1 2005), pp. i468–i477. ISSN: 1367-4803. DOI: 10.1093/BIOINFORMATICS/BTI1012.
- [281] C. Peterson et al. “Inferring metabolic networks using the Bayesian adaptive graphical lasso with informative priors”. In: *Statistics and its interface* 6 (4 2013), p. 547. ISSN: 19387989. DOI: 10.4310/SII.2013.V6.N4.A12.
- [282] A. Pérez-Martí et al. “Reducing lipid bilayer stress by monounsaturated fatty acids protects renal proximal tubules in diabetes”. In: *eLife* 11 (2022). DOI: 10.7554/ELIFE.74391.
- [283] M. Borgmeyer et al. “Multiomics of synaptic junctions reveals altered lipid metabolism and signaling following environmental enrichment”. In: *Cell Reports* 37 (1 2021), p. 109797. ISSN: 2211-1247. DOI: 10.1016/J.CELREP.2021.109797.
- [284] B. Bao et al. “Correcting for sparsity and interdependence in glycomics by accounting for glycan biosynthesis”. In: *Nature Communications* 12 (1 2021), pp. 1–14. ISSN: 2041-1723. DOI: 10.1038/s41467-021-25183-5.
- [285] L. Zhang et al. “Clinical lipidomics in understanding of lung cancer: Opportunity and challenge”. In: *Cancer Letters* 470 (2020), pp. 75–83. ISSN: 0304-3835. DOI: 10.1016/J.CANLET.2019.08.014.
- [286] M. Buszewska-forajta et al. “Lipidomics as a diagnostic tool for prostate cancer”. In: *Cancers* 13 (9 2021), p. 2000. ISSN: 20726694. DOI: 10.3390/CANCERS13092000.
- [287] E. G. Armitage and A. D. Southam. “Monitoring cancer prognosis, diagnosis and treatment efficacy using metabolomics and lipidomics”. In: *Metabolomics* 12 (10 2016), pp. 1–15. ISSN: 15733890. DOI: 10.1007/S11306-016-1093-7.
- [288] P. J. Meikle et al. “Lipidomics: Potential role in risk prediction and therapeutic monitoring for diabetes and cardiovascular disease”. In: *Pharmacology & Therapeutics* 143 (1 2014), pp. 12–23. ISSN: 0163-7258. DOI: 10.1016/J.PHARMTHERA.2014.02.001.

- [289] D. Kopczynski et al. "Goslin 2.0 Implements the Recent Lipid Shorthand Nomenclature for MS-Derived Lipid Structures". In: *Analytical Chemistry* 94 (16 2022), pp. 6097–6101. ISSN: 15206882. DOI: 10.1021/ACS.ANALCHEM.1C05430.
- [290] E. Fahy et al. "LIPID MAPS online tools for lipid research". In: *Nucleic Acids Research* 35 (suppl_2 2007), W606–W612. ISSN: 0305-1048. DOI: 10.1093/NAR/GKM324.
- [291] L. Aimo et al. "The SwissLipids knowledgebase for lipid biology". In: *Bioinformatics* 31 (17 2015), pp. 2860–2866. ISSN: 1367-4803. DOI: 10.1093/BIOINFORMATICS/BTV285.
- [292] T. Hyötyläinen et al. "Lipidomics in biomedical research-practical considerations". In: *Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids* 1862 (8 2017), pp. 800–803. ISSN: 1388-1981. DOI: 10.1016/J.BBALIP.2017.04.002.
- [293] S. M. Lam, H. Tian, and G. Shui. "Lipidomics, en route to accurate quantitation". In: *Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids* 1862 (8 2017), pp. 752–761. ISSN: 1388-1981. DOI: 10.1016/J.BBALIP.2017.02.008.
- [294] H. C. Köfeler et al. "Recommendations for good practice in MS-based lipidomics". In: *Journal of Lipid Research* 62 (2021), p. 100138. ISSN: 0022-2275. DOI: 10.1016/J.JLR.2021.100138.
- [295] T. Baba et al. "Dissociation of Biomolecules by an Intense Low-Energy Electron Beam in a High Sensitivity Time-of-Flight Mass Spectrometer". In: *Journal of the American Society for Mass Spectrometry* 32 (8 2021), pp. 1964–1975. ISSN: 18791123. DOI: 10.1021/JASMS.0C00425.
- [296] *scverse*. URL: <https://scverse.org/> (visited on 08/11/2022).
- [297] J. K. Pauling and E. Klipp. "Computational Lipidomics and Lipid Bioinformatics: Filling In the Blanks". In: *Journal of Integrative Bioinformatics* 13 (1 2016), pp. 34–51. ISSN: 1613-4516. DOI: 10.1515/JIB-2016-299.
- [298] T. Baba et al. "Quantitative structural multiclass lipidomics using differential mobility: electron impact excitation of ions from organics (EIEIO) mass spectrometry". In: *Journal of Lipid Research* 59 (5 2018), pp. 910–919. ISSN: 15397262. DOI: 10.1194/JLR.D083261.
- [299] D. A. Los and N. Murata. "Structure and expression of fatty acid desaturases". In: *Biochimica et Biophysica Acta (BBA) - Lipids and Lipid Metabolism* 1394 (1 1998), pp. 3–15. ISSN: 0005-2760. DOI: 10.1016/S0005-2760(98)00091-5.
- [300] M. K. Andersen et al. "Spatial differentiation of metabolism in prostate cancer tissue by MALDI-TOF MSI". In: *Cancer & Metabolism* 9 (1 2021), pp. 1–13. ISSN: 2049-3002. DOI: 10.1186/S40170-021-00242-Z.
- [301] J. Tanevski et al. "Explainable multiview framework for dissecting spatial relationships from highly multiplexed data". In: *Genome Biology* 23 (1 2022), pp. 1–31. ISSN: 1474760X. DOI: 10.1186/S13059-022-02663-5.
- [302] H. Teng, Y. Yuan, and Z. Bar-Joseph. "Clustering spatial transcriptomics data". In: *Bioinformatics* 38 (4 2022), pp. 997–1004. ISSN: 1367-4803. DOI: 10.1093/BIOINFORMATICS/BTAB704.

- [303] A. Rao et al. "Exploring tissue architecture using spatial transcriptomics". In: *Nature* 596 (7871 2021), pp. 211–220. ISSN: 1476-4687. DOI: 10.1038/s41586-021-03634-9.
- [304] D. Sun et al. "STRIDE: accurately decomposing and integrating spatial transcriptomics using single-cell RNA sequencing". In: *Nucleic Acids Research* 50 (7 2022), e42–e42. ISSN: 0305-1048. DOI: 10.1093/NAR/GKAC150.
- [305] L. Capolupo et al. "Sphingolipids control dermal fibroblast heterogeneity". In: *Science* 376 (6590 2022). ISSN: 0036-8075. DOI: 10.1126/SCIENCE.ABH1623.
- [306] H. Liu et al. "Entropy-based consensus clustering for patient stratification". In: *Bioinformatics* 33 (17 2017), pp. 2691–2698. ISSN: 1367-4803. DOI: 10.1093/BIOINFORMATICS/BTX167.
- [307] P. Rajpurkar et al. "AI in health and medicine". In: *Nature Medicine* 28 (1 2022), pp. 31–38. ISSN: 1546-170X. DOI: 10.1038/s41591-021-01614-0.
- [308] K. B. Johnson et al. "Precision Medicine, AI, and the Future of Personalized Health Care". In: *Clinical and Translational Science* 14 (1 2021), pp. 86–93. ISSN: 1752-8062. DOI: 10.1111/CTS.12884.
- [309] J. C. Lindon and J. K. Nicholson. "The emergent role of metabolic phenotyping in dynamic patient stratification". In: *Expert Opinion on Drug Metabolism & Toxicology* 10 (7 2014), pp. 915–919. ISSN: 17447607. DOI: 10.1517/17425255.2014.922954.
- [310] J. R. Everett, R. L. Loo, and F. S. Pullen. "Pharmacometabonomics and personalized medicine". In: *Annals of Clinical Biochemistry* 50 (6 2013), pp. 523–545. ISSN: 17581001. DOI: 10.1177/0004563213497929.
- [311] H. Li, J. He, and W. Jia. "The influence of gut microbiota on drug metabolism and toxicity". In: *Expert Opinion on Drug Metabolism & Toxicology* 12 (1 2015), pp. 31–40. ISSN: 17447607. DOI: 10.1517/17425255.2016.1121234.
- [312] J. A. Kirwan et al. "Biobanking for Metabolomics and Lipidomics in Precision Medicine". In: *Clinical Chemistry* 65 (7 2019), pp. 827–832. ISSN: 0009-9147. DOI: 10.1373/CLINCHEM.2018.298620.
- [313] M. J. Sheller et al. "Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data". In: *Scientific Reports* 10 (1 2020), pp. 1–12. ISSN: 2045-2322. DOI: 10.1038/s41598-020-69250-1.
- [314] X. Han. "Lipidomics for precision medicine and metabolism: A personal view". In: *Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids* 1862 (8 2017), pp. 804–807. ISSN: 1388-1981. DOI: 10.1016/J.BBALIP.2017.02.012.
- [315] V. B. O'Donnell et al. "Lipidomics: Current state of the art in a fast moving field". In: *Wiley Interdisciplinary Reviews: Systems Biology and Medicine* 12 (1 2020), e1466. ISSN: 1939-005X. DOI: 10.1002/WSBM.1466.