Technische Universität München
TUM School of Life Sciences

Graph-Based Methods for the Integration of Lipidome and Metabolome Data into the Omics-Landscape

Nikolai Köhler

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

Prüfende der Dissertation:

1. TUM Junior Fellow Dr. Josch K. Pauling
2. Prof. Dr. Dimitrij Frischmann
3. Jun. Prof. Dr. Anne-Christin Hauschild

Die Dissertation wurde am 09.02.2024 bei der Technischen Universität München eingereicht und durch die TUM School of Life Sciences am 15.07.2024 angenommen.

Acknowledgments

First and foremost, I want to thank Dr. Josch Pauling for taking a chance on me. From early on in my doctorate, you gave me the freedom to explore my interests and integrate them into my work. I want to thank you for all the support, not only regarding our projects but also my personal growth. The last three years taught me lots and none of this would have been possible without you giving me this opportunity in the first place.

While this thesis is, of course, my work, science is nothing without a great environment to work in. Big thanks to everyone who has been part of LipiTUM and ExBio in the last three years for making this possible! Special thank you goes to Tim for the great time we had working on LINEX and discussing anything science-related. Another great thank you to Vivian, Cemil, and Maria, not only for all the fun in the office and at conferences but also for listening to all my complaints while finishing this thesis. At last, thank you so much, Martina Rüttger, for keeping so much paperwork away from us and always making sure we can focus on science.

Beyond our group, I was fortunate enough to collaborate with many experimental and computational researchers. Thank you all for introducing me to a variety of interesting science and allowing me to be a part of it. A special thank you to Smita Krishnaswamy and her lab, especially Edward, Dhananjay, Arman, Holly, and Scott, for welcoming me so warmly during my visit and teaching me so many things. You truly made the time special for me. Another special thanks to Bastian Grossenbacher-Rieck for many fun and inspirational discussions on science and your continued support.

A big thanks also to those who helped improve my thesis, in particular Edward, Olga, Tim, and Vivian.

Finally, I want to thank my parents and brother for always supporting me on my journey, no matter how spontaneous and unforeseeable my decisions may have been, and never losing trust in me. I cannot put into words how grateful I am to my wife, Jessi, not only for her love and support but also for her patience in sitting through my "exciting" science talks, long working hours, and stressed-out moods. Without you, I am sure I would not be at this point. Last but certainly not least, thank you to my daughter Ellie. You may not be able to communicate much at this point in your life, yet nothing in this world can give me the peace you do.

Abstract

Modern molecular biology experiments produce vast amounts of data. In a biomedical context, such data can be leveraged to improve our understanding of complex disease patterns and, thus, enable the development of more effective ways to diagnose and treat patients.

With changes in lifestyle in the modern world, non-communicable and metabolism-related diseases are posing a major burden to global health. To analyze the state of metabolism, the metabolome, the quantitative description of a large number of small biomolecules (metabolites) in an organism, can be used. Most commonly, the metabolome is measured by mass spectrometry. Building on the progress in mass spectrometry in the past decades, the metabolic phenotype can nowadays be captured better than ever. Despite the possibilities for such detailed measurements, computational methods for interpreting the resulting metabolome data are largely missing, creating a major bottleneck for basic and translational research. In the following publication-based doctoral thesis, computational solutions filling this gap in metabolomics data interpretation are introduced.

The first publication aims at the subdiscipline of metabolomics analyzing lipids - a special group of hydrophobic and amphiphilic metabolites - namely lipidomics. It introduces the *Lipid Network Explorer* (*LINEX*), a computational approach to generate biochemical lipid networks specific to a given dataset. Using these networks, it enables the analysis of global lipidome changes by combining them with statistical measures and enhances their functional interpretation.

In the subsequent publication, introducing *LINEX*^{2}, the network generation concept of *LINEX* is extended to comprise database-based metabolic reactions and connections to other omics disciplines. Furthermore, this publication presents a novel method for generating hypotheses on enzymatic dysregulation using lipid metabolic networks. Such hypotheses provide starting points for the mechanistic interpretation of lipidomics data and its integration with other types of molecular ("omics") data.

Finally, a preprint is presented introducing an algorithmic approach, termed *mantra*, to analyze metabolomics data with a focus on metabolic reactions. By providing a metric to approximate changes in metabolic reaction activity between biological conditions, it allows to focus on the alterations in metabolism that lead to differences in metabolome composition. The proposed metric is designed such that it can be integrated with other types of molecular data enabling functional hypotheses on the origin of dysregulated metabolic states.

In closing, this dissertation introduces new computational methods to analyze lipidomics and metabolomics data in a functional manner that also facilitates the integration of other omics disciplines. Ultimately, these improvements in functional interpretation significantly speed up the extraction of knowledge from biological and biomedical data. Thereby, they provide the basis for novel and more precise diagnostic and treatment procedures to combat the global disease burden.

Zusammenfassung

Moderne molekularbiologische Experimente produzieren immense Mengen und Daten. Im biomedizinischen Kontext können diese Daten dazu genutzt werden, um unser Verständnis komplexer Krankheitsmuster zu verbessern und ermöglichen damit die Entwicklung effektiverer Wege zur Diagnose und Behandlung von Patienten.

Durch Veränderungen des Lebensstils spielen nicht-übertragbare und Metabolismus-bezogene Krankheiten eine wichtige Rolle für die globale Krankheitslast in der modernen Welt. Um den Zustand des Metabolismus zu analysieren, kann das Metabolom, die quantitative Beschreibung einer großen Anzahl kleiner Biomoleküle (sogenannter Metabolite) in einem Organismus, genutzt werden. Typischerweise wird das Metabolom durch Massenspektrometrie gemessen. Durch den technischen Fortschritt in der Massenspektrometrie in den letzten Jahrzehnten kann der metabolische Phänotyp heutzutage besser denn je bestimmt werden. Trotz der Möglichkeiten für solch detailierte Messungen fehlen rechnergestützte Methoden zur Interpretation der erlangten Metabolom-Daten. Dadurch entsteht ein Nadelöhr für Grundlagen- und translationale Forschung. In der folgenden publikationsbasierten Dissertation werden rechnergestützte Methoden zur Interpretation von Metabolom-Daten vorgestellt, die diese Lücke füllen.

Die erste Publikation zielt auf die Lipidomik, eine Unterdisziplin der Metabolomik, die Lipide - eine spezielle Gruppe von hydrophoben und amphiphilen Molekülen - untersucht, ab. Sie stellt den *Lipid Network Explorer (LINEX)*, einen computergestützten Ansatz zur Generierung biochemischer Lipidnetzwerke spezifisch für einen gegebenen Datensatz, vor. Er erlaubt die Analyse globaler Lipidom-Änderungen durch Nutzung dieser Netzwerke in Kombination mit statistischen Metriken und verbessert deren funktionale Interpretation.

In der darauffolgenden Publikation, die *LINEX*^{2} vorstellt, wird das Konzept der Netzwerkerstellung aus LINEX durch den Einbezug datenbankgestützter metabolischer Reaktionen und Verbindungen zu anderen Omik-Disziplinen. Darüber hinaus präsentiert diese Publikation eine neue Methode zur Generierung von Hypothesen zu enzymatischer Dysregulation mithilfe von lipid-metabolischen Netzwerken. Solche Hypothesen dienen als Startpunkt für die mechanistische Interpretation von Lipidom-Daten und deren Integration mit andere Omik-Daten.

Als Drittes wird ein Preprint vorgestellt, der einen Algorithmus namens *mantra* zur Analyse von Metabolom-Daten mit einem Fokus auf metabolischen Reaktionen präsentiert. Indem er eine Metrik zur Approximation von Änderungen der Aktivität metabolischer Reaktionen zwischen biologischen Gruppen bietet, erlaubt er den Fokus auf die Veränderungen zu legen, die zu Unterschieden in der Metabolom-Komposition führen. Diese Metrik ist so gestaltet, dass sie direkt mit anderen molekularen Daten integriert werden kann, sodass eine funktionelle Hypothese zu den Ursprüngen dysregulierter metabolischer Stadien erstellt werden kann.

Zusammenfassend stellt diese Dissertation neue rechnergestützte Methoden zur Analyse von Lipidom- und Metabolom-Daten in einer funktionellen Art und Weise vor, die die Integration anderer Omik-Disziplinen ermöglicht. Diese Verbesserungen in der funktionellen Interpretation beschleunigen letztlich die Generierung von Wissen aus biologischen und biomedizinischen Daten. Dadurch bilden sie die Basis für neue, präzisere diagnostische und therapeutische Verfahren zur Bekämpfung der globalen Krankheitslast.

Publication Record

Peer-Reviewed Publications

- 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

- TD Rose[†] and N Köhler[†] et al. “Lipid network and moiety analysis for revealing enzymatic dysregulation and mechanistic alterations from lipidomics data”
  Briefings in Bioinformatics 2023, 24 (1), bbac572;
  doi: 10.1093/bib/bbac572

- 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

- TD Rose, T Bechtler, OA Ciora, KA Lilian Le, F Molnar, N Köhler et al. “MoSBi: Automated signature mining for molecular stratification and subtyping”
  Proceedings of the National Academy of Sciences 2022, 119 (16), e2118210119;
  doi: 10.1073/pnas.2118210119

- T Damiani, S Bonciarelli, GG Thallinger, N Köhler et al. “Software and Computational Tools for LC-MS-Based Epilipidomics: Challenges and Solutions”
  Analytical Chemistry 2023, 95 (1), 287-303;
  doi: 10.1021/acs.analchem.2c04406

- EM Haberl, TS Weiss, G Peschel, K Weigand, N Köhler et al. “Liver lipids of patients with hepatitis B and C and associated hepatocellular carcinoma”
  International Journal of Molecular Sciences 2021, 22 (10), 5297;
  doi: 10.3390/ijms22105297

- S Dieckmann, A Strohmeyer, M Willershäuser, SF Maurer, W Wurst, S Marschall, M Hrabe de Angelis, R Kühn, A Worthmann, MM Fuh, J Heeren, N Köhler et al. “Susceptibility to diet-induced obesity at thermoneutral conditions is independent of UCP1”
  American Journal of Physiology-Endocrinology and Metabolism 2022, 322 (2), E85-E100;
  doi: 10.1152/ajpendo.00278.2021

Preprints

- N Köhler et al. “Identification and Integration of Key-Metabolic Reactions from Untargeted Metabolomics Data”
  bioRxiv 2023
  doi: 10.1101/2023.05.15.540613

- F Hoheneder[†], CE Steidele[†], M Messerer, K Mayer, N Köhler et al. “Barley shows reduced Fusarium Head Blight under drought and modular expression of differential expressed genes under combined stress”
  bioRxiv 2023
  doi: 10.1101/2023.02.15.528674

- OI Coleman[†], A Sorbie[†], S Bierwirth, J Kövilein, M von Stern, N Köhler et al. “ATF6 activation alters colonic lipid metabolism causing tumor-associated microbial adaptation”
  bioRxiv 2023
  doi: 10.1101/2023.11.03.565267

- D Häcker[†], K Siebert[†], BJ Smith, N Köhler et al. “Exclusive Enteral Nutrition Initiates Protective Microbiome Changes to Induce Remission in Pediatric Crohn’s Disease”
  medRxiv 2023
  doi: 10.1101/2023.12.21.23300351

[†] These authors contributed equally

Contents

1 Motivation

Through progress in (bio)medical research, life expectancy has increased by around ten years from 1970 to 2021 in Germany [1]. However, in recent years the growth of life expectancy in Germany has fallen behind compared to other high-income countries, except for the U.S., where stagnation is observed despite the highest per-capita health expenditures [2, 3]. One aspect fueling this lag in longevity is underperforming primary care and disease prevention [2]. Consequently, improving early diagnosis and intervention treatments is a critical factor in enabling *affordable* long-term health.

Caused by shifts in lifestyle, especially diet and physical activity, non-infectious diseases nowadays account for the largest share of the disease burden, being responsible for 3 out of 4 deaths [4, 5]. One major class of non-infectious diseases are so-called metabolic diseases [6]. These are diseases involving dysregulation of metabolism, such as Type 2 Diabetes (T2D), hypertension, Non-Alcoholic Fatty Liver Disease (NAFLD), and obesity. Due to the high relevance of metabolic diseases and the need to improve prevention, diagnostics, and treatment, studying dysregulations of metabolism is an important factor for (bio)medical research. In particular, an in-depth understanding of how metabolic alterations contribute to the progression of diseases is essential for reducing the burden caused by metabolic diseases.

In order to get such mechanistic insights that allow treating causes instead of symptoms, molecular data describing the metabolic state, best reflected by the metabolome, is necessary [7]. Owing to progress in analytical technology, such measurements can nowadays be done in a high-throughput fashion. Consequently, the amounts of metabolic data available are far too large to be analyzed manually and computational methods extracting patterns and mechanistic hypotheses are a key factor for tackling the prevention, diagnosis, and treatment of metabolic diseases [8]. Despite this urgent need, algorithms concerned with the interpretation of metabolomics data lack the level of maturity present in algorithms designed for other types of molecular data. Therefore, the overall theme of this thesis is to develop methods that allow biological and biomedical researchers to better extract key insights from metabolomics data and speed up the process of generating validatable hypotheses that ultimately enable the discovery of new biomarkers and treatment targets.

In particular, this publication-based dissertation targets two specifically understudied areas of computational metabolomics: the functional interpretation of metabolomics data and multi-omics integration.

In the first publication introduced in this thesis, summarized in Section 4.1, I introduce a tool named Lipid Network Explorer (LINEX) [9] focusing on generating biochemical networks representing the metabolic connections between lipid species. Generating such networks enables a structured analysis of lipidomics data. By combining them with statistical measures and complex, interactive visualizations LINEX gives rise to quantitative lipidomics analyses that incorporate the biochemical connections between lipid species.

The second publication, summarized in Section 4.2, introduces an extension of the idea from [9] termed LINEX^2 [10]. Using metabolic reactions between lipid species, it allows the identification of metabolic reactions with the highest change in activity between biological conditions. This is especially relevant for biological and clinical researchers to enable faster and more interpretable interpretation of their data.

Together, these two publications offer a solution for the objective of delivering accessible computational methods that allow mechanistic interpretation of metabolism-related data and are thus a first step towards improved diagnostic and treatment options.

The third project, a preprint included as unpublished work (Section 5.1), presents Metabolic Network Reaction Analysis (mantra), a framework for the functional analysis of metabolic reaction activity [11]. This functional aspect enables the identification of metabolic reactions which may serve as drug target candidates or subtype-biomarkers. Furthermore, mantra is designed to integrate data from other omics disciplines, such as transcriptomics or metagenomics, to generate more fine-grained hypotheses on the mechanisms behind altered metabolic activity. Thereby, it targets both objectives of making metabolomics data more functionally interpretable and integratable.

Outline of the Thesis

The following chapter will introduce general concepts of molecular biology and computer science to provide the background necessary to understand the methods of this thesis and the greater context of their relevance. The first section starts with an introduction to the layers of molecular biology, which are integrated with multi-omics approaches, and the basics of metabolism, crucial to understanding the design of both LINEX and mantra. Subsequently, I will give an overview of the analytical techniques used to produce the data analyzed in this thesis. This knowledge is particularly important when discussing the limitations originating from them and how future developments will expand the applicability of my work. The second section provides the background on computational and mathematical techniques used in the publications that this thesis is based on.

The subsequent chapters provide an overview of the methods I developed during my doctorate and the publications they are presented in.

Finally, the last chapter is composed of a critical classification of the presented methods with an in-depth discussion of how they advance the field and how they open up new possibilities for future research. To conclude, I provide an outlook on where the field is headed and how new developments in other areas of computational biology and computer science will impact computational metabolomics in the coming years.

2 Background

Computational biology is, as already obvious from its composed name, an inherently interdisciplinary area. As a result, a certain familiarity with biology *and* computer science concepts is necessary to comprehend methods in computational biology and understand their degree of novelty and impact. Therefore, the following chapter gives an introduction to the biological and computational concepts that were either used directly in or as inspiration for the work presented in this thesis and the background required to understand them. The first part will cover essential parts of molecular biology, including analytical techniques to collect high throughput data. Following is an introduction to common computer science and mathematical approaches used in computational biology, focusing on those relevant to computational metabolomics and lipidomics.

2.1 Metabolism and Molecular Biology

Molecular biology is a subdiscipline of biology concerned with processes happening on a microscopic level. It studies mechanisms involving all "layers" of cellular molecules and their interactions. The main layer this thesis is concerned with is metabolism, the extraction of energy from food, and the synthesis of so-called metabolites, molecules used in cellular processes.

While molecular biology nowadays lays the foundation for many breakthroughs in, e.g., biomedicine and food production, the field itself only evolved in the last century. This section briefly introduces major aspects of the field and different levels of information flow and signaling relevant to metabolomics.

2.1.1 The Central Dogma of Molecular Biology

The central dogma of molecular biology (Figure 2.1), first proclaimed in 1958 by Francis Crick (and published in 1970 [12]), is a key model describing the information flow in molecular biology. Based on the knowledge available at the time, the central dogma states that information can be transmitted from Deoxyribonucleic Acid (DNA) and Ribonucleic Acid (RNA), but not proteins[1]. More precisely, Crick divides all theoretically possible types of information transfer between different polymers (DNA, RNA, and proteins) into three categories:

I common with evidence

II rarely occurring (without evidence)

III not occurring

[1] The description of the dogma provided here is based on the article from 1970 [12], which clarifies the proposal made in a lecture in 1958, including new findings made in the meantime.

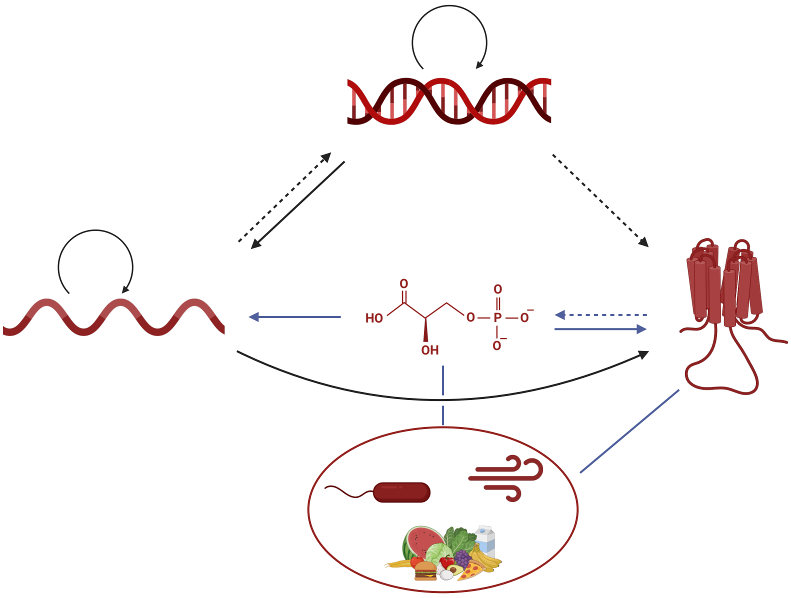


Figure 2.1: The central dogma of molecular biology as it was originally proposed by Crick [12] (black arrows) and additions based on modern knowledge of metabolism and the environment (blue lines). Solid black lines represent flows of information deemed as "common" while dashed lines represent "special" types of flow. Figure created with BioRender.

While different modifications and additions to the original dogma have been suggested [13, 14, 15, 16, 17], its original form is still widely known. Here, I will use the dogma to introduce the basic "layers" of molecular biology and their interactions before expanding the view to entities not covered by Crick. In light of this thesis' aim of developing methods that help to understand the molecular origin of diseases better, the dogma has a central role, as this goal can also be seen as trying to reconstruct aberrant information flows.

2.1.1.1 DNA

Deoxyribonucleic Acid (DNA) is a biomolecule made up of two chains of so-called *polynucleotides*. Nucleotides are (deoxy)ribose molecules with one to three phosphate groups at the carbon atom at position five ("C5") and a so-called *nucleobase* at the C1 position. DNA contains four different nucleobases: Adenin (A) and Thymine (T), and Cytosine (C) and Guanine (G). Two nucleobases each, called *complementary*, form a base pair through hydrogen bonds; A with T and C with G. Nucleotides, especially their triphosphate versions, will also be important players in metabolism, as we will see later. Each DNA strand is formed by polymerization from a condensation reaction between the phosphate group of one nucleotide at the C5 - the so-called 5' end - and the hydroxy group at the C3 - the 3' (pronounced 3 prime) end - of another nucleotide. While the exact positions may seem unimportant at this point, they are crucial for understanding the sequencing technologies introduced in Section 2.1.4 (Sequencing-Based Technologies).

The two strands, called forward and reverse strand, are oriented in opposite directions, and the bases at each position (with respect to the inverse directionalities) are complementary. The information in a DNA section is encoded by the sequence of base pairs. It can either be transferred in the classical sense of the central dogma, by replication or transcription, or through regulatory functions, such as forming binding sites for proteins that regulate transcriptional processes. While replication simply "copies" DNA molecules, transcription ("gene expression") transforms the information into RNA. As shown in Figure 2.1, there is also an inverse transformation to transcription, which turns RNA code into DNA. This process is called reverse transcription, a process utilized by retroviruses, which carry their genetic information as RNA and insert it into host cells as DNA.

The transcriptional process is tightly regulated, as the transcriptional program of a cell is extremely specific for each cell type, condition, signal, etc.. One aspect of regulation are the so-called transcription factors. These are proteins that promote or suppress gene expression by binding to specific DNA regions. Another way cells regulate transcription is by *epigenetic* factors. One of them is the methylation of cytosines, which can prevent regulators of transcription from binding [18] and takes part in mediating chromatin formation [19, 20]. Chromatin refers to a complex of DNA and specific proteins called histones, around which DNA strands are wrapped. While its main function is to provide a more densely packed form, post-translational modifications of histones are the second way of regulating gene expression epigenetically. Modifications causing tighter packaging in some areas ("Heterochromatin") make genes inaccessible, while less dense packaging produces open regions ("Euchromatin") available for replication and transcription.

2.1.1.2 RNA

Ribonucleic Acid (RNA) is a (mostly[2]) single-stranded polynucleotide, in contrast to the double-stranded DNA, that is also made up of a four-nucleotide "alphabet". Another biochemical difference to DNA is that Thymine is replaced with Uracil (U). This represents an interesting evolutionary trade-off, as Uracil is more energy efficient to produce than Thymine, the methylated form of Uracil, at the expense of having undetectable C → U mutations through spontaneous Cytosine-deamination [22]. Since a single RNA copy only leads to a limited number of translations, such occasional mutations are acceptable. DNA, on the other hand, needs to be as stable as possible as mutations persist and are even inherited at cell division[3]. Therefore, the additional cost of forming Thymine is justifiable for DNA but not for RNA.

While the main function of RNA is to carry the information for protein synthesis (the so-called messenger RNA (mRNA)), it has a diverse range of additional functions executed by different flavors of RNA, canonically referred to as non-coding RNA. Some of these functions come from the single-stranded nature that enables a diverse range of structures through base pairings within the same strand. To translate mRNA into proteins, transfer RNA (tRNA) and ribosomal RNA (rRNA) are required, making them essential players in the central dogma of molecular biology. Even though mRNA is often considered the main function of RNA, the cellular RNA content is heavily dominated by rRNA, comprising around 80%. Together with ribosomal proteins, the latter forms the so-called ribosome, a complex "translating" genetic code from a RNA sequence

[2] Double-stranded RNA (dsRNA) molecules also occur in nature, e.g., in viral infections, which the innate immune system can even recognize through dsRNA receptors [21]. However, single-stranded RNA is much more common.
[3] This is, of course, only relevant for non-terminally differentiated cells.

to a peptide sequence. This function is carried out by recruiting tRNA molecules, which carry a specific peptide on one end and a corresponding three-nucleotide sequence called “anticodon” on the other. The genetic code encoded by mRNA is divided into nucleotide triplets called “codons”, to which the anticodons are complementary. This way, the mRNA sequence encodes a specific peptide sequence.

In addition to translation-related non-coding functions, RNA can also mediate signaling, for example, through RNA interference (RNAi) - small interfering RNA (siRNA) molecules (20 to 24 bp long) adhering to complementary RNA sections forming dsRNA that is subsequently used as a marker for digestion [23].

2.1.1.3 Proteins

The terminal end of information flow in Crick’s central dogma is the level of proteins. Unlike nucleic acids, they are not made up of nucleotides, but instead of (20[4]proteinogenic) α-Amino Acids (AAs). These are molecules in which a carbon atom connects an amino group (“N-terminus”), a carboxy group (“C-terminus”), and a specific residue. AAs can be chained to peptides via a condensation reaction between the N-terminus of one and the C-terminus of another AA.

A protein’s linear AA sequence is referred to as the primary structure. Based on chemical and physical interactions between AAs, secondary structures are formed. Generally, secondary structures are divided into α-helices and β-sheets plus linker and unstructured regions. The 3-dimensional assembly of these secondary structures, a crucial factor for protein function, is referred to as the tertiary structure. Sometimes, multiple proteins associate and act as one large complex. This assembly of tertiary structures is then called quaternary structure.

Certain motifs occur frequently despite the immense diversity of protein sequences and structures within and between organisms. Usually, such motifs exhibit specific functions, such as anchoring proteins in membranes or catalyzing specific reactions like the phosphorylation of specific residues of other proteins or the breakage of specific molecular bonds. Thus, they can also be used to classify proteins into functional categories.

2.1.1.4 Metabolites

An entity not covered in the central dogma are metabolites. They are a chemically diverse class of (small) molecules generated by metabolism. Though never clearly stated by Crick[5], leaving metabolites out is a logical step when considering the sole idea of characterizing *genetic* information flow. Nevertheless, metabolites play an essential role. For one, they are required to build all three entities of the dogma and provide the energy for their synthesis. Furthermore, they can act as signaling molecules influencing the “expression” of DNA, RNA, and proteins, as well as their interactions.

The functions of metabolites are as diverse as their chemical and structural properties. Often considered as the main function of metabolism, metabolites are intermediates and storage

[4] In addition to the 20 “standard” proteinogenic AAs a special 21^{st} α-AA, called selenocysteine exists.

[5] To the best of my knowledge, there is no record of Francis Crick commenting on metabolites and metabolism as a potential part of the dogma.

products of energy taken up from food in the form of nucleoside phosphates, such as Adenosine Triphosphate (ATP) or Guanosine Triphosphate (GTP), Glycogen or Triacylglycerols (TGs). Other examples of metabolite functions are signal transduction in various ways, defense mechanisms (e.g., antimicrobial peptides), and even communication between individual organisms, such as pheromones or molecules involved in quorum sensing. A special class of hydrophobic and amphiphilic metabolites are lipids. They are essential for life as a whole due to their ability to form vesicles and membranes, which are used for compartmentalization to separate intra-cellular space from the outside. For a more detailed examination of metabolism and the role of metabolites and lipids, refer to Section 2.1.2 (Metabolism in Health and Disease).

2.1.1.5 Environmental Interactions

Taking the idea of adding new entities even further, one can also consider the influence of environmental factors. Such features can be both living organisms as well as nutrients or toxins taken up from the environment.

Interactions with the environment lead to mostly temporary, non-heritable responses. For example, the metabolites produced by the gut microbiome have regulatory effects on the host's immune system and even the nervous system[6][24, 25, 26]. Also, the recognition of (pathogenic) microbial antigens directly triggers an immune response [27], including alterations in genetic information flow.

Similar to microbial products, metabolites from nutrition can alter human gene expression [28, 29]. In addition to such short-term effects, it has been shown that conditions like malnutrition can lead to long-term alteration of the genetic information flow, e.g., through sustained differential methylation [30]. Even transgenerational, thus inheritable, alterations caused by starvation - the absence of essential metabolites - were observed in *Caenorhabditis elegans* [31], showing that environmental factors have a substantial influence on the genetic information used. One might add mutagenic molecules as a class that can directly alter genetic code. However, since these are usually unwanted effects organisms try to evade as much as possible, they are not considered here.

This expanded view on the central dogma of molecular biology assigns a central role to the metabolome as the level reflecting the phenotype most closely and an organism's interface with the environment. Thus the *metabotype* is an important reflector of the status of an organism.

2.1.2 Metabolism in Health and Disease

In general, metabolism is mostly seen as the process of breaking down nutrients from food and extracting the energy they contain, as well as synthesizing small molecules required to maintain a dynamic equilibrium (homeostasis). Metabolic reactions are biochemical processes modifying metabolites by adding or removing certain other metabolites. These reactions are usually *not* spontaneous, i.e. they require some form of catalyzation. This makes them energy-consuming but, at the same time, more controllable. To have an efficient way of controlling them, organisms use enzymes, proteins - or protein complexes - that bind to the respective metabolite(s) and catalyze the reaction in a so-called active site.

[6] The so-called gut-brain axis

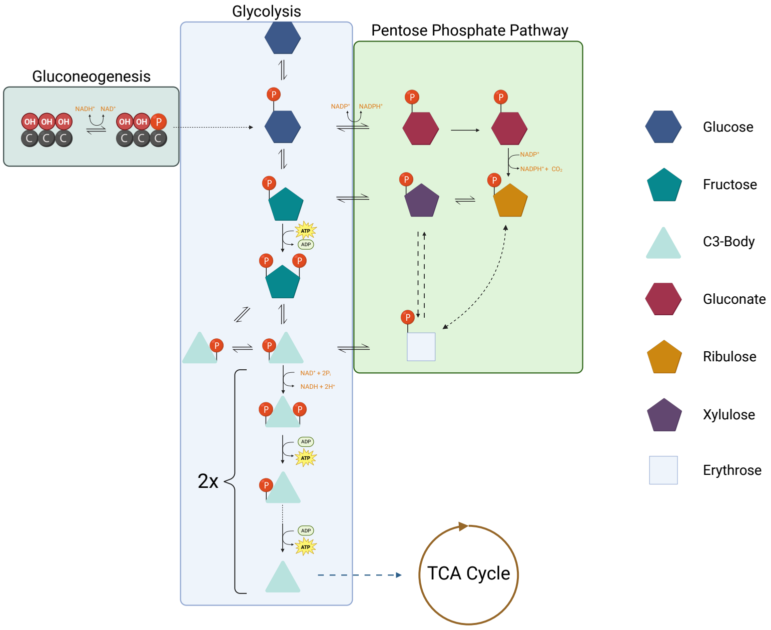


Figure 2.2: Overview of carbohydrate metabolism. The light blue box in the center represents Glycolysis; hexagons represent glucose, pentagons fructose and triangles C3-bodies. The green box on the right schematically shows Pentose Phosphate Pathway. The left box shows Gluconeogeneis. Figure created with BioRender.

Enzymes are ubiquitously present and we can observe their impact in everyday life, whether we chew some bread for too long and it suddenly starts tasting sweet or yogurt tastes bitter after being mixed with kiwi fruit or papaya. The former is caused by salivary amylases breaking down starch into its components - sugars. The latter are cysteine proteases digesting milk proteins. While these examples are single reactions, metabolic processes are usually complex, tightly controlled chains of reactions - referred to as pathways. Deregulation of these can have dramatic effects, leading to disorders and diseases. The deregulation of cellular energy metabolism, for example, is recognized as an emerging hallmark of cancer, a characteristic that *any* cancer needs for rapid growth [32].

The following section will give a more detailed introduction to the basic pathways of (energy-)metabolism and their relevance for the development of certain diseases. It is divided into three parts, one for each class of macronutrients.

When reading this section, please bear in mind that my descriptions are only scratching the surface of each pathway and one could write an entire thesis about researching a single metabolic reaction, let alone the interaction and co-regulation between pathways happening in each and every cell. That said, the following pages might be a good point to appreciate all the astounding capabilities of our bodies and the incredibly complex, fine-tuned machinery it is.

2.1.2.1 Carbohydrate Metabolism

Carbohydrates, organic compounds composed of chains of monosaccharides (sugar molecules), are a main source of energy. Mammals can both catabolize and synthesize these molecules. Catabolic activity, on the one hand, uses the energy stored in carbohydrate molecules to generate directly usable units of energy such as ATP or GTP and electron donors, e.g. Nicotinamide Adenine Dinucleotide (NADH) and Flavin Adenine Dinucleotide ($FADH_2$). The energy in ATP and GTP is stored in the phosphoric anhydride bonds. Hydrolysis of these bonds releases energy that can be used to fuel reactions with a positive free energy balance, i.e. energy-consuming reactions, to synthesize metabolites. The energy from electron transfer is used, for example, in electron transport chains such as the mitochondrial respiratory chain to generate a proton gradient between two sides of a membrane. This gradient can then power different processes, e.g. the synthesis of ATP. Anabolic activity, on the other hand, invests energy to produce mono- and polysaccharides.

Glucose Catabolism One of the major pathways to extract energy from carbohydrates is *glycolysis*. It converts one glucose molecule, a monosaccharide, into 2 pyruvate molecules and reduces 2 NAD+ to 2 NADH while generating 2 ATP molecules (net)[7].

Monosaccharides other than glucose can also be processed through this pathway by converting them into glucose- or fructose-6-phosphate, the first intermediates of the glycolysis pathway generated via hexokinase phosphorylation and glucose-6-Phosphate isomerase interconversion. Despite the positive net balance, the first step and third steps each consume one molecule of ATP per molecule of glucose. The resulting molecule, Fructose-1,6-biphosphate, can be broken down into 2 Glyceraldehyde-3-Phosphate (GAP), one of them via an isomerase reaction from Dihydroxyacetone, which can also be converted to glycerol. Subsequently, *each* GAP is phosphorylated with a *free* phosphate, reducing NAD+ and releasing a free proton. This bound phosphate group can then be used to generate the first molecule of ATP, and after an enolase reaction, the second phosphate group - coming from the first step of glycolysis - is transferred onto a molecule of ATP, essentially recovering the investment from the first/third step.

While glycolysis is generally *catabolic*, all but the phosphorylation steps are reversible. Three intermediates of the pathway, glucose-6-phosphate, fructose-6-phosphate, and GAP, are shared with the *Pentose Phosphate Pathway* (PPP). It is an *anabolic* pathway producing, among other molecules, ribose-5-phosphate, the basis for synthesizing nucleotides, and erythrose-4-phosphate, the basis for aromatic amino acid metabolism [33].

Under aerobic conditions, the product of glycolysis - pyruvate - is oxidized to acetyl-Coenzyme A (CoA), again reducing NAD+, and passed into the *Tricarboxylic Acid Cycle* (TCA cycle). As part of aerobic respiration, the TCA cycle is also one of the links between carbohydrate, lipid, and amino acid metabolism [34]. While it also produces 2 ATP per molecule glucose[8], the main aspect for energy, however, is the reduction of 6 NAD+ and 2 FAD. The resulting electron donors - 6

[7] Glycolysis produces 4 molecules of ATP per molecule of glucose in total. However, it also consumes 2 ATP in the first steps of the pathway.

[8] Technically, the TCA cycle produces 1 GTP per acetyl-CoA. However, GTP can easily be used to generate ATP by transferring one of its phosphate groups onto Adenosine Diphosphate (ADP), and each glucose molecule yields 2 acetyl-CoA, hence the 2 molecules of ATP.

NADH and 2 $FADH_2$ - can subsequently be used in the *Oxidative Phosphorylation* (OxPhos) to generate up to 34 molecules of ATP by building a proton gradient with the energy released from electron transfer, which then powers an ATP synthase.

Blood Glucose Regulation Both very high as well as very low blood glucose levels can be detrimental to humans. Therefore, glucose homeostasis tightly regulates the balance between pathways metabolizing glucose, such as glycolysis and glycogenesis, and those synthesizing glucose, such as glycogenolysis and gluconeogenesis [35]. *Glycogenesis* is the process of turning glucose monomers into glycogen, a polysaccharide to store glucose. *Glycogenolysis* refers to the reverse reaction. Both processes can occur in the liver, which acts as body-wide glucose storage by performing glycogenesis in high blood glucose situations and glycogenolysis when blood glucose is low, and the muscles, which exclusively use the generated glucose for themselves.

When its glycogen stores are depleted, for example, after a longer period of fasting, the liver can also activate *gluconeogenesis*, an anabolic pathway generating glucose from lactate, glycerol, or amino acids, thus connecting carbohydrate metabolism with lipid and amino acid metabolism. It allows the liver to maintain blood glucose levels by digesting lipids and proteins.

Central Carbon Metabolism in Disease Central Carbon Metabolism (CCM) - glycolysis, PPP, and the TCA cycle - alterations are associated with a wide range of disease conditions, such as cancer, Parkinson's Disease, or COVID-19 [36, 37, 38]. Arguably, the most prominent case of a metabolic alteration (potentially) paving the way for disease progression is the Warburg effect. Named after its discoverer, Noble Prize winner Otto Warburg, it describes an increase in glucose degradation and lactate production through fermentation, thus avoiding the TCA cycle and OxPhos, in tumor cells compared to "normal" cells [39, 40]. Under anaerobic conditions, with which tumors are often faced, this poses an advantage as oxidative stress is avoided. Surprisingly, at first sight, this change in metabolic activity is observed regardless of the presence of oxygen, even though it reduces the efficiency of ATP production. When considering more recent findings, especially with respect to the metabolic differences between quiescent and proliferating cells [41], it becomes clear that tumor cells also mimic the behavior of "normal" proliferating cells to maintain their high growth rates [42]. The reason why proliferation is accompanied by a bypass of the TCA cycle and OxPhos is that their anabolic activity uses intermediates of glycolysis as well as *cytosolic* acetyl-CoA[9]to synthesize lipids (Section 2.1.2.2), nucleotides, and amino acids (Section 2.1.2.3) [42].

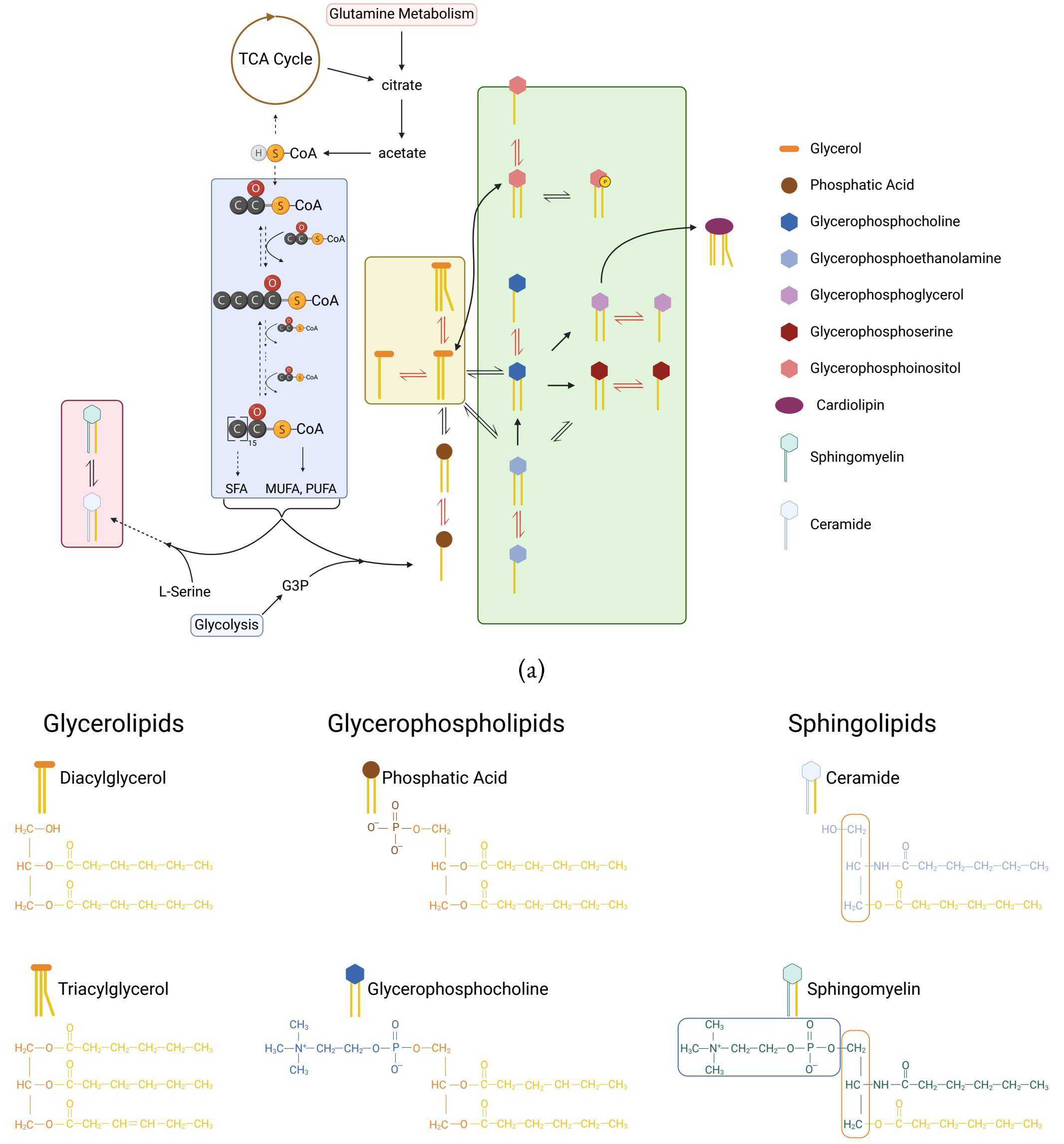
In breast cancer, an application case for the mantra paper presented in this thesis (Section 5.1), alteration in CCM have also been reported in a similar way as explained for tumors in general [43]. In addition, Richardson et al. [43] also report an increase in nucleotide hexoses, such as Uridine Diphosphate-Glucoronate (UDP-GlcN), synthesized from intermediates of glycolysis, the PPP, and glutamine degradation (see Section 2.1.2.3), in tumor cells. This is an interesting showcase of the intersection of altered metabolic activity and signaling, as UDP-hexoses are precursors for Hyaluronic Acid, which has numerous signaling functions and is important for metastatic activity [44].

[9] in contrast to *mitochondrial* acetyl-CoA in the TCA cycle

2.1.2.2 Lipid Metabolism

Lipids form a specific sub-class of metabolites that play a major role in numerous physiological processes. They are usually characterized as hydrophobic or amphiphilic metabolites. Despite their structural diversity, lipids are less diverse compared to the extremely high diversity *all* metabolites have as a whole.

Above all, they enable the main prerequisite for life - compartmentalization - due to their ability to form biological membranes in the form of mono- and bilayers. These are mostly composed of amphiphilic lipids, namely glycerophospholipids, sphingomyelins, and cholesterol (Figure 2.3a c and d). Another main function of lipids, arguably the most prominent one, is energy storage in the form of TGs. These are stored in lipid droplets inside the cytoplasm and can be used to generate energy and glucose when necessary. Dedicated tissue types exist to handle lipid storage, the so-called adipose tissue, which also constitutes our (not so beloved) body fat. Furthermore, lipids can also be important players in signaling cascades.



(a)

(b)

Figure 2.3: Schematic overview of lipid metabolism and examples of common complex lipid structures. *(a)* Overview of lipid metabolism. The blue depicts Fatty Acid metabolism from acetyl-CoA to Palmitic Acid (C16). Together with Glycerophosphate from Glycolysis, fatty acids are turned into Lyso-Phosphatidic Acid and Phosphatidic Acid. The yellow box shows Glycerolipid metabolism with Diacylglycerol at its center and connecting to Glycerophospholipid metabolism (green box). Cardiolipins are generated from combining two Phosphatidylglycerols and are thus also connected to Glycerophospholipid metabolism. Sphingolipid metabolism is shown in the red box on the left. Connections to carbohydrate and amino acid metabolism are shown at the top and the bottom. *(b)* Examples of complex lipid structures.

Figure 2.3: The first column shows Glycerolipids, Diacylglycerol and Triacylglycerol, respectively, with the glycerol group colored in orange (the color scheme matches a) and fatty acids in yellow. In the second column Phosphatidic Acid and Phosphatidylcholine are shown. Glycerol is again colored orange, the phosphate group in Phosphatidic Acid (PA) in brown, and the phosphocholine group of Phosphatidylcholine (PC) in blue. The last column shows Ceramide (Cer) and Sphingomyelin (SM) with their sphingosine colored in light blue and light green, respectively. The orange box indicates the carbon hydrogens originating from glycerol. The phosphocholine group in Sphingomyelin (SM) is indicated by the blue box. Figures created with BioRender.

Structurally, complex lipids are composed of a backbone, an (optional) head group, and one or multiple Fatty Acids (FAs) (Figure 2.3b). In reflection of this composition, the next paragraphs will first outline FA and, subsequently complex lipid metabolism.

Fatty Acid Metabolism The precursor of FA synthesis is acetyl-CoA, usually coming from mitochondria-exported citrate, that was produced in the TCA cycle, or cytoplasmic acetate (Figure 2.3a a). In the first step of FA biosynthesis, it is transformed to malonyl-CoA by carboxylation. The malonyl-moiety is subsequently elongated in steps of 2 carbon atoms by transferring the hydrocarbons from acetyl-CoA until palmitate is reached [45]. The elongation process is also the reason that FAs with an even number of carbon atoms are more common, as odd-chain FAs need to start synthesis from propionyl-CoA [46]. Palmitate is commonly annotated as "16:0" due to its hydrocarbon chain containing 16 carbon atoms and zero Double Bonds (DBs) between them. DBs are structurally very important properties, as they induce bends in the 3D structure of FAs. Such non-linear structures reduce the density in which FAs/complex lipids containing DBs can be packed, impacting, for example, membrane fluidity[10]. Furthermore, the altered molecule geometry causes differences in binding affinities, which make the number and position of DBs an important criterion for signaling cascades [47] and FA-preferences of certain types of complex lipid metabolism enzymes [48]. As FAs with no double bonds are called "saturated", the process of introducing DBs is referred to as "desaturation". In every desaturation step, two hydrogen atoms are transferred onto FAD, introducing a single DB. Because only certain FAs can be desaturated[11], some FAs need to be taken up by diet. The two prominent families of such FAs are ω-3 and ω-6 fatty acids, such as eicosapentaenoic (EPA) and docosahexaenoic acid (DHA) [50]. They are typically found in fish and vegetable oils and play an important role in the regulation of inflammatory processes.

To use lipids for energy storage, the human body also needs to be able to break down FAs into molecules that can be used to generate ATP. The main way of degrading FAs is β-Oxidation [51]. The process, happening predominantly in mitochondria, similar to FA synthesis, is a sequential removal of C2-bodies in the form of acetyl-CoA until only acetyl-CoA itself (or propionyl-CoA) is

[10] For example, the lipid "packing" in butter is high, due to a high degree of saturated FAs, making it solid at room temperature. Plant oil, on the contrary, is liquid at room temperature as it contains a higher rate of unsaturated FAs.

[11] For example, the KEGG [49] pathway for unsaturated FA synthesis lists only a total of 22 possible desaturation reactions.

left[12](Figure 2.3a a). β-Oxidation thereby fuels ATP production via the TCA cycle and oxidative phosphorylation [51]. Furthermore, in low-glucose scenarios, acetyl-CoA can also be used to produce ketone bodies, which can serve as a substitute for glucose in e.g. the brain and skeletal muscles [52, 53].

Lipid Class Metabolism The first steps in the *de-novo* synthesis of complex lipids, the so-called Kennedy pathway, are the transfers of acyl moieties from acyl-CoA onto Glycerol-3-Phosphate (G3P) followed by another acyl-transfer onto the reaction product, Lyso-Phosphatidic Acid (LPA), resulting in a PA [54]. Note that this is a generic *acyl* transfer instead of a specific fatty acyl transfer. This initial reaction links both fatty acid metabolism, which produces acyl-CoA, and carbohydrate metabolism, which produces dihydroxyacetone phosphate (see Section 2.1.2.1), a precursor for G3P, to complex lipid metabolism. PA serves as the starting point for synthesizing two major lipid categories: glycero- and glycerophospholipids.

Glycerolipid anabolism via G3P begins with the dephosphorylation of PA to Diacylglycerol (DG) [55]. Additionally, DG can be synthesized via acylation of Monoacylglycerol (MG), usually coming from the breakdown of dietary nutrients [56, 57, 58]. Since glycerol carries three hydroxy groups and DG consists of two FAs esterified to glycerol, another FA can be bound, which results in a TG. For a more comprehensive review of glycerolipid synthesis, see [54] and [59].

Two glycerophospholipid classes, namely PC and Phosphatidylethanolamine (PE), are also formed from DG by transferring their headgroups - phosphocholine and phosphoethanolamine, respectively - onto the free hydroxy group of DG [60, 61, 62]. While Phosphatidylserine (PS) can be synthesized from PC and PE [63, 64], a second route via cytidine diphospho-DAG (CDP-DAG) exists in yeast [65]. CDP-DAG, obtained via a transfer of cytidine diphosphate onto PA, is the precursor for PS, Phosphatidylinositol (PI), and Phosphatidylglycerol (PG) [66]. All three are formed by transferring the respective characteristic head group moiety onto the DG-bound phosphate group and releasing cytidine monophosphate.

Sphingolipids form the third major category of lipids. The most abundant sphingolipid classes, Ceramide (Cer) and SM, are structurally similar to DG and PC, respectively. However, they are synthesized in a separate pathway as they don't have a glycerol backbone but a so-called long chain base. It is formed during the synthesis of Cer via a palmitoyl transfer onto serine followed by a reduction step [67, 68, 69]. In a condensation reaction, a fatty acyl is added to the amino group. A subsequent desaturation yields Cer [70, 71]. By adding a phosphocholine onto the hydroxy group of the sphingosine backbone, Cer is converted to SM.

Lipid Nomenclature A major issue when analyzing lipid data from a bioinformatics point of view is the naming of specific lipid species, as the names following the International Union of Pure and Applied Chemistry (IUPAC) nomenclature can get extremely long and thus shorthand conventions are typically used. Multiple shorthand notation schemes exist, e.g. from Liebisch et al. [72] or Pauling et al. [73], that describe the combination of lipid class and fatty acid, including the level of resolution with which the respective lipid species could be identified. For example, PC(16:0/18:1) describes a Phosphatidylcholine (PC) with a saturated FA with 16 carbon atoms

[12] This is also possible for unsaturated FAs through auxiliary enzymes [51].

at the *sn-1* position and a FA with 18 carbon atoms and one double at the *sn-2* position[13] in the nomenclature from Liebisch et al. [72] [14]. This level of identification is often referred to as *sn-specific*. If it is unclear in which order the fatty acyls are located, the lipid species is referred to as molecular lipid species and denoted as PC(16:0_18:1) for our example. The most coarse-grained identification level is the sum lipid species level: PC(34:1). In this case, only the total sum of carbon atoms and double bonds, not the exact fatty acyl composition, could be inferred. An explanation of how these identifications work and why coarse-grained annotations are still common is given later in Section 2.1.3.

In addition to the mentioned FA properties, the exact double bond position and orientation can also be specified. However, these properties cannot be identified at a larger throughput with the current state-of-the-art analytics.

Complex Lipids The above example shows that lipids are a combination of fatty acids and their backbones/head groups. While fatty acid modifications only happen on free fatty acids (as acyl-CoAs), lipid class reactions happen on complex lipids. Depending on the enzyme catalyzing a specific reaction, the fatty acid composition of a complex lipid often influences the conversion rate. Often, multiple enzymes catalyze the same reaction but with different fatty acid preferences. Unfortunately, such preferences are rarely known, let alone quantified, making it hard to incorporate them into bioinformatic pipelines. In fact, not even metabolic networks for lipid species are established, preventing a variety of computational approaches (see Section 2.2) from being applicable to lipidomics data. Taking a first step in this direction is one of the main projects presented in this thesis.

Lipid Metabolism in Disease Due to the fundamental roles of lipids, lipid metabolism is involved in many types of diseases, among them cardiovascular diseases, diabetes, obesity, and cancer [74].

As already teased in the previous sections, lipids play a crucial role in tumors as they are required for proliferation and signaling. Consequently, different oncogenes can activate *de-novo* FA synthesis and lipogenesis [75]. For example, the Hypoxia-Inducible Factor 1 increases FA synthesis expression, and hypoxia in cancer can lead to increased synthesis of *cytoplasmic* acetyl-CoA from acetate providing the necessary precursors for FA synthesis [76, 77, 75]. Despite these effects of tumorigenesis on lipid metabolism, lipids themselves can also drive cancer development. One of the ways this influence is manifested can be seen in obesity, where high lipid content - especially of Cer and DG - causes defective insulin signaling, triggering a release of insulin and Insulin-like Growth Factor, both tumor-promoting factors [78, 79, 75].

Obesity is, by definition - an excess of body fat [80] - directly linked to lipids. To accommodate for the increased volume of lipids stored, adipocytes increase in size (“hypertrophic obesity”) or number (“hyperplastic obesity”) during the development of obesity [81]. Due to both hypertrophy and hyperplasia requiring specific lipids for cell membrane remodeling or organization, cells need to adapt their lipid metabolism to account for these needs. Furthermore, *de-novo* FA synthesis

[13] The sn-position indicates to which carbon atom the hydroxy group, to which the fatty acid is esterified, is connected.

[14] For the specific example, the number of hydroxylations is implicitly given as 0, as only carbon and double bond numbers are given, but they can also be specified if present. A single hydroxylation of the FA 16:0 would be denoted as PC(16:0;OH/18:1).

Essential	Conditionally Essential	Nonessential
Histidine	Arginine	Alanine
Isoleucine	Cysteine	Asparagine
Leucine	Glutamine	Aspartate
Lysine	Glycine	Glutamate
Methionine	Proline	
Phenylalanine	Tyrosine	
Threonine	Serine	
Tryptophan		
Valine		

Table 2.1: Amino Acids categorized by essentiality. Adapted from Figure 1 in Chandel [85].

and lipogenesis are regulated depending on the degree of obesity. One hypothesis, based on the current literature, is that they are upregulated in earlier stages, when body fat is built up, and downregulated in later stages, putatively to restrict the total fat content [82, 83, 84]. The changes in lipid metabolism in obesity also play an important role in one of the main papers Appendix A.2 presented in this thesis when demonstrating the capabilities of LINEX^2 .

2.1.2.3 Amino Acid Metabolism

Human AAs are α-AAs, which indicates that the functional groups are linked through a central carbon atom. Depending on the type of residue, specifically its charge or hydrophobicity, they can be classified into different chemical categories.

Furthermore, proteinogenic AAs can be separated into three groups depending on their role in metabolism:

1. essential
2. conditionally essential
3. nonessential

By definition, the first group, spanning nine AAs, cannot be synthesized by humans and needs to be taken up through the diet or be produced by the gut microbiome [86], as plants and bacteria are able to synthesize all AAs [87]. The seven conditionally essential AAs can generally be synthesized by humans and only become essential under certain conditions, such as low-birth-weight infants [88]. In contrast, nonessential AAs do not have to be taken up, as the human body can synthesize them on its own.

AA synthesis generally builds on intermediates of the CCM pathways. However, humans are incapable of PPP-based amino acid biosynthesis. Therefore, Tyrosine, for example, can only be synthesized directly from the essential AA Phenylalanine (by mammals). The biosynthesis of AAs is not only important for protein synthesis but also for other metabolic processes, as they serve as precursors for neurotransmitters and nucleotides, among others [85].

Excess AAs are degraded to intermediates that can enter the TCA cycle or be used for gluconeogenesis. AA breakdown can be separated into two parts: metabolization of the amino group at the α position and processing of the carbon backbone. α-amino groups are removed by first converting an (α-)AA[15] to glutamate, yielding an α-keto acid as a side product. These reactions are catalyzed by a family of enzymes called transaminases. This group of enzymes will reappear again later in Section 5.1. Glutamate can then be dehydrogenated to serve as a substrate for the urea cycle, which is used to excrete excess nitrogen in the form of urea. Interestingly, the urea cycle is linked to the TCA cycle and both were discovered by Hans Krebs[16] [85]. Some AAs can also be deaminated directly without requiring the transaminase step.

Subsequently, the remaining carbon skeleton is processed. AAs that can be directly converted to acetyl- or acetoacetyl-CoA are called ketogenic because acetyl-CoA is used to produce ketone bodies. Glucogenic AAs are processed to pyruvate or TCA cycle intermediates and can be used to fuel gluconeogenesis. Only two AAs are purely ketogenic - Leucine and Lysine - 14 are exclusively glucogenic and four can be both [90].

Amino Acid Metabolism in Disease AA degradation, especially of glutamine, is of great importance for the energy metabolism of tumor cells by producing citrate as a precursor for lipogenesis in hypoxia [91], showcasing how crucial the interplay of CCM, lipid and AA metabolism for disease progression is. As glutamine deprivation induces apoptosis, tumor cells further modify amino acid metabolism to increase the production of cellular asparagine to suppress their death [92]. Despite their “main” function as building blocks of proteins and their role in energy and lipid metabolism, AAs also serve as precursors for nucleotides and contribute to epigenetic regulation and immunosuppression in tumors [93].

AAs and their metabolism are also known to impact Inflammatory Bowel Disease (IBD) on both the host as well as the microbiome side [94]. Indeed, dietary supplementation of some AAs is even used as a treatment for some forms of IBD, yet other AAs can also have a pro-inflammatory role and dietary intake of these should thus be restricted [94, 95]. Although IBD and the influence of the microbiome have been studied extensively, many aspects of disease progression, the (potentially causal) role of the microbiome, and its interaction with host metabolism remain unclear [96]. In my work on developing an approach to infer alterations of metabolic reaction activity from metabolomics data and integrating it with microbiome data, I demonstrate how mantra can be used to gain further insights into the metabolic changes under IBD.

While the CCM is often considered the most fundamental part of metabolism, all introduced pathways are essential for human life. As already highlighted, they are all connected through shared metabolites, mostly via CCM. From a biological perspective, this means that limited resources can introduce trade-offs between different pathways. Thus, tight regulation is required, and changes in one part of metabolism can affect metabolism as a whole. From a bioinformatics perspective, this gives rise to interesting properties of biological systems that can be exploited for certain applications but also complicate inference tasks (more on this in Section 2.2). The introduced pathways are generally well understood under healthy conditions. However, many questions regarding regulatory mechanisms, the interaction with the environment, and especially the role of dysregulation in disease progression remain unanswered in many cases. Especially mechanistic hypotheses, which

[15] Any amino acid except for Lysine can be *directly* transaminated [89].

[16] Therefore, an alternative name for the TCA cycle is “Krebs cycle”.

aim at not only identifying aberrant phenotypes but also pinpointing possible modes of action, are needed to move the field forward.

2.1.3 Measuring Small Molecules

Analytical chemists have been trying to accurately identify and quantify molecules for a long time, with the earliest approaches dating back to the 18th century [97]. The technologies that enabled modern metabolite analyses are Nuclear Magnetic Resonance (NRM) and Mass Spectrometry (MS). Using a magnetic field NRM induces chemical shift and spin-spin coupling, yielding spectra characteristic for a given molecule. In contrast, MS is concerned with measuring the mass-to-charge ratio (m/z) of ionized molecules. While NRM is still used for structural elucidation and analysis of living samples, among others, the standard for high-throughput metabolomics is nowadays MS. One of the major reasons for this is the 10 to 100-fold increased sensitivity of MS, leading to a much higher number of possible identifications [98]. Nowadays, the acquisition of hundreds to thousands of molecules can be achieved with workflows lasting 30 minutes or less [98]. While the acquisition of lipids follows the same general workflow as for metabolites, the individual steps are typically optimized for either lipids or metabolites.

Züllig, Trötzmüller, and Köfeler [99] divide a typical lipidomics (or metabolomics) workflow into four steps:

1. Sample Preparation
2. Data Acquistion
3. Data Processing
4. Data Interpretation

This section provides an introduction to the first two steps, while steps three and four are covered in Section 2.2.

2.1.3.1 Molecule Extraction

While it is possible to directly analyze complex, even living, samples with NRM, analysis via MS requires specific extraction protocols. These aim at isolating specific classes of metabolites and removing unwanted molecules, such as polynucleotides and proteins. Naturally, due to the structural diversity of metabolites, different extraction protocols favor different types of metabolites.

Two major components in metabolite extraction are the choice of solvent(s) and the technique chosen [100]. Solvents should typically have a high solubilization strength and a low selectivity towards metabolite classes. For metabolites, typical solvents are methanol and acetonitrile [101]. For lipids, which have more distinct properties, methanol, choloroform, and methyl-tert-butyl ether (MTBE) are commonly used solvents [102, 103, 104]. A technique applicable to both metabolites and lipids is the so-called *liquid-liquid extraction* [105, 106]. It uses two solvents with different polarity and/or water solubility to separate molecules based on their chemical properties [99]. By retaining only one of the two phases, unwanted molecules are removed. Depending on the type of pre-separation (see Section 2.1.3.3) and the categories of metabolites to target, derivatization after extraction can enhance analytical capabilities [107].

In addition to the above-explained considerations, sample handling prior to extraction is critical. Due to short turnover times and the highly dynamic nature of metabolism, sampling time should be kept as short as possible [108, 100]. Subsequent storage conditions need to prevent spontaneous chemical modifications, such as oxidation. Furthermore, all best practices for experiment design and sampling in molecular biological experiments, for example, adequate randomization and sufficient controls, apply to metabolomics experiments, too.

Once metabolites/lipids are extracted, samples are ready to be analyzed. Generally, the measurement is divided into 5 steps:

1. Pre-Separation
   a) Chromatography
   b) Ion-Mobility Spectrometry
2. Ionization
3. Measuring intact molecules
4. Fragmenting molecules
5. Measuring fragments

with steps 1, as well as 4 and 5 (together), being optional.

2.1.3.2 Separation by Mass

To understand how metabolites are identified and quantified in modern high-throughput experiments, a brief introduction to the workings of mass spectrometers is necessary.

While different types of mass spectrometers exist, they are all unified by the idea of ionizing molecules (i.e. giving them a charge) and subsequently using the acquired charge for detecting their mass-to-charge ratio by manipulating their trajectory with electromagnetic forces. The process of ionization is happening in a so-called *ion source*. Different ways of achieving ionization have been developed, with the most common nowadays being Electrospray Ionization (ESI) [109, 110][17].

Once ionized molecules enter a drift tube, guiding them via electromagnetic forces. To measure the mass of an "intact" molecule, the so-called precursor ion, analytes enter a mass analyzer. All mass analyzers measure molecules dependent on both their mass and their charge. Therefore, mass spectrometers report the mass-to-charge ratio (m/z). The charge of any ion can be inferred using what is referred to as its isotopic pattern, and thus, molecule masses can (easily) be computed. Two common types of mass analyzers are Time-of-Flight (TOF) and orbitrap. The principle behind TOF, as its name already suggests, is to use the proportional relationship between an ion's m/z and the time it takes to cover a specific distance in the flight tube after acceleration by a voltage difference. At the end of the flight tube, a detector releases electrons when hit by an ion. The number of ions can be quantified, as it is proportional to the number of electrons released. The reported "intensity" cannot be directly used to calculate concentrations or absolute quantities of molecules unless specific internal standards are included. This is due to unknown molecule-specific

[17] ESI is the most widely employed method for liquid samples. For applications such as MS imaging, different ionization modes have to be used.

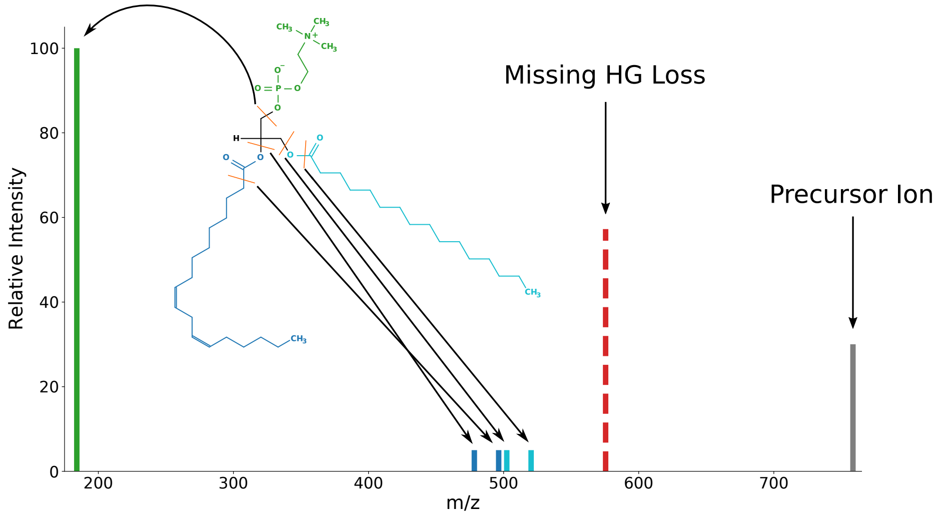


Figure 2.4: Schematic overview of the fragmentation pattern of PC(16:0_18:2) in positive mode. Experimental spectrum data was obtained from the MassBank of North America (MoNA) (https://mona.fiehnlab.ucdavis.edu/spectra/display/LipidBlast060656; accessed on 5 September 2023). Molecule depiction was generated from the SMILES string of PC(16:0/18:2(9Z,12Z) from the LipidMaps Structure Database [113] using OpenBabel. The expected fragments were taken from the ALEX^{123} database with H+ as adduct [73].

ionization efficiencies. In contrast to TOF analyzers, where ions are destroyed when hitting the detector, orbitraps use non-destructive detectors. This is possible because they determine m/zs by circling ions around an electrode[18], creating axial oscillations[19][111]. The axial oscillations lead to an image current [111], which can be Fourier-transformed into frequencies and translated to m/z values. Intensities are determined through the strength of the signal.

Obtaining m/zs for precursor ions and their corresponding charge state allows for the identification of metabolites to a certain degree. Generally, these masses can be matched against databases to obtain one or multiple candidate molecules - a process termed accurate mass search. As expected, the higher the accuracy, the further the number of candidate molecules can be narrowed down. However, many metabolites are isobaric, i.e. their masses are identical. For example, Opialla, Kempa, and Pietzke [112] showed that less than half of all metabolites in the KEGG database [49] have unique masses. In lipidomics, this problem is even more dramatic, as all sum species (recall the definition from Section 2.1.2.2) have the same precursor mass. Consequently, accurate mass search can only get us so far in terms of identification, and additional ways of telling isobaric compounds apart are required.

Fragmentation One way to resolve isobaric compounds is by fragmenting compound ions. To get “clean” fragmentation patterns, single precursor m/zs are isolated using mass filters, most

[18] The orbitrap also contains two outer electrodes in addition to the inner electrode around which ions are oscillating.
[19] Additionally, radial movement is induced, and balancing the electric fields is a delicate balance [111]

commonly quadrupoles, and guided into a fragmentation chamber. A typical fragmentation technique used to identify small molecules is Collision-Induced Dissociation (CID) [114, 115]. It works by accelerating ions inside a neutral-gas-filled collision cell such that precursor ions colliding with the chemically inert gas physically break them apart. The exact position and proportion of bond breakage are dependent on the energetic state [116]. After passing through the fragmentation chamber, the *product ions* are guided to the mass analyzer to determine their m/zs. A fragmentation event in metabolomics typically leads to one charged product ion and one uncharged (neutral) ion, whose m/z cannot be detected. When distinguishing isobaric or even isomeric molecules, the fragmentation pattern serves as a fingerprint specific to each compound [115]. However, in practice, oftentimes not all product ions are observed, and certain types of isomers, such as cis-trans-isomers, are indistinguishably in state-of-the-art high-throughput experiments [20] [118]. Therefore, each molecule identification comes with its own degree of uncertainty.

To get even more fine-grained identification with MS only, it is also possible - with certain instrumental setups - to pass fragment ions into the fragmentation chamber one or multiple times [119]. Doing experiments with one round of fragmentation is referred to as MS/MS, tandem MS, or MS2, while higher orders of fragmentation are denoted as MSn, where n is the number of fragmentation steps + 1. Accordingly, MS experiments without fragmentation are called MS1.

To showcase what fragmentation patterns look like and how they influence the level of compound annotation, PC(16:0_18:2) is going to serve as an example. Figure 2.4 depicts the fragments measured/expected in positive mode with CID. It shows that PCs fragment at bonds connecting the components of complex lipids - head group and FAs. While this principle is true for both positive and negative charge mode, the exact fragments measured differ between the charge modes.

Characteristic for PCs is the peak at around 184.07 (green peak), representing the protonated Phosphocholine breaking off from the glycerol backbone (for a compilation of major Phospholipid head group fragments, see Table 2.2). Additionally, a loss of the Phosphocholine (uncharged) would be expected but is not observed in this spectrum (red dotted line). Because the lost moiety weighs[21]around 183 Da (184.07 minus the mass of a proton), the peak would be expected at the precursor mass minus 183 (≈ 575).

In addition to these lipid class-specific peaks, FA-specific peaks can be observed. These allow the identification of molecular lipid species. Each FA can produce two distinct peaks from losses in positive mode because both bonds surrounding the ester bond break frequently. For this particular example, this results in peaks at 478 and 496 Da for the 18:2 FA (blue peaks) and 502 and 520 Da for the 16:0 FA (cyan peaks). In negative mode, FA fragments instead of losses would be observed.

The remaining peak in the spectrum (grey) represents unfragmented precursor ions. Generally, the higher the collision energy, i.e. the kinetic energy of the ions in the collision cell, the lower the intensity of the precursor ion in the MS2 spectrum.

Because fragmentation patterns allow to distinguish between isobaric and some isomeric compounds, metabolomics experiments can be run by directly injecting sample extracts into a mass spectrometer. In lipidomics, this approach is commonly referred to as shotgun mass spectrom-

[20] The recent revival of alternative fragmentation techniques such as Electron-Activated Dissociation (EAD) [117], together with improvements in pre-separation, promises progress in isomer-differentiation in the future.

[21] Note that the 183 is *not* a m/z as it describes a neutral molecule, which thus has no charge.

Lipid Class	Head Group Fragment m/z values	
	Positive Mode (+H+)	Negative Mode (-H+)
PC	184.07	
PE		140.01, 196.03
PG		153.00, 171.01
PI		153.00, 223.00, 241.01
PS		153.00

Table 2.2: Characteristic head group fragments for five major glycerophospholipid classes as contained in the ALEX123 database [73]. Fragment m/z values are rounded to the second digit after the comma. All values in positive mode are for proton adducts and negative mode for proton loss. Neutral losses of head group fragments are excluded.

etry[22]. While it saves time to use direct infusion, it can come at the cost of sensitivity[23], due to a phenomenon known as ion suppression [99]. It describes a decreased efficiency of ionization leading to low abundant analytes becoming potentially undetectable [121, 99]. One factor thought to influence is the amount of (especially) poorly- or non-volatile compounds[120, 121]. Naturally, shotgun mass spectrometry leads to large amounts of compounds at the same time, as the entire sample is continuously injected. One possibility to reduce these effects is to use pre-separation techniques.

2.1.3.3 Separation by Chemical Properties

Chromatography Chromatography is a general technique in analytical chemistry to separate compounds on the basis of chemical properties. Roughly speaking, chromatography guides analytes through a column in which they chemically interact with the material present on the inside of the column. More technically speaking, chromatography uses two components for separation, a *mobile* phase, which carries molecules through the column, and a *stationary* phase, the material on the inside of the column. While traveling through the column, analytes chemically interact with both the mobile and the stationary phase, for example, through van-der-Waals forces. The time spent inside the column, the Retention Time (RT), of a compound depends on how strong the association with the mobile phase is relative to the association with the stationary phase (the higher, the shorter). Two common types of chromatography for lipidomics are Liquid Chromatography (LC) and Gas Chromatography (GC), which have a liquid and a gaseous mobile phase, respectively.

In exploratory metabolomics experiments, a common approach is to have two runs (for each MS charge mode) with different Liquid Chromatography (LC) columns separating based on different chemical forces. One example of a column combination is Reversed Phase Chromatography (RP) and Hydrophilic Interaction Chromatography (HILIC) [122]. HILIC columns have a polar stationary phase and a more organic mobile phase[24], thus separating hydrophilic molecules well

[22] Note that the use of the word “shotgun” in lipidomics is distinct from the use in proteomics.

[23] The degree of ion suppression also depends on the ionization technique, with ESI being rather susceptible [120].

[24] To elute hydrophilic molecules, the hydrophobicity of the mobile would be gradually decreased throughout the

[123]. For the separation of hydrophobic molecules, such as lipids, RP columns are used. These contain an organic stationary phase, and the mobile will go from hydrophilic to hydrophobic during the course of a sample run.

Ion Mobility Spectrometry In contrast to chromatography, Ion Mobility Spectrometry (IMS) is a technology that does not use chemical properties of ions for separation, but size and shape [124]. Different *drift-times* can be achieved through different IMS architectures, e.g. by using an asymmetric electrical field or a neutral gas flow [124, 125]. While it is not fully orthogonal to mass spectrometry, as mass also plays a role for the drift time [126], IMS has shown to improve the identification of isomers, e.g. in lipidomics [127, 128, 129, 130, 125]. Because IMS needs ionized compounds, it can either be used as a sole pre-separation step or follow a chromatographic separation prior to injection into the mass spectrometer.

Both chromatography and IMS not only improve the detection of compounds inside the mass spectrometer but also serve as additional filtering steps for identification [131, 132]. For example, reference RT windows, from *in-silico* prediction or experimentally determined, can be used to filter out unlikely compound candidates. Similarly, the drift-time from IMS can be used to compute an ions Collisional Cross-Section (CCS), which can then be compared to the expected CCS of candidate molecules.

Nowadays, a number of software suites for the identification of metabolites from mass spectra exist. Some prominent examples are MS-DIAL [133], OpenMS [134], and MetaboAnalyst [135]. These platforms can handle all typical combinations of pre-separation and mass spectrometer variations. While they are also capable of identifying lipids, the special requirements of Lipidomics led to various specialized tools for lipid identification, such as ALEX^{123} [73], LipidXplorer [136], LipidMatch [137], and LipidHunter [138].

Although the work I will present in the following chapters is not focused on identifying small molecules, the approaches presented for both lipidomics [9, 10] and metabolomics [11] rely on accurate (and large-scale) identifications. Furthermore, some of the student projects I supervised that do not appear in this thesis aim at improving lipid identification.

A major downside of current MS-based metabolomics workflows - irrespective of instrument type, analysis mode, and pre-separation - is that it is hardly possible to multiplex. With acquisition times of up to 30 minutes, this displays a clear disadvantage compared to sequencing technologies (Section 2.1.4) or MS-based proteomics, where multiplexing approaches such as tandem mass tag (TMT) labeling are available [139].

2.1.4 Sequencing-Based Technologies

Nucleotide sequencing has become a standard technique in molecular biology in the last decades. The earliest report of sequencing dates back to 1968 [140], but methods efficient enough to decode more than a few tens of nucleotides were not developed until 1976 [141]. In this year, two methods with different approaches were developed by Sanger, Nicklen, and Coulson [142] and Maxam and

analysis time.

Gilbert [143]. The first method is based on synthesizing copies of a gene, but instead of supplying exclusively "regular" nucleotides, small fractions of modified nucleotides missing a hydroxy group[25] thus ending elongation are added [142]. This way, one partial fragment for each base is produced - in small fractions. Subsequent sorting of fragments by length via electrophoresis with one lane per type of dideoxy nucleotide allows to "read off" the sequence. Sequencing with chain-terminating bases followed by electrophoresis separation was later coined as *Sanger sequencing*. Maxam and Gilbert [143] approached sequencing from the "other" side - doing restriction instead of elongation. By having cleavages that are specific to each nucleotide with labeled fragments, the sequences can be read off in a similar manner to the method by Sanger, Nicklen, and Coulson [142]. These first approaches used radioactive labeling. Later, fluorescence labeling versions of Sanger sequencing were developed that allowed to sequence thousands of base pairs in an automated fashion [141]. This way of sequencing is nowadays often referred to as First Generation Sequencing.

Next Generation Sequencing (NGS), also referred to as massively parallel sequencing, describes a new approach that does not require electrophoresis and allows to highly multiplex. The basic principle is to ligate adapters to the end of each fragment, serving as both an identifier and to mount the fragment to sequences followed by synthesizing copies in a way that emits a light signal at every iteration. Three general flavors of this so-called *sequencing-by-synthesis* approach were proposed - pyrosequencing [144], ligase-mediated label transfer [145], and polymerase-mediated labeling [146, 147]. Nowadays, the most widely used method is polymerase-mediated synthesis [141]. It adds a single base per cycle by using fluorescence-labeled nucleotides that terminate elongation. Subsequently, the labeling for each fragment is read off, and the fluorescence tag, as well as the group preventing further elongation, are removed so the next cycle can start.

Recently, a third generation of methods has been developed. In contrast to first and second-generation methods, which require cutting DNA into small fragments, these methods aim at measuring large DNA strands. One such approach uses nanopores, small transmembrane pores, through which one single strand of DNA is guided. During this process, nucleotides (3 to 7) passing through the pore lead to a characteristic electric signal [148].

Sequencing Microorganism One aspect of health that caught attention in the last decades is the microbiome [149]. Initially, the analytics of quantifying bacteria focussed on detecting a specific subunit of bacterial rRNA - the 16S unit - with DNA sequencing [150]. Thus, this type of analysis is often referred to as *16S sequencing*. By using specific primers "spanning"[26] the 16S gene, small mutations within the sequence can be used to bin them into so-called Operational Taxonomic Units (OTUs) based on the similarity of their 16S sequence [151]. While this has been the most widely used strategy for years, it comes with the great limitation of not being able to resolve microbial species and identify non-bacterial microorganisms at all. Therefore, with the decreasing cost of NGS, Metagenomic Sequencing (MGS) became more popular as it allows species, and possibly strain-level, resolution [153]. Furthermore, it opens up the possibility to sequence individual genes, which can give a more functional view of microbial composition, for

[25] The dideoxy nucleotides used for chain termination lack the hydroxy group at the C3. Recalling the mode of polymerization of nucleotides from Section 2.1.1.1, the 3' hydroxy group is the one forming the ester bond with the phosphate group of another nucleotide.

[26] Different strategies for designing primers exist, and the choice heavily affects the outcome of OTU grouping and quantification [150, 152]

example, by determining the metabolic capacities of communities [154]. In a nutshell, MGS is done by sequencing sites of the microbiome, such as stool or mucosa from different body sites, and mapping the resulting sequences against a database of known microbial genomes.

While I've worked on different projects integrating 16S with metabolomics data that are not part of this thesis for brevity (e.g. [155]), MGS plays an important role in supporting the results of the mantra algorithm[11] - one of my main publications presented here. Additionally, mantra is developed to improve the interpretation of multi-omics data, especially of MS-based metabolomics and sequencing data.

2.2 Computational Biology and Bioinformatics

Bioinformatics is a subdiscipline of computer science concerned with the development of algorithms for the analysis of biological data, while computational biology is the application of computational methods to biological data [156]. The first bioinformatics methods, working on *de-novo* assembly of peptide sequences, date back to the 1960s [157]. While the field was originally mostly dealing with analyzing sequencing data, predicting protein structures, and generating databases, it has since evolved to cover a vast bandwidth of applications [157]. One notable demonstration of the need for computational biology was the initial sequencing of the human genome [158], as the vast amount of sequencing data produced could not have been aligned without the use of computational algorithms. A more recent example of a computationally driven breakthrough in biology is the introduction of AlphaFold2 [159], which is able to deliver accurate predictions of protein structures for many cases. Especially with the recent advent of machine and deep learning, a broader data science community has found interest in biological applications.

Computational metabolomics is the sub-field analyzing metabolomics-related experiments, all the way from processing and annotating mass spectra to interpreting and integrating metabolomics data. Despite this broad definition, the main focus of the field still lies in the annotation of mass spectra. Although this rather narrow focus of the community has started to shift, approaches for metabolomics data analysis, integration, and interpretation are still lagging behind compared to proteomics or transcriptomics. Parts of this apparent gap are addressed by the work in my thesis.

In the following, I will introduce aspects of computer science relevant to understanding the methods presented in the next chapter as well as understanding the discussion of my work in a broader context.

2.2.1 Graphs in Biology

A central part of this thesis is leveraging prior knowledge in the form of metabolic information to guide data interpretation methods for metabolomics. Like many other types of prior knowledge in biology, metabolism can be represented as a *graph*. A graph $G = (V, E)$ is a mathematical structure defined as a pair of two sets: the set of vertices (or "nodes") V and the set of edges E. Vertices can represent one or multiple entities, for example, proteins in Protein-Protein Interaction networks (PPI networks) or metabolites and metabolic reactions in metabolic networks. Depending on whether E consists of ordered or unordered 2-tuples, a graph is said to be *directed* or *undirected*.

Edges can also be assigned weights representing e.g. the strength of association of two vertices[27]. PPI networks, on the one hand, are typically undirected as they represent the interaction between two proteins, which is a symmetric relationship. Chemical reaction systems, on the other hand, are inherently directed as each reaction has a specific direction[28]. If two types of vertices are contained in a graph and *every* edge $e \in E$ represents a connection between a vertex of one type with a vertex of the other type, the graph is referred to as *bipartite*. Metabolic networks are typically represented as (directed) bipartite graphs, as only connections between metabolites and reactions exist. Computationally graphs can be represented with different types of data structures. The choice of data structure is usually dependent on the application, as each allows for certain operations to be implemented efficiently. For example, determining the degree - the number of direct neighbors - of a vertex in an unweighted adjacency matrix, a binary $n \times n$ matrix where 1 indicates edge existence and 0 edge absence, requires to sum over the ith row (or column). With an adjacency list - a list containing the list of neighbors for each vertex - the degree can be determined by taking the length of the respective neighbor list.

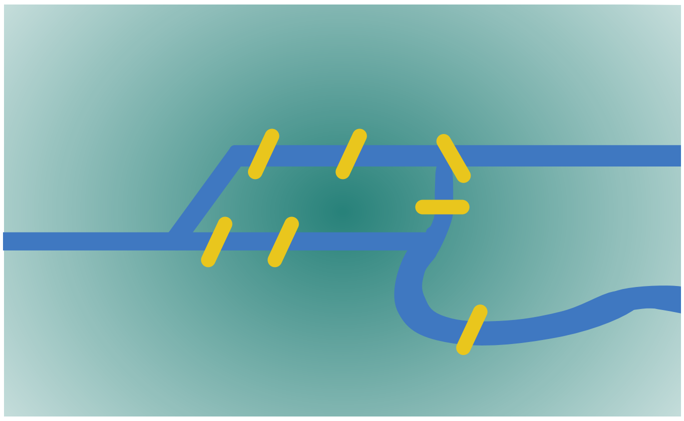
In addition to representing direct interactions of molecules, graphs are also used to define knowledge representations, such as ontologies. Gene Ontology (GO), for example, is a hierarchical grouping of genes in functional categories, that models the hierarchy between classes as a Directed Acyclic Graph (DAG) [160]. Another specific case of an acyclic graph frequently occurring in computational biology is a tree. In a tree, any pair of vertices $u, v \in V$ is connected by not more than one path. Trees are commonly used to show the evolutionary relationship between organisms in a so-called *phylogenetic tree*. These can be useful for visualizing evolutionary distances or for finding common ancestors and partitioning groups of organisms, e.g. by tree-based clustering [161].

Graphs are not only useful for intuitively representing systems and relationships but they can also be used as data structures to efficiently solve various types of questions. A famous example of an early use case is the problem of the seven bridges of Königsberg. It asks whether it is possible to walk through the city of Königsberg, crossing each of its 7 bridges exactly once (Figure 2.5a). Leonhard Euler famously solved the problem by removing all geometric information and only considering the connectivity between parts of the city - modeled as a graph (Figure 2.5b). Using the vertex degrees, Euler went on to determine that *no* such walk exists [162]. He stated that every vertex in the graph must have an even degree for a walk traversing every edge exactly once to exist, giving a solution to a generalization of the specific problem. Paths traversing every edge of a graph exactly once are now referred to as *Eulerian paths*. In contrast, a path traversing each *vertex* exactly once is called a *Hamiltonian path*. Despite the seemingly abstract nature of the solution, Eulerian paths and cycles are relevant for computational biology, for example, for traversing de-Bruijn graphs in genome assembly [163]. With his representation of the problem, Euler initiated not only the field of graph theory but also laid the foundation for algebraic topology[29][164] through the idea of considering the connectivity irrespective of the exact geometry.

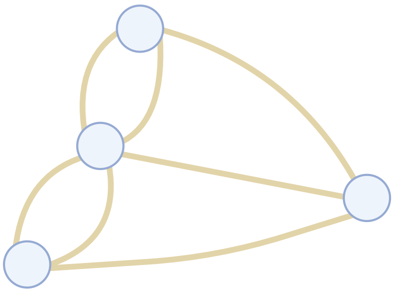
[27] In the case of weighted edges, they can be represented as 3- instead of 2-tuples, where the third value indicates the edge weight. Alternatively, a weighted graph can be represented as $G = (V, E, W)$ where each $w_i \in W$ gives the weight of edge $e_i \in E$.

[28] Reversible reactions are represented with two directed edges, one for the forward and one for the reverse reaction.

[29] Among other contributions, Euler is famous for the *Euler characteristic* (χ), a fundamental topological invariant describing the relation between the number of vertices, edges, and faces (F) in a polyhedron ($\chi = |V| - |E| + |F|$)



(a)



(b)

Figure 2.5: Illustration of the problem of the seven bridges of Königsberg. *(a)* Schematic of the old city of Königsberg where the river is indicated in blue, bridges in yellow, and city grounds in green. *(b)* Graph representation of the problem with every vertex representing a part of the city and every edge indicating a bridge connecting the two respective parts. Figures created with BioRender

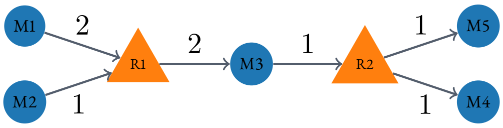
Nowadays, graph theory is almost ubiquitously used in computer science from navigation systems [168] and web searches [169] all the way to databases [170]. The first application uses a common class of algorithms trying to compute the shortest paths between vertices. A shortest path between two vertices is defined as the path containing the minimal sum of edge weights[30]. Multiple algorithms tackling the problem of finding the shortest paths exist, covering different scenarios concerning edge weights, e.g. non-negativity, sparsity, or single-source vs. all-pairs. In bioinformatics, shortest path algorithms have been used to infer Gene Regulatory Network (GRN) [171], for example. Because graphs induce a metric space together with the shortest path distance, it is also useful to transfer concepts from geometry, such as curvature [172, 173], to a discrete setting. Recently, curvature has gained interest in applications such as Geometric Deep Learning (GDL) [174], an interesting field for the future of computational metabolomics. Furthermore, Ollivier-Ricci Curvature (ORC) [173], one specific extension of Ricci curvature to discrete settings, was originally developed on Markov models. Markov models are one natural way of representing reaction networks with the transition probabilities as the probability that one metabolite is converted into another metabolite [175]. Together with extensions of ORC to hypergraphs [176] - another way of representing metabolic networks [177] - such concepts hold interesting possibilities for prior-knowledge driven metabolomics methods in the future (see Chapter 6 for an in-depth discussion).

Closely connected to Markov Models are so-called random walks on graphs. Random walk matrices are stochastic square matrices where each element defines the probability of "walking" from vertex A to vertex B with a single step. A nice property of such matrices is that n-step walks can easily be simulated by raising the matrix to the n^{th} power. The principle of random walks is, for example, used by PageRank [169], the algorithm initially used by Google for web page recommendation. Frainay et al. [178] have adapted the concept of PageRank to develop a recommendation system for metabolic fingerprints.

A common use-case of graphs in bioinformatics and computational biology outside computational metabolomics is active module identification with many specific methods developed over the years [179, 180, 181, 182, 183, 184, 185, 186, 187]. Active modules are connected subgraphs with the highest change in biological signal between conditions [188]. The idea behind such modules is that they pinpoint to main areas of biological "processes", similar to pathways, affected by a knock-out, disease, or other condition. To identify active modules, an NP-hard problem [189], approximation techniques from combinatorial optimization, such as Ant Colony Optimization [190], Genetic Algorithms [191], or Local Search with Simulated Annealing [192], are used. While many methods for active module identification are doing so-called *de-novo* pathway enrichment [193] - because they define "disease pathways" - making them unbiased towards pathway-definitions, they are still highly dependent on the quality of biological networks they use. Lazareva et al. [194] have shown that this is especially problematic for approaches relying on generic PPI networks, which usually have a high degree of bias. Because PPI networks suffer from both technical [195] as well as study bias [196], this displays a crucial limitation for many of the proteomics and transcriptomics analyses based on this type of method.

[165]. It is nowadays still used to develop new approaches for biological [166] or machine-learning [167] applications, for example.

[30] For an unweighted graph, this amounts to the path containing the least edges.



(a)

$$\begin{array}{c|cc} & R1 & R2 \\ \hline M1 & -2 & 0 \\ M2 & -1 & 0 \\ M3 & 2 & -1 \\ M4 & 0 & 1 \\ M5 & 0 & 1 \end{array}$$

(b)

Figure 2.6: Representations of Metabolism. (*a*) Weighted Bipartite Graph. Metabolites are represented as blue circles, metabolic reactions as triangles. Edge weights represent the number of molecules of each metabolite. (*b*) Stoichiometric Matrix. Each row represents a metabolite; each column a metabolic reaction. Negative values indicate the number of molecules required as substrate, positive values the number of molecules as products of a given reaction.

Metabolic networks, in contrast, are generated very differently and are much less prone to false positive reactions, although it is known that current networks are incomplete [197]. Therefore, developing enrichment methods for metabolic networks holds great potential to give more noise-robust and interpretable results. This idea is used in two of the main publications of this thesis, LINEX^2 [10], which also tackles the limited availability of lipids in metabolic networks, and mantra [11].

In addition to graph theory-inspired analyses, metabolic networks can also be used to formulate a mathematical representation of metabolism, typically by defining a stoichiometric matrix. A stoichiometric matrix is a matrix in which every row represents a metabolite, every column a reaction and the value of a cell x_{ij} defines whether metabolite i takes part in a reaction j ($x_{ij} \neq 0$) and if yes how many of this molecule act as substrate ($x_{ij} < 0$) or product ($x_{ij} > 0$) (Figure 2.6). Together with the law of mass conservation - no mass is lost or gained during a reaction[31] - this gives rise to a system of linear functions that, can be used to find the flux values of each reaction to maximize a given objective under certain constraints [198]. However, this approach is limited to reporting fluxes in steady-states but cannot give information with respect to metabolite concentrations [199]. Dynamic modeling, which uses highly parametrized differential equation systems, allows for insights into concentrations and dynamics, yet it is limited by the parameterization required [199, 200]. Consequently, despite their ability to accurately model dynamical systems, constraint-based and dynamic modeling are limited in their practical applicability [10]. Hence, there is a lack of methods that allow researchers to learn about changes in metabolic reactions from metabolomics data, especially in a data-driven way. Both LINEX^2 [10] and mantra [9] address this gap.

Despite the confidence of annotated metabolic reactions, consideration of the possible biases introduced by integrating prior knowledge into data interpretation algorithms should always be taken into account when designing them and interpreting their results. Another drawback[32]

[31] Omitting the mass defect, i.e. the mass gained or lost due to changes in binding energy.

[32] In fact, this can also be seen as another form of bias, as different instrumentation, reference databases, etc., lead to

of such methods is that they omit features that cannot be mapped onto databases. To include unknown features or have completely unbiased analyses, data-driven approaches can be used.

One option to still apply graph-based methods is to use correlation networks. In these graphs, every edge indicates an absolute correlation above a chosen cutoff between the two connected metabolites. Despite the common use of correlation networks in metabolomics data analysis [201], the topology of these graphs can be highly sensitive to the choice of parameters [202]. For conducting purely data-driven analyses without the use of graph-like structures, statistical and machine-learning algorithms are commonly used nowadays.

2.2.2 Statistical & Machine Learning

The most basic ways to analyze metabolome and lipidome data in a hypothesis-free manner are still methods from univariate statics, most prominently statistical hypothesis tests and fold-change analyses [203]. These methods are sufficient to identify "significantly" changing metabolites. However, they are not able to take into account the interactions between variables and identify more complex patterns.

A typical first analysis step in computational biology is to visualize the - inherently high-dimensional - data in two or three dimensions. Dimensionality reduction techniques employed for this purpose are matrix factorization and manifold learning approaches. Matrix factorization methods decompose the original matrix to obtain a lower dimensional representation of the original data, e.g. using eigendecomposition [204]. In Principal Component Analysis (PCA) [205], which belongs to this class of methods, principal components are computed by the dot product between the original data and the loading matrix - the eigenvectors of the covariance matrix [206]. Each principal component is a *linear* combination of the observed dimensions and principal components are orthogonal to each other. Because the eigenvalues of each principal component, relative to the sum of all eigenvalues, correspond to the relative variance of the original data it explains the principal components can be sorted in a meaningful way. Furthermore, due to the principal components being linear combinations of the original feature vector, the contribution of each feature to a given principal component is known. This is an advantage over manifold learning methods.

To not be limited by the assumption of Euclidean spaces [207], non-linear methods from manifold learning, such as diffusion maps [208], t-distributed Stochastic Neighbor Embedding (t-SNE) [209], Uniform Manifold Approximation and Projection (UMAP) [210], or Potential of Heat diffusion for Affinity-based Transition Embedding (PHATE) [211], can be used. These methods generally don't assume that the underlying manifold of the data is a Euclidean space and try to approximate the manifold structure. Especially in the analysis of sequencing data, t-SNE, and UMAP have nowadays replaced PCA as the default method for dimensionality reduction. The basis of manifold learning methods is the so-called manifold hypothesis [212]. This hypothesis states that observed high-dimensional data is typically coming from a lower-dimensional manifold. It is generally assumed that due to interactions and redundancies between features, the observed biological data is sampled from such a lower-dimensional manifold [213].

Another more recent way of reducing dimensionality is by employing autoencoders. Autoencoders are models consisting of an encoder that "encodes" the original data into a latent space,

biases in metabolite identification.

which is typically low dimensional, and a decoder that reconstructs the original features from the encoded data [214]. The latent space thus contains all relevant information, as long as the decoder is able to sufficiently reconstruct the original data. Autoencoders have been shown to meaningfully compress metabolomics data across cohorts [215].

During my doctorate, dimensionality reduction techniques played a vital role in visualizing the results of the substructure analysis of LINEX^2 [10] and the reaction activity change in mantra [11], as both lead to a latent space. Furthermore, I employed (multi-omics) dimensionality reduction methods in most of my co-author papers.

Machine Learning (ML) methods can be broadly categorized by the degree to which they use sample labels during training. *Supervised* ML methods are models using sample labels or values and thus solve two types of tasks:

1. Classification: predicting *labels*
2. Regression: predicting *values*

In the typical notation, the sample data X is classically referred to as the *independent* variables, and the values to predict Y are called the *dependent* variables. The goal of supervised models is to learn the parameters θ of a function f that describes the relationship between the dependent and the independent variables, i.e. $Y = f_\theta(X)$. "Learning" the relationships between X and Y can have two rationales: discovering them *per se*, e.g. to uncover regulatory effects, or predicting *unseen* samples with the learned model.

The samples used for training parameters are called *training* set, while the *test* and *validation* set are used to evaluate model robustness

One of the most basic classifiers is the k-Nearest Neighbor (kNN) classifier. For a given sample x_i, it predicts the class label c_i as the most frequent class among the k-closest samples in the training data.

The simplest way of predicting continuous dependent variables is linear regression. When looking at linear regression from a geometric point of view it can be formulated as $\hat{y} = x^T\beta + \epsilon$ - essentially learning the parameters of a straight line[33] β [216][34]plus independent random noise ϵ. A typical assumption is that noise follows a Gaussian distribution with 0 mean and variance σ^2, i.e. $\epsilon \sim \mathcal{N}(0, \sigma^2)$. The probability of predicting a class label y given the independent variables and the model can be defined as $P(y|x, \beta) = \mathcal{N}(y|x^T\beta, \sigma^2)$ (if following the Gaussian distribution assumption) and thus a conditional probability density [217].

To describe the variance unexplained by the model, the *residual*, the difference between the predicted value $\hat{y}_i$ and the actual value y_i, is used. An aggregated statistic based on the residuals is the Residual Sum of Squares (RSS), defined as

$$RSS = \sum_i^N (y_i - x_i^T\beta)^2 = \sum_i^N |\epsilon_i|^2$$

[33] Strictly speaking an n-dimensional plane.

[34] Formulated as a line equation, linear regression can also be written as $y_i = \beta_0 + x_i^T\beta_1$, where β_0 is the intercept (scalar value) and β_1 is a scalar if x_i is one otherwise it is a vector of the same dimensionality as x_i.

[217]. Dividing the RSS by the number of samples N gives the Mean Squared Error (MSE), a common metric for evaluating regression tasks in ML. A common metric for evaluating the model fit of a linear regression model is the coefficient of determination R^2. It is defined as $R^2 = 1 - \frac{RSS}{TSS}$, where TSS is the total sum of squares $\sum_i^N (y_i - \bar{y}_i)^2$. As RSS describes the variance unexplained by the model and TSS describes the variance of the dependent variable, $\frac{RSS}{TSS}$ simply quantifies the proportion of unexplained variance.

Because of the first fact, linear models can also be used to correct for covariates. To perform such correction, for each feature in the dataset, a linear model is fitted with the feature as the dependent and all covariates as the independent variables. The residuals of these models then display the variance that cannot be explained by the covariates, i.e. the data corrected for these covariates.

The residuals and explained variance of linear models between the substrates and products of metabolic reactions play a crucial role in the inference of the reaction activity in one of the main publications introduced in this thesis [11].

To fit the parameters β of a linear regression, the residuals are also used as the objective of the so-called Ordinary Least Squares (OLS) method which minimizes the RSS (or equivalently the MSE). This is a very intuitive way to define the regression objective, as it simply tries to minimize the distance between the regression line and the observed data points. Interesting, it leads to the same solution as Maximum Likelihood Estimation (MLE), a procedure to fit the parameters of a distrubtion that maximizes the likelihood given some observed data [217]. This can easily be seen when defining the log-likelihood of a linear regression model (that MLE tries to maximize)

$$\begin{aligned} \ell(X, Y, \beta) &= \sum_i^N log\, p(y_i | x_i, \beta) \\ &= \sum_i^N log \left(\frac{1}{\sigma\sqrt{2\pi}} exp \left[-\frac{1}{2} \left(\frac{y_i - x_i^T \beta}{\sigma} \right)^2 \right] \right) \\ &= -\frac{N}{2}\, log(\sigma^2 2\pi) - \frac{1}{2\sigma^2} RSS \end{aligned} \tag{2.1}$$

which shows that minimizing the RSS maximizes the log-likelihood - thereby maximizing the likelihood of the model.

Linear models and the theory behind residuals have a pivotal role in the idea of approximating changes in metabolic reaction activity in mantra [11] (Section 3.2, which, intuitively speaking, uses a metric derived from the residuals to measure how reaction activity deviates between sample groups, such as healthy and disease.

According to the Gauss-Markov theorem, the OLS estimator is the best *unbiased* estimator of β [218, 219][35]. However, when data is noisy or collinear[36], introducing bias to reduce the variance can be advantageous. The problem of fitting parameters to the random noise present in training

[35] Briefly, a model's MSE can be decomposed into the sum of variance and squared bias - the so-called bias-variance tradeoff. By showing that the OLS minimizes the variance, it can be concluded that only biased estimators may result in a smaller MSE.

[36] Arguably the default for biological data

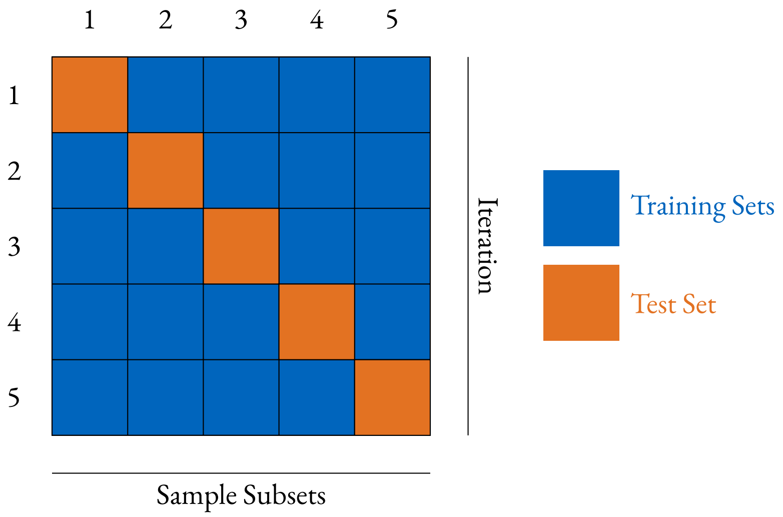


Figure 2.7: Schematic of the k-fold Cross-Validation (CV) procedure. Each row indicates one iteration, each column one set of samples in the overall training data set. In every iteration, all blue sample subsets are used for training the model and only samples in the orange subset are used to evaluate model performance.

data is referred to as *overfitting* [220]. In general, the higher dimensional (relative to the number of samples) the data and the more parameters a model has the more likely it is to run into overfitting [220]. For regression models, Ridge and Lasso regression, both shrinkage methods that aim to maximize the product of the log-likelihood and a prior distribution over the model parameters, are commonly used to reduce overfitting [217]. In Ridge regression, a Gaussian prior is used, leading to the penalty being the squared 2-norm of the parameters (ℓ_2 regularization) $||\beta||_2^2$, whereas in Lasso, the prior distribution is a Laplace distribution and the penalty is simply the sum of the absolute parameter values (ℓ_1 regularization) $||\beta||_1$ [216][37], [38].

A more general way to assess a model's degree of overfitting is CV. In k-fold CV, for example, the *training* samples are split into k subsets (see Figure 2.7). The model to assess is trained and evaluated k times, each time leaving out a different subset of samples. Left-out samples are used to evaluate the model's performance on *unseen* data after training. Evaluation metrics or curves obtained this way can be used to estimate the actual performance of the model [216]. Naturally, CV cannot rule out all factors rendering model evaluations incorrect, and correct data processing and splitting is crucial to avoid the rather widespread phenomenon of leakage [221].

Especially for my contributions to analyzing the clinical data in Häcker and Siebert et al. [222] CV played an important role in demonstrating the robustness of the results on (noisy) patient data.

Having established the most basic concepts of regression, we now move on to prediction tasks

[37] The added penalties represent a form of bias, thus using shrinkage methods trade in variance for bias in the hope of reducing the MSE.

[38] In contrast to the normal distribution, in which $y_i - x_i^T\beta$ is squared in the exponential, the probability density of the Laplace distribution only uses the absolute difference. This leads to a higher probability for μ in the Laplace distribution and thus sparser β when using ℓ_1 regularization.

starting with a simple example of binary classification with the classes labeled 0 and 1, i.e. $y \in \{0, 1\}$. In fact, feeding the results of a regression into a sigmoid function and treating the output as a probability for class 1 one obtains a prediction model. The log-likelihood can be derived analogous to Equation (2.1) with $p(y_i|x_i, \beta) = Ber(y_i|sigm(x_i^T\beta))$, where $Ber(\cdot)$ indicates the Bernoulli distribution and $sigm(\cdot)$ the sigmoid function, as

$$\begin{aligned} \ell(X, Y, \beta) &= \sum_i^N log\, p(y_i|x_i, \beta) \\ &= \sum_i^N y_i \ \log \ sigm(x_i^T\beta) + (1 - y_i) log(1 - sigm(x_i^T\beta)) \end{aligned} \tag{2.2}$$

The negative version of this function is called the *cross-entropy* - a very common loss function for prediction tasks [217]. Unlike linear regression, logistic regression does not have a closed-form solution. One way of learning its parameters β is by *gradient ascent/descent*, a simple optimization method that uses the gradient of a function with respect to its parameters for stepwise optimization. The parameters in the n^{th} iteration are updated as $\beta_{n+1} = \beta_n + \eta\frac{\partial \ell}{\partial \beta}$[39].

While linear models have not been used for prediction tasks in my main projects, they are the basis for a number of analysis pipelines developed and used in co-author publications, such as Coleman and Sorbie et al. [155] and Häcker and Siebert et al. [222].

While different types of ML algorithms such as Random Forests [223], Gradient Boosting [224], or Support Vector Machines [225] are commononly used, a large focus nowadays lies on Artificial Neural Networks (ANNs). The earliest networks developed were so-called Multilayer Perceptrons (MLPs) - multiple layers of regressions with non-linear "activation" functions, such as the sigmoid [217]. Already in 1989, Hornik, Stinchcombe, and White [226] showed that such networks, if arbitrarily deep, are theoretically capable of approximation *any* function - the so-called Universal Approximation Theorem. To train MLPs, a procedure termed backpropagation is used. It is based on a simple idea: because an MLP is composed of individual functions - regressions and non-linearities - the gradient for any layer can be computed by consecutively applying the chain rule to the loss function, and thus gradient descent be employed for training. Nowadays, a plethora of computational frameworks capable of automatic differentiation, such as PyTorch [227], Tensorflow [228], and JAX [229] have been developed, greatly facilitating the implementation of ANNs.

Expanding on the MLP, numerous general architectures, such as Convolutional or Recurrent Neural Networks, were developed to tackle different types of problems. More recently, GDL, focussing on the analysis of non-Euclidean data [230], has gained traction holding interesting promises for the analysis of structured (biological) data. Even more prominently, the introduction of the Transformer architecture in 2017 by Vaswani et al. [231] has allowed for great progress not only in natural language processing [232] but also bioinformatics tasks, such as protein structure prediction [159]. While these architectures have not been directly used in this thesis, they will play

[39] Note that the formula is showing gradient *ascent*, because Equation (2.2) defines the log-likelihood, which requires maximization. In typical ML literature gradient descent ($\theta_{n+1} = \theta - \eta\frac{\partial \mathcal{L}}{\partial \theta}$) is described, where $\mathcal{L}$ describes the loss function (e.g. for logistic regression the cross-entropy loss) to be minimized.

		Predicted Label	
		Positive	Negative
True Label	Positive	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

Table 2.3: Layout of the confusion matrix describing typical performance evaluation metrics.

Metric	Formula
Sensitivity	$\frac{TP}{TP+FN}$
Specificity	$\frac{TN}{TN+FP}$
Precision	$\frac{TP}{TP+FP}$
Recall	$\frac{FN}{TP+FN}$

Table 2.4: Performance metrics for classification tasks based on the confusion matrix. The selection presented is based on the scores used to assess generalization performance in the mantra paper [11] (Section 5.1).

a key role in the discussion of the future direction of the field later on.

Irrespective of the architecture of a ML model, it is crucial to have good evaluation metrics. For a binary classification task, a simple metric is the accuracy - the proportion of correct predictions. To define it more formally, we can use the elements of the (binary) confusion matrix (Table 2.3):

$$Acc = \frac{TP + TN}{TP + FP + TN + FN}$$

While this looks like a good first guess, the accuracy can be tricked very easily. Consider the case where 95% of all samples have a positive label. A classifier that simply always predicts a positive label will have 95% accuracy, despite not having learned any relationship between features and labels.

A better metric to use is, for example, the Area under the Receiver Operating Characteristic Curve (ROC-AUC) [233]. The Receiver Operating Characteristic (ROC) curve is based on two metrics that can be computed directly from the confusion matrix: sensitivity and specificity (see Table 2.4). The former describes the proportion of positive cases detected, i.e., the true positive rate, while the latter specifies the fraction of negative samples classified correctly, i.e., the true negative rate. Intuitively it is easy to see that both these metrics should be high. Take the example of COVID-19 diagnostic tests: a test should detect as many infected people as possible - a high sensitivity - while reporting as few uninfected people as "positive" as possible - a high specificity - to avoid unnecessary restrictions.

In contrast to an actual binary test, a classification model will typically give a probability to which a cutoff has to be applied to get a class label. The ROC curve is generated by computing sensitivity and specificity for "all" cutoffs between 0 and 1 [234]. This way, the ROC curve also shows how confident classifier predictions are. To get a single metric out of a ROC curve, the Area Under the Curve (AUC) is computed. It is a value in the range $[0, 1]$ where 0.5 is considered random classification (for a binary task) and 1 is a perfect classifier. Besides ROC, Precision Recall (PR) (Table 2.4) is another commonly used curve, computed analogously to the procedure described above.

While there are many more evaluation metrics with properties suited for different scenarios, the selection presented in this thesis is based on those metrics relevant for the evaluations in Köhler et al. [11], which is presented in more detail in Section 3.2 and Section 5.1, as well as the results of my contribution in Häcker and Siebert et al. [222].

Approaches *not* using training labels, such as dimensionality reduction and clustering, are called *unsupervised*. With respect to clustering, a large range of approaches has been developed.

Algorithm 1 Pseudocode to compute k-means clustering

```
Input
Data points to cluster: $X = \{x_0, x_1, \ldots, x_{N-1} | x_i \in \mathbb{R}^n\}$
Number of clusters: $k \in \mathbb{Z}^+$
Output
Cluster assignment for each data point:
$c = \{c_0, c_1, \ldots, c_{N-1} | c_i \in \mathbb{Z}^+_{<k}\}$
$\mu_0, \mu_1, \ldots, \mu_{k-1} \in \mathbb{R}^n \leftarrow random()$
while not converged do
    for all $i < N$ do
        $c_i = \arg\min_j \|x_i - \mu_j\|$
    end for
    for all $j < k$ do
        $m = \sum_i^N [c_i = j]$
        $\mu_i = \frac{1}{m} \sum_{i<N; c_i=j} x_i$
    end for
end while
```

In general, they all try to aggregate similar samples into the same cluster and dissimilar ones into different clusters. One very simple example of a clustering algorithm is k-means [235] - a short overview of its workings is given in Algorithm 1. Two downsides of k-means are that the number of clusters k has to be decided *a-priori*, which is not always trivial, and that it is non-deterministic due to the randomly initialized centroids. Hierarchical clustering, in contrast, is an example of a deterministic clustering method. Different types of clustering exist, for example, simple step-wise merging or separating samples by similarity in a hierarchical fashion [236] or using the idea that clusters are more dense regions [237, 238].

While clustering groups either samples or features based on their similarity, biclustering is an extension of "1D" clustering to simultaneously group subsets of samples and subsets of features. Biclustering is a cornerstone of two publications I co-authored during my doctorate ([239] and [240]).

2.3 Objectives

With this chapter having provided an introduction to the topics relevant to understanding the work in this thesis, we can now proceed to define what its objectives are and where they come from. The overarching goal underlying this thesis is the apparent gap between the availability of metabolomics and lipidomics experiments paired with their relevance for understanding socially and economically important disease conditions and the lack of computational methods for efficient interpretation and mechanistic hypothesis generation. Not only is this lack of algorithmic solutions hindering the progression of the field itself it also slows down the integration with multi-omics data and translational research. Therefore, the main objective of my work is to introduce algorithmic solutions that make lipidomics and metabolomics data interpretation more readily available to non-computational researchers and reduce the workload required to get to testable and mechanistic hypotheses.

The theme unifying all projects is the use of graph-theoretic approaches, in particular metabolic networks. As the availability of networks for lipid metabolism is very limited, the first main aspect of my work focused on making lipid metabolic networks available and analyzable in two consecutive first-author publications introducing the Lipid Network Explorer (LINEX) platform [9, 10]. The first publication lays the foundation by describing a method for generating lipid metabolic networks specific to a given data lipidomics data set. To demonstrate the usefulness of such graphs, it also shows that including quantitative measurements by superimposing statistical metrics on them allows for the direct extraction of biological insights. The second publication extends the network generation procedure to a database-driven scheme, providing a starting point to mechanistically incorporate lipidomics with, for example, proteomics. Furthermore, this paper introduces a novel enrichment algorithm for extracting mechanistic hypotheses from lipidomics data. Their underlying methodology is described in Section 3.1.1 and Section 3.1.2 while the publications are summarized in Section 4.1 and Section 4.2.

The third first-author publication of my thesis focuses on generalizing the idea of using graphs to identify aberrant metabolic reactions [11]. It introduces a metric that approximates how metabolic reactions change their activity between conditions in a sample-specific manner allowing for the integration with genome, transcriptome, proteome, and microbiome data. Together with a novel multi-omics network enrichment method, the publication tackles the unavailability of methods for interpreting metabolomics data with respect to alterations in metabolic reactions. The approximation and network enrichment algorithm are presented in Section 3.2, and the publication is summarized in Section 5.1.

3 Methods

The following chapter will provide an overview of the methods developed during my dissertation, with the first part focussing on lipid network analysis in two sequential publications (for full publications, see Appendix A.1 and Appendix A.2). The second part introduces the mantra framework for analyzing metabolic reactions in a multi-omics context (full publication in Appendix A.3). Both parts provide a comprehensive summary of the algorithms; readers interested in the full details are referred to the respective publications.

3.1 Lipid Network Analysis

3.1.1 Rule-Based Lipid Network Analysis for Functional Lipidomics Analysis

The Lipid Network Explorer (LINEX) [9] is a framework for generating and analyzing lipid metabolic networks specific to the lipids measured in a given lipidomics experiment. An overview of its workflow is provided in Figure 3.1.

Naturally, the first step of the pipeline is generating the network. However, due to the non-standardized nomenclature in lipidomics, obtaining a unified lipid species representation before computationally processing lipid names is crucial. For this purpose, LINEX internally converts all lipid names into the same nomenclature style using LipidLynxX [241].

The procedure LINEX utilizes builds on the fact that lipid metabolism is essentially composed of Fatty Acid (FA) and complex lipid metabolism (recall Section 2.1.2.2). For every possible pair of lipid species, the network contains an edge if one of the following mutually exclusive conditions is satisfied:

1. (**Lipid Class Connection**) Both lipids have the same *FA composition*, and the reaction rules allow a conversion between the two lipid classes.
2. (**Fatty Acid Connection**) Both lipids have the same *lipid class*, all but one FAs are identical, and there exists a reaction rule that allows the conversion between the non-identical FAs.

For lipid class connections, one "special" case is lipid species pairs with the same head group but different numbers of FAs, e.g., Lyso-Phosphatidylcholine (LPC) and PC (1 and 2 FAs) or DG and TG (2 and 3 FAs). In this case, the lipid species with one more FA must contain all FAs that the lipid species with one less FA[1]contains. The "missing" FA must be in a list of user-defined FAs that should be specific to the organism from which the analyzed samples originate.

[1] A special exception is the case of MG and TG, which have the same head group but their number of FAs differs by 2. In this case, LINEX does not infer metabolic connection.

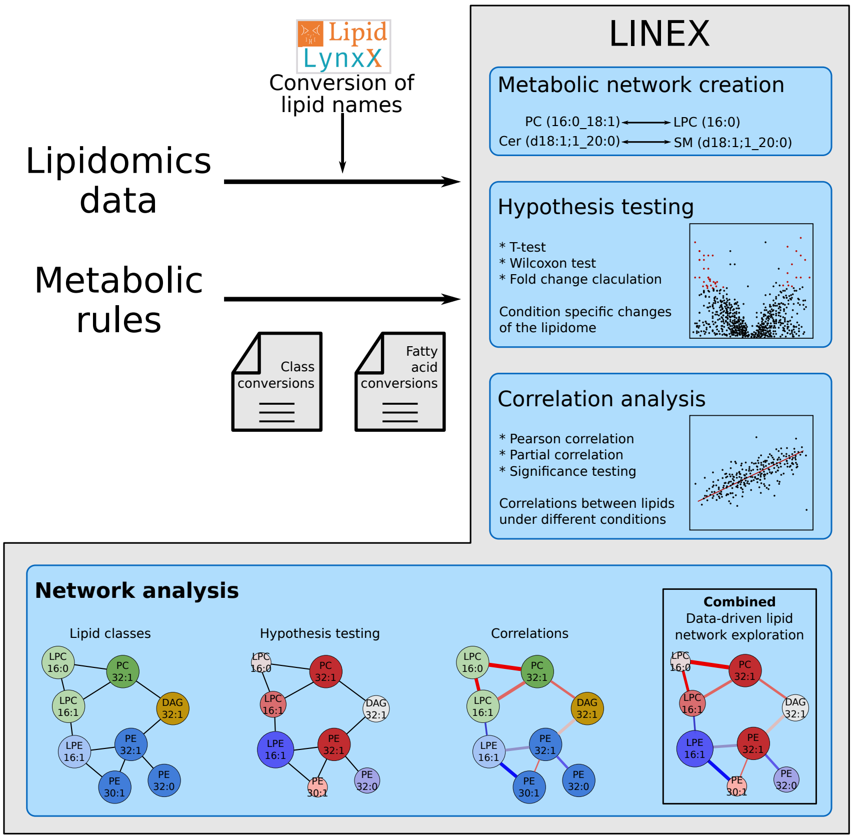


Figure 3.1: Overview of the LINEX workflow as described in [9]. The top left shows the input required: lipidomics data and metabolic rules used to generate the lipid metabolic network (top right). Subsequently, metrics for condition-specific changes for all nodes (lipid species) and associations between metabolically associated lipid species under biological conditions are computed. A combined visualization can be achieved by superimposing these metrics onto the network (bottom panel). The figure is originally from Köhler and Rose et al. [9] (unmodified) and distributed under the CC BY license.

A second special case that can occur due to current technical limitations of lipid identification is that two lipid species were measured on different levels of resolution (recall Section 2.1.3), i.e., one is a sum (only the sum of FAs is known) and the other a molecular lipid species (the exact FAs composition is known). The coarser identification level is used in this case, as it can be easily inferred for the more detailed lipid species.

For fatty acid connections, there are three possible reactions (by default):

1. Elongation
2. Desaturation
3. Hydroxylation

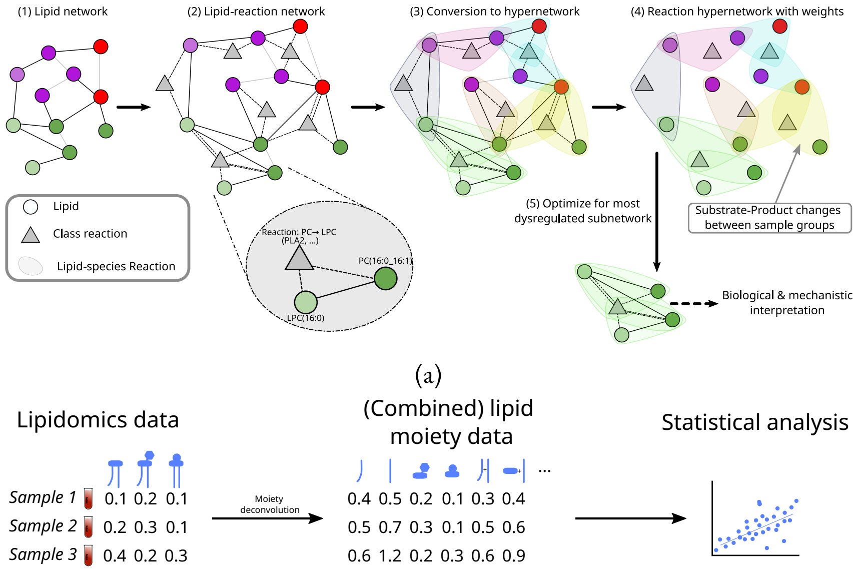
Elongation refers to increasing the chain length by 2, desaturation to increasing the number of double bonds by 1, and hydroxylation increasing the number of hydroxy groups by 1. In the case of sum species lipids, the sums are used as proxies for the FAs. This is justifiable under the assumption that the FA pool is the same for both lipid species. It is important to note that fatty acid connections are not directly happening, as only free fatty acids are modified, and thus de- and re-acetylation are required. Nevertheless, adding these types of reactions increases the network's connectivity in a biochemically meaningful way, improving the visualization of biological effects and downstream analyses. As all pairwise combinations of measured lipid species need to be evaluated to generate the full metabolic network, $n * (n - 1)$ checks of the above rules are required. Nevertheless, network generation for typical lipidomics datasets is fast, even on standard laptop hardware.

While the lipid network generated is specific to the measured data, it does contain any information regarding changes between biological conditions. To enable such analyses quantitatively, LINEX computes commonly used statistical metrics, most prominently p-values from (parametric or non-parametric) hypothesis tests and log-Fold Changes (logFCs) for all vertices (lipids) and (partial) correlations per group and their changes between groups (pairwise) for each edge. By superimposing these measures onto the network, biological changes are easily visualized *together* with the biochemical connections between lipids.

3.1.2 Enzyme-Focussed Lipid Network Enrichment to Infer Mechanistic Hypotheses

Because of its customizability, which is especially interesting for research on less-studied organisms, LINEX does not contain connections to databases and thus cannot draw direct connections to proteins. Therefore, LINEX^2 [10] is not based on reaction rules for lipid class connections but instead uses metabolic reactions from Rhea [242] and Reactome [243] to infer lipid metabolic networks. Additionally, the procedure of drawing FA connections between pairs of lipids involving at least one sum species was modified. Instead of comparing these lipids on the sum level, all possible molecular species combinations are enumerated based on the list of expected FAs. This increases the runtime, especially for datasets with a high number of TG and Cardiolipins (CLs) sum species, which carry three and four FAs, respectively, but allows for more detailed qualitative inference.

With the database-based lipid class connections, including enzyme annotations for each reaction, the basis for algorithms inferring mechanistic hypothesis on *enzyme level* using lipidomics data


(1) Lipid network
(2) Lipid-reaction network
(3) Conversion to hypernetwork
(4) Reaction hypernetwork with weights
Lipid
Class reaction
Lipid-species Reaction
Reaction: PC→ LPC (PLA2, ...)
PC(16:0_16:1)
LPC(16:0)
(5) Optimize for most dysregulated subnetwork
Substrate-Product changes between sample groups
Biological & mechanistic interpretation
Lipidomics data
(Combined) lipid moiety data
Statistical analysis
Sample 1 0.1 0.2 0.1
Sample 2 0.2 0.3 0.1
Sample 3 0.4 0.2 0.3
Moiety deconvolution
...
0.4 0.5 0.2 0.1 0.3 0.4
0.5 0.7 0.3 0.1 0.5 0.6
0.6 1.2 0.2 0.3 0.6 0.9

(a)

(b)

Figure 3.2: Schematic of novel lipid-specific algorithms introduced by LINEX^2 [10]. *(a)* Schematic of the Lipid-Enzyme enrichment algorithm. Starting with a lipid-lipid graph (1), a lipid-enzyme graph is generated by connecting each lipid to an enzyme node if it catalyzes a reaction involving it. All substrate-product-enzyme n-tuples are converted into hyperedges (3), which then serve as the nodes of a "reaction graph" in which every node is a hyperedge and every edge indicates a shared lipid species (4). The substrate-product ratio is computed for every reaction node in this graph, serving as a proxy for the change in enzymatic activity. Using combinatorial optimization, a subgraph (sub-optimally) maximizing the change in substrate-product ratio between two biological conditions is identified and reported as a mechanistic hypothesis. The figure is originally a subfigure from Rose and Köhler et al. [10] (unmodified) and is distributed under the CC BY license. *(b)* Schematic of the substructure analysis. All individual substructures in the original lipidomics data are identified, and all pairwise combinations are enumerated. For each substructure (combination), the total abundance of all lipid species containing it is calculated per sample, leading to a new feature matrix that can be used as input for downstream analyses. The figure is originally from the supplementary material of Rose and Köhler et al. [10] (unmodified) and is distributed under the CC BY license.

only is laid. Therefore, LINEX^2 introduces a network enrichment algorithm specifically tailored to the peculiarities of lipid metabolism. First and foremost, it utilizes the fact that each lipid class reaction is (typically) observed more than once, as many pairs of lipid species display substrates and products of the same reaction. For example, *any* PE species will be transformed into the respective PC species via tri-methylation. To leverage this *multispecificity*, the first step of the proposed enrichment algorithm is to turn the lipid-lipid network into a hypergraph-like structure, where each hyperedge is a n-tuple of the substrate and product species as well as the class-reaction node, that contains information about the enzyme catalyzing the reaction (steps 2 and 3 in Figure 3.2a).

Subsequently, the ratio of substrate to product concentrations for each hyperedge is computed while accounting for the number of substrates and products. The change in this ratio between biological conditions can then be used as a proxy for the degree of enzymatic dysregulation. To obtain a mechanistic hypothesis on the enzymatic changes between conditions, a local search together with simulated annealing aiming to maximize the mean change in substrate-product ratio is used. Local search is a local optimization approach that iteratively performs single modifications to the current solution, i.e., the subgraph that improves a given objective function. Because of its local nature, it can easily get stuck in local optima. To increase the chances of finding the globally optimal solution, simulated annealing randomly allows accepting solutions with a worse objective score.

In addition to this network enrichment algorithm, which focuses on identifying changes in the relationships of lipids, LINEX^2 also contains a complementary substructure analysis that aims at identifying changes in the abundances of lipid substructures (head group, backbone, sum composition, etc.). It is based on a method for analyzing glycomics data from Bao et al. [244]. Essentially, for each combination of substructures, the abundances of all lipids containing them are summed, generating a sample × substructure matrix (Figure 3.2b). These newly generated features are then analyzed statistically to find the most changing substructures.

Both LINEX versions are fully open-source and available in a web tool, including interactive result visualization, to make it accessible to the entire lipidomics community regardless of programming knowledge. LINEX^2 is additionally available as a `pip`-installable python package on PyPI.

3.2 Inferring Changes in Metabolic Reaction Activity and Reaction-Centered Multi-Omics Integration

The third main publication of my thesis introduces a method called Metabolic Network Reaction Analysis (mantra) [11] that approximates changes in reaction activity between biological conditions and allows their integration in a multi-omics context.

The first step in the mantra pipeline (Figure 3.3) is the generation of data-specific metabolic networks. In order to have a comprehensive knowledge basis covering mammalian and microbial metabolism, a merged database combining KEGG [49], Reactome [245], and Virtual Metabolic Human [246] was generated. Before computing the data-specific subgraph, the metabolomics data given as input needs to be mapped to the identifiers used in the database. For this purpose, internal identifier maps, if input metabolites already have database identifiers, or the MetaboAnalyst [247] name conversion API, if the input is metabolite names, are employed.

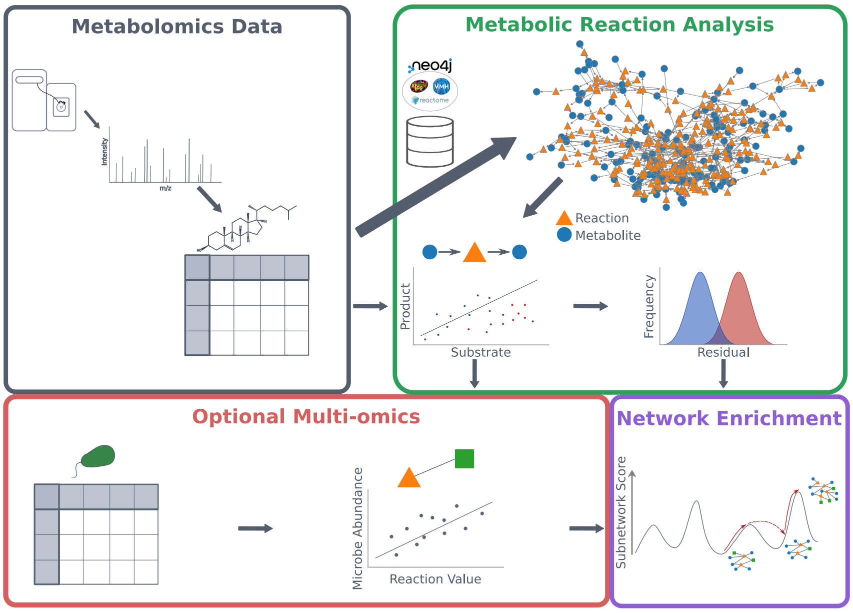


Figure 3.3: Overview of the mantra [11] workflow.The input data, metabolomics data containing identified metabolite abundances (grey box), is mapped onto a combined database to obtain a bipartite metabolic network only consisting of metabolic reactions with at least one measured substrate and one measured product (green box, top). For each reaction in the subgraph, a linear model is fitted to approximate the change in reaction activity (green box, bottom). Using the distributions of these estimates, a network enrichment algorithm identifies subgraphs of reactions that maximally change between biological conditions (purple box). Given multi-omics data, correlations between metabolic reactions and the multi-omics features can be computed (red box). These correlations can also be used for multi-omics network enrichment. The figure is originally from Köhler et al. [11] (unmodified) and is distributed under the CC BY license.

The metabolic network contained in the database is a bipartite, directed graph (recall the definition from Section 2.2.1) containing metabolite and reaction nodes. In addition, the graph contains information on which genes encode the enzymes catalyzing a given reaction and which organisms carry these genes. To subset this graph for the given input metabolites, all metabolites, as well as all "reactions for which *at least* one substrate and one product are measured" [11], are extracted.

With the assumption that quantitative relations between substrate and product abundances change when metabolic reactions change their activity, we can quantify the degree of activity change by the degree to which substrate-product relations change. Therefore, for every reaction in the subgraph, a linear model with substrate abundances as independent and product abundances as dependent variables is computed using the samples of one group. Subsequently, the change in reaction activity for *all* samples is approximated by the change in residuals, which quantify the unexplained variance (recall Section 2.2.2) and thus the degree to which the relationships change. To obtain a relative value independent of the number of products, the final estimate is defined as

$$a = 1 - \frac{residual}{TSS}$$

with TSS being the total sum of squares. This way of approximation yields a sample × reaction change matrix. For testing how statistically significant a change in the activity of a given reaction is, a p-value from a Wilcoxon rank sum test is reported as well.

While the obtained p-values are sufficient to rank metabolic reactions, deriving hypotheses from such lists without knowing the relations between them can be difficult. Therefore, mantra also introduces a network enrichment for providing mechanistic hypotheses in the form of metabolic subgraphs. Since the method is focused on the change in metabolic reaction activity, the bipartite, directed graph is first converted into an undirected graph only containing reaction nodes, where an edge indicates two reactions sharing *at least* one metabolite. The enrichment algorithm uses local search with simulated annealing similar to the enrichment of LINEX^2 [10] with adaptions to deal with the differences between general metabolic and lipid-specific metabolic networks. The objective to maximize is the change of reaction activity per node in the subgraph.

Besides metabolomics-only analyses, sample-wise reaction estimates enable the integration with other omics data, such as microbiome or proteome data. mantra leverages this fact by computing correlation coefficients between reaction activity estimates and multi-omics data. Furthermore, genome, transcriptome, proteome, and microbiome data can be mapped onto the database to restrict non-zero correlation coefficients to known connections. In this case, the proposed framework also includes a modified version of the enrichment algorithm with the objective of finding a subgraph maximizing the correlation between the multi-omics features and metabolic reactions simultaneous to maximizing the change in reaction activity.

mantra is open-source and available as a python package on PyPI. The database is provided as a download for local usage or can be queried online via a REST API.

4 Publication Summaries

4.1 Investigating Global Lipidome Alterations with the Lipid Network Explorer

Citation

Nikolai Köhler[†], Tim Daniel Rose[†], Lisa Falk and Josch Konstantin Pauling "Investigating global lipidome alterations with the lipid network explorer" *Metabolites* 2021, 11 (8), 488; doi: 10.3390/metabo11080488

[†] These authors contributed equally

Summary

The analytical side of lipidomics has seen great progress in the last decade, and lipidomics experiments are popular in biological and biomedical studies. Nevertheless, data interpretation algorithms for lipidomics are still understudied. Comprehensive lipid metabolic networks, in particular, are unavailable due to the discrepancy between measurements being on species-level but databases separating lipid class and fatty acid metabolism (recall Section 2.1.2.2).

To bridge this gap and allow lipidomics data interpretation to use graph-theoretic approaches for focussing on metabolic reaction, we introduce the Lipid Network Explorer (LINEX), a framework for generating lipid metabolic networks based on metabolic rules specific to the lipids measured in an experiment. The method uses metabolic rules for both lipid class and fatty acid metabolism to infer direct metabolic connections between all lipid species measured in an experiment. By superimposing statistical metrics for lipid species and correlations for all metabolically associated pairs of lipids onto the network, LINEX visualizations enable the derivation of quantitative functional associations.

To showcase its capabilities, we used LINEX to analyze publicly available lipidomics data from three independent publications that span a range of common experimental setups. The first dataset measured lipids on sum-species level in cancerous and normal mucosa from colorectal cancer patients. We show that while the original paper did not find globally changing patterns in lipid abundances, the LINEX network highlights that a few highly connected subgraphs exist in which the correlation between lipid species changes significantly between "healthy" and cancer mucosa. On a second dataset with much lower coverage, we demonstrate how the incorporation of fatty acid connections facilitates interpretation and how missing lipid classes (or even species) can lead to unconnected components. Lastly, we illustrate how the novel visualizations highlight changes in metabolically connected areas.

Because LINEX is made available as a web service, it offers a new way of analyzing lipidomics data complementary to existing approaches to all researchers in the lipidomics community regardless of their programming knowledge.

Contributions

Together with co-first author Tim Daniel Rose I developed and implemented the method for network generation, all visualizations, and the web service, as well as contributing to the analysis presented, writing, and revision of the manuscript.

Contributions as stated in the original 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. All authors have read and agreed to the published version of the manuscript.” [9]

Availability, Rights & Permission

The publication is available in Appendix A.1, including links to the source code and web application.

Rights & Permission as stated in the original publication: “©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/).” [9]

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

Citation

Tim Daniel Rose†, Nikolai Köhler†, Lisa Falk, Lucie Klischat, Olga Lazareva and Josch Konstantin Pauling "Lipid network and moiety analysis for revealing enzymatic dysregulation and mechanistic alterations from lipidomics data" *Briefings in Bioinformatics* 2023, 24 (1), bbac572; doi: 10.1093/bib/bbac572

† These authors contributed equally

Summary

Despite some recent progress in lipidomics data interpretation on the side of pathway enrichment, no *de-novo* network enrichment methods are available. With the development of LINEX, we laid the foundation for such approaches, yet without introducing a specific network-analysis algorithm with it.

Therefore, we developed LINEX^2 , an extended version of LINEX addressing limitations in network generation that restrict multi-omics connections by using database information for lipid class reactions. Furthermore, we introduce a novel enrichment algorithm specifically designed for the peculiarities of lipid metabolism, aiming at generating hypotheses on changes in enzyme activity.

As a proof of concept, we applied the enrichment algorithm to a lipidomics dataset containing wild-type and MBOAT7 knockout mice. Our results show that LINEX^2 successfully identifies the reaction catalyzed by MBOAT7 as the most dysregulated, even indicating the known fatty acid preferences of MBOAT7. To further showcase the applicability to clinical data, we applied our enrichment method to white adipose tissue lipidomics data from lean and obese humans, with the resulting hypothesis indicating changes in membrane composition for adipocyte expansion. Additionally, we analyzed this data with a substructure analysis method introduced in LINEX^2 . It revealed particular patterns of increased abundances in neutral lipids depending on the sum composition.

Our results showed that LINEX^2 can recover knocked-out reactions purely from lipidomics data and thus provide reasonable hypotheses on the dysregulation of lipid metabolic enzymes. In combination with the complementary analysis of lipid substructures, the framework can identify different types of influences on the lipidome providing valuable interpretations.

Contributions

Together with Tim Daniel Rose (co-first author) I developed the web service and, additionally with Olga Lazareva, the network enrichment method. Furthermore, I ran parts of the evaluation and co-wrote and revised the manuscript.

Contributions as stated in the original publication: "J.K.P. supervised the project and secured the funding. N.K., T.D.R. and J.K.P. planned and conceptualized the work. N.K. and T.D.R.

developed the web service. N.K., O.E.L. and T.D.R. designed and implemented the network enrichment procedure. L.F., L.K. and T.D.R. parsed and curated the reaction databases, and implemented the network extension. N.K. and T.D.R. applied, validated and interpreted the approach on lipidomics data. N.K., O.E.L., T.D.R. and J.K.P. wrote the manuscript. All authors read, reviewed and accepted the manuscript in its final form." [10]

Availability, Rights & Permission

The publication is available in Appendix A.2, including links to the source code, analysis code, and web application.
Rights & Permission as stated in the original publication: "©The Author(s) 2023. Published by Oxford University Press. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited." [10]

5 Unpublished Results

5.1 Identification and Integration of Key-Metabolic Reactions from Untargeted Metabolomics Data

At the time of submission, the following manuscript was still undergoing peer-review. However, it reflects a significant contribution to my research as a doctoral candidate and is therefore included in this thesis. This chapter summarizes the manuscript, which is published as a preprint [11] and available in full in Appendix A.3, as well as my contributions.

Citation

Nikolai Köhler, Vivian Würf, Tim Daniel Rose and Josch Konstantin Pauling "Identification and Integration of Key-Metabolic Reactions from Untargeted Metabolomics Data" *bioRxiv* 2023; doi: 10.1101/2023.05.15.540613

Summary

Despite the popularity of metabolomics in biological and biomedical research, computational methods for data interpretation are still a (if not the) major bottleneck in the metabolomics workflow. Available methods commonly focus on computing significances and incorporate (metabolic) relations between features only downstream, if at all.

In this publication, we introduce Metabolic Network Reaction Analysis (mantra), an approach to estimate how metabolic reactions change between biological conditions by exploiting the relationships between their substrates and products. mantra allows the identification of metabolic subnetworks corresponding to areas of high metabolic changes as well as integrating metabolomics data and metabolic reactions with multi-omics data.

To demonstrate the ability of mantra to approximate how metabolic reactions change their activity, we applied the activity estimation method to untargeted metabolomics data from a study of breast biopsies of Triple-Negative Breast Cancer patients and non-Breast Cancer subjects. Our results showed that our proposed method was able to recover metabolic reactions whose catalyzing enzymes are either associated with breast cancer risk or prognostic biomarkers.

We also presented results of the enrichment algorithm on stool metabolomics comparing non-Inflammatory Bowel Disease and Crohn's Disease patients. These showed that the subnetwork identified contains reactions whose catalyzing enzymes are significantly differentially abundant in metagenomics results of the same samples, except for one reaction. Furthermore, we could show that a simple random forest classifier trained on the reaction activity estimates of the discovery cohort is able to classify patients from an independent validation cohort. To showcase the ability of our framework to integrate reaction activity changes with multi-omics data, we further showed how correlation analyses with microbial species abundances yield interesting patterns connecting microbes and metabolic reactions.

In conclusion, we have shown that mantra is able to generate valid hypotheses on the change in activity of metabolic reaction and its effectiveness in providing a mechanistic approach for the multi-omics integration of metabolomics data.

Contributions

I designed and implemented the methods and the python package containing it, ran the evaluations, wrote the manuscript draft, and revised it. Furthermore, I supervised the parsing and merging of the metabolic databases.

Contributions as stated in the original publication: "NK and JKP planned the work. NK designed and implemented the reaction activity estimation method. NK and TDR designed the network enrichment procedure and NK implemented it. NK ran the evaluations. VW parsed and merged the reaction databases. NK, JKP, VW and TDR wrote and reviewed the manuscript. JKP supervised the project and secured the funding. All authors accepted the manuscript in its final form." [11]

6 Discussion

Owing to improvements in Mass Spectrometry (MS)-based analytics, metabolomics and lipidomics are suitable for studying large-scale experiments and clinical cohorts nowadays. However, their use, both in isolation and integrated with other omics data, is limited by the availability of computational methods to interpret the vast amounts of data generated. This restricts not only the progress of the field itself but also the speed by which insights from biological and biomedical research can be acquired. Therefore, alleviating the bottleneck of data interpretation is critical to fulfilling the promises both fields hold [248, 249]. To address this gap in computational metabolomics, this thesis provides new approaches to interpret metabolomics and lipidomics data mechanistically and integrate them with other omics disciplines.

Designing such methods in a robust way requires dealing with multiple peculiarities of metabolomics and lipidomics data. Some of these are general to all biological data, particularly noisiness, sparsity, and high variance. At the same time, the (comparably) small number of samples per experiment, due to a lack of *true* multiplexing methods and single-cell approaches, is more drastic than in other omics disciplines. This leads to an especially challenging starting point as smaller sample numbers make it harder to estimate noise in data and, thus, approximate underlying distributions or manifolds. Furthermore, combined with the high dimensionality of (untargeted) experiments, the curse of dimensionality [250] presents a major issue, requiring careful evaluation of potential overfitting in supervised settings. When incorporating prior knowledge, lipidomics suffers from a lack of available metabolic models. For metabolomics, the main limitation factor is still the identification of metabolites and the mapping of identified metabolites onto database identifiers.

The previous chapters have provided an overview of the computational and biological background and presented the methods and results originating from my doctorate. In the following sections, I will put these into the context of the current state-of-the-art of the field, discuss how future developments can build on them, and conclude with an outlook on the open problems of the field.

6.1 Mechanistic Lipidomics Analysis via Biochemical Graphs

Mechanistic lipidomics data interpretation is largely done manually by experts reviewing coarse-grained statistical results summarizing changes in lipid species, lipid class, or FA-related abundances. Automatic identification of hypotheses requires incorporating prior knowledge, for example, in the form of metabolic networks. Besides the LINEX framework introduced in this thesis, only one other freely available method capable of generating biochemical lipid networks, termed BioPAN [251, 252], exists.

As discussed in Rose and Köhler et al. [10], LINEX^2 includes reactions not included in BioPAN and enables the generation of more detailed lipid networks. In particular, LINEX^2 networks are capable of utilizing molecular lipid species resolution and can easily be modified to cover sn-species resolution and modified lipids (oxidized, sulfonated, etc.) [253], so-called epilipids [1]. Current improvements in analytical lipidomics, such as the re-introduction of Electron-Activated Dissociation (EAD) [254] and Electron Impact Excitation of Ions from Organics (EIEIO) [255], are promising techniques to improve the structural resolution of high-throughput lipidomics experiments in the future [2]. Such developments will enable more fine-grained LINEX^2 networks and analysis results, thus facilitating the knowledge that can be extracted from lipidomics experiments.

With the availability of such comprehensive networks, LINEX^2 also opens up two largely uncharted areas in computational lipidomics: graph theory and metabolic modeling, both of which have been successfully studied in other contexts of computational biology [256, 257].

One common use case of graph theory in bioinformatics is finding disease signatures [258]. Typically, such methods use combinatorial optimization to identify subgraphs in which the measured vertices, representing molecules, "significantly" (e.g., via p-values) differ between experimental conditions. Applying this paradigm to lipidomics data neglects the multispecificity of lipid metabolic enzymes and the non-mechanistic nature, as the observed changes in lipid concentrations are usually the consequence of changes in metabolic activity. Exclusively using statistical metrics such as p-values in this context also prevents the utilization of quantitative relationships between lipid species, which are potential indicators for changes in underlying lipid metabolic reactions. By focusing on these associations between lipid substrates and products (Section 3.1.2), LINEX^2 enables mechanistic interpretations, making it more interpretable than enrichment algorithms developed for PPI networks or GRNs, which can (mostly[3]) be applied to lipid metabolic networks as well.

Furthermore, LINEX^2 leverages the multispecificity ubiquitous in lipid metabolism in contrast to BioPAN [251, 252], which ranks individual "reaction chains". Multispecificity is explicitly encoded in the objective function used by LINEX^2 , which includes a penalty for the number of reaction nodes, thereby incentivizing solutions containing lipid species reactions representing the same lipid class reaction. This approach is sufficient for the enrichment analysis of LINEX^2 but does not allow incorporating multispecificity into non-enrichment analyses. Future research could develop this approach further by computing substrate-product ratio distributions for each lipid species reaction instead of averaging them per group. This would enable the definition of new metrics characterizing lipid species reactions in the context of their "parent" class reaction. Furthermore, lipid species reactions from the same class reaction can be clustered into subgroups using distributional similarities. This subgrouping can potentially uncover differences in FA specificity within the same lipid class reaction, for example. To generate hypotheses on the FA specificity of different lipid metabolic enzymes, which are largely unknown [259] and resource-

[1] During my doctorate, I also contributed to a review about the current status and future of computational methods for epilipidomics [253].

[2] Among other projects, I have supervised the development of an automated pipeline for lipid identification from EAD mass spectra.

[3] For example, the DOMINO algorithm from Levi, Elkon, and Shamir [182] cannot be used for lipid-species network enrichment as it relies on GO terms.

intensive to experimentally measure, additional omics data is required. For example, correlating the expression or phosphorylation status of enzymes catalyzing a given lipid class reaction - acquired via proteomics analysis - with the substrate-product ratios of each subgroup could be a first approach to computationally obtain hypotheses of FA enzyme specificity.

Because of their connections between lipid metabolic reactions and enzymes LINEX^2 networks enable the integration of multi-omics data, in particular genomics, transcriptomics, and proteomics, beyond conjecturing about FA specificity. Unlike general-purpose multi-omics methods, for example, those embedding different modalities into one latent space [260, 261], such integration methods would focus on mechanistically explaining connections between omics-layers instead of aiming to find signatures or axes of variation shared between them. Different classes of methods from other omics disciplines [262] could be adapted to analyze multi-layer networks centered around lipid metabolism.

In addition to multi-omics integration, multi-layer networks - now containing data from different steps of the lipidomics pipeline instead of different omics data - can also inform lipid identification with data interpretation and vice versa. Building experimental networks representing mass-spectral similarity or mass differences, e.g., through Global Natural Products Social Molecular Networking (GNPS) [263], and combining them with lipid metabolic networks can, therefore, help to improve both lipid identification as well as network generation itself [201]. The reason for this approach being especially promising for lipids lies in the fragmentation patterns of lipids, which are fairly analogous between lipid classes (Section 2.1.3.2) and mostly differ in their head group fragment and loss masses. Therefore, most types of reactions covered by LINEX^2 represent lipid modifications that (under optimal analytical conditions) result in spectrum shifts, e.g., through head group modifications or FA elongation or desaturation. Due to the limited number of mass shifts that can occur for a given lipid species, such approaches are promising candidates for advancements in computational lipidomics.

Lipid species have been largely excluded from metabolic modeling so far due to the difficulty of matching lipid species onto (genome-scale) metabolic models [264]. With its ability to generate lipid species-level metabolic networks - including non-lipid metabolites - LINEX^2 can easily be extended to generate stoichiometric matrices. The availability of such matrices enables constraint-based modeling approaches, such as Flux Balance Analysis (FBA), to be extended to cover lipid species in addition to “regular” metabolites contained in genome-scale metabolic models.

6.2 Metabolic Reaction Analysis

In contrast to lipids, metabolic networks for non-lipid metabolites have been available for longer. Therefore, computational methods utilizing these have been developed in the past [265], targeting various use cases and underlying assumptions. One vast area of research is constraint-based and dynamic modeling, which is used for metabolic engineering [266], *in-silico* analysis of microbial communities [267], or simulation of metabolic alterations in tumors [268]. While modeling can give exact results, it relies heavily on model assumptions and complex parametrization [269], which are often inferred in lab conditions and may vary under differing environmental conditions [270]. Therefore, usability is often limited to cases where vast amounts of data for parameter estimation are available under the conditions intended to study [271].

Frainay et al. [178] use genome-scale metabolic models in a different way to recommend metabolic fingerprints based on an initial seed, e.g., metabolites with significantly altered abundances between conditions, using a graph-theoretic approach. Furthermore, approaches utilizing multiple levels of knowledge, for example, the KEGG [49] hierarchy, have been used to predict metabolic entities affected by biological conditions based on metabolite significances [272]. By building upon estimates of statistical significance, these methods are limited to discovering changes in metabolic abundances and making qualitative statements about associations.

The Metabolic Network Reaction Analysis (mantra) method, developed as part of my doctorate, is, to the best of my knowledge, the first method that aims to overcome these limitations by directly using the *quantitative* relationships between substrates and products when dealing with (untargeted) metabolomics data. By enabling a fine-grained functional interpretation of metabolomics experiments, the results presented in Köhler et al. [11] show that mantra can generate clinically relevant hypotheses on changes in metabolic reaction activity. Because it only requires the abundances of identified metabolites as input, it can significantly speed up the process from initial (exploratory) measurements to hypothesis validation. Furthermore, the approximation procedure of mantra is designed in a way that allows the correlation of multi-omics data with the inferred relative reaction activity. Therefore, it provides a way to perform step-wise *functional* integration of metabolomics into a multi-omics context by constructing multi-layer networks with a hybrid knowledge-data approach.

Although mantra requires less detailed biochemical prior knowledge than modeling approaches, its current major limitation is the mapping of measured metabolites onto the metabolic network. For one, identifiers (IDs) are not unified between databases, for example, due to different hierarchy structures, making it hard to map IDs [273]. This is especially problematic, as metabolites might not be available in all databases at the identified level of resolution. Furthermore, metabolites can have many synonyms following different naming conventions [274] and, thus, cannot be easily matched via common names. While some available tools, such as MetaboAnalyst [247], are tackling this issue, the comprehensive mapping of metabolites remains an open problem.

In addition to mapping database IDs, the number of metabolites is affected by the rate of feature annotation in MS-based metabolomics. With state-of-the-art technology, around a thousand metabolites are identified in a Tandem Mass Spectrometry (MS/MS) experiment, while thousands of features remain unannotated. Recently, approaches using deep learning have been developed to improve feature annotation [275] and *in-silico* spectrum prediction[276]. Simultaneously, multi-level networks are used to combine mass spectra-based networks with prior knowledge networks to infer correct annotations [201]. As feature annotation is still the most researched area in computational metabolomics by far, one can expect drastic improvements in metabolite identification in the future. This progress in feature annotation will boost the comprehensive applicability and performance of mantra. Furthermore, mantra itself can be a starting point to enhance metabolite annotation. For example, its reaction activity models could be used to rank multiple possible annotations of one mass spectral feature by the fit of the model for each annotation candidate.

An advantage of metabolic networks over other biological networks, e.g., PPI networks, is their low false positive rate. However, they suffer from a certain level of incompleteness, making manual curation necessary [177, 277, 278]. Furthermore, computational metabolomics methods using metabolic networks, such as mantra, can only be applied to data from organisms for which such networks are available. While this is no issue for biomedical applications, it limits their applicability to other research areas and organisms.

These shortcomings are addressed by methods that aim at speeding up and reducing the manual curation needed during the reconstruction process [279] and methods to fill gaps in metabolic networks [177, 280, 281, 282]. For the latter, both "classic" flux-based, as well as ML-based link prediction methods exist. One shortcoming of these methods is that they mainly focus on adding new reactions to metabolic networks but not metabolites. Therefore, these cannot be used to add metabolites that could not be matched to any database ID, not to mention unidentified features. Including methods with this ability into the pre-processing pipeline of mantra would eliminate its two major limitations.

In other fields of bioinformatics research, data-driven inference of regulatory or signaling networks is a well-studied problem. For example, many recent methods for GRN inference have been proposed [283, 284, 285, 286, 287, 288, 289] following different ideas that could be adapted to the setting of metabolomics. Metabolic network inference, however, is arguably more difficult. One aspect is that regulatory mechanisms are represented as pairwise interactions in GRNs, whereas metabolic reactions are biochemical relations between one or more substrates and one or more products. Therefore, instead of inferring (pairwise) edges, metabolic network inference requires predicting higher-order interactions. Additionally, GRN inference can leverage large amounts of data[4] due to the multiplexing capabilities of sequencing technologies as well as the advanced single-cell technology in transcriptomics. This is especially problematic when considering that the inference of higher-order interactions is an inherently more complex problem than pairwise interactions and, thus, would require an *increased* number of training data points. Although some approaches to include unmatched or unidentified metabolites, such as the strategy proposed by Benedetti et al. [202], into metabolic networks, true data-driven *de-novo* inference remains an unsolved problem that has the potential to not only improve biomedical data interpretation but generalizes to all areas in which metabolomics experiments are used.

The reaction estimation of mantra, as well as its downstream methods, currently require sample group annotations. While this supervised case is a common setting, metabolomics is also used for disease subtyping [290, 291, 292]. In addition to its current implementation, the reaction estimation procedure can easily be adapted to work in an unsupervised setting. The new feature space (sample × reaction change matrix) resulting from this can be directly used as input for (bi-)clustering algorithms, giving rise to a method for metabolic-reaction subtyping. This approach, however, does not have the mechanistic nature of the mantra enrichment algorithm, as it disregards the biochemical connections between metabolic reactions. To allow for mechanistic subtyping, the objective function would be rewritten to represent a clustering metric based on the reactions (and potential multi-omics associations) within a given subgraph. Alternatively, other graph-based approaches for subtyping, such as the one developed by Lazareva et al. [293] for gene expression data, could be adapted to the specific structure of metabolic networks. A mechanistic metabolomics subtyping algorithm would be an important step forward for Systems Medicine, as many diseases have unknown metabolic subtypes [294, 295]. Directly understanding their underlying mechanism enables the effective development of treatments specific to each subtype. Transferring the mantra

[4] In this case, "large amounts of data" refers to both a high number of dimensions *and* a high number of samples.

algorithm to an unsupervised setting, as described above, is a promising approach for future research to fill this gap.

The second main contribution of mantra, in addition to mechanistic metabolomics analysis, is the ability to perform knowledge-based multi-omics integration of metabolomics data. In contrast to previously developed approaches, which are either overrepresentation analysis-like, which suffer from a lack of the reaction-mechanistic component as they rely on p-values or constraint-based modeling, which requires strong assumptions [273], mantra is able to integrate metabolic reactions without requiring metabolite concentrations. This is especially advantageous as it offers wet-lab researchers with a way of analyzing their data that directly provides them with a mechanistic hypothesis on which multi-omics features are associated with the underlying changes in metabolism. Furthermore, it gives a starting point for multi-omics subtyping, which can reveal more fine-grained patterns compared to separate omics-analyses [296]. Given the steadily increasing popularity of multi-omics experiments to study diseases [297], mantra is an important step towards better understanding metabolic alterations under disease conditions and developing treatment approaches targeting them instead of symptomatic treating.

The attentive reader may have noticed that so-far lipidomics, albeit a subset of metabolomics, has been treated separately. Because of the structural and biochemical particularities of lipids, compared to general metabolites, lipidomics data analysis uses specialized methods that are advantageous over universal metabolomics methods. In terms of knowledge-based analyses, the difficulty of integrating lipid measurements with metabolomics data is the discrepancy between lipid metabolism in databases, which separates FA and lipid classes, and the actual lipid species measured. Due to the ability of LINEX^2 to generate metabolic networks on the level of lipid species, it paves the way for integrating lipidomics into metabolic networks. Since untargeted metabolomics experiments also measure lipid species, combining metabolomics and lipidomics in knowledge-based approaches will be a highly demanded analysis approach. The main contributions of this thesis, mantra and LINEX, offer a solid starting point for such integration. In fact, joint networks can already be computed by first generating a mantra network for all non-lipid metabolites and subsequently using the non-lipid reaction participants of the lipid-specific LINEX^2 network to connect both graphs. Methods analyzing such a joint network can also build on the ideas for scoring metabolic dysregulation presented in this thesis. However, they will require careful design to avoid introducing biases toward lipid or non-lipid reactions due to the multispecificity present in lipid metabolism but absent in other parts of metabolism.

6.3 The Future of Metabolomics Data Interpretation and Multi-Omics Integration

Both computational metabolomics and lipidomics data interpretation methods are still in their infancy compared to the current standings of computational methods for other omics disciplines. While this fact is becoming more widely recognized and the community is starting to give more attention to the aspect of data interpretation, several challenges lie ahead for the field to catch up - some of them shared between lipidomics and metabolomics and some specific to each.

Currently, both fields suffer from a lack of standardized naming conventions limiting the automated processing for generating data-specific metabolic networks, such as the ones in LINEX and mantra, or other knowledge-based approaches. For lipids, two tools, LipidLynxX [241], and GOSLIN [298], unifying multiple nomenclature styles have been developed. Even though they do not cover all available ways of annotating lipids, they are a major improvement and enable a wider and more user-friendly user interface for LINEX. I expect them to cover more nomenclature approaches in the future, enabling standardization across different lipidomics platforms. For metabolomics, however, incompatibilities between databases, and thus the naming conventions, are larger and less straightforward to overcome due to different hierarchies that are more difficult to resolve. Even though molecules can, theoretically, be matched via their structure, e.g., in the form of canonical SMILES (Simplified Molecular Input Line Entry System) [299] strings or InChI (International Chemistry Identifier) representations [300], the different levels of isomer resolution in databases can prevent 1-to-1 matching. Therefore, it may not be possible to have a one-approach-fits-all solution, and thus, it will be important for the field to agree on standardized approaches that researchers can follow.

Knowledge-based approaches, which are the focus of my thesis, rely not only on database matching but also on confident identifications. Despite the currently rather low rates of metabolite identification, the improvements in molecular networking [301, 302] and ML-based approaches [275, 276] will boost them, consequently also enhancing the capabilities of methods relying on them.

In addition to metabolite annotation, advances in ML/Deep Learning (DL) will serve as the foundation for novel data-driven solutions to analyze metabolomics and lipidomics data. Arguably, the most widely discussed area, Natural Language Processig (NLP), has already proven useful in scraping knowledge from publication databases, allowing the generation of biological knowledge graphs [303] and possibly metabolic networks in the future. Due to the fast pace with which such methods keep improving, it is not far-fetched to expect their resulting models will be used heavily. Therefore, a crucial task will be to establish mechanisms to ensure their validity on a large scale.

Besides NLP, Geometric Deep Learning (GDL) has seen great progress since the term was coined [230]. Various applications of GDL, such as RNA structure prediction [304], antimicrobial peptide prediction [305], drug discovery [306], and representation learning on molecules [307], have shown its potential for biological applications. As these examples show, a major focus of these approaches lies in structural biology problems. However, many other types of biological problems can profit from progress in GDL due to the non-euclidean nature of biological data and the manifold uses of graphs, such as metabolic networks. Due to their close connection with (first order) Markov models and random walks (Section 2.2.1), which are extensively studied in the context of graph neural networks, biochemical reaction networks are a natural application case for GDL. Furthermore, when interpreting information propagation of message-passing neural networks [308] as the metabolic "information" transmitted through a reaction, one may be able to leverage a phenomenon known as over-squashing. Over-squashing refers to messages not being diffused to far away nodes because the information from many nodes gets "squashed" in bottlenecks [309]. While such bottlenecks are problematic for ML tasks focusing on large-range interactions between nodes [308], they may represent interesting properties in the case of metabolic networks, for example, critical reactions that serve to regulate certain parts of metabolism (i.e. "gatekeepers" of subgraphs). Therefore, using properties such as graph curvature to detect bottleneck edges [174] may be a useful surrogate when analyzing metabolic reaction networks.

Aside from viewing metabolic networks from a graph point of view, advancements in ML also open up new possibilities for analyzing metabolomics data from a dynamical system position (recall Section 2.2.1 for why metabolic networks describe dynamical systems). Particularly interesting developments in this direction have happened in the field of Neural Differential Equations (NDEs). They use some form of neural networks to approximate differential equations subsequently evaluated by a solver and can handle continuous time-series data [310]. In real-world applications, this allows, for example, to forecast how the state of a system, such as a cell, will evolve in the future and can be used for tasks such as predicting drug response effects [311]. Especially interesting for clinical applications, where data cannot always be sampled at the same time points, specific models to deal with irregular time series have been developed [312, 313]. Neural Ordinary Differential Equations, for example, have already been used to model gene expression dynamics [311]. Training neural networks, of course, requires relatively large amounts of data, which may be difficult in a metabolomics setting. However, it is also possible to prime NDEs with inductive bias by encoding certain properties of the system to study (i.e., metabolism) [314]. Therefore, NDEs have the potential to bridge the gap between classic modeling and “purely” data-driven approaches and thus enable a more fine-grained understanding of the dynamics of metabolism. Additionally, such inductive bias can reduce the number of model parameters and thus the need for large amounts of training samples.

Nonetheless, data availability remains a major roadblock for computational metabolomics data interpretation, especially when developing specialized approaches for lipid data. Despite the availability of multiple databases, such as MetaboLights [315], METASPACE [316], and MassBank [317], public availability of published metabolomics data is still much less common than in other omics disciplines [318, 319]. This is not only problematic in the context of science’s reproducibility crisis but also hinders computational research, as exclusively computational labs often have to rely on experimental collaborators to join their projects to validate or even develop their methods. For less established/connected labs, such factors might almost act like a gatekeeping mechanism. From a larger perspective, having large amounts of data publicly available would also fuel methods that aim to transfer patterns from one experimental platform (e.g., organism, analytical platform, etc.) to another to enhance analyses of understudied settings.

An additional point to be considered in future computational metabolomics research is the fact that metabolism is a rapidly changing dynamic system, yet we only measure *static* snapshots. For control-case (e.g., healthy vs. disease) study designs, where the goal is to identify how the case group differs from the control group on a mechanistic level during disease progression, the dynamic nature of the system cannot accurately be captured with a single time point measured for each group. Consequently, to truly understand how metabolism changes during disease progression, time series data is required to learn the overall dynamics. Therefore, while both LINEX and mantra have been designed to work with the currently (by far) most common experimental setups, extending their methodology to consider temporal dynamics will be crucial as such data becomes more available.

Due to the limited availability of metabolomics data, especially in the highly sought-after single-cell context, a currently popular approach is to use (single-cell) RNA sequencing (RNAseq) data to model metabolic states. These methods are typically based on classic modeling approaches, such as FBA or kinetic modeling, and offer great potential to be extended by DL [320]. For example, Wagner et al. [321] proposed a method termed Compass, which uses single-cell gene expression data to penalize metabolic reactions in FBA solvers. While these approaches mitigate the data availability issues in metabolomics, they suffer from the same limitations as classic modeling. Furthermore, even though the expression of enzyme-encoding genes may be proportional to enzymatic activity in some cases, many other factors can reduce their correlation. One such factor is the translation rate influencing the correlation between gene expression and protein abundance, as single-cell transcriptome data has been reported to have higher variability than single-cell proteome data [322]. Even for enzymes, for which gene expression patterns are correlated with protein abundances, post-translational modifications could regulate the activity and, thus, uncouple transcript abundance and enzymatic activity [323].

With advances in multi-omics data acquisition and processing, especially on the side of proteomics, it is likely that these limitations will be overcome to a certain extent. Additionally, recent progress in protein structure prediction [159] is a promising starting point for incorporating structural aspects when analyzing metabolism, for example, by considering the effects of (non-silent) mutations or alternative splicing events. Nevertheless, a crucial aspect often overlooked when trying to model metabolism entirely without metabolite abundances, is the effect metabolite concentrations can have on metabolic flux, e.g., through allosteric regulation [324]. Considering this form of metabolite-guided modulation, accurate modeling of metabolism will require a certain degree of metabolomics data. By finding ways to rank metabolites by their importance for predicting metabolic states, one may be able to get more precise approximations while staying within the limitations of current (targeted) metabolomics capabilities. Nonetheless, integrating high-throughput metabolomics data will eventually result in the most accurate representation once analytical techniques have caught up.

6.4 Conclusion

With the progress in MS-based metabolomics and lipidomics in the last decade, the need for computational methods to aid their interpretation has emerged. Yet tools tackling this need are mostly absent. Having such methods available offers great value for biological, biomedical, and clinical research.

In this dissertation, I aimed to address this gap by introducing graph-driven methods for interpreting lipidome and metabolome data in a way that allows the inclusion of biochemical prior knowledge and integration with other omics data.

For lipidomics data interpretation, this thesis introduces the Lipid Network Explorer (LINEX), an approach to generate lipid-metabolic networks describing the biochemical reactions between lipids measured in an experiment. These networks enable the functional analysis of lipidomics data and the identification of dysregulated lipid-metabolic enzymes. Thereby, LINEX provides a link between changes in lipidome composition and other omics-layers, in particular proteomics, transcriptomics, and genomics.

Currently, a major issue when using prior knowledge in lipidomics data analysis is the poor standardization of lipid naming conventions. With the community actively tackling the standardization and translation between nomenclature styles, I expect to see fewer and fewer such issues in the next years. Another current limitation of lipid metabolic networks is the resolution of lipid structures, especially with respect to fatty acid structures. Due to recent improvements in alternative fragmentation approaches, lipid-metabolic network analysis could become an even more powerful tool, given that the technological promises of approaches, such as EAD, hold.

Regardless of this progress, the work presented in this thesis already demonstrated how lipid network analysis boosts the interpretability of lipidomics experiments and enables the identification of enzymes potentially responsible for the biological mechanisms behind altered lipidome composition. It, therefore, helps researchers in identifying lipid-related disease mechanisms, ultimately improving time and specificity in the development of diagnostic and therapeutic procedures.

The second major framework presented in this thesis is Metabolic Network Reaction Analysis (mantra), an approach to approximate changes in metabolic reaction activity. Its design allows to directly and quantitatively link multi-omics data to the change in activity of a given reaction. As shown in this dissertation, mantra thereby contributes a faster derivation of mechanistic hypotheses that can be used, for example, in translational research.

Even more drastic than for lipidomics is the issue of mapping between metabolite annotations and onto metabolic networks. While there certainly will be progress in this area, I see great potential in hybrid network approaches that combine mass spectral similarity and metabolic networks, thus combining identification and interpretation into a single step. Furthermore, with the open data-sharing mindset spreading further in the community, it could be possible to eventually learn models predicting metabolic network structure, either purely from data or in hybrid approaches combining real-world and synthetic data. Having such models then eliminates the need to identify metabolites when conducting prior knowledge-driven interpretation. Instead, identification is then only required for a selected subset of metabolites.

Despite the aforementioned limitations, the work introduced in this dissertation already shows how mantra enables the identification of metabolic subgraphs representing hypotheses of the biological mechanisms. These can then serve as a starting point in biomarker or drug discovery and thus aid translational research.

Overall, the amount of biological and biomedical data available will grow even faster with improvements in experimental technology. The computational interpretation of biological data will greatly profit from this trend, as it opens the world of omics data analysis to a large number of data-driven approaches. Therefore, biological problems will continue to attract researchers (especially) from the ML/DL community. Combining expertise in this area with expertise in computational biology has been fruitful in the past decade and is likely the most promising path forward for the field in general.

In particular, for disciplines such as metabolomics and lipidomics, for which data is still limited, a crucial factor will be how data-driven approaches can be adapted to scarcer settings. Whether the solution will lie in combining experimental with synthetic data, altered model designs and training schemes, including prior knowledge, or a combination of these remains to be seen.

A Appendix

A.1 Investigating Global Lipidome Alterations with the Lipid Network Explorer

metabolites

MDPI

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 availableas a web-application and as a publicly available docker container.

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

check for updates

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

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 KeyPathwayMiner [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 glycero-, 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.

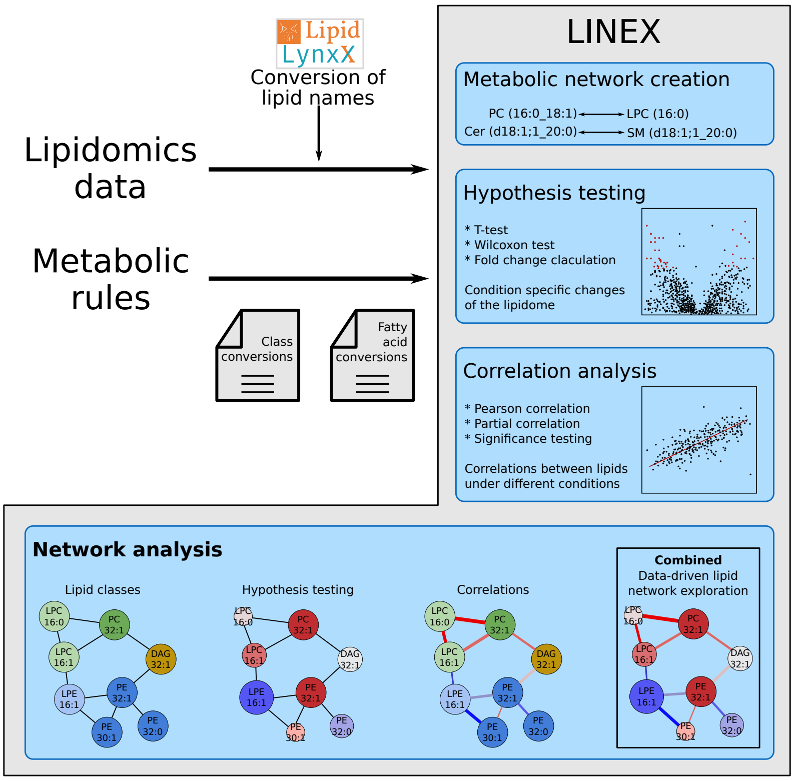


Figure 1. Workflow of the LINEX approach. Lipidomics data and optionally customized metabolic rules are uploaded by the user. The data are used to generate an experiment-specific lipid network, which can be visualized together with statistical measures such as correlation and fold change.

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.

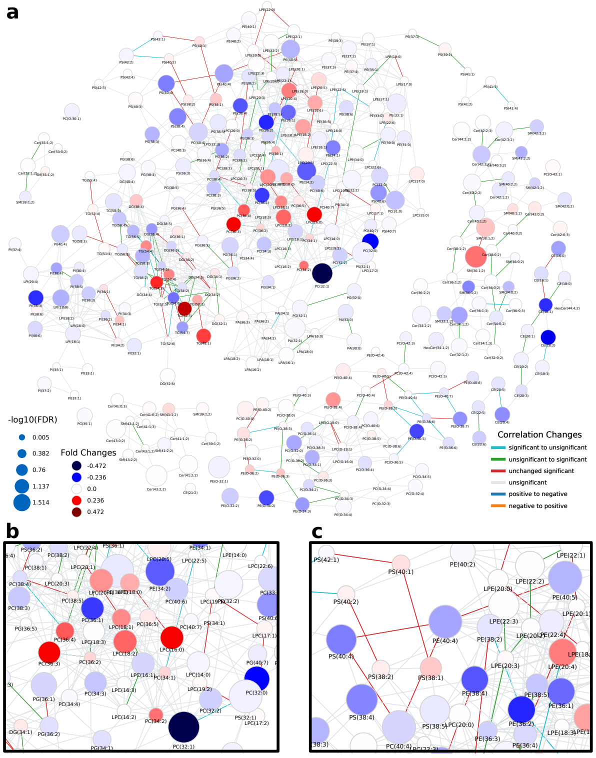


Figure 2. Lipid network of colorectal cancer lipidomics data from Wang et al. [16]. (**a**) Full lipid network with node size scaled by negative log10 of *p*-values for comparison between healthy and cancer tissue. Lipids are colored by log fold change between healthy and cancer tissue. Blue colors indicate 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 intraconnected 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).

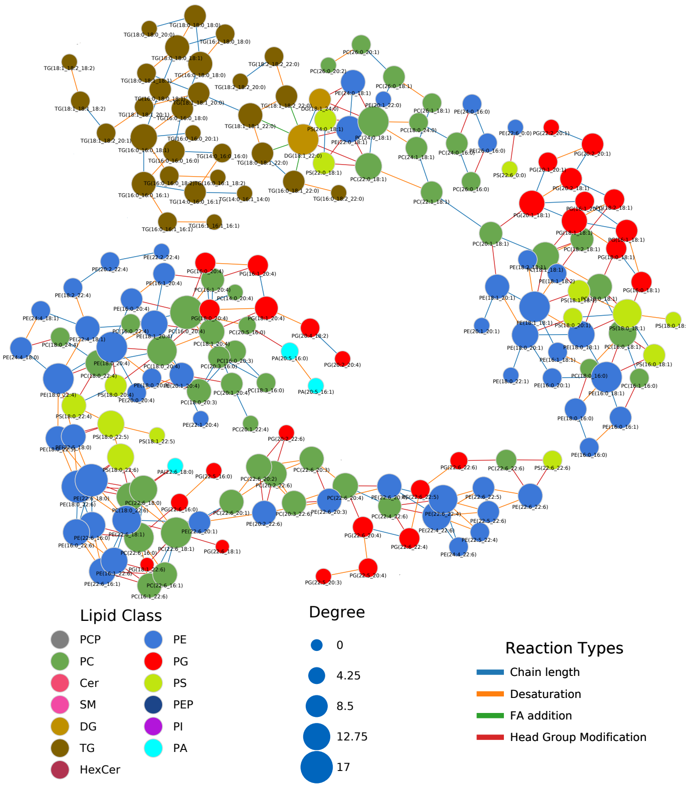


Figure 3. Part of the lipid network of the lipidomics data from Tu et al. [19]. Shown are the two main components of the GPL metabolism. Nodes are colored by lipid class, and edges are colored by reaction type. Node sizes represent the degree.

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).

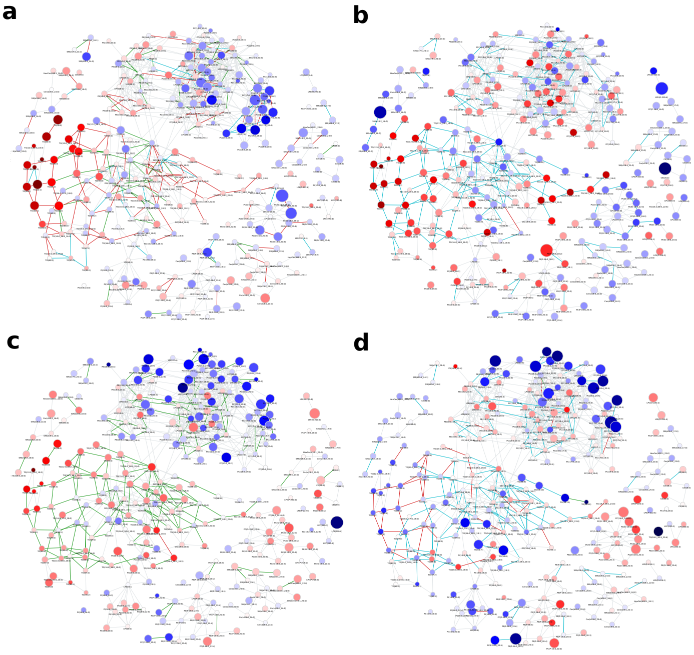


Figure 4. Global age-related plasma lipidome changes in a healthy human reference population from Kyle et al. [22]. Node colors represent log fold-changes with blue being negative, i.e., lower in the first condition, and red being positive. Node sizes are proportional to -log10(FDR) values. Edge colors 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 ("ssignificant 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 LynxX [20] to convert and standardize lipid names in order to recognize them. All lipids recognized by Lipid LynxX 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 ether 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 note, 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: MTBLS562 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

Table A1. *Cont.*

Saturated FAs	Monounsaturated FAs	Polyunsaturated FAs
		22:5 22:6 24:6

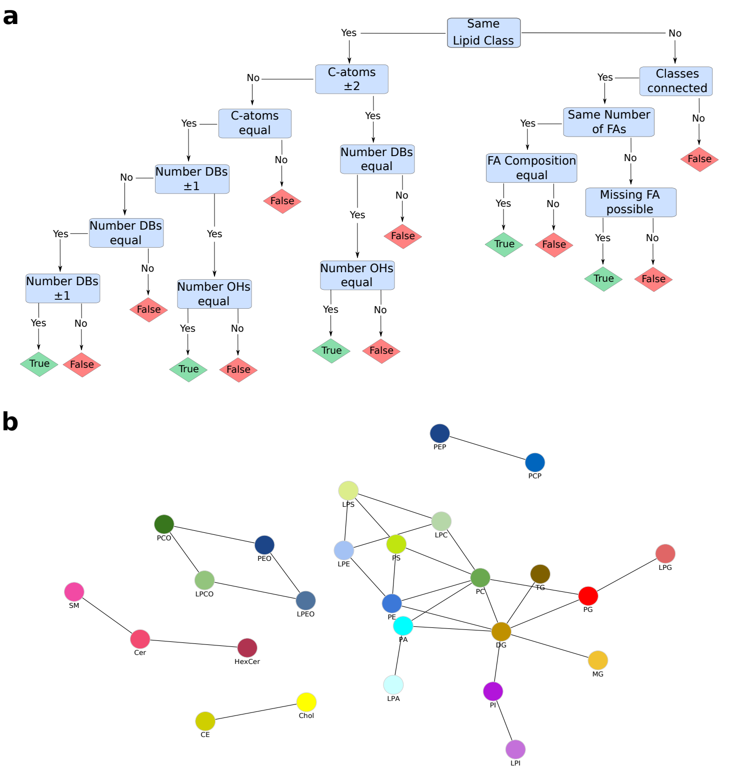
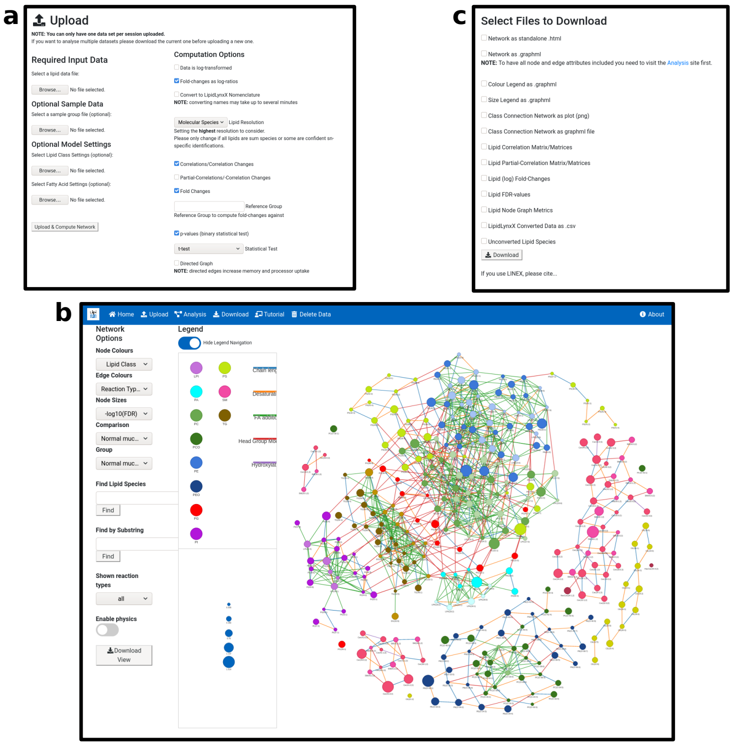


Figure A1. LINEX Default Reaction Rules. (**a**) Decision workflow for lipid connections with default fatty acid reaction rules. Due to the internal logic, lipid classes with different numbers of fatty acids have to have the same head group if they are connected. (**b**) Default lipid class connections. PEP: PE—Plasmalogen; PCP: PC—Plasmalogen.



Upload

NOTE: You can only have one data set per session uploaded.
If you want to analyse multiple datasets please download the current one before uploading a new one.

Required Input Data

Select a lipid data file:

Browse... No file selected.

Optional Sample Data

Select a sample group file (optional):

Browse... No file selected.

Optional Model Settings

Select Lipid Class Settings (optional):

Browse... No file selected.

Select Fatty Acid Settings (optional):

Browse... No file selected.

Upload & Compute Network

Computation Options

☐ Data is log-transformed

☑ Fold-changes as log-ratios

☐ Convert to LipidLynxX Nomenclature
NOTE: converting names may take up to several minutes

Molecular Species Lipid Resolution
Setting the **highest** resolution to consider.
Please only change if all lipids are sum species or some are confident sn-specific identifications.

☑ Correlations/Correlation Changes

☐ Partial-Correlations/-Correlation Changes

☑ Fold Changes

Reference Group
Reference Group to compute fold-changes against

☑ p-values (binary statistical test)

t-test Statistical Test

☐ Directed Graph
NOTE: directed edges increase memory and processor uptake

Select Files to Download

☐ Network as standalone .html

☐ Network as .graphml
NOTE: To have all node and edge attributes included you need to visit the Analysis site first.

☐ Colour Legend as .graphml

☐ Size Legend as .graphml

☐ Class Connection Network as plot (png)

☐ Class Connection Network as graphml file

☐ Lipid Correlation Matrix/Matrices

☐ Lipid Partial-Correlation Matrix/Matrices

☐ Lipid (log) Fold-Changes

☐ Lipid FDR-values

☐ Lipid Node Graph Metrics

☐ LipidLynxX Converted Data as .csv

☐ Unconverted Lipid Species

Download

If you use LINEX, please cite...

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.

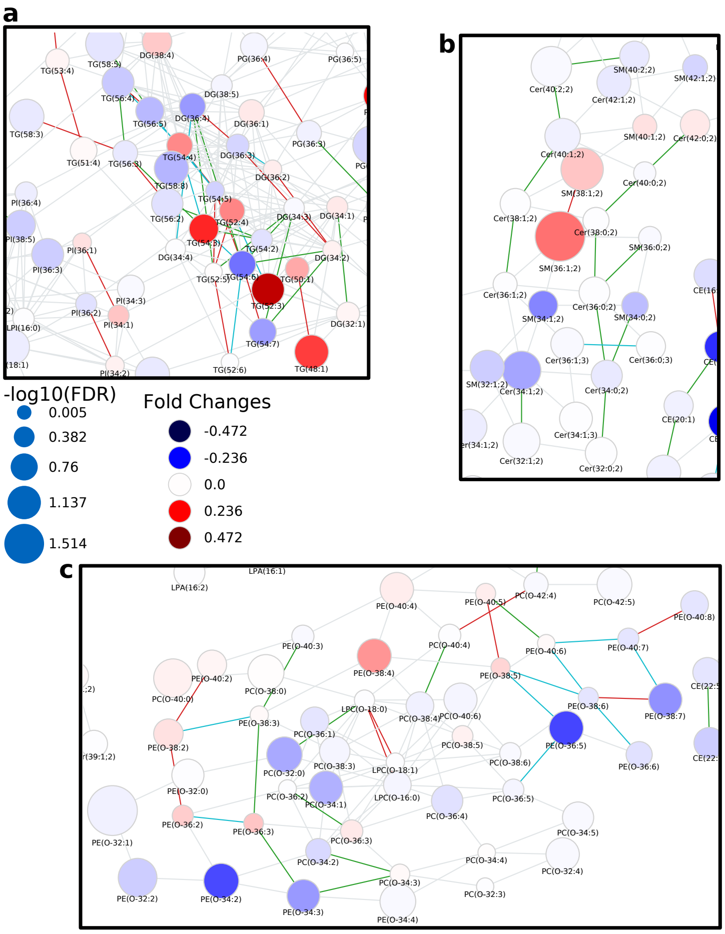


Figure A3. Detailed views on subnetworks of the lipidomics data of Wang et al. [16] showing the metabolism of (**a**) TG and DG, (**b**) ether lipids, and (**c**) sphingolipids. The full network can be seen in Figure 2. Nodes are colored by fold change and node size is scaled by−log10 of multiple testing corrected *p*-value. Edges are colored by correlation changes (see Figure 2).

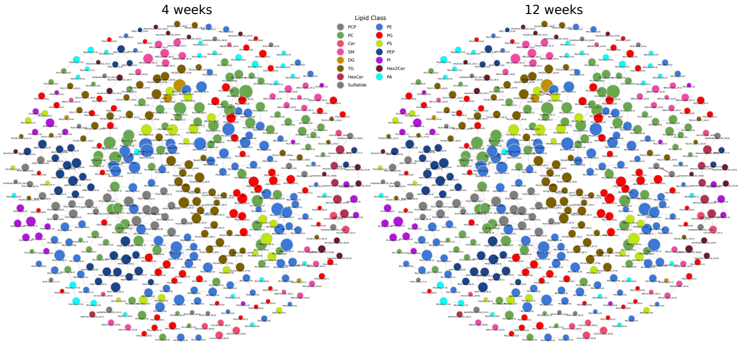


Figure A4. Lipid networks of the lipidomics data from Tu et al. [19]. Nodes are colored by lipid class and edges show correlations between lipids for each mouse age group. Significant and negative correlations are blue, significant and positive correlations red, and insignificant correlations gray. Other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

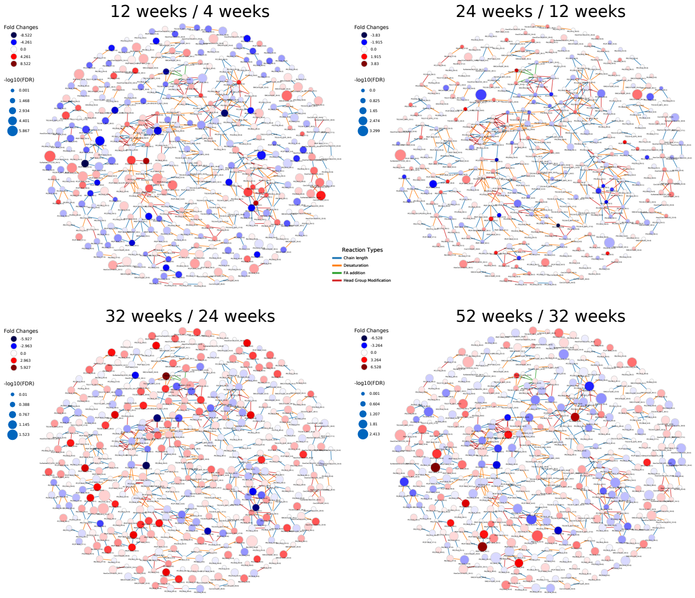


Figure A5. Fold changes of lipids visualized on lipid networks of the lipidomics data from Tu et al. [19]. Node size scaled by negative log10 of the *p*-values for comparison between healthy and cancer tissue. Lipids are colored by log fold change between mouse brain age groups. Blue indicates negative 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).

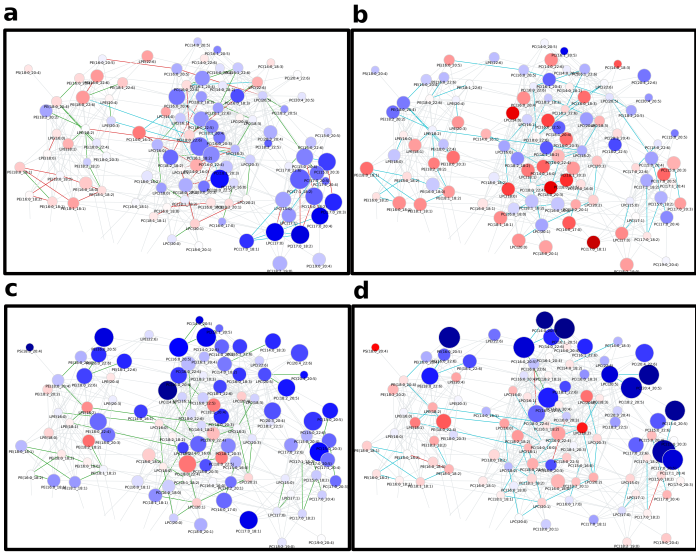


Figure A6. Detailed view on the PC/PE subnetwork from Kyle et al. [22] comparing (**a**) Toddler to Children, (**b**) Children to Teenager, (**c**) Teenager to Young Adults and (**d**) Young Adults to Older Adults.

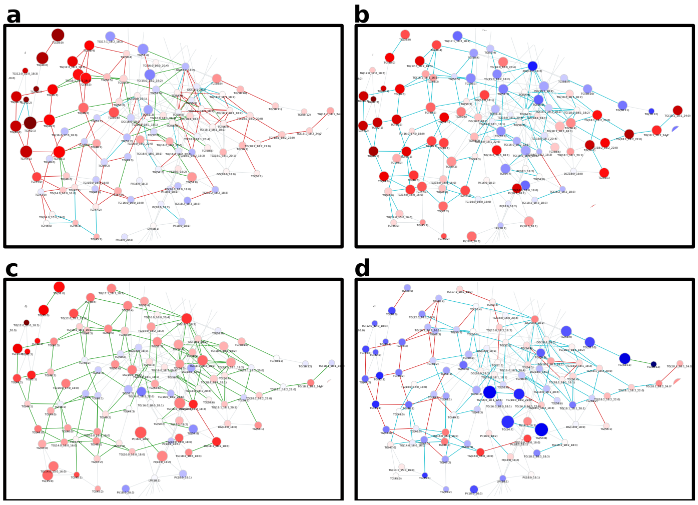


Figure A7. Neutral lipid subnetwork based on Kyle et al. [22] comparing (**a**) Toddler to Children, (**b**) Children to Teenager, (**c**) Teenager to Young Adults and (**d**) Young Adults to Older Adults.

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.

A Appendix

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

1. 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]
2. Mohamed, A.; Hill, M.M. LipidSuite: Interactive web server for lipidomics differential and enrichment analysis. *Nucleic Acids Res.* **2021**, *49*, W346–W351. [CrossRef]
3. 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]
4. 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]
5. 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]
6. 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]
7. 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]
8. 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]
9. 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]
10. 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]
11. 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]
12. Benedetti, E.; Pučić-Baković, M.; Keser, T.; Gerstner, N.; Büyüközkan, 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]
13. 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 metabonomics. *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.2 Lipid network and moiety analysis for revealing enzymatic dysregulation and mechanistic alterations from lipidomics data

OXFORD

Briefings in Bioinformatics, 2023, **24(1)**, 1–11
https://doi.org/10.1093/bib/bbac572
Problem Solving Protocol

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 and Josch K. Pauling

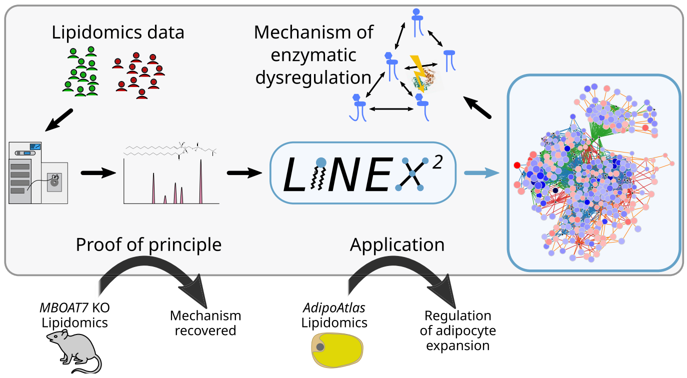
Corresponding author. Josch K. Pauling, LipiTUM, Chair of Experimental Bioinformatics, TUM School of Life Sciences, Technical University of Munich, 85354 Freising, Germany. Tel: +49 8161 71 2762; E-mail: josch.pauling@tum.de

†Tim D. Rose and Nikolai Köhler contributed equally

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 lipid network explorer (LINEX^2), 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 multispecificity 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.

Graphical Abstract

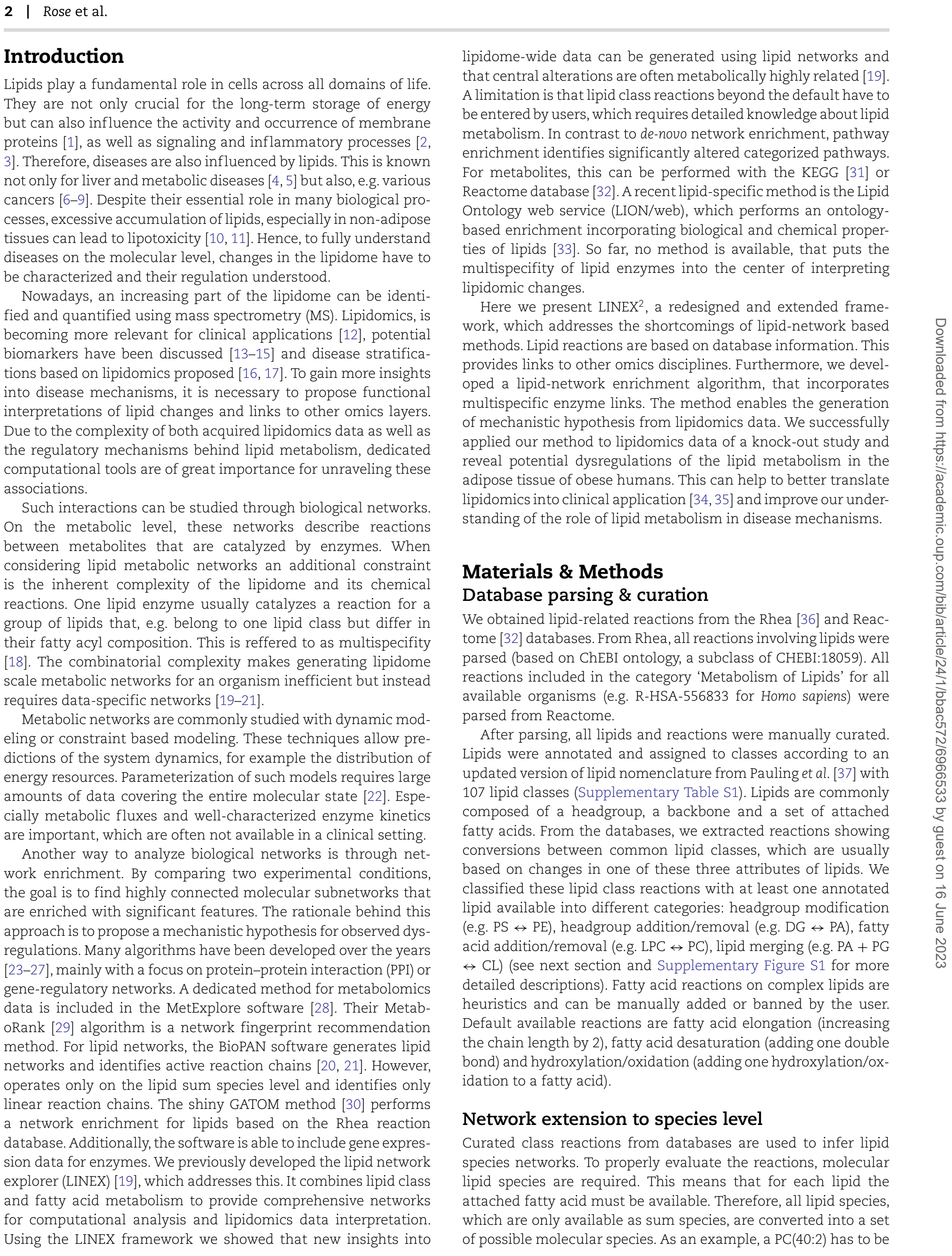


Lipid network explorer (LINEX^2) is a framework to visualize and analyze quantitative lipidomics data. The included algorithms offer new perspectives on the lipidome and can propose potential mechanisms of dysregulation.

Keywords: network enrichment, lipid metabolic networks, lipidomics, disease mechanisms

Tim D. Rose was a PhD student in the research group LipiTUM, TUM School of Life Sciences, Technical University of Munich. He is currently a postdoctoral fellow at the European Molecular Biology Laboratory (EMBL).
Nikolai Köhler is a PhD student in the research group LipiTUM, TUM School of Life Sciences, Technical University of Munich.
Lisa Falk was a master student in the research group LipiTUM in the program Molecular Biotechnology at the Technical University of Munich.
Lucie Klischat was a bachelor student in the research group LipiTUM in the program Bioinformatics at the Technical University of Munich
Olga E. Lazareva was a PhD student in the Chair of Experimental Bioinformatics at the TUM School of Life Sciences, Technical University of Munich. She is currently a postdoctoral fellow at the German Cancer Research Center (DKFZ) and the European Molecular Biology Laboratory (EMBL).
Josch K. Pauling is a group leader of the research group LipiTUM at the TUM School of Life Sciences, Technical University of Munich. He obtained his PhD from the University of Southern Denmark.

Received: August 26, 2022. **Revised:** November 10, 2022. **Accepted:** November 24, 2022



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 signaling 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 [6–9]. 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). Lipidomics, is becoming more relevant for clinical applications [12], potential biomarkers have been discussed [13–15] and disease stratifications based on lipidomics proposed [16, 17]. To gain more insights into disease mechanisms, it is necessary to propose 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 metabolic networks an additional constraint is the inherent complexity of the lipidome and its chemical reactions. One lipid enzyme usually catalyzes a reaction for a group of lipids that, e.g. belong to one lipid class but differ in their fatty acyl composition. This is reffered to as multispecifity [18]. The combinatorial complexity makes generating lipidome scale metabolic networks for an organism inefficient but instead requires data-specific networks [19–21].

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 features. The rationale behind this approach is to propose a mechanistic hypothesis for observed dysregulations. Many algorithms have been developed over the years [23–27], mainly with a focus on protein–protein interaction (PPI) or gene-regulatory networks. A dedicated method for metabolomics data is included in the MetExplore software [28]. Their MetaboRank [29] algorithm is a network fingerprint recommendation method. For lipid networks, the BioPAN software generates lipid networks and identifies active reaction chains [20, 21]. However, operates only on the lipid sum species level and identifies only linear reaction chains. The shiny GATOM method [30] performs a network enrichment for lipids based on the Rhea reaction database. Additionally, the software is able to include gene expression data for enzymes. We previously developed the lipid network explorer (LINEX) [19], which addresses 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 showed that new insights into lipidome-wide data can be generated using lipid networks and that central alterations are often metabolically highly related [19]. A limitation is that lipid class reactions beyond the default have to be entered by users, which requires detailed knowledge about lipid metabolism. In contrast to *de-novo* network enrichment, pathway enrichment identifies significantly altered categorized pathways. For metabolites, this can be performed with the KEGG [31] or Reactome database [32]. 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 [33]. So far, no method is available, that puts the multispecifity of lipid enzymes into the center of interpreting lipidomic changes.

Here we present LINEX^2, 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 [34, 35] and improve our understanding of the role of lipid metabolism in disease mechanisms.

Materials & Methods

Database parsing & curation

We obtained lipid-related reactions from the Rhea [36] and Reactome [32] 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] with 107 lipid classes (Supplementary Table S1). 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 S1 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).

Network extension to species level

Curated class reactions from databases are used to infer lipid species networks. 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 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. A detailed explanation can be found in the Supplementary Methods.

Network enrichment

We developed a novel network enrichment algorithm for lipid networks. The methodology involves 1) building a reaction network from a LINEX^2 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 Substrate-product changes are calculated using the lipidomics data matrix $L = \mathbb{R}^{l \times n}$ consisting of l lipids and n samples. Samples are assigned either to the disease condition $D = \{d_1..d_x\}$ or to the control condition $C = \{c_1..c_z\}$. The score is calculated independently for each reaction r_i for all reactions R. A reaction r_i is a subsets of lipids that participate in the reaction as substrates $S(r_i)$ or products $P(r_i)$. The absolute substrate product difference Γ^a for reaction r_i for of the disease samples D is calculated as:

$$\Gamma^{a,D}_{r_i} = \frac{\sum_{d \in D} \left(\frac{1}{|P(r_i)|} \sum_{p \in P(r_i)} L_{p,d} - \frac{1}{|S(r_i)|} \sum_{s \in S(r_i)} L_{s,d} \right)}{|D|}.$$

Similarly, the relative substrate–product difference Γ^r:

$$\Gamma^{r,D}_{r_i} = \frac{\sum_{d \in D} \left((\prod_{p \in P(r_i)} L_{p,d})^{\frac{1}{|P(r_i)|}} / (\prod_{s \in S(r_i)} L_{s,d})^{\frac{1}{|S(r_i)|}} \right)}{|D|}.$$

Within the calculation, the mean or root is used to correct for bias towards reactions with multiple products or substrates. The final score for each reaction node in the network is then calculated as follows:

$$\text{Score}(r_i) = \frac{\left| \Gamma^D_{r_i} - \Gamma^C_{r_i i} \right|}{\Gamma^C_{r_i}}.$$

As previously explained, reactions of the fatty acid metabolism or ether 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 optimization 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 [38] 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 $CR(V^*)$, 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.

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.

Lipid moiety analysis

The (combined) abundance of lipid features was implemented inspired by the glycan substructure method by Bao *et al.* [39]. 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 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 S2.

Analyzed data sets

The lipidomics data for MOBAT7 WT and knockout mice were taken from Thangapandi *et al.* [40]. The data contain 16 knockout samples and 14 WT samples with 244 lipids from 18 lipid classes. No further processing was done and the data were analyzed as provided by the authors. Data for the Adipo Atlas were used as provided in the supplement of Lange *et al.* [41]. It contains 18 samples, 6 lean and 12 obese, with 674 lipids from 16 lipid classes. The comparison of mesenchymal stem cells (MSC) to adipogenic cells is coming from the supplement of Levental *et al.* [42]. Lipid species measured in less than 50% of all samples were removed before analysis with LINEX^2. The processed data contains 577 lipid species from 21 lipid classes and 4 samples per analyzed sample group (undifferentiated PM untreated and adipogenic PM untreated).

For all data sets analyzed with LINEX^2, HTML files with the LINEX^2 output are available at https://doi.org/10.6084/m9.figshare.20508870.

Results

A framework for lipid network creation and analysis

The workflow of a lipidomics experiment can be divided into five steps: sampling, sample preparation, data acquisition, data processing and data interpretation [43]. LINEX^2 is aiming at the biological interpretation of lipidomics data (Figure 1). The LINEX^2 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 fatty acids commonly observed in lipid classes (e.g. DG(16:0_18:1), DG(16:1_18:0) or DG(14:0_20:1) for DG(34:1)). These can be adapted for each experimental setup. If molecular species are identified but not quantified they can be used instead of inferring fatty acyl sets, as reported in other studies [16]. An example for the network extension is the lipid class reaction between a Phosphatidylcholine (PC) and a Diacylglycerol (DG) (PC → DG), where the phosphocholine headgroup is cleaved off, is applied to the molecular lipid species PC(16:0_18:1) → 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)) → 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 network analysis.

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 [32] databases (Figure 2A). 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 *H. sapiens* (HSA), *R. norvegicus* (RNO) and *M. musculus* (MMU) (Figure 2B). 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 LINEX^2 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 LINEX^2 can be updated in a community effort to enhance support for less studied parts of the lipidome.

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. 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 LINEX^2-networks (as implemented in the network extension) every edge between two lipid species

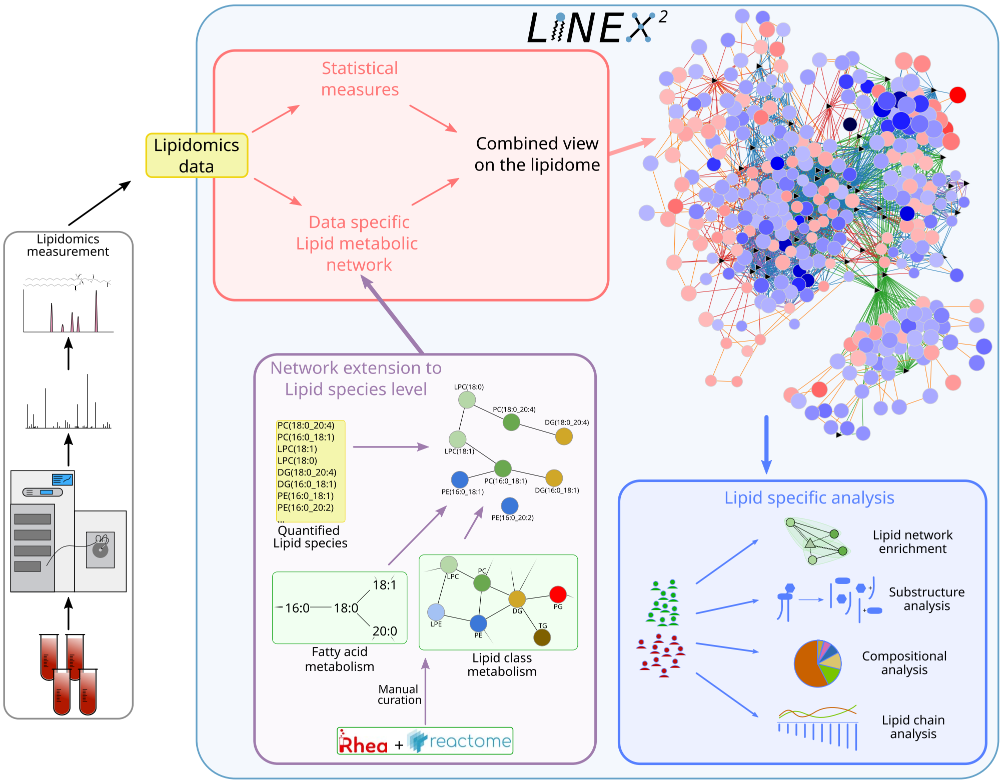


Figure 1. Lipidomics data are used as an input to LINEX^2. The lipids are then utilized to perform network extension that converts lipid class and fatty acid metabolic networks to lipid species, which are then visualized together with statistical measures such as t-tests or correlations. The network is also used as a basis for lipid substructure, compositional and lipid chain analysis. A lipid network enrichment algorithm, which takes enzymatic multispecificity into account, can be used to generate hypotheses for enzymatic dysregulation.

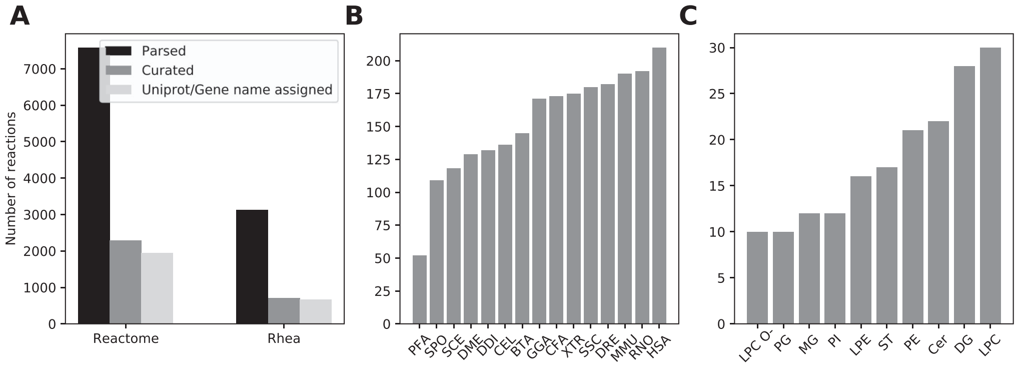


Figure 2. **(A)** Number of lipid-reactions parsed from Reactome and Rhea databases (black), after curation for available lipid classes and number of curated reactions (dark-gray), for which Uniprot or gene name annotations were available (light-gray). **(B)** Curated reactions per organism from the Reactome database (Rhea does not list details about organisms). Legend: HSA - *Homo sapiens*, RNO - *Rattus norvegicus*, MMU - *Mus musculus*, DRE - *Danio rerio*, SSC - *Sus scrofa*, XTR - *Xenopus tropicalis*, CFA - *Canis familiaris*, GGA - *Gallus gallus*, BTA - *Bos taurus*, CEL - *Caenorhabditis elegans*, DDI - *Dictyostelium discoideum*, DME - *Drosophila melanogaster*, SCE - *Saccharomyces cerevisiae*, SPO - *Schizosaccharomyces pombe*, PFA - *Plasmodium falciparum*. **(C)** Top 10 lipid classes with the most curated class reactions.

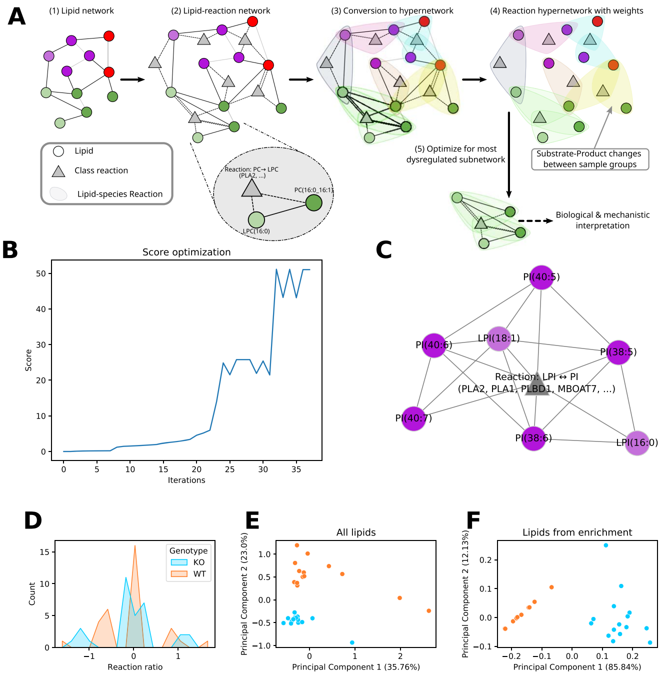


Figure 3. **(A)** Description of network enrichment workflow. In brief, the lipid network is converted into a hypernetwork, in which hyperedges correspond to lipid species reactions. Based on the computed dysregulation per hyperedge, an optimization algorithm finds the subnetwork with the maximum dysregulation. **(B)** Optimal subnetwork predicted by the enrichment algorithm for mice liver lipidomics data by [40]. 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.

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. Based on this representation, the enrichment algorithm can 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 LINEX^2-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 species 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 [44]. 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 to increase connectivity. Additionally, the number of class reactions in the network can be penalized to favor parsimonious solutions with a simple mechanistic explanation.

Inferring known enzymatic dysregulation from a knock-out study

As a proof of principle for the enrichment, we selected data from Thangapandi *et al.* [40]. 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 → PI + CoA with a specific preference for Arachidonic acid (20:4(ω-6), AA) [45]. The data from Thangapandi *et al.* [40] are 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 interactive network available at https://doi.org/10.6084/m9.figshare.20508870). It consists only of PI, LPI species and one

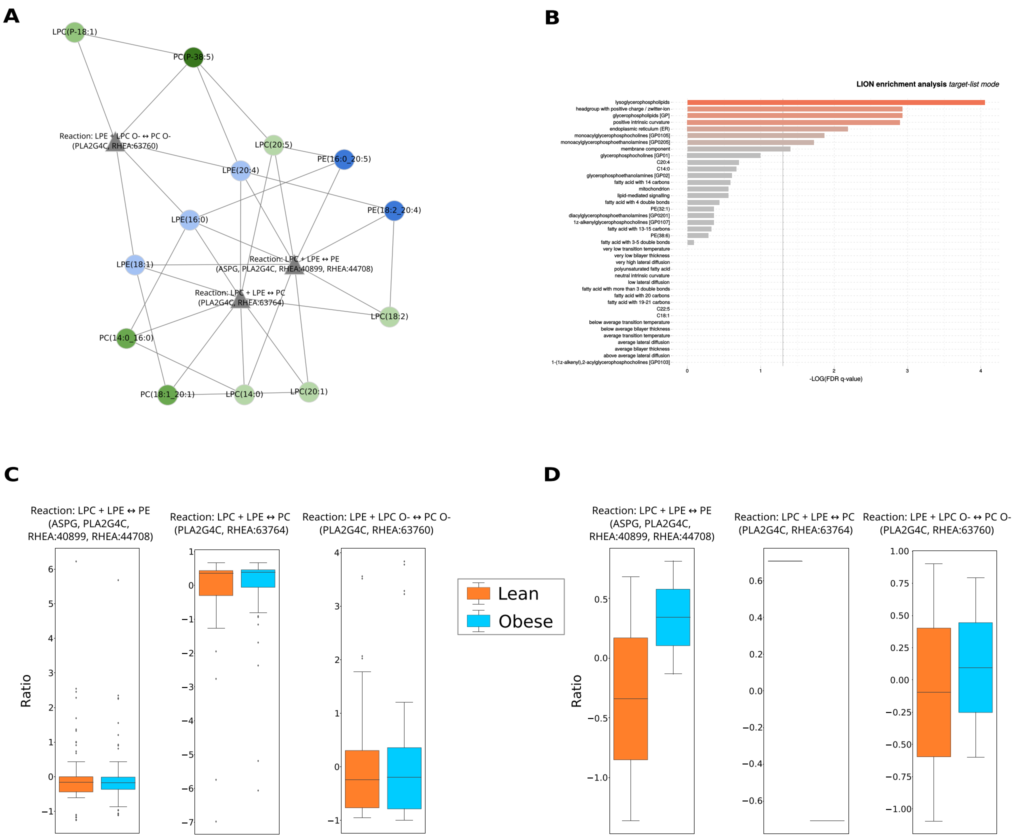


Figure 4. LINEX^2 application on the AdipoAtlas data. **(A)** Subnetwork returned by the introduced enrichment algorithm. The enriched subnetwork contains three reaction nodes, all representing fatty acid transfer between lysophospho- and phospholipids. Furthermore, the network shows a preference for 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. In both C and D ratios are shown as per-reaction z-scores.

class reaction. This class reaction represents the transformation between LPI and PI. LINEX^2 cannot differentiate between the exact enzyme for this reaction. However, in contrast to e.g. PLA2, MBOAT7 only catalyzes LPI → PI class reactions. Additionally, MBOAT7 is known for a higher affinity for AA [45]. 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 S3). A recently published preprint by [46] elucidated the structure and catalytic mechanism of MBOAT7 and found that saturated acyl-CoAs are less likely to bind, supporting our results. To further showcase how our method prefers the right fatty acid preference, we plotted the distribution of all lipid species for MBOAT7 (Supplementary Figure S4A) and the same distributions for only those lipid species reactions in the subnetwork (Supplementary Figure S4B). While the former plot shows very similar distributions between WT and KO, the selected species plot clearly shows that while the WT reactions are staying close to zero, the KO distributions show high absolute ratios in a bimodal fashion.

While LINEX^2 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). 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 ↔ 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.

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.

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 [47]. This preference is reflected in the subnetwork. Literature research shows that PLA2G4C has been reported to be differentially expressed in obese individuals [48, 49] and products of (c) PLA2 activity are known mediators of adipose tissue metabolism [50].

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 Phosphocholine headgroup to lipids with a Phosphoethanolamine headgroup influences the membrane curvature [51, 52]. This property is important because adipocytes expand in obesity [53]. A change in this ratio has also been associated with altered membrane integrity and fluidity [54, 55]. We confirmed this with a Lipid Ontology (LION) enrichment analysis [33], 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 [42] (Supplementary Figure S5B). These insights further support the practical feasibility of our reaction enrichment approach.

Lipid moieties show alterations in neutral lipid composition

Despite changes in the Glycerophospholipid composition that are an indication for adipocyte expansion, synthesis and 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 S6). Concerning chain length, we observe that TG species with a sum length >30 and <57 are accumulated in obese samples (Supplementary Figure S7A). Since TGs and DGs are synthesized de-novo, they were not picked up by the network enrichment as strong alteration between lipid classes, 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 S2). Especially lipid species with a sum length >45 and 2–3 double bonds show a sharp increase in obesity, predominantly TG species with a length of 49 and 53 (Supplementary Figure S7B). Also Sterol esters show significant changes in disease progression. The observed changes in the TG composition are in accordance with previously published results [56]. This analysis can provide additional insights into the lipid metabolism and complement the network analysis.

LINEX^2 software

The LINEX^2 software framework for analysis and visualization of lipid networks is available as a web service at https://exbio.wzw.tum.de/linex2. Lipidomics data can be used to perform not only network enrichment and visualization, but also summarizing statistics, lipid chain analysis [57], 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/). The details of the implementation can be found in the Supplementary Materials.

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.

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 LINEX^2 enrichment differs from BioPAN, by searching for dysregulation of multispecific enzymes

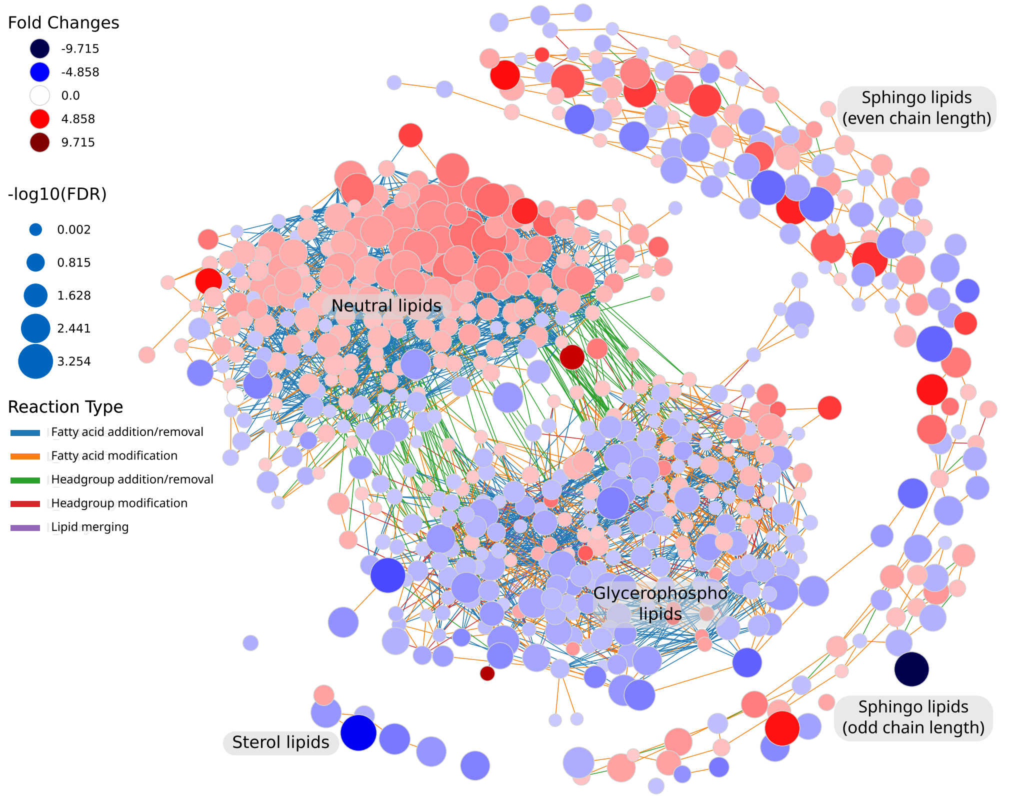
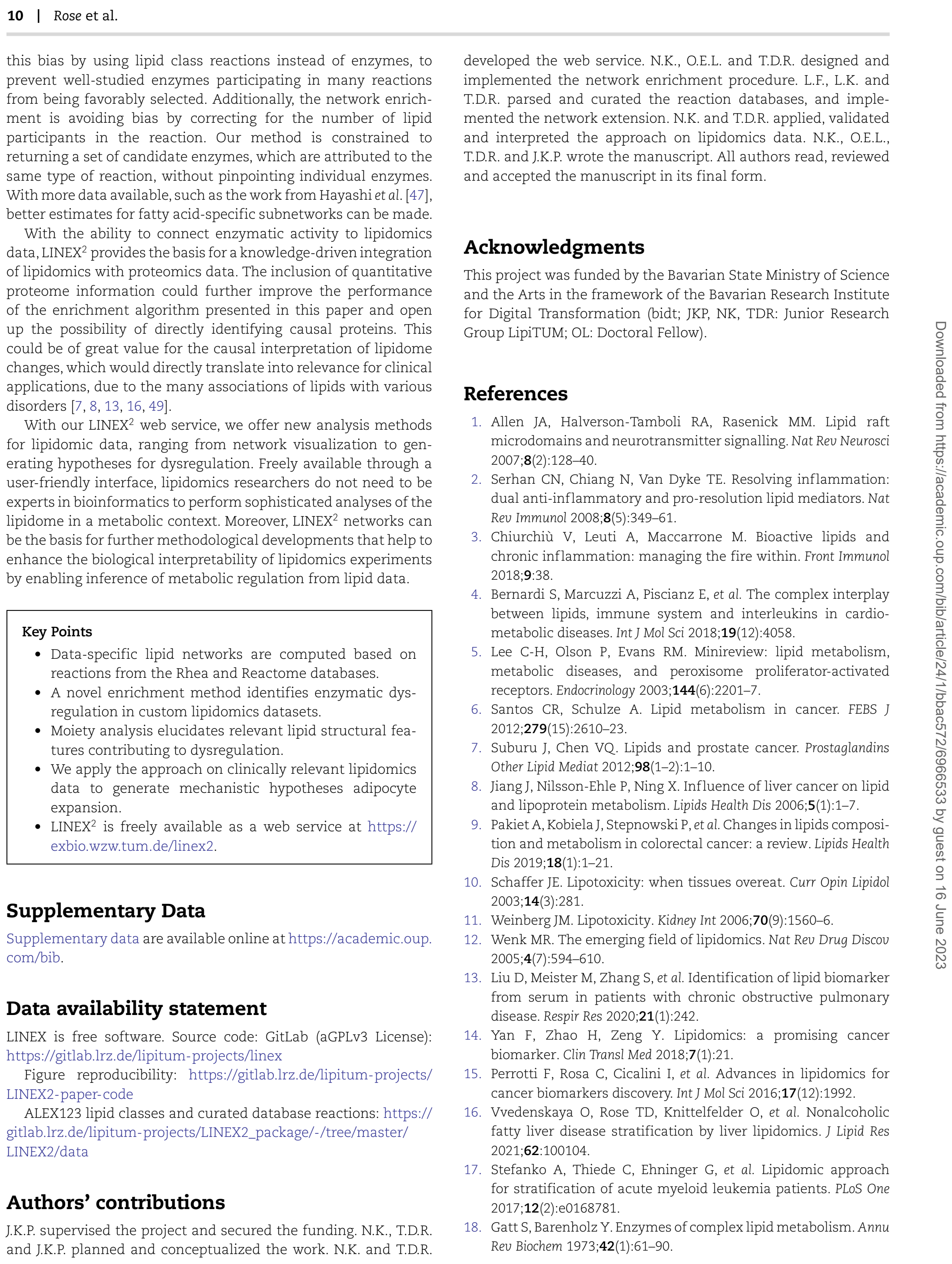


Figure 5. Lipidomics data from the AdipoAtlas visualized with LINEX. In the network lipids are represented as circular nodes. The red color of lipid nodes represents a positive fold change from lean to obese condition, and blue a negative fold change. Edge color indicates the type of reaction connecting two nodes. An interactive version of the network as well as all other analyses conducted with LINEX are available in an HTML file at https://doi.org/10.6084/m9.figshare.20508870.

that likely affect lipids of the same class with different sets of fatty acyls. Another difference is in the network computation. LINEX^2 includes fatty acyl addition/removal, enabling insights such as the MBOAT7 example we show in this work. To illustrate how LINEX^2 compares to BioPAN [20], we computed the BioPAN network (Supplementary Figure S8A) 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 S8B). Therefore a hypothesis on MBOAT7 dysregulation cannot be drawn from this method. Similarly for the application of Shiny GATOM [30] on the same data (Supplementary Figure S9). The subnetwork contains many reactions but misses reactions of PI with more than 40 carbon atoms and does not attribute reactions between PI and LPI to MBOAT7. [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, LINEX^2 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.

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 [58] 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 favorably selected. Additionally, the network enrichment is avoiding bias by correcting for the number of lipid participants in the reaction. 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.* [47], better estimates for fatty acid-specific subnetworks can be made.

With the ability to connect enzymatic activity to lipidomics data, LINEX^2 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 [7, 8, 13, 16, 49].

With our LINEX^2 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, LINEX^2 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.

Key Points

- Data-specific lipid networks are computed based on reactions from the Rhea and Reactome databases.
- A novel enrichment method identifies enzymatic dysregulation in custom lipidomics datasets.
- Moiety analysis elucidates relevant lipid structural features contributing to dysregulation.
- We apply the approach on clinically relevant lipidomics data to generate mechanistic hypotheses adipocyte expansion.
- LINEX^2 is freely available as a web service at https://exbio.wzw.tum.de/linex2.

Supplementary Data

Supplementary data are available online at https://academic.oup.com/bib.

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 reactions: https://gitlab.lrz.de/lipitum-projects/LINEX2_package/-/tree/master/LINEX2/data

Authors' contributions

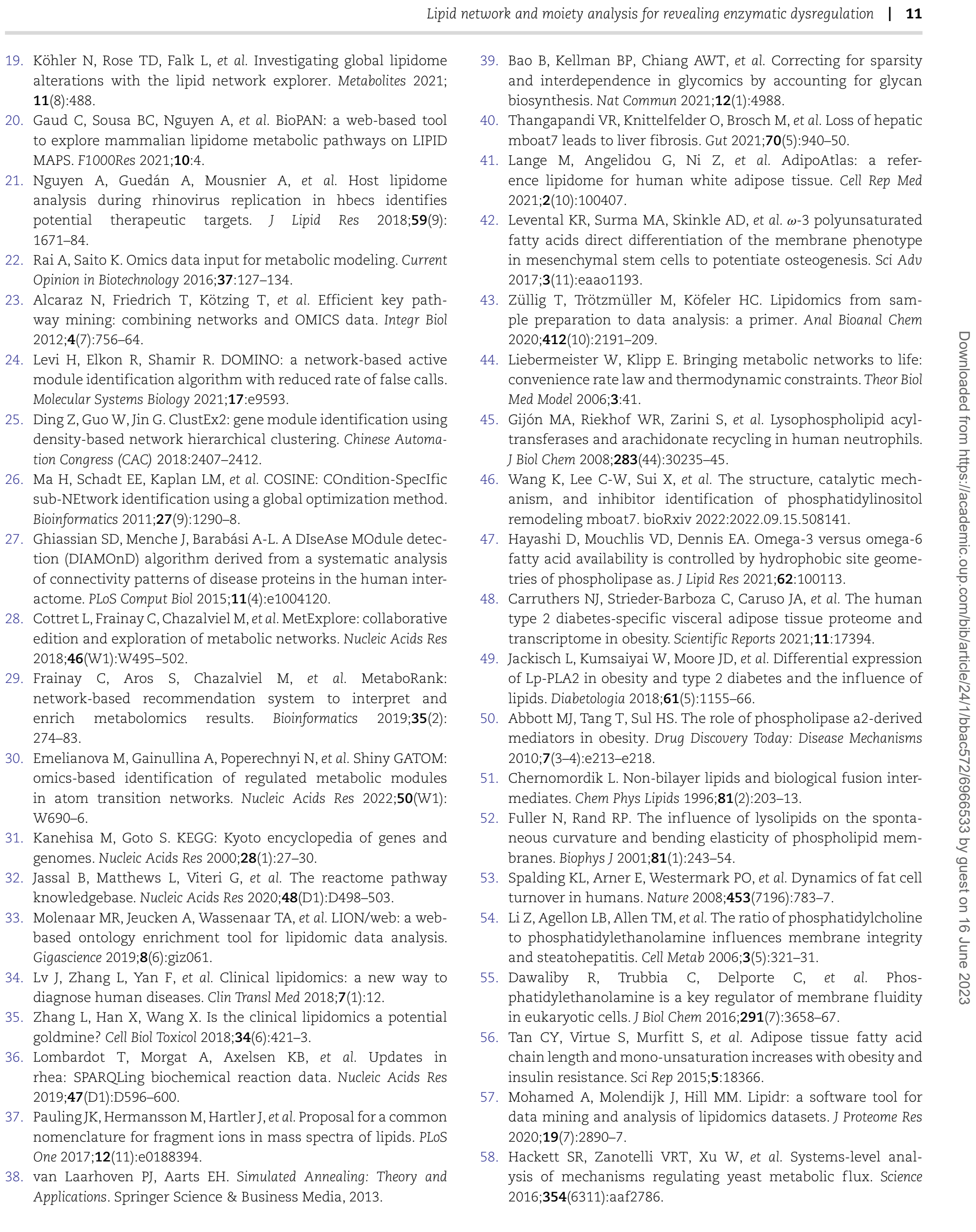
J.K.P. supervised the project and secured the funding. N.K., T.D.R. and J.K.P. planned and conceptualized the work. N.K. and T.D.R. developed the web service. N.K., O.E.L. and T.D.R. designed and implemented the network enrichment procedure. L.F., L.K. and T.D.R. parsed and curated the reaction databases, and implemented the network extension. N.K. and T.D.R. applied, validated and interpreted the approach on lipidomics data. N.K., O.E.L., T.D.R. and J.K.P. 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 (bidt; JKP, NK, TDR: Junior Research Group LipiTUM; OL: Doctoral Fellow).

References

1. Allen JA, Halverson-Tamboli RA, Rasenick MM. Lipid raft microdomains and neurotransmitter signalling. *Nat Rev Neurosci* 2007;**8**(2):128–40.
2. Serhan CN, Chiang N, Van Dyke TE. Resolving inflammation: dual anti-inflammatory and pro-resolution lipid mediators. *Nat Rev Immunol* 2008;**8**(5):349–61.
3. Chiurchiù V, Leuti A, Maccarrone M. Bioactive lipids and chronic inflammation: managing the fire within. *Front Immunol* 2018;**9**:38.
4. Bernardi S, Marcuzzi A, Piscianz E, *et al.* The complex interplay between lipids, immune system and interleukins in cardio-metabolic diseases. *Int J Mol Sci* 2018;**19**(12):4058.
5. Lee C-H, Olson P, Evans RM. Minireview: lipid metabolism, metabolic diseases, and peroxisome proliferator-activated receptors. *Endocrinology* 2003;**144**(6):2201–7.
6. Santos CR, Schulze A. Lipid metabolism in cancer. *FEBS J* 2012;**279**(15):2610–23.
7. Suburu J, Chen VQ. Lipids and prostate cancer. *Prostaglandins Other Lipid Mediat* 2012;**98**(1–2):1–10.
8. Jiang J, Nilsson-Ehle P, Ning X. Influence of liver cancer on lipid and lipoprotein metabolism. *Lipids Health Dis* 2006;**5**(1):1–7.
9. Pakiet A, Kobiela J, Stepnowski P, *et al.* Changes in lipids composition and metabolism in colorectal cancer: a review. *Lipids Health Dis* 2019;**18**(1):1–21.
10. Schaffer JE. Lipotoxicity: when tissues overeat. *Curr Opin Lipidol* 2003;**14**(3):281.
11. Weinberg JM. Lipotoxicity. *Kidney Int* 2006;**70**(9):1560–6.
12. Wenk MR. The emerging field of lipidomics. *Nat Rev Drug Discov* 2005;**4**(7):594–610.
13. Liu D, Meister M, Zhang S, *et al.* Identification of lipid biomarker from serum in patients with chronic obstructive pulmonary disease. *Respir Res* 2020;**21**(1):242.
14. Yan F, Zhao H, Zeng Y. Lipidomics: a promising cancer biomarker. *Clin Transl Med* 2018;**7**(1):21.
15. Perrotti F, Rosa C, Cicalini I, *et al.* Advances in lipidomics for cancer biomarkers discovery. *Int J Mol Sci* 2016;**17**(12):1992.
16. Vvedenskaya O, Rose TD, Knittelfelder O, *et al.* Nonalcoholic fatty liver disease stratification by liver lipidomics. *J Lipid Res* 2021;**62**:100104.
17. Stefanko A, Thiede C, Ehninger G, *et al.* Lipidomic approach for stratification of acute myeloid leukemia patients. *PLoS One* 2017;**12**(2):e0168781.
18. Gatt S, Barenholz Y. Enzymes of complex lipid metabolism. *Annu Rev Biochem* 1973;**42**(1):61–90.



19. Köhler N, Rose TD, Falk L, *et al.* Investigating global lipidome alterations with the lipid network explorer. *Metabolites* 2021; **11**(8):488.
20. Gaud C, Sousa BC, Nguyen A, *et al.* BioPAN: a web-based tool to explore mammalian lipidome metabolic pathways on LIPID MAPS. *F1000Res* 2021;**10**:4.
21. Nguyen A, Guedán A, Mousnier A, *et al.* Host lipidome analysis during rhinovirus replication in hbecs identifies potential therapeutic targets. *J Lipid Res* 2018;**59**(9): 1671–84.
22. Rai A, Saito K. Omics data input for metabolic modeling. *Current Opinion in Biotechnology* 2016;**37**:127–134.
23. Alcaraz N, Friedrich T, Kötzing T, *et al.* Efficient key pathway mining: combining networks and OMICS data. *Integr Biol* 2012;**4**(7):756–64.
24. Levi H, Elkon R, Shamir R. DOMINO: a network-based active module identification algorithm with reduced rate of false calls. *Molecular Systems Biology* 2021;**17**:e9593.
25. Ding Z, Guo W, Jin G. ClustEx2: gene module identification using density-based network hierarchical clustering. *Chinese Automation Congress (CAC)* 2018:2407–2412.
26. Ma H, Schadt EE, Kaplan LM, *et al.* COSINE: COndition-SpecIfic sub-NEtwork identification using a global optimization method. *Bioinformatics* 2011;**27**(9):1290–8.
27. Ghiassian SD, Menche J, Barabási A-L. A DIseAse MOdule detection (DIAMOnD) algorithm derived from a systematic analysis of connectivity patterns of disease proteins in the human interactome. *PLoS Comput Biol* 2015;**11**(4):e1004120.
28. Cottret L, Frainay C, Chazalviel M, *et al.* MetExplore: collaborative edition and exploration of metabolic networks. *Nucleic Acids Res* 2018;**46**(W1):W495–502.
29. Frainay C, Aros S, Chazalviel M, *et al.* MetaboRank: network-based recommendation system to interpret and enrich metabolomics results. *Bioinformatics* 2019;**35**(2): 274–83.
30. Emelianova M, Gainullina A, Poperechnyi N, *et al.* Shiny GATOM: omics-based identification of regulated metabolic modules in atom transition networks. *Nucleic Acids Res* 2022;**50**(W1): W690–6.
31. Kanehisa M, Goto S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 2000;**28**(1):27–30.
32. Jassal B, Matthews L, Viteri G, *et al.* The reactome pathway knowledgebase. *Nucleic Acids Res* 2020;**48**(D1):D498–503.
33. Molenaar MR, Jeucken A, Wassenaar TA, *et al.* LION/web: a web-based ontology enrichment tool for lipidomic data analysis. *Gigascience* 2019;**8**(6):giz061.
34. Lv J, Zhang L, Yan F, *et al.* Clinical lipidomics: a new way to diagnose human diseases. *Clin Transl Med* 2018;**7**(1):12.
35. Zhang L, Han X, Wang X. Is the clinical lipidomics a potential goldmine? *Cell Biol Toxicol* 2018;**34**(6):421–3.
36. Lombardot T, Morgat A, Axelsen KB, *et al.* Updates in rhea: SPARQLing biochemical reaction data. *Nucleic Acids Res* 2019;**47**(D1):D596–600.
37. Pauling JK, Hermansson M, Hartler J, *et al.* Proposal for a common nomenclature for fragment ions in mass spectra of lipids. *PLoS One* 2017;**12**(11):e0188394.
38. van Laarhoven PJ, Aarts EH. *Simulated Annealing: Theory and Applications*. Springer Science & Business Media, 2013.
39. Bao B, Kellman BP, Chiang AWT, *et al.* Correcting for sparsity and interdependence in glycomics by accounting for glycan biosynthesis. *Nat Commun* 2021;**12**(1):4988.
40. Thangapandi VR, Knittelfelder O, Brosch M, *et al.* Loss of hepatic mboat7 leads to liver fibrosis. *Gut* 2021;**70**(5):940–50.
41. Lange M, Angelidou G, Ni Z, *et al.* AdipoAtlas: a reference lipidome for human white adipose tissue. *Cell Rep Med* 2021;**2**(10):100407.
42. Levental KR, Surma MA, Skinkle AD, *et al.* ω-3 polyunsaturated fatty acids direct differentiation of the membrane phenotype in mesenchymal stem cells to potentiate osteogenesis. *Sci Adv* 2017;**3**(11):eaao1193.
43. Züllig T, Trötzmüller M, Köfeler HC. Lipidomics from sample preparation to data analysis: a primer. *Anal Bioanal Chem* 2020;**412**(10):2191–209.
44. Liebermeister W, Klipp E. Bringing metabolic networks to life: convenience rate law and thermodynamic constraints. *Theor Biol Med Model* 2006;**3**:41.
45. Gijón MA, Riekhof WR, Zarini S, *et al.* Lysophospholipid acyltransferases and arachidonate recycling in human neutrophils. *J Biol Chem* 2008;**283**(44):30235–45.
46. Wang K, Lee C-W, Sui X, *et al.* The structure, catalytic mechanism, and inhibitor identification of phosphatidylinositol remodeling mboat7. bioRxiv 2022:2022.09.15.508141.
47. Hayashi D, Mouchlis VD, Dennis EA. Omega-3 versus omega-6 fatty acid availability is controlled by hydrophobic site geometries of phospholipase as. *J Lipid Res* 2021;**62**:100113.
48. Carruthers NJ, Strieder-Barboza C, Caruso JA, *et al.* The human type 2 diabetes-specific visceral adipose tissue proteome and transcriptome in obesity. *Scientific Reports* 2021;**11**:17394.
49. Jackisch L, Kumsaiyai W, Moore JD, *et al.* Differential expression of Lp-PLA2 in obesity and type 2 diabetes and the influence of lipids. *Diabetologia* 2018;**61**(5):1155–66.
50. Abbott MJ, Tang T, Sul HS. The role of phospholipase a2-derived mediators in obesity. *Drug Discovery Today: Disease Mechanisms* 2010;**7**(3–4):e213–e218.
51. Chernomordik L. Non-bilayer lipids and biological fusion intermediates. *Chem Phys Lipids* 1996;**81**(2):203–13.
52. Fuller N, Rand RP. The influence of lysolipids on the spontaneous curvature and bending elasticity of phospholipid membranes. *Biophys J* 2001;**81**(1):243–54.
53. Spalding KL, Arner E, Westermark PO, *et al.* Dynamics of fat cell turnover in humans. *Nature* 2008;**453**(7196):783–7.
54. Li Z, Agellon LB, Allen TM, *et al.* The ratio of phosphatidylcholine to phosphatidylethanolamine influences membrane integrity and steatohepatitis. *Cell Metab* 2006;**3**(5):321–31.
55. Dawaliby R, Trubbia C, Delporte C, *et al.* Phosphatidylethanolamine is a key regulator of membrane fluidity in eukaryotic cells. *J Biol Chem* 2016;**291**(7):3658–67.
56. Tan CY, Virtue S, Murfitt S, *et al.* Adipose tissue fatty acid chain length and mono-unsaturation increases with obesity and insulin resistance. *Sci Rep* 2015;**5**:18366.
57. Mohamed A, Molendijk J, Hill MM. Lipidr: a software tool for data mining and analysis of lipidomics datasets. *J Proteome Res* 2020;**19**(7):2890–7.
58. Hackett SR, Zanotelli VRT, Xu W, *et al.* Systems-level analysis of mechanisms regulating yeast metabolic flux. *Science* 2016;**354**(6311):aaf2786.

A.3 Identification and Integration of Key-Metabolic Reactions from Untargeted Metabolomics Data

Identifying and Multi-Omics Integration of key-Metabolic Reactions from Untargeted Metabolomics

Nikolai Köhler[1,*] **Vivian Würf**[1] **Tim D. Rose**[1,2] **Josch K. Pauling**[1,*]

[1] LipiTUM, Chair of Experimental Bioinformatics, TUM School of Life Sciences, Technical University of Munich, 85354 Freising, Germany
[2] Structural and Computational Biology Unit, European Molecular Biology Laboratory, Heidelberg, Germany

[*] Correspondence to `nikolai.koehler@tum.de` and `josch.pauling@tum.de`

Abstract

Metabolomics has become increasingly popular in biological and biomedical research, especially for multi-omics studies, due to the many associations of metabolism with diseases. This development is driven by improvements in metabolite identification and generating large amounts of data, increasing the need for computational solutions for data interpretation. In particular, only few computational approaches directly generating mechanistic hypotheses exist, making the biochemical interpretation of metabolomics data difficult. We present *mantra*, an approach to estimate how metabolic reactions change their activity between biological conditions without requiring absolute quantification of metabolites. Starting with a data-specific metabolic network we utilize linear models between substrates and products of a metabolic reaction to approximate deviations in activity. The obtained estimates can subsequently be used for network enrichment and integration with other omics data. By applying *mantra* to untargeted metabolomics measurements of Triple-Negative Breast Cancer biopsies, we show that it can accurately pinpoint biomarkers. On a dataset of stool metabolomics from Inflammatory Bowel Disease patients, we demonstrate that predictions on our proposed reaction metric generalize to an independent validation cohort and that it can be used for multi-omics network integration. By allowing mechanistic interpretation we facilitate knowledge extraction from metabolomics experiments.

***K*eywords** Metabolic networks · Multi-omics Integration · Network Enrichment · Disease mechanisms

1 Introduction

Metabolites display the product of metabolism and thereby the metabolic state of an organism. Their chemical structures are as diverse as their functions, ranging from pure energy metabolism to immune modulation and environmental sensing [1]. Owing to the essential nature of many of these functions, metabolism is tightly regulated through the control of enzymatic activity. This regulation can happen on different levels, such as the amount of enzyme, post-translational modifications, or allosteric regulation by other metabolites [2, 3, 4, 5, 6]. Since metabolite concentrations are also highly influenced by environmental factors such as diet, medication, or a host's microbiome [7], the metabolic phenotype is the highly complex result of internal metabolic processes and environmental factors. The resulting metabolic phenotype is often referred to as the "metabotype" [8].

Metabolomics, the large-scale study of metabolites, is used to characterize the changes in metabolite levels. Due to the importance of metabolic processes for almost any aspect of life, metabolomics is becoming increasingly popular for biological and biomedical research. Even though the chemical analysis of metabolites, most commonly via Mass Spectrometry (MS), has made great progress in the last decade confident large-scale identification remains difficult, and especially reliable absolute quantification is only possible for a small set of target metabolites. These shortcomings make it particularly challenging to apply metabolomics for exploratory purposes with no clear hypothesis when trying to understand the molecular mechanisms behind different biological conditions.

In addition to such analytical challenges, computational metabolomics, outside molecule identification, is still a small field. Consequently, a rather small number of methods for the computational interpretation of metabolomics data are available. The most commonly used analyses are "classical" univariate statistical tests and fold-changes as well as multivariate approaches such as Principal Component Analysis (PCA) and Partial Least Squares(-Discriminant Analysis) (PLS(-DA)) [9]. While these methods allow it to extract significantly altered metabolites and get an overview of how different the metabolome in different conditions is, they do not allow for direct biochemical interpretation of the results. Instead, specific over-representation or pathway enrichment methods, e.g. MSEA [10] or IMPaLA [11], are used to obtain high-level summaries. Despite delivering biochemically more coarse-grained and comprehensible results, they don't allow for the generation of mechanistic hypotheses on the level of *de-novo* pathways or quantitatively for individual reactions.

To computationally propose such mechanistic interpretation, metabolic networks can be utilized. They can be represented as directed bipartite graphs in which metabolites and reactions are nodes connected by (directed) edges - substrate or product relations - which are catalyzed by specific enzymes [12]. Such networks are available for many organisms nowadays, e.g. from KEGG [13] or BioCyc [14], but only cover parts of the entire metabolome, therefore limiting the scope to known metabolic reactions.

One way to leverage such networks is metabolic modeling, more precisely kinetic or constraint-based modeling [15]. The advantage of such methods is that they are able to make precise predictions on how metabolism behaves, given that the underlying model assumptions are valid. However, this dependence is also a major weak point since, particularly for eukaryotes, the correct parameterization of these models is hard, yet critical for the predicted outcome to the extent of yielding possibly contrasting results [16].

Another strategy to incorporate prior-knowledge networks into data analysis is to use graph theoretic approaches. Especially network enrichment, which aims at identifying subgraphs characterized by high changes between conditions, has been extensively used and studied in genomics, transcriptomics, and proteomics [17]. For metabolomics, only a few such methods are available [12]. One of them is the MetaboRank algorithm [18]. It turns the metabolic network into a Markov model by defining transition probabilities between substrate-product pairs on the basis of the proportion of mapped atoms. The "metabolic fingerprint", as the authors call it, is then computed by performing random walks via a variation of Personalized Page Rank [19]. Another method also using random walks/diffusion to assign relevance to entities in a metabolic graph was introduced by Picart-Armada et al. [20]. In contrast to using metabolic networks, this approach uses a KEGG graph including different hierarchies from compounds down to the level of modules and pathways. This makes it one of the few methods that gives results on metabolic reactions directly. To compute the relevance of each node in the graph, heat diffusion is simulated with only significantly altered metabolites being able to introduce heat into the system. Each node's relevance is then determined by the heat flow

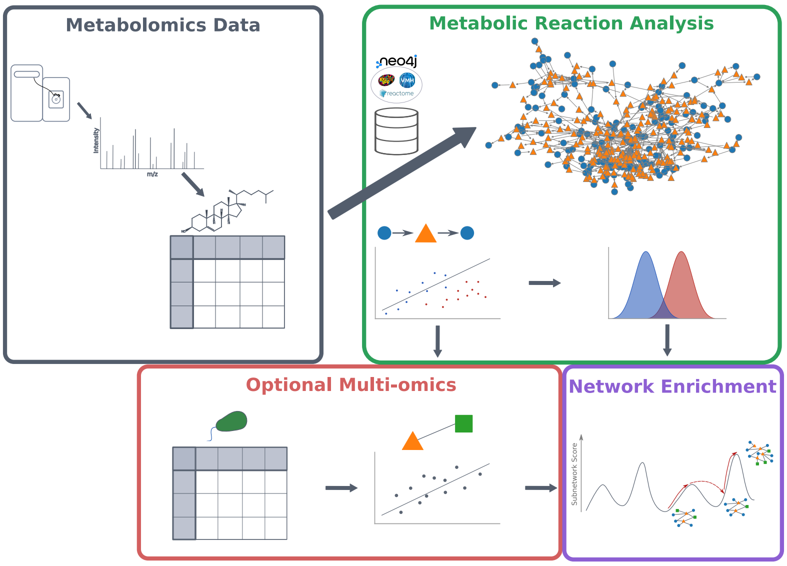


Figure 1: Overview of the *mantra* workflow. Starting with identified (untargeted) metabolomics data (*gray box*), a metabolic network containing only measured metabolites is constructed. It is a directed bipartite graph with metabolites depicted as blue circles and metabolic reactions as orange triangles. For each metabolic reaction in this network, the reaction activity relative to the activity in the control group is estimated using a linear model between the substrate and product abundances (*green box*). These activities can either be directly used for network enrichment (*purple box*) or together with multi-omics data (*red box*). In the latter case, the estimated per-sample activity values are correlated with the expression of each feature in the multi-omics data.

going through or the proportion of random walks including it and then compared to the outcome of a null model in which significances are randomly permuted to avoid structure-based biases.

One drawback of the available methods is, that they either only consider the network-topological properties of metabolites or incorporate only significances, which solely rely on univariate tests and a p-value threshold. In this work, we tackle these shortcomings by presenting an approach we named *mantra* (**M**et**a**bolic **N**e**t**work **R**eaction **A**nalysis), to estimate how the activity of individual metabolic reactions changes between biological conditions on the basis of relative metabolite intensities. In contrast to existing graph-based metabolomics methods, our approach does not rely on computing univariate statistics for all metabolites but uses metabolite abundances to obtain samplewise estimates. It thereby avoids the need for a significance threshold prior to enrichment and allows for the integration of additional omics layers via their biochemical connections.

Using two independent clinical studies, we demonstrate how our approach conserves a significant proportion of the original variation while enabling the generation of hypotheses on the metabolic mode of action. Furthermore, we show the capability of our method to improve the integration of metabolomics into a multi-omics context and how it can be used to define *de-novo* disease pathways. These characteristics make it a promising concept for advancing the mechanistic interpretation of metabolomics data in a clinical context. The idea presented can be used as a basis to develop a new class of approaches for metabolic network analysis.

2 Results

We first start by briefly giving an overview of our approach before presenting its application on independent clinical data sets. A more detailed explanation of each step of the workflow is given in Materials & Methods.

The input to the proposed method is simply a table of normalized metabolite intensities and, optionally, a metabolic network. If no metabolic network is given, the metabolites are first mapped onto internal identifiers to generate a network from a custom database built from KEGG [13], Reactome [21] and Virtual Metabolic Human (VMH) [22]. Its structure is schematically shown in the top right of Figure 1. To estimate the activity of each reaction in the network, a linear model with the substrates as the predicting and the products as the dependent variables is computed. Next, the explained variance for each sample is computed from the distance of the predicted product values to the actual product values. The intuition behind this step is that the higher the activity of a metabolic reaction, the higher the dependence structure between substrate and product concentrations (and thus the measured intensities) will be. If a reaction now changes its activity, the association persists, but the coefficients describing it will likely change. Therefore, this alteration can be observed in a change of residuals. These values can now directly be used to perform network enrichment where the objective is to find a connected subgraph in which the difference between residual distribution is maximized for two given sample groups. Since our proposed metric is per-sample, it can also be used to correlate them to features from other omics layers, for example, microbial abundances. The associations can also be restricted to known interactions if provided with the metabolic network or provided at network generation when using our custom database. To avoid assumptions on distributions these interactions are computed using Spearman's rank correlation.

2.1 *mantra* Recovers Known Key Reactions in Triple-Negative Breast Cancer

To demonstrate the capabilities of our approach to metabolomics, we chose a data set by Xiao et al. [23]. The processed data contains 330 Triple-Negative Breast Cancer (TNBC) and 149 control samples, and 594 identified metabolites with their corresponding intensities. Mapping the metabolites onto our database yielded a network with 173 metabolites and 254 reactions (Figure 2a). The reduced number of metabolites is due to metabolites not being matched to database identifiers present in our database or not being connected to any reaction for which at least one substrate and one product were measured.

Most nodes in the network have between 2 and 5 connections (Supplementary Figure S1). Metabolic reaction nodes do not range higher than this, which is expected, considering that reactions generally don't have a large number of substrates and products. For metabolite nodes, however, there are some hub nodes, which take part in up to 50 metabolic reactions. Metabolites with a node degree above 10 are listed in Supplementary Table S1.

To evaluate how well our proposed metric for relative reaction activity conserves the variance contained in the original metabolomics data, their PCA plots are shown in Figure 2b (metabolome data) and c (reaction data). Sub-figure b shows a separation of the control and the TNBC samples along PC1, which explains around 47% of the total variance. In comparison, PC1 of the reaction data in Figure 2c explains a little less variance. Although the variance between control samples is low, there are few outliers on the far right. The TNBC group, on the other hand, shows a much higher variation than in the general metabolome data. Nevertheless, Figure 2c shows that our approach is able to retain biological variance and separate the clinical conditions.

Going into a more detailed analysis on which metabolic reactions are identified as the most changing between conditions, 18 reaction nodes appear as highly significant (Figure 2d). Some of these reaction nodes represent more than one metabolic reaction, as some reactions have the same measured substrates and products, making the possibly involved catalytic enzymes indistinguishable. The details of the reactions from Figure 2d are given in Supplementary Table S2.

Most of these reactions involve nucleotides, most prominently Uridine Mono-Phosphate (UMP) and Uridine Di-Phosphate (UDP), and different glycosylation reactions. Both of these reaction classes, but especially UDP-related reactions have been identified as key players in breast cancer [24, 25, 26, 27, 28, 29]. Notably, enzymes catalyzing 4 of the significant reactions - UDP-glucuronosyltransferase (UGT), UMP-CMP kinase (CMPK1) and UDP-glucose-dehydrogenase (UGDH) - are known to

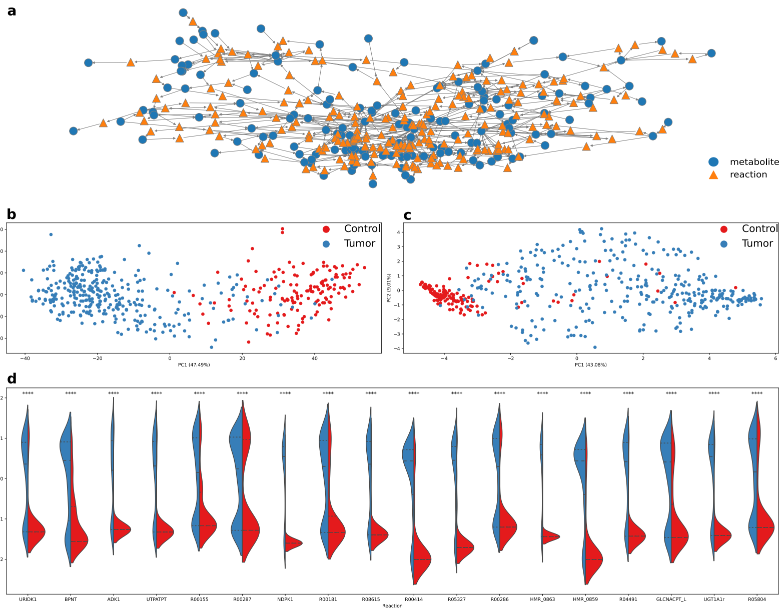


Figure 2: *mantra* results on Control vs. Triple-Negative Breast Cancer (TNBC). **a** Metabolic network build with the metabolites measured and identified in Xiao et al. [23]. Metabolites are shown as blue circular nodes, reactions as orange triangles. **b** PCA of the processed metabolite data with samples colored by condition. Despite a few overlapping samples, the groups are clearly separated by PC1, which explains almost half of the variance in the data. **c** PCA of the reaction activity estimates calculated by *mantra* with samples colored by condition. Similar to *b*, sample groups are separated by PC1, with only a minor reduction in the fraction of explained variance. In contrast to the original metabolome data, control samples show a reduced within-group variance, while tumor samples show a higher within-group variance. **d** Distributions of activity estimates for the most significantly changing reactions. Significance values were computed with a Wilcoxon rank sum test and Bonferroni correction. **** indicates a corrected p-value < 0.001.

be associated with breast cancer risk or are significant prognostic biomarkers [24, 19, 26, 27]. Furthermore, Adenylate Kinase 4 was found to regulate resistance to Tamoxifen treatment [29], which is used to treat Estrogen Receptor (ER) positive breast cancer patients. This is especially interesting since UGT enzymes conjugate ER ligands, forming a direct link between these reactions on a signaling level. These findings indicate that our approach is able to generate both usable and testable hypotheses on changes in metabolic activity.

In addition to this non-metabolic link between significant reactions, the reactions from Figure 2d form a connected subgraph of the metabolic network in Figure 2a, shown in Supplementary Figure S2. This supports our hypothesis, that in addition to metabolic networks being small-world networks, changes in metabolic activity are often constrained to a smaller subgraph of metabolic reactions, despite changes on metabolite level usually being observed throughout a larger part of the graph. Linking dysregulation of individual reactions back to broader metabolic mechanisms additionally allows one to elucidate mechanistic relations on a higher level. Because the computed activity values are computed for each sample, it is also possible to identify sub-populations within the group of disease samples. In combination with the mechanistic character of our method, this enables insights into metabolic alterations down to a patient-specific level making it a promising tool for precision medicine applications.

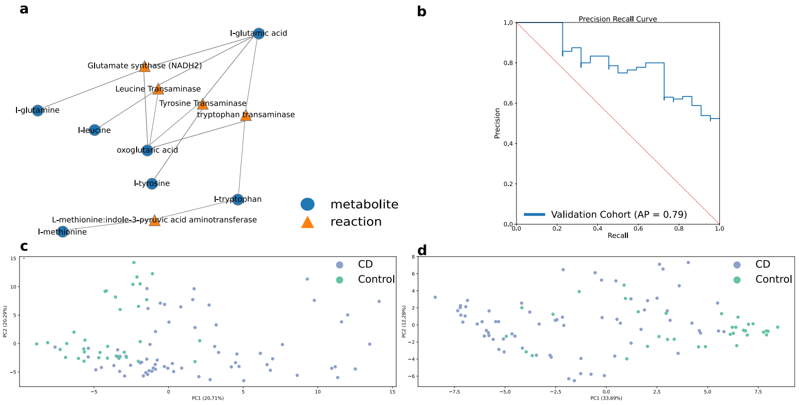


Figure 3: *mantra* results on Control vs. Crohn's Disease stool samples from Franzosa et al. [30] **a** Subgraphs identified by the local search-based enrichment, repeated 5 times with different random starts. Two disconnected subgraphs are identified, one representing a part of cholic acid metabolism and the other one a part of amino acid metabolism. **b** Precision Recall (PR)-curve showing the predictive performance of a random forest model on the validation cohort trained on the PRISM cohort (both from Franzosa et al. [30]). The Expected performance by a random model is depicted by the red dashed line, the blue curve indicates the performance of the reaction activity data. With a PR-Area Under the Curve (AUC) of 0.79 the reaction estimate-based prediction seems to generalize well to the validation cohort. **c** PCA of the processed metabolite data. Samples colored by condition, which are mainly separated by PC1, explaining 20% of the variance in the dataset. Generally, control samples are more similar to each other than CD samples. **d** PCA on the basis of the reaction activity values computed with *mantra* with samples colored by condition. Even though PC1 explains a larger proportion of variance, the separation between sample groups is less clear than for the metabolome data.

2.2 Application to Inflammatory Bowel Disease and Multi-omics Integration

The second analysis we present here is applying the *mantra* approach to data from the PRISM cohort from Franzosa et al. [30]. It consists of 155 stool samples, 34 control and 121 Inflammatory Bowel Disease (IBD) samples, which are grouped into 68 Crohn's Disease (CD) and 53 Ulcerative Colitits (UC) samples. For the presented analysis only control and CD samples were used. After filtering and mapping the metabolites and microbial species onto the internal database (for details see Materials & Methods), 138 metabolites and 70 microbial species were retained and 108 metabolic reactions were included in the metabolic network.

Linking Metabolomics-based Reaction Activity to Differential Enzymatic Potential An overview of the variance in the processed metabolome data and the computed reaction activity data are shown using PCA in Figure 3 c and d, respectively. Generally, both spaces seem to discriminate the patient groups well. Sub-figure c shows a good separation of CD and control samples along PC1, which explains around 20% of the total variance. Notably, CD samples appear to be metabolically more diverse than control samples. While the separation between CD and control along PC1 in based sub-figure d, which is based on our proposed reaction activity estimates, is a little less clear with some control outliers, the explained variance along this component is increased to around 33%.

To find a metabolically connected subgraph with high changes in reaction estimates between control and CD samples, we employed Simulated Annealing (SA) together with a local search, where the objective is to maximize the difference in distributions between control and CD samples (for details see Materials & Methods). This strategy is essentially an extension of the analysis presented in the previous section. Figure 3a, depicting the subgraph identified by the enrichment analysis, shows

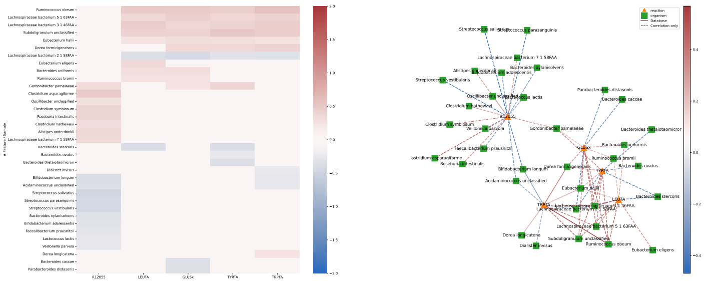


Figure 4: Spearman's rank correlation-based associations between microbial species and enriched metabolic reactions. Reaction IDs matching the following reactions in Figure 3a: GLUSx, GLUSy, GLFRDOi: Glutamate synthase; LEUTA, LEUTAm: Leucine Transaminase; TYRTA, TYRTAm: Tyrosine Transaminase; TRPTA: tryptophan transaminase; R12055: methionine:indole-3-pyruvic acid aminotransferase. **a** Correlation matrix showing the difference in correlation coefficient between control and Crohn's Disease (CD) samples for the reactions identified in the network enrichment and all microbial species. All correlation coefficients with a p-value > 0.05 are set to zero prior to computing the differences. **b** Correlation network resulting from the correlation matrix in a. Node color and shape indicated node type, edge color indicates the respective difference in correlation coefficient.

that exclusively amino acid interconversion reactions, mostly involving glutamate, are identified. Since Franzosa et al. [30] also performed metagenomics to quantify the differences in enzymatic capabilities in the gut lumen of control and CD samples, we checked whether any enzymes catalyzing the reactions in the subgraph are found to be significantly different (Supplementary Dataset 7 in [30]). Indeed, for 4 out of the 5 reactions, at least one enzyme capable of driving them is found to be differentially abundant between conditions (Supplementary Table S3). The only non-significant reaction, matched to Branched-chain-amino-acid transaminase, has a corrected q-value of 0.054, making it just barely missing the q-value cutoff of 0.05 used by the authors of [30]. With the general discussion about the choice of p-values and cutoffs in mind, one can conclude that the metabolic reactions identified with *mantra* are well reflected by the results of the metagenomics analysis.

The identified conversion between glutamine and glutamate is especially well-connected to IBD literature due to the role of glutamine-based signaling in the regulation of tissue integrity, inflammatory processes, and apoptosis [31]. The regulation of apoptosis is directly influenced by the conversion of glutamine to glutamate, which is then used to produce the reduced form of glutathione (GSH) together with cysteine and glycine [32, 33]. GSH is then used to regulate the redox potential via the binding of 2 GSH, resulting in the oxidized form of glutathione (GSSG). While only glutamate synthase was previously directly associated with IBD, tryptophan, tyrosine, and methionine metabolism have also been linked to CD [34, 35]. The results of this analysis in combination with the fact that all reactions are represented by differentially abundant enzymes, suggest that the subgraph identified by the local search algorithm on the basis of our proposed reaction activity metric yields a hypothesis consistent with additional measurements on expected enzyme levels and existing literature.

In addition to the PRISM cohort, which was used in the analyses presented above, Franzosa et al. [30] also introduce a validation cohort from a different hospital containing 22 control and 20 CD samples. We used this cohort to evaluate how well a predictive model trained on our reaction activity metric can generalize to a different cohort. The PR curve of this evaluation is shown in Figure 3b. The AUC of 0.79 demonstrates that the reaction activity estimation introduced is also generalizing across independent cohorts, making it suitable for clinical analyses. Additionally, the Receiver Operating Characteristic (ROC) curve for the same evaluation (AUC of 0.74) is shown in Supplementary Figure S3.

Associating Metabolic Reaction Activity with Microbial Species Abundances A major advantage of using the residual variance instead of e.g. correlation coefficient is that they give per-sample estimates. Hence, they can be used to compute associations of metabolic reaction activity with features from other modalities. Computing all pairwise Spearman's correlation coefficients between reaction values and microbial species resulted in 249 significant associations (q-value < 0.05) in the control group and 300 in the CD group, with 25 shared associations between the two groups. While these numbers appear rather low, the proportion of significant metabolite-microbe associations found by Franzosa et al. [30] is also below 10%. The correlation matrices and networks over all features are shown in Supplementary Figure S4. Since interpreting these networks without further analyses is a tedious task given their size and complex structure, we continue our analysis of multi-omics associations using the amino acid-reaction subgraph identified in the previous subsection (Figure 3a). More specifically, we look at the differences in correlations between control and CD patients. The rationale behind taking this approach is that a change in correlation potentially indicates a difference in the association between the features under the two conditions. In total, 35 microbial species have at least one significant association with a metabolic reaction that is different between the sample groups (Figure 4a). A large proportion of associations that are more positive in the control group are between a set of species and the three transaminase reactions, as well as the glutamate synthase reaction. The species with significant correlation changes to the methionine-related reaction, in contrast, are almost exclusively correlated with this reaction. Positively valued species are dominated by *Clostridium*, whereas those with negative values are dominated by *Streptococcus*.

3 Discussion

In this work, we introduced an approach to estimate the changes in reaction activity between biological conditions using untargeted metabolomics data and metabolic networks. We demonstrate the ability of the proposed heuristic to recover known key reactions in biopsies of healthy and TNBC tissue and in stool samples of CD and non-IBD patients as a proof of concept. Furthermore, we showcase the possibility of computing multi-omics associations to these relative reaction activities via correlation metrics.

Despite the increasing use of metabolomics to study biological and biomedical phenomena, computational methods for a mechanistic interpretation of metabolomics data are rare. Especially metabolic reactions are only targeted when using metabolic modeling or when doing enrichment analyses using metabolite p-values (e.g. in [20]). The former is exact but often not feasible due to the requirements for data and model parameterization [36]. The latter option considers the biochemical connections between metabolites but disregards their quantitative relations and discards valuable information by discretizing with a threshold value. Inferring metabolic reaction activity without absolute quantification is hard, as values (i.e. MS intensities) are typically on metabolite-specific scales. Hence, even with a complete metabolic model, exact modeling is not possible. The approach we introduce circumvents these shortcomings by investigating the relative change of metabolite relations between sample groups on the basis of linear models. Based on the rationale that metabolites participating in the same active metabolic reaction show correlating properties [37], captured by the linear model, we propose that a change in the coefficients, describing the substrate-product relations, can be used as an indicator for a change in reaction activity. Following this idea, we use the residuals of two conditions for the same model as a proxy for this change of coefficients. The general applicability of our approach to generate mechanistic hypotheses is demonstrated by the results depicted in Figure 2d and Figure 3a and b, where we show that it pinpoints reactions by only using untargeted metabolomics data that were previously validated as key players in more complex experiments. Nevertheless, our approach may be limited with respect to non-linear behavior in reactions [37].

An advantage of using linear models is that correction of covariates, especially relevant in clinical studies, as well as the usage of regularized models such as elastic net [38] is simple and already implemented in the published code (see Code Availability Statement). Since each reaction is described by a separate model, our method considers each metabolic reaction in isolation, whereas in reality, reactions are connected through molecules participating in both reactions and many changes are propagated through the network. In cases where a consecutive sequence of reactions is actually changing, this might lead to the method only picking up the flanking reactions of this sequence. Additionally, in studies where the effect of external metabolite administration is "directly" given to the tissue/body site of sampling, like dietary interventions paired with gut/stool metabolomics, our method can be biased toward reactions in which the administered metabolites participate. Despite many disease-unrelated metabolic changes constantly happening, it is unlikely that such changes are picked up by the model. Especially in clinical studies, where high inter-sample heterogeneity is common, these effects will mostly remain as noise, whereas the true underlying changes are more constant across samples.

While this manuscript only evaluates application cases in a supervised form, the proposed idea can also be used in an unsupervised context. Therefore, applications such as *de-novo* subtyping need to be evaluated with respect to the robustness of similar metrics in future work. While the main focus of this work is the evaluation of a metabolic reaction activity metric, we also introduce a local search-based enrichment method to perform subgraph enrichment. Despite the good results, this approach is influenced by hyperparameter selection, such as temperature and allowed solution sizes for SA, and can become slow depending on the size of the network and the parameter settings. For example, the initial temperature T_0 in SA controls the probability of accepting a random solution and thus the degree of exploration across the objective function landscape. Consequently, the choice of initial temperature is a trade-off between exploration and exploitation, and inappropriate temperature settings can lead to unstable or considerably sub-optimal solutions.

The participation of some metabolites in a large number of metabolic reactions may also be problematic from a graph-topological point of view due to their high connectivity within the metabolic network. Since our analysis does not include the neighborhood of a reaction, this does not affect the metric but only possible downstream applications acting on the network. However, it is not possible to distinguish reactions with the same substrates and products, as these will result in the same linear model, as well as the directionality of the reaction. In practical applications, this can mean that the hypothesis can include a larger number of possible enzyme candidates and, thus, more laborious hypothesis validation, even when the size of the results themselves is rather small.

Despite these limitations, the demonstrated ability of our method to accurately propose mechanistic hypotheses makes it a promising approach to improve the functional interpretation of metabolomics data in many experimental setups. By providing a metric for the quantitative approximation of reaction activity changes, it paves the way for a novel class of metabolomics data analysis methods. Given the wide-spread associations between functional metabolic changes in diseases [39, 40, 41, 42], we believe these developments can directly impact clinical research. In addition, the presented results give rise to the development of new strategies for prior knowledge-guided functional multi-omics integration to further strengthen biological and biomedical research on the level of metabolism.

4 Materials & Methods

4.1 Network Generation & ID mapping

To have a comprehensive database for human and microbial metabolism, we use a custom database. It contains the merged information from the Virtual Metabolic Human project [22], the KEGG [13], and the Reactome [21] database. The database is available through the provided Python package either through a public API or locally via a docker application provided (Code Availability Statement).

For a given metabolomics dataset, the identified metabolites are first mapped onto the internal database, either by directly using database identifiers, if given or by using the Metaboanalyst [43] name conversion API (`http://api.xialab.ca/mapcompounds`). Subsequently, the metabolic subnetwork containing all measured and mapped metabolites and all metabolic reactions for which at least one substrate and one product are measured is extracted. Metabolites not connected to any reaction in the subnetwork are removed.

4.2 Estimating Changes in Reaction Activity

The estimation of relative activity is based on reaction-wise linear models. This is based on findings from Krumsiek et al. [37] who showed that metabolites involved in active metabolic reactions have correlating properties. Prior to computing these models, all metabolites are mean-centered and scaled to unit variance to avoid scale-based biases in the residuals for each reaction. The input data is assumed to be normalized and transformed to follow (approximately) a normal distribution. Intuitively, our model tries to capture the change in relations between the substrates and products of a metabolic reaction between two conditions. The underlying assumption is that if reaction activity stays the same, so should the model coefficients, while they are expected to change if activity changes. Hence, while there can be similar correlation strength in both sample groups, the quantitative relation describing this correlation can change.

For each reaction in the metabolic network, the substrate abundances (S) are the predicting (i.e. independent) variables, and the product abundances (P) are the dependent variables. Therefore, each reaction model can be a multiple and/or multivariate regression, depending on the number of substrates and products that were measured for a particular reaction. The model for each reaction is formally defined as

$$P = \beta S$$

where β are the coefficients of the model and P and S are the intensities of the products and substrates of the respective reaction. Reactions with the same substrates and products, such as transport reactions, are skipped because $P = S$. The fitting of the model coefficients is happening in one of the sample groups. To avoid corruption of models by outliers, for each reaction, Cook's distance [44] is computed for every sample. It is defined as the change of prediction relative to the model's error when a sample is left out. For a sample i its Cook's distance D_i is computed as

$$D_i = \frac{\sum_{j=1}^{n} \left(\hat{y}_j - \hat{y}_{j(i)}\right)^2}{ps^2} \tag{1}$$

with $\hat{y}_j$ and $\hat{y}_{j(i)}$ being the predicted product intensities of the total model and the model without sample i respectively, for each product j of a reaction, s^2 representing the model's mean square error, and p the rank of the model. Generally, a high Cooks's distance indicates a high influence of a sample and thus makes it more likely to be an outlier. If any sample has a distance value above a user-defined threshold, the model is fitted again without those samples. Alternatively, if no distance threshold is defined, we use the survival function of an F-distribution to provide p-values for selecting outliers to remove. Subsequently, the coefficient of determination (R^2) is used to filter out all models that fail to describe the relation between substrate and product values adequately. Under the assumption that some metabolic reactions may only be active in one of the sample groups, this might mean that reactions are removed, which potentially describes a major difference between conditions. Therefore, we provide the option to re-compute models that fail in one group on samples from another group.

The reaction-wise models computed with the procedure explained above are then used to compute the reaction values for all samples. More specifically, the reaction activity estimates are calculated as the (normed) proportion of explained variance. For a sample i and a reaction r this estimate is defined as

$$a_{ir} = 1 - \frac{RSS_{ir}}{TSS_r} = 1 - \frac{\left(y_{ir} - \hat{y}_{ir}\right)^2}{\sum_{j=1}^{n} \left(y_{jr} - \bar{y}_r\right)^2} \tag{2}$$

where $\hat{y}$ represents the predicted product values and $\bar{y}$ the mean product values. The summation in the denominator over j represents the total sum of squares (TSS), i.e. the residual sum of squares summed over all samples. Subsequently, the estimates are mean-centered and scaled to unit variance.

To obtain p-values describing how statistically different the activity estimate distributions between groups are, a Wilcoxon rank sum test is calculated for each reaction on the basis of the activity estimates computed as described before. The reported p-values are family-wise error rate (Bonferroni) corrected.

4.3 Multi-Omics Associations

To compute multi-omics associations, we use Spearman's rank correlation implemented in the scipy package [45]. Corresponding p-values above a user-defined threshold are used to set the respective correlation coefficients to zero. Subsequently, all reactions and multi-omics features without a single significant correlation coefficient are removed. Correlations can either be computed for each group individually or over all samples.

4.4 Network Enrichment

The combinatorial optimization algorithm used for the network enrichment is an adapted version of a local search approach used in Rose and Köhler et al. [46]. Local search generally examines a search space in a greedy manner by iteratively testing local candidate solutions for the one with an optimal objective function. Candidate solutions are generated by applying one of three operations: node insertion, deletion, and substitution to the solution from the last iteration or a randomly selected subgraph in the first iteration.

While this procedure is sufficient to find local maxima, it cannot escape them, and thus, the resulting optimal solution is highly dependent on the initial starting point and the landscape of the objective function over the entire graph. Therefore, we use SA [47] to avoid stagnation and increase the chances of finding a global maximum. SA allows accepting non-optimal subgraphs with a probability decreasing with the number of iterations. This decrease is implemented through a temperature parameter T exponentially decaying by a rate α with the number of iterations n. Whenever the local search reaches a local maximum, a sub-optimal solution is accepted if

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

where o_n and o_{n-1} are the values of the objective function in the current and last iteration.

The objective function is defined as

$$o = \frac{1}{|SG|} \sum_{r \in SG} 1 - W(a_{r,x}, a_{r,y})$$

where SG indicates the set of vertices forming the current solution and W is the p-value of a Welch's test comparing group x against group y. It is defined such that the local search maximizes the difference between the reaction activity estimate distributions between the two sample groups.

Even when using SA, finding a global optimum is not guaranteed and in some cases, multiple global optima might exist. Hence, the enriched subgraph is dependent on the randomly chosen initial solution. Therefore, our method runs the enrichment algorithm multiple times with different seeds and returns the union of all solutions. Another option instead of using the union would be the intersection of all repeats (an option in the package). In an intuitive sense, the union subgraph would give a lower probability of "false negatives" at the expense of a higher chance of including "false positives". In practice, the choice also depends on the settings of other hyperparameters, such as the allowed solution size and the downstream experiments that the enrichment results should be used for.

4.5 Data Processing and Experiments

Common to both experiments presented in this manuscript is the usage of the networkx [48] and matplotlib [49] packages for handling network visualization and the scikit-learn library [50] for generating PCAs.

4.5.1 Triple-Negative Breast Cancer Data Set

The metabolomics data provided by Xiao et al. [23] gives already normalized metabolite data. Therefore, only missing-value imputation with half of the feature-wise minimum, mean centering, and unit variance scaling was applied. To map the measured metabolites to our internal database, the HMDB [51] and KEGG [13] identifiers provided with the feature annotation were used together with a mapping database provided with the python package (see Code Availability Statement). Following the name mapping step, the data-specific metabolic network was also generated with package-provided functions using a neo4j database that we made publicly available (see Code Availability Statement).

For analyzing and visualizing the residual distributions, the scipy [45] and seaborn [52] libraries were used.

4.5.2 Inflammatory Bowel Disease Data Set

For the metabolomics data from Franzosa et al. [30] missing values were imputed using half the feature-wise minimum. Subsequently, samples were quotient normalized [53] to account for sample-specific dilution effects and log-transformation was applied, such that features are approximately normally distributed. All features were then mean-centered and scaled to unit variance. To avoid leakage between the PRISM (discovery) and the validation cohort, the PRISM cohort was processed and parameters for each (parameter-dependent) step were retained. The validation cohort was then processed with the parameters from the discovery cohort.

Since the available data did not contain database identifiers, we used the Metaboanalyst [43] name conversion API (`http://api.xialab.ca/mapcompounds`) to obtain HMDB and KEGG IDs for all metabolites. The remaining name mapping and network generation steps are the same as described above for the other analyzed data set.

Microbial species data was imputed by setting all zero values to the total minimum divided by the number of zero values before applying centered log-ratio transformation [54].

For assessing the generalization of the predictive model using our reaction activity metric, the random forest implementation from the scikit-learn package [50] was used with the default parameter (no hyperparameter optimization was done). Reaction models were computed on the discovery, followed by training the classifier on the resulting values. Subsequently, PR and ROC curves were computed on the validation samples with reaction values estimated using the reaction models fitted on the discovery cohort control samples. The curves shown were also generated with scikit-learn and matplotlib [49] functions.

Multi-omics correlations were computed based on the computed reaction metrics and the processed microbial species data using Spearman's correlation coefficient.

Data Availability Statement

All data used in the experiments is publicly available from the referenced articles by Xiao et al. [23] and Franzosa et al. [30].

Code Availability Statement

Source code including all presented experiments: `https://github.com/lipitum/pymantra`
Python package: `https://pypi.org/project/pymantra`
Database and API: `https://github.com/lipitum/pymantraAPI`

Author Contributions

NK and JKP planned the work. NK designed and implemented the reaction activity estimation method and the network enrichment procedure. VW and NK parsed and merged the reaction databases. NK ran the evaluations. NK, JKP, VW and TDR wrote and reviewed the manuscript. JKP secured the funding. All authors read and accepted the manuscript in its final form.

Acknowledgments

We thank Felix Niedermaier for helping with the database API setup and Emma Schonner for internal evaluations that are not part of this manuscript. 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; JKP, VW, TDR, NK: Junior Research Group LipiTUM). TDR is supported through state funds approved by the State Parliament of Baden-Württemberg for the Innovation Campus Health + Life Science Alliance Heidelberg Mannheim.

Conflicts of interest

The authors declare no conflicts of interest.

Abbreviations

AUC	Area Under the Curve
CD	Crohn's Disease

CMPK1	UMP-CMP kinase
CV	Cross Validation
ER	Estrogen Receptor
IBD	Inflammatory Bowel Disease
MS	Mass Spectrometry
PCA	Principal Component Analysis
PLS(-DA)	Partial Least Squares(-Discriminant Analysis)
PR	Precision Recall
ROC	Receiver Operating Characteristic
SA	Simulated Annealing
TNBC	Triple-Negative Breast Cancer
UC	Ulcerative Colitits
UDP	Uridine Di-Phosphate
UGDH	UDP-glucose-dehydrogenase
UGT	UDP-glucuronosyltransferase
UMP	Uridine Mono-Phosphate

References

[1] David S. Wishart. "Metabolomics for Investigating Physiological and Pathophysiological Processes". In: *Physiological Reviews* 99.4 (Oct. 2019). Publisher: American Physiological Society, pp. 1819–1875. DOI: `10.1152/physrev.00035.2018`.

[2] Yue Xiong and Kun-Liang Guan. "Mechanistic insights into the regulation of metabolic enzymes by acetylation". In: *The Journal of Cell Biology* 198.2 (July 2012), pp. 155–164. DOI: `10.1083/jcb.201202056`.

[3] Abhisha Sawant Dessai, Poonam Kalhotra, Aaron T. Novickis, and Subhamoy Dasgupta. "Regulation of tumor metabolism by post translational modifications on metabolic enzymes". en. In: *Cancer Gene Therapy* (Aug. 2022). Publisher: Nature Publishing Group, pp. 1–11. DOI: `10.1038/s41417-022-00521-x`.

[4] Vincent J Hilser. "An ensemble view of allostery". In: *Science* 327.5966 (2010), pp. 653–654.

[5] Luca Gerosa and Uwe Sauer. "Regulation and control of metabolic fluxes in microbes". In: *Current opinion in biotechnology* 22.4 (2011), pp. 566–575.

[6] Janet E Lindsley and Jared Rutter. "Whence cometh the allosterome?" In: *Proceedings of the National Academy of Sciences* 103.28 (2006), pp. 10533–10535.

[7] Jeremy K Nicholson. "Global systems biology, personalized medicine and molecular epidemiology". In: *Molecular Systems Biology* 2.1 (2006), p. 52. DOI: `https://doi.org/10.1038/msb4100095`.

[8] Claire L Gavaghan, Elaine Holmes, Eva Lenz, Ian D Wilson, and Jeremy K Nicholson. "An NMR-based metabonomic approach to investigate the biochemical consequences of genetic strain differences: application to the C57BL10J and Alpk: ApfCD mouse". In: *FEBS letters* 484.3 (2000), pp. 169–174.

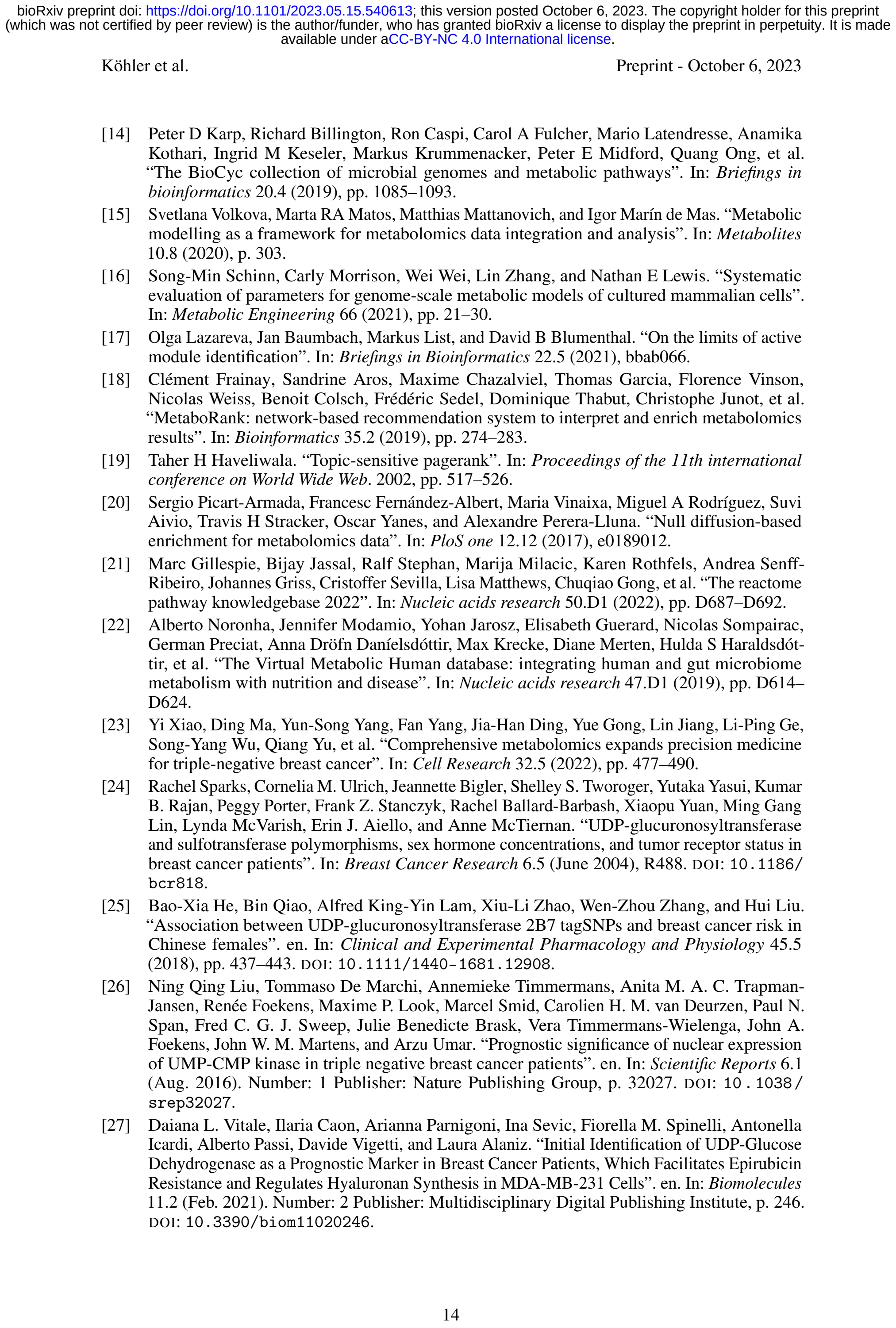
[9] Yang Chen, En-Min Li, and Li-Yan Xu. "Guide to Metabolomics Analysis: A Bioinformatics Workflow". en. In: *Metabolites* 12.4 (Apr. 2022). Number: 4 Publisher: Multidisciplinary Digital Publishing Institute, p. 357. DOI: `10.3390/metabo12040357`.

[10] Jianguo Xia and David S. Wishart. "MSEA: a web-based tool to identify biologically meaningful patterns in quantitative metabolomic data". In: *Nucleic Acids Research* 38 (May 2010), W71–W77. DOI: `10.1093/nar/gkq329`.

[11] Atanas Kamburov, Rachel Cavill, Timothy MD Ebbels, Ralf Herwig, and Hector C Keun. "Integrated pathway-level analysis of transcriptomics and metabolomics data with IMPaLA". In: *Bioinformatics* 27.20 (2011), pp. 2917–2918.

[12] Adam Amara, Clément Frainay, Fabien Jourdan, Thomas Naake, Steffen Neumann, Elva María Novoa-Del-Toro, Reza M Salek, Liesa Salzer, Sarah Scharfenberg, and Michael Witting. "Networks and graphs discovery in metabolomics data analysis and interpretation". In: *Frontiers in Molecular Biosciences* (2022), p. 223.

[13] Minoru Kanehisa, Yoko Sato, Masayuki Kawashima, Miho Furumichi, and Mao Tanabe. "KEGG as a reference resource for gene and protein annotation". In: *Nucleic acids research* 44.D1 (2016), pp. D457–D462.



[14] Peter D Karp, Richard Billington, Ron Caspi, Carol A Fulcher, Mario Latendresse, Anamika Kothari, Ingrid M Keseler, Markus Krummenacker, Peter E Midford, Quang Ong, et al. "The BioCyc collection of microbial genomes and metabolic pathways". In: *Briefings in bioinformatics* 20.4 (2019), pp. 1085–1093.

[15] Svetlana Volkova, Marta RA Matos, Matthias Mattanovich, and Igor Marín de Mas. "Metabolic modelling as a framework for metabolomics data integration and analysis". In: *Metabolites* 10.8 (2020), p. 303.

[16] Song-Min Schinn, Carly Morrison, Wei Wei, Lin Zhang, and Nathan E Lewis. "Systematic evaluation of parameters for genome-scale metabolic models of cultured mammalian cells". In: *Metabolic Engineering* 66 (2021), pp. 21–30.

[17] Olga Lazareva, Jan Baumbach, Markus List, and David B Blumenthal. "On the limits of active module identification". In: *Briefings in Bioinformatics* 22.5 (2021), bbab066.

[18] Clément Frainay, Sandrine Aros, Maxime Chazalviel, Thomas Garcia, Florence Vinson, Nicolas Weiss, Benoit Colsch, Frédéric Sedel, Dominique Thabut, Christophe Junot, et al. "MetaboRank: network-based recommendation system to interpret and enrich metabolomics results". In: *Bioinformatics* 35.2 (2019), pp. 274–283.

[19] Taher H Haveliwala. "Topic-sensitive pagerank". In: *Proceedings of the 11th international conference on World Wide Web*. 2002, pp. 517–526.

[20] Sergio Picart-Armada, Francesc Fernández-Albert, Maria Vinaixa, Miguel A Rodríguez, Suvi Aivio, Travis H Stracker, Oscar Yanes, and Alexandre Perera-Lluna. "Null diffusion-based enrichment for metabolomics data". In: *PloS one* 12.12 (2017), e0189012.

[21] Marc Gillespie, Bijay Jassal, Ralf Stephan, Marija Milacic, Karen Rothfels, Andrea Senff-Ribeiro, Johannes Griss, Cristoffer Sevilla, Lisa Matthews, Chuqiao Gong, et al. "The reactome pathway knowledgebase 2022". In: *Nucleic acids research* 50.D1 (2022), pp. D687–D692.

[22] Alberto Noronha, Jennifer Modamio, Yohan Jarosz, Elisabeth Guerard, Nicolas Sompairac, German Preciat, Anna Dröfn Daníelsdóttir, Max Krecke, Diane Merten, Hulda S Haraldsdóttir, et al. "The Virtual Metabolic Human database: integrating human and gut microbiome metabolism with nutrition and disease". In: *Nucleic acids research* 47.D1 (2019), pp. D614–D624.

[23] Yi Xiao, Ding Ma, Yun-Song Yang, Fan Yang, Jia-Han Ding, Yue Gong, Lin Jiang, Li-Ping Ge, Song-Yang Wu, Qiang Yu, et al. "Comprehensive metabolomics expands precision medicine for triple-negative breast cancer". In: *Cell Research* 32.5 (2022), pp. 477–490.

[24] Rachel Sparks, Cornelia M. Ulrich, Jeannette Bigler, Shelley S. Tworoger, Yutaka Yasui, Kumar B. Rajan, Peggy Porter, Frank Z. Stanczyk, Rachel Ballard-Barbash, Xiaopu Yuan, Ming Gang Lin, Lynda McVarish, Erin J. Aiello, and Anne McTiernan. "UDP-glucuronosyltransferase and sulfotransferase polymorphisms, sex hormone concentrations, and tumor receptor status in breast cancer patients". In: *Breast Cancer Research* 6.5 (June 2004), R488. DOI: 10.1186/bcr818.

[25] Bao-Xia He, Bin Qiao, Alfred King-Yin Lam, Xiu-Li Zhao, Wen-Zhou Zhang, and Hui Liu. "Association between UDP-glucuronosyltransferase 2B7 tagSNPs and breast cancer risk in Chinese females". en. In: *Clinical and Experimental Pharmacology and Physiology* 45.5 (2018), pp. 437–443. DOI: 10.1111/1440-1681.12908.

[26] Ning Qing Liu, Tommaso De Marchi, Annemieke Timmermans, Anita M. A. C. Trapman-Jansen, Renée Foekens, Maxime P. Look, Marcel Smid, Carolien H. M. van Deurzen, Paul N. Span, Fred C. G. J. Sweep, Julie Benedicte Brask, Vera Timmermans-Wielenga, John A. Foekens, John W. M. Martens, and Arzu Umar. "Prognostic significance of nuclear expression of UMP-CMP kinase in triple negative breast cancer patients". en. In: *Scientific Reports* 6.1 (Aug. 2016). Number: 1 Publisher: Nature Publishing Group, p. 32027. DOI: 10.1038/srep32027.

[27] Daiana L. Vitale, Ilaria Caon, Arianna Parnigoni, Ina Sevic, Fiorella M. Spinelli, Antonella Icardi, Alberto Passi, Davide Vigetti, and Laura Alaniz. "Initial Identification of UDP-Glucose Dehydrogenase as a Prognostic Marker in Breast Cancer Patients, Which Facilitates Epirubicin Resistance and Regulates Hyaluronan Synthesis in MDA-MB-231 Cells". en. In: *Biomolecules* 11.2 (Feb. 2021). Number: 2 Publisher: Multidisciplinary Digital Publishing Institute, p. 246. DOI: 10.3390/biom11020246.

[28] Satu Tiainen, Sanna Oikari, Markku Tammi, Kirsi Rilla, Kirsi Hämäläinen, Raija Tammi, Veli-Matti Kosma, and Päivi Auvinen. "High extent of O-GlcNAcylation in breast cancer cells correlates with the levels of HAS enzymes, accumulation of hyaluronan, and poor outcome". en. In: *Breast Cancer Research and Treatment* 160.2 (Nov. 2016), pp. 237–247. DOI: 10.1007/s10549-016-3996-4.

[29] Xiaochuan Liu, Gwendolyn Gonzalez, Xiaoxia Dai, Weili Miao, Jun Yuan, Ming Huang, David Bade, Lin Li, Yuxiang Sun, and Yinsheng Wang. "Adenylate Kinase 4 Modulates the Resistance of Breast Cancer Cells to Tamoxifen through an m6A-Based Epitranscriptomic Mechanism". English. In: *Molecular Therapy* 28.12 (Dec. 2020). Publisher: Elsevier, pp. 2593–2604. DOI: 10.1016/j.ymthe.2020.09.007.

[30] Eric A Franzosa, Alexandra Sirota-Madi, Julian Avila-Pacheco, Nadine Fornelos, Henry J Haiser, Stefan Reinker, Tommi Vatanen, A Brantley Hall, Himel Mallick, Lauren J McIver, et al. "Gut microbiome structure and metabolic activity in inflammatory bowel disease". In: *Nature microbiology* 4.2 (2019), pp. 293–305.

[31] Min-Hyun Kim and Hyeyoung Kim. "The Roles of Glutamine in the Intestine and Its Implication in Intestinal Diseases". en. In: *International Journal of Molecular Sciences* 18.5 (May 2017). Number: 5 Publisher: Multidisciplinary Digital Publishing Institute, p. 1051. DOI: 10.3390/ijms18051051.

[32] Erich Roth, Rudolf Oehler, Nicole Manhart, Ruth Exner, Barbara Wessner, Eva Strasser, and Andreas Spittler. "Regulative potential of glutamine—relation to glutathione metabolism". In: *Nutrition* 18.3 (2002), pp. 217–221.

[33] AG Hall. "The role of glutathione in the regulation of apoptosis." In: *European journal of clinical investigation* 29.3 (1999), pp. 238–245.

[34] Almut Heinken, Johannes Hertel, and Ines Thiele. "Metabolic modelling reveals broad changes in gut microbial metabolism in inflammatory bowel disease patients with dysbiosis". en. In: *npj Systems Biology and Applications* 7.1 (May 2021). Number: 1 Publisher: Nature Publishing Group, pp. 1–11. DOI: 10.1038/s41540-021-00178-6.

[35] Xochitl C. Morgan, Timothy L. Tickle, Harry Sokol, Dirk Gevers, Kathryn L. Devaney, Doyle V. Ward, Joshua A. Reyes, Samir A. Shah, Neal LeLeiko, Scott B. Snapper, Athos Bousvaros, Joshua Korzenik, Bruce E. Sands, Ramnik J. Xavier, and Curtis Huttenhower. "Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment". In: *Genome Biology* 13.9 (Sept. 2012), R79. DOI: 10.1186/gb-2012-13-9-r79.

[36] Amit Rai and Kazuki Saito. "Omics data input for metabolic modeling". In: *Current opinion in biotechnology* 37 (2016), pp. 127–134.

[37] Jan Krumsiek, Karsten Suhre, Thomas Illig, Jerzy Adamski, and Fabian J Theis. "Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data". In: *BMC systems biology* 5 (2011), pp. 1–16.

[38] Hui Zou and Trevor Hastie. "Regularization and variable selection via the elastic net". In: *Journal of the royal statistical society: series B (statistical methodology)* 67.2 (2005), pp. 301–320.

[39] Khushboo G Upadhyay, Devendra C Desai, Tester F Ashavaid, and Alpa J Dherai. "Microbiome and metabolome in inflammatory bowel disease". In: *Journal of Gastroenterology and Hepatology* 38.1 (2023), pp. 34–43.

[40] Daniel R Schmidt, Rutulkumar Patel, David G Kirsch, Caroline A Lewis, Matthew G Vander Heiden, and Jason W Locasale. "Metabolomics in cancer research and emerging applications in clinical oncology". In: *CA: a cancer journal for clinicians* 71.4 (2021), pp. 333–358.

[41] Vivian Tounta, Yi Liu, Ashleigh Cheyne, and Gerald Larrouy-Maumus. "Metabolomics in infectious diseases and drug discovery". In: *Molecular Omics* 17.3 (2021), pp. 376–393.

[42] Christopher B Newgard. "Metabolomics and metabolic diseases: where do we stand?" In: *Cell metabolism* 25.1 (2017), pp. 43–56.

[43] Jasmine Chong, David S Wishart, and Jianguo Xia. "Using MetaboAnalyst 4.0 for comprehensive and integrative metabolomics data analysis". In: *Current protocols in bioinformatics* 68.1 (2019), e86.

[44] R Dennis Cook. "Detection of influential observation in linear regression". In: *Technometrics* 19.1 (1977), pp. 15–18.

[45] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python". In: *Nature Methods* 17 (2020), pp. 261–272. DOI: `10.1038/s41592-019-0686-2`.

[46] Tim D Rose, Nikolai Köhler, Lisa Falk, Lucie Klischat, Olga E Lazareva, and Josch K Pauling. "Lipid network and moiety analysis for revealing enzymatic dysregulation and mechanistic alterations from lipidomics data". In: *Briefings in Bioinformatics* 24.1 (2023), bbac572.

[47] Peter JM Van Laarhoven, Emile HL Aarts, Peter JM van Laarhoven, and Emile HL Aarts. *Simulated annealing*. Springer, 1987.

[48] Aric Hagberg, Pieter Swart, and Daniel S Chult. *Exploring network structure, dynamics, and function using NetworkX*. Tech. rep. Los Alamos National Lab.(LANL), Los Alamos, NM (United States), 2008.

[49] J. D. Hunter. "Matplotlib: A 2D graphics environment". In: *Computing in Science & Engineering* 9.3 (2007), pp. 90–95. DOI: `10.1109/MCSE.2007.55`.

[50] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

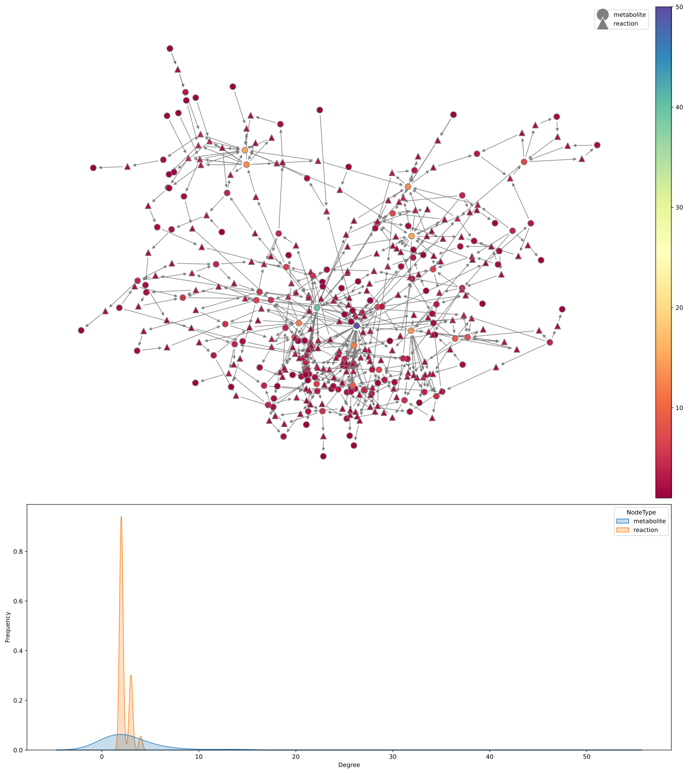
[51] David S Wishart, AnChi Guo, Eponine Oler, Fei Wang, Afia Anjum, Harrison Peters, Raynard Dizon, Zinat Sayeeda, Siyang Tian, Brian L Lee, et al. "HMDB 5.0: the human metabolome database for 2022". In: *Nucleic Acids Research* 50.D1 (2022), pp. D622–D631.

[52] Michael L. Waskom. "seaborn: statistical data visualization". In: *Journal of Open Source Software* 6.60 (2021), p. 3021. DOI: `10.21105/joss.03021`.

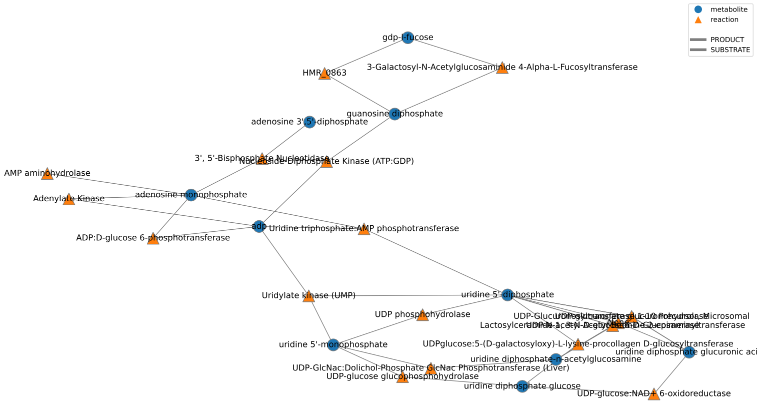
[53] Frank Dieterle, Alfred Ross, Götz Schlotterbeck, and Hans Senn. "Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in 1H NMR metabonomics". In: *Analytical chemistry* 78.13 (2006), pp. 4281–4290.

[54] John Aitchison. "The statistical analysis of compositional data". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 44.2 (1982), pp. 139–160.

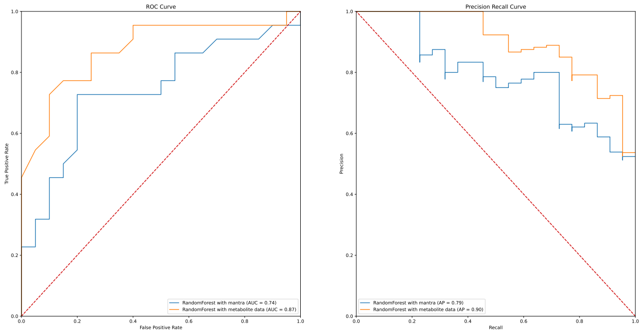
Supplementary Figures



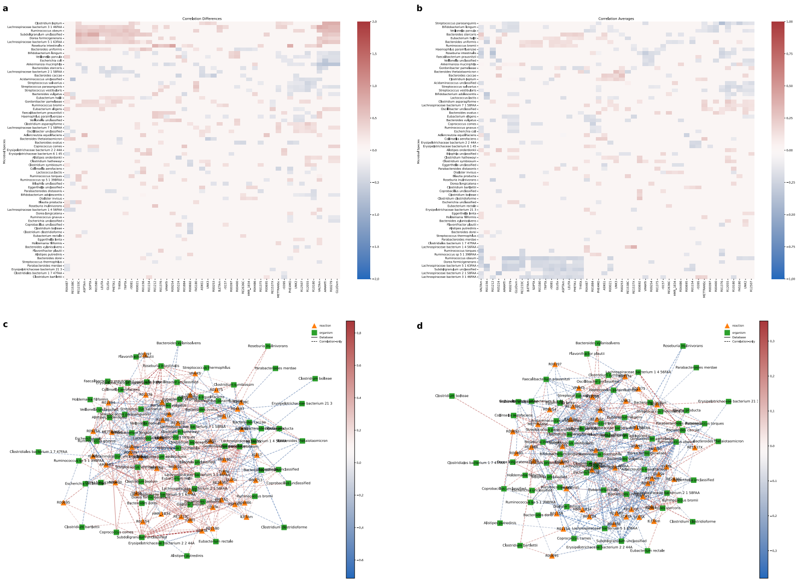
Supplementary Figure S1: Node degrees of the metabolic network based on the data from Xiao et al. [23] **a** Metabolic network presented in Figure 2a showing the node degree by color. Node shapes indicate whether a node is a metabolite or a metabolic reaction. **b** Distributions of node degrees from a for metabolite and reaction nodes separately.


metabolite
reaction
PRODUCT
SUBSTRATE
gdp-l-fucose
HMR_0863
3-Galactosyl-N-Acetylglucosaminide 4-Alpha-L-Fucosyltransferase
guanosine diphosphate
adenosine 3',5'-diphosphate
3', 5'-Bisphosphate Nucleotidase
Nucleoside-Diphosphate Kinase (ATP:GDP)
AMP aminohydrolase
Adenylate Kinase
adenosine monophosphate
adp
Uridine triphosphate:AMP phosphotransferase
ADP:D-glucose 6-phosphotransferase
Uridylate kinase (UMP)
uridine 5'-diphosphate
UDP phosphohydrolase
UDP-Glucuronosyltransferase 1-10 Precursor, Microsomal
None
UDPglucose:5-(D-galactosyloxy)-L-lysine-procollagen D-glucosyltransferase
uridine 5'-monophosphate
uridine diphosphate glucuronic aci
uridine diphosphate-n-acetylglucosamine
UDP-GlcNAc:Dolichol-Phosphate GlcNac Phosphotransferase (Liver)
UDP-glucose glucophosphohydrolase
uridine diphosphate glucose
UDP-glucose:NAD+ 6-oxidoreductase

Supplementary Figure S2: Induced subgraph using the reactions identified as significantly altered in activity in Figure 2d.



Supplementary Figure S3: Receiver Operating Characteristic (ROC) and Precision Recall (PR) curves for the prediction of CD labels from the validation cohort of Franzosa et al. [30] on the processed metabolome data (orange) and the *mantra* estimates (blue).



Supplementary Figure S4: Significant correlations between metabolic reactions and microbial species in the PRISM cohort [30]. **a** Correlation matrix between metabolic reactions and microbial species showing the differences between non-IBD and CD. **b** Correlation matrix between metabolic reactions and microbial species showing the averages of non-IBD and CD. **c** Correlation network from the correlation differences in a. **d** Correlation network from the correlation averages in b.

Supplemetary Tables

Label	Name	Degree
ahcys	s-adenosylhomocysteine	13
gdp	guanosine diphosphate	12
adp	adp	50
pyr	pyruvic acid	15
udp	uridine 5'-diphosphate	14
amp	adenosine monophosphate	40
amet	s-adenosylmethionine	15
gly	glycine	13
asp_L	l-aspartic acid	12

Supplementary Table S1: Node degrees of metabolites with a degree of at least 10 from the Triple-Negative Breast Cancer data set ([23])

Short Name	Formula	Description
URIDK1	atp[c] + ump[c] -> adp[c] + udp[c]	Uridylate kinase (UMP)
BPNT	h2o[c] + pap[c] -> amp[c] + pi[c]	3', 5'-Bisphosphate Nucl
ADK1	amp[c] + atp[c] <=> 2.0 adp[c]	Adenylate Kinase
UTPATPT	amp[c] + utp[c] <=> adp[c] + udp[c]	Uridine triphosphate:AM
R00155	UDP + H2O <=> UMP + Orthophosphate	UDP phosphohydrolase
R00287	UDP-glucose + H2O <=> UMP + D-Glucose 1-phosphate	UDP-glucose glucophosp
NDPK1	atp[c] + gdp[c] <=> adp[c] + gtp[c]	Nucleoside-Diphosphate
GMPR	gmp[c] + 2.0 h[c] + nadph[c] -> imp[c] + nadp[c...	GMP Reductase
R00181	AMP + H2O <=> IMP + Ammonia	AMP aminohydrolase
R08615	UDP-glucuronate + H2O <=> UDP + D-Glucuronate	UDP-glucuronate glucur
R00414	UDP-N-acetyl-alpha-D-glucosamine + H2O <=> N-Ac...	UDP-N-acetyl-D-glucosa
R00589	L-Serine <=> D-Serine	serine racemase
HMR_0873	fucgalfucgalacglcgalgluside_hs[c] + gdp[c] + h[...	3-Galactosyl-N-Acetylgl
R05327	n UDP-N-acetyl-alpha-D-glucosamine + n UDP-gluc...	None
R00286	UDP-glucose + H2O + 2 NAD+ <=> UDP-glucuronate ...	UDP-glucose:NAD+ 6-o
HMR_0863	galacglcgalgluside_hs[c] + gdpfuc[c] -> gdp[c] ...	HMR_0863
HMR_0859	galgluside_hs[c] + uacgam[c] -> acglcgalgluside...	Lactosylceramide 1, 3-N
R04491	UDP-glucose + 5-(D-Galactosyloxy)-L-lysine-proc...	UDPglucose:5-(D-galact
GLCNACPT_L	0.1 dolp_L[c] + uacgam[c] -> 0.1 naglc2p_L[c] +...	UDP-GlcNac:Dolichol-P
UGT1A1r	estrone[r] + udpglcur[r] -> estroneglc[r] + udp...	UDP-Glucuronosyltransf
R05804	ADP + D-Glucose <=> AMP + D-Glucose 6-phosphate	ADP:D-glucose 6-phosp

Supplementary Table S2: Detailed descriptions of the significant reactions from F
The table is also available as a .csv file in the additional supplements.

Sub-graph reaction	Matched EC Number	q-value
Glutamate syntase	1.4.1.14	0.00497
Tyrosine Transaminase	2.6.1.57	0.00678
tryptophan transaminase	2.6.1.57	0.00678
methionine:indole-3-pyruvic acid aminotransferase	2.6.1.88	0.0126
Leucine Transaminase	2.6.1.42	0.0545

Supplementary Table S3: Comparing the reactions from the network enrichment results (Figure 2a) to metagenomics-based functional profiling. The table shows the name of the reaction displayed in the original subgraph plot together with the respective EC number and its respective corrected p-value (q-value) in Supplementary Dataset 7 in Franzosa et al. [30].

List of Figures

Glossary

accurate mass search	Identifying compound candidates by searching a measured m/z against reference databases
molecular lipid species	A complex lipid for which the lipid class and the fatty acyls are known, but the positions of fatty acyls are unknown
precursor ion	The total intact analyte ion, usually subjected to fragmentation
shotgun mass spectrometry	Direct infusion of sample extracts into the mass spectrometer without pre-separation
sn-position	Position of the hydroxy group of the glycerol backbone of a lipid to which a fatty acid is bound
sum lipid species	A complex lipid for which only the lipid class and the total number of C atoms, double bonds, and hydroxy groups are known

Abbreviations

AA	Amino Acid
ADP	Adenosine Diphosphate
ANN	Artificial Neural Network
ATP	Adenosine Triphosphate
CCM	Central Carbon Metabolism
CCS	Collisional Cross-Section
CD	Crohn's Disease
CDP-DAG	cytidine diphospho-DAG
Cer	Ceramide
CID	Collision-Induced Dissociation
CL	Cardiolipin
CRC	Colorectal Cancer
CV	Cross-Validation
CVD	Cardio Vascular Diseases
DAG	Directed Acyclic Graph
DB	Double Bond
DG	Diacylglycerol
DL	Deep Learning
DNA	Deoxyribonucleic Acid
EAD	Electron-Activated Dissociation
EIEIO	Electron Impact Excitation of Ions from Organics
ESI	Electrospray Ionization
FA	Fatty Acid
$FADH_2$	Flavin Adenine Dinucleotide
FBA	Flux Balance Analysis
G3P	Glycerol-3-Phosphate
GAP	Glyceraldehyde-3-Phosphate
GC	Gas Chromatography
GDL	Geometric Deep Learning
GO	Gene Ontology
GRN	Gene Regulatory Network
GTP	Guanosine Triphosphate
HCC	Hepatocellular Carcinoma
HILIC	Hydrophilic Interaction Chromatography
IBD	Inflammatory Bowel Disease
IMS	Ion Mobility Spectrometry
LC	Liquid Chromatography
LINEX	Lipid Network Explorer
logFC	log-Fold Change
LPA	Lyso-Phosphatidic Acid
LPC	Lyso-Phosphatidylcholine
LPE	Lyso-Phosphatidylethanolamine
LPG	Lyso-Phosphatidylglycerol
LPI	Lyso-Phosphatidylinositol

LPS	Lyso-Phosphatidylserine
m/z	mass-to-charge ratio
mantra	Metabolic Network Reaction Analysis
MetS	Metabolic Syndrome
MG	Monoacylglycerol
MGS	Metagenomic Sequencing
ML	Machine Learning
MLE	Maximum Likelihood Estimation
MLP	Multilayer Perceptron
mRNA	messenger RNA
MS	Mass Spectrometry
MS/MS	Tandem Mass Spectrometry
MSE	Mean Squared Error
NADH	Nicotinamide Adenine Dinucleotide
NAFLD	Non-Alcoholic Fatty Liver Disease
NASH	Non-Alcoholic Steatohepatitis
NDE	Neural Differential Equation
NGS	Next Generation Sequencing
NLP	Natural Language Processig
NRM	Nuclear Magnetic Resonance
OLS	Ordinary Least Squares
ORC	Ollivier-Ricci Curvature
OTU	Operational Taxonomic Unit
OxPhos	Oxidative Phosphorylation
PA	Phosphatidic Acid
PC	Phosphatidylcholine
PCA	Principal Component Analysis
PE	Phosphatidylethanolamine
PG	Phosphatidylglycerol
PI	Phosphatidylinositol
PPI network	Protein-Protein Interaction network
PPP	Pentose Phosphate Pathway
PR	Precision Recall
PS	Phosphatidylserine
RNA	Ribonucleic Acid
RNAi	RNA interference
RNAseq	RNA sequencing
ROC	Receiver Operating Characteristic
ROC-AUC	Area under the Receiver Operating Characteristic Curve
RP	Reversed Phase Chromatography
rRNA	ribosomal RNA
RSS	Residual Sum of Squares
RT	Retention Time
siRNA	small interfering RNA

SM	Sphingomyelin
T2D	Type 2 Diabetes
TCA cycle	Tricarboxylic Acid Cycle
TG	Triacylglycerol
TOF	Time-of-Flight
tRNA	transfer RNA
UC	Ulcerative Colitis
UDP-GlcN	Uridine Diphosphate-Glucoronate

Bibliography

1. United Nations, Department of Economic and Social Affairs, Population Division. *Unicef Data Warehouse*. 2022. URL: https://population.un.org/wpp/ (visited on 11/16/2023).

2. S. H. Woolf. "Falling behind: the growing gap in life expectancy between the United States and other countries, 1933–2021". *American Journal of Public Health* 0, 2023, e1–e11.

3. D. Jasilionis, A. A. van Raalte, S. Klüsener, and P. Grigoriev. "The underwhelming German life expectancy". *European journal of epidemiology*, 2023, pp. 1–12.

4. W. H. Organization et al. "Invisible numbers: The true extent of noncommunicable diseases and what to do about them", 2022.

5. A. Weber, K. Kroiss, L. Reismann, P. Jansen, G. Hirschfelder, A. M. Sedlmeier, M. J. Stein, P. Bohmann, M. F. Leitzmann, and C. Jochem. "Health-Promoting and Sustainable Behavior in University Students in Germany: A Cross-Sectional Study". *International Journal of Environmental Research and Public Health* 20:7, 2023, p. 5238.

6. N. W. Chew, C. H. Ng, D. J. H. Tan, G. Kong, C. Lin, Y. H. Chin, W. H. Lim, D. Q. Huang, J. Quek, C. E. Fu, et al. "The global burden of metabolic disease: Data from 2000 to 2019". *Cell Metabolism* 35:3, 2023, pp. 414–428.

7. C. Nogales, Z. M. Mamdouh, M. List, C. Kiel, A. I. Casas, and H. H. Schmidt. "Network pharmacology: curing causal mechanisms instead of treating symptoms". *Trends in Pharmacological Sciences* 43:2, 2022, pp. 136–150.

8. M. Y. Lee and T. Hu. "Computational methods for the discovery of metabolic markers of complex traits". *Metabolites* 9:4, 2019, p. 66.

9. N. Köhler†, T. D. Rose†, L. Falk, and J. K. Pauling. "Investigating global lipidome alterations with the lipid network explorer". *Metabolites* 11:8, 2021, p. 488. DOI: 10.1093/bib/bbac572.

10. T. D. Rose†, N. Köhler†, L. Falk, L. Klischat, O. E. Lazareva, and J. K. Pauling. "Lipid network and moiety analysis for revealing enzymatic dysregulation and mechanistic alterations from lipidomics data". *Briefings in Bioinformatics* 24:1, 2023, bbac572. DOI: 10.3390/metabo11080488.

11. N. Köhler, V. Würf, T. D. Rose, and J. K. Pauling. "Identification and Integration of Key-Metabolic Reactions from Untargeted Metabolomics Data". *bioRxiv*, 2023, pp. 2023–05. DOI: 10.1101/2023.05.15.540613.

12. F. Crick. "Central dogma of molecular biology". *Nature* 227:5258, 1970, pp. 561–563. DOI: 10.1038/227561a0.

13. E. V. Koonin. "Does the central dogma still stand?" *Biology direct* 7, 2012, pp. 1–7. DOI: 10.1186/1745-6150-7-27.

14. C. L. Tan and E. Anderson. “The New Central Dogma of Molecular Biology”. *Resonance* 14:3, 2020, pp. 1–32.

15. S. L. Schreiber. “Small molecules: the missing link in the central dogma”. *Nature chemical biology* 1:2, 2005, pp. 64–66. DOI: 10.1038/nchembio0705-64.

16. P. E. Saw, X. Xu, J. Chen, and E.-W. Song. “Non-coding RNAs: The new central dogma of cancer biology”. *Science China Life Sciences* 64, 2021, pp. 22–50. DOI: 10.1007/s11427-020-1700-9.

17. G. Costa dos Santos, M. Renovato-Martins, and N. M. de Brito. “The remodel of the “central dogma”: A metabolomics interaction perspective”. *Metabolomics* 17, 2021, pp. 1–15. DOI: 10.1007/s11306-021-01800-8.

18. P. A. Wade. “Methyl CpG-binding proteins and transcriptional repression”. *Bioessays* 23:12, 2001, pp. 1131–1137. DOI: 10.1002/bies.10008.

19. M. M. Suzuki and A. Bird. “DNA methylation landscapes: provocative insights from epigenomics”. *Nature reviews genetics* 9:6, 2008, pp. 465–476. DOI: 10.1038/nrg2341.

20. B. Jin, Y. Li, and K. D. Robertson. “DNA methylation: superior or subordinate in the epigenetic hierarchy?” *Genes & cancer* 2:6, 2011, pp. 607–617. DOI: 10.1177/1947601910393957.

21. S. Hur. “Double-stranded RNA sensors and modulators in innate immunity”. *Annual review of immunology* 37, 2019, pp. 349–375. DOI: 10.1146/annurev-immunol-042718-041356.

22. T. Lindahl and B. Nyberg. “Heat-induced deamination of cytosine residues in deoxyribonucleic acid”. *Biochemistry* 13:16, 1974, pp. 3405–3410. DOI: 10.1021/bi00713a035.

23. A. Fire, S. Xu, M. K. Montgomery, S. A. Kostas, S. E. Driver, and C. C. Mello. “Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans”. *nature* 391:6669, 1998, pp. 806–811. DOI: 10.1038/35888.

24. C. Li, Y. Liang, and Y. Qiao. “Messengers from the gut: Gut microbiota-derived metabolites on host regulation”. *Frontiers in Microbiology* 13, 2022, p. 1339. DOI: 10.3389/fmicb.2022.863407.

25. N. Sudo, Y. Chida, Y. Aiba, J. Sonoda, N. Oyama, X.-N. Yu, C. Kubo, and Y. Koga. “Postnatal microbial colonization programs the hypothalamic–pituitary–adrenal system for stress response in mice”. *The Journal of physiology* 558:1, 2004, pp. 263–275. DOI: 10.1113/jphysiol.2004.063388.

26. J. Appleton. “The gut-brain axis: Influence of microbiota on mood and mental health”. *Integrative Medicine: A Clinician’s Journal* 17:4, 2018, p. 28.

27. D. Zheng, T. Liwinski, and E. Elinav. “Interaction between microbiota and immunity in health and disease”. *Cell research* 30:6, 2020, pp. 492–506. DOI: 10.1038/s41422-020-0332-7.

28. R. J. Cousins. “Nutritional regulation of gene expression”. *The American journal of medicine* 106:1, 1999, pp. 20–23. DOI: 10.1016/S0002-9343(98)00342-8.

29. D. Haro, P. F. Marrero, and J. Relat. “Nutritional regulation of gene expression: Carbohydrate-, fat-and amino acid-dependent modulation of transcriptional activity”. *International journal of molecular sciences* 20:6, 2019, p. 1386. DOI: 10.3390/ijms20061386.

30. B. T. Heijmans, E. W. Tobi, A. D. Stein, H. Putter, G. J. Blauw, E. S. Susser, P. E. Slagboom, and L. Lumey. “Persistent epigenetic differences associated with prenatal exposure to famine in humans”. *Proceedings of the National Academy of Sciences* 105:44, 2008, pp. 17046–17049. DOI: 10.1073/pnas.0806560105.

31. O. Rechavi, L. Houri-Ze’evi, S. Anava, W. S. S. Goh, S. Y. Kerk, G. J. Hannon, and O. Hobert. “Starvation-induced transgenerational inheritance of small RNAs in C. elegans”. *Cell* 158:2, 2014, pp. 277–287. DOI: 10.1016/j.cell.2014.06.020.

32. D. Hanahan and R. A. Weinberg. “Hallmarks of cancer: the next generation”. *cell* 144:5, 2011, pp. 646–674.

33. A. Stincone, A. Prigione, T. Cramer, M. M. Wamelink, K. Campbell, E. Cheung, V. Olin-Sandoval, N.-M. Grüning, A. Krüger, M. Tauqeer Alam, et al. “The return of metabolism: biochemistry and physiology of the pentose phosphate pathway”. *Biological Reviews* 90:3, 2015, pp. 927–963.

34. M. Inigo, S. Deja, and S. C. Burgess. “Ins and outs of the TCA cycle: the central role of anaplerosis”. *Annual review of nutrition* 41, 2021, pp. 19–47.

35. D. H. Wasserman. “Four grams of glucose”. *American Journal of Physiology-Endocrinology and Metabolism* 296:1, 2009, E11–E21.

36. R. J. DeBerardinis and N. S. Chandel. “Fundamentals of cancer metabolism”. *Science advances* 2:5, 2016, e1600200.

37. A. Anandhan, M. S. Jacome, S. Lei, P. Hernandez-Franco, A. Pappa, M. I. Panayiotidis, R. Powers, and R. Franco. “Metabolic dysfunction in Parkinson’s disease: bioenergetics, redox homeostasis and central carbon metabolism”. *Brain Research Bulletin* 133, 2017, pp. 12–30.

38. S. Krishnan, H. Nordqvist, A. T. Ambikan, S. Gupta, M. Sperk, S. Svensson-Akusjärvi, F. Mikaeloff, R. Benfeitas, E. Saccon, S. M. Ponnan, et al. “Metabolic perturbation associated with COVID-19 disease severity and SARS-CoV-2 replication”. *Molecular & Cellular Proteomics* 20, 2021.

39. O. Warburg, K. Posener, and E. Negelein. “Über den stoffwechsel der carcinomzelle”. *Naturwissenschaften* 12:50, 1924, pp. 1131–1137.

40. O. Warburg, F. Wind, and E. Negelein. “The metabolism of tumors in the body”. *The Journal of general physiology* 8:6, 1927, p. 519.

41. M. G. Vander Heiden, L. C. Cantley, and C. B. Thompson. “Understanding the Warburg effect: the metabolic requirements of cell proliferation”. *science* 324:5930, 2009, pp. 1029–1033.

42. N. N. Pavlova and C. B. Thompson. “The emerging hallmarks of cancer metabolism”. *Cell metabolism* 23:1, 2016, pp. 27–47.

43. A. D. Richardson, C. Yang, A. Osterman, and J. W. Smith. "Central carbon metabolism in the progression of mammary carcinoma". *Breast cancer research and treatment* 110, 2008, pp. 297–307.

44. D. Vigetti, E. Karousou, M. Viola, S. Deleonibus, G. De Luca, and A. Passi. "Hyaluronan: biosynthesis and signaling". *Biochimica et Biophysica Acta (BBA)-General Subjects* 1840:8, 2014, pp. 2452–2459.

45. S. J. Wakil. "Mechanism of fatty acid synthesis". *Journal of Lipid Research* 2:1, 1961, pp. 1–24.

46. B. Jenkins, J. A. West, and A. Koulman. "A review of odd-chain fatty acid metabolism and the role of pentadecanoic acid (C15: 0) and heptadecanoic acid (C17: 0) in health and disease". *Molecules* 20:2, 2015, pp. 2425–2444.

47. C. Mao, P. Xiao, X.-N. Tao, J. Qin, Q.-T. He, C. Zhang, S.-C. Guo, Y.-Q. Du, L.-N. Chen, D.-D. Shen, et al. "Unsaturated bond recognition leads to biased signal in a fatty acid receptor". *Science* 380:6640, 2023, eadd6220.

48. K. Wang, C.-W. Lee, X. Sui, S. Kim, S. Wang, A. B. Higgs, A. J. Baublis, G. A. Voth, M. Liao, T. C. Walther, et al. "The structure of phosphatidylinositol remodeling MBOAT7 reveals its catalytic mechanism and enables inhibitor identification". *Nature Communications* 14:1, 2023, p. 3533.

49. M. Kanehisa, Y. Sato, M. Kawashima, M. Furumichi, and M. Tanabe. "KEGG as a reference resource for gene and protein annotation". *Nucleic acids research* 44:D1, 2016, pp. D457–D462.

50. N. Kaur, V. Chugh, and A. K. Gupta. "Essential fatty acids as functional components of foods-a review". *Journal of food science and technology* 51, 2014, pp. 2289–2303.

51. S. M. Houten, S. Violante, F. V. Ventura, and R. J. Wanders. "The biochemistry and physiology of mitochondrial fatty acid β-oxidation and its genetic disorders". *Annual review of physiology* 78, 2016, pp. 23–44.

52. Y. Zhang, Y. Kuang, J. C. LaManna, and M. A. Puchowicz. "Contribution of brain glucose and ketone bodies to oxidative metabolism". In: *Oxygen Transport to Tissue XXXIV*. Springer. 2013, pp. 365–370.

53. M. Evans, K. E. Cogan, and B. Egan. "Metabolism of ketone bodies during exercise and training: physiological basis for exogenous supplementation". *The Journal of physiology* 595:9, 2017, pp. 2857–2871.

54. R. A. Coleman and D. P. Lee. "Enzymes of triacylglycerol synthesis and their regulation". *Progress in lipid research* 43:2, 2004, pp. 134–176.

55. H. Wang, M. V. Airola, and K. Reue. "How lipid droplets "TAG" along: Glycerolipid synthetic enzymes and lipid storage". *Biochimica et Biophysica Acta (BBA)-Molecular and Cell Biology of Lipids* 1862:10, 2017, pp. 1131–1145.

56. X. Pan and M. M. Hussain. "Gut triglyceride production". *Biochimica et Biophysica Acta (BBA)-Molecular and Cell Biology of Lipids* 1821:5, 2012, pp. 727–735.

57. C.-L. E. Yen, D. W. Nelson, and M.-I. Yen. “Intestinal triacylglycerol synthesis in fat absorption and systemic energy metabolism”. *Journal of lipid research* 56:3, 2015, pp. 489–501.

58. K. Reue and H. Wang. “Mammalian lipin phosphatidic acid phosphatases in lipid synthesis and beyond: metabolic and inflammatory disorders”. *Journal of lipid research* 60:4, 2019, pp. 728–733.

59. K. Athenstaedt and G. Daum. “The life cycle of neutral lipids: synthesis, storage and degradation”. *Cellular and Molecular Life Sciences CMLS* 63, 2006, pp. 1355–1369.

60. A. L. Henneberry, G. Wistow, and C. R. McMaster. “Cloning, genomic organization, and characterization of a human cholinephosphotransferase”. *Journal of Biological Chemistry* 275:38, 2000, pp. 29808–29815.

61. M. M. Wright and C. R. McMaster. “PC and PE synthesis: mixed micellar analysis of the cholinephosphotransferase and ethanolaminephosphotransferase activities of human choline/ethanolamine phosphotransferase 1 (CEPT1)”. *Lipids* 37:7, 2002, pp. 663–672.

62. A. L. Henneberry, M. M. Wright, and C. R. McMaster. “The major sites of cellular phospholipid synthesis and molecular determinants of fatty acid and lipid head group specificity”. *Molecular biology of the cell* 13:9, 2002, pp. 3148–3161.

63. K. Saito, M. Nishijima, and O. Kuge. “Genetic evidence that phosphatidylserine synthase II catalyzes the conversion of phosphatidylethanolamine to phosphatidylserine in Chinese hamster ovary cells”. *Journal of Biological Chemistry* 273:27, 1998, pp. 17199–17205.

64. S. Tomohiro, A. Kawaguti, Y. Kawabe, S. Kitada, and O. Kuge. “Purification and characterization of human phosphatidylserine synthases 1 and 2”. *Biochemical Journal* 418:2, 2009, pp. 421–429.

65. S. A. Henry, S. D. Kohlwein, and G. M. Carman. “Metabolism and regulation of glycerolipids in the yeast Saccharomyces cerevisiae”. *Genetics* 190:2, 2012, pp. 317–349.

66. Y. Yang, M. Lee, and G. D. Fairn. “Phospholipid subcellular localization and dynamics”. *Journal of Biological Chemistry* 293:17, 2018, pp. 6230–6240.

67. M. M. Nagiec, J. A. Baltisberger, G. B. Wells, R. L. Lester, and R. C. Dickson. “The LCB2 gene of Saccharomyces and the related LCB1 gene encode subunits of serine palmitoyltransferase, the initial enzyme in sphingolipid synthesis.” *Proceedings of the National Academy of Sciences* 91:17, 1994, pp. 7899–7902.

68. K. Hanada, M. Nishijima, M. Kiso, A. Hasegawa, S. Fujita, T. Ogawa, and Y. Akamatsu. “Sphingolipids are essential for the growth of Chinese hamster ovary cells. Restoration of the growth of a mutant defective in sphingoid base biosynthesis by exogenous sphingolipids.” *Journal of Biological Chemistry* 267:33, 1992, pp. 23527–23533.

69. A. H. Merrill. “De novo sphingolipid biosynthesis: a necessary, but dangerous, pathway”. *Journal of Biological Chemistry* 277:29, 2002, pp. 25843–25846.

70. J. Rother, G. van Echten, G. Schwarzmann, and K. Sandhoff. “Biosynthesis of sphingolipids: dihydroceramide and not sphinganine is desaturated by cultured cells”. *Biochemical and biophysical research communications* 189:1, 1992, pp. 14–20.

71. M. Levy and A. H. Futerman. “Mammalian ceramide synthases”. *IUBMB life* 62:5, 2010, pp. 347–356.

72. G. Liebisch, E. Fahy, J. Aoki, E. A. Dennis, T. Durand, C. S. Ejsing, M. Fedorova, I. Feussner, W. J. Griffiths, H. Köfeler, et al. “Update on LIPID MAPS classification, nomenclature, and shorthand notation for MS-derived lipid structures”. *Journal of lipid research* 61:12, 2020, pp. 1539–1555.

73. J. K. Pauling, M. Hermansson, J. Hartler, K. Christiansen, S. F. Gallego, B. Peng, R. Ahrends, and C. S. Ejsing. “Proposal for a common nomenclature for fragment ions in mass spectra of lipids”. *PLoS One* 12:11, 2017, e0188394.

74. V. Natesan and S.-J. Kim. “Lipid metabolism, disorders and therapeutic drugs–review”. *Biomolecules & therapeutics* 29:6, 2021, p. 596.

75. C. R. Santos and A. Schulze. “Lipid metabolism in cancer”. *The FEBS journal* 279:15, 2012, pp. 2610–2623.

76. E. Furuta, S. K. Pai, R. Zhan, S. Bandyopadhyay, M. Watabe, Y.-Y. Mo, S. Hirota, S. Hosobe, T. Tsukada, K. Miura, et al. “Fatty acid synthase gene is up-regulated by hypoxia via activation of Akt and sterol regulatory element binding protein-1”. *Cancer research* 68:4, 2008, pp. 1003–1011.

77. Y. Yoshii, T. Furukawa, H. Yoshii, T. Mori, Y. Kiyono, A. Waki, M. Kobayashi, T. Tsujikawa, T. Kudo, H. Okazawa, et al. “Cytosolic acetyl-CoA synthetase affected tumor cell survival under hypoxia: the possible function in tumor acetyl-CoA/acetate metabolism”. *Cancer science* 100:5, 2009, pp. 821–827.

78. V. T. Samuel and G. I. Shulman. “Mechanisms for insulin resistance: common threads and missing links”. *Cell* 148:5, 2012, pp. 852–871.

79. S. A. Rosenzweig and H. S. Atreya. “Defining the pathway to insulin-like growth factor system targeting in cancer”. *Biochemical pharmacology* 80:8, 2010, pp. 1115–1124.

80. H. Rosen. “Is obesity a disease or a behavior abnormality? Did the AMA get it right?” *Missouri medicine* 111:2, 2014, p. 104.

81. C. Weyer, J. Foley, C. Bogardus, P. Tataranni, and R. Pratley. “Enlarged subcutaneous abdominal adipocyte size, but not obesity itself, predicts type II diabetes independent of insulin resistance”. *Diabetologia* 43, 2000, pp. 1498–1506.

82. E. Guiu-Jurado, T. Auguet, A. Berlanga, G. Aragonès, C. Aguilar, F. Sabench, S. Armengol, J. A. Porras, A. Martí, R. Jorba, et al. “Downregulation of de novo fatty acid synthesis in subcutaneous adipose tissue of moderately obese women”. *International Journal of Molecular Sciences* 16:12, 2015, pp. 29911–29922.

83. F. J. Ortega, D. Mayas, J. M. Moreno-Navarrete, V. Catalán, J. Gómez-Ambrosi, E. Esteve, J. I. Rodriguez-Hermosa, B. Ruiz, W. Ricart, B. Peral, et al. “The gene expression of the main lipogenic enzymes is downregulated in visceral adipose tissue of obese subjects”. *Obesity* 18:1, 2010, pp. 13–20.

84. O. Poulain-Godefroy, C. Lecoeur, F. Pattou, G. Fruhbeck, and P. Froguel. “Inflammation is associated with a decrease of lipogenic factors in omental fat in women”. *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology* 295:1, 2008, R1–R7.

85. N. S. Chandel. “Amino acid metabolism”. *Cold Spring Harbor Perspectives in Biology* 13:4, 2021, a040584.

86. R. Lin, W. Liu, M. Piao, and H. Zhu. “A review of the relationship between the gut microbiota and amino acid metabolism”. *Amino acids* 49, 2017, pp. 2083–2090.

87. J. M. Berg, J. L. Tymoczko, G. J. Gatto jr., and L. Stryer. “Biosynthesis of Amino Acids”. In: *Stryer Biochemistry*. 8th ed. WH Freeman, 2015. Chap. 24. ISBN: 978-1464126109.

88. T. J. de Koning. “Amino acid synthesis deficiencies”. *Journal of inherited metabolic disease* 40:4, 2017, pp. 609–620.

89. J. T. Brosnan. “Glutamate, at the interface between amino acid and carbohydrate metabolism”. *The Journal of nutrition* 130:4, 2000, 988S–990S.

90. J. M. Berg, J. L. Tymoczko, G. J. Gatto jr., and L. Stryer. “Protein Turnover and Amino Acid Catabolism”. In: *Stryer Biochemistry*. 8th ed. WH Freeman, 2015. Chap. 23. ISBN: 978-1464126109.

91. C. M. Metallo, P. A. Gameiro, E. L. Bell, K. R. Mattaini, J. Yang, K. Hiller, C. M. Jewell, Z. R. Johnson, D. J. Irvine, L. Guarente, et al. “Reductive glutamine metabolism by IDH1 mediates lipogenesis under hypoxia”. *Nature* 481:7381, 2012, pp. 380–384.

92. J. Zhang, J. Fan, S. Venneti, J. R. Cross, T. Takagi, B. Bhinder, H. Djaballah, M. Kanai, E. H. Cheng, A. R. Judkins, et al. “Asparagine plays a critical role in regulating cellular adaptation to glutamine depletion”. *Molecular cell* 56:2, 2014, pp. 205–218.

93. E. L. Lieu, T. Nguyen, S. Rhyne, and J. Kim. “Amino acids in cancer”. *Experimental & molecular medicine* 52:1, 2020, pp. 15–30.

94. K. Sugihara, T. L. Morhardt, and N. Kamada. “The role of dietary nutrients in inflammatory bowel disease”. *Frontiers in immunology* 9, 2019, p. 3183.

95. X. C. Morgan, T. L. Tickle, H. Sokol, D. Gevers, K. L. Devaney, D. V. Ward, J. A. Reyes, S. A. Shah, N. LeLeiko, S. B. Snapper, et al. “Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment”. *Genome biology* 13:9, 2012, pp. 1–18.

96. K. G. Upadhyay, D. C. Desai, T. F. Ashavaid, and A. J. Dherai. “Microbiome and metabolome in inflammatory bowel disease”. *Journal of Gastroenterology and Hepatology* 38:1, 2023, pp. 34–43.

97. M. Giera, O. Yanes, and G. Siuzdak. “Metabolite discovery: biochemistry’s scientific driver”. *Cell metabolism* 34:1, 2022, pp. 21–34.

98. A.-H. Emwas, R. Roy, R. T. McKay, L. Tenori, E. Saccenti, G. N. Gowda, D. Raftery, F. Alahmari, L. Jaremko, M. Jaremko, et al. “NMR spectroscopy for metabolomics research”. *Metabolites* 9:7, 2019, p. 123.

99. T. Züllig, M. Trötzmüller, and H. C. Köfeler. “Lipidomics from sample preparation to data analysis: a primer”. *Analytical and bioanalytical chemistry* 412, 2020, pp. 2191–2209.

100. M. Y. Mushtaq, Y. H. Choi, R. Verpoorte, and E. G. Wilson. "Extraction for metabolomics: access to the metabolome". *Phytochemical Analysis* 25:4, 2014, pp. 291–306.

101. A. Lindahl, S. Sääf, J. Lehtiö, and A. Nordström. "Tuning metabolome coverage in reversed phase LC–MS metabolomics of MeOH extracted samples using the reconstitution solvent composition". *Analytical chemistry* 89:14, 2017, pp. 7356–7364.

102. J. Folch, M. Lees, G. H. Sloane Stanley, et al. "A simple method for the isolation and purification of total lipids from animal tissues". *J biol Chem* 226:1, 1957, pp. 497–509.

103. E. G. Bligh and W. J. Dyer. "A rapid method of total lipid extraction and purification". *Canadian journal of biochemistry and physiology* 37:8, 1959, pp. 911–917.

104. V. Matyash, G. Liebisch, T. V. Kurzchalia, A. Shevchenko, and D. Schwudke. "Lipid extraction by methyl-tert-butyl ether for high-throughput lipidomics". *Journal of lipid research* 49:5, 2008, pp. 1137–1146.

105. R. Liu, J. Chou, S. Hou, X. Liu, J. Yu, X. Zhao, Y. Li, L. Liu, and C. Sun. "Evaluation of two-step liquid-liquid extraction protocol for untargeted metabolic profiling of serum samples to achieve broader metabolome coverage by UPLC-Q-TOF-MS". *Analytica Chimica Acta* 1035, 2018, pp. 96–107.

106. C. Z. Ulmer, C. M. Jones, R. A. Yost, T. J. Garrett, and J. A. Bowden. "Optimization of Folch, Bligh-Dyer, and Matyash sample-to-extraction solvent ratios for human plasma-based lipidomics studies". *Analytica chimica acta* 1037, 2018, pp. 351–357.

107. S. Zhao and L. Li. "Chemical derivatization in LC-MS-based metabolomics study". *TrAC Trends in Analytical Chemistry* 131, 2020, p. 115988.

108. C. J. Bolten, P. Kiefer, F. Letisse, J.-C. Portais, and C. Wittmann. "Sampling for metabolome analysis of microorganisms". *Analytical chemistry* 79:10, 2007, pp. 3843–3849.

109. J. B. Fenn, M. Mann, C. K. Meng, S. F. Wong, and C. M. Whitehouse. "Electrospray ionization for mass spectrometry of large biomolecules". *Science* 246:4926, 1989, pp. 64–71.

110. M. Wilm. "Principles of electrospray ionization". *Molecular & cellular proteomics* 10:7, 2011.

111. R. A. Zubarev and A. Makarov. *Orbitrap mass spectrometry*. 2013.

112. T. Opialla, S. Kempa, and M. Pietzke. "Towards a more reliable identification of isomeric metabolites using pattern guided retention validation". *Metabolites* 10:11, 2020, p. 457.

113. M. Sud, E. Fahy, D. Cotter, A. Brown, E. A. Dennis, C. K. Glass, A. H. Merrill Jr, R. C. Murphy, C. R. Raetz, D. W. Russell, and S. Shankar. "Lmsd: Lipid maps structure database". *Nucleic acids research* 35:suppl_1, 2007, pp. D527–D532. DOI: `10.1093/nar/gkl838`.

114. J. M. Wells and S. A. McLuckey. "Collision-induced dissociation (CID) of peptides and proteins". *Methods in enzymology* 402, 2005, pp. 148–185.

115. M. Gerlich and S. Neumann. "MetFusion: integration of compound identification strategies". *Journal of Mass Spectrometry* 48:3, 2013, pp. 291–298.

116. A. R. Johnson and E. E. Carlson. *Collision-induced dissociation mass spectrometry: a powerful tool for natural product structure elucidation*. 2015.

117. T. Baba, P. Ryumin, E. Duchoslav, K. Chen, A. Chelur, B. Loyd, and I. Chernushevich. "Dissociation of biomolecules by an intense low-energy electron beam in a high sensitivity time-of-flight mass spectrometer". *Journal of the American Society for Mass Spectrometry* 32:8, 2021, pp. 1964–1975.

118. X. Zhang, X. Ren, K. Chingin, J. Xu, X. Yan, and H. Chen. "Mass spectrometry distinguishing C= C location and cis/trans isomers: A strategy initiated by water radical cations". *Analytica Chimica Acta* 1139, 2020, pp. 146–154.

119. F.-F. Hsu. "Complete structural characterization of ceramides as [M- H]- ions by multiple-stage linear ion trap mass spectrometry". *Biochimie* 130, 2016, pp. 63–75.

120. R. King, R. Bonfiglio, C. Fernandez-Metzler, C. Miller-Stein, and T. Olah. "Mechanistic investigation of ionization suppression in electrospray ionization". *Journal of the American Society for Mass Spectrometry* 11:11, 2000, pp. 942–950.

121. T. M. Annesley. "Ion suppression in mass spectrometry". *Clinical chemistry* 49:7, 2003, pp. 1041–1044.

122. D.-Q. Tang, L. Zou, X.-X. Yin, and C. N. Ong. "HILIC-MS for metabolomics: An attractive and complementary approach to RPLC-MS". *Mass spectrometry reviews* 35:5, 2016, pp. 574–600.

123. B. Buszewski and S. Noga. "Hydrophilic interaction liquid chromatography (HILIC)—a powerful separation technique". *Analytical and bioanalytical chemistry* 402, 2012, pp. 231–247.

124. R. Cumeras, E. Figueras, C. Davis, J. I. Baumbach, and I. Gracia. "Review on ion mobility spectrometry. Part 1: current instrumentation". *Analyst* 140:5, 2015, pp. 1376–1390.

125. C. G. Vasilopoulou, K. Sulek, A.-D. Brunner, N. S. Meitei, U. Schweiger-Hufnagel, S. W. Meyer, A. Barsch, M. Mann, and F. Meier. "Trapped ion mobility spectrometry and PASEF enable in-depth lipidomics from minimal sample amounts". *Nature communications* 11:1, 2020, p. 331.

126. R. Cumeras, E. Figueras, C. Davis, J. I. Baumbach, and I. Gracia. "Review on ion mobility spectrometry. Part 2: hyphenated methods and effects of experimental parameters". *Analyst* 140:5, 2015, pp. 1391–1410.

127. T. P. Lintonen, P. R. Baker, M. Suoniemi, B. K. Ubhi, K. M. Koistinen, E. Duchoslav, J. L. Campbell, and K. Ekroos. "Differential mobility spectrometry-driven shotgun lipidomics". *Analytical chemistry* 86:19, 2014, pp. 9662–9669.

128. G. Paglia, P. Angel, J. P. Williams, K. Richardson, H. J. Olivos, J. W. Thompson, L. Menikarachchi, S. Lai, C. Walsh, A. Moseley, et al. "Ion mobility-derived collision cross section as an additional measure for lipid fingerprinting and identification". *Analytical chemistry* 87:2, 2015, pp. 1137–1144.

129. P. D. Rainville, I. D. Wilson, J. K. Nicholson, G. Isaac, L. Mullin, J. I. Langridge, and R. S. Plumb. "Ion mobility spectrometry combined with ultra performance liquid chromatography/mass spectrometry for metabolic phenotyping of urine: Effects of column length, gradient duration and ion mobility spectrometry on metabolite detection". *Analytica chimica acta* 982, 2017, pp. 1–8.

130. I. Blaženović, T. Shen, S. S. Mehta, T. Kind, J. Ji, M. Piparo, F. Cacciola, L. Mondello, and O. Fiehn. "Increasing compound identification rates in untargeted lipidomics research with liquid chromatography drift time–ion mobility mass spectrometry". *Analytical chemistry* 90:18, 2018, pp. 10758–10764.

131. R. Aalizadeh, M.-C. Nika, and N. S. Thomaidis. "Development and application of retention time prediction models in the suspect and non-target screening of emerging contaminants". *Journal of Hazardous materials* 363, 2019, pp. 277–285.

132. A. Celma, R. Bade, J. V. Sancho, F. Hernandez, M. Humphries, and L. Bijlsma. "Prediction of Retention Time and Collision Cross Section (CCSH+, CCSH−, and CCSNa+) of Emerging Contaminants Using Multiple Adaptive Regression Splines". *Journal of chemical information and modeling* 62:22, 2022, pp. 5425–5434.

133. H. Tsugawa, T. Cajka, T. Kind, Y. Ma, B. Higgins, K. Ikeda, M. Kanazawa, J. VanderGheynst, O. Fiehn, and M. Arita. "MS-DIAL: data-independent MS/MS deconvolution for comprehensive metabolome analysis". *Nature methods* 12:6, 2015, pp. 523–526.

134. H. L. Röst, T. Sachsenberg, S. Aiche, C. Bielow, H. Weisser, F. Aicheler, S. Andreotti, H.-C. Ehrlich, P. Gutenbrunner, E. Kenar, et al. "OpenMS: a flexible open-source software platform for mass spectrometry data analysis". *Nature methods* 13:9, 2016, pp. 741–748.

135. J. Xia and D. S. Wishart. "Web-based inference of biological patterns, functions and pathways from metabolomic data using MetaboAnalyst". *Nature protocols* 6:6, 2011, pp. 743–760.

136. R. Herzog, K. Schuhmann, D. Schwudke, J. L. Sampaio, S. R. Bornstein, M. Schroeder, and A. Shevchenko. "LipidXplorer: a software for consensual cross-platform lipidomics". *PloS one* 7:1, 2012, e29851.

137. J. P. Koelmel, N. M. Kroeger, C. Z. Ulmer, J. A. Bowden, R. E. Patterson, J. A. Cochran, C. W. Beecher, T. J. Garrett, and R. A. Yost. "LipidMatch: an automated workflow for rule-based lipid identification using untargeted high-resolution tandem mass spectrometry data". *BMC bioinformatics* 18, 2017, pp. 1–11.

138. Z. Ni, G. Angelidou, M. Lange, R. Hoffmann, and M. Fedorova. "LipidHunter identifies phospholipids by high-throughput processing of LC-MS and shotgun lipidomics datasets". *Analytical Chemistry* 89:17, 2017, pp. 8800–8807.

139. A. Thompson, J. Schäfer, K. Kuhn, S. Kienle, J. Schwarz, G. Schmidt, T. Neumann, and C. Hamon. "Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS". *Analytical chemistry* 75:8, 2003, pp. 1895–1904.

140. R. Wu and A. Kaiser. "Structure and base sequence in the cohesive ends of bacteriophage lambda DNA". *Journal of molecular biology* 35:3, 1968, pp. 523–537.

141. J. Shendure, S. Balasubramanian, G. M. Church, W. Gilbert, J. Rogers, J. A. Schloss, and R. H. Waterston. "DNA sequencing at 40: past, present and future". *Nature* 550:7676, 2017, pp. 345–353.

142. F. Sanger, S. Nicklen, and A. R. Coulson. "DNA sequencing with chain-terminating inhibitors". *Proceedings of the national academy of sciences* 74:12, 1977, pp. 5463–5467.

143. A. M. Maxam and W. Gilbert. "A new method for sequencing DNA." *Proceedings of the National Academy of Sciences* 74:2, 1977, pp. 560–564.

144. P. Nyrén, B. Pettersson, and M. Uhlén. "Solid phase DNA minisequencing by an enzymatic luminometric inorganic pyrophosphate detection assay". *Analytical biochemistry* 208:1, 1993, pp. 171–175.

145. S. Brenner, M. Johnson, J. Bridgham, G. Golda, D. H. Lloyd, D. Johnson, S. Luo, S. McCurdy, M. Foy, M. Ewan, et al. "Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays". *Nature biotechnology* 18:6, 2000, pp. 630–634.

146. I. Braslavsky, B. Hebert, E. Kartalov, and S. R. Quake. "Sequence information can be obtained from single DNA molecules". *Proceedings of the National Academy of Sciences* 100:7, 2003, pp. 3960–3964.

147. R. D. Mitra, J. Shendure, J. Olejnik, G. M. Church, et al. "Fluorescent in situ sequencing on polymerase colonies". *Analytical biochemistry* 320:1, 2003, pp. 55–65.

148. P. E. Warburton and R. P. Sebra. "Long-Read DNA Sequencing: Recent Advances and Remaining Challenges". *Annual Review of Genomics and Human Genetics* 24, 2023.

149. A. B. Shreiner, J. Y. Kao, and V. B. Young. "The gut microbiome in health and in disease". *Current opinion in gastroenterology* 31:1, 2015, p. 69.

150. J. Tremblay, K. Singh, A. Fern, E. S. Kirton, S. He, T. Woyke, J. Lee, F. Chen, J. L. Dangl, and S. G. Tringe. "Primer and platform effects on 16S rRNA tag sequencing". *Frontiers in microbiology* 6, 2015, p. 771.

151. N.-P. Nguyen, T. Warnow, M. Pop, and B. White. "A perspective on 16S rRNA operational taxonomic unit clustering using sequence similarity". *NPJ biofilms and microbiomes* 2:1, 2016, pp. 1–8.

152. I. Abellan-Schneyder, M. S. Matchado, S. Reitmeier, A. Sommer, Z. Sewald, J. Baumbach, M. List, and K. Neuhaus. "Primer, pipelines, parameters: issues in 16S rRNA gene sequencing". *Msphere* 6:1, 2021, pp. 10–1128.

153. N. Segata. "On the road to strain-resolved comparative metagenomics". *MSystems* 3:2, 2018, pp. 10–1128.

154. L. Chistoserdova. "Functional metagenomics: recent advances and future challenges". *Biotechnology and Genetic Engineering Reviews* 26:1, 2009, pp. 335–352.

155. O. I. Coleman†, A. Sorbie†, S. Bierwith, J. Kövilein, M. von Stern, N. Köhler, J. Wirbel, C. Schmidt, T. Kacprowski, A. Dunkel, et al. "ATF6 activation alters colonic lipid metabolism causing tumor-associated microbial adaptation". *bioRxiv*, 2023, pp. 2023–11.

156. D. Fenstermacher. “Introduction to bioinformatics”. *Journal of the American Society for Information Science and Technology* 56:5, 2005, pp. 440–446.

157. J. Gauthier, A. T. Vincent, S. J. Charette, and N. Derome. “A brief history of bioinformatics”. *Briefings in bioinformatics* 20:6, 2019, pp. 1981–1996.

158. U. D. J. G. I. H. T. 4. B. E. 4. P. P. 4. R. P. 4. W. S. 4. S. T. 4. D. N. 4. C. J.-F. 4. O. A. 4. L. S. 4. E. C. 4. U. E. 4. F. M 4, R. G. S. C. S. Y. 9. F. A. 9. H. M. 9. Y. T. 9. T. A. 9. I. T. 9. K. C. 9. W. H. 9. T. Y. 9. T. T. 9, Genoscope, C. U.-8. W. J. 1. H. R. 1. S. W. 1. A. F. 1. B. P. 1. B. T. 1. P. E. 1. R. C. 1. W. P. 10, I. o. M. B. R. A. 1. P. M. 1. N. G. 1. T. S. 1. R. A. 1. Department of Genome Analysis, G. S. C. S. D. R. 1. D.-S. L. 1. 11, B. G. I. G. C. Y. H. 1. Y. J. 1. W. J. 1. H. G. 1. G. J. 15, et al. “Initial sequencing and analysis of the human genome”. *nature* 409:6822, 2001, pp. 860–921.

159. J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, et al. “Highly accurate protein structure prediction with AlphaFold”. *Nature* 596:7873, 2021, pp. 583–589.

160. G. O. Consortium. “The Gene Ontology (GO) database and informatics resource”. *Nucleic acids research* 32:suppl_1, 2004, pp. D258–D261.

161. M. Balaban, N. Moshiri, U. Mai, X. Jia, and S. Mirarab. “TreeCluster: Clustering biological sequences using phylogenetic trees”. *PloS one* 14:8, 2019, e0221068.

162. L. Euler. “Solutio problematis ad geometriam situs pertinentis”. *Commentarii academiae scientiarum Petropolitanae*, 1741, pp. 128–140.

163. P. E. Compeau, P. A. Pevzner, and G. Tesler. “How to apply de Bruijn graphs to genome assembly”. *Nature biotechnology* 29:11, 2011, pp. 987–991.

164. L. Debnath. “A brief historical introduction to Euler’s formula for polyhedra, topology, graph theory and networks”. *International Journal of Mathematical Education in Science and Technology* 41:6, 2010, pp. 769–785.

165. A. Smith and V. M. Zavala. “The Euler characteristic: A general topological descriptor for complex data”. *Computers & Chemical Engineering* 154, 2021, p. 107463.

166. A. Smith, S. Runde, A. K. Chew, A. S. Kelkar, U. Maheshwari, R. C. Van Lehn, and V. M. Zavala. “Topological Analysis of Molecular Dynamics Simulations using the Euler Characteristic”. *Journal of Chemical Theory and Computation* 19:5, 2023, pp. 1553–1567.

167. K. V. Nadimpalli, A. Chattopadhyay, and B. Rieck. “Euler Characteristic Transform Based Topological Loss for Reconstructing 3D Images from Single 2D Slices”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 571–579.

168. D. R. Lanning, G. K. Harrell, and J. Wang. “Dijkstra’s algorithm and Google maps”. In: *Proceedings of the 2014 ACM Southeast Regional Conference*. 2014, pp. 1–3.

169. L. Page, S. Brin, R. Motwani, and T. Winograd. *The pagerank citation ranking: Bring order to the web*. Technical report. Technical report, stanford University, 1998.

170. R. Angles and C. Gutierrez. “Survey of graph database models”. *ACM Computing Surveys (CSUR)* 40:1, 2008, pp. 1–39.

171. Y.-K. Shih and S. Parthasarathy. "A single source k-shortest paths algorithm to infer regulatory pathways in a gene network". *Bioinformatics* 28:12, 2012, pp. i49–i58.

172. Forman. "Bochner's method for cell complexes and combinatorial Ricci curvature". *Discrete & Computational Geometry* 29, 2003, pp. 323–374.

173. Y. Ollivier. "Ricci curvature of Markov chains on metric spaces". *Journal of Functional Analysis* 256:3, 2009, pp. 810–864.

174. J. Topping, F. Di Giovanni, B. P. Chamberlain, X. Dong, and M. M. Bronstein. "Understanding over-squashing and bottlenecks on graphs via curvature". *arXiv preprint arXiv:2111.14522*, 2021.

175. D. F. Anderson and T. G. Kurtz. "Continuous time Markov chain models for chemical reaction networks". In: *Design and analysis of biomolecular circuits: engineering approaches to systems and synthetic biology*. Springer, 2011, pp. 3–42.

176. C. Coupette, S. Dalleiger, and B. Rieck. "Ollivier-Ricci Curvature for Hypergraphs: A Unified Framework". *arXiv preprint arXiv:2210.12048*, 2022.

177. C. Chen, C. Liao, and Y.-Y. Liu. "Teasing out missing reactions in genome-scale metabolic networks through hypergraph learning". *Nature Communications* 14:1, 2023, p. 2375.

178. C. Frainay, S. Aros, M. Chazalviel, T. Garcia, F. Vinson, N. Weiss, B. Colsch, F. Sedel, D. Thabut, C. Junot, et al. "MetaboRank: network-based recommendation system to interpret and enrich metabolomics results". *Bioinformatics* 35:2, 2019, pp. 274–283.

179. Z. Ding, W. Guo, and J. Gu. "ClustEx2: gene module identification using density-based network hierarchical clustering". In: *2018 Chinese Automation Congress (CAC)*. IEEE. 2018, pp. 2407–2412.

180. H. Ma, E. E. Schadt, L. M. Kaplan, and H. Zhao. "COSINE: COndition-SpecIfic sub-NEtwork identification using a global optimization method". *Bioinformatics* 27:9, 2011, pp. 1290–1298.

181. S. D. Ghiassian, J. Menche, and A.-L. Barabási. "A DIseAse MOdule Detection (DIAMOnD) algorithm derived from a systematic analysis of connectivity patterns of disease proteins in the human interactome". *PLoS computational biology* 11:4, 2015, e1004120.

182. H. Levi, R. Elkon, and R. Shamir. "DOMINO: a network-based active module identification algorithm with reduced rate of false calls". *Molecular systems biology* 17:1, 2021, e9593.

183. R. Breitling, A. Amtmann, and P. Herzyk. "Graph-based iterative Group Analysis enhances microarray interpretation". *BMC bioinformatics* 5, 2004, pp. 1–10.

184. Ş. Nacu, R. Critchley-Thorne, P. Lee, and S. Holmes. "Gene expression network analysis and applications to immunology". *Bioinformatics* 23:7, 2007, pp. 850–858.

185. M. A. Reyna, M. D. Leiserson, and B. J. Raphael. "Hierarchical HotNet: identifying hierarchies of altered subnetworks". *Bioinformatics* 34:17, 2018, pp. i972–i980.

186. G. Barel and R. Herwig. "NetCore: a network propagation approach using node coreness". *Nucleic acids research* 48:17, 2020, e98–e98.

187. S. Choobdar, M. E. Ahsen, J. Crawford, M. Tomasoni, T. Fang, D. Lamparter, J. Lin, B. Hescott, X. Hu, J. Mercer, et al. "Assessment of network module identification across complex diseases". *Nature methods* 16:9, 2019, pp. 843–852.

188. D. Li, Z. Pan, G. Hu, Z. Zhu, and S. He. "Active module identification in intracellular networks using a memetic algorithm with a new binary decoding scheme". *BMC genomics* 18:2, 2017, pp. 1–9.

189. T. Ideker, O. Ozier, B. Schwikowski, and A. F. Siegel. "Discovering regulatory and signalling circuits in molecular interaction networks". *Bioinformatics* 18:suppl_1, 2002, S233–S240.

190. N. Alcaraz, H. Kücük, J. Weile, A. Wipat, and J. Baumbach. "KeyPathwayMiner: detecting case-specific biological pathways using expression data". *Internet Mathematics* 7:4, 2011, pp. 299–313.

191. E. M. Novoa-del-Toro, E. Mezura-Montes, M. Vignes, M. Térézol, F. Magdinier, L. Tichit, and A. Baudot. "A multi-objective genetic algorithm to find active modules in multiplex biological networks". *PLoS computational biology* 17:8, 2021, e1009263.

192. O. C. Martin and S. W. Otto. "Combining simulated annealing with local search heuristics". *Annals of operations research* 63:1, 1996, pp. 57–75.

193. R. Batra, N. Alcaraz, K. Gitzhofer, J. Pauling, H. J. Ditzel, M. Hellmuth, J. Baumbach, and M. List. "On the performance of de novo pathway enrichment". *NPJ systems biology and applications* 3:1, 2017, p. 6.

194. O. Lazareva, J. Baumbach, M. List, and D. B. Blumenthal. "On the limits of active module identification". *Briefings in Bioinformatics* 22:5, 2021, bbab066.

195. K. B. Stibius and K. Sneppen. "Modeling the two-hybrid detector: experimental bias on protein interaction networks". *Biophysical Journal* 93:7, 2007, pp. 2562–2566.

196. M. H. Schaefer, L. Serrano, and M. A. Andrade-Navarro. "Correcting for the study bias associated with protein–protein interaction measurements reveals differences between protein degree distributions from different cancer types". *Frontiers in genetics* 6, 2015, p. 260.

197. N. C. Duarte, S. A. Becker, N. Jamshidi, I. Thiele, M. L. Mo, T. D. Vo, R. Srivas, and B. Ø. Palsson. "Global reconstruction of the human metabolic network based on genomic and bibliomic data". *Proceedings of the National Academy of Sciences* 104:6, 2007, pp. 1777–1782.

198. J. D. Orth, I. Thiele, and B. Ø. Palsson. "What is flux balance analysis?" *Nature biotechnology* 28:3, 2010, pp. 245–248.

199. D. Machado, R. S. Costa, E. C. Ferreira, I. Rocha, and B. Tidor. "Exploring the gap between dynamic and constraint-based models of metabolism". *Metabolic engineering* 14:2, 2012, pp. 112–119.

200. A. Rai and K. Saito. "Omics data input for metabolic modeling". *Current opinion in biotechnology* 37, 2016, pp. 127–134.

201. A. Amara, C. Frainay, F. Jourdan, T. Naake, S. Neumann, E. M. Novoa-del-Toro, R. M. Salek, L. Salzer, S. Scharfenberg, and M. Witting. “Networks and graphs discovery in metabolomics data analysis and interpretation”. *Frontiers in Molecular Biosciences* 9, 2022, p. 841373.

202. E. Benedetti, M. Pučić-Baković, T. Keser, N. Gerstner, M. Büyüközkan, T. Štambuk, M. H. Selman, I. Rudan, O. Polašek, C. Hayward, et al. “A strategy to incorporate prior knowledge into correlation network cutoff selection”. *Nature communications* 11:1, 2020, p. 5153.

203. Y. Chen, E.-M. Li, and L.-Y. Xu. “Guide to metabolomics analysis: a bioinformatics workflow”. *Metabolites* 12:4, 2022, p. 357.

204. J. P. Cunningham and Z. Ghahramani. “Linear dimensionality reduction: Survey, insights, and generalizations”. *The Journal of Machine Learning Research* 16:1, 2015, pp. 2859–2900.

205. K. Pearson. “LIII. On lines and planes of closest fit to systems of points in space”. *The London, Edinburgh, and Dublin philosophical magazine and journal of science* 2:11, 1901, pp. 559–572.

206. D. DeMers and G. Cottrell. “Non-linear dimensionality reduction”. *Advances in neural information processing systems* 5, 1992.

207. B. Raducanu and F. Dornaika. “A supervised non-linear dimensionality reduction approach for manifold learning”. *Pattern Recognition* 45:6, 2012, pp. 2432–2444.

208. R. R. Coifman and S. Lafon. “Diffusion maps”. *Applied and computational harmonic analysis* 21:1, 2006, pp. 5–30.

209. L. Van der Maaten and G. Hinton. “Visualizing data using t-SNE.” *Journal of machine learning research* 9:11, 2008.

210. L. McInnes, J. Healy, and J. Melville. “Umap: Uniform manifold approximation and projection for dimension reduction”. *arXiv preprint arXiv:1802.03426*, 2018.

211. K. R. Moon, D. van Dijk, Z. Wang, S. Gigante, D. B. Burkhardt, W. S. Chen, K. Yim, A. v. d. Elzen, M. J. Hirn, R. R. Coifman, et al. “Visualizing structure and transitions in high-dimensional biological data”. *Nature biotechnology* 37:12, 2019, pp. 1482–1492.

212. C. Fefferman, S. Mitter, and H. Narayanan. “Testing the manifold hypothesis”. *Journal of the American Mathematical Society* 29:4, 2016, pp. 983–1049.

213. G. Huguet, D. S. Magruder, A. Tong, O. Fasina, M. Kuchroo, G. Wolf, and S. Krishnaswamy. “Manifold interpolating optimal-transport flows for trajectory inference”. *Advances in Neural Information Processing Systems* 35, 2022, pp. 29705–29718.

214. D. Bank, N. Koenigstein, and R. Giryes. “Autoencoders”. *Machine Learning for Data Science Handbook: Data Mining and Knowledge Discovery Handbook*, 2023, pp. 353–374.

215. D. P. Gomari, A. Schweickart, L. Cerchietti, E. Paietta, H. Fernandez, H. Al-Amin, K. Suhre, and J. Krumsiek. “Variational autoencoders learn transferrable representations of metabolomics data”. *Communications Biology* 5:1, 2022, p. 645.

216. T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Vol. 2. Springer, 2009.

217. K. P. Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

218. C.-F. Gauss. *Theoria combinationis observationum erroribus minimis obnoxiae*. Henricus Dieterich, 1823.

219. F. A. Graybill. "Theory and application of the linear model". *(No Title)*, 1976.

220. E. W. Steyerberg and E. W. Steyerberg. "Overfitting and optimism in prediction models". *Clinical prediction models: A practical approach to development, validation, and updating*, 2019, pp. 95–112.

221. S. Kapoor and A. Narayanan. "Leakage and the reproducibility crisis in machine-learning-based science". *Patterns*, 2023.

222. D. Häcker†, K. Siebert†, B. J. Smith, N. Köhler, H. Heimes, A. Metwaly, A. Mahapatra, H. L. Hoelz, F. De Zen, J. Heetmeyer, et al. "Exclusive Enteral Nutrition Initiates Protective Microbiome Changes to Induce Remission in Pediatric Crohn's Disease". *medRxiv*, 2023, pp. 2023–12.

223. T. K. Ho. "Random decision forests". In: *Proceedings of 3rd international conference on document analysis and recognition*. Vol. 1. IEEE. 1995, pp. 278–282.

224. J. H. Friedman. "Greedy function approximation: a gradient boosting machine". *Annals of statistics*, 2001, pp. 1189–1232.

225. C. Cortes and V. Vapnik. "Support-vector networks". *Machine learning* 20:3, 1995, pp. 273–297.

226. K. Hornik, M. Stinchcombe, and H. White. "Multilayer feedforward networks are universal approximators". *Neural networks* 2:5, 1989, pp. 359–366.

227. A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al. "Pytorch: An imperative style, high-performance deep learning library". *Advances in neural information processing systems* 32, 2019.

228. Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Y. Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. Software available from tensorflow.org. 2015.

229. J. Bradbury, R. Frostig, P. Hawkins, M. J. Johnson, C. Leary, D. Maclaurin, G. Necula, A. Paszke, J. VanderPlas, S. Wanderman-Milne, and Q. Zhang. *JAX: composable transformations of Python+NumPy programs*. Version 0.3.13. 2018.

230. M. M. Bronstein, J. Bruna, Y. LeCun, A. Szlam, and P. Vandergheynst. “Geometric deep learning: going beyond euclidean data”. *IEEE Signal Processing Magazine* 34:4, 2017, pp. 18–42.

231. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. “Attention is all you need”. *Advances in neural information processing systems* 30, 2017.

232. T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. “Language models are few-shot learners”. *Advances in neural information processing systems* 33, 2020, pp. 1877–1901.

233. A. Luque, A. Carrasco, A. Martín, and A. de Las Heras. “The impact of class imbalance in classification performance metrics based on the binary confusion matrix”. *Pattern Recognition* 91, 2019, pp. 216–231.

234. J. N. Mandrekar. “Receiver operating characteristic curve in diagnostic test assessment”. *Journal of Thoracic Oncology* 5:9, 2010, pp. 1315–1316.

235. J. MacQueen et al. “Some methods for classification and analysis of multivariate observations”. In: *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*. Vol. 1. 14. Oakland, CA, USA. 1967, pp. 281–297.

236. S. C. Johnson. “Hierarchical clustering schemes”. *Psychometrika* 32:3, 1967, pp. 241–254.

237. M. Ester, H.-P. Kriegel, J. Sander, X. Xu, et al. “A density-based algorithm for discovering clusters in large spatial databases with noise”. In: *kdd*. Vol. 96. 34. 1996, pp. 226–231.

238. L. McInnes, J. Healy, and S. Astels. “hdbscan: Hierarchical density based clustering.” *J. Open Source Softw.* 2:11, 2017, p. 205.

239. T. D. Rose, T. Bechtler, O.-A. Ciora, K. Anh Lilian Le, F. Molnar, N. Köhler, J. Baumbach, R. Röttger, and J. K. Pauling. “MoSBi: Automated signature mining for molecular stratification and subtyping”. *Proceedings of the National Academy of Sciences* 119:16, 2022, e2118210119.

240. N. Köhler, M. Höring, B. Czepukojc, T. D. Rose, C. Buechler, T. Kröhler, J. Haybaeck, G. Liebisch, J. K. Pauling, S. M. Kessler, et al. “Kupffer cells are protective in alcoholic steatosis”. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease* 1868:6, 2022, p. 166398.

241. Z. Ni and M. Fedorova. “LipidLynxX: a data transfer hub to support integration of large scale lipidomics datasets”. *Biorxiv*, 2020, pp. 2020–04.

242. R. Alcántara, K. B. Axelsen, A. Morgat, E. Belda, E. Coudert, A. Bridge, H. Cao, P. De Matos, M. Ennis, S. Turner, et al. “Rhea—a manually curated resource of biochemical reactions”. *Nucleic acids research* 40:D1, 2012, pp. D754–D760.

243. B. Jassal, L. Matthews, G. Viteri, C. Gong, P. Lorente, A. Fabregat, K. Sidiropoulos, J. Cook, M. Gillespie, R. Haw, et al. “The reactome pathway knowledgebase”. *Nucleic acids research* 48:D1, 2020, pp. D498–D503.

244. B. Bao, B. P. Kellman, A. W. Chiang, Y. Zhang, J. T. Sorrentino, A. K. York, M. A. Mohammad, M. W. Haymond, L. Bode, and N. E. Lewis. "Correcting for sparsity and interdependence in glycomics by accounting for glycan biosynthesis". *Nature communications* 12:1, 2021, p. 4988.

245. M. Gillespie, B. Jassal, R. Stephan, M. Milacic, K. Rothfels, A. Senff-Ribeiro, J. Griss, C. Sevilla, L. Matthews, C. Gong, et al. "The reactome pathway knowledgebase 2022". *Nucleic acids research* 50:D1, 2022, pp. D687–D692.

246. A. Noronha, J. Modamio, Y. Jarosz, E. Guerard, N. Sompairac, G. Preciat, A. D. Daníelsdóttir, M. Krecke, D. Merten, H. S. Haraldsdóttir, et al. "The Virtual Metabolic Human database: integrating human and gut microbiome metabolism with nutrition and disease". *Nucleic acids research* 47:D1, 2019, pp. D614–D624.

247. J. Chong, D. S. Wishart, and J. Xia. "Using MetaboAnalyst 4.0 for comprehensive and integrative metabolomics data analysis". *Current protocols in bioinformatics* 68:1, 2019, e86.

248. J. C. Alarcon-Barrera, S. Kostidis, A. Ondo-Mendez, and M. Giera. "Recent advances in metabolomics analysis for early drug development". *Drug discovery today* 27:6, 2022, pp. 1763–1773.

249. A. Kvasnička, L. Najdekr, D. Dobešová, B. Piskláková, E. Ivanovová, and D. Friedecký. "Clinical lipidomics in the era of the big data". *Clinical Chemistry and Laboratory Medicine (CCLM)* 61:4, 2023, pp. 587–598.

250. R. Bellman. "Dynamic programming". *Science* 153:3731, 1966, pp. 34–37.

251. C. Gaud, B. C. Sousa, A. Nguyen, M. Fedorova, Z. Ni, V. B. O'Donnell, M. J. Wakelam, S. Andrews, and A. F. Lopez-Clavijo. "BioPAN: a web-based tool to explore mammalian lipidome metabolic pathways on LIPID MAPS". *F1000Research* 10, 2021.

252. A. Nguyen, S. A. Rudge, Q. Zhang, and M. J. Wakelam. "Using lipidomics analysis to determine signalling and metabolic changes in cells". *Current opinion in biotechnology* 43, 2017, pp. 96–103.

253. T. Damiani, S. Bonciarelli, G. G. Thallinger, N. Koehler, C. A. Krettler, A. K. Salihoglu, A. Korf, J. K. Pauling, T. Pluskal, Z. Ni, et al. "Software and computational tools for LC-MS-based epilipidomics: Challenges and solutions". *Analytical chemistry* 95:1, 2023, pp. 287–303.

254. M. Pearson, C. Hunter, and T. Baba. "Complete structural elucidation of lipids in a single experiment using electron activated dissociation (EAD)". *SCIEX technical note*, 2021.

255. J. L. Campbell and T. Baba. "Near-complete structural characterization of phosphatidylcholines using electron impact excitation of ions from organics". *Analytical chemistry* 87:11, 2015, pp. 5837–5845.

256. G. A. Pavlopoulos, M. Secrier, C. N. Moschopoulos, T. G. Soldatos, S. Kossida, J. Aerts, R. Schneider, and P. G. Bagos. "Using graph theory to analyze biological networks". *BioData mining* 4, 2011, pp. 1–27.

257. C. Gu, G. B. Kim, W. J. Kim, H. U. Kim, and S. Y. Lee. "Current status and applications of genome-scale metabolic models". *Genome biology* 20, 2019, pp. 1–18.

258. W. Zhang, J. Chien, J. Yong, and R. Kuang. “Network-based machine learning and graph theory algorithms for precision oncology”. *NPJ precision oncology* 1:1, 2017, p. 25.

259. T. B. Ware, C. E. Franks, M. E. Granade, M. Zhang, K.-B. Kim, K.-S. Park, A. Gahlmann, T. E. Harris, and K.-L. Hsu. “Reprogramming fatty acyl specificity of lipid kinases via C1 domain engineering”. *Nature chemical biology* 16:2, 2020, pp. 170–178.

260. R. Argelaguet, B. Velten, D. Arnol, S. Dietrich, T. Zenz, J. C. Marioni, F. Buettner, W. Huber, and O. Stegle. “Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets”. *Molecular systems biology* 14:6, 2018, e8124.

261. A. Singh, C. P. Shannon, B. Gautier, F. Rohart, M. Vacher, S. J. Tebbutt, and K.-A. Lê Cao. “DIABLO: an integrative approach for identifying key molecular drivers from multi-omics assays”. *Bioinformatics* 35:17, 2019, pp. 3055–3062.

262. F. E. Agamah, J. R. Bayjanov, A. Niehues, K. F. Njoku, M. Skelton, G. K. Mazandu, T. H. Ederveen, N. Mulder, E. R. Chimusa, and P. A. t Hoen. “Computational approaches for network-based integrative multi-omics analysis”. *Frontiers in Molecular Biosciences* 9, 2022, p. 1214.

263. M. Wang, J. J. Carver, V. V. Phelan, L. M. Sanchez, N. Garg, Y. Peng, D. D. Nguyen, J. Watrous, C. A. Kapono, T. Luzzatto-Knaan, et al. “Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking”. *Nature biotechnology* 34:8, 2016, pp. 828–837.

264. N. Poupin, F. Vinson, A. Moreau, A. Batut, M. Chazalviel, B. Colsch, L. Fouillen, S. Guez, S. Khoury, J. Dalloux-Chioccioli, et al. “Improving lipid mapping in Genome Scale Metabolic Networks using ontologies”. *Metabolomics* 16, 2020, pp. 1–11.

265. C. Frainay and F. Jourdan. “Computational methods to identify metabolic sub-networks based on metabolomic profiles”. *Briefings in bioinformatics* 18:1, 2017, pp. 43–56.

266. M. R. Antoniewicz. “A guide to metabolic flux analysis in metabolic engineering: Methods, tools and applications”. *Metabolic engineering* 63, 2021, pp. 2–12.

267. A. Heinken, J. Hertel, and I. Thiele. “Metabolic modelling reveals broad changes in gut microbial metabolism in inflammatory bowel disease patients with dysbiosis”. *NPJ Systems Biology and Applications* 7:1, 2021, p. 19.

268. N. Berndt, A. Egners, G. Mastrobuoni, O. Vvedenskaya, A. Fragoulis, A. Dugourd, S. Bulik, M. Pietzke, C. Bielow, R. van Gassel, et al. “Kinetic modelling of quantitative proteome data predicts metabolic reprogramming of liver cancer”. *British Journal of Cancer* 122:2, 2020, pp. 233–244.

269. W. Wiechert and S. Noack. “Mechanistic pathway modeling for industrial biotechnology: challenging but worthwhile”. *Current opinion in biotechnology* 22:5, 2011, pp. 604–610.

270. E. Vasilakou, D. Machado, A. Theorell, I. Rocha, K. Nöh, M. Oldiges, and S. A. Wahl. “Current state and challenges for dynamic metabolic modeling”. *Current opinion in microbiology* 33, 2016, pp. 97–104.

271. O. D. Kim, M. Rocha, and P. Maia. "A review of dynamic modeling approaches and their application in computational strain optimization for metabolic engineering". *Frontiers in microbiology* 9.

272. S. Picart-Armada, F. Fernández-Albert, M. Vinaixa, M. A. Rodríguez, S. Aivio, T. H. Stracker, O. Yanes, and A. Perera-Lluna. "Null diffusion-based enrichment for metabolomics data". *PloS one* 12:12, 2017, e0189012.

273. M. A. Wörheide, J. Krumsiek, G. Kastenmüller, and M. Arnold. "Multi-omics integration in biomedical research–A metabolomics-centric review". *Analytica chimica acta* 1141, 2021, pp. 144–162.

274. N. Pham, R. G. van Heck, J. C. van Dam, P. J. Schaap, E. Saccenti, and M. Suarez-Diez. "Consistency, inconsistency, and ambiguity of metabolite names in biochemical databases used for genome-scale metabolic modelling". *Metabolites* 9:2, 2019, p. 28.

275. M. A. Stravs, K. Dührkop, S. Böcker, and N. Zamboni. "MSNovelist: de novo structure generation from mass spectra". *Nature Methods* 19:7, 2022, pp. 865–870.

276. M. Murphy, S. Jegelka, E. Fraenkel, T. Kind, D. Healey, and T. Butler. "Efficiently predicting high resolution mass spectra with graph neural networks". *arXiv preprint arXiv:2301.11419*, 2023.

277. I. Thiele and B. Ø. Palsson. "A protocol for generating a high-quality genome-scale metabolic reconstruction". *Nature protocols* 5:1, 2010, pp. 93–121.

278. C. J. Norsigian, X. Fang, Y. Seif, J. M. Monk, and B. O. Palsson. "A workflow for generating multi-strain genome-scale metabolic models of prokaryotes". *Nature protocols* 15:1, 2020, pp. 1–14.

279. D. Machado, S. Andrejev, M. Tramontano, and K. R. Patil. "Fast automated reconstruction of genome-scale metabolic models for microbial species and communities". *Nucleic acids research* 46:15, 2018, pp. 7542–7553.

280. J. Zimmermann, C. Kaleta, and S. Waschina. "gapseq: Informed prediction of bacterial metabolic pathways and reconstruction of accurate metabolic models". *Genome biology* 22:1, 2021, pp. 1–35.

281. N. Yadati, V. Nitin, M. Nimishakavi, P. Yadav, A. Louis, and P. Talukdar. "NHP: Neural hypergraph link prediction". In: *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 2020, pp. 1705–1714.

282. G. Sharma, P. Patil, and M. N. Murty. "C3MM: Clique-closure based hyperlink prediction". In: *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*. 2021, pp. 3364–3370.

283. P. Badia-i-Mompel, L. Wessels, S. Müller-Dott, R. Trimbour, R. O. Ramirez Flores, R. Argelaguet, and J. Saez-Rodriguez. "Gene regulatory network inference in the era of single-cell multi-omics". *Nature Reviews Genetics*, 2023, pp. 1–16.

284. Q. Wang, M. Guo, J. Chen, and R. Duan. "A gene regulatory network inference model based on pseudo-siamese network". *BMC bioinformatics* 24:1, 2023, pp. 1–18.

285. V. A. Huynh-Thu, A. Irrthum, L. Wehenkel, and P. Geurts. "Inferring regulatory networks from expression data using tree-based methods". *PloS one* 5:9, 2010, e12776.

286. T. Moerman, S. Aibar Santos, C. Bravo González-Blas, J. Simm, Y. Moreau, J. Aerts, and S. Aerts. "GRNBoost2 and Arboreto: efficient and scalable inference of gene regulatory networks". *Bioinformatics* 35:12, 2019, pp. 2159–2161.

287. Y. Yuan and Z. Bar-Joseph. "Deep learning for inferring gene relationships from single-cell expression data". *Proceedings of the National Academy of Sciences* 116:52, 2019, pp. 27151–27158.

288. M. Zhao, W. He, J. Tang, Q. Zou, and F. Guo. "A hybrid deep learning framework for gene regulatory network inference from single-cell transcriptomic data". *Briefings in bioinformatics* 23:2, 2022, bbab568.

289. D. Bhaskar, S. Magruder, E. De Brouwer, A. Venkat, F. Wenkel, G. Wolf, and S. Krishnaswamy. "Inferring dynamic regulatory interaction graphs from time series data with perturbations". *arXiv preprint arXiv:2306.07803*, 2023.

290. Y. Zhang, J. Li, X. Zhang, D. Song, and T. Tian. "Advances of mechanisms-related metabolomics in Parkinson's disease". *Frontiers in Neuroscience* 15, 2021, p. 614251.

291. D. Dai, C. He, S. Wang, M. Wang, N. Guo, and P. Song. "Toward personalized interventions for psoriasis vulgaris: molecular subtyping of patients by using a metabolomics approach". *Frontiers in Molecular Biosciences* 9, 2022, p. 945917.

292. I. Martínez-Arranz, C. Bruzzone, M. Noureddin, R. Gil-Redondo, I. Mincholé, M. Bizkarguenaga, E. Arretxe, M. Iruarrizaga-Lejarreta, D. Fernández-Ramos, F. Lopitz-Otsoa, et al. "Metabolic subtypes of patients with NAFLD exhibit distinctive cardiovascular risk profiles". *Hepatology* 76:4, 2022, pp. 1121–1134.

293. O. Lazareva, S. Canzar, K. Yuan, J. Baumbach, D. B. Blumenthal, P. Tieri, T. Kacprowski, and M. List. "BiCoN: Network-constrained biclustering of patients and omics data". *Bioinformatics* 37:16, 2021, pp. 2398–2404.

294. Y. Xiao, D. Ma, Y.-S. Yang, F. Yang, J.-H. Ding, Y. Gong, L. Jiang, L.-P. Ge, S.-Y. Wu, Q. Yu, et al. "Comprehensive metabolomics expands precision medicine for triple-negative breast cancer". *Cell research* 32:5, 2022, pp. 477–490.

295. O. Vvedenskaya†, T. D. Rose†, O. Knittelfelder, A. Palladini, J. A. H. Wodke, K. Schuhmann, J. M. Ackerman, Y. Wang, C. Has, M. Brosch, et al. "Nonalcoholic fatty liver disease stratification by liver lipidomics". *Journal of lipid research* 62, 2021.

296. N. Rappoport and R. Shamir. "Multi-omic and multi-view clustering algorithms: review and cancer benchmark". *Nucleic acids research* 46:20, 2018, pp. 10546–10562.

297. J. E. Stanton, S. Malijauskaite, K. McGourty, and A. M. Grabrucker. "The metallome as a link between the "omes" in autism spectrum disorders". *Frontiers in Molecular Neuroscience* 14, 2021, p. 695873.

298. D. Kopczynski, N. Hoffmann, B. Peng, and R. Ahrends. "Goslin: a grammar of succinct lipid nomenclature". *Analytical Chemistry* 92:16, 2020, pp. 10957–10960.

299. N. M. O'Boyle. "Towards a Universal SMILES representation-A standard method to generate canonical SMILES based on the InChI". *Journal of cheminformatics* 4, 2012, pp. 1–14.

300. S. R. Heller, A. McNaught, I. Pletnev, S. Stein, and D. Tchekhovskoi. "InChI, the IUPAC international chemical identifier". *Journal of cheminformatics* 7:1, 2015, pp. 1–34.

301. S. Baskiyar, C. Ren, K. L. Heck, A. M. Hall, M. Gulfam, S. Packer, C. D. Seals, and A. I. Calderón. "Bioactive Natural Products Identification Using Automation of Molecular Networking Software". *Journal of Chemical Information and Modeling* 62:24, 2022, pp. 6378–6385.

302. N. J. Morehouse, T. N. Clark, E. J. McMann, J. A. van Santen, F. J. Haeckl, C. A. Gray, and R. G. Linington. "Annotation of natural product compound families using molecular networking topology and structural similarity fingerprinting". *Nature Communications* 14:1, 2023, p. 308.

303. S. Sang, Z. Yang, L. Wang, X. Liu, H. Lin, and J. Wang. "SemaTyP: a knowledge graph based literature mining method for drug discovery". *BMC bioinformatics* 19, 2018, pp. 1–11.

304. R. J. Townshend, S. Eismann, A. M. Watkins, R. Rangan, M. Karelina, R. Das, and R. O. Dror. "Geometric deep learning of RNA structure". *Science* 373:6558, 2021, pp. 1047–1051.

305. F. C. Fernandes, M. H. Cardoso, A. Gil-Ley, L. V. Luchi, M. G. da Silva, M. L. Macedo, C. de la Fuente-Nunez, and O. L. Franco. "Geometric deep learning as a potential tool for antimicrobial peptide prediction". *Frontiers in Bioinformatics* 3, 2023.

306. Z. Zhang, J. Yan, Q. Liu, and E. Che. "A Systematic Survey in Geometric Deep Learning for Structure-based Drug Design". *arXiv preprint arXiv:2306.11768*, 2023.

307. K. Atz, F. Grisoni, and G. Schneider. "Geometric deep learning on molecular representations". *Nature Machine Intelligence* 3:12, 2021, pp. 1023–1032.

308. J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl. "Neural message passing for quantum chemistry". In: *International conference on machine learning*. PMLR. 2017, pp. 1263–1272.

309. U. Alon and E. Yahav. "On the bottleneck of graph neural networks and its practical implications". *arXiv preprint arXiv:2006.05205*, 2020.

310. R. T. Chen, Y. Rubanova, J. Bettencourt, and D. K. Duvenaud. "Neural ordinary differential equations". *Advances in neural information processing systems* 31, 2018.

311. R. Erbe, G. Stein-O'Brien, and E. J. Fertig. "Transcriptomic forecasting with neural ordinary differential equations". *Patterns* 4:8, 2023.

312. E. De Brouwer, J. Simm, A. Arany, and Y. Moreau. "GRU-ODE-Bayes: Continuous modeling of sporadically-observed time series". *Advances in neural information processing systems* 32, 2019.

313. P. Kidger, J. Morrill, J. Foster, and T. Lyons. "Neural controlled differential equations for irregular time series". *Advances in Neural Information Processing Systems* 33, 2020, pp. 6696–6707.

314. P. Kidger. “On neural differential equations”. *arXiv preprint arXiv:2202.02435*, 2022.

315. K. Haug, K. Cochrane, V. C. Nainala, M. Williams, J. Chang, K. V. Jayaseelan, and C. O’Donovan. “MetaboLights: a resource evolving in response to the needs of its scientific community”. *Nucleic acids research* 48:D1, 2020, pp. D440–D444.

316. T. Alexandrov, K. Ovchinnikova, A. Palmer, V. Kovalev, A. Tarasov, L. Stuart, R. Nigmetzianov, D. Fay, K. M. contributors, M. Gaudin, et al. “METASPACE: A community-populated knowledge base of spatial metabolomes in health and disease”. *BioRxiv*, 2019, p. 539478.

317. H. Horai, M. Arita, S. Kanaya, Y. Nihei, T. Ikeda, K. Suwa, Y. Ojima, K. Tanaka, S. Tanaka, K. Aoshima, et al. “MassBank: a public repository for sharing mass spectral data for life sciences”. *Journal of mass spectrometry* 45:7, 2010, pp. 703–714.

318. R. A. Spicer and C. Steinbeck. “A lost opportunity for science: journals promote data sharing in metabolomics but do not enforce it”. *Metabolomics* 14, 2018, pp. 1–4.

319. M. Witting. “(Re-) use and (re-) analysis of publicly available metabolomics data”. *Proteomics*, 2023, p. 2300032.

320. K. Hrovatin, D. S. Fischer, and F. J. Theis. “Toward modeling metabolic state from single-cell transcriptomics”. *Molecular metabolism* 57, 2022, p. 101396.

321. A. Wagner, C. Wang, J. Fessler, D. DeTomaso, J. Avila-Pacheco, J. Kaminski, S. Zaghouani, E. Christian, P. Thakore, B. Schellhaass, et al. “Metabolic modeling of single Th17 cells reveals regulators of autoimmunity”. *Cell* 184:16, 2021, pp. 4168–4185.

322. A.-D. Brunner, M. Thielert, C. Vasilopoulou, C. Ammar, F. Coscia, A. Mund, O. B. Hoerning, N. Bache, A. Apalategui, M. Lubeck, et al. “Ultra-high sensitivity mass spectrometry quantifies single-cell proteome changes upon perturbation”. *Molecular systems biology* 18:3, 2022, e10798.

323. M. Kokkinidis, N. M. Glykos, and V. E. Fadouloglou. “Catalytic activity regulation through post-translational modification: The expanding universe of protein diversity”. *Advances in Protein Chemistry and Structural Biology* 122, 2020, pp. 97–125.

324. A. Wegner, J. Meiser, D. Weindl, and K. Hiller. “How metabolites modulate metabolic flux”. *Current opinion in biotechnology* 34, 2015, pp. 16–22.

[†] These authors contributed equally