HELMHOLTZ
ASSOCIATION
Alliance on Systems Biology

HelmholtzZentrum münchen
German Research Center for Environmental Health

TECHNISCHE
UNIVERSITÄT
MÜNCHEN

Computational modeling of metabolite dependencies: From metabolomics data to biochemical networks

Jan Krumsiek

2012

TECHNISCHE UNIVERSITÄT MÜNCHEN

Lehrstuhl für Genomorientierte Bioinformatik

Computational modeling of metabolite dependencies: From metabolomics data to biochemical networks

Jan Krumsiek

Vollständiger Abdruck der von der Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften

genehmigten Dissertation.

Vorsitzender:	Univ.-Prof. Dr. M. Hrabě de Angelis
Prüfer der Dissertation:	1. Univ.-Prof. Dr. H.-W. Mewes 2. Univ.-Prof. Dr. Dr. F. J. Theis 3. Univ.-Prof. Dr. Th. Dandekar Julius-Maximilians-Universität Würzburg (nur schriftliche Beurteilung) apl. Prof. Dr. J. Adamski (mündliche Prüfung)

Die Dissertation wurde am 15.05.2012 bei der Technischen Universität München eingereicht und durch die Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt am 12.10.2012 angenommen.

Danksagung

An dieser Stelle möchte ich mich bei einigen Personen bedanken.

Zunächst bei Fabian Theis für seine uneingeschränkte Unterstützung über inzwischen mehr als vier Jahre. Bei Hans-Werner Mewes dafür, dass er uns alle am IBIS zusammengebracht und das Projekt möglich gemacht hat. Bei Karsten Suhre und Gabi Kastenmüller für die intensive Zusammenarbeit in zahllosen Stunden.

Bei den vielen Kollegen aus den unterschiedlichen Kollaborationen im Metabolomics-Projekt, insbesondere bei Kirstin, Ann-Kristin, Janina, Carolin, Elisabeth, Gitti, Nikola, Anke Meyer-Baese, Thomas Illig, Jerzi Adamski und Christian Gieger.

Bei Ferdi und Jörg, die zu dem Metabolomics-Projekt dazu gestoßen sind und mit denen ich seit einiger Zeit täglich gerne zusammenarbeite.

Vielen Dank auch an die anderen Kollegen und guten Freunde aus der CMB Gruppe. Ohne die nette Arbeitsatmosphäre hätte es nur halb so viel Spaß gemacht.

Weiterhin möchte ich mich bei der Studienstiftung des deutschen Volkes und dem BMBF für die finanzielle Unterstützung meiner Doktorarbeit bedanken.

Mein ganz besonderer Dank gilt meinen Eltern und Großeltern, für die große Unterstützung und das ständige Interesse an meiner Forschungsarbeit.

Und schließlich Tini. Vielen Dank, dass du die manchmal überlangen Arbeitstage ausgehalten hast und immer für mich da warst. ⋆

Abstract

Metabolites are intermediate molecules of metabolic processes such as sugars, amino acids, fatty acids or vitamins, which are nowadays measured in a high-throughput manner. The term ‘metabolomics’ was coined for this new discipline around ten years ago, taking up the popular trend of ‘omics’ words for large-scale measurements. Since metabolites are strongly interconnected in a biochemical reaction network, the measured metabolite concentrations are not independent. Statistical associations between metabolites can be assessed by measures like the Pearson correlation coefficient. Intuitively, if two molecules are biochemically connected, then high concentrations of one metabolite tend to coincide with high concentrations of the other metabolite, and vice versa. This thesis provides an in-depth investigation of such statistical relationships between metabolites in large-scale measurements. Our major aim is to verify whether metabolite correlation structures carry a detectable footprint of the underlying metabolic reaction networks.

A major drawback of regular Pearson correlations is their inability to distinguish between direct and indirect associations. Even for distantly related molecule pairs, we regularly observe profoundly high correlation coefficients that are hardly distinguishable from those correlations of biochemically related species. To address the issue of unspecific correlation coefficients and indirect effects, we use Gaussian graphical models (GGMs). GGMs are based on so-called partial correlation coefficients and represent a specific class of probabilistic graphical models, which encode the conditional independence structure between measured entities. The partial correlation of two metabolites is given by the pairwise correlation after the effects of all other metabolites have been removed. In this thesis, we focused on evaluating whether GGMs on metabolomics data correspond to real metabolic networks.

First, we evaluate whether GGMs have the capacity to reconstruct the topology of a biochemical network by using computer-simulated reaction systems. This forward simulation approach has the advantage of knowing the correct topology beforehand. The reconstruction succeeds for all reaction topologies, except for few scenarios like entirely irreversible reaction chains or specific types of negative feedback loops. Interestingly, reconstruction quality of the GGM is better for stronger variation of the input reaction into the system. This indicates that strong variability as observed in a human metabolomics dataset is rather beneficial than detrimental for the reconstruction of metabolic reactions with GGMs.

Second, we calculate a GGM on a real metabolomics datasets comprising 1000-3000 human serum samples with several hundreds of measured metabolites from a German population cohort. Global inspection of the resulting GGM network reveals a modular structure with respect to the underlying metabolic classes. That is, metabolites that belong to the same class, like amino acids or sphingomyelins, tend to be more connected than molecules from different classes. A manual investigation of subnetworks with high partial correlations reveals known biochemical reaction cascades from the beta-oxidation pathway and fatty acid biosynthesis. We systematically validate this observation by comparing partial correlation coefficients with network distances from a manually curated fatty acid pathway model. The analysis reveals significantly higher partial correlations for metabolite pairs with a pathway distance of exactly one. While this result was expectable, it had never been systematically shown before. Our findings suggest that indeed GGMs are able to recover biochemical reaction steps from human blood metabolomics data.

Third, we extend metabolomics GGMs by large-scale SNP genotyping data in order to tackle a fundamental problem of the experimental field: Metabolomics measurements usually generate a substantial amount of signals that are reliably detected in the samples, but for which the biochemical identity of the respective compound remains unknown. By combining metabolite-SNP associations, GGMs and publically available reaction lists, we derive pathway classifications for a large fraction of these unknown metabolites. Specifically, GGM edges with known metabolites or genetic associations with loci encoding for a metabolic protein point us towards specific parts of the metabolic pathway in which the unknown compound might be involved. For several cases, this even allows for concrete identity predictions which are then validated experimentally. As an additional result of the analysis, we find seven previously unreported loci of metabolic individuality, i.e. loci where genetic variation coincides with changes in blood metabolite concentrations.

Fourth, we biologically exploit the data-driven metabolic networks reconstructed by the GGM approach in three different directions: (1) We introduce the concept of 'effect networks', where we annotate a GGM with results from statistical analyses. For instance, we color each metabolite node with relative metabolite differences between male and female probands in the cohort. This graphical illustration then allows to identify specific effects of gender differences within the metabolic pathway. The effect network approach is furthermore applied to fat-free mass and Type-D personality as analyzed phenotypes. (2) We develop a 'differential' GGM on the lipidome of glioblastoma cells

under varying conditions. By running the estimation procedure on different subsets of the available samples, we are able to identify specific effects on the lipidome of a combined treatment with a chemotherapeutic agent and a gene construct. (3) We use the GGM to define biologically meaningful groups of metabolites for two different biological applications. Intuitively, since partial correlations are closely connected to biochemical reaction systems, a clustering of the GGM will result in groups of biochemically related metabolites.

Finally, we shift the focus to independent component analysis (ICA). Covariance-based methods like GGMs miss higher-order statistical dependencies, which may contain additional information on the underlying relationships. We introduce a Bayesian, noisy ICA framework and discuss the application to the blood serum metabolomics data set. The recovered statistically independent components each contain strong signatures of individual metabolic pathways, including amino acid metabolism, lipid metabolism, and energy metabolism. Moreover, the strength of one independent component (primarily containing branched-chain amino acids) in the probands displays a stronger association with plasma HDL concentrations than any metabolite.

Taken together, we demonstrated that GGMs and ICA are able to reconstruct pathway signatures from high-throughput metabolomics data. Interestingly, our results could be obtained from metabolomics data in human blood samples. This suggests that blood metabolites not only represent products of leaking from larger metabolically active organs into the vascular system, but carry a full footprint of the metabolic pathways. Moreover, we showed that GGMs improved the detection of metabolome-phenotype associations and possible pathological dysfunctions in blood samples. In summary, this thesis provided new insights and bioinformatical analysis methods for the statistical relationships between metabolites, forming a more comprehensive picture of human metabolism.

Zusammenfassung

Unter ‘Metaboliten’ versteht man Zwischenprodukte des Stoffwechsels, wie zum Beispiel Zucker, Aminosäuren, Fettsäuren oder Vitamine. Seit einigen Jahren können Metaboliten in großer Anzahl mit Hochdurchsatzmethoden gemessen werden. Diese neue Disziplin nennt sich ‘Metabolomics’ und knüpft damit an den beliebten Trend der ‘omics’ Begriffe für systemweite Messungen an. Da Metaboliten in einem komplexen, biochemischen Reaktionsnetzwerk verknüpft sind, stellen die entsprechenden Metabolitenkonzentrationen keine unabhängigen Signale dar, sondern weisen starke Assoziationen miteinander auf. Solche Zusammenhänge werden in der Regel durch statistische Methoden, wie dem Pearson Korrelationskoeffizienten, erfasst. Eine biochemische Verbindung zwischen zwei Metaboliten führt zu positiver Korrelation, d.h. hohe Konzentrationen des einen Metaboliten gehen mit hohen Konzentrationen des jeweils anderen einher und umgekehrt. In dieser Arbeit wird ein detaillierter Analyseansatz solcher statistischer Zusammenhänge zwischen Metaboliten in großen Messdatensätzen entwickelt. Das Hauptziel ist dabei, zu prüfen, ob die Korrelationsstrukturen zwischen Metaboliten einen messbaren ‘Abdruck’ des darunterliegenden Stoffwechselnetzwerkes beinhalten.

Ein entscheidendes Problem von herkömmlichen Pearson-Korrelationen ist die Unfähigkeit, zwischen direkten und indirekten Interaktionen in den Daten zu unterscheiden. Selbst für lediglich entfernt verwandte Metabolitenpaare beobachtet man häufig hohe Korrelationen, welche sich kaum von denen von direkt verknüpften Metabolitenpaaren unterscheiden lassen. In dieser Arbeit werden Gaußsche grafische Modelle (GGMs) eingesetzt, um dieses Problem der unspezifischen Korrelationen gezielt zu bearbeiten. GGMs basieren auf partiellen Korrelationskoeffizienten und gehören zu einer bestimmten Klasse probabilistischer grafischer Modelle, welche die bedingten Unabhängigkeitsstrukturen zwischen gemessenen Variablen abbilden. Die partielle Korrelation zwischen zwei Metaboliten errechnet sich aus der herkömmlichen paarweisen Korrelation, nachdem die Effekte aller anderen Metaboliten entfernt worden sind. Ein Ziel dieser Arbeit ist es, zu prüfen, ob die Kanten in einem GGM tatsächlichen biochemischen Reaktionen entsprechen.

Im ersten Schritt werden wir die Anwendbarkeit von GGMs zur Rekonstruktion metabolischer Netzwerke auf computersimulierten Reaktionssystem evaluieren. Da in solchen Simulationsansätzen die tatsächliche Netzwerk-Topologie bereits bekannt ist, eignen sie sich besonders zur Auswertung von Netzwerkrekonstruktionsverfahren wie den GGMs. Für die Mehrzahl der simulierten Systeme rekonstruiert das GGM die Netzwerktopo-

logie korrekt. Lediglich einige wenige Szenarien mit irreversiblen Reaktionsketten oder bestimmten Formen der negativen Rückkopplung können nicht korrekt erkannt werden. Weiterhin zeigen wir, dass stärkere Variationen der Eingangsreaktionen in das System zu einer verbesserten Rekonstruktionsqualität führen. Dieses Ergebnis suggeriert, dass starke Schwankungen, wie wir sie in menschlichen Metabolomicsdaten beobachten, für den Rekonstruktionsprozess eher vorteilhaft als problematisch sind.

Im zweiten Schritt berechnen wir ein GGM auf Metabolomicsdaten einer großen deutschen Populationskohorte mit mehreren hundert gemessener Metaboliten in 1000-3000 Proben. Das rekonstruierte Netzwerk weist eine modulare Struktur in Bezug auf die zugrundeliegenden metabolischen Klassen auf. Dies bedeutet, dass Metaboliten tendenziell mit anderen Metaboliten derselben Klasse verbunden sind und eher wenige Verbindungen zu anderen Klassen aufweisen. Weiterhin können wir bestimmte Teilnetzwerke mit hohen partiellen Korrelationen bereits bekannten Stoffwechselwegen (‘Pathways’), wie zum Beispiel der Beta-Oxidation und der Fettsäuresynthese, zuweisen. Diesen Befund können wir durch einen systematischen Vergleich aller partiellen Korrelationskoeffizienten mit Netzwerkdistanzen aus einem Pathway-Modell des Fettsäurestoffwechsels weiter belegen. Metabolitenpaare, die durch eine biochemische Reaktion direkt verbunden sind, weisen signifikant höhere partielle Korrelation auf nicht direkt verbundene Paare. Zwar war dies grundsätzlich zu erwarten, wurde aber in keiner Arbeit zuvor systematisch gezeigt. Dieses Ergebnis belegt, dass GGMs tatsächlich biochemische Reaktionen in Metabolomicsdaten aus menschlichem Blut ermitteln können.

Im nächsten Schritt wird der Metabolomics GGM Ansatz um SNP Genotypisierungsdaten erweitert. Das Ziel dieses Ansatzes ist die Bearbeitung eines grundsätzlichen Problemes des Metabolomics Felds: In Metabolomicsexperimenten werden üblicherweise eine erhebliche Menge reproduzierbarer Signale erzeugt, für welche die biochemische Identität der zugrundliegenden Substanz noch nicht aufgeklärt werden konnte. Durch die Kombination von Metabolit-SNP Assoziationen, GGMs und Reaktionslisten aus öffentlichen Datenbanken können wir Pathway Klassifikationen für eine große Anzahl dieser ‘unbekannten’ Metaboliten erstellen. Sowohl GGM Kanten zwischen unbekannten und bekannten Metaboliten als auch genetische Assoziationen zwischen unbekannten Metaboliten und genetischen Loci liefern entsprechende Hinweise auf die Stoffwechselwege, in welchen der unbekannte Metabolit eine Rolle spielen könnte. In einigen Fällen können auf diesem Weg sogar konkrete Vorhersagen über die biochemischen Identitäten der Unbekannten hergeleitet werden, welche anschließend experimentell getestet werden. Als

Nebenprodukt unserer Analyse können wir sieben neue genetische Loci identifizieren, welche mit Konzentrationsveränderungen von Metaboliten im Blut einhergehen.

Die von den GGMs rekonstruierten, datengestützen metabolischen Netzwerke werden im Folgenden in drei Ansätzen für biologische Fragestellungen verwendet. (1) Wir führen sogenannte 'Effect networks' ein, d.h. GGMs welche mit den Ergebnissen von differentiellen statistischen Analysen annotiert werden. Beispielsweise werden die Knoten in einem GGM mit geschlechtsspezifischen Unterschieden in den Metabolitenkonzentrationen angefärbt. Durch diese grafische Darstellung können wir anschließend spezifische Unterschiede in den Stoffwechselwegen zwischen Männern und Frauen identifizieren. Weiterhin wird der Effect network Ansatz für die Analyse von Stoffwechseleffekten des Fettfreie-Masse-Index und der Type D Persönlichkeit in den Probanden eingesetzt. (2) Wir entwickeln einen 'differenziellen' GGM Ansatz und wenden ihn auf Lipidomics Daten einer Glioblastom Zelllinie unter verschiedenen experimentellen Bedingungen an. Durch Berechnung von GGMs auf verschiedenen Teilmengen der gemessenen Proben können wir spezifische Effekte eines Chemotherapeutikums und einer speziellen Gentherapie in den Stoffwechselprofilen detektieren. (3) Weiterhin werden die GGMs genutzt, um biologisch sinnvolle Metabolitengruppen zu definieren. Da wir zuvor zeigen konnten, dass partielle Korrelationen eine direkte Verbindung zu biochemischen Reaktionen haben, wird ein entsprechendes Clustering der GGM Netzwerke folglich Gruppen von biologisch verwandten Molekülen produzieren.

Im letzten Ergebniskapitel beschäftigen wir uns mit Independent Component Analysis (ICA). Kovarianz-basierte Methoden wie GGMs können statistische Abhängigkeiten höherer Ordnung, welche zusätzliche Informationen über die zugrundeliegenden Zusammenhänge liefern könnten, nicht erfassen. Wir stellen einen bayesschen ICA Ansatz mit Fehlerterm vor und diskutieren die Anwendung auf die Blut-Metabolomicsdaten. Die geschätzten Independent Components weisen starke Effekte von bekannten Stoffwechselwegen, wie zum Beispiel dem Aminosäurestoffwechsel, dem Lipidstoffwechsel, oder dem Energiestoffwechsel, auf. Weiterhin können wir zeigen, dass die Stärke einer bestimmten Independent Component in den Probanden (welche primär verzweigtkettige Aminosäuren beinhaltet) stärkere Assoziationen mit Plasma HDL Konzentrationen aufweist als die reinen Metabolitenkonzentrationen.

Zusammenfassend konnten wir zeigen, dass GGMs und ICA tatsächlich Teile von Stoffwechselwegen aus Hochdurchsatz-Datensätzen rekonstruieren können. Von besonderem Interesse ist hierbei, dass diese Ergebnisse aus Metabolomicsdaten von menschlichem

Blut gewonnen werden konnten. Metaboliten im Blut scheinen daher nicht lediglich Transport- und Abfallprodukte von metabolisch aktiven Organen zu sein, welche sich im Gefäßsystem wiederfinden, sondern enthalten vielmehr einen systematischen Abdruck der zugrundliegenden Stoffwechselwege. In der vorliegenden Arbeit wurden neue Erkenntnisse und bioinformatische Methoden zur Analyse der statistischen Zusammenhänge zwischen Metaboliten vorgestellt, welche das bisherige Wissen über den menschlichen Metabolismus erweitern.

Scientific publications

The following list shows peer-reviewed publications and patents relevant for each chapter of this thesis.

Chapter 1

- ⋆ **Krumsiek, J.**, Stückler, F., Kastenmüller, G., and Theis, F.J. Systems Biology meets Metabolism. In K. Suhre, editor, *Genetics Meets Metabolomics.* Springer, 2012.

Chapters 4 & 5

- ⋆ **Krumsiek, J.**, Suhre, K., Illig, T., Adamski, J., and Theis, F.J. Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data. *BMC Syst Biol*, 5(1):21, 2011

Chapter 6

- ⋆ **Krumsiek, J.**, Suhre, K., Evans, A.M., Mitchell, M.W., Mohney, R.P., Milburn, M.V., Wägele, B., Römisch-Margl, W., Illig, T., Adamski, J., Gieger, C., Theis, F.J., and Kastenmüller, G. Mining the unknown: A systems approach to metabolite identification. *PLoS Genetics*, 8(10):e1003005, 2012.
- ⋆ Identity Elucidation of Unknown Metabolites. U.S. Patent Application No. 61503673 Unpublished, filing date Jul. 1, 2011. (Michael Milburn, applicant)

Chapter 7

- ⋆ Mittelstrass, K., Ried, J.S., Yu, Z., **Krumsiek, J.**, Gieger, C., Prehn, C., Roemisch-Margl, W., Polonikov, A., Peters, A., Theis, F.J., Meitinger, T., Kronenberg, F., Weidinger, S., Wichmann, H.E., Suhre, K., Wang-Sattler, R., Adamski, J., and Illig, T. Discovery of Sexual Dimorphisms in Metabolic and Genetic Biomarkers. *PLoS Genetics*, 7(8):e1002215, 2011.
- ⋆ Jourdan, C., Petersen, A.K., Gieger, C., Döring, A., Illig, T., Wang-Sattler, R., Meisinger, C., Peters, A., Adamski, J., Prehn, C., Suhre, K., Altmaier, E., Kastenmüller, G., Römisch-Margl, W., Theis, F.J., **Krumsiek, J.**, Wichmann, H.E., and Linseisen, J. Association between Fat Free Mass and Serum Metabolite Profile in a Population-Based Study at Two Points in Time. *PLoS ONE*, 7(6):e40009, 2012.
- ⋆ Altmaier, E., Emeny, R., **Krumsiek, J.**, Lacruz, E., Lukaschek, K., Haefner, S., Kastenmüller, G., Römisch-Margl, W., Prehn, C., Mohney, R.P., Milburn, M.V., Illig, T., Adamski, J., Theis, F.J., Suhre, K., and Ladwig, K.H. Metabolomic profiles in individuals with negative affectivity and social inhibition: a population-based study of Type D personality. *Psychoneuroendocrinology*, in press.

- ⋆ Mueller, N.S., **Krumsiek, J.**, Theis, F.J., Böhm, C., and Meyer-Baese, A. Gaussian graphical modeling reveals specific lipid correlations in glioblastoma cells. volume 8058, page 805819. SPIE, 2011.
- ⋆ Petersen, A.K., **Krumsiek, J.**, Wägele, B., Theis, F.J., Wichmann, H.E., Gieger, C., and Suhre, K. On the hypothesis-free testing of metabolite ratios in genome-wide and metabolome-wide association studies. *BMC Bioinformatics*, 13:120, 2012.

Chapter 8

- ⋆ **Krumsiek, J.**, Suhre, K., Illig, T., Adamski, J., and Theis, F.J. Bayesian Independent Component Analysis recovers pathway signatures from blood metabolomics data. *Journal of Proteome Research*, 11(8):41204131, 2012.

Further publications

During the course of my PhD student time, I was involved in further projects which are not specifically discussed in this thesis.

The following publications represent collaboration projects in the metabolomics field besides Gaussian graphical modeling:

- ⋆ Gutch, H., **Krumsiek, J.**, and Theis, F. An ISA Algorithm With Unknown Group Sizes Identifies Meaningful Clusters in Metabolomics Data. EURASIP, 2011.
- ⋆ Krug, S., Kastenmüller, G., Stückler, F., Rist, M.J., Skurk, T., Sailer, M., Raffler, J., Römisch-Margl, W., Adamski, J., Prehn, C., Frank, T., Engel, K.H., Hofmann, T., Luy, B., Zimmermann, R., Moritz, F., Schmitt-Kopplin, P., **Krumsiek, J.**, Kremer, W., Huber, F., Oeh, U., Theis, F.J., Szymczak, W., Hauner, H., Suhre, K., and Daniel, H. The dynamic range of the human metabolome revealed by challenges. *FASEB J*, 2012.
- ⋆ Köttgen, A.*, Albrecht, E.*, Teumer, A.*, Vitart, V.*, **Krumsiek, J.***, GUGC Consortium, Ciullo, M., Fox, C., Caulfield, M., Bochud, M., and Gieger, C. Genome-wide association analyses identify 18 new loci associated with serum urate concentrations. *Nature Genetics*, in press.
- ⋆ Winkler, C., Lempainen, J., **Krumsiek, J.**, Achenbach, P., Grallert, H., Giannopoulou, E., Bunk, M., Theis, F.J., Bonifacio, E., and Ziegler, A.G. A strategy for combining minor genetic susceptibility genes to improve prediction of disease in type 1 diabetes. *Genes and Immunity*, 13:549-555, 2012.

We wrote two papers as follow-up studies of my diploma thesis about the computational modeling of hematopoietic differentiation:

⋆ **Krumsiek, J.***, Marr, C.*, Schroeder, T., and Theis, F.J. Hierarchical differentiation of myeloid progenitors is encoded in the transcription factor network. *PLoS ONE*, 6(8):e22649, 2011.

⋆ Schwarzfischer, M., Marr, C., **Krumsiek, J.**, Hoppe, P.S., Schroeder, T., and Theis, F.J. Efficient fluorescence image normalization for time lapse movies. In *ICSB 2011 workshop: Microscopic Image Analysis with Applications in Biology*. 2011.

* = equal contributions

Finally, there were several side projects in the group of Computational Modeling in Biology at our institute:

⋆ Wittmann, D.M., **Krumsiek, J.**, Saez-Rodriguez, J., Lauffenburger, D.A., Klamt, S., and Theis, F.J. Transforming Boolean models to continuous models: methodology and application to T-cell receptor signaling. *BMC Syst Biol*, 3:98, 2009.

⋆ Lutter, D., Marr, C., **Krumsiek, J.**, Lang, E.W., and Theis, F.J. Intronic microRNAs support their host genes by mediating synergistic and antagonistic regulatory effects. *BMC Genomics*, 11:224, 2010.

⋆ **Krumsiek, J.**, Pölsterl, S., Wittmann, D.M., and Theis, F.J. Odefy – from discrete to continuous models. *BMC Bioinformatics*, 11:233, 2010.

⋆ **Krumsiek, J.**, Wittmann, D.M., and Theis, F.J. From Discrete to Continuous Gene Regulation Models A Tutorial Using the Odefy Toolbox. In *Applications of MATLAB in Science and Engineering*. 2011.

Contents

1 Introduction **1**

1.1 Metabolomics 2

1.2 Metabolism and Systems Biology 5

1.3 Bioinformatics resources for metabolic research 7

1.4 Exploiting biological variation 8

1.5 Indirect effects 11

1.6 Research questions 13

1.7 Overview of this thesis 14

2 Materials **17**

3 Gaussian graphical models (GGMs) **21**

3.1 Independence, conditional distributions and conditional independence . 21

3.2 Second moment-based statistics 23

3.3 Partial correlations 25

3.4 From correlations to graphical models 28

3.5 The small n, large p problem 31

4 GGMs on computer-simulated reaction systems **37**

4.1 Modeling metabolic reaction systems by ODEs 38

4.2 GGMs reconstruct direct relationships 39

4.3 Reconstruction stability 42

4.4 Input noise 46

4.5 Enzyme kinetic rate laws only marginally affect GGM estimation 48

4.6 Negative feedback might compromise discrimination 49
4.7 All systems exhibit unique steady states 51
4.8 Conclusion 52

5 GGMs on metabolomics data 55
5.1 Network modularity calculation 56
5.2 Network properties 58
5.3 GGM edges represent pathway interactions 61
5.4 A fatty acid pathway model 64
5.5 Directly and indirectly connected metabolites 66
5.6 Low-order partial correlations 68
5.7 Conclusion 70

6 Mining the unknown 73
6.1 Methods 77
6.2 Genetic associations 78
6.3 GGM with unknown metabolites 81
6.4 Combining GGMs and GWAS 86
6.5 Experimental validation 89
6.6 Discussion of novel genetic associations 95
6.7 Conclusion 97

7 The potential of metabolomics GGMs 99
7.1 Effect networks 100
7.1.1 Gender-specific differences of blood metabolites 101
7.1.2 Metabolome associations with fat free mass 106
7.1.3 Effects of Type D personality on the metabolome 109
7.2 Metabolism in glioblastoma cells 112
7.3 Phenotype set enrichment analysis 116
7.4 Metabolite ratios, genetic networks and GGMs 118

8 Higher-order dependencies 123

8.1 Methods . . . 127
8.2 Bayesian noisy ICA on metabolomics data . . . 129
8.3 Manual investigation of independent components . . . 131
8.4 Systematic analysis and statistical enrichment . . . 134
8.5 Comparison with PCA and k-means clustering . . . 136
8.6 Analyzing the mixing matrix — associations with HDL . . . 138
8.7 Conclusion . . . 141

9 Summary & Outlook 143

Chapter 1

Introduction

The investigation of human metabolism, and particularly the investigation of metabolic disorders, are among the oldest research fields of mankind. For instance, diabetes mellitus was already recognized by the Egyptians around 1500 BC as a disorder of ‘too great emptying of the urine’ [1]. Indians termed this disease ‘honey urine’ due to the observation that the urine of affected individuals attracted ants and flies. While obviously the Indians could not determine the mechanistic reasons for their finding, this urine test can be regarded as one of the first occurrences of an empirically determined disease biomarker. It was not before the discovery of enzymes at the end of the 19th century [2] when scientists were allowed to gain direct mechanistic insights into metabolic processes for the first time. This marked the beginning of the field of biochemistry, giving rise to a rapid development of experimental methods to monitor biochemical processes. With the possibility to determine precise concentrations of a substance in a given biosample, researchers started to collect molecular biomarkers for various pathological states. A classical example from modern medicine is the case of the phenylketonuria (PKU) disorder. The most common form of PKU is caused by a loss-of-function mutation of *phenylalanine hydroxylase*, an enzyme responsible for the conversion of phenylalanine to tyrosine [3]. PKU is nowadays readily detected in newborn screenings by an increased phenylalanine-to-tyrosine ratio in the blood. Hence, in addition to a mere biomarker of the disease, this represents an early example of a metabolic readout that is directly linked to the respective underlying pathway mechanism. Furthermore, PKU demonstrates a well-defined interplay between the genetic makeup of individuals and their metabolism. Novel experimental techniques nowadays shift the focus from the selected investigation of specific phenotypes to large-scale metabolic screenings of many individuals and multiple disease phenotypes. The main focus of this thesis will be the statistical analysis of large sets of metabolic markers in human population cohorts, and the thorough investigation of the biochemical relationships between these markers. Furthermore, we will demonstrate how to use these relationships to analyze phenotypic traits, e.g. a disease state, the gender, or the body fat content, in a human population.

1.1 Metabolomics: the new field of large-scale small molecule measurements

With the advent of advanced measurement methods for small molecules at the end of the 20th century, the new field of *metabolomics* was arising. Its goal is to measure ideally all endogenous 'metabolites', i.e. metabolic intermediates like sugars, amino acids and fatty acids, in a given biological sample [4, 5]. The term metabolomics was first mentioned independently by Tweeddale et al. [6] and Oliver et al. [7] in 1998. They referred to a 'global metabolite pool (metabolome) analysis', thereby taking up the popular trend of 'omics' words for whole-system measurements. Metabolomics analyses are predominantly performed using either mass spectrometry (MS) or nuclear magnetic resonance spectroscopy (NMR) [8–10]. We will not go into the technical details of these measurement techniques here, since in this work we rather analyze the final concentration data than dealing with experimental particularities. Specific details on the MS-based identification of metabolites in a heterogeneous sample will be given in Chapter 6.

Considering the information flow in biological systems, from DNA to RNA, proteins and enzymes which finally act on the metabolites (Figure 1.1), the metabolome provides a readout of the integrated response of cellular processes to genetic and environmental factors [8]. It has therefore also been referred to as the 'link between genotype and phenotype' [11]. It is to be noted that this can be seen as both an advantage and as a pitfall of metabolomics measurements. On the one hand, metabolic profiles cover a wide range of effects which allow us to capture genetic effects, health and disease states, and nutritional habits. On the other hand, this heterogeneity might render the determination of the possible sources of an effect hard to impossible. Nevertheless, important insights into metabolism on both physiological and cellular scale have been gained in the past few years.

Similar to transcriptomics and proteomics, possible applications of metabolomics approaches are manifold. An economically important branch, which was among the first

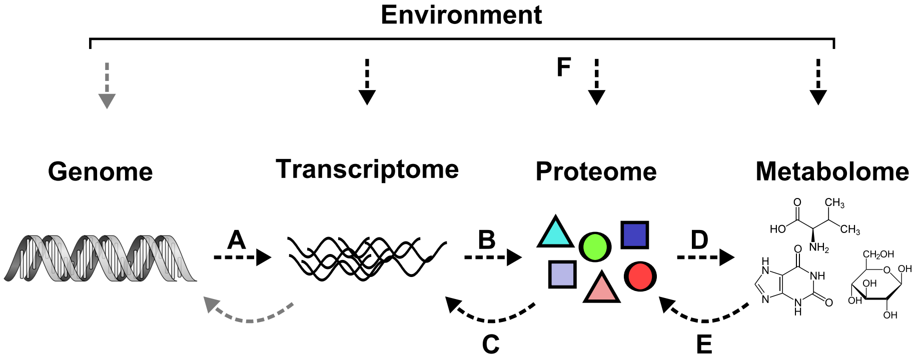


Figure 1.1: Metabolomics in the context of biological information flow. RNAs are transcribed from the genomic DNA (**A**), which are then translated to proteins (**B**). Signaling cascades and transcription factors regulate transcriptional activity (**C**). Metabolic enzymes and transporters drive the biochemical pathways and thus directly affect metabolite levels (**D**), which can in turn act as activity regulators on the proteins (**E**). Transcriptome, proteome and metabolome are influenced by environmental factors, which include nutritional effects, health states, life style, and environmental exposure. (**F**). Gray arrows represent effects from the environment on the genome (through mutations) or from the transcriptome (through reverse transcription). These mechanisms, however, can be considered relevant only on evolutionary scales or during specific pathological processes.

to use this new technology, is the field of plant research. Studies go from fundamental genome-metabolome interactions [12], over plant organ-specific metabolic investigations [13] and host-pathogen interactions [14], to the quality assessment of wine [15]. Independent of a specific taxonomic branch, metabolomics measurements can be used to explore basic cell biological mechanisms. For instance, Fendt et al. [16] investigated optimal enzyme concentration ranges that maintain metabolic homeostasis. Prominent applications of metabolomics in human physiology are nutritional interventions, where the intake of, for instance, sugar-rich or lipid-rich diets, might reveal metabolic system properties not visible in the resting state alone [17, 18]. Such studies are usually performed on small- to medium-sized groups of study probands.

A particularly important study type which has attracted wide interest in the past few years are large-scale population studies in epidemiological settings. Due to substantial technical advances in all ‘omics’ fields, it is now possible to obtain (mostly blood) metabolite profiles, genotypes, transcriptomics and proteomics measurements for thousands of study participants. While certainly the statistical power for a random sample from the general population is substantially lower than for classical case-control experiments, such a dataset can be applied to a variety of different research questions. For instance, population-based metabolomics analyses yielded biomarkers for diverse disease phenotypes, including eating disorders [19], osteoarthritis [20] and diabetes type II [21, 22]. Moreover, due to the generic nature of population studies, questions like the impact of smoking on the metabolome [23], coffee consumption [24] or gender-specific differences [25] could be answered on the same data sets.

The combination of genome-wide association studies (GWAS) with large-scale metabolomics measurements is a promising new approach [26–29]. GWAS search for statistical associations between a given phenotypic trait, like a disease state or a metabolite concentration, and genetic variation in a population cohort. Such studies identified a series of novel genetic loci that could be associated with human metabolic individuality.

An intuitive but important finding was the link between genetic variation in metabolic transporters and enzymes with functionally associated metabolites. For instance, genetic variation in a locus coding for an enzyme might lead to concentration changes of substrates or products of the reactions the enzyme catalyzes. Furthermore, as discussed above, metabolites are particularly interesting traits for GWAS, since the metabolome represents an integrated phenotype. It is influenced by all regulatory layers, from genetic mutations to nutrition-induced modulation of the metabolism. A comprehensive genetic analysis like a GWAS is only possible due to the large sample sizes provided by population-based studies. High-throughput genotyping methods currently only capture genetic variation with a high frequency in the population (minor allele frequency of at least 5%). The effects of such common variants on phenotypic traits have been shown to be rather small and thus only detectable with highly powered statistical analyses [30]. The reason for this two-fold: First, if a genetic variant excerts a strong phenotypic effect, it will not be a common variant due natural selection mechanisms [31]. Second, rare variants that show a strong effect will usually vanish in the overall population if not specifically selected for. Capturing and analyzing rare variants is only possible with extended profiling methods, like next-generation sequencing as shown in the 1,000 genomes project [32].

In this thesis, we primarily focus on metabolomics and genotyping data from the KORA cohort [33], in combination with several phenotypes like gender or a disease state. KORA (Kooperative Gesundheitsforschung in der Region Augsburg) is a research platform in southern Germany with a primary focus on cardiovascular diseases, diabetes mellitus type 2, and genetic epidemiology. The KORA cohort provides data on several thousand participants with metabolomics measurements on diverse platforms, genotyping data, transcriptomics, as well as a questionnaire-based survey of medication, disease state, nutritional habits, life style parameters, clinical chemistry, and basic anthropometric parameters. A more detailed introduction on the KORA cohort will be given in Chapter 2.

1.2 Metabolism and Systems Biology

Despite the tremendous progress in both biochemical research and later in the high-throughput measurement of metabolites, understanding the functional relationships between metabolite concentrations and physiological traits remains a challenging task. Metabolic research was therefore early on combined with ideas of systems level analysis, which later lead to the field of systems biology. In fact, metabolic pathways were among the first systems from molecular biology where rigorous mathematical modeling was applied. The most famous early metabolic modeling framework is certainly *metabolic control analysis* (MCA), which was developed by Kacser and Burns [34] in 1973. Originally referring to 'The Control of Flux', the authors developed a specific type of sensitivity analysis, which investigates the impact of changes in dynamic parameters on certain properties of the system (i.e. steady state concentrations or molecule oscillation amplitudes). Today there are hundreds of scientific publications using or building upon the MCA approach, ranging from drug target discovery [35] over plant metabolism [36] to biotechnological engineering [37].

Another branch of systems analyses in metabolic systems, mainly inspired by biotechnological research in microorganisms, was *constraint-based modeling* [38]. This methodological framework originally works on a list of biochemical reactions along with the respective stoichiometry of each substrate and product. It introduced the concept of metabolic *flux*, i.e. the number of molecules flowing through each reaction per unit of time. The central assumption is a constant equilibrium of internal metabolites in a system. Enzymatic reactions are considered to be fast in comparison to the physiological or chemical changes that drive the system from the outside. Consequently, the system is assumed to be in steady state: despite constant mass flow through the system, the actual metabolite concentrations remain unchanged. By only considering combinations which maintain this required steady state, the number of possible flux distributions in the system is drastically reduced. There are numerous applications of the constraint-based metabolic modeling approach. For example, it was used to consolidate and refine genome-scale metabolic network reconstructions, to predict minimal growth media, to determine robustness of metabolic networks, and to find optimal flux distributions for bacterial growth [39]. It is important to note that all methods which employ the constraint-based modeling approach do not take into account actual molecule concentrations.

With the availability of metabolomics datasets, new systems biological approaches were developed that included metabolite concentration data into the analysis. Classical dynamic modeling and parameter fitting could then be applied to metabolic systems. For example, Gupta et al. [40] derived a dynamic model of ceramide (a specific type of sphingolipid) biosynthesis in activated macrophages. Time-course metabolomics and transcriptomics data were used for parameter calibration of the model. The fitted parameters are then discussed to gain further insights into the system, in this case e.g. the apparently sub-maximal activity of certain enzymes in the pathway. In another study, we [18] developed a simplified model of fatty acid β-oxidation based on fasting time-course metabolomics data in 15 healthy subjects. The estimated model parameters could then be shown to improve statistical associations with anthropometric parameters in comparison to the raw metabolite concentrations.

The majority of systems biological studies based on metabolomics data does not include a dynamic modeling component. They primarily focus on multivariate statistical methods for high-dimensional data analysis, coupled with a systematic knowledge-based result evaluation. For example, Hirai et al. [41] projected changes of glucosinolate metabolism in *Arabidopsis thaliana* to known metabolic pathways. This allowed the authors to detect specific metabolic responses to sulfur and nitrogen deficiency. As another example, Xiao et al. [42] performed singular value decomposition on metabolomics data from the prefrontal cortex, and subsequently determined which known metabolic pathways displayed significantly changed metabolite concentrations upon drug treatment. The model class proposed for metabolomics analysis in this thesis, Gaussian graphical models, also represents a member of this group of systems biological approaches.

It is important to acknowledge that systems biological models always represent an abstraction of the actual underlying mechanisms. The real biological system is obviously more complicated than suggested by a formalized model. There are numerous processes and general aspects, like physiological and cellular compartmentalization, transport mechanisms, certain thermodynamical constraints and external factors that can either not be observed or are too complex to be directly included in a model. Nevertheless, if we are aware of this abstraction and carefully interpret the results produced by a systems biological model, we may gain valuable insights into the underlying biological system.

1.3 Bioinformatics resources for metabolic research

The evaluation of high-throughput data on a systematic scale heavily relies on publically available pathway databases. Since the main concept of a systems biological analysis is the automatic analysis of an entire dataset, the respective biological knowledge going into the analysis must be represented in a computer-readable format as well. Currently, public databases usually focus on a specific subset of biochemical interactions. For instance, the popular KEGG database [43] collects metabolic pathways and several signaling pathways, whereas STRING [44] captures protein-protein interactions of various types. Several projects attempt systematic integration of various molecule interaction databases, such as ConsensusPathDB [45] – which however does not grant full access to the underlying data – or the commercial Ingenuity Pathway Analysis software (www.ingenuity.com). To the best of our knowledge, a free and comprehensive database including multiple types of biological interactions has not yet been published.

It is furthermore important to acknowledge that all databases will show a substantial amount of both false positives and false negatives due to misannotations and missing experiments. Even more severe, a strong research bias can be expected for all datasets, where well-studied biological pathways have a better coverage than less studied ones. Moreover, the setup of organism- or even tissue-specific pathways sets is far from trivial. For example, it was not before 2002 when metabolic reconstruction for a specific cell type, the human erythrocyte, was published [46]. For more complex metabolic systems, like the human hepatocyte, the first metabolic reconstructions were published within the last years [47].

In this thesis, we will make use of appropriate, mostly metabolic pathway databases. However, we have seen that we cannot (yet) consider any set of interactions derived from a public databases to be actually *complete*. Issues of incompleteness and bias always need to be kept in mind when performing systematic, knowledge-based data analyses and the subsequent result interpretations.

1.4 Exploiting biological variation

Metabolomics measurements from multiple biological samples usually contain a substantial amount of biological variation. This holds true for all biological domains, from biological replicates of bacterial colonies, over tissue samples from clonal mice populations, up to blood metabolomics samples from different subjects in a population cohort. On the biochemical level, metabolite concentrations are determined by a set of specific metabolic enzymes. Variabilities in enzyme activities, enzyme concentrations and metabolite exchange rates - induced by a continuous spectrum of metabolic states throughout measured samples - give rise to characteristic patterns in the metabolite profiles which are directly linked the to underlying biochemical reaction network. We will make use of the fact that metabolite concentrations do not represent independent signals in the data, but display strong correlations which are a direct consequence of the wiring of the underlying metabolic network. For example, if two molecules are connected through a biochemical reaction, then high concentrations of one metabolite will coincide with high concentrations of the other metabolite, and vice versa.

Furthermore, we assume that stronger variability in the data will lead to more profound statistical dependencies between metabolites (Figure 1.2A). A single snapshot of metabolite concentrations cannot provide any information about the wiring of the underlying network. Only if there is a substantial amount of biological variation in the measured dataset, there will be a statistically detectable footprint of the biochemical network in the data. In other words, measuring a biological system in heterogeneous, distinct metabolic states will reveal its biochemical wiring. Chapter 4 will provide concrete evidence for this hypothesis. Importantly, we assume the different biological samples to have *identical* underlying biochemical reaction systems. If strong differences in the underlying network are expected, a *differential* evaluation of statistical dependencies might be favorable (cf. Chapter 7.2).

Statistical relationships are commonly estimated using second-order dependencies like the correlation coefficient, a measure of the linear association between two entities. Recently, several studies attempted to elucidate the origins of such metabolite-metabolite correlations in metabolomics data. We will discuss two of these studies in the following.

An early example on how to systematically investigate variation in metabolic systems has been published by Steuer et al. [48] in 2003. The authors assumed stochastic fluctuations of metabolites inside and outside of cells which are in identical states otherwise

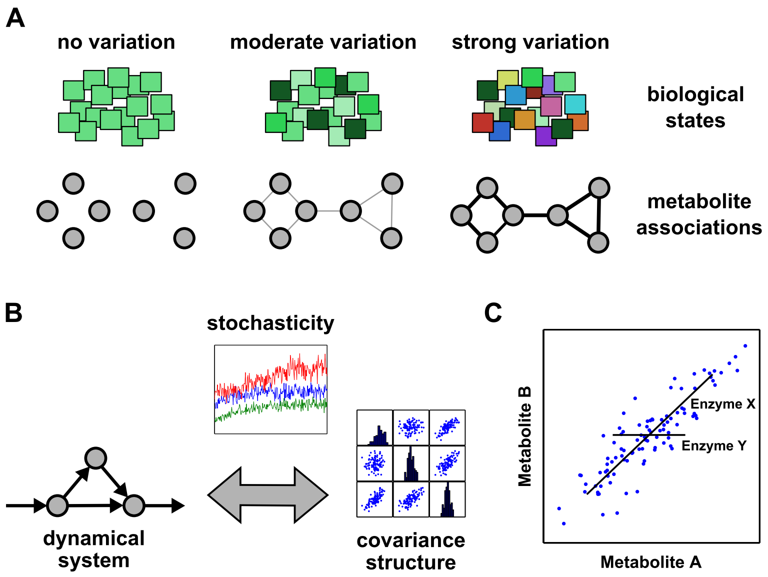


Figure 1.2: Connecting metabolite level variations with the underlying metabolic system. **A:** Stronger variation in samples with identical underlying biochemical networks will lead to a more profound establishment of statistical dependencies between metabolites. **B:** Steuer et al. [48] devised a mathematical framework based on stochastic differential equations, which establishes a direct connection between dynamical systems (represented as the corresponding Jacobian matrix) and the observed pairwise covariances. **C:** In a later study, Camacho et al. [49] explained covariance between metabolites using co-response profiles. Each enzyme introduces a specific direction of covariance between metabolites, the overlaying of which results in the finally observed correlation. Panels **B** and **C** were adapted from Krumsiek et al. [50].

(biological replicates). Two cells with an identical internal state will fall into qualitatively the same steady state after a given amount of time, but the actually measured steady state concentrations of biochemical molecules might differ slightly. The main contribution of the study was the derivation of a mathematical relationship between metabolite covariance and the Jacobian matrix of the underlying dynamical system (Figure 1.2B). The Jacobian matrix can be understood as a combination of the network topology with specific rates for each reaction. In this framework, given a metabolic network with given reaction rates, one can immediately derive the covariances between all pairs of metabolites. Moreover, given measured covariance values between metabolites, one can obtain information about the dynamics of the metabolic network acting on the metabolite pools. The paper provided a first link between variation in measured metabolite concentrations and properties of the underlying biochemical system.

A later study by Camacho et al. [49] shifted the focus from intrinsic fluctuations of the metabolite levels to actual differences in enzyme levels, thus directly affecting reaction rates in the system. This scenario can be termed *extrinsic* variation; the states between different cells actually differ and variations are not only due to stochastic fluctuations. The main methodological concept of this study was the investigation of so-called 'co-response profiles', which are related to the above-mentioned metabolic control analysis. For fixed enzyme concentrations, the system will fall into a single, unique steady state that can be represented as one dot in a 2D phase plane. Varying the concentration of one enzyme at a given time will create a co-response profile for this enzyme in a certain direction in metabolic space (solid lines in Figure 1.2C). The mixture of co-response profiles of all enzymes in the system then produces the co-variation we observe between metabolites (scatter plot in Figure 1.2C). The study thus provides a systematic definition of the origins of pairwise correlations in metabolomics data given changes in enzyme concentrations. Importantly, the paper also describes limitations of correlation-based approaches. For example, if co-response profiles of similar strength are orthogonal, the mutual covariance is canceled out and no correlation will be observed. Such issues have to be kept in mind when attempting to reconstruct metabolic reaction networks from steady state data.

In summary, both studies aimed to determine the origins of correlations on metabolomics data, but used conceptually different methodological approaches. While the Steuer et al. study focused on intrinsic, stochastic fluctuations of the metabolite levels, Camacho et al. studied the effects of varying enzyme levels.

In this thesis, we aim to find an approximation of the biological variation between human individuals in a population cohort. Our modeling approach (Chapter 4) represents a combination of both studies introduced above. Due to substantial differences in nutritional habits, lifestyle and the current metabolic state between individuals, we will allow for both variation in the enzyme concentrations as well as changes of metabolite concentrations outside of the modeled system. Each sampled data point then resembles one individual in the cohort. With respect to the illustration in Figure 1.2C, we are certainly in the *strong variation* scenario when analyzing population cohort metabolomics data.

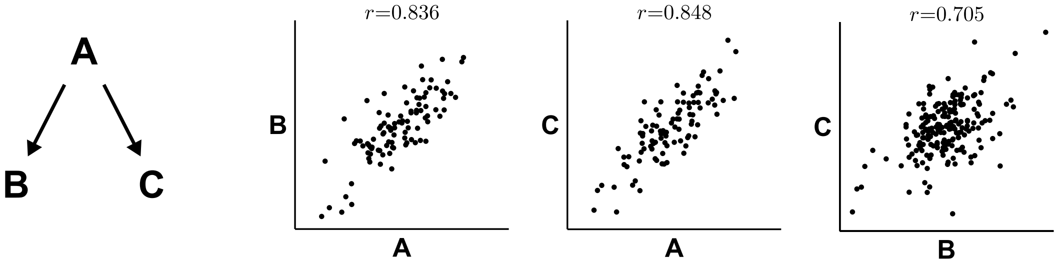


Figure 1.3: Example of indirect effects and spurious correlations. Since variable A coordinately affects B and C, the two variables will also be highly correlated, albeit not directly connected. r represents the Pearson correlation coefficient between the respective variables.

1.5 The issue of indirect effects, spurious correlations and false causality

A major drawback of correlation-based analysis is the inability to distinguish between direct and indirect associations. Correlation coefficients are generally high in large-scale *omics* data sets, suggesting a plethora of indirect and systemic associations. For example, transcriptional coregulation among many genes will give rise to indirect interaction effects in mRNA expression data [51]. Similar effects can be observed in metabolic systems which, in contrast to genetic networks, contain fast biochemical reactions in an open mass-flow system. Metabolite levels are supposed to be in quasi-steady state compared to the time scales of upstream regulatory processes [52]. That is, metabolites will follow changes in gene expression and physiological processes on the order of minutes and hours, but will appear unchanged on the order of seconds. These properties, even though substantially different from mRNA expression mechanisms, also give rise to indirect, system-wide correlations between distantly connected metabolites.

Unspecifically high correlations between two variables can usually be attributed to the presence of further variables that were not accounted for in the pairwise analysis. Consider the example with three variables shown in Figure 1.3, where by construction B and C are directly linked to A[1]. Given some variation in the data, A and B as well as A and C will be highly correlated – as expected. However, B and C will also be highly correlated due to the shared influence of A, even though they are not directly connected in the underlying network. A is called a *confounding factor* or *confounding variable* with respect to the correlation between B and C [53]. The effect created by such

[1] A mathematically precise formulation of this three-variable scenario will be given in Chapter 3. Briefly, B and C are set to A plus a certain amount of normal noise.

a confounding factor is then referred to as *spurious correlation.* Examples of spurious correlations can be found throughout all areas of life. For instance, investigating house fires in San Francisco, there is a profound correlation between the number of fire engines that were sent to a fire and the amount of damage the fire caused [54]. Obviously, it is not the firemen who do the damage (a falsely inferred causality), but rather a missing confounding factor, namely the actual size of the fire, which should have been taken into account. Analogously to the example in Figure 1.3, the size of the fire is the causal factor for both number of fire engine and damage of the fire.

Gaussian graphical models (GGMs) circumvent indirect association effects by evaluating *conditional* dependencies in multivariate Gaussian distributions [51]. A GGM is an undirected graph in which each edge represents the pairwise correlation between two variables conditioned against the correlations with all other variables (also denoted as *partial* correlation coefficients). GGMs have a simple interpretation in terms of linear regression techniques. When regressing two random variables A and B on the remaining variables in the data set, the partial correlation coefficient between A and B is given by the Pearson correlation of the residuals from both regressions. Intuitively speaking, we remove the (linear) effects of all other variables on A and B and compare the remaining signals. If the variables are still correlated, the correlation is directly determined by the association of A and B and not mediated by the other variables. A detailed introduction to GGMs will be given in Chapter 3. Partial correlations have recently been applied to biological data sets for the inference of association networks from mRNA expression data [55–58], and for the elucidation of relationships between genomic features in the human genome [59]. One previous study used partial correlations between genetic associations to elucidate genetically determined relations between metabolites [12].

Note that confounding effects are the major reason why correlation is never to be confused with causation. Causation will induce (some kind of) correlation, but whether or not a high correlation also represents a direct causative effect needs to be carefully evaluated [60]. This also holds true for the partial correlations, since further non-measured confounding factors might be present. Moreover, the directionality of causation cannot immediately be obtained from pairwise correlations. Nevertheless, when applied and evaluated appropriately, statistical association measures, and particularly partial correlations, can provide substantial insights into a biological system. This is the major focus of this thesis.

1.6 Research questions

The main goal of this thesis is to determine to which extent it is possible to recover footprints of biochemical pathways from metabolomics data. Specifically, we will focus on partial correlations and Gaussian graphical models on metabolomics data for the reconstruction of metabolic pathways. For gene regulatory systems, the connection between cellular processes and mRNA or protein correlations is often rather obvious – direct transcriptional activation or common regulators result in positive correlation, antagonistic processes lead to negative correlations. In contrast, for mass-flow systems like metabolism, the nature of pairwise correlations is far from trivial. Small changes at one point in the system might potentially propagate throughout the whole metabolic network, without actual regulatory changes in between. This thesis will investigate how biochemical reaction systems give rise to correlation structures of the respective metabolite concentrations.

We will then ask how the resulting GGMs can be used in biological applications. Having established metabolomics GGMs as a tool for the recovery of direct biochemical relationships, we can use these unbiased, data-driven metabolic networks for functional analysis. For instance, metabolomics GGMs can be used to further elucidate phenotypic differences (e.g. gender or a disease state) in the population, or to transfer functional classifications for insufficiently annotated metabolites.

Another important question specifically addressed in this thesis is the extent to which metabolic pathways are reflected in the human blood. Most applications presented here are based on metabolomics measurements from human serum samples. Blood can easily and uninvasively be obtained in large population cohorts, but it represents a heterogeneous mixture of nutritional effects, transport mechanisms and disposal processes of various organs and cell types. As discussed above, there are numerous studies which successfully linked blood markers with cellular metabolic processes (recall for instance the PKU disease). It has still been an open question, whether the impact of cellular metabolism on the blood metabolome is rather localized and sporadic, or whether there is a systematic signature of metabolic pathways in the blood. This thesis will give substantial evidence for the latter scenario.

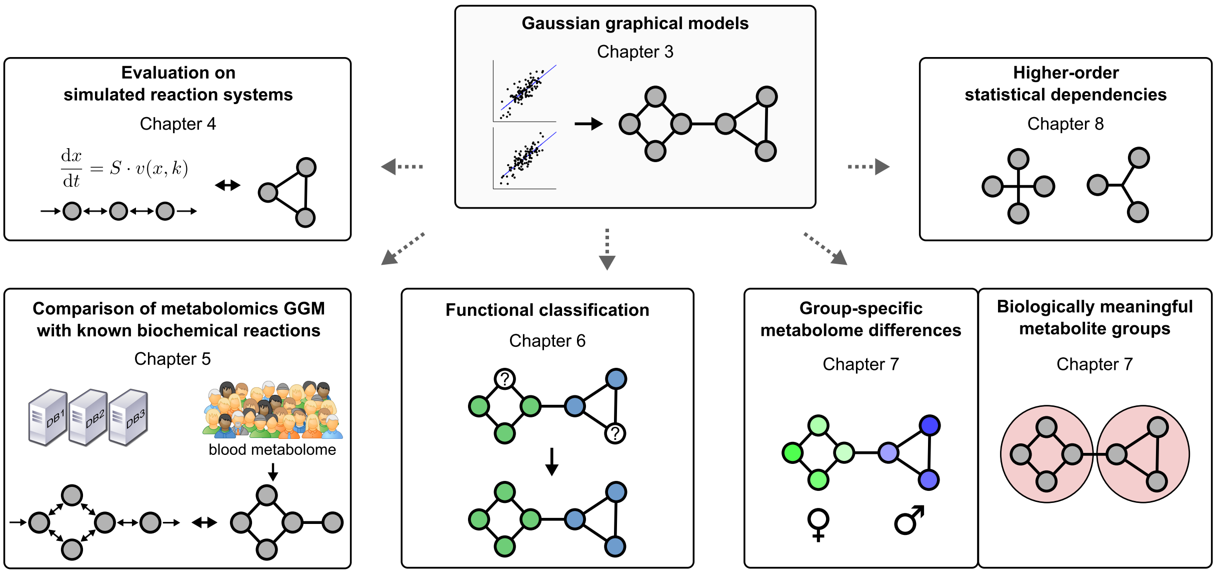

Gaussian graphical models
Chapter 3
Evaluation on simulated reaction systems
Chapter 4
$\frac{\mathrm{d}x}{\mathrm{d}t} = S \cdot v(x,k)$
Higher-order statistical dependencies
Chapter 8
Comparison of metabolomics GGM with known biochemical reactions
Chapter 5
DB1
DB2
DB3
blood metabolome
Functional classification
Chapter 6
?
?
Group-specific metabolome differences
Chapter 7
♀
♂
Biologically meaningful metabolite groups
Chapter 7

Figure 1.4: Overview of the thesis. Starting from Gaussian graphical models (GGMs) as a tool to elucidate the second-order conditional dependence structure of a dataset, we will discuss various applications and extensions of the approach. Chapter 3 introduces the mathematical backgrounds of GGMs. In Chapters 4 and 5, we evaluate GGMs as a tool to reconstruct biochemical reaction networks on computer-simulated systems and real metabolomics data from a population cohort. Chapter 6 then introduces an approach to combine GGMs with genetics data in order to provide functional classifications of unknown metabolites. In Chapter 7 we discuss applications of the GGM approach to elucidate group-specific metabolome differences (e.g. between males and females), and to determine biologically meaningful metabolite groups. Finally, Chapter 8 introduces independent component analysis (ICA) as a statistical tool which extends the covariance-based analysis of GGMs.

1.7 Overview of this thesis

In the following, we will briefly outline the content of this thesis. A graphical overview is given in **Figure 1.4**.

Chapter 2 introduces the KORA population and the different datasets used throughout this thesis. We will discuss the ‘Biocrates’ and ‘Metabolon’ metabolomics measurements as well as the genotyping data and further covariates.

In the introductory **Chapter 3**, we provide an overview of independence, conditional distributions, covariance, correlations, partial correlations and Gaussian graphical models. A detailed derivation of the connection between partial correlation coefficients and pairwise conditional independence for multivariate Gaussian distributions will be given.

We discuss various mathematical properties of GGMs and review published estimation algorithms.

Chapter 4 will then present an application of the GGM methodology to computer-simulated reaction systems. We model biological variation by a log-normal noise model on the metabolic reaction rates and create *in silico* metabolomics measurements by forward simulation and steady state determination. In most scenarios, a GGM properly reconstructs the correct network topology, while regular correlation coefficients fail to distinguish direct from indirect relationships.

In **Chapter 5**, we apply the GGM methodology to a human metabolomics dataset of 151 measured metabolites, most of which are lipid species, in 1020 fasting serum samples from the KORA F4 population. Applying both manual investigation and a systematic analysis of the resulting metabolomics GGM, we find that, to a significant extend, connected metabolites in the model indeed correspond to real biochemical reactions. This finding demonstrates GGMs as a suitable tool for the unbiased reconstruction of metabolic pathways from high-throughput metabolomics data.

Chapter 6 introduces a specific application of GGMs for the identification of unknown metabolites. Untargeted metabolomics measurements frequently generate signals where a certain substance can be reliably detected in the sample, but the precise biochemical identity of the compound remains to be elucidated. By combining the GGM methodology with genome-wide association studies, we are able to derive pathway predictions for a series of such *unknown* metabolites. For a number of cases, this even allows for a concrete pathway classification, which is then experimentally validated in the lab. Furthermore, we identify seven genetic loci that were previously unreported to associate with blood metabolite concentrations.

Several further applications of metabolomics GGMs to specific biological questions are then demonstrated in **Chapter 7**. In three projects, we integrate the GGMs with results from differential concentration analyses of gender-specific differences, influences of the fat-free body mass, and the type D personality on the metabolome. Using this specific combination of classical statistical methods with the network-based GGM approach ('effect networks'), we can pinpoint specific changes in the metabolic pathways for the respective phenotypic traits. Another application introduces the concept of *differential* GGMSs, which we use to delineate specific metabolic changes in a glioblastoma cell line under varying drug treatments. Finally, we will present two projects where GGMs were used to define biologically meaningful metabolite groups. In one project, GGMs were used to further validate a novel enrichment algorithm, the other project provides an in-depth analysis of metabolite *ratios* which have previously been shown to be particularly useful in genome-wide association studies.

In **Chapter 8** we extend the purely covariance-based analysis of metabolite dependencies by integrating higher-order statistical moments. Specifically, we use a Bayesian variant of independent component analysis on the KORA metabolomics data. We can show that the reconstructed statistically independent metabolite profiles contain strong signatures of specific metabolic pathways, including amino acid metabolism, lipid metabolism, and energy metabolism. Furthermore, the strength of a specific independent component in the study participants represents a strong biomarker for blood HLD (high density lipoprotein) levels.

The final **Chapter 9** will discuss the scientific contributions in the context of the field and discuss possible extensions and potential future projects.

Chapter 2

Materials

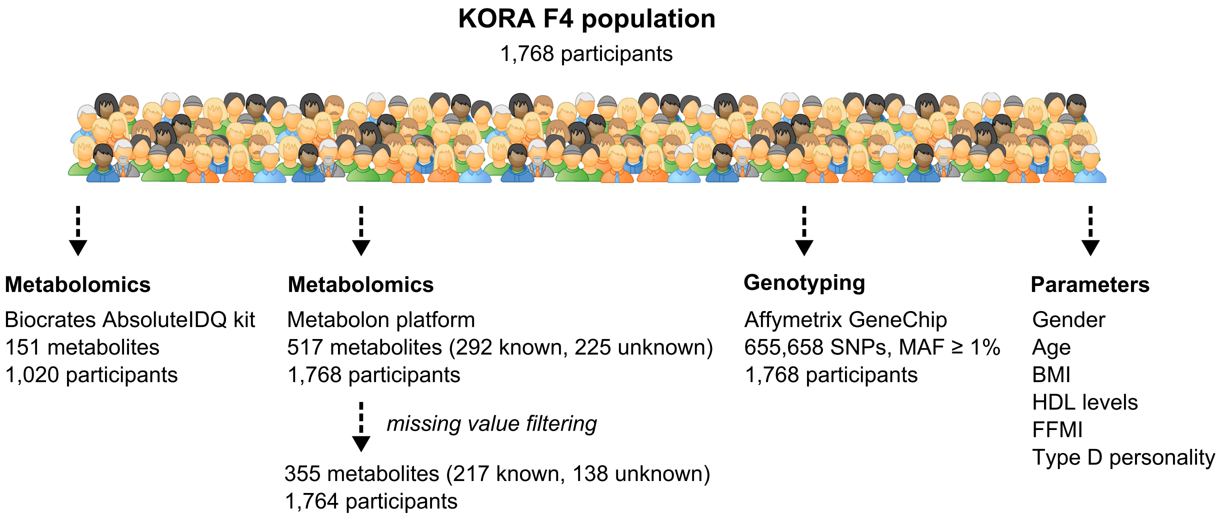


Figure 2.1: Datasets analyzed in this thesis. We used metabolomics data measured on two different experimental platforms, genotyping data and six general parameters from the KORA F4 population. The *missing value filtering* step is required in order to get a full data matrix for GGM calculation.

KORA (Kooperative Gesundheitsforschung in der Region Augsburg) is a research platform in southern Germany with a primary focus on cardiovascular diseases, diabetes mellitus type 2, and genetic epidemiology [33]. In four independent health surveys (termed S1 to S4) between 1984 and 2001, data from a total of 18,000 participants were collected. Two ten- and seven-year follow-up surveys for S3 and S4, termed F3 and F4, were conducted to introduce a longitudinal component into the study. During the visits

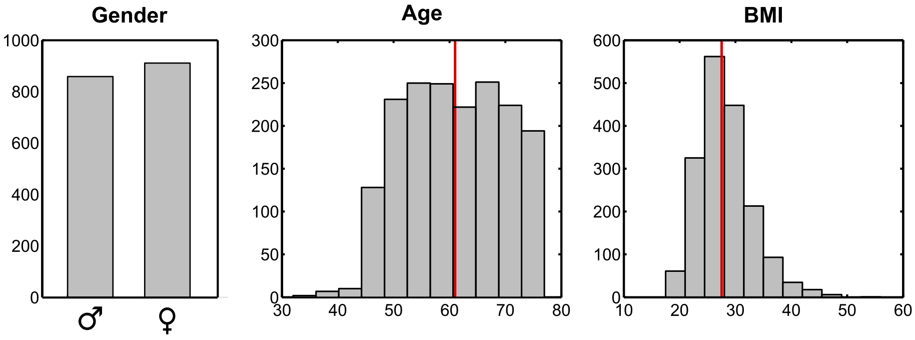


Figure 2.2: General characteristics of the 1,768 study participants from KORA F4 we mainly worked with in this thesis. Vertical red lines indicate median values.

information on medical history, risk factors and life style (smoking, physical activities, etc.), blood pressure, and anthropometric parameters (weight, body mass index, fat-free mass, height, etc.) were gathered. Furthermore, for S3, S4, F3 and F4, we now have genotyping data and metabolomics data from blood under fasting conditions.

In this thesis, we will primarily focus metabolomics and genotyping data from a subset of 1,768 participants of the F4 survey (Figure 2.1). Of these participants, 858 were male and 910 were female, the median age was 61, and the median body mass index (BMI) was 27.5 (Figure 2.2). In addition to gender, age and BMI, we will use three further parameters in our analyses: (1) Plasma high-density lipoprotein (HDL) levels, a lipid-carrying particle class in the blood. (2) The fat-free mass index (FFMI), a height-independent measure of fat-free mass based on the body fat percentage. (3) Information on a type D personality, which can be understood as a general liability to psychological distress [61].

Metabolomics data were measured on two different experimental platforms. First, a total of 151 metabolites were measured by electrospray ionization tandem mass spectrometry (ESI-MS/MS) with the Biocrates AbsoluteIDQ kit. Details on the experimental procedures can be found in Illig et al. [27]. The metabolite panel comprises 14 amino acids including 13 proteinogenic amino acids and ornithine; hexose (sugars with 6 carbon atoms, e.g. glucose and fructose); 23 acylcarnitines [Cx:y-carn] (with x carbon atoms and y double bonds), 7 hydroxy-acylcarnitines [Cx:y-OH-carn], 6 dicarboxy-acylcarnitines [Cx:y-DC-carn], and 2 methylated dicarboxy-acylcarnitines variants [Cx:y-M-DC-carn]; 9 sphingomyelins [SM Cx:y] and 5 hydroxy-sphingomyelins [SM Cx:y-OH]; and 87 phosphatidylcholines (PC). These glycerophospholipids are further subdivided with respect to the presence of ester and ether bonds of fatty acid residues with the glycerol moiety. The set contains 36 diacyl-PCs with two esterified fatty acid residues [PC aa Cx:y], 38 acyl-alkyl-PCs with one ether-bond at the sn-2 position [PC ae Cx:y] and 13 lyso-PCs with only one esterified fatty acid residue at the sn-1 or sn-2 position [lysoPC a Cx:y]. The mass spectrometry technology cannot distinguish between the side chains of diacyl-phospholipids. The measured compounds are thus associated with the sum of carbon atoms and double bounds for both fatty acid residues. We used a subset of 1,020 samples for this analysis, which represents the first batch of samples measured at our local metabolomics platform.

In addition to the Biocrates platform, serum samples were measured by Metabolon Inc., NC, USA. Briefly, metabolic profiling was done using ultrahigh-performance liquid-phase chromatography and gas-chromatography separation, coupled with tandem mass spectrometry. Details of the experimental procedures can be found in Suhre et al. [29]. The dataset contains a total of 292 known compounds and 225 unknown compounds. An unknown represents a reproducible signal in an untargeted metabolomics approach, whose precise biochemical identity has not been elucidated yet. Unknown metabolites and their functional characterization will be the main topic of Chapter 6.

In contrast to the lipid-centered Biocrates metabolite panel, the Metabolon panel covers a wide range of metabolic processes. The known metabolites are subdivided into eight ‘super-pathways’, including ‘Lipid’, ‘Carbohydrate’, ‘Amino acid’, ‘Xenobiotics’, ‘Nucleotide’, ‘Energy’, ‘Peptide’ and ‘Cofactors and vitamins’. In addition, each metabolite is associated with a more fine-grained ‘sub-pathway’ like ‘Oxidative phosphorylation’, ‘Carnitine metabolism’ or ‘Valine, leucine and isoleucine metabolism’. Fatty acid-based lipids are described by the number of carbon atoms, double bonds and, if applicable, position of the last double bond. For instance, ‘fatty acid 18:2(n-6)’ denotes a fatty acid with 18 carbon atoms and two double bonds, the last of which lies at the n-6 position (between carbon atom 12 and 13). Phospholipids are named by their headgroup and the fatty acids in both side chains. For example, PI(20:4(n-6)/0:0) represents a phosphatidylinositol containing an arachidonate residue (20 carbon atoms, four double bonds, n-6) at the sn-1 position. PC(0:0/18:0) contains a 18:0 fatty acid at the sn-2 position. Note that the current metabolite panel only measures lyso-phospholipids, that is phospholipids with only one fatty acid chain.

The Metabolon dataset contains a substantial amount of missing values (178,325 missing values out of 914,056 total values), which occur either due to measurement errors or signals below the detection limit. For the GGM calculation, we require a full data matrix without missing values. We therefore first excluded metabolites with more than 20% missing values, and then samples with more than 10% missing values. The filtered data matrix still contained n=1764 samples with 355 metabolites (217 knowns and 138 unknowns). Remaining missing values were imputed with the 'mice' R package [62].

SNP genotyping was carried out using the Affymetrix GeneChip array 6.0 [29]. For our analyses, we only considered autosomal SNPs passing the following criteria: call rate $> 95\%$, Hardy-Weinberg-Equilibrium p-value p(HWE) $> 10^{-6}$, minor allele frequency MAF $> 1\%$. In total, 655,658 SNPs were left after filtering. Genotypes are represented by 0, 1, and 2 for major allele homozygous, heterozygous, and minor allele homozygous with respect to the general population.

Chapter 3

Gaussian graphical models (GGMs)

This chapter introduces the basic concepts of independence, conditional distributions, covariance, correlation coefficients and the subsequent definition of partial correlation coefficients. We will discuss how this descriptive statistical measure gives rise to a specific probabilistic model, a *Gaussian graphical model* (GGM), and determine implications for the overall dependency structure between random variables. We will put a particular focus on *conditional* independence between two random variables, which has a direct relationship to the inverse of a covariance matrix in case of a multivariate Gaussian distribution. Several methods for the estimation of GGMs will be introduced, both for the well-defined case where we have more samples than variables, but also algorithms suitable for datasets with less samples than variables (the more common case in large-scale 'omics' analyses). The following sections lay the mathematical groundwork for Chapters 4 through 7.

3.1 Independence, conditional distributions and conditional independence

In the following, we will briefly introduce independence, conditional distributions and conditional independence. The concepts will be used throughout this chapter to estimate Gaussian graphical models from the covariance matrix of a random vector. For a detailed introduction to random variables and statistical fundamentals, we refer the interested reader to Grimmett and Stirzaker [63] or related literature. Let X and Y be two continuous random variables, $f_X(x), f_Y(y)$ their respective marginal probability density functions and $f_{X,Y}(x,y)$ their joint density. The variables X and Y are *independent* if and only if the joint density can be factorized as

$$f_{X,Y}(x,y) = f_X(x)f_Y(y). \tag{3.1}$$

Intuitively, independence implies that knowledge of the value of one variable provides no information about the value of the other variable. We write $X \perp\!\!\!\perp Y$ to denote independence between the two random variables X and Y.

The *conditional* probability density function of X given $Y = y$ is defined as

$$f_X(x \mid Y = y) := \frac{f_{X,Y}(x,y)}{f_Y(y)} \quad \forall\{y \mid f_Y(y) > 0\}. \tag{3.2}$$

It represents the distribution of X when Y takes a value of y. Comparing equations (3.1) and (3.2), the independence relation can be reformulated to $X \perp\!\!\!\perp Y \Leftrightarrow f_X(x \mid Y = y) = f_X(x)$. This notation directly reflects the above-mentioned concept that one variable does not influence the probability density of the other variable. The conditional probability density function for the joint distribution of two variables X and Y given a third variable Z is defined analogously as

$$f_{X,Y}(x,y \mid Z = z) := \frac{f_{X,Y,Z}(x,y,z)}{f_Z(z)} \quad \forall\{z \mid f_Z(z) > 0\}, \tag{3.3}$$

where $f_{X,Y,Z}(x,y,z)$ represents the joint density of X, Y and Z. Independence can then be extended to *conditional* independence of X and Y given a third variable Z:

$$(X \perp\!\!\!\perp Y) \mid Z \;\Leftrightarrow\; f_{X,Y}(x,y \mid Z = z) = f_X(x \mid Z = z)f_Y(y \mid Z = z) \quad \forall\{z \mid f_Z(z) > 0\}, \tag{3.4}$$

where $(X \perp\!\!\!\perp Y) \mid Z$ states that X and Y are conditionally independent given Z. Moreover, the concept can directly be generalized to more than one conditioning variable by introducing further variables into the conditional probability density function.

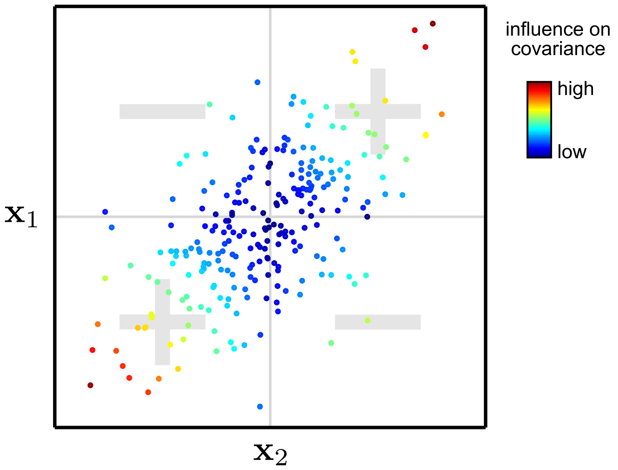


Figure 3.1: Illustration of covariance for bivariate normally distributed variables. Each quadrant contributes negatively or positively to the total covariance value. Colors indicate the strength of this effect. The plot shows 250 samples from a normal distribution with zero mean and covariance matrix $\Sigma = \left(\begin{smallmatrix} 1 & 0.65 \\ 0.65 & 1 \end{smallmatrix}\right)$.

3.2 Second moment-based statistics

Let $\mathbf{X} = (X_1, \ldots, X_p)$ be a p-dimensional random vector of continuous random variables with finite second moments. The covariance between two variables X_i and X_j with $i, j \in \{1, \ldots, p\}$ is defined as

$$\mathrm{Cov}(X_i, X_j) = \sigma_{ij} = \mathrm{E}[(X_i - \mathrm{E}[X_i])(X_j - \mathrm{E}[X_j])].$$

The matrix $\Sigma = (\sigma_{ij})$ of covariance values is referred to as the *covariance matrix*. Covariance provides a measure of the linear associations between the involved variables. Intuitively, positive values indicate that if values of X_i are above the mean, values of X_j tend to be above the mean as well, and vice versa (Figure 3.1). By normalizing the covariance with the respective standard deviations of the random variables, we obtain the population correlation coefficient:

$$\mathrm{Corr}(\mathrm{X_i}, \mathrm{X_j}) = \rho_{ij} = \frac{\mathrm{Cov}(X_i, X_j)}{\sqrt{\mathrm{Var}(X_i)}\sqrt{\mathrm{Var}(X_j)}}, \tag{3.5}$$

which is also referred to as the *Pearson product-moment correlation coefficient* [64]. Obviously, $-1 \leq \rho_{ij} \leq 1$, with 1 representing perfect linear correlation, and -1 perfect anticorrelation.

Now let $\mathbf{x} = (x_{ki}) \in \mathbb{R}^{n \times p}$ be a realization of the random vector with n samples, then the respective sample correlation coefficient r_{ij} is defined as

$$r_{ij} = \frac{\sum_{k=1}^{n} (x_{ki} - \overline{\mathbf{x}}_i)(x_{kj} - \overline{\mathbf{x}}_j)}{\sqrt{\sum_{k=1}^{n} (x_{ki} - \overline{\mathbf{x}}_i)^2} \sqrt{\sum_{k=1}^{n} (x_{kj} - \overline{\mathbf{x}}_j)^2}},$$

where $\overline{\mathbf{x}}_i = \frac{1}{n} \sum_{k=1}^{n} x_{ki}$ represents the sample mean value of the data column $\mathbf{x}_{\cdot i}$, for $i = 1, \ldots, p$. In the case of a normally distributed random vector $\mathbf{X}$, sample means $\overline{\mathbf{x}}_i$ represent the maximum likelihood (ML) estimators for the mean values $\mathrm{E}[X_i]$, the sample covariance $\frac{1}{n} \sum_{k=1}^{n} (x_{ki} - \overline{\mathbf{x}}_i)(x_{kj} - \overline{\mathbf{x}}_j)$ is the ML estimator for $\mathrm{Cov}(X_i, X_j)$ and, subsequently, r_{ij} represents the ML estimator for ρ_{ij} [65]. Note that in order to obtain an *unbiased* estimator for the covariance whose expected value precisely equals the covariance, the normalization term $\frac{1}{n-1}$ instead of $\frac{1}{n}$ needs be be applied. For the large sample sizes used our metabolomics analysis, the differences between both estimators will be marginal.

A particularly important property for this thesis is the connection between decorrelation and independence of two variables (see Section 3.1). In the multivariate Gaussian case, two variables are (marginally) independent if and only if they are uncorrelated [66]:

$$X_i \perp\!\!\!\perp X_j \; \Leftrightarrow \; \rho_{ij} = 0 \;\; \forall i, j \in \{1, \ldots, p\}. \tag{3.6}$$

Since the shape a normal distribution is fully parametrized by the covariance matrix, the marginal density of each variable is invariant to the value of the respective other variable if the pairwise covariance (or correlation) is zero, cf. equations (3.1) and (3.2). Note that for arbitrary distributions, only the direction $X_i \perp\!\!\!\perp X_j \Rightarrow \rho_{ij} = 0$ holds. Statistical independence always causes zero correlation, but uncorrelated variables might still be dependent. The same holds true if the two investigated variables are marginally normally distributed, but not jointly normally distributed [67].

A statistical test for non-zero correlation coefficients r_{ij} can be constructed using the Fisher transformation [68]. The transformation is defined as:

$$z(r_{ij}) = \frac{1}{2} \ln \left(\frac{1 + r_{ij}}{1 - r_{ij}} \right).$$

For a true correlation value of zero and multivariate normally distributed $\mathbf{X}$, the quantity z is approximately normally distributed with mean 0 and variance $\frac{1}{n-3}$. A two-sided p-value of r_{ij} being significantly different from zero can thus be obtained by

$$\text{p-val}(r_{ij}) = \left(1 - \phi\left(\sqrt{n-3} \cdot z(|r_{ij}|)\right)\right) \cdot 2, \tag{3.7}$$

where ϕ stands for the cumulative distribution function of the standard normal distribution. Note that this procedure can easily be adapted to test the correlation coefficient for difference from a non-zero value r_0. In this thesis, however, we are only interested in correlations being zero or not, i.e. absent or present statistical dependencies between the corresponding biological molecules.

3.3 Partial correlations

As already discussed in Chapter 1.5, a major drawback for the application of correlation-based measures in practice is their inability to distinguish between direct and indirect effects. Partial correlation coefficients are able to circumvent this problem by estimating the relationship between two variables conditioned against a given set of other variables. The simplest and most intuitive derivation of partial correlation coefficients is based on linear regression analysis. Let $V = \{1, \dots, p\}$ be the index set of all variables $\mathbf{x} \in \mathbb{R}^{n \times p}$, and $Q \subset V$ be an index subset with $1 \leq |Q| \leq p-2$. Let $\mathbf{x}_Q := (\mathbf{x}_i \mid i \in Q)$ be the corresponding subset of random variable realizations arranged in a matrix of column vectors. We regress two variables $\mathbf{x}_i$, $\mathbf{x}_j$ with $i, j \in V \setminus Q$ on $\mathbf{x}_Q$ separately. For example, the linear regression model of $\mathbf{x}_i$ on $\mathbf{x}_Q$ reads

$$\mathbf{x}_i = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \beta_{i0} + \mathbf{x}_Q \boldsymbol{\beta}_i + \epsilon_i,$$

where β_{i0} is the coefficient of the intercept, $\boldsymbol{\beta}_i = \left(\beta_{i1}, \dots, \beta_{i|Q|}\right)^T$ represents a column vector of regression coefficients and ϵ_i represents a normally distributed error with zero mean [69]. We then fit the coefficients as

$$\left(\hat{\beta}_{i0}, \hat{\boldsymbol{\beta}}_i\right) = \arg \min_{\beta_{i0}, \boldsymbol{\beta}_i} \sum_{k=1}^{n} \left(x_{ki} - \beta_{i0} - \sum_{l=1}^{|Q|} \beta_{il} \cdot (\mathbf{x}_Q)_{kl} \right)^2, \tag{3.8}$$

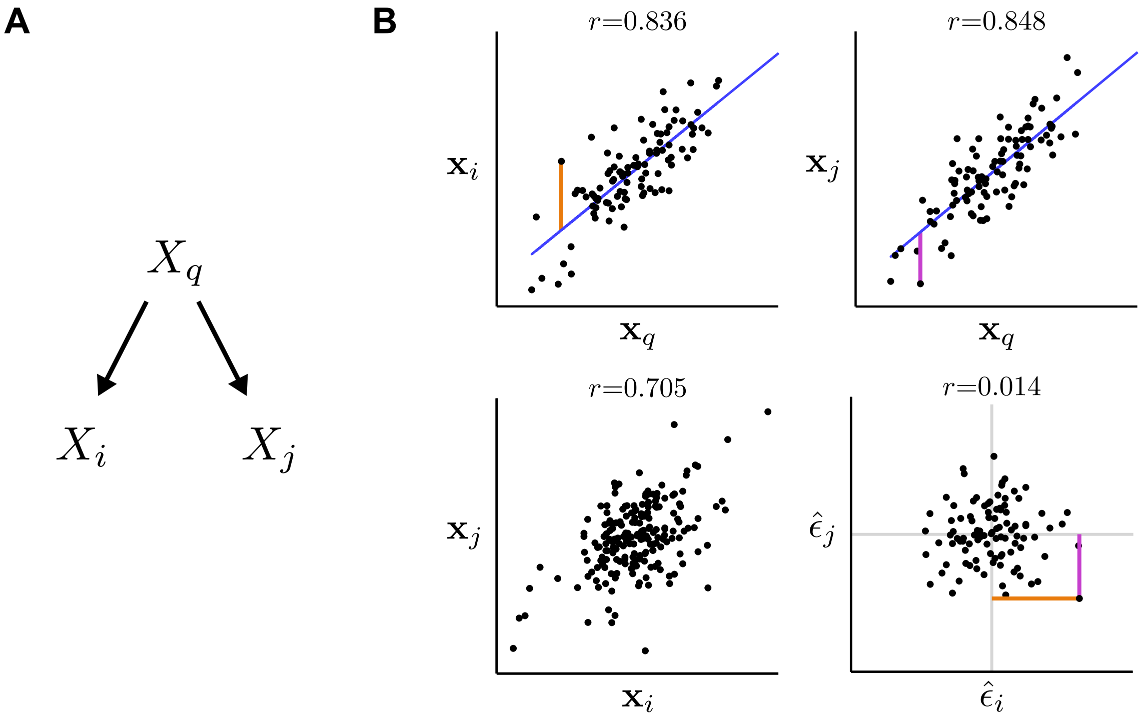


Figure 3.2: Derivation of partial correlation coefficients by linear regression. **A:** Influence diagram of three random variables X_q, X_i and X_j. The relationship was modeled as $X_q \sim \mathcal{N}(0,1)$ and $X_i := X_q + e_1,\ X_j := X_q + e_2$ with $e_1, e_2 \sim \mathcal{N}(0, 0.4)$. **B:** Realizations $\mathbf{x}_q$, $\mathbf{x}_i$ and $\mathbf{x}_j$ of the random variables with $n = 100$ samples. $\hat{\epsilon}_i$ and $\hat{\epsilon}_j$ represent the residuals from linear regression, see text. As expected, all three variables are strongly positively correlated. However, when comparing the residual vectors of the regression of $\mathbf{x}_i$ and $\mathbf{x}_j$ on $\mathbf{x}_q$ (bottom right), no remaining correlation can be detected. The variables X_i and X_j are thus uncorrelated given X_q. Colored lines mark the residuals for one selected data point.

where $\hat{\beta}_{i0}$ and $\hat{\boldsymbol{\beta}}_i$ are the least square error estimates of β_{i0} and $\boldsymbol{\beta}_i$, respectively, and $(\mathbf{x}_Q)_{kl}$ represents the entry of $\mathbf{x}_Q$ at the position k, l. The residuals are then defined as $\hat{\epsilon}_i = \sum_{k=1}^{n} \left(x_{ki} - \hat{\beta}_{i0} - \sum_{l=1}^{|Q|} \hat{\beta}_{il} \cdot (\mathbf{x}_Q)_{kl} \right)$. We regress $\mathbf{x}_j$ analogously, yielding a residual vector $\hat{\epsilon}_j$.

The $|Q| - order$ partial correlation coefficient $r_{ij|Q}$ of $\mathbf{x}_i$ and $\mathbf{x}_j$ given $\mathbf{x}_Q$ is then defined as the correlation between the respective residuals $\hat{\epsilon}_i$ and $\hat{\epsilon}_j$ from these regressions:

$$r_{ij|Q} = \mathrm{Corr}(\hat{\epsilon}_\mathrm{i}, \hat{\epsilon}_\mathrm{j})\,. \tag{3.9}$$

Figure 3.2 provides a detailed illustration of the idea for one conditioning variable. Note that the regression calculation is only possible if $\mathbf{x}_Q$ has full column rank, since otherwise the problem is ill-conditioned – we cannot estimate more independent coefficients in the model than we have data points. This issue will be discussed in more detail in Section 3.5.

For multivariate normally distributed **X**, partial correlations have a direct relationship to conditional distributions (see Section 3.1). Let $R \subset V$ and $Q \subset V$ be two distinct subsets of variables in **X**. Then the covariance matrix $\Sigma_{R|Q}$ (the covariance of $\mathbf{X}_R$ given $\mathbf{X}_Q$) of the conditional density is given by

$$\Sigma_{R|Q} = \Sigma_{RR} - \Sigma_{RQ}\Sigma_{QQ}^{-1}\Sigma_{QR}, \tag{3.10}$$

where Σ_{RQ} represents the covariance matrix reduced to the rows and columns of the index sets R and Q, respectively, and Σ_{QQ}^{-1} represents the inverse covariance matrix with according subsetting. Note that the covariance of this conditional density is independent of the actual value of the variables which are conditioned for. Now let $(\omega_{ij}) = \Omega = \Sigma^{-1}$ be the inverse covariance matrix, also called *precision matrix* or *concentration matrix* of **X**. Using properties of the inverse of partitioned matrices (Lauritzen [70], p.243f.), we obtain

$$\Omega_{RR}^{-1} = \Sigma_{RR} - \Sigma_{RQ}\Sigma_{QQ}^{-1}\Sigma_{QR}. \tag{3.11}$$

Comparing equations (3.10) and (3.11), we see that $\Sigma_{R|Q} = \Omega_{RR}^{-1}$. We now assume $R = \{i, j\}$ and $Q = V \setminus R$, i.e. we investigate the conditional distribution of two variables given all remaining variables. We can then reformulate this equality as

$$\Sigma_{R|Q} = \Omega_{RR}^{-1} = \frac{1}{\det \Omega_{RR}} \begin{pmatrix} \omega_{ii} & -\omega_{ij} \\ -\omega_{ji} & \omega_{jj} \end{pmatrix}. \tag{3.12}$$

Hence, the variables X_i and X_j are conditionally independent (zero covariance in the conditional distribution) if and only if the respective entry ω_{ij} in the inverse covariance matrix is zero. Similar to marginal independence, in the Gaussian case the value of one variable does not influence the conditional probability density of the respective other variable, see equation (3.4).

The full-order partial correlation coefficient matrix $Z = (\zeta_{ij}) := r_{ij|Q}$ can now be derived by a single matrix inversion step with subsequent normalization (Lauritzen [70], p.129f.):

$$Z = (\zeta_{ij}) = -\omega_{ij}/\sqrt{\omega_{ii}\omega_{jj}}. \tag{3.13}$$

The inversion is only well-defined if Σ has full rank, which almost always the case if the number of sample rows is equal to or larger than the number of variables p. This can easily be seen when expressing the covariance calculation as a matrix operation. Without loss of generality, assume the samples **x** to have zero mean (by definition, covariance and correlation are invariant under translation). Then the respective covariance matrix is given by $\mathrm{Cov}(\mathbf{x}) = \frac{1}{n}\mathbf{x}^T \cdot \mathbf{x}$. Since the rank of a matrix product is smaller than or equal to the ranks of the respective multiplied matrices, it follows that $\mathrm{rank}(\mathrm{Cov}(\mathbf{x})) \leq \mathrm{rank}(\mathbf{x})$.

The inverse covariance matrix Σ^{-1} was first introduced by Dempster [71] in 1972 as a parameterization of multivariate Gaussian distributions. The procedure was originally termed *covariance selection* and developed to reduce the number of parameters required to describe the distribution. Note that due to the normalization operation, we could also use the inverse Pearson correlation matrix instead of the inverse covariance matrix in equation (3.13). Furthermore, the entire procedure can analogously be performed on other types of correlation coefficients, for instance Spearman's rank correlation coefficient which is able to detect arbitrary monotonic relationships [66, 72].

As seen above, partial correlations extend the marginal independence concept from equation (3.6) to *conditional* independence (see Section 3.1) given the variables which are corrected for:

$$(X_i \perp\!\!\!\perp X_j) \mid X_{V\setminus\{i,j\}} \Leftrightarrow \zeta_{ij} = 0 \;\; \forall i, j \in \{1, \ldots, p\}. \tag{3.14}$$

The variables X_i and X_j are conditionally independent given all remaining variables if and only if their respective full-order partial correlation is zero. Again, under non-Gaussianity the forward direction always holds, but the reverse direction might not be true (partially uncorrelated variables could be conditionally dependent).

The statistical test for non-zero partial correlations is constructed analogously to the regular Pearson correlations, see equation (3.7). The variance of the Fisher-transformed partial correlations is now given by $\frac{1}{n-|Q|-3}$, resulting in the following equation for two-sided p-value:

$$z(\zeta_{ij}) = \frac{1}{2}\ln\left(\frac{1+\zeta_{ij}}{1-\zeta_{ij}}\right), \;\; \text{p-val}(\zeta_{ij}) = 2 \cdot \left(1 - \phi\left(\sqrt{n-|Q|-3} \cdot z(\zeta_{ij})\right)\right). \tag{3.15}$$

Note that in the full-order scenario $|Q| = p - 2$.

3.4 From correlations to graphical models

We now introduce *graphical models*, which represent a combination of concepts from probability theory and graph theory. Graphical models represent a class of probabilistic models where

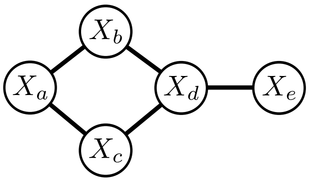


Figure 3.3: Hypothetical graphical model of five random variables. Any two nodes that do not share an edge in the graph are conditionally independent given the remaining variables.

conditional independence relationships between variables are represented by a graph [73]. The central idea is that two random variables might be independent given another set of variables, and thus their relationship needs not to be explicitly modeled. As an example, one prominent type of graphical models in biomedical research are *Bayesian networks*, where the joint probability of random variables is expressed by a lower-dimensional factorization of conditional distributions [74]. One example application of Bayesian networks is the inference of gene regulatory networks from expression data [75]. Importantly, Bayesian networks are based on directed, acyclic graphs, which prohibits the modeling of circular relationships.

In this thesis, we are particularly interested in *Markov random fields*, a specific type of graphical models based on undirected graphs [76]. Let $G = (V, E)$ be an undirected graph[1] with nodes V and edges $E \subseteq \{\{u, v\} \mid u, v \in V\}$, and $\mathbf{X} = (X_v)_{v \in V}$ a random vector of arbitrary distribution. Then the random variables X_v form a Markov random field with respect to G, if they satisfy the *pairwise Markov property*:

$$\{i, j\} \notin E \ \Rightarrow \ X_i \perp\!\!\!\perp X_j | X_{V \setminus \{i,j\}} \quad \forall i, j \in V. \tag{3.16}$$

That is, if no edge between two nodes is present in the graph, then the respective random variables are conditionally independent given all other variables. In other words, the variables that cannot be statistically separated by conditioning against the other variables induce the edges of G. Consider the example graph in Figure 3.3 and assume that $\mathbf{X}$ is a Markov random field with respect to this graph. Then the pairwise Markov property states that, for instance, X_a and X_d are conditionally independent given the remaining three variables, since they do not share an edge in the graph.

[1] For a detailed introduction to graph theory, we refer the reader to specialized text books, e.g. Bondy and Murty [77]

There are two additional, equivalent Markov properties which further clarify the connection between conditional dependencies and the underlying graph. First, the *local* Markov property at i:

$$X_i \perp\!\!\!\perp X_{V\setminus(\mathrm{ne}(i)\cup\{i\})}|X_{\mathrm{ne}(i)}, \tag{3.17}$$

where $\mathrm{ne}(i)$ represents the set of graph neighbors of node i. For our example scenario, this indicates that X_e will already be conditionally independent of X_a, X_b and X_c by only conditioning against X_d. Second, the *global* Markov property:

$$X_A \perp\!\!\!\perp X_B|X_C, \tag{3.18}$$

where $A, B, C \subseteq V$ represent distinct node subsets, and C separates A from B in the graph. In Figure 3.3, $\{X_b, X_c\}$ separate $\{X_a\}$ from $\{X_d, X_e\}$, and thus X_a and X_d as well as X_a and X_e are conditionally independent given X_b and X_c.

If **X** follows a multivariate normal distribution, then the corresponding random field is called a *Gaussian Markov random field* or *Gaussian graphical model* [78]. For the remainder of this thesis, we will refer to the second term, Gaussian graphical model, or its abbreviation 'GGM'. From equation (3.14) and the pairwise Markov property we can immediately see when a normal distribution is Markov with respect to a graph $G = (V, E)$:

$$\{i, j\} \notin E \;\Rightarrow\; \zeta_{ij} = \Sigma_{ij}^{-1} = 0.$$

Whenever no edge is present between the two nodes in the graph, the respective full-order partial correlation coefficient must be zero. This relationship provides a straightforward approach to construct a GGM if the covariance matrix can be properly inverted. We simply construct the graph which contains an edge for each non-zero entry of Σ^{-1}:

$$\{i, j\} \in E \;\Leftrightarrow\; \zeta_{ij} \neq 0 \quad \forall i, j \in V. \tag{3.19}$$

The statistical challenge lies in determining whether an entry ζ_{ij} should be considered zero or non-zero. Throughout this thesis, we will employ the statistical test of non-zero partial correlation coefficients, including appropriate adjustments for type I errors (i.e. false positives), introduced in equation (3.15). This simple GGM estimation approach has also been referred to as *edge exclusion* [79]. Starting from a fully connected graph, we remove as many edges as possible given the conditional distribution of each variable pair. More involved GGM estimation procedures are usually only required if the covariance matrix cannot be properly inverted, see Section 3.5.

At this point it is important to understand that the pairwise Markov property is not an equivalence relation. That is, it requires conditional independence for pairs of nodes without an edge, but not conditional dependence for nodes that do share an edge. For example, any multivariate distribution is Markov with respect to a fully connected graph. Therefore, GGM estimation should include a minimality constraint, in order to derive a smallest possible graph such that the Markov property still holds (cf. Castelo et al. [80]). The graph constructed using equation (3.19) merely represents one possible GGM for the underlying distribution which might still contain edges that could be removed. Specifically, conditioning against all other variables might not be appropriate in all cases. The covariance of two variables conditioned against all other variables is non-zero, but there exists a smaller set of variables where this covariance would turn zero. One then needs to decide which type of conditional association one considers to be 'correct'. This issue needs to be kept in mind when applying the approach to real data later. In practice, we will employ the edge exclusion approach from equation (3.19).

A graph-based representation of the conditional dependency structure between random variables has several advantages. First and most obvious, a graph represents a convenient mathematical structure suitable for visualization and manual interpretation. This will become particularly important when analyzing GGMs with hundreds of variables later in this thesis. Second, the graph provides systematic insights into the variable relationships. Inspecting the example graph in Figure 3.3, for instance, demonstrates not only that X_a and X_d are conditionally independent given all other variables, but also that the conditioning against X_b and X_c would have sufficed to separate these variables. Third, graph theory represents a well-established scientific field with a plethora of algorithms and methods that can immediately be applied. To this end, several GGM estimation methods specifically exploit the Markov properties in order to reconstruct the dependency structures in a graph-based fashion (see Section 3.5 for examples of such approaches).

Taken together, the graph of a Markov random field contains conditional independence information of the underlying distribution.The Markov properties represent the foundation for a rigorous theory and a plethora of methods and applications, which will not be discussed in further detailed here. We refer the interested reader to the pertinent literature [70, 80, 81].

3.5 The small n, large p problem

In the previous sections we have seen that full-order partial correlations cannot be properly calculated if the rank of the data matrix is lower than the number of variables p (which in particular is the case when we have less samples n than variables). Microarray studies yield tens of thousands of measured variables, and even for the metabolomics setups used in this thesis we can by now measure on the order of 500 different compounds. Only for large-scale population studies like KORA, with around 1000 to 3000 measured samples we can match the number of variables for the metabolomics experiments, let alone reaching sufficiently many samples for mRNA expression measurements. Several methods, especially motivated by questions from *omics* analyses, employ regularization approaches to circumvent the $n < p$ issue and nevertheless allow for an estimate of partial correlations or, more general, conditional independence. Although most applications presented in this thesis do not require such methods due to the high number of measured samples, investigating the respective algorithms in detail yields further insights into the nature of partial correlation coefficients and GGMs.

Even for situations where we do have sufficiently many samples to invert the covariance matrix, the obtained estimate might not reflect the correct conditional independence relationships. While the empirical covariance always represents the maximum likelihood estimate of the covariance matrix in the data, it might substantially deviate from the true covariance underlying the measurements for low sample sizes [82, 83]. All methods introduced in the following directly or indirectly tackle the problem of covariance instability, either by statistical means or by exploiting the graphical model properties introduced in Section 3.4.

Graph-based methods

A specific class of GGM estimation methods suitable for small n, large p scenarios are graph-based methods that exploit the properties introduced in Section 3.4. We will review three example algorithms in the following. First, de la Fuente et al. [55] introduced an approach which only calculates up to third-order partial correlations, i.e. $|Q| \leq 3$, for the elucidation of associations in genomics data.

Such low-order partial correlations can even be calculated for very small sample sizes. More specifically, the calculation only requires $\text{rank}(\mathbf{x}) \geq |Q| + 1$ (fitting one regression coefficient for each covariate, and one for the intercept). For each pair of variables, the algorithm iterates through all possible combinations of either one, two or three conditioning variables. If the respective partial correlation drops below significance level at least once, the edge is removed from the graph. Recall the global Markov property introduced on page 30. A k-order partial correlation approach will reconstruct the correct model whenever any two nodes in the (true) underlying dependency network can be separated by removing maximally k other nodes. In the example graph from Figure 3.3, first-order partial correlations would not have been sufficient to properly separate X_a and X_d, whereas already second-order partial correlations would have reconstructed the correct graph. Whether or not the removal of only three nodes will be sufficient for real biological networks might be debated. Importantly, the authors state that the primary goal of the approach was not to entirely reconstruct the underlying network correctly, but rather to correct for the 'most active' paths in the network and generate new hypotheses on biochemical interactions. On the other hand, we will see in Chapter 5 that this approach works reasonably well for a lipid-focused metabolomics dataset.

The second algorithm, published by Castelo et al. [80], follows an approach quite similar to the de la Fuente et al. study. Instead of enumerating all possible combinations for small k values, they randomly draw conditioning variable sets in the sense of a Monte Carlo approach, followed by computation of a *non-rejection rate*. This rate represents the number of times the null hypothesis of zero partial correlation was rejected. Lower values here indicate a higher probability of a present edge in the true underlying graph. The method allows larger sets of conditioning variables, on the order of $k = 10$ or $k = 20$ (given that $n > k$). The approach is developed on a rich statistical and graph theoretical framework, representing a valuable contribution on its own. For instance, the authors formalize the above-mentioned node separation concept into graph property called *outer connectivity*. A major drawback of the approach as such is the vast number of variable combinations which arise for larger values of k. For example, from a measured set of 150 metabolites, we can draw around 10^{15} subsets of size $k = 10$, a number which can obviously not even be remotely reached by a Monte Carlo sampling approach. The probability of actually drawing a suitable set of conditioning variables that separates the two variables under investigation, or at least sufficiently lowers their partial correlation to create a detectable signal, might be unfeasibly low. The approach is evaluated on toy data in the original publication, and was later applied to reverse-engineer regulatory networks from E.coli expression data [84].

Another graph-based GGM estimation approach employing the Markov properties has been published by Peña [79] in 2008. The concept of the algorithm is based on the local Markov property (equation 3.17). Instead of the 'neighborhood' of a node, the author refers to the *Markov boundary* (MB) of a node, which is equivalent to the Markov property concept. The MB is reconstructed from the data for each variable separately. The algorithm alternates between (1) adding new variables to the Markov boundary which display the highest, significant partial correlations to the variable under investigation given the current MB, and (2) removing variables from the boundary whose partial correlations have vanished given the current MB. A particularity of this approach is the integration of false-discovery rate control directly into the algorithm. The advantages of this method are calculation speed and good reconstruction performance for low sample sizes. Problems might arise due to the 'greedy' character of the algorithm which might not take into account complex interactions between variables. In the original publication, the method has been used to reconstruct interaction networks from yeast expression data (300 samples, 6316 transcripts). A specific high-scoring subnetwork around iron homeostasis is briefly discussed, where four iron transporter transcripts present central hubs in the network.

Bootstrap aggregation

A simple way of generating an estimate of the inverse covariance matrix, which we require to compute full-order partial correlations, is to use the Moore-Penrose pseudoinverse. It represents a generalization of the standard matrix inverse, can be calculated for any singular matrix and reduces to the standard inverse if possible [85]. In an *empirical Bayes approach* termed algorithm, Schäfer and Strimmer [57] used the pseudoinverse in combination with bootstrap aggregation (*bagging*) in order to generate a stable estimate of partial correlations in a low sample size scenario. For bagging, bootstrap samples are drawn from the original dataset, the desired statistic is calculated for each of these samples independently, and finally a mean (bagged) estimate of the statistic is obtained. This procedure reduces the variance of the estimated statistic. The authors report two approaches, one where a bagged estimate of the correlation matrix is obtained and only one pseudoinverse is calculated, and one where the pseudoinverse is subject to bagging. The former variant is reported to be more suitable for the $n < p$ situation, whereas the latter one displays small error and good statistical properties when n is on the order of p (which they call the *critical n zone*). Note that statistical testing is more involved for this approach, since the earlier introduced Fisher transformation model, see Equation (3.15), is not appropriate for this scenario. As a biological application, the authors estimate a GGM from breast cancer expression data, followed by a manual investigation of high-scoring subnetworks. Interestingly, the authors later admitted this method to be computationally too demanding for very large p (e.g. above 1000) [83], and actually removed it from the corresponding GGM estimation R package[2].

Shrinkage

As an alternative method to variance reduction of the covariance estimator using bagging, the same authors suggested a *shrinkage*-based covariance estimation in the same year [83]. This algorithm will be used in Chapter 7.2 to estimate GGMs from lipidomics data with very low sample sizes. Similar to bagging, the general idea of shrinkage methods (also called *biased estimation*) is to *shrink* the variance of an estimator, yet however with a substantially different

[2] `http://cran.r-project.org/web/packages/GeneNet/GeneNet.pdf`

statistical approach. The basic concept is as follows: In addition to the actual model $\boldsymbol{U}$ to be estimated (in our case, the full covariance matrix) which might be statistically unstable, one defines a lower-dimensional submodel $\boldsymbol{T}$, which usually contains a substantial estimation bias but is easy to calibrate due to a small amount of parameters. A shrinkage estimator is then defined as $\boldsymbol{U}^{\star} = \lambda\boldsymbol{T} - (1-\lambda)\boldsymbol{U}$, where the shrinkage parameter λ defines the 'mixture' of true model and submodel. The particular challenge is then to find an optimal value for the shrinkage parameter λ, such that the difference between estimated and true parameters is minimal.

The shrinkage target $\boldsymbol{T}$ can have various forms, depending on the respective problem. For covariance estimation, the authors suggest a 'diagonal, unequal variance' target, i.e. all off-diagonal elements are zero and each variable is allowed to have a different variance. The model then reduces from $p \cdot (p+1)/n$ to only p parameters to be estimated. Furthermore, the authors demonstrate how to determine the optimal shrinkage intensity λ for this target model analytically. While obviously $\boldsymbol{T}$ does not represent a proper estimation of the true covariance matrix, it is statically stable even for small sample sizes. Furthermore, the authors show that in fact any target $\boldsymbol{T}$ will lead to a reduction of the variance of the estimator. The reduction might however be neglectable for a strongly misspecified target. With a statistically stable and reasonably accurate estimate of the true covariance matrix, one can then obtain the full-order partial correlation matrix by matrix inversion.

The shrinkage approach including subsequent partial correlation estimation was then applied to expression data from E.coli under stress conditions, on a total of p=102 preselected transcripts under n=8 experimental conditions. The authors describe a specific high-scoring subnetwork around the genes lacA, lacZ and lacY, i.e. transcripts around the lac operon which they functionally link to the respective experimental conditions. For method comparison, the authors also generated a 'relevance network' based on shrinkage-based (common) correlation coefficients and a GGM based on a graphical Lasso approach. They claim that the correlation-based approach is not suitable for recovering biologically reasonable associations, and that the approach should merely be used to assess marginal statistical independence, cf. equation (3.6). For the lasso GGM, they report a structural bias, since this method implements sparsity per node instead of for the whole network. This creates a structural constraint on the network which might prohibit estimation of the correct dependency structure in biological networks. The authors conclude by a critical statement on the usage of correlation-based measures in the bioinformatics field. Correlations are often applied 'rather blindly' to datasets with many variables and only few samples. As discussed, these estimators can then perform very poor and should be replaced by a statistically more robust variant.

Further approaches

In addition to the graph-based and estimator variance-reducing approaches mentioned above, there are numerous further GGM estimation methods suitable for $n < p$ scenarios. For instance, several algorithms employ L_1 (Lasso) regularization for GGM estimation [86–88]. The basic idea of Lasso regression is a penalty term for non-zero coefficients, which then automatically induces sparsity of the model. This approach is not to be confused with the linear regression given in equation (3.9). Rather, variables are regressed onto each other with Lasso penalty, and if at least one (or alternatively both) variable has a non-zero coefficient on the other variable, the estimated value in Σ^{-1} will be non-zero. Other methods employ Bayesian ideas for GGM estimation, which allows to define prior distributions for the partial correlation matrix. For instance, Wong et al. [89] set up a Bayesian framework for covariance selection, with a Markov Chain Monte Carlo (MCMC) algorithm to sample from the posterior distribution. It is to be noted that the priors employed in this study do not represent prior knowledge, e.g. from biochemical pathways, but rather represent structural properties like sparsity of the estimated graph. A further branch of estimation approaches deals with *robust* estimation of Gaussian graphical models, i.e. account for effects of outliers in the data. We will not go into detail here and refer the reader to Miyamura and Kano [90] as an exemplary study.

Taken together, there are numerous methods for GGM estimation based on different statistical and graph theoretical findings for the inverse covariance matrix. They provide both, valuable alternatives to the simple matrix inversion steps in cases where we have less samples than variables, but also interesting insights into the properties of this class of graphical models.

Chapter 4

GGMs on computer-simulated reaction systems

The forward simulation of artificially constructed models is a valuable tool for the evaluation of network inference methods, which we will use prior to their application to real metabolomics data sets. The goal of this approach is to determine the general capabilities of GGMs to distinguish direct from indirect biochemical reactions, and to discover possible problems and pitfalls. Specifically, we dynamically model biochemical reactions by ODEs in order to evaluate the metabolite correlation structures that arise in metabolic systems. As discussed in Chapter 1.4, previous works focused on the modeling of biological replicates with intrinsic noise on the metabolite levels [48], or varying enzyme concentrations [49]. In contrast, we here investigate the effects of variation of enzymatic activity and metabolic states in a human population cohort. Such variation might be genetically determined or, more likely, be the result of distinct regulatory effects and metabolic states between individuals. All reaction systems were implemented as ordinary differential equations (ODEs) with simple mass-action kinetics rate laws and reversible Michaelis-Menten-type enzyme kinetics. In order to account for the enzymatic variability, we applied a log-normal noise model, which has been previously described to be a reasonable approximation of cellular rate parameter distributions [91].

The results reported in this chapter are part of the following publication:

⋆ **Krumsiek, J.**, Suhre, K., Illig, T., Adamski, J., and Theis, F.J. Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data.. *BMC Syst Biol*, 5(1):21, 2011

4.1 Modeling metabolic reaction systems by ODEs

We now introduce the methodological backgrounds for the modeling and simulation of metabolic reaction systems.

Let $x = (x_1, \dots, x_r)$ be a vector of metabolite concentrations and $S \in \mathbb{Z}^{m \times r}$ the stoichiometry matrix of a dynamical system with m metabolites and r reactions. Each column in S represents the compound stoichiometry of a single reaction, with negative values for the educts of a reaction and positive values for its products (cf. Palsson [92]). Furthermore, we define an *educt stoichiometry matrix* S^e, which only contains the negative values from S. The reaction rate laws v can be written as $v(x, k) = \mathrm{diag}(k)c(x)$, where $k := (k_1, \dots, k_r)$ represents a vector of elementary rate constants and $c_j(x) := \prod_{i=1}^{m} x_i^{-S^e_{ij}}$, $j = 1, \dots, r$ contains the products of substrate concentrations according to the law of mass action [93]. For example, for the reaction $x_1 + x_2 \to x_3$ we obtain $c = x_1 x_2$, and $2x_1 + 3x_2 \to 2x_3$ yields $c = {x_1}^2 {x_2}^3$.

For enzyme-catalyzed reactions i, the corresponding entries in v are formulated using reversible Michaelis-Menten-type kinetics [94, 95] instead of the mass-action term above:

$$v_i = \frac{\frac{V_{\max}^{+}}{K_M^s} \cdot [S] - \frac{V_{\max}^{-}}{K_M^p} \cdot [P]}{1 + \frac{[S]}{K_M^s} + \frac{[P]}{K_M^p}},$$

where $V_{\max}^{+}$ and $V_{\max}^{-}$ are the product and substrate formation constants, respectively, K_M^s and K_M^p represent the Michaelis constants for substrate and product, $[S]$ represents the substrate concentration and $[P]$ represents the product concentration. Note that we omitted reaction-specific parameter indices for simplicity here. Allosteric regulation was modeled using a mixed inhibition mechanism, which extends the rate law from equation (4.1) as follows:

$$v_i = \frac{\frac{V_{\max}^{+}}{K_M^s} \cdot [S] - \frac{V_{\max}^{-}}{K_M^p} \cdot [P]}{1 + \frac{[I]}{K_i} + \left(\frac{[S]}{K_M^s} + \frac{[P]}{K_M^p}\right)\left(1 + \frac{[I]}{K_{ii}}\right)},$$

with $[I]$ being the inhibitor concentration, K_i the binding rate of the inhibitor to the enzyme and K_{ii} the binding rate of the inhibitor to the substrate-enzyme (or product-enzyme) complex. In a simple mixed (*non-competitive*) inhibition scenario, we assume $K_i = K_{ii}$.

The ordinary differential equations (ODEs) describing the temporal evolution of the system are now given by

$$\frac{\mathrm{d}x}{\mathrm{d}t} = S \cdot v(x,k). \tag{4.1}$$

In order to introduce variability, each parameter is subject to fluctuations according to a log-normal distribution with mean 1 and changing variances: $k_i \sim \mathrm{LogN}(1, \sigma_i^2)$. For fixed S and k, Pearson and partial correlations are calculated by drawing the vector k 5000 times from the parameter distribution, calculating the corresponding metabolite steady state concentrations and logarithmizing the obtained values. If the system contains only zeroth-order and first-order reactions (i.e. input reactions and reactions with only one substrate), the steady state concentrations for a given k can be readily computed by equating (4.1) to zero and solving for c using linear algebra techniques. On the other hand, if higher order reactions are present, the ODEs are integrated numerically and simulated until equilibrium to get corresponding steady states. For this purpose, a variable-order solver for stiff differential equations (`ode15s`) from MATLAB was used [96].

4.2 GGMs reconstruct direct relationships in artificial reaction systems

In the following, we will first discuss the general reconstruction capabilities of Gaussian graphical models on varying biochemical network topologies. The subsequent sections will then focus on particular features of these systems, including subtle topology changes, input noise, enzyme-catalyzed reactions and negative feedback.

The default standard deviation σ for the simulations was set to 0.2 for the underlying normal distribution. For each of the 5000 parameter samples, we calculated the metabolite steady state concentrations on log-scale, and subsequently estimated the GGM by calculating partial correlation coefficients.

The first network we analyzed consists of a linear chain of three metabolites with different variants of reaction reversibility (Figure 4.1A-C). We observe high pairwise correlations for metabolites in mutual equilibrium due to reversible reactions (Figure 4.1A). This is in accordance with previous findings from Camacho et al. [49], where correlation-generating mechanisms in metabolic reaction networks were identified. Furthermore, this simple example demonstrates how partial correlation coefficients in GGMs discriminate between directly and indirectly related

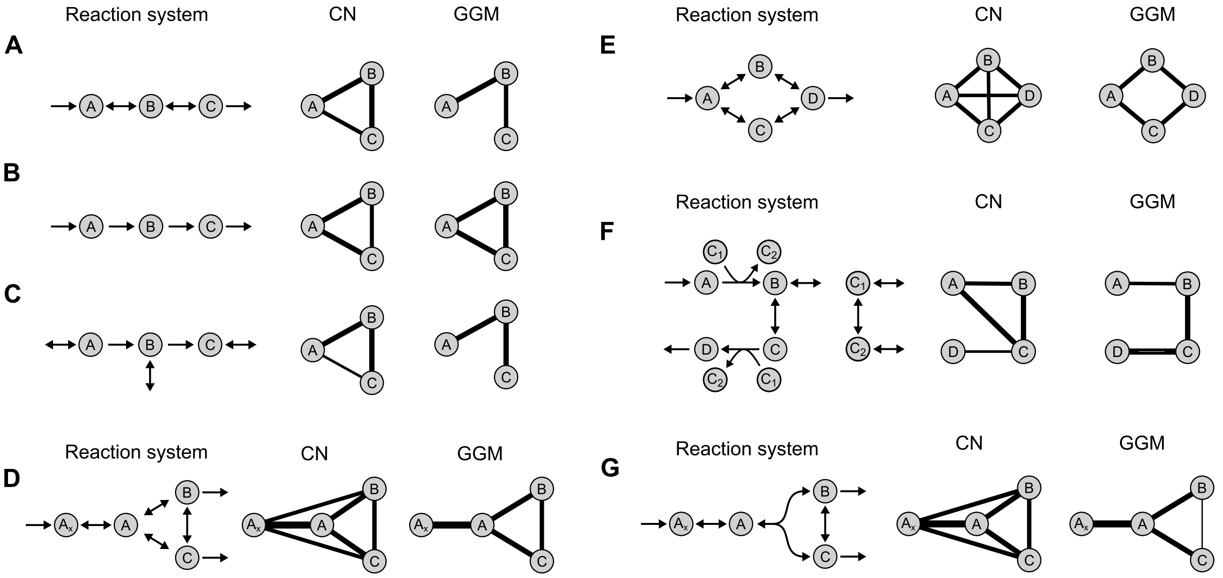


Figure 4.1: Evaluation of correlation networks (CN) and Gaussian graphical models (GGM) on artificial systems. Line widths represent relative edge weights in the respective networks (scaled to the strongest edges). **A:** Linear chain of three metabolites with reversible intermediate reactions. While the standard Pearson correlation network (CN) is fully connected, implying an overall high correlation of all metabolites, the GGM correctly discriminates between direct and indirect interactions. **B:** Linear chain with irreversible intermediate reactions. Neither CN nor GGM can distinguish direct from indirect effects, as metabolite A equally determines the levels of both B and C in our steady state scenario. **C:** Linear chain with irreversible reactions and input/output reactions for each metabolite. Although the edge weights for both CN and GGM are generally lower, the GGM now correctly predicts the network topology. **D+E:** Branched-chain first-order networks are correctly reconstructed by the GGM. **F:** Cofactor-driven network resembling the first three reactions from the glycolysis pathway. The correlation network fails to predict the correct pathway relationships. **G:** Non-linear system with a bi-molecular reaction. The GGM predicts only a only weak interaction between B and C. This is due to counterantagonistic processes of isomerization and substrate participation in the same reaction.

metabolites. If only irreversible reactions are employed in the chain, neither regular correlation networks nor GGMs can distinguish between direct and indirect effects (Figure 4.1B). Species A is the only input metabolite in the system, and thus completely determines the levels of both B and C. This leads to generally high and non-distinguishable correlations between the three metabolites. However, if we introduce exchange reactions for all species, the GGM again correctly describes the network connectivity (Figure 4.1C). Such exchange mechanisms are likely to be present for most intracellular metabolites, which usually participate in multiple metabolic pathways (see e.g. KEGG PATHWAY online). Note that for this third case both regular and partial correlation values are notably lower than for the first two chain variants. In addition to linear chains, pathway modules consisting of branched topologies with first-order, reversible reactions are correctly reconstructed in the GGM (Figure 4.1D+E).

Next, we studied the influence of cofactor-driven reactions on the reconstruction. Cofactors are ubiquitous substances usually involved in the transfer of certain molecular moieties or redox potentials [97]. We investigated such cofactor-coupled reactions (a) because they introduce non-linearity in the simulated dynamical systems, and (b) because cofactors are usually involved in many reactions and thus generate network-wide metabolite dependencies. We set up a network resembling the first three reactions from the glycolysis pathway. It consists of four metabolites and two energy transfer-related cofactors, ATP and ADP, involved in two phosphorylation reactions [98]. Again the GGM precisely describes metabolite connectivity in the system, whereas a regular correlation graph leads to false interpretations of the network topology (Figure 4.1F). Cofactors were modeled with input and output reactions to the rest of the metabolic system in order to account for the above-mentioned participation of cofactors in various reactions of the system. Again, it makes a substantial difference whether such exchange reactions are included in the model or not. Since our toy model only represents a small part of a larger system, missing exchange reaction for cofactors would create a false mass conservation relation that compromises correlation calculation.

Finally, we investigated the effects of rate laws with non-linear substrate dependencies in the absence of cofactors. We modeled a reversible, bimolecular split reaction with isomerization of the two substrates (Figure 4.1G). An example of such a reaction network can be found in the glycolysis pathway between *fructose-1,6-bisphosphate*, *glyceraldehyde-3-phosphate* and *dihydroxyacetone phosphate*. Our simulations demonstrate that again a regular Pearson correlation network cannot delineate direct from indirect relationships in the pathway. The GGM only detects a weak association between B and C. This is due to counterantagonistic processes in this reaction setup: isomerization and other reversible reactions generally induce positive correlations, whereas coparticipation as substrates in the same reaction induces negative correlations.

Such effects of correlation-generating mechanisms which cancel each other out have been described before [49] and pose a problem to all reconstruction approaches which rely on linear dependencies.

4.3 Analysis of reconstruction stability in selected first-order networks

In addition to the detailed analysis of the seven networks from Figure 4.1, we systematically investigated the discrimination stability of GGMs on various first-order reaction systems (Figure 4.2). These include: (1) Three metabolites connected in a row (*Chain 3*), with all 8 variants of reversible inner reactions and reversible external reactions. (2) Five metabolites connected in a row (*Chain 5*), with all 8 variants of reversible inner reactions and reversible external reactions. (3) A branching pathway (*Split*), with all 8 variants of reversible inner reactions and reversible external reactions. (4) An irreversible feed-forward loop motif (*FFL*), and one variant with external reactions. (5) An irreversible branching and merging motif (*Diamond*), and one variant with external reactions. (6) A densely interconnected network of six player with several subvariants (*Dense*).

In order to objectively evaluate the discrimination between directly and indirectly connected metabolites, we calculated sensitivity and specificity as:

$$\text{sens} := \frac{\text{TP}}{\text{TP}+\text{FN}} \text{ and } \text{spec} := \frac{\text{TN}}{\text{TN}+\text{FP}},$$

with TP true positives, FP false positives, TN true negatives, FN false negatives [99].

A metabolite pair is considered true positive if it exhibits a partial correlation above the threshold and has a direct pathway connection; a false positive represents a metabolite pair also above the threshold but with no direct pathway connection; a false negative pair lies below the threshold but does have a direct pathway connection; and finally a true negative pair lies below the threshold and also has no direct pathway connection. The F_1 score was calculated as the harmonic mean of both quantities [100]:

$$F_1 := 2 \cdot \frac{\text{sens} \cdot \text{spec}}{\text{sens} + \text{spec}}.$$

We generated samples of 1000 simulations (i.e. 1000 measured data points) with log-normal noise and subsequent partial correlation computation. This procedure was repeated 100 times. For each of these 100 runs we calculated the discriminatory power according to the area under the ROC curve. The 100 F_1 values for each network for both regular and partial correlations are visualized in Figure 4.3.

In general, partial correlations display a profoundly higher discrimination quality than Pearson correlations, further confirming our findings from Section 4.2. As expected, Pearson correlations usually yield perfect sensitivity (all true interactions are captured), but a poor specificity (‘false positives’ introduced by indirect associations). For instance, the reversible reaction chains with and without exchange reactions are perfectly reconstructed by partial correlations (median $F_1 = 1.0$), whereas Pearson correlations show F_1 scores of 0.8. Similar effects can be observed for the more complex network topologies ‘Big split’, ‘FFL’, ‘Diamond’ and ‘Dense’. Only for one case, ‘Dense, all irrev., no intermed., in+out’, Pearson correlations produce a slightly better F_1 score than partial correlations (F_1=0.75 and F_1=0.71, respectively). However, this network cannot be considered properly reconstructed, as also shown in Figure 4.4 (bottom right) in the following section.

We observed the following reconstruction features for partial correlations: (1) Networks with reversible reactions show perfect discrimination (except for a few cases in the *Big split, rev, all out* network which we attribute to parameter outliers). (2) Irreversible reactions generally impair the discrimination quality. For the chains, there is no discrimination at all between directly and indirectly connected metabolites. (3) Input and especially output reactions improve the quality and make discrimination possible even for the irreversible straight chains. (4) GGMs can delineate intricate relationships as seen in the *FFL* and *Diamond* networks. (5) Discrimination works acceptable for all variants of the *Dense* network, but never perfect due to irreversible reactions and missing input/output mechanisms. (6) Since multiple runs produce different AUC results (widths of boxes in Figure 4.3), there is a certain stability issue for GGM calculation. It is to be noted however, that for systems with very instable discrimination results (e.g. ‘Chain 3, irrev’), this is due only to subtle changes in the correlation values.

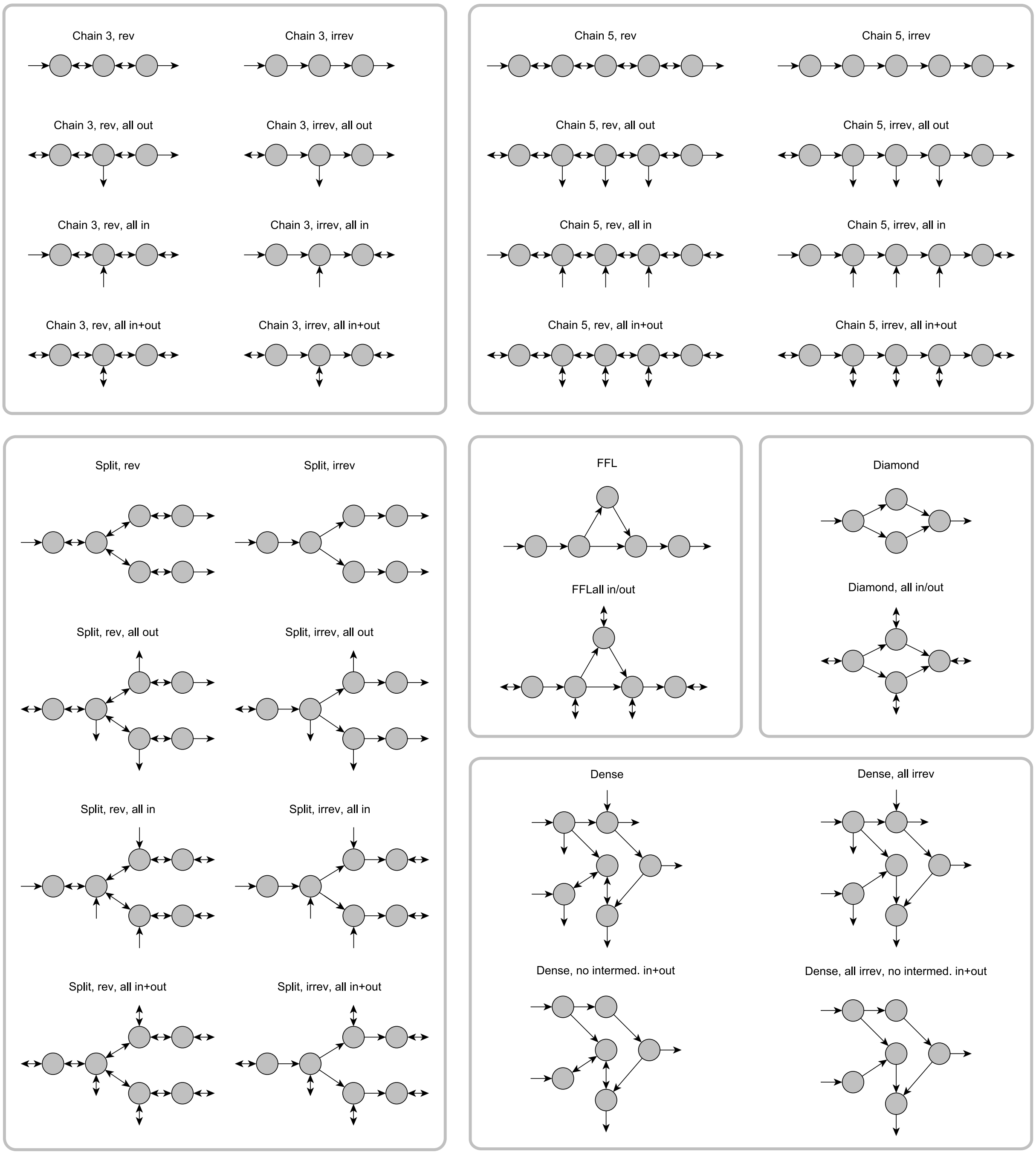


Figure 4.2: Variants of first-order networks. We investigated a total of 32 reaction systems, including linear chains with combinations of reaction reversibility and boundary reactions, a split motif, a feed forward loop, a diamond-shaped network and a densely interconnected reaction system.

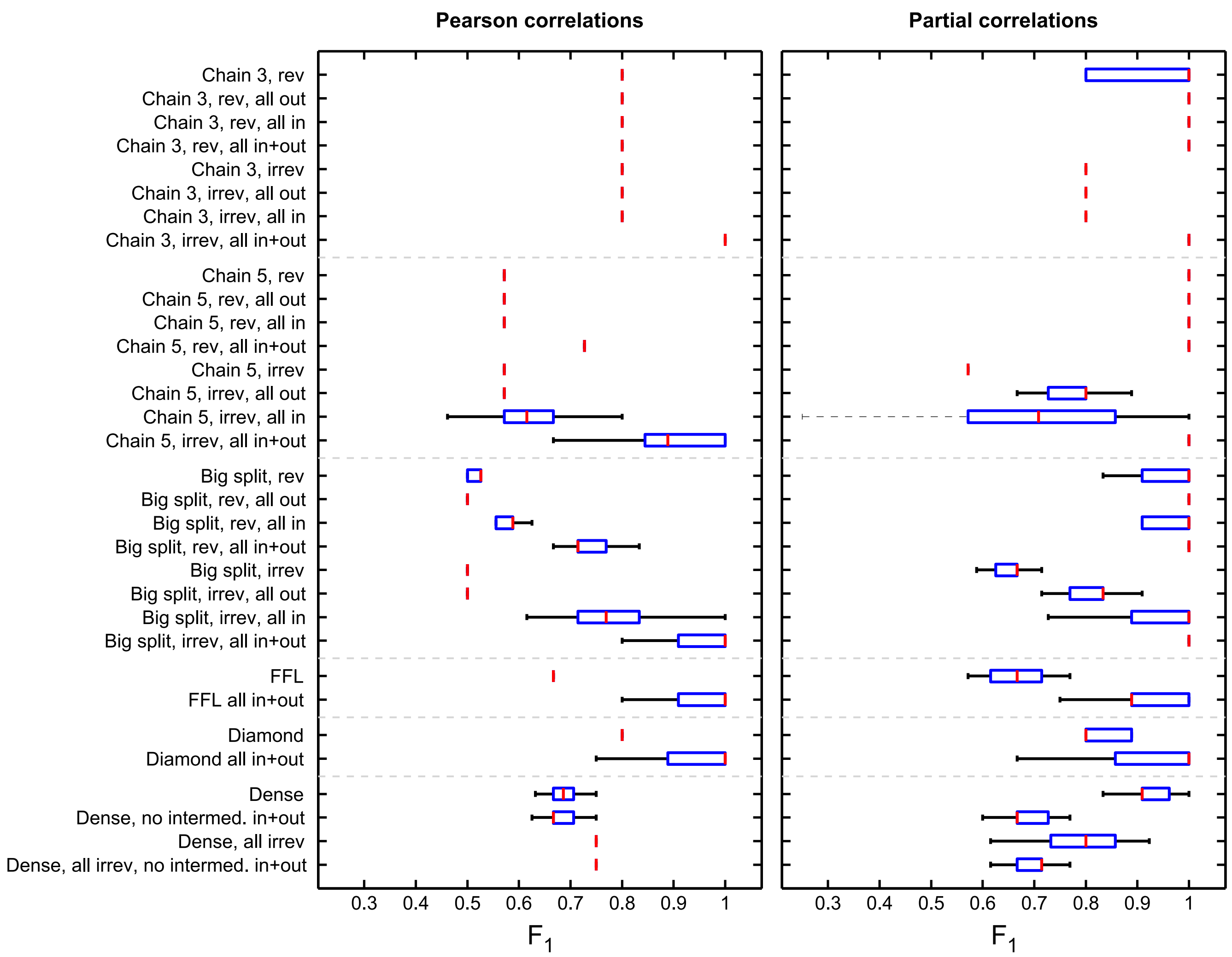


Figure 4.3: Discrimination quality between directly and indirectly connected metabolites for the 32 simulated networks for a total of 100 runs with 1000 sampled steady states each. F_1 represents the harmonic mean of sensitivity and specificity. Values close to 1.0 indicate perfect discrimination. As expected, we observe a profoundly higher discrimination quality for partial correlations due to the poor specificity of Pearson correlations (caused by indirect associations).

4.4 Stronger input noise generally improves discrimination quality

In the next step, we analyzed increasing fluctuation strengths for the input reaction of the first metabolite in all toy networks. Applying the log-normal noise model, we ranged the standard deviation of the underlying normal distribution from 0 to 10 for the input reaction of the first metabolite. Standard deviations for all other reactions were kept constant at a value of 0.5.

In order to actually quantify the discrimination properties, we here investigated raw partial correlation coefficients rather than analyzing the F_1 score (Figure 4.4). For all networks where discrimination is generally possibly, we observe an increase in discrimination strength for higher strengths of the input noise. We exemplarily discuss the ‘Chain 3, irrev, all in’ network. For low values of the input noise mean value, the GGM cannot distinguish between directly and indirectly connected metabolites; an effect of the above-mentioned irreversibility of reactions. For an input noise strength of 1 or higher, however, the system is capable of reconstructing the topology correctly. The analysis also demonstrates that irreversible reaction chains without any exchange reactions can never be delineated, irrespective of the input noise. Other scenarios like the reversible reaction chains can always be properly reconstructed.

The relationship between input noise and reconstruction quality has important implications for the analysis of real metabolomics datasets. In a heterogeneous human population study like KORA, we expect a variety of different metabolic states. Therefore, the availability of substrates for a given pathway might drastically vary between two probands, and thus the ‘input noise’ in the real population is most certainly high. Compared with our simulations, we conclude that a high heterogeneity in the metabolomics samples might be beneficial rather than problematic for the GGM reconstruction approach. This is in accordance with our hypothesis from Chapter 1.4.

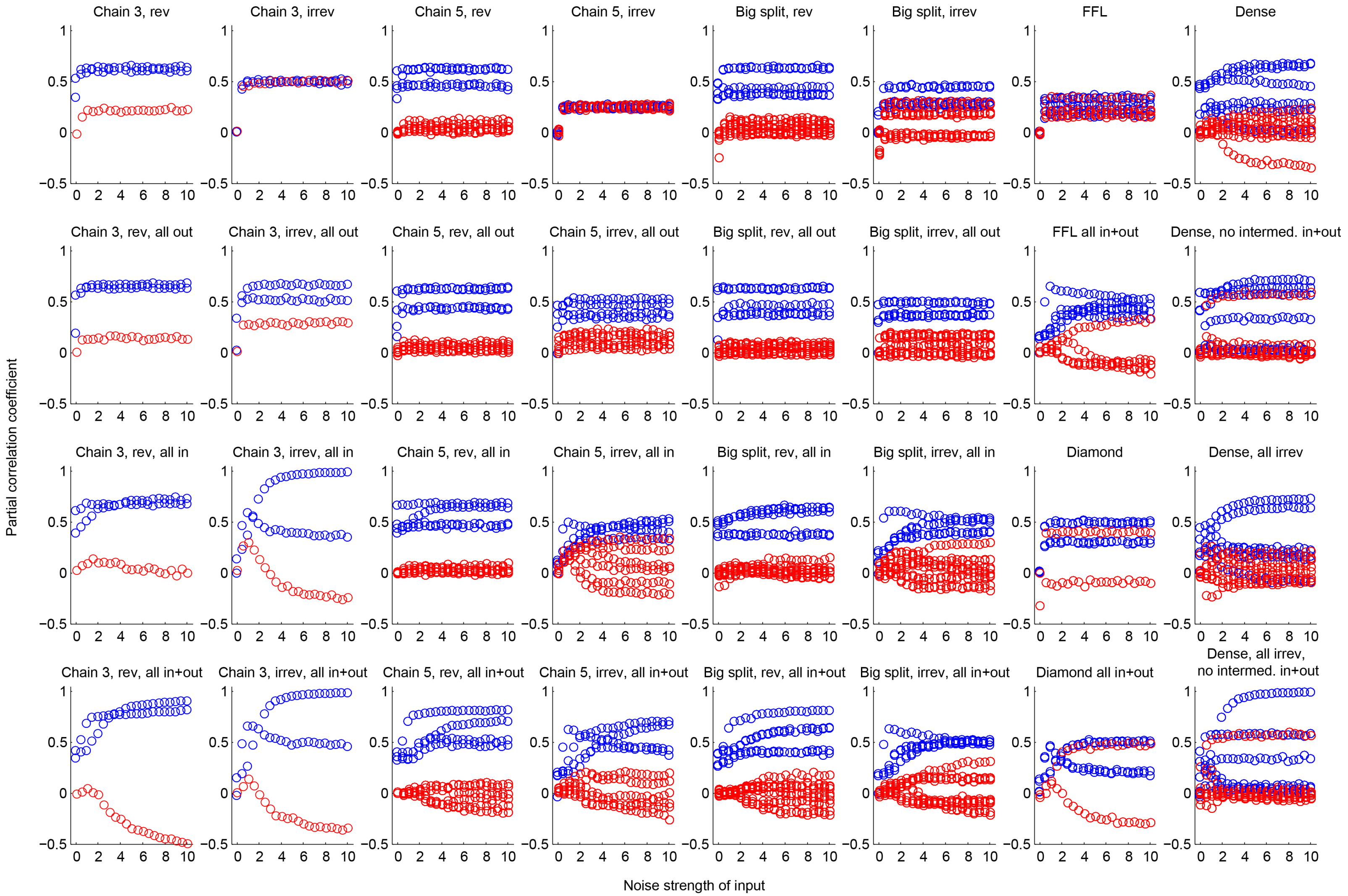


Figure 4.4: Effects of input noise on GGM estimation. The x-axis represents the strength of noise on the input reaction of the system; partial correlation coefficients are plotted along the y-axis. Blue circles represent metabolite pairs that are directly connected in the underlying network, whereas red circles stand for unconnected metabolite pairs. For many scenarios, we observe an increase in discrimination quality for stronger input noise (e.g. 'Chain 3, irrev, all in+out', last row, second column).

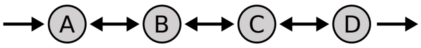


Figure 4.5: Linear reaction chain of four metabolites with reversible internal reactions.

4.5 Enzyme kinetic rate laws only marginally affect GGM estimation

The next question we addressed with artificial reaction networks was the influence of enzyme-catalyzed reactions and variations in the respective parameters on GGM estimation. We set up reaction chains with four metabolites (Figure 4.5) incorporating reversible enzymatic reactions.

The log-normal noise assumption of cellular rate parameters can be interpreted as a log-normal variation of $V_{\max}$ parameters in the Michaelis-Menten case. Since K_M is supposed to be an intrinsic property of the respective enzyme-substrate interaction, this parameter was kept constant throughout all simulations.

Similar to the first-order case described above, one simulation consists of drawing 1000 parameter sets, calculating a steady state for each parameter sample and subsequent GGM estimation using the obtained 1000 steady states. We performed simulations for (a) varying *mean* values of $V^{+}_{\max}$, while $V^{-}_{\max}$ was kept constant in order to investigate different degrees of reaction reversibility; (b) varying *constant* values of both K_M parameters to introduce different levels of response linearity; (c) varying *mean* values of the zeroth-order input reaction carrying A into the system; (d) varying levels of overall noise strengths, i.e. the standard deviation of the underlying log-normal distribution. All parameters means were again set to 1 by default, except for K_M where we chose a constant value of 0.01. We only accepted parameter combinations that reach a stable steady state. In contrast to mass-action kinetics, Michaelis-Menten kinetics introduce an upper bound to the reaction rate (namely $V_{\max}$). Specifically, if the constant influx into the system exceeds the net rate from A to B, A will grow infinitely large. These scenarios are biochemically not viable and have thus been ignored.

For all variations of $V^{+}_{\max}$ the GGM distinguishes direct from indirect interactions (Figure 4.6A). Only if the forward reaction rate exceeds the backward reaction by far, e.g. $\log10(V^{+}_{\max}) = 2$, that is $V^{+}_{\max} = 100$, the discrimination quality is impaired. This is in line with the observation that purely irreversible reactions cannot be distinguished in the mass-action case. For the cases $\log10(V^{+}_{\max}) = 3$ and 4, unusually high regular Pearson correlations occur (mean overall correlation >0.985), and thus the seemingly improved GGM reconstructions cannot be consid-

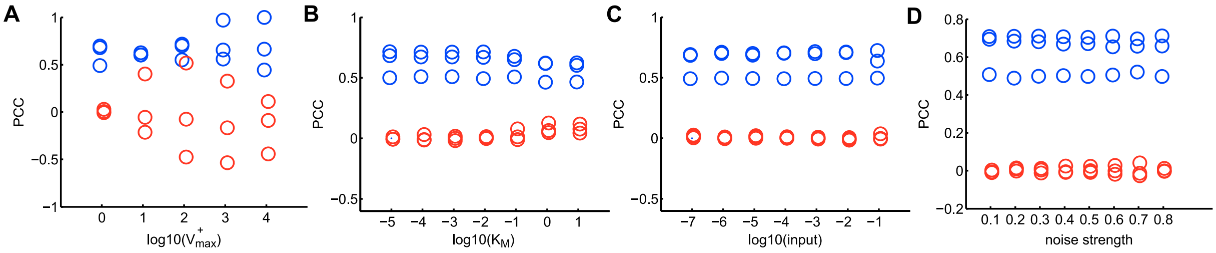


Figure 4.6: GGM discrimination quality for Michaelis-Menten kinetics with varying parameter settings. **A:** PCC changes for the maximal forward reaction rate $V_{\max}^{+}$. **B:** Changing the constant saturation parameter K_M. **C:** Input noise strength, i.e. variation of the reaction which produces the first metabolite in the cascade. **D:** Overall noise strength (σ of log-normal distribution). Blue circles represent metabolite pairs that are directly connected in the underlying network, whereas red circles stand for unconnected metabolite pairs. Generally, the reconstruction capabilities of GGMs are not strongly influenced by enzyme rate law parameters. PCC = partial correlation coefficient.

ered meaningful for these cases. Note that $V_{\max}^{-}$ was kept at a constant value of 1, and thus the parameter value in this plot represents the ratio between forward and backward reaction rate.

Other enzyme kinetics parameters did not display significant impacts on GGM calculation (Figure 4.6B-D). This is particularly interesting for the Michaelis constants K_M, which adjust the degree of saturation in the activation curve. Low values in this parameter cause a quick saturation towards the respective $V_{\max}$ value. However, since our approach does not investigate actual reaction speeds but rather the steady state levels at equilibrium, the effect on reconstruction quality is neglectable.

4.6 Negative feedback might compromise discrimination

Another important aspect of enzyme-catalyzed reactions are allosteric regulation mechanisms, like end-product inhibition for instance, which constitutes a negative feedback from the end to the beginning of a pathway [101]. We set up a linear reaction chain with enzyme kinetics and an inhibition of the last metabolite to onto the first reaction of the cascade (Figure 4.7). For the initial analysis, forward maximal reaction rates $V_{\max}$ were set twice as fast as the backward reactions in order to ensure a directed mass flow. The reconstruction results differ depending on whether exchange reactions are included in the system or not (Figure 4.7A). If the inhibitory module represents a closed system (no external fluxes except for the first and last metabolite), the regulatory interaction does not influence GGM calculation. The net metabolite turnover

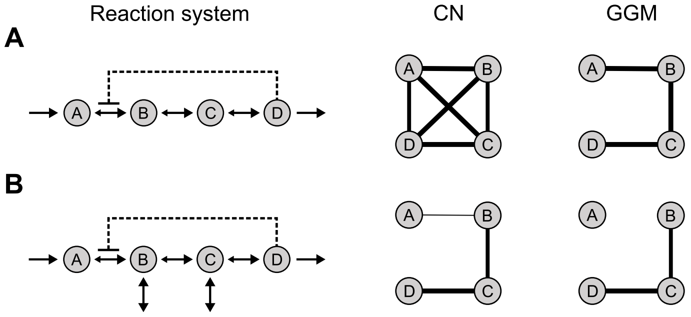


Figure 4.7: End-product inhibition modules without (**A**) and with (**B**) exchange reactions. When modeled as an open system (with exchange reactions), A is decoupled from the other metabolites and reconstruction fails at this point. Dashed lines mark enzyme inhibition interactions, larger arrows to the right indicate faster forward than backward reactions.

speed might be drastically affected, but the topological effects of this reaction chain on the correlation structure remain unchanged. In contrast, when exchange reactions are introduced (Figure 4.7B), the inhibition decouples A from the other metabolites and the reconstruction fails for the connectivity of this metabolites.

To further investigate this relationship, we performed simulations using standard parameter values for enzyme kinetics with ranging values of the inhibition parameter K_i, once without and once including exchange fluxes for intermediate metabolites (Figure 4.8). As discussed above, if no exchange fluxes are present, the inhibition strength does not significantly affect discrimination quality. Mass-flow has to be routed through the metabolite chain, independent of any feedback effects. Note that for K_i values below 10^{-3}, the system reaches steady states only after a very long simulation time and A grows unusually large. We do not expect such situations to occur in a real biological system, since a very strong inhibition essentially corresponds to a full blocking of the pathway at the respective reaction. Without further exchange reactions, metabolite A would drastically accumulate in the system which will in most cases not be desirable. If exchange fluxes are introduced, the inhibitory influence decouples A from the remaining metabolites if K_i falls below a certain threshold. In the plot, we observe one of the blue circles (representing the partial correlation between A and B) reaching noise levels for $K_i = 10^{-1}$ and below. Conclusively, we need to keep in mind that strong inhibitory feedbacks might impair the reconstruction process and lead to false negative results.

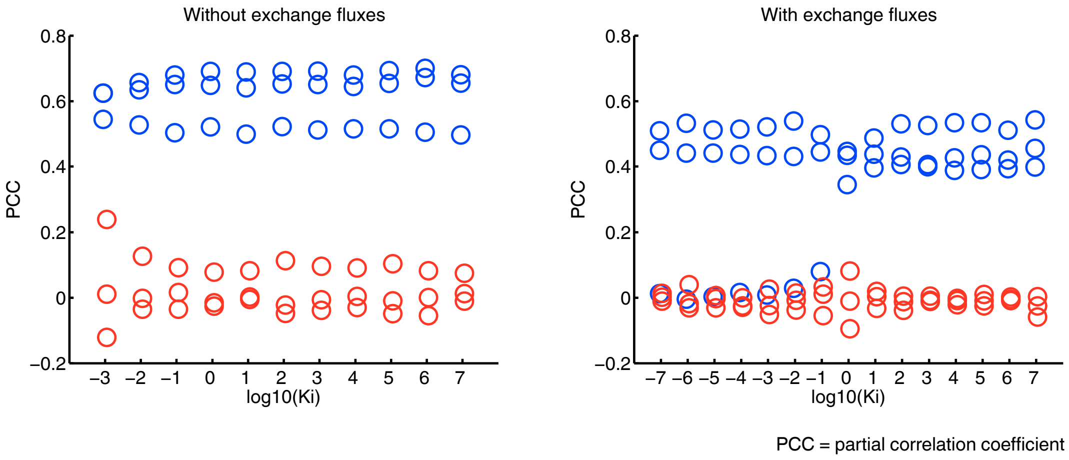


Figure 4.8: Influence of the inhibition strength on metabolic network reconstruction. Blue circles represent directly connected metabolites, red circles indicate distant metabolites in the underlying network. If no exchange fluxes are present (left) the inhibition does not affect reconstruction at all. With exchange fluxes (right), reconstruction will be impaired if the inhibition is sufficiently strong (low K_i).

4.7 All systems exhibit unique steady states

For all systems used in our study it was necessary to verify that they can only only exhibit a single stable steady state. Multistability is an import aspect in many biological systems, see Tyson et al. [102], Craciun et al. [103], Huang et al. [104] to name but just a few. In our case, however, multiple equilibria could result in to false correlation patterns, possibly leading to misinterpretations of the reconstructed pathway reactions. Therefore we used the ERNEST toolbox [105] to structurally verify uniqueness of a single steady state independent of actual parameter assignments. In the following, we discuss the results for all biochemical networks. For a detailed discussion of the deficiency theorems and *strong sign determination*, we refer the reader to the respective original publications cited below.

- All first-order networks in our study show a deficiency of zero [106] and thus cannot exhibit more than one positive steady states.

- For the enzyme-driven networks with reversible Michaelis Menten kinetics and the bimolecular split network, the deficiency zero theorem does not hold. However, the deficiency one theorem [106] applies in this case, and the stoichiometric matrix is *strongly sign determined* (SSD) [107], so the networks do not have the capacity for multiple equilibria.

- The stoichiometry matrix of the cofactor network (Figure 4.1F) is SSD, so this system also gives rise to a single steady state for any given parameter combination.

- The mixed-inhibition enzyme mechanisms are known to constitute bistability [103]. However, in the two parameter combinations reported in this study, several parameters differ by several orders of magnitude. Our toy models do not cover such a huge range of parameters and, more importantly, such parameter differences are also not expected to occur in a human population cohort with similar metabolic states (namely fasting). In order to further ensure monostability for our parameter ranges, we let the system run from 1000 different initial states each, for 100 parameter combinations in reasonable parameter ranges. Initial values were uniformly drawn between zero and 1000 times the metabolite concentrations of the detected steady state. We did not encounter a single case where a system with identical parameters ended up in a different steady state for a different initial value setting.

4.8 Conclusion

In this chapter, we set up a series of biochemical reaction systems in order to evaluate the reconstruction capabilities of GGMs. The advantage of such toy systems is that we know both the input and the expected output of the method, and can thus objectively assess which dynamical systems a GGM can reconstruct and where possible problems might occur.

We deduced a set of important aspects to be considered when interpreting partial correlation coefficients in reaction systems: (a) Metabolites in equilibrium due to reversible reactions can readily be recovered, whereas irreversible reactions might pose a substantial problem for correlation-based reconstruction attempts (in accordance with Camacho et al. [49]). (b) Input and output reactions for intermediate metabolites, however, improve the reconstruction accuracy. Such exchange reactions are likely to be present for most naturally occurring metabolites due to highly interconnected metabolic pathways. (c) Metabolite connectivity in cofactor-driven networks can be accurately reconstructed. The presence of exchange reactions for cofactors, as they are likely to be present in real systems, has substantial impact on the reconstruction quality. The connectivity of the cofactors themselves, however, remains spurious. (d) Non-linear rate laws and antagonistic, correlation-generating mechanisms might impair reconstruction quality. (e) With an increasing amount of fluctuations on the input reaction, the partial correlation difference between direct and indirect interactions increases for certain network topologies (e.g. for the irreversible linear metabolite chains). This indicates that a high heterogeneity of metabolic states in a population data set like the KORA cohort might be beneficial rather than problematic for our approach. (f) Saturation effects in enzyme-catalyzed reactions do not pose a problem for the reconstruction process. However, inhibitory influences in metabolic modules that include exchange reactions might decouple metabolites and lead to false negative results.

Taken together, the results on simulated biochemical reaction systems are promising and encourage GGM application to real metabolomics datasets. We generated a general overview of which network wirings can be reconstructed, and where problems are to be expected. Furthermore, we have seen that stronger variation on the system inputs is rather beneficial for the reconstruction than impairing it. This represents the modeling evidence for the 'stronger variation, stronger association' hypothesis we postulated in Chapter 1.4. In the following chapters, we will generate GGMs on real metabolomics datasets from the KORA population cohort, along with biological interpretations of the generated models.

Chapter 5

GGMs reconstruct pathway reactions from high-throughput metabolomics data

After evaluating the general capability of Gaussian graphical models (GGMs) to distinguish between direct and indirect reactions, we now focus on a real metabolomics data set. In the following, we estimate a GGM using targeted metabolomics data from the German population study KORA [33] ("Kooperative Gesundheitsforschung in der Region Augsburg"), see Chapter 2. We here used the dataset measured using the Biocrates AbsoluteIDQ kit, containing 1020 targeted metabolomics fasting blood serum measurements with 151 quantified metabolites. The metabolite panel includes acyl-carnitines, four classes of phospholipid species, amino acids and hexoses.

We will see that the GGM is sparse in comparison to the corresponding Pearson correlation network, displays a modular structure with respect to different metabolite classes, and is stable towards changes in the underlying data set. We demonstrate that top-ranking metabolite pairs and further densely connected subgraphs in the GGM can be attributed to known reactions in the human fatty acid biosynthesis and degradation pathways. In order to systematically verify this finding, we map partial correlation coefficients to the number of reaction steps between all metabolite pairs based on a literature-curated fatty acid pathway model. We observe statistically significant discriminatory features of GGMs to distinguish between directly and non-directly interacting metabolites in the metabolic network. In addition, low-order partial correlations represent a suitable alternative to full-order GGMs for the present dataset. Finally, we will summarize and discuss the relevance of GGMs for metabolomics data sets, point out limitations of the method and suggest future steps.

All results reported in this chapter are part of the following publication:

⋆ **Krumsiek, J.**, Suhre, K., Illig, T., Adamski, J., and Theis, F.J. Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data. *BMC Syst Biol*, 5(1):21, 2011

5.1 Network modularity calculation

In the following, we describe the network modularity calculation procedure required in Section 5.2. We define the adjacency matrix ξ_{ij} of a new unweighted, undirected graph induced by all significantly positive partial correlations in ζ_{ij}:

$$\xi_{ij} := \begin{cases} 1, & \text{if } p(\zeta_{ij}) \geq \tilde{\alpha} \\ 0, & \text{else} \end{cases},$$

where $\tilde{\alpha}$ represents the significance level after multiple testing correction. Now let $(V_1, \ldots, V_6)$ be the partitioning of the metabolites into the six metabolite classes: acyl-carnitines, diacyl-PCs, lyso-PCs, acyl-alkyl-PCs, sphingomyelins and amino acids (the hexose is left out as only a single metabolite belongs to that class). We calculate the *relative out-degree* $R_{ij} \in \mathbb{R}^{6\times 6}$ from each class to the other classes (i.e. the proportion of edges each class shares with the other classes) as:

$$R_{ij} := \frac{\mathcal{A}(V_i, V_j)}{\mathcal{A}(V_i, V)},$$

where $\mathcal{A}(V', V'') = \sum_{i \in V', j \in V''} \xi_{ij}$ represents the total number of edges between V' and V'', and $V = \bigcup V_i$ contains all metabolites in the network. The total network modularity Q of the network can be quantified according to White and Smyth [108] as:

$$Q := \sum_{i=1}^{6} \left[\frac{\mathcal{A}(V_i, V_i)}{\mathcal{A}(V, V)} - \left(\frac{\mathcal{A}(V_i, V)}{\mathcal{A}(V, V)} \right)^2 \right]. \tag{5.1}$$

Intuitively, this measure compares the within-class edges with the edges to the rest of the network. The more edges there are within each class in comparison to the other classes, the higher

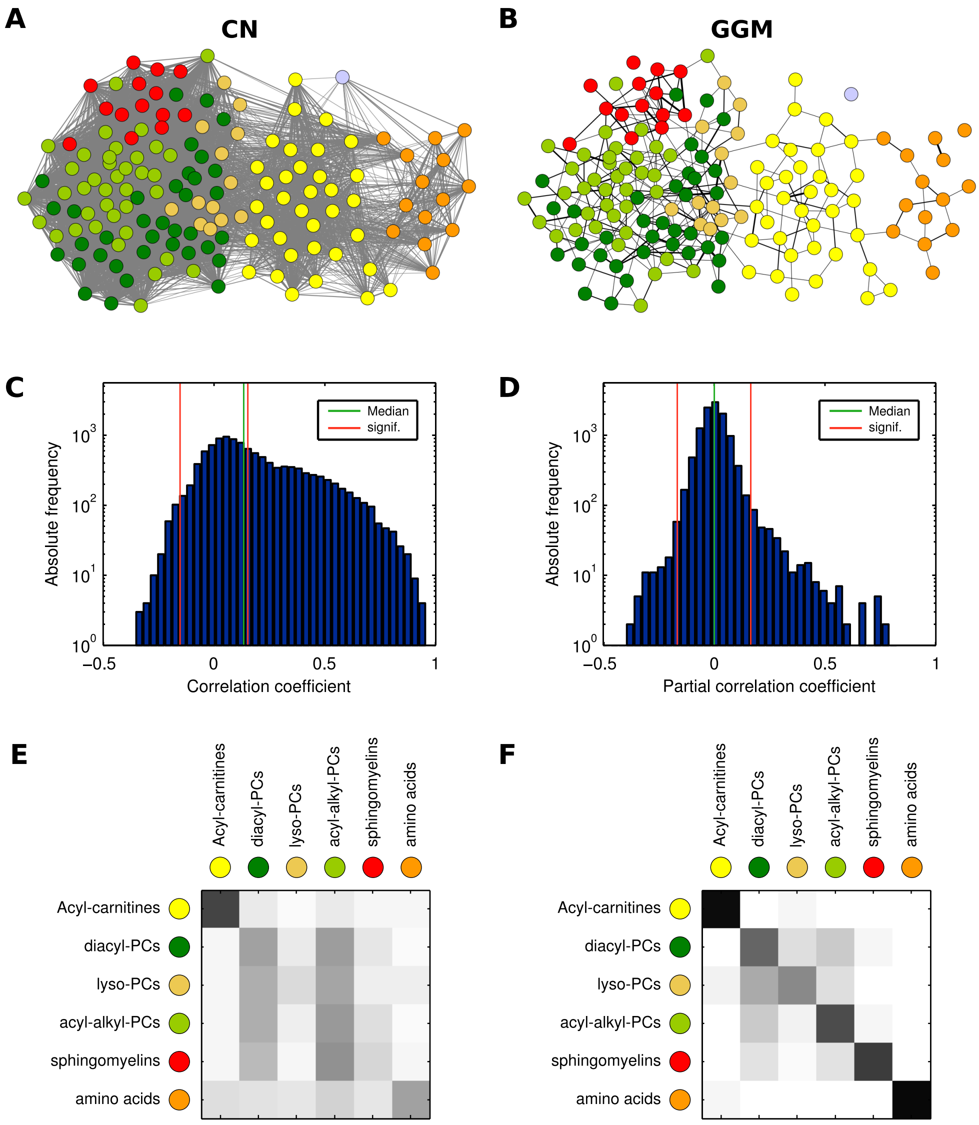


Figure 5.1: Network properties of the correlation network (CN) and Gaussian graphical model (GGM) inferred from the KORA Biocrates data set (1020 participants, 151 quantified metabolites). **A+B:** Graphical depiction of significantly positive edges in both networks, emphasizing local clustering structures. Each circle color represents a single metabolite class. **C+D:** Histograms of $\binom{151}{2} = 11325$ pairwise correlation coefficients (i.e. edge weights) for both networks. Green lines indicate the median values, red lines denote a significance level of 0.01 with Bonferroni correction. The CN displays a general bias towards positive correlations throughout all metabolites. For the GGM, the median value lies around zero and we observe a shift towards significantly positive values. **E+F:** Modularity between metabolite classes measured as the relative out-degree from each class (rows) to all other classes (columns). The GGM (right) shows a clear separation of metabolite classes, with some overlaps for the different phospholipid species diacyl-PCs, lyso-PCs, acyl-alkyl-PCs and sphingomyelins. Values range from white (0.0 out-degree towards this class) to black (1.0). PCs = phosphatidylcholines.

Q will be. Note that equation (5.1) can also be applied to weighted graphs. To assess the significance of the observed value, we perform graph randomization by edge rewiring [109, 110] and subsequent calculation of Q. During the rewiring process we randomly pick two edges from the network and exchange the target nodes of each edge. In order to achieve sufficient randomization, this operation is repeated $5 \cdot e$ times, where e represents the number of edges in the graph. To perform edge reshuffling on weighted graphs, we decided on a neighbor-preserving variant as described in Hartsperger et al. [111].

5.2 The GGM displays a sparse, modular and robust structure

Both regular Pearson correlation coefficients and partial correlation coefficients (see Chapters 3 and 4) were calculated on logarithmized metabolite concentrations. A manual inspection of QQ plots revealed that metabolite concentrations are usually closer to a log-normal than to a normal distribution[1]. All edges corresponding to correlation values significantly different from zero induce the networks displayed in Figure 5.1A+B.

In order to exclude correlation effects generated by genetic variation in the study cohort, we investigated the influence of the 15 SNPs reported in Illig et al. [27] on GGM calculation. If a SNP coordinately affects the concentrations of two metabolites, and the SNP is not included in the GGM analysis, a spurious correlation between the metabolites could occur (cf. Chapter 1.5). However, we found genetic effects on the resulting partial correlation coefficients to be neglectable (Figure 5.2). This indicates that Gaussian graphical models recover intrinsic properties of the metabolic system, and that effects of natural genetic variation are neglectable for our calculations.

Pearson correlation coefficients show a strong bias towards positive values in our data set (Figure 5.1C); a typical feature of high-throughput data sets also observed e.g. in microarray expression data, which can be attributed to unspecific or indirect interactions [51]. We obtain 5479 correlation values significantly different from zero with $\tilde{\alpha} = 8.83 \cdot 10^{-7}$ ($\alpha = 0.01$ after Bonferroni correction), yielding an absolute significance correlation cutoff value of 0.1619. In contrast, the GGM shows a much sparser structure with 417 significant partial correlations after Bonferroni correction (Figure 5.1D). Most values center around a partial correlation coefficient of zero, whereas we observe a clear shift towards positive significant values. Note that

[1] QQ plots can be downloaded from `http://helmholtz-muenchen.de/cmb/ggm`

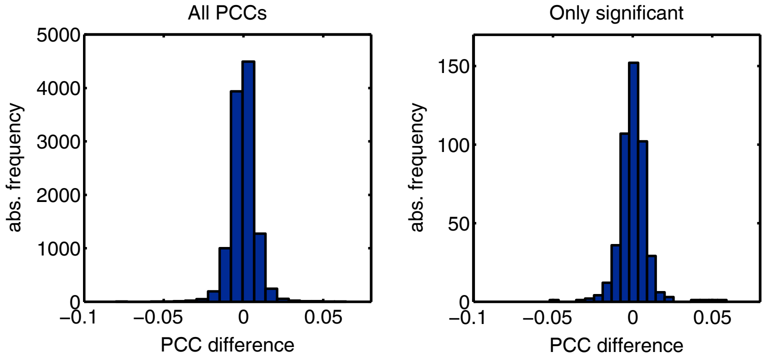


Figure 5.2: Effects of genetic variation on GGM calculation. SNP data were integrated with the metabolite concentrations by appending 15 further columns to the data set, thus extending the 1020 × 151 data matrix to a 1020 × 166 matrix. We then compared the 151 × 151 metabolite sub-matrix with the original partial correlation matrix without SNPs. Changes are generally small for all partial correlations ($-1.28 \cdot 10^{-5} \pm 8.09 \cdot 10^{-3}$, left histogram), and also when only investigating significant partial correlations ($3.8 \cdot 10^{-4} \pm 1.02 \cdot 10^{-2}$, right histogram).

negative partial correlations provide particular information that will be discussed at the end of Section 5.3.

The GGM displays a modular structure with respect to the seven metabolite classes in our panel, while the class separation in the correlation network appears rather blurry (Figure 5.1E+F). We observe a clear separation of the amino acids and acyl-carnitines from all other classes. The four groups of phospholipids (diacyl-PCs, lyso-PCs, acyl-alkyl-PCs, and sphingomyelins) still show locally clustered structures, but are strongly interwoven in the network. This is probably an effect of the dependence of all phospholipids on a similar fatty acid pool and, subsequently, the biosynthesis pathway acting on this substrate pool. In order to get an objective quantification of this observation, we calculated the group-based modularity Q on all significantly positive GGM edges according to Newman and Girvan [112]. The same measure was calculated for 10^5 randomized GGM networks (random edge rewiring). For the original GGM we obtain a modularity of $Q = 0.488$, and the random networks yield $Q = 0.118 \pm 0.016$, resulting in a highly significant z-score of $z = 23.49$. Furthermore, the modularity value induced by using the metabolite classes was compared to a partitioning optimized by simulated annealing. The optimized modularity is only slightly higher with $Q = 0.557$ and the resulting partitioning is very similar to the metabolite classes (results not shown). Performing the modularity analysis with the full, weighted partial correlation matrix produces equivalent results.

An important question for a multivariate statistical measure such as partial correlations is the robustness with respect to changes in the underlying data set. Furthermore, the dependence

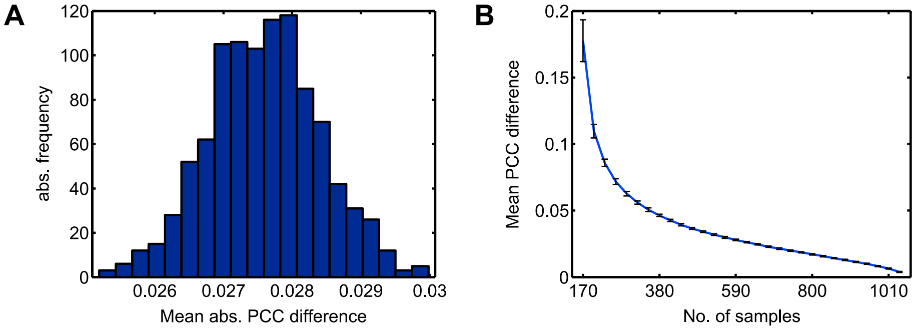


Figure 5.3: **A:** Mean differences of partial correlation coefficients from the values calculated for the original dataset, obtained by generating 1000 bootstrap samples and then calculating the mean absolute distance to the original correlations. Deviations from the original values are relatively low, indicating a high stability of PCC with respect to changes in the dataset. **B:** Mean differences from the original dataset for varying sample sizes. For each tested dataset size, the respective number of samples was randomly drawn from the original dataset 100 times (standard deviations are plotted in black).

of the measure on the size of the data set needs to be addressed. To answer these questions, we performed two types of perturbations of our data set. First, we applied sample bootstrapping with 1000 repetitions and compared the resulting partial correlations to the original data set (Figure 5.3A). We observe small mean absolute differences with low standard deviation ($0.03 \pm 8.2 \cdot 10^{-4}$). This indicates that for a large data set with $n = 1020$ samples, GGMs are robust against the choice of samples. We assume that each distinct metabolic state in the cohort is captured by a bootstrap sample, and thus all information required to calculate the GGM is contained. In addition to the bootstrap analysis, we estimated partial correlations for continuously decreasing sample sizes (Figure 5.3B). For each data set size we randomly picked samples from the original data set and repeated the procedure 100 times. The analysis shows that the GGM is stable even under decrease of the sample number. For instance, for a data set containing only around half of the original samples ($n = 530$) we obtain a mean absolute partial correlation difference of $0.03 \pm 6.9 \cdot 10^{-4}$. Only when the number of samples gets close to the number of variables ($m = 151$) the correlation matrix becomes ill-conditioned and strong differences from the original partial correlations occur. These problems of smaller metabolomics studies could be dealt with by regularization approaches or the usage of low-order partial correlations (Chapter 3.5). Taken together, these results indicate that the analyzed metabolomics data set is sufficient to robustly elucidate the statistical relationships between the measured metabolites.

Metabolite 1	Metabolite 2	PCC	Comment
Val	xLeu	0.821	Branched-chain amino acids
SM C18:0	SM C18:1	0.767	SCD/SCD5 desaturation
SM C16:1	SM C18:1	0.765	ELOVL6
PC ae C34:2	PC ae C36:3	0.752	2 reaction steps
SM (OH) C22:1	SM (OH) C22:2	0.743	sphingolipid-specific desaturation?
PC aa C34:2	PC aa C36:2	0.735	ELOVL1/ELOVL6 elongation
C10:0-carn	C8:0-carn	0.735	β-oxidation step
lysoPC a C16:0	lysoPC a C18:0	0.731	ELOVL6 elongation
PC aa C38:6	PC aa C40:6	0.709	ACOX1/3 + various ELOVLs
SM (OH) C14:1	SM (OH) C16:1	0.686	sphingolipid-specific elongation?
PC aa C36:4	PC aa C38:4	0.672	ACOX1/3 + various ELOVLs
PC aa C32:1	lysoPC a C16:1	0.661	C16:0/C16:1 phospholipid association
PC aa C38:5	PC aa C40:5	0.653	various ELOVLs
PC ae C34:3	PC ae C36:5	0.607	at least 3 reaction steps
PC aa C36:5	PC aa C38:5	0.596	ACOX1/3 + various ELOVLs
SM C24:0	SM C24:1	0.577	sphingolipid-specific desaturation?
PC ae C32:1	PC ae C32:2	0.574	SCD/SCD5 desaturation
SM (OH) C22:2	SM C24:1	0.567	possible elongation intermediate
C18:1-carn	C18:2-carn	0.561	β-oxidation intermediate

Table 5.1: Top 20 positive GGM edge weights (i.e. partial correlation coefficients, PCC) in our data set along with proposed metabolic pathway explanations. Most metabolite pairs can be directly linked to reactions in the fatty acid biosynthesis pathway, the β-oxidation pathway or amino acid-associated pathways.

5.3 Strong GGM edges represent known metabolic pathway interactions

The next step in our analysis was the manual investigation of metabolite pairs displaying strong partial correlation coefficients. Remarkably, we are able to provide pathway explanations for most metabolite pairs in the top 20 positive partial correlations (Table 5.1). In the following, we will specifically discuss interesting, high-scoring metabolite pairs along with their responsible enzymes in the metabolic pathways.

The highest partial correlation in the data set with $\zeta = 0.821$ is found for the two branched-chain amino acids valine and xLeucine, where the latter compound represents both leucine and isoleucine (which have equal masses and are not distinguishable by the mass-spectrometry approach used for this study). The three metabolites are in close proximity in the metabolic network concerning their biosynthesis and degradation pathways. Further related amino acid pairs that display significant partial correlations are histidine and glutamine ($\zeta = 0.383$), glycine and serine ($\zeta = 0.326$) as well as threonine and methionine ($\zeta = 0.298$).

Clear-cut signatures of the desaturation and elongation of long chain fatty acids can be seen for various sphingomyelins and lyso-PCs (Figure 5.4A). For example, SM C18:0 and SM C18:1 strongly associate with $\zeta = 0.767$, most probably representing the initial $\Delta 9$ desaturation step of the polyunsaturated fatty acid biosynthesis pathway from C18:0 to C18:1-$\Delta 9$ by SCD (*Steaoryl-CoA desaturase*). The similarly high partial correlation between SM C16:1 and SM C18:1 ($\zeta = 0.765$) as well as lysoPC a C16:1 and lysoPC a C18:1 ($\zeta = 0.315$) can be attributed to the ELOVL6-dependent elongation from C16:1-$\Delta 9$ to C18:1-$\Delta 11$. Interestingly, this reaction is not contained in the public reaction databases but has been previously described by Matsuzaka et al. [113].

We identify a variety of strong GGM edges between diacyl-PC (lecithins, PC aa) and acyl-alkyl-PC (plasmalogens, PC ae) metabolite pairs (Figure 5.4B). For instance, PC aa C34:2 and PC aa C36:2 associate strongly with $\zeta = 0.735$, and PC aa C36:4 and PC aa C38:4 show a partial correlation of $\zeta = 0.672$. While the first pair can be precisely explained by an elongation from C16:0 to C18:0 by ELOVL6, different combinatorial variants come into play for the PC aa C36:4/PC aa C38:4 pair. Our mass-spectrometry technique only measures *brutto* compositions, that is the bulk side chain carbon content and total degree of desaturation. Depending on the exact composition of both fatty acid residues in the respective lipids, this association could be caused by long-chain elongations (C14 to C16 and C16 to C18 through fatty acid synthase and ELOVL6, respectively), by very-long-chain elongations (C22:4 to C24:4 through ELOVL2 or ELOVL5) and even by peroxisomal β-oxidation of fatty acids (through ACOX1 or ACOX3). An interesting situation arises for the phospholipids PC ae C34:2, PC ae C36:3 and PC ae C36:2. From its brutto formula the latter species could represent an intermediate step between the other two metabolites. However, it associates poorly with both other phospholipids, which in turn display a strong partial correlation ($\zeta = 0.752$). This finding can be explained by distinct fatty acid side chain compositions, showing differential incorporation of C18:0, C18:1 and C18:2 (Figure 5.4B, bottom).

For the acyl-carnitine group we observe a remarkably high partial correlation of $\zeta = 0.735$ for C8-carn and C10-carn and further acyl-carnitine pairs with a carbon atom difference of two (Figure 5.4C). These associations can be attributed to the β-oxidation pathway, i.e. the catabolic breakdown of fatty acids in the mitochondria [97]. During this degradation process, C_2 units are continuously split off from the shrinking fatty acid chain. Four *acyl-CoA dehydrogenases*, ACADS, ACADM and ACADL, ACADVL, catalyze the rate limiting reactions of β-oxidation for different fatty acid chain lengths [43, 114]. Our interpretation of acyl-carnitine correlations as signatures of mitochondrial β-oxidation is in accordance with Illig et al. [27], where asso-

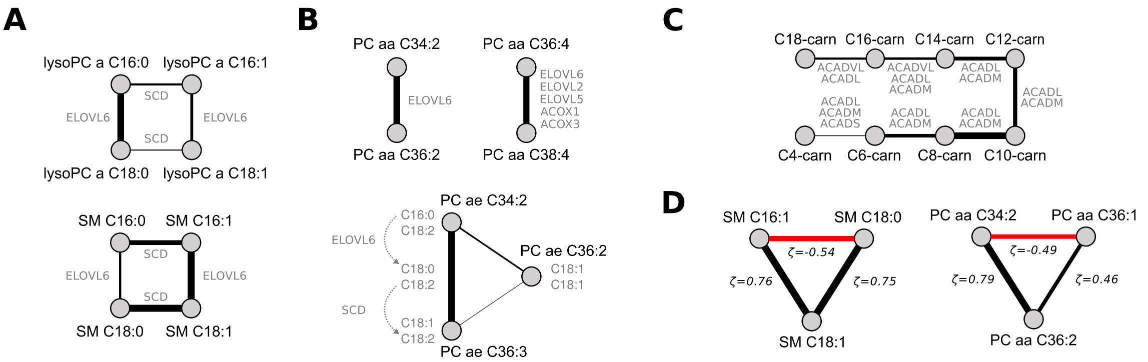


Figure 5.4: Biochemical subnetworks identified by the GGM. Line widths correspond to partial correlation coefficients. **A:** Elongation and desaturation signatures, most likely mediated by ELOVL6 and SCD, for C16 and C18 fatty acids incorporated in lyso-PCs and sphingomyelins. **B:** Top: Diacyl-phosphatidylcholine (PC aa) species with elongation and peroxisomal β-oxidation associations. Several combinatorial variants of side chain compositions are possible for C36:4 and C38:4, and thus different enzymes could mediate this connection. Bottom: Alkyl-acyl-phosphatidylcholines (PC ae) with supposedly distinct side chain composition, giving rise to a low association with a directly connected species (C36:2). **C:** Recovered β-oxidation pathway from C18 down to C4. Four enzymes with overlapping substrate specificities catalyze the rate-limiting reactions of this pathway. **D:** Two high-scoring triads, where metabolite pairs with a pathway distance of two constitute strong partial correlations. This feature of partial correlations aids in the reconstruction of the network topology beyond the direct neighborhood of each metabolite.

ciations between C8+C10, C12 and C4 with genetic variation in the ACADM, ACADL and ACADS loci, respectively, were identified.

We observe several associations that are not directly attributable to enzymatic interactions in the fatty acid biosynthesis or degradation pathways. For instance, lysoPC a 18:1 and lysoPC a 18:2 share a strong GGM edge ($\zeta = 0.543$) although the Δ12-desaturation step from oleic acid to linoleic acid is known to be missing in humans [115]. This missing reaction gives rise to the *essentiality* of fatty acids in the ω-6 unsaturated fatty acid pathway. A functional explanation could be a systemic equilibrium between the two fatty acids or remodeling processes specific for the lyso-PC metabolite class. Further examples are high partial correlations between the hydroxy sphingomyelins SM (OH) C22:1 and SM (OH) C22:2 ($\zeta = 0.743$) as well as the sphingomyelins SM C24:0 and SM C24:1 ($\zeta = 0.577$). To the best of our knowledge, there is no evidence for such fatty acid desaturation reactions in humans. The detected associations might therefore represent novel pathway interactions recovered by the Gaussian graphical model.

Negative values play a particular role in the interpretation of partial correlations coefficients. On the one hand, they obviously occur whenever regular negative correlations are involved. Mechanisms giving rise to negative correlations are, for example, coparticipation in the same reaction (cf. Figure 4.1E), mass conservation relations [49] or opposing regulatory effects. It is to be noted, however, that negative correlations are rare in our specific metabolomics data set (cf. Figure 5.1C). On the other hand, due to the mathematical properties of partial correlation coefficients, negative partial correlation coefficients occur whenever two metabolites A and B have a strong correlation with a third metabolite C, but do not share a high correlation value with each other. Two examples from our data set are shown in Figure 5.4D. First, SM C18:0 is negatively partially correlated with SM C16:1, and both of these in turn are highly positively partially correlated with SM C18:1. The fatty acids C16:1 and C18:0 have no direct connection in the pathway, causing the strong negative partial correlation value. A similar situation can be found for three diacyl-PCs: PC aa C34:2 and PC aa C36:1 show a high partial correlation with PC aa C36:2, but a negative partial correlation with each other. Again, there is no possible direct reaction from a C34:2 lipid species to a C36:1 species. Not all metabolite triads in the network show such a one-negative/two-positive motif. But if present, they provide another step in the reconstruction of metabolic pathways (beyond the direct neighborhood of each metabolite) by detecting metabolites which are exactly two steps apart. Furthermore, the *d-separation* rule from graphical modeling theory suggests that for these cases we can infer directionality of the associations. The only directed topology compatible with this correlation structure is from the two uncorrelated variables towards the third one. For more information on the methodology, we refer the reader to Freudenberg et al. [59].

5.4 Establishment of a literature-curated fatty acid pathway model

The analyses from the previous section strengthened our conception that a GGM inferred from blood serum metabolomics data represents true metabolite associations. To systematically assess how GGM edges and pathway proximity between our lipid metabolites are related, we generated a literature-based model of fatty acid biosynthesis (Figure 5.5A). Pathway reactions of the human fatty acid metabolism were drawn from three independent databases: (1) *H. sapiens Recon 1* from the BiGG databases (confidence score of at least 4) [116], (2) the Edinburgh Human Metabolic Network reconstruction [117] and (3) the KEGG PATHWAY database [43] as of July 2010. The reaction set was subdivided into two groups: (1) Fatty acid biosynthesis reactions which apply to the metabolite classes lyso-PC, diacyl-PC, acyl-alkyl-PC and sphingomyelins. (2) β-oxidation reactions representing fatty acid degradation to model reactions between the acyl-carnitines. The β-oxidation model consists of a linear chain of C2 degradation steps (C10→C8→C6 etc.). Fatty acid residues with identical masses, that cannot be

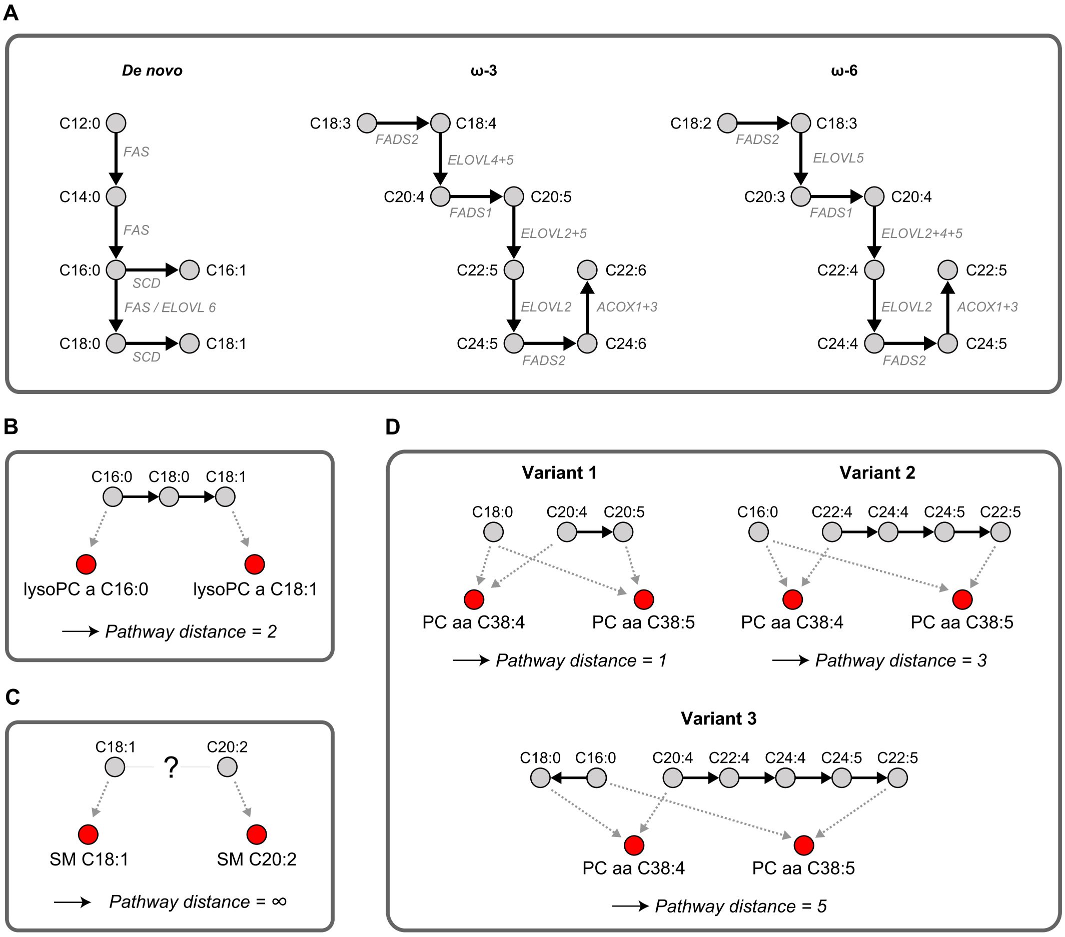


Figure 5.5: Fatty acid biosynthesis model and pathway distance calculation method. **A:** *De novo* synthesis of fatty acids with initial SCD-dependent desaturations (left), and the ω-3 and ω-6 poly-unsaturated fatty acid pathways (middle and right). Note that we omitted the specific positions of each double-bond since the mass-spectrometry technique in our study does not resolve positional information. **B:** Exemplary distance calculation on two lyso-PCs. We project lipid side chain compositions onto the respective fatty-acid biosynthesis reactions. Reaction reversibility is not taken into account in our calculation, i.e. distances are always symmetric. **C:** If no known pathway connection between two fatty acids exists, we assign a formal distance of infinity. **D:** For phospholipids that contain two fatty acid residues we need to take into account all combinatorial variants. We here show three variants for the connection between PC aa C38:4 and PC aa C38:5. In these examples, PC aa C38:4 could either consist of C18:0+C20:4 or C16:0+C22:4, while PC aa C38:5 could be C18:0+C20:5 or C16:0+C22:5. The shortest possible distance, one in this case, will be used for further calculations.

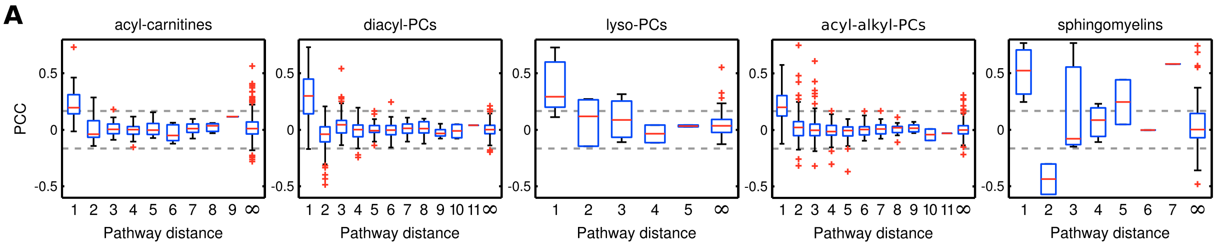
distinguished by our mass-spectrometry technology, are merged into a single metabolite in the reaction set. For instance, the polyunsaturated fatty acids C20:4Δ8,11,14,17 from the ω-3 pathway and C20:4Δ5,8,11,14 from the ω-6 pathway have identical numbers of carbon atoms and double bonds and are thus merged into a single metabolite C20:4.

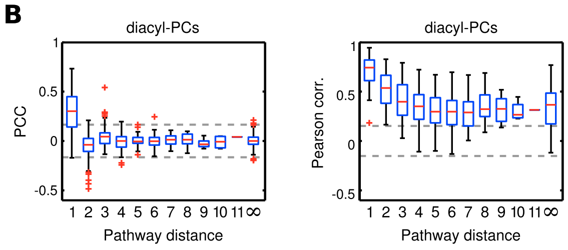
In the next step, we mapped the partial correlation coefficients from the KORA data set onto the minimal number of reaction steps between each pair of metabolites (*pathway distance*). Since our metabolite panel contains fatty-acid based lipids, we project the respective lipid compositions onto the fatty acid biosynthesis pathway (Figure 5.5B-D).

5.5 Partial correlation coefficients discriminate between directly and indirectly connected metabolites

We observe a strong tendency towards significantly positive partial correlations for a pathway distance of one, i.e. directly connected metabolite pairs, for all five metabolite classes (Figure 5.6A). In total, 86 out of 130 partial correlations (66%) for a pathway distance of one are significantly positive. For instance, for the lyso-PC class nearly all partial correlation coefficients for a pathway distance of one are above significance level, whereas most values for a distance of two or larger remain insignificant. Some outliers from this observation, however, require closer inspection: First, for some metabolite classes we observe negative partial correlation values for metabolite pairs that are exactly two steps apart in the metabolic pathway: 10 of 73 partial correlations in the diacyl-PC class and 2 of 2 partial correlations in the sphingomyelin class are significantly negative for a distance of two. These negative values are effects of the coregulated metabolite triads described previously in this chapter. Second, we find 91 of 932 ($\sim$ 9.8%) unconnected metabolite pairs (pathway distance = ∞) with a partial correlation above significance level. These pairs represent potentially novel pathway predictions, missing interactions in the model or effects upstream of the metabolic network like enzyme coregulation.

A direct comparison of both partial and Pearson correlation coefficients for the diacyl-phosphatidylcholine class is shown in Figure 5.6B. As described earlier in this chapter, we observe a general over-abundance of significant Pearson correlations independent of the actual pathway distance. Even for the metabolites without a known pathway connection, 1394 of a total of 1569 Pearson correlations are significant (88.85%, over all classes), in contrast to 131 out of 1569 for the partial correlations (8.35%).





C

Class	p-value	sens.	spec.	F_1
acyl-carn.	$1.85 \cdot 10^{-8}$	0.824	0.893	0.857
diacyl-PC	$4.49 \cdot 10^{-21}$	0.809	0.943	0.871
lyso-PC	$1.57 \cdot 10^{-3}$	0.800	0.753	0.776
acyl-alkyl-PC	$6.10 \cdot 10^{-19}$	0.759	0.936	0.838
sphingomyelins	$1.38 \cdot 10^{-2}$	1.000	0.705	0.827

Figure 5.6: Systematic evaluation of partial correlation coefficients versus pathway distances. Dashed lines in A and B indicate a significance level of 0.01 with Bonferroni correction. **A:** Pathway distances from our consensus model against partial correlation coefficients for the five lipid-based metabolite classes in our data set. We observe an enrichment of high partial correlations for a pathway distance of one, which rapidly drops for an increasing number of pathway steps. **B:** Comparison of partial correlation coefficients and Pearson correlation coefficients. Pearson correlation coefficients are generally high, independent of the actual pathway distance, indicating for systemic coregulation effects throughout the lipid metabolism. **C:** Wilcoxon rank sum test p-values between the partial correlation distributions of directly and indirectly connected pairs, and sensitivity/specificity/F_1 values measuring the discriminatory power to distinguish direct from indirect pairs.

The significantly different correlation value distributions between directly and indirectly linked metabolites (Figure 5.6A+B) barely provide a good quantification of the actual discrimination accuracy of this feature. Therefore, we assessed the discriminative power of partial correlations to tell apart direct from indirect interactions by means of *sensitivity* and *specificity*. The sensitivity evaluates which fraction of directly connected metabolites in the pathway are recovered by significant GGM edges, whereas the specificity states how many of the significant edges actually represent a direct connection. A commonly used tradeoff measure between sensitivity and specificity is the F_1 score, which is defined as the harmonic mean of both quantities, see Chapter 4.3. Figure 5.6C lists sensitivity, specificity and F_1 for all 5 metabolite classes along with an evaluation of partial correlation distribution differences between directly and indirectly linked metabolites (determined by Wilcoxon's ranksum test). F_1 values over 0.75 and significant p-values for the ranksum test indicate a strong discrimination effect of partial correlation coefficients concerning direct vs. indirect pathway interactions. Possible reasons for non-perfect sensitivity and specificity values will be discussed in detail at the end of this chapter.

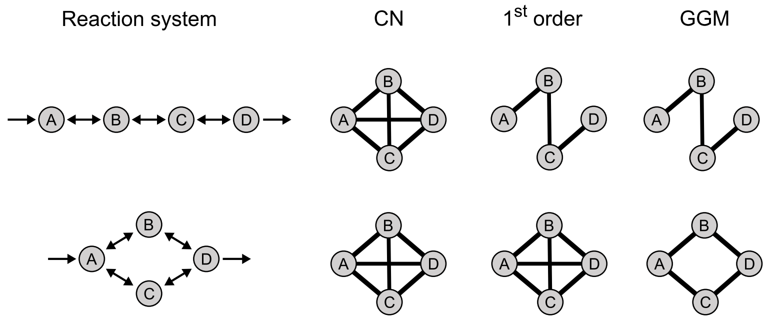


Figure 5.7: Illustration of low-order partial correlations in metabolic reaction systems. If there is more than one path between two nodes in the underlying network, first-order partial correlations cannot correctly reconstruct the topology.

5.6 Low-order partial correlations

The data set from our present study contained enough samples to calculate full-order partial correlations, i.e. pairwise correlations conditioned against all other n-2 metabolites. However, previous studies demonstrated that low-order partial correlation approaches can already be sufficient to elucidate direct interactions [55, 59].

We will exemplarily discuss the case of first-order partial correlations. Reconstruction results will be correct for two given nodes whenever the removal of one other node is sufficient to separate the two nodes in the underlying (true) graph. Removing nodes from the network is the graphical depiction of conditioning against variables in the underlying statistical dependence structure (recall the *Markov properties* described in Chapter 3). If multiple paths through the graph are possible between two nodes, conditioning against just one further node cannot be enough to rule out indirect effects. Consider the two example scenarios given in Figure 5.7. For the first network both first-order partial correlations and the GGM (in this case identical to second-order partial correlations) correctly reconstruct the network topology. For instance, A and D can be separated by conditioning on either B or C (or both). In the second network, however, removing just one node from the network is not sufficient to separate the indirectly connected nodes, and thus first-order partial correlations fail to reconstruct the correct topology. The same principles hold true for higher-order partial correlations. In general, in order for n-*th*-order partial correlations reconstruct the network topology correctly, any indirectly connected pair of nodes in the underlying graph must be separable by the removal of n nodes. Since we do not know the true dependency structure, the usage of GGMs is a simple and unbiased

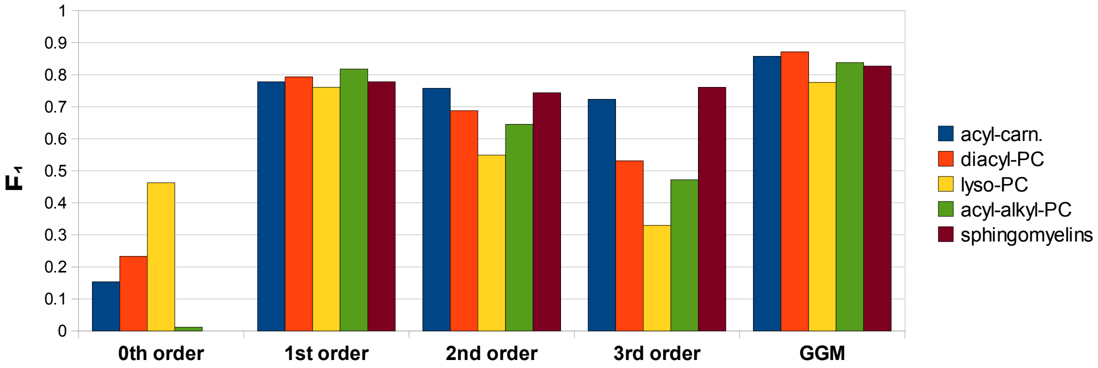


Figure 5.8: Comparison of low-order partial correlation approaches. As expected, 0th order partial correlations (regular Pearson correlations) provide only weak discrimination abilities between directly and indirectly connected metabolites. First-, second- and third-order partial correlations, however, perform reasonably well in comparison to the full-order GGM.

reconstruction approach; given that they can be properly estimated from the given amount of samples in the data set.

In order to assess how low-order partial correlations perform in comparison to the full-order GGM, we calculated first-, second- and third order partial correlations from the metabolomics data set using the approach described in de la Fuente et al. [55] (Figure 5.8). Surprisingly, especially first-order partial correlations worked remarkably well in discriminating direct from indirect interactions in the real data. This result provides two valuable pieces of information. First, low-order partial correlation approaches, which require much less samples to obtain stable estimates, appear to be a suitable alternative to GGMs for the metabolite panel used in this study. Second, the high relative scoring of first-order partial correlations provides insights into the correlation structures in the data set. In particular, this result indicates that the underlying metabolic pathways are primarily composed of acyclic, linear chains, which fits well to the fatty acid pathways dominating our measured lipid species.

We did not consider the application of further '$n<p$' GGM estimation algorithms introduced in Chapter 3 here. First, other algorithms working on the Markov properties will most likely generate very similar results to those obtained with the low-order partial correlations. Second, the covariance matrix will be well-estimated in a scenario with 1020 samples and 151 metabolites, so the benefit of shrinkage approaches will be marginal.

5.7 Conclusion

In this chapter, we addressed the reconstruction of metabolic pathway reactions from high-throughput targeted metabolomics measurements. Previous reconstruction approaches employed pairwise association measures, primarily standard Pearson correlation coefficients, to infer network topology information from metabolite profiles [48, 49, 118, 119]. We here demonstrated the usefulness of Gaussian graphical models and their ability to distinguish direct from indirect associations by estimating the conditional independence structures between variables. GGMs are based on partial correlation coefficients, that is the Pearson correlation between two metabolites corrected for the correlations with all other metabolites.

We inferred both a GGM and a regular correlation network from a large-scale metabolomics data set with 1020 standardized samples from overnight fasting individuals. We investigated the influence of the 15 genome-wide-significant SNPs from this study on our GGM and demonstrated that genetic variation in the general population is neglectable for partial correlation calculation. We found that the GGM displays a much sparser structure than regular correlation networks. Only around 400 partial correlation values were above significance level ($\sim$3.6%), whereas half of all Pearson correlation values were significant after Bonferroni correction. This depicted the nature of partial correlation coefficients to neglect indirect associations between distantly related metabolites. We detected a strongly modular structure in the GGM with respect to the different metabolite classes, except for the four types of phospholipids which appear slightly interwoven. This provides a unique picture of the separation of metabolic pathways (synthesis, degradation and amino acid metabolism), but also the interaction between different lipid classes dependent on a single intracellular fatty acid pool. Finally, GGMs were stable with respect to both choice and number of samples in the data set. Even a smaller data set with only a few hundred samples would have been sufficient to achieve the results from this study. The estimation of GGMs for data sets with less samples than metabolites is possible (see Chapter 3.5), but deviations from the true partial correlation coefficients have to be expected.

Manual investigation of high-scoring substructures in the GGM revealed groups of metabolites that could be directly attributed to reaction steps from the human fatty acid biosynthesis and degradation pathways. We detected effects of ELOVL-mediated elongations and FADS-mediated desaturations of fatty acids as well as signatures of the catabolic β-oxidation pathway. For instance, our method successfully recovered a direct elongation from C16:1 to C18:1, which has been experimentally shown by Matsuzaka et al. [113] but is not present in the public reaction databases. Furthermore, we identified highly negative partial correlations as an indication for a pathway distance of two, serving as a further hint in the reconstruction of metabolic network topology. In order to systematically evaluate whether high partial correlations represent direct interactions, we generated a consensus model of fatty acid biosynthesis reactions from three publically available reaction databases. By mapping partial correlation coefficients to the number of reaction steps between two metabolites we detected a statistically significant enrichment of high values for a pathway distance of one. We calculated a high accuracy for partial correlations to discriminate between directly and indirectly associated metabolites, as measured by sensitivity, specificity and the F_1 measure. Interestingly, we could show that the discrimination quality of low-order partial correlations [55], especially the first-order variants, is close to the full-order GGM. Even though this might be a feature specific to the metabolite panel used in this study, low-order partial correlations represent a suitable alternative especially for studies with only few samples. If more samples than variables are available, full-order GGMs as an unbiased approach conditioning against as many parameters as possible should be preferred.

Further analyzing those edges in the GGM that could *not* be explained by known pathway interactions represents a promising task for future analyses. Specifically, the edges that were considered *false positive* in our modeling framework, i.e. high partial correlations without a known reaction evidence, will be of particular interest. Several cases were exemplarily discussed in this chapter, e.g. the strong GGM edge between lysoPC a 18:1 and lysoPC a 18:2, which cannot be explained by the common fatty acid pathway, or the association between SM C24:0 and SM C24:1, for which there is no evidence in the known human metabolic network. Removing all explainable edges from our GGM will generate a list of pathway hypotheses which could then be subjected to further examination and validation.

Interestingly, the metabolomics data used in this study originated from human blood, while we could infer strong signatures of intracellular and even inner-mitochondrial processes. Previous studies on blood plasma samples detected similar relationships with cellular processes based on genetic associations [27] and case/control drug trials [120]. In this work we could now show that blood metabolite profiles alone are sufficient to capture the dynamics of metabolic pathways.

However, GGMs can never provide a perfect reconstruction of the underlying system. There are several factors that lead to the absence of high partial correlations between interacting metabolites, that is false negative edges in the GGM: (a) Counterantagonistic correlation-generating processes and bimolecular reactions (see Chapter 4.2) might lead to the elimination of pairwise association; cf. Camacho et al. [49]. (b) The respective enzyme might not be active in the current metabolic state, or its effects on the respective metabolite pools are neglectable. (c) Contrary to our general finding that even blood plasma metabolites carry strong signatures of metabolic pathways, the signal might be diminished for certain types of metabolites. Furthermore, the actual origins of blood plasma metabolites, e.g. in terms of cell type or tissue activity leading to the detected metabolite signals, still remain to be unraveled. The above-mentioned mechanisms are possible explanations for the non-perfect sensitivity values observed in Figure 5.6C.

In conclusion, this initial study presented Gaussian graphical models as a valuable tool for the recovery of biochemical reactions from high-throughput targeted metabolomics data. The following two chapters of this thesis will use the GGM approach for the functional annotation of metabolites as well as the delineation of group-specific metabolome differences.

Chapter 6

Mining the unknown: A systems approach to metabolite identification

In this chapter, we present an approach to utilize Gaussian graphical models in combination with data on genetic variation, in order to derive functional annotations of unknown metabolites. Recently, genome-wide association studies (GWAS) on metabolic quantitative traits have proven valuable tools to uncover the genetically determined metabolic individuality in the general population [26–29, 121]. Interestingly, a great portion of the genetic loci that were found to significantly associate with levels of specific metabolites are within or in close proximity to metabolic enzymes or transporters with known disease or pharmaceutical relevance. Moreover, compared to GWAs with clinical endpoints, the effect sizes of the genotypes are exceptionally high.

While these previous GWAS focused on metabolic features with known identity, untargeted metabolomics approaches additionally provide quantifications of so-called ‘unknown metabolites’. An unknown metabolite is a small molecule that can reproducibly be detected and quantified in a metabolomics experiment, but whose chemical identity has not been elucidated yet. In an experiment using liquid chromatography (LC) coupled to MS, such an unknown would be defined by a specific retention time, one or multiple masses (e.g. from adducts), and a characteristic fragmentation pattern of the primary ion(s). An unknown observed by NMR spectroscopy would correspond to a pattern in the chemical shifts. Unknowns may constitute previously undocumented small molecules, such as rare xenobiotics or secondary products of metabolism, or they may represent molecules from established pathways which could not be assigned using current libraries of MS fragmentation patterns [122, 123] or NMR reference spectra [124].

The impact of unknown metabolites for biomedical research has been shown in recent metabolomics-based discovery studies of novel biomarkers for diseases and various disease-causing conditions. This includes studies investigating altered metabolite levels in blood for insulin resistance [22], type 2 diabetes [21], and heart disorders [125]. A considerable number of high-ranking hits reported in these biomarker studies represent unknown metabolites. As long as their chemical identities are not clarified, the usability of unknown metabolites as functional biomarkers for further investigations and clinical applications is rather limited.

In mass-spectrometry-based metabolomics approaches, the assignment of chemical identity usually involves the interpretation and comparison of experiment-specific parameters, such as accurate masses, isotope distributions, fragmentation patterns, and chromatography retention times [126–128]. Various computer-based methods have been developed to automate this process. For example, Rasche and colleagues [129] elucidated structural information of unknown metabolites in a mass-spectrometry setup using a graph-theoretical approach. Their approach attempts to reconstruct the underlying fragmentation tree based on mass-spectra at varying collision energies. Other authors excluded false candidates for a given unknown by comparing observed and predicted chromatography retention times [130, 131], or by the automatic determination of sum formulas from isotope distributions [132]. Furthermore, Gipson et al. [133] and Weber and Viant [134] integrated public metabolic pathway information with correlating peak pairs in order to facilitate metabolite identification. However, these methods might not be applicable for high-throughput metabolomics datasets that have been produced in a in fee-for-service manner, since the mass spectra as such might not be readily available.

Approaching the problem from a conceptually different perspective, we here present a novel functional metabolomics method to predict the identities of unknown metabolites using a systems biological framework. By combining high-throughput genotyping data, metabolomics data, and literature-derived metabolic pathway information, we generate testable hypotheses on the metabolite identities based solely on the obtained metabolite quantifications (Figure 6.1). No further experiment-specific data such as retention times, isotope and fragmentation pattern are required for this analysis.

The concept of our approach is based on the following observations: As discussed above, GWAS with metabolic traits can reveal functional relationships between genetic loci encoding metabolic enzymes and metabolite concentration levels in the blood. Moreover, we have seen in the previous chapters that GGMs can identify biochemically related metabolites from high-throughput metabolomics data alone. These observations suggest that if an unknown compound displays a similar statistical association with a genetic locus in a GWAS or a known metabolite in a GGM, then this may provide specific information of where it is located in the metabolic network. Based on this information we can then derive testable hypotheses on the biochemical identity of the unknown metabolite. This annotation idea parallels classical concepts from functional genomics, where for instance co-expression between RNA transcripts is used to predict the function of poorly characterized genes [44, 135], or protein functions are inferred from protein-protein interactions [136].

In the following, we first conduct a full genome-wide association study on 655,658 genotyped SNPs with concentrations of 292 known and 225 unknown metabolites in fasting blood serum samples from the KORA F4 population (Metabolon data set, cf. Chapter 2). This dataset is less lipid-centered than the Biocrates dataset used in Chapter 5 and provides a broad coverage of metabolic pathways, including central energy metabolism, steroid hormones and xenobiotics. We then compute a Gaussian graphical model including both known and unknown metabolites. In a third step, we integrate the results of the GWAS and GGM computations and combine them with metabolic pathway information from public databases to derive predictions for a total of 106 unknown metabolites. In order to validate the approach, we investigate six distinct cases in detail. We derive specific identity predictions for a total of nine unknown metabolites, which we then confirm experimentally. Finally, we discuss the relevance of newly discovered genetic loci and unknown identity predictions in the context of existing disease biomarker discovery and pharmacogenomics studies.

All results reported in this chapter are part of the following publication:

⋆ **Krumsiek, J.**, Suhre, K., Evans, A.M., Mitchell, M.W., Mohney, R.P., Milburn, M.V., Wägele, B., Römisch-Margl, W., Illig, T., Adamski, J., Gieger, C., Theis, F.J., and Kastenmüller, G. Mining the unknown: A systems approach to metabolite identification. *PLoS Genetics*, 8(10):e1003005, 2012.

Furthermore, a patent application of the method has been filed:

⋆ Identity Elucidation of Unknown Metabolites. U.S. Patent Application No. 61503673 Unpublished, filing date Jul. 1, 2011. (Michael Milburn, applicant)

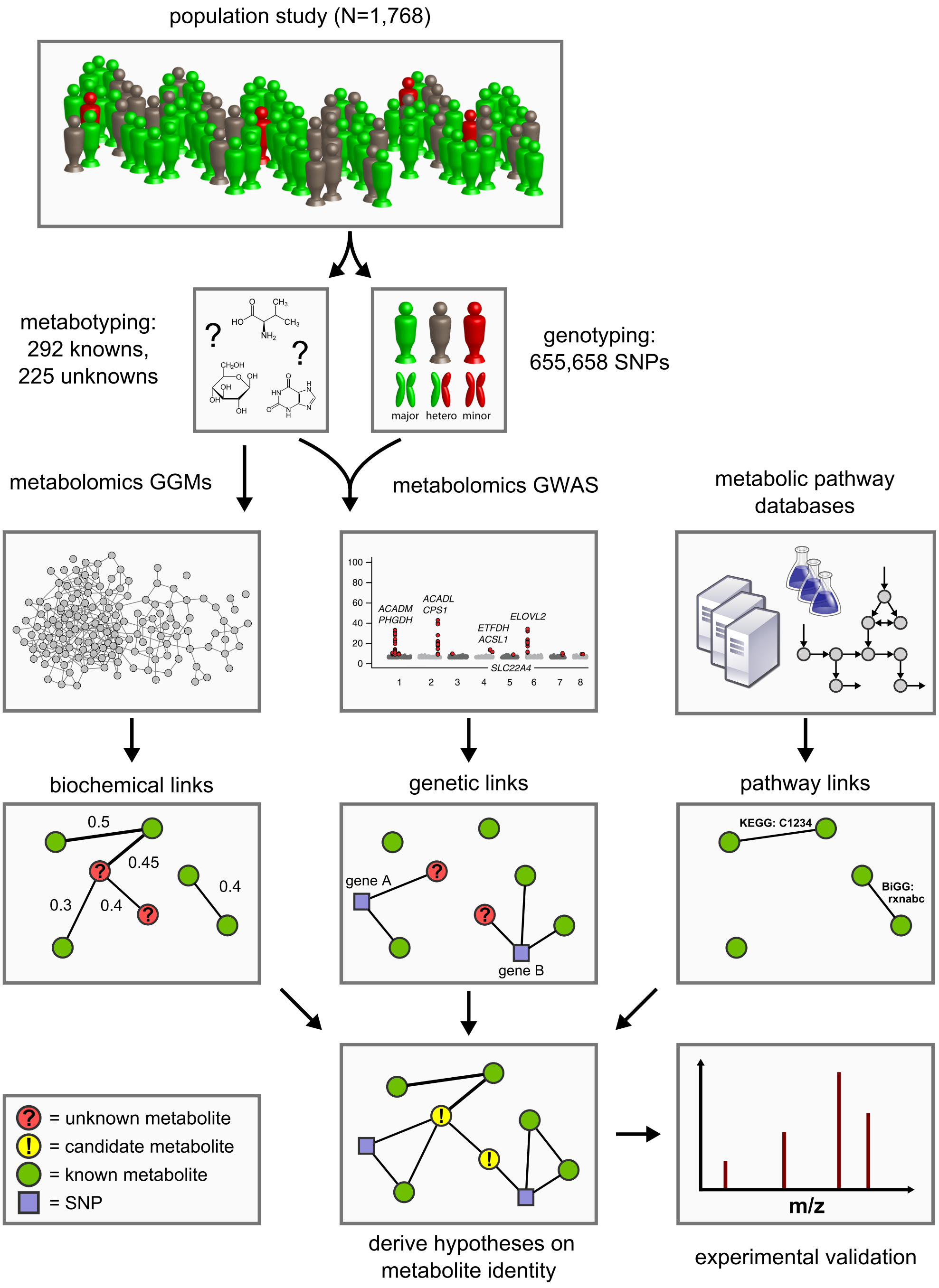


Figure 6.1: Data integration workflow for the systematic classification of unknown metabolites. We combine high-throughput metabolomics and genotyping data in Gaussian graphical models (GGMs) and in genome-wide association studies (GWAS) in order to produce testable predictions of the unknown metabolites' identities. These hypotheses are then subject to experimental verification by mass-spectrometry. Six such cases have been fully worked through and are presented in Table 6.3.

6.1 Methods

Gaussian graphical models

To ensure log-normality, we compared QQ-plots against normal distributions for both non-logarithmized and logarithmized metabolite concentrations (analogously to Chapter 5)[1]. All distributions were closer to log-normality than to regular normality, so we logarithmized the metabolite concentrations for the following analysis steps.

Age, gender and SNP effects were removed by adding the respective variables and SNPs states to the data matrix. Recall that for each pairwise correlation, GGMs automatically correct for all remaining variables in the data matrix. SNP states were coded as numerical values of 0, 1 and 2, such that linear regression calculation corresponds to an additive genetic model (see next section). Note that age, gender and SNPs were not investigated as an actual node in the network but merely used for the correction procedure, an inherent effect when adding variables to the GGM. For the later analysis steps, we then only considered metabolite-metabolite edges in the network.

Genome-wide associations

In order to avoid spurious false positive associations due to small sample sizes, only metabolic traits with at least 300 non-missing values were included and data-points of metabolic traits that lay more than 3 standard deviations off the mean were excluded by setting them to ‘missing’ in the analysis (leaving 273 known and 213 unknown metabolites). Genotypes are represented by 0, 1, and 2 for major allele homozygous, heterozygous and minor allele homozygous individuals, respectively. We employed a linear model to test for associations between a SNP and a metabolite assuming an additive mode of inheritance. Statistical tests were carried out using the PLINK software (version 1.06) [137] with age and gender as covariates. Based on a conservative Bonferroni correction, associations with p-values $< 1.6 \cdot 10^{-10}$ meet genome-wide significance, corresponding to a significance level of α=0.05. A SNP was associated with a gene whenever there was at least one other SNP lying in the transcribed region of this gene (from 5’UTR to 3’UTR) that displays an LD $\geq$ 0.8 with the query SNP. A detailed description of the GWAS procedure can be found in Suhre et al. [29].

[1] QQ plots can be downloaded from `http://helmholtz-muenchen.de/cmb/ggm`

Metabolic pathway model and functional annotations

Metabolic reactions were imported from three independent human metabolic pathway resources: (1) H. sapiens Recon 1 from the BiGG databases [116], (2) the Edinburgh Human Metabolic Network (EHMN) reconstruction [117] and (3) the KEGG PATHWAY database [43] as of January 2012. We attempted to create a highly accurate mapping between the different metabolite identifiers of the respective databases, in order to ensure the identity of each compound in our list. Entries referring to whole groups of metabolites, such as 'phospholipid', 'fatty acid residue' or 'proton acceptor' were excluded from our study. Furthermore, we did not consider metabolic cofactors such as 'ATP', 'CO_2', and 'SO_4' etc. in our analysis, since such metabolites unspecifically participate in a plethora of metabolic reactions. For each enzyme catalyzing one or more reactions in our pathway model, we retrieved functional annotations from two independent sources: (i) GO-Terms from the Gene Ontology [138] and (ii) enzyme pathway annotations from the KEGG PATHWAY database [43].

6.2 Genetic associations link unknown metabolites to functionally related genes

In the first step of our analysis, we conducted a GWAS with the concentrations of known and unknown metabolites, testing a total of 655,658 genotyped SNPs from the KORA cohort for association. In total, we observe 34 distinct loci that display metabolite associations at a genome-wide significance level (Figure 6.2). Out of these 34 loci, 15 associate with at least one unknown compound. From the 213 unknown metabolites analyzed here, 28 show at least one genome-wide significant hit. For 12 loci, an unknown compound constitutes the strongest association of all tested compounds. Seven of these loci (SLC22A2, COMT, CYP3A5, CYP2C18, GBA3, UGT3A1, rs12413935) have not been described in GWAS with metabolic traits previously and thus represent new genetic loci of metabolic individuality.

The genome-wide significant genetic associations that include at least one unknown compound are presented in Table 6.1. Based on the observation that metabolites associating with genetic variants in or near enzymes are likely to be functionally linked to these proteins, we used the GWAS data to derive hypotheses on the potential identity of the respective unknowns. For instance, the SNP rs296391 in close proximity to the SULT2A1 gene (*sulfotransferase family, cytosolic, 2A, dehydroepiandrosterone DHEA-preferring*) strongly associates with the concentrations of the unknown metabolites X-11440 and X-11244 ($p = 1.7 \cdot 10^{-43}$ and $p = 2.1 \cdot 10^{-26}$,

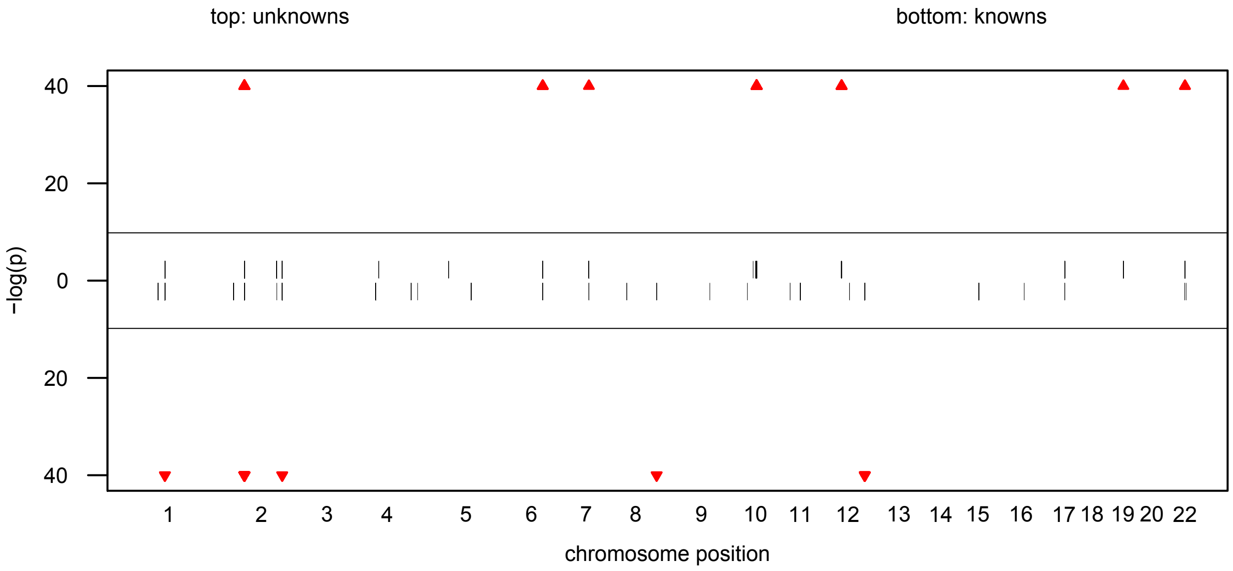


Figure 6.2: Manhattan plot of genetic associations. The strength of association for known (bottom) and unknown (top) metabolites is indicated as the negative logarithm of the p-value for the linear model. Only metabolite-SNP associations with p-values below 10^{-6} are plotted (grey circles). Red triangles represent metabolite-SNP associations with p-values below 10^{-40}. Horizontal lines indicate the threshold for genome-wide significance ($\hat{\alpha} = 1.6 \cdot 10^{-10}$) corresponding to $\alpha = 0.05$ after Bonferroni correction); vertical dashes indicate loci at which this threshold is attained.

Locus	Locus Info	Lead-SNP	Metabolite	p-value	Published associations	References
PYROXD2	pyridine nucleotide-disulfide oxidoreductase domain 2	rs4488133	X-12092 X-12093	$2.2 \cdot 10^{-281}$ $1.4 \cdot 10^{-27}$	trimethylamine (urine) / dimethylamine (plasma)	[121]
SLCO1B1	organic anion transporter family, bile acids	rs4149056	X-11529 X-11538 X-13429 X-12063 X-12456 X-14626	$3.3 \cdot 10^{-81}$ $1.4 \cdot 10^{-37}$ $4.9 \cdot 10^{-22}$ $5.2 \cdot 10^{-20}$ $8.4 \cdot 10^{-17}$ $2.1 \cdot 10^{-13}$	eicosenoate / tetradecanedioate	[29]
SLC22A2	solute carrier family 22 (organic cation transporter), member 2	rs316020	X-12798	$1.7 \cdot 10^{-72}$	*New*	-
NAT8	N-acetyltransferase 8	rs7598396	X-12510 X-11787 X-12093	$1.5 \cdot 10^{-56}$ $3.0 \cdot 10^{-37}$ $8.9 \cdot 10^{-22}$	N-acetylornithine	[29]
COMT	catechol-O-methyltransferase	rs4680	X-11593 X-01911	$1.1 \cdot 10^{-48}$ $5.8 \cdot 10^{-11}$	*New*	-
CYP3A5	cytochrome P450, family 3, subfamily A, polypeptide 5	rs10242455	X-12063	$1.5 \cdot 10^{-45}$	*New*	-
SULT2A1	sulfotransferase family, cytosolic, dehydroepiandrosterone-preferring	rs296391	X-11440 X-11244	$1.7 \cdot 10^{-43}$ $2.1 \cdot 10^{-26}$	dehydroepiandrosterone sulfates	[139]
UGT1A	UDP glucuronosyltransferase 1 family, polypeptide A complex locus	rs6742078	X-11530 X-11441 X-11793 X-11442	$2.1 \cdot 10^{-38}$ $5.6 \cdot 10^{-30}$ $2.6 \cdot 10^{-26}$ $1.2 \cdot 10^{-25}$	bilirubin (E,E) / oleoylcarnitine	[29]
ACADL	acyl-CoA dehydrogenase, long-chain	rs2286963	X-13431	$2.7 \cdot 10^{-33}$	C9 / C10:2	[27]
ACADM	acyl-CoA dehydrogenase, medium-chain	rs12134854	X-11421	$1.9 \cdot 10^{-27}$	C12 / C10, hexanoylcarnitine / oleate	[26, 27, 29]
CYP2C18	cytochrome P450, family 2, subfamily C, polypeptide 18	rs7896133	X-11787	$4.0 \cdot 10^{-26}$	*New*	-
GBA3	glucosidase, beta, acid 3 (cytosolic)	rs358231	X-11799	$2.9 \cdot 10^{-17}$	*New*	-
ACE	angiotensin I converting enzyme (peptidyl-dipeptidase A) 1	rs4343	X-14189 X-14208 X-14205 X-14304	$1.5 \cdot 10^{-16}$ $4.6 \cdot 10^{-15}$ $4.0 \cdot 10^{-14}$ $2.7 \cdot 10^{-12}$	aspartylphenylalanine	[5]
UGT3A1	UDP glycosyltransferase 3 family, polypeptide A1	rs13358334	X-11445	$2.4 \cdot 10^{-12}$	*New*	-
–	[no known gene locus]	rs12413935	X-06226	$4.0 \cdot 10^{-11}$	*New*	-

Table 6.1: Genome-wide significant associations ($p < 1.6 \cdot 10^{-10}$) involving unknown metabolites. We observe associations at 15 genetic loci that involve genes from various biological processes. Note that most of these genes code for proteins that are related to metabolic activities in the body, thereby providing information that allows to derive concrete hypotheses on the biochemical identity of each unknown. Previously published associations with known metabolites provide further evidence on specific parts of a pathway in which the unknown might be involved.

respectively). The enzyme encoded by SULT2A1, a bile salt sulfotransferase, converts steroids and bile acids into water-soluble sulfate conjugates for excretion [140]. Thus, we may speculate that X-11440 and X-11244 are biochemically related to steroids, bile acids, or water-soluble sulfate conjugates. Additional insights can be gained from genetic associations that involve both known and unknown metabolites. For instance, X-12510, X-11787, X-12093 and N-acetylornithine strongly associate with genetic variation at the NAT8 locus. NAT8 encodes the protein N-acetyltransferase 8. In this case, we may speculate that the unknowns represent similar substrates or products of the N-acetylation processes linked to this enzyme. Finally, we can link the results obtained here with results from other GWAS on metabolic traits. For example, the unknown metabolite X-13431 associates with a genetic variant in the ACADL (acyl-CoA dehydrogenase, long-chain) gene. This locus does not associate with any other metabolite in the present study, but was previously reported to associate with the medium-chain length carnitines C9 and C10:1 [26, 27]. Proteins from the ACAD family catalyze rate-limiting reactions in the β-oxidation pathway which generally associate with carnitines. This observation suggests that X-13431 may be a member of this medium-chain length carnitine family. These examples demonstrate that concrete information on the biochemical identity of unknown metabolites can be derived from our experimental dataset by using the GWAS approach.

6.3 Gaussian graphical modeling provides a biochemical context for unknown metabolites

In the second step of our analysis we focused solely on intrinsic relations between the measured metabolites and, in particular, on associations between known and unknown compounds. To this end, we again calculated a metabolomics Gaussian graphical model (see Chapter 5). In order to obtain a dataset that is independent of our genetic analysis, and to avoid circular arguments, co-variations in metabolite concentrations that are due to association with genetic variants (SNPs) were specifically removed from the data by adding the SNPs to the data matrix. A partial correlation was included in the model if it is significantly different from zero with α=0.05 after Bonferroni correction, yielding a corrected significance level of $\hat{\alpha} = 7.9 \cdot 10^{-7}$ and an absolute partial correlation cutoff of 0.178. The resulting GGM consists of a total of 399 out of 62,835 theoretically possible edges (0.64% connectivity, Figure 6.3A). In line with our previous observations from Chapter 5, metabolites tend to be strongly connected within their respective metabolic class, while links between different classes are rare. We obtained a modularity of $Q = 0.389$ and a randomized modularity of $Q = -0.0041 \pm 0.0222$, resulting in a z-score of $z = 17.71$ (compared to $Q = 0.488$ and $z = 23.49$ for the Biocrates data GGM in

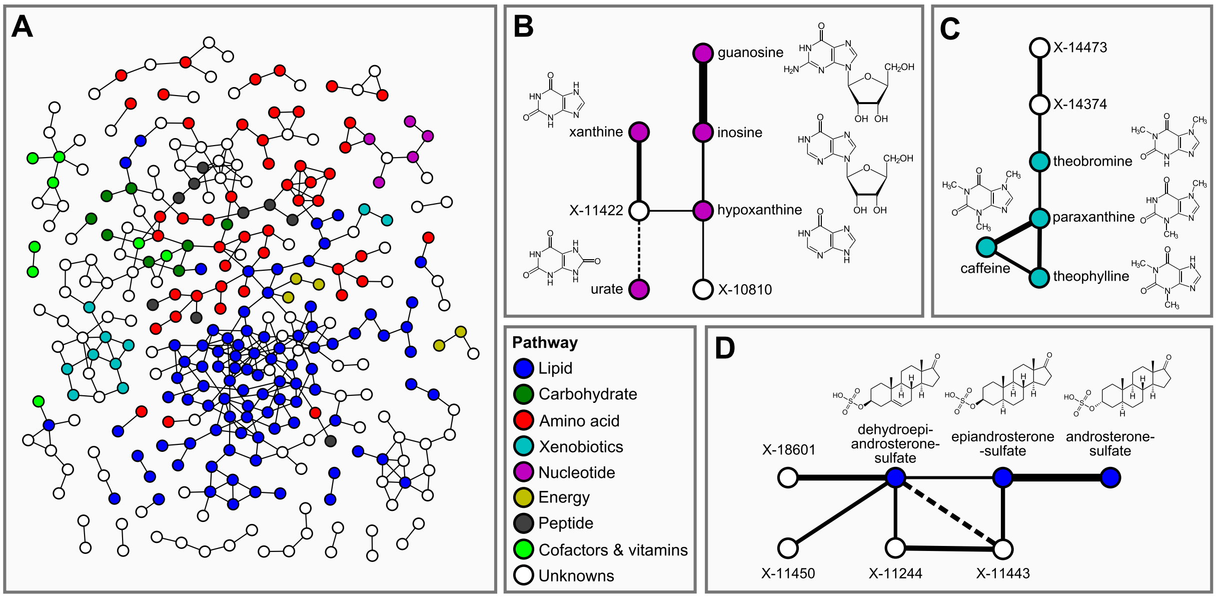


Figure 6.3: Gaussian graphical modeling. GGMs embed unknown metabolites into their biochemical context. **A:** Complete network presentation of partial correlations that are significantly different from zero at α=0.05 after Bonferroni correction. The unknown metabolites are spread over the entire network and are involved in various metabolic pathways. **B-D**: Selected high-scoring sub-networks. We observe that GGM edges directly correspond to chemical reactions which alter specific chemical groups (e.g. carbonyl groups and methyl groups). Solid lines denote positive partial correlations, dashed lines indicate negative partial correlations. Line widths represent partial correlation strengths.

Metabolite 1	**Metabolite 2**	ζ	**Interpretation**
X-11847	X-11849	0.901	biochemical link between two unknowns
3-indoxyl sulfate	X-12405	0.84	*tryptophan metabolism*
X-11452	X-12231	0.832	biochemical link between two unknowns
X-12094	X-12095	0.822	biochemical link between two unknowns
guanosine	inosine	0.798	nucleosides
X-11441	X-11442	0.76	biochemical link between two unknowns
androsterone sulfate	epiandrosterone sulfate	0.755	steroid sulfates
X-11537	X-11540	0.753	biochemical link between two unknowns
X-02269	X-11469	0.734	biochemical link between two unknowns
X-11204	X-11327	0.706	biochemical link between two unknowns
decanoylcarnitine	octanoylcarnitine	0.689	β-oxidation signatures
linoleamide (18:2n6)	oleamide	0.654	C18:1/C18:2 acylamides
3-methyl-2-oxovalerate	4-methyl-2-oxopentanoate	0.646	branched-chain amino acid degradation
catecholsulfate	X-12217	0.601	*catechol metabolism*
X-14189	X-14304	0.593	biochemical link between two unknowns
1,5-anhydroglucitol (1,5-AG)	X-12696	0.58	*sugar metabolism*
dehydroisoandrosterone sulfate	X-18601	0.575	*steroid hormones*
PE(20:4(5Z,8Z,11Z,14Z)/0:0)	X-12644	0.57	*phospholipids (PE)*
X-14208	X-14478	0.558	biochemical link between two unknowns
caffeine	paraxanthine	0.554	caffeine metabolism
X-11423	X-12749	0.549	biochemical link between two unknowns
PC(18:2(9Z,12Z)/0:0)	PC(0:0/16:0)	0.544	phospholipids (PC)
piperine	X-01911	0.526	*amino acid-derived alkaloids*
2-hydroxypalmitate	2-hydroxystearate	0.523	hydroxy fatty acids
X-14056	X-14057	0.519	biochemical link between two unknowns
3-methyl-2-oxovalerate	isoleucine	0.514	isoleucine degradation
X-11244	X-11443	0.51	biochemical link between two unknowns
urea	X-09706	0.506	*urea metabolism*
isoleucine	leucine	0.506	branched-chain amino acids
PE(20:4(n-6)/0:0)	PE(18:2(9Z,12Z)/0:0)	0.502	phospholipids (PE)

Table 6.2: Interpretation of top-ranking partial correlation coefficients (PCC$\geq$0.5). Connections between two known metabolites indicate a direct metabolic relationship, e.g. between purines (guanosine/inosine) or steroid hormones (androsterone sulfate/epiandrosterone sulfate). A link between a known and an unknown compound therefore provides evidence for a shared metabolic pathway. For instance, the link between 3-indoxylsulfate and X-12405 suggests a role of this unknown in tryptophan metabolism. Abbreviations: PC=phosphatidylcholine, PE=phosphatidylethanolamine, ζ=partial correlation coefficient. Italic text represents hypothetical known-unknown connections.

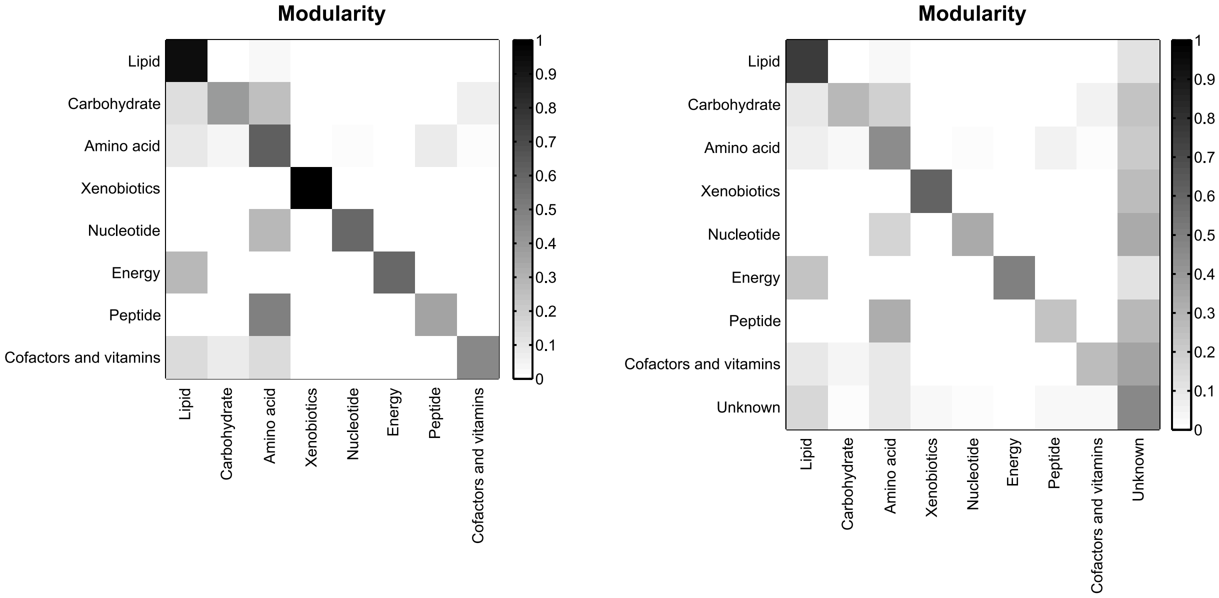


Figure 6.4: Class-wise modularity of the GGM without (left) and with unknowns (right). Colors encode the relative out-degree from each class (rows) to all other classes (columns). Again, we observe strong links within each class and rather weak partial correlations between classes. Some overlaps between related classes (e.g. amino acids and peptides) can be observed. Moreover, the rightmost column in the right-hand figure demonstrates that unknowns are tightly integrated in the GGM.

Chapter 5). The class-wise modularity again shows strong edges within each class, and some overlaps between related groups like “Amino acid” and “Peptide” (Figure 6.4, left).

Inspecting the GGM in detail, we observe that the unknowns are tightly integrated within the network and connected to known compounds of various metabolic classes. This is reflected both in the overall network (Figure 6.3A; Figure 6.4, right) and in the top list of high-scoring GGM edges (Table 6.2), where 18 of the 30 strongest partial correlations comprise at least one unknown metabolite. The second-highest partial correlation in the dataset actually involves a known-unknown metabolite pair, namely 3-indoxylsulfate and the unknown metabolite X-12405 ($\zeta = 0.840$). For pairs of known metabolites, we consistently observe associations of biochemically related metabolites from various metabolic pathways, such as the metabolites inosine and guanosine ($\zeta = 0.798$), which are involved in nucleotide metabolism, or androsterone sulfate and epiandrosterone sulfate ($\zeta = 0.755$), which represent related steroid hormone metabolites. Other pathways with related metabolite pairs include amino acid metabolism, lipid metabolism, bile acid metabolism, and xanthine metabolism. Following our line of reasoning, correlating pairs of a known and an unknown metabolite then directly point to specific pathways on which the

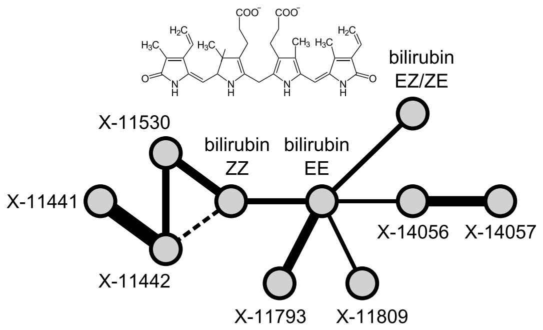


Figure 6.5: High-scoring GGM subnetwork around three bilirubin stereoisoforms containing 7 unknown metabolites.

unknown metabolite may lie. The investigation of the network structure around the unknown compounds provides an additional biochemical context.

We selected four high-scoring sub-networks in the GGM to show that this concept is indeed applicable to real data. The first two of these sub-networks consist of a series of intermediate compounds from purine metabolism, including guanosine, inosine, xanthine derivatives and urate (Figure 6.3B+C). In these cases, one can actually follow the addition and removal of chemical groups by following the edges in the GGM network: Most edges in these sub-networks correspond to the change of either a single methyl group at the purine double-ring structure or to the removal of a ribose residue in the reaction from nucleosides to xanthine variants. While the compounds in both sub-networks appear structurally similar, the distinction into two groups by the GGM is indeed biochemically sound. The metabolites in Figure 6.3B correspond to endogenous substances in the nucleoside pathway, whereas the molecules in Figure 6.3C relate to signals from xenobiotic metabolism of drugs and caffeine. Here, the unknown metabolites X-11422 and X-10810, as well as X-14473 and X-14374 are prominently placed in the networks, making them direct targets for closer inspection with respect to endogenous xanthines and xenobiotics, respectively.

The third sub-network comprises three androsterone sulfate variants, which belong to the class of steroid hormones (Figure 6.3D). We observe direct GGM links between the unknowns X-11450, X-11244 and X-11443 with both dehydroepiandrosterone sulfate (DHEAS) and epiandrosterone sulfate, suggesting androsterone derivatives as likely candidates for these three metabolites (note that the systematic search in the next section will provide further evidence for the steroid hypothesis). The fourth sub-network involves different stereoisomers of bilirubin, which is the degradation product of the oxygen transporter hemoglobin [97] (Figure 6.5). In this sub-network, we observe high partial correlations between the bilirubin variants and a

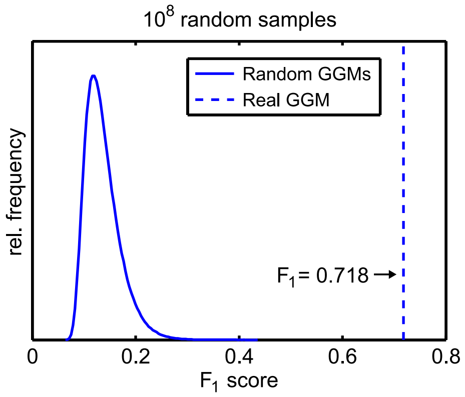


Figure 6.6: Positive control: Classification of known compounds by majority voting amongst their respective GGM neighbors. No random sample achieved an F_1 score equal to or greater than the real GGM, yielding an empirical p-value below 10^{-8}.

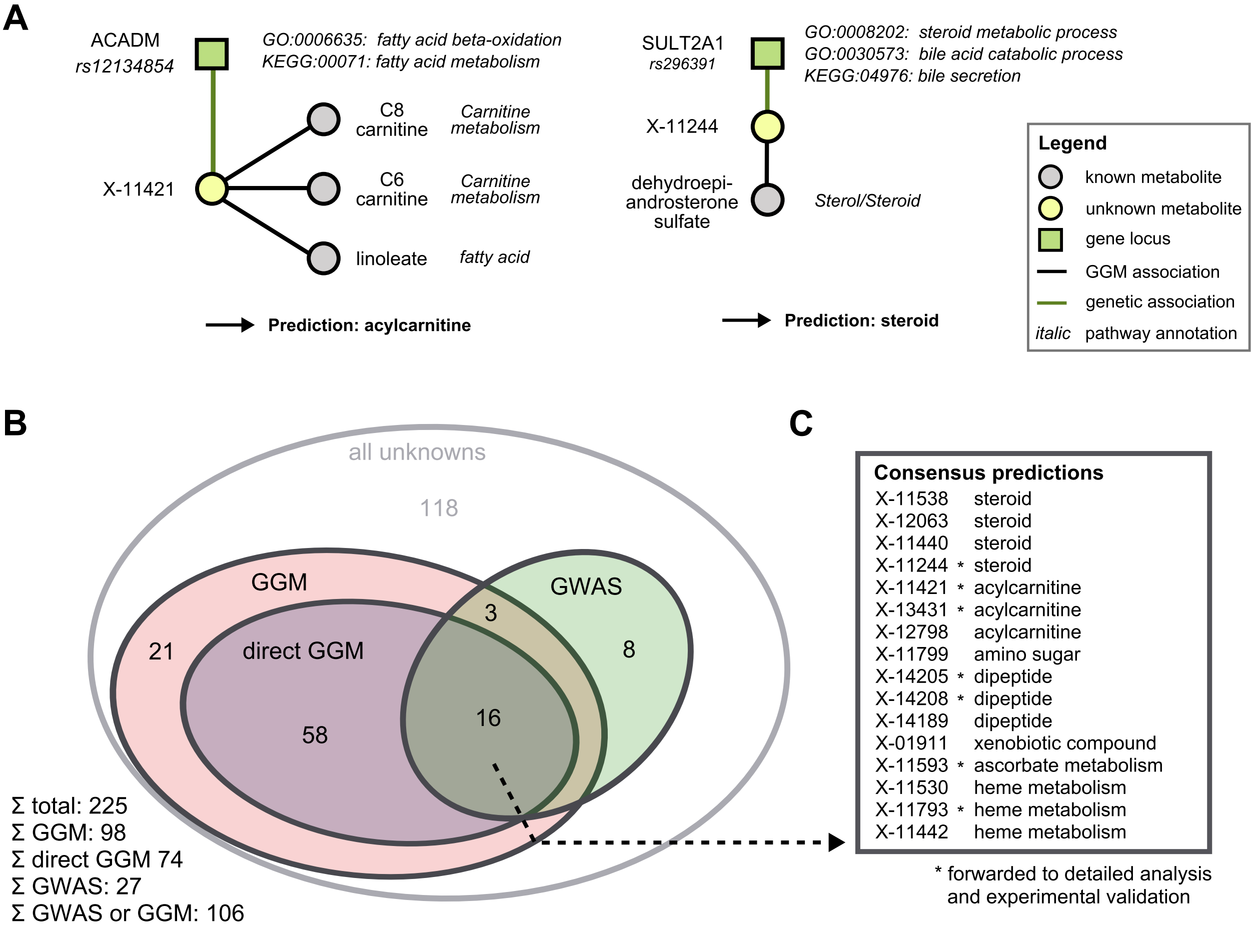
series of unknown metabolites (X-11441, X-11530, X-11442, X-11793, X-11809, X-14056, and X-14057). The seven unknown compounds in this GGM sub-network are thus likely to be involved in hemoglobin degradation processes.

Taken together, the examples confirm that concrete information on the biochemical identity of unknown metabolites can be derived from the present experimental dataset by using the GGM approach.

6.4 Combining GGMs and GWAS allows deriving specific pathway annotations for unknown metabolites

The next step in our analysis was the integration of the GGM and GWAS approaches with general pathway information from external databases, in order to generate concrete predictions for the unknowns' metabolic pathway memberships. As a feasibility test, we first asked whether the local neighborhood of a *known* metabolite in the GGM can be used to correctly predict its metabolic class. Each metabolite is annotated with one out of eight super-pathway annotations: Carbohydrate, Lipid, Nucleotide, Amino acid, Xenobiotics, Energy, Peptide and Cofactors and vitamins. A majority voting approach was implemented, where each known metabolite is assigned to the pathway that occurs most frequently amongst its GGM neighbors. We then determine whether this indeed corresponds to the true pathway of the metabolite. This approach yields a classifier quality of F_1=0.718. Recall that F_1 can be regarded as a quantitative tradeoff between sensitivity and specificity of a classifier (see Chapter 4.3). In order to objectively evaluate classification performance, we generated 10^8 randomly rewired GGM networks and recalculated the majority predictions. No random sample achieved an F_1 score equal to or greater than the real GGM, revealing classification abilities far beyond random ($p < 10^{-8}$, Figure 6.6). It is to be noted at this point, that the actual quality of our classifier might be even higher, since GGM connections between different classes should not always be considered false positive. As an example, metabolites assigned to different inherently related classes such as "amino acid" and "peptide" might actually belong to the same pathway. A classical hypergeometric enrichment analysis [141] among the neighbor classes of a node in the network is not appropriate, since the inherent sparseness of a GGM is not compatible with the null model behind an enrichment approach. While obviously majority voting is amongst the simplest possible classifiers, it is easy to implement and performs well for the task at hand.

We then combined functional annotations for both GGM neighbors and GWAS hits for each unknown in order to derive specific pathway classifications. For unknowns that did not have a known metabolite neighbor in the GGM, we also investigated the 2- and 3-neighborhoods. Since these hits certainly represent weaker evidence than a direct GGM neighbor, we distinguish between 'GGM hit' and 'direct GGM hit' in the following. Functional annotations were obtained from three sources: (1) The sub-pathway assignment provided for each known metabolite in the GGM neighborhood, (2) the GO functional terms for the associated gene of all genome-wide significant GWAS hits, and (3) the KEGG pathways on which the associated genes lie. To the best of our knowledge, there is presently no consistent mapping between annotations from different functional classification schemes available, so we here had to perform the only non-automatic step in the analysis: By manual interpretation of different functional classes (Figure 6.7A), we derive a single consensus pathway annotation for a total of 106 of the unknown metabolites (Figure 6.7B). For 98 unknowns, we obtained annotations from the GGM network, with 74 of these hits representing direct GGM hits. From the 28 genetics hits introduced in the section before, 27 were in a known genetic region with functional annotation. Overlaying the direct edge GGM set and the GWAS set, we obtained 16 unknowns with both biochemical and genetic evidence (Figure 6.7C). From this set of high confidence predictions, we then selected several unknowns which were forwarded to detailed analysis and experimental validation.



Consensus predictions

X-11538	steroid
X-12063	steroid
X-11440	steroid
X-11244 *	steroid
X-11421 *	acylcarnitine
X-13431 *	acylcarnitine
X-12798	acylcarnitine
X-11799	amino sugar
X-14205 *	dipeptide
X-14208 *	dipeptide
X-14189	dipeptide
X-01911	xenobiotic compound
X-11593 *	ascorbate metabolism
X-11530	heme metabolism
X-11793 *	heme metabolism
X-11442	heme metabolism

* forwarded to detailed analysis and experimental validation

Figure 6.7: Semi-automatic prediction of unknown metabolite identities. **A:** Examples of how to determine pathway classifications based on the functional annotations of GGM and GWAS hits. We present two metabolites, X-11421 and X-11244, whose GGM and GWAS associations clearly point into carnitine and steroid metabolism, respectively. **B:** Overview of unknowns functionally annotated by both GGMs and the GWAS approach. 'GGM' refers to an unknown metabolite which is three or less steps away from the unknown in the GGM, whereas 'direct GGM' represents direct neighbors in the network. **C:** Pathway predictions for the 16 unknowns with both direct GGM and GWAs annotations. Unknowns marked with a star were subjected to in-depth analysis followed by experimental validation in the following.

Scenario name	Unknowns	Evidence used	Prediction	Validated as
DIPEPTIDE	X-14208	GGM, genetics	Phe-Ser or Ser-Phe	Phe-Ser
	X-14205		Glu-Tyr or Try-Glu	α-Glu-Tyr
	X-14778		Phe-Phe	Phe-Phe
STEROID	X-11244	GGM, genetics	sulfated androsterone	androstene disulfate
HETE	X-12441	GGM, pathway	hydroxy-arachidonate (HETE)	12-HETE
CARNITINE	X-11421	GGM, genetics, pathway	carnitine species, with 6 to 10 carbon atoms	cis-4-decenoyl-carnitine
	X-13431			nonanoyl carnitine *
BILIRUBIN	X-11793	GGM, genetics	oxidized bilirubin variant	oxidized bilirubin variant *
ASCORBATE	X-11593	GGM, genetics, pathway	O-methylascorbate	O-methylascorbate *

Table 6.3: Six specific scenarios and their experimental validations. Predictions marked by * are supported by exact mass, fragmentation pattern and chromatographic retention time; however, validation using a pure standard compound as a reference is pending since these compounds are presently commercially unavailable in pure form.

6.5 Experimental validation of nine predictions in six distinct scenarios

In total, we investigated six metabolic scenarios in-depth and attempted experimental confirmation of the respective predictions (Table 6.3). In the analysis of these scenarios we used all available evidence, the metabolite correlations, genetic associations, biochemical data, and in addition the molecular masses reported with the known and unknown compounds.

Scenario 1: Our first scenario, DIPEPTIDE, represents the prediction and successful validation of three unknown metabolites involved in short-peptide metabolism (Figure 6.8, left). In the GGM, we observe X-14205, X-14208 and X-14478 in close proximity to various dipeptides, to glutathione derivatives, and to two longer fibrinogen-related peptides. The primary pieces of genetic evidence for this case are the GGT1 locus, which shows a strong association to S-gluthathionyl-L-cysteine, and the ACE locus, which connects to aspartyl-phenylalanine, X-14205, and X-14208. GGT1 encodes for the protein γ-glutamyl transpeptidase, which transfers glutamyl-residues from glutathione in order to generate short-chain peptides [142]. This fits well into the network picture, since GGT1 is connected to the glutathione derivative, which in turn shares a GGM edge with γ-glutamyl-glutamine. ACE, on the other hand, encodes the angiotensin I converting enzyme, a peptidase that cleaves dipeptide fragments from angiotensin precursors and other functional oligopeptides. Since the biochemical and genetic evidence pointed us to short peptides, and dipeptides in particular, we enumerated all possible 400 (=20x20) combinatorial variants of dipeptides and checked the mass against the masses of the three unknowns under investigation. As an example, we shortened the list of candidates for X-14208 from 2,732 (ChemSpider search) to only 8 molecules, respectively. For

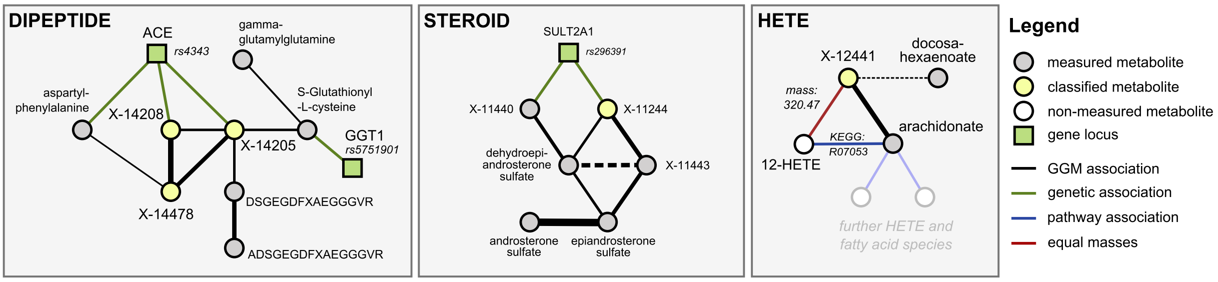


Figure 6.8: Detailed investigation of three scenarios (DIPEPTIDE, STEROID, and HETE). In order to generate concrete hypotheses on the unknowns' identities, we assembled all available information for each scenario. This includes biochemical edges from the GGM, genetic associations from the GWAS, pathway annotations as well as mass information. For details of the predicted identities, see Table 6.3 and main text.

experimental validation, we first checked the plausibility of the candidates with respect to the fragmentation spectra and determined the exact masses. The accurate mass determined for X-14208 is 252.11172 ± 0.001 Da, supporting the chemical formula $C_{12}H_{16}N_2O_4$. While the formula still matches more than 1,200 molecular structures, the prediction of this unknown as a dipeptide leaves only two candidate molecules, namely phenylalanylserine (Phe-Ser) and serylphenylalanine (Ser-Phe). Both variants were obtained from a commercial source and run on the LC-MS/MS platform. The retention index [143] and the fragmentation spectrum received for Phe-Ser matched the index and spectrum of X-14208, whereas Ser-Phe produced a clearly different spectrum (Figure 6.9). Thus, the identity of X-14208 was experimentally confirmed as the dipeptide phenylalanylserine. Importantly, using our integrated approach, we were able to identify X-14208 by only testing two candidate molecules. The other two unknowns, X-14205 and X-14478, were identified through similar experiments as α-glutamyltyrosine (α-Glu-Tyr) and phenylalanylphenylalanine (Phe-Phe), respectively.

Scenario 2: In the second scenario, STEROID, we investigated an unknown metabolite (X-11244) for which both GGM and GWAS data strongly indicate an identity related to steroid-hormone compounds: X-11244 is tightly linked via GGM edges to dehydroepiandrosterone sulfate and two other unknowns, which in turn connect to epiandrosterone sulfate and androsterone sulfate (Figure 6.8, middle). Furthermore, X-11244 displays a highly significant genetic association ($p = 2.1 \cdot 10^{-26}$) with rs296391, which lies in strong LD in the SULT2A1 gene locus. SULT2A1 encodes for a member of the sulfotransferase family 2A, dehydroepiandrosterone-preferring, further strengthening the metabolic context. Based on the GGM and GWAS results, we hypothesized that X-11244 is a steroid sulfate related to androstane. Experimentally, the primary loss of a fragment with a nominal mass of 98 and the presence of an ion at 97 m/z ob-

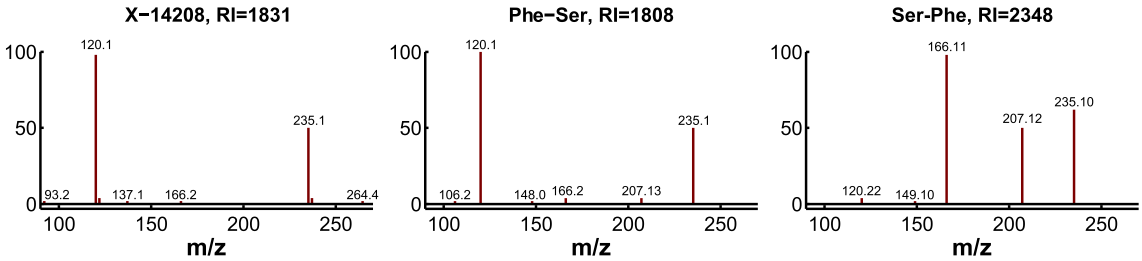


Figure 6.9: Experimental confirmation of X-14208 as phenylalanylserine. Two possible dipeptide variants were predicted and consequently tested. The fragmentation spectrum of pure Phe-Ser matches that of the unknown compound, whereas the spectrum for pure Ser-Phe differs visibly. Moreover, the retention index (RI) of Phe-Ser is similar to the RI of X-14208, whereas that of Ser-Phe is significantly different.

servable in the fragmentation spectrum of X-11244 indicate the presence of at least one sulfate group in this unknown. The exact mass determined for X-11244 supports the chemical formula $C_{19}H_{30}O_8S_2$. Querying ChemSpider for this chemical formula yields only four results, one of which corresponds to an androstene disulfate variant (ChemSpider ID 21403154). Analysis of several disulfated androstenes demonstrated similar retention times and fragmentation spectra. Among the tested variants, 4-androsten-3β,17β-disulfate showed the best match. Given that other isomers are also possible, which cannot necessarily be chromatographically resolved, we annotated X-11244 more generically as ‘androstene disulfate’.

Scenario 3: In the third scenario, HETE, we made explicit use of known biochemical interactions derived from three publically available pathway databases. We searched for cases where an unknown shows GGM connections to known compounds for which a direct pathway interaction with a metabolite having the same mass as the unknown exists. The unknown metabolite X-12441 does not show any genome-wide significant SNP hits and only a single GGM neighbor: cis-5,8,11,14-eicosatetraenoic acid (arachidonate, Figure 6.8, right). Arachidonate constitutes pathway connections to several other lipid-related metabolites, including a variety of hydroxy-arachidonate variants (HETEs). These variants have the chemical formula $C_{20}H_{32}O_3$ with a molecular weight of 320.2351 Da, matching the mass of the unknown. We thus hypothesized that X-12441 represents a HETE species. Experimentally, the determination of the exact mass of the unknown further supported our hypothesis, as the accurate mass determined for X-12441 matches the chemical composition of HETE to a precision of 0.002 Da. A number of HETE isoforms were experimentally tested, including the 5, 8, 9, 11, 12 and 15 isoforms. All isoforms produced unique fragmentation spectra that permitted the precise identification of the unknown X-12441 as the 12-HETE isoform.

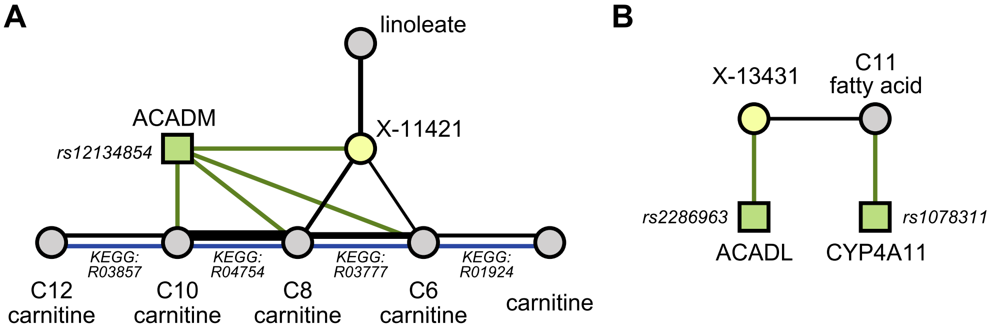


Figure 6.10: Scenario CARNITINE. **A:** X-11421 associates with ACADM, which catalyzed a rate-limiting step in β-oxidation and has previously been shown to link to plasma acyl-carnitines. Furthermore, the unknown is connected to C6 and C8 carnitines in the GGM. X-11421 has been verified as cis-4-decenoyl-carnitine. **B:** X-13431 is associated with ACADM (another variant of the beta-oxidation enzymes) and C11 fatty acid. In accordance with previous associations between ACADL and C9 carnitines, X-13431 has been confirmed as nonanoyl (C9) carnitine.

Scenario 4: In the CARNITINE scenario, we investigated two specific unknowns that, on the one hand, display associations with fatty acid derivatives (in particular with acylcarnitines) and, on the other hand, associate with enzymes of the acyl-coenzyme A dehydrogenase (ACAD) class. Acylcarnitines represent a transport form of fatty acids tagged for mitochondrial transport and subsequent β-oxidation [144]. In Chapter 5 we already demonstrated strong GGM edges between carnitine species with a carbon atom difference of two. Furthermore, previous metabolomics GWAS revealed genetic associations between various acylcarnitines and loci encoding for β-oxidation-related ACAD enzymes (e.g. Illig et al. [27]).

The first unknown metabolite, X-11421, shares significant GGM edges with C8 and C6 carnitines and further associates with ACADM, the ACAD enzyme for medium-chain length fatty acyl residues (Figure 6.10A). In the context of our previous findings and considering the mass peak of X-11421 (314.2 m/z, pos. mode), we therefore hypothesized that X-11421 is a medium-chain length carnitine with 10 carbon atoms. Matching our computational prediction, this unknown has indeed been experimentally identified as cis-4-decenoyl-carnitine, a carnitine with 10 carbon atoms and an ω-6 double bond, by testing the pure compound. It has to be noted that carnitines shift elution times dramatically in relation to their RI markers on the analytical platform used in this work. The cis-4-form was confirmed in a spiking experiment in a well characterized human plasma sample, which was run with original samples.

The second unknown metabolite, X-13431, is linked to a C11 free fatty acid in the GGM and associates with the ACADL locus (Figure 6.10B). In a previous study, this locus has been shown to associate with C9 carnitine levels [27]. This observation together with the molecule mass peak

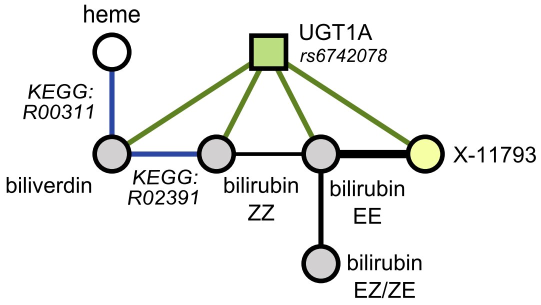


Figure 6.11: Scenario BILIRUBIN. The unknown X-11793, several bilirubin isomers and biliverdin are tightly connected via both the GGM and the UGT1A locus. Experimentally, there is strong evidence that X-11793 represents an epoxidized bilirubin variant.

detected for the unknown (302.3 m/z, positive mode) makes C9 carnitine a good candidate for X-13431. Our prediction is experimentally confirmed by the fragmentation of X-13431 as the molecule produces fragments shared by mid- and long-chain acylcarnitines and several neutral losses (loss of 59 m/z and 161 m/z) that are highly diagnostic of carnitines. With respect to chromatography, X-13431 elutes between C8 and C10 carnitines, thus further supporting the hypothesized C9 carnitine. The accurate mass of 301.22476±0.0015 Da determined for X-13431 corresponds to the molecular formula $C_{16}H_{31}NO_4$, which also matches C9 carnitine. Due to the lack of a commercial source for pure C9 carnitine, the final confirmation of the predicted chemical identity by testing the pure compound is still pending.

Scenario 5: In the BILIRUBIN scenario, we focused on the unknown metabolite X-11793, which shares a GGM edge with a specific bilirubin stereoisomer (EE) and associates with the UGT1A locus encoding for the enzyme UDP glucuronosyltransferase 1 family, polypeptide A. The bilirubin stereoisomers, which show close proximity in the GGM, are degradation products of heme, the oxygen-carrying prosthetic group contained in hemoglobin [145]. For further metabolization and excretion, the very insoluble bilirubin must be transformed into soluble derivatives. Glucuronidation of bilirubin represents the main mechanism for this transformation in the human metabolism. The reaction is mainly catalyzed by an enzyme encoded at the UGT1A1 locus [146, 147], matching the observations in our data that X-11793 and three of the four degradation products display genetic associations with the UGT1A locus.

Since X-11793 is embedded in the biochemical and genetic network of bilirubin derivatives and also shares their association with the UGT1A locus, we assumed that X-11793 represents a further bilirubin derivative. Moreover, the mass difference between bilirubin and X-11793 is 15.9, which might correspond to the addition of oxygen. We therefore predicted X-11793 to be an (ep)oxidized bilirubin as a possible result of bilirubin oxidation mediated by cytochrome

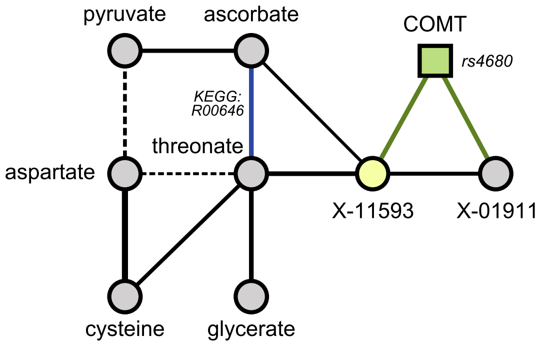


Figure 6.12: Scenario ASCORBATE. X-11593 associates with the COMT locus on the one hand, and is connected to the ASCORBATE pathway on the other hand. We thus hypothesized that the unknown represents an O-methylated ascorbate, which is in accordance with the observed mass spectrum and chromatography retention time.

P450. Such oxidation processes have previously been suggested as alternative routes for the metabolization of bilirubin besides glucuronidation [148].

From an experimental perspective, the neutral accurate mass of X-11793 of 600.25859 Da, corresponding to $C_{33}H_{36}N_4O_7$, perfectly matches the formula for the predicted (ep)oxidized bilirubin variant. The fragmentation pattern produced by the unknown molecule further supports the hypothesis: Bilirubin generates fragments with 285 m/z and 299 m/z corresponding to a cleavage of the central C-C bond of the molecule. If the hypothesized ep(oxidized) bilirubin broke at the same position, it would produce fragments with 299 m/z and 301 m/z accordingly, which both occur in the fragmentation spectrum of X-11793. The final confirmation of the prediction by running pure epoxidized bilirubin is still pending due to the lack of commercial sources of the pure substances.

X-11793 identified as (ep)oxidized bilirubin might represent an interesting additional biomarker for the efficacy of heme degradation processes, which plays an important role in various diseases. Serum concentrations of bilirubin as well as the UGT1A locus are not only associated with bilirubin turnover-related syndromes such as jaundice, but also with different cancer variants and coronary heart disease (CHD) [146, 149–151]. While jaundice is caused by high bilirubin concentrations, bilirubin has proven to be an effective antioxidant [152], which might explain the association found between reduced risk of CHD and various forms of cancer with higher bilirubin concentrations.

Scenario 6: In the ASCORBATE scenario, we investigated the unknown X-11593, which is close to threonate, ascorbate and related substances in the GGM. These metabolites are tightly interconnected in the ascorbate (vitamin C) pathway. Furthermore, we found significant associations of X-11593 with SNPs in the gene encoding catechol-O-methyltransferase (COMT), an enzyme relevant for the inactivation and degradation of many drugs. COMT O-methylates molecules with catechol-like structures.

Since, according to the GWAS, X-11593 is probably a substrate or a product of O-methylation, we determined the mass differences to the known metabolites neighboring X-11593, namely ascorbate and threonate. While the mass difference of X-11593 and threonate is 54, X-11593 and ascorbate show a mass difference of 14, which corresponds to the addition of a methyl moiety. Moreover, in ascorbate, the double bond within the 5-ring with its two hydroxyl moieties could 'mimic' the corresponding planar substructure in catechol, on which COMT is usually working. Finally, the methylation of ascorbate through the catalysis of COMT has already been shown experimentally [153]. These observations make O-methylated ascorbate derivatives (most probably 2-O-methylascorbate) good candidates for X-11593. From an experimental perspective, our hypothesis is supported by the accurate neutral mass of 190.04787 Da determined for X-11593. Based on the accurate mass, the molecular formula for X-11593 is $C_7H_{10}O_6$, matching our prediction. The retention time of X-11593 shows a slight shift compared to the time for ascorbate. This shift matches the shift expected for adding a methyl group. Moreover, X-11593's primary fragment loss is 60, which is the same as for ascorbate. The loss of 15, also seen for X-11593, is typical for phenols substituted with a -OH and -OCH_3. Due to the lack of a commercial source for 2-O-methylascorbate the confirmation through the spectrum of the pure substance is still pending

6.6 Discussion of novel genetic associations

We developed and validated a novel integrative approach for the biochemical characterization of 'unknown metabolites' from high-throughput metabolomics and genotyping datasets. Our method allows for the functional annotation of previously unknown metabolites and, as a consequence, enhances the interpretability of metabolomics data in genome-wide association studies and biomarker discovery. For the first time, we systematically evaluated genetic associations of unknown metabolites, thereby discovering seven new loci of metabolic individuality. By classifying a series of unknown metabolites, we gained new insights into the functional interplay between genetic variation and the metabolome both for previously reported and new loci. Furthermore, several of the unknown compounds that we identified as well as their newly associated loci were independently reported in disease-related studies. In the following, we discuss three recently published studies.

COMT & hepatic detoxification

The first example is a recent biomarker study, where Milburn et al. [154] reported an association of X-11593 with hepatic detoxification. In our GWAS, we find a strong association of X-11593 with the COMT locus, which encodes the catechol-O-methyltransferase enzyme. COMT is responsible for the inactivation of catecholamines such as L-dopa and various neuroactive drugs by O-methylation [155]. Following our identification approach, we experimentally confirmed the identity of X-11593 as O-methylascorbate. Notably, O-methylascorbate is a known product of ascorbate (vitamin C) O-methylation by COMT [153, 156]. Thus, our observations establish a link between O-methylascorbate blood levels, common genetic variation in the COMT locus and COMT-mediated liver detoxification processes.

ACE & hypertension

The second study relates to the ACE gene locus, which is a known risk locus for cardiovascular disease, hypertension and kidney failure. The protein encoded by the ACE locus, angiotensin-converting enzyme, is an exopeptidase which cleaves dipeptides from vasoactive oligopeptides, and plays a central role in the blood pressure-controlling renin-angiotensin system [157]. Moreover, the ACE protein is a target for various pharmaceuticals, especially in the treatment of hypertension [158]. Steffens et al. [125] recently published a study on metabolic differences between depressed and non-depressed individuals suffering from heart failure. They reported two differentially regulated unknown metabolites measured using the same metabolomics platform as the one used here. These two potential biomarkers, X-11805 and X-03094, were also analyzed in the present study. In our GGM network, X-11805 is in close proximity to angiotensin-related peptides. It may thus be involved in blood pressure control processes. The second unknown metabolite, X-03094, is directly connected to cholesterol in the GGM and therefore may represent a metabolic intermediate of cholesterol metabolism. Blood cholesterol levels in turn are a major risk factor for coronary heart disease [159]. Thus, our predictions suggest a potential biological link for both unknowns to the associations reported by Steffens et al.

UGT1A/ACADM & insulin resistance

The third example is an explorative study to detect biomarkers for insulin sensitivity. Gall et al. [22] reported several known metabolites (most prominently α-hydroxybutyrate) as biomarkers for insulin resistance. They also reported a series of unknown metabolites among their top hits. Here, we investigated three of these unknowns: X-11793 associates with UGT1A (UDP glucuronosyltransferase 1) and most likely represents a bilirubin-related substance. Moreover, we experimentally validated X-11421 and X-13431, which display a strong association with ACADM (*acyl-Coenzyme A dehydrogenase, C-4 to C-12 straight chain*), as acylcarnitines containing 10 and 9 carbon atoms, respectively. The identification of these latter two unknown metabolites as medium-chain length acylcarnitines is coherent with reports by Adams et al. [160]. The authors found elevated blood plasma acylcarnitine levels in women with type 2 diabetes. Functionally, they attributed this finding to incomplete β-oxidation. Thus, our identification of X-11421 and X-13431 now suggests incomplete β-oxidation as an explanation for the associations found by Gall et al. and implies that acylcarnitines containing 10 and 9 carbon atoms are potential biomarkers for insulin resistance.

6.7 Conclusion

In summary, we integrated high-throughput metabolomics and genotyping data from a large population cohort for elucidating the biochemical identities of unknown metabolites. To this end, we applied metabolomics genome-wide association studies and Gaussian graphical modeling in order to link these unknown metabolites with known metabolic classes and biological processes. For six specific scenarios, we went from systematic hypothesis generation over detailed investigation and identity prediction to direct experimental confirmation. Similar validations may now be undertaken for the remaining predictions that we report in Figure 6.7. Finally, we demonstrated the benefit of our method by discussing several of these newly identified metabolites in the context of existing biomarker discovery studies on liver detoxification, hypertension and insulin resistance.

Our present approach can be extended in several directions. It can be combined with method-specific, automated techniques that further reduce the search space of candidate metabolites. Previously mentioned methods relying on mass-spectra [129] or chromatographic properties [131] are suitable candidates here. Furthermore, the biochemical context provided by the GWAS might be used in more detailed analyses, i.e. by taking into account the specific chemical transformation a given enzyme catalyzes. The method can be directly transferred to other types of metabolomics datasets not specifically originating from MS experiments, such as NMR-based metabolomics.

Beyond the application to metabolite identification, our study demonstrates the general potential of functional metabolomics in the context of genome-wide association studies. The comprehensive metabolic picture provided by GGMs in combination with GWAS allows for the detailed analysis of metabolic functions, chemical classes, enzyme-metabolite relationships and metabolic pathways.

Chapter 7

The potential of metabolomics GGMs: Further applications

In the previous three chapters, we demonstrated the ability of Gaussian graphical models to reconstruct pathway reactions from metabolomics data. Importantly, this data-driven metabolic network reconstruction is not dependent on existing knowledge and able to embed metabolites with weak pathway evidence (like the phospholipids in Chapter 5) and even unknown metabolites (in Chapter 6). We next asked how to further exploit these metabolic networks biologically. A major focus of this chapter will be concepts from *differential network biology* [161], where group-specific differences (e.g. between healthy and diseased) are investigated in a network context. Most of the projects were performed in collaborations with other research groups from the Institute of Epidemiology, the Genome Analysis Center and the Institute of Bioinformatics and Systems Biology at the Helmholtz Zentrum München.

In Section 7.1, we introduce the concept of 'effect networks', where GGMs are combined with results from statistical analyses. We present applications from three collaboration projects: (1) Gender-specific differences in the KORA study participants, (2) associations between the fat-free body mass and the metabolome and (3) the impact of a Type-D personality on the metabolome. A complementary approach of using sample groups is then introduced in Section 7.2. We developed a *differential* GGM approach which directly uses the experimental design in order to elucidate specific perturbations introduced by a chemotherapeutic treatment of the U87 glioblastoma (brain cancer) cell line. In Section 7.3 we introduce a study where GGMs were used to define biologically meaningful metabolite groups for a novel enrichment algorithm called *phenotype set enrichment analysis*. Finally, in Section 7.4, GGM edges are used to

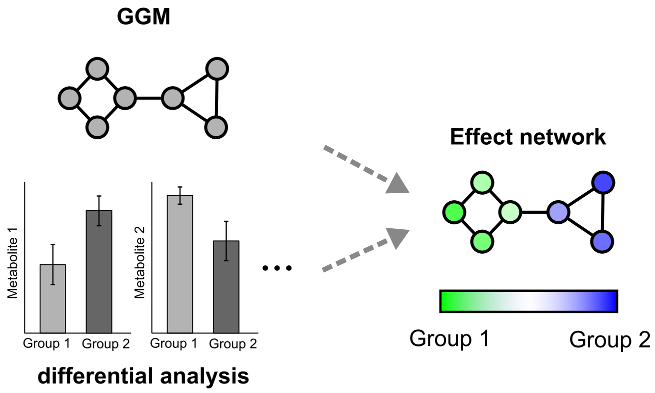


Figure 7.1: The concept of 'effect networks'. We combine GGMs with the results from a differential statistical analysis, e.g. between two groups like gender or healthy/disease. The resulting network then combines both biochemical relationships and the metabolic effect of the group under investigation.

shed light on the role of metabolite ratios and *p-gains*. P-gains are specific measure of statistical association for metabolite pairs introduced in metabolomics genome-wide association studies.

Most of the projects discussed in the following are based on data from the KORA cohort (cf. Chapter 2) either with measurements from the Biocrates platform (Chapter 5) or the Metabolon platform (Chapter 6). Since different research groups tend to apply different quality control and missing values treatment procedures, or had access to only a limited subset of the data, the number of samples and metabolites used for each particular project differ slightly.

7.1 Group-specific metabolome differences: Effect networks

In the following, we present three projects where we combined results from statistical analyses with Gaussian graphical models ('effect networks', Figure 7.1). In all projects, metabolite concentration differences were investigated with respect to a specific grouping or phenotype in the data (male/female, fat-free body mass, Type-D personality). A statistical model usually yields a coefficient which represents the strength of association. For instance, in a linear regression model, the coefficients from the fitted model represent a quantitative measure of the association strength. These values can then be color-coded and visualized in the GGM network. Inspection of the resulting networks allows to evaluate the statistical signal in the context of their biochemical interactions. A major advantage of GGMs over publically available pathway networks is that they can readily be calculated on any data matrix, without the need for high-quality functional annotations. In this thesis, we will only manually investigate the effect networks; future projects will extend this approach by computational methods like graph clustering and statistical enrichment.

Note that our effect network approach belongs to the class of *differential network biology* methods, which have recently been attributed a central role for upcoming high-throughput data analyses [161].

7.1.1 Gender-specific differences of blood metabolites

It is well-acknowledged that gender differences, i.e. being male or being female, are substantial and can be detected throughout various levels of biological organization. For example, considerable sexual dimorphisms have been reported for behavioral traits [162], brain morphometry [163], mental disorders [164], and fat metabolism [165], to name but just a few examples. Obviously, such differences may be a problem for population-based studies, which might be hampered by sex bias [166], but are in parallel an interesting field for fundamental research on its own [167]. Again using data from the KORA study, we sought to determine differences in the blood serum metabolome between males and females. The study was published in

⋆ Mittelstrass, K., Ried, J.S., Yu, Z., **Krumsiek, J.**, Gieger, C., Prehn, C., Roemisch-Margl, W., Polonikov, A., Peters, A., Theis, F.J., Meitinger, T., Kronenberg, F., Weidinger, S., Wichmann, H.E., Suhre, K., Wang-Sattler, R., Adamski, J., and Illig, T. Discovery of Sexual Dimorphisms in Metabolic and Genetic Biomarkers. *PLoS Genetics*, 7(8):e1002215, 2011.

The analysis was based on a total of 3,300 individuals from the KORA F3 and F4 cohorts, and used a total of 131 measured metabolites measured using the AbsoluteIDQ$^{\text{TM}}$ kit. Note that this set differs from the dataset introduced in Chapter 2, having more samples (the study was conducted after our initial GGM analysis) and a slightly different quality control leaving only 131 metabolites.

Linear regressions were carried out with metabolites as dependent variables, gender as the explanatory variable, and age and BMI as covariates for correction. That is, we fitted linear models of the form:

$$m_i = \beta_{0i} \cdot \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} + \beta_{1i} \cdot \text{gender} + \beta_{2i} \cdot \text{age} + \beta_{3i} \cdot \text{BMI} + \epsilon_i \qquad \text{for } i = (1, \ldots, p), \quad (7.1)$$

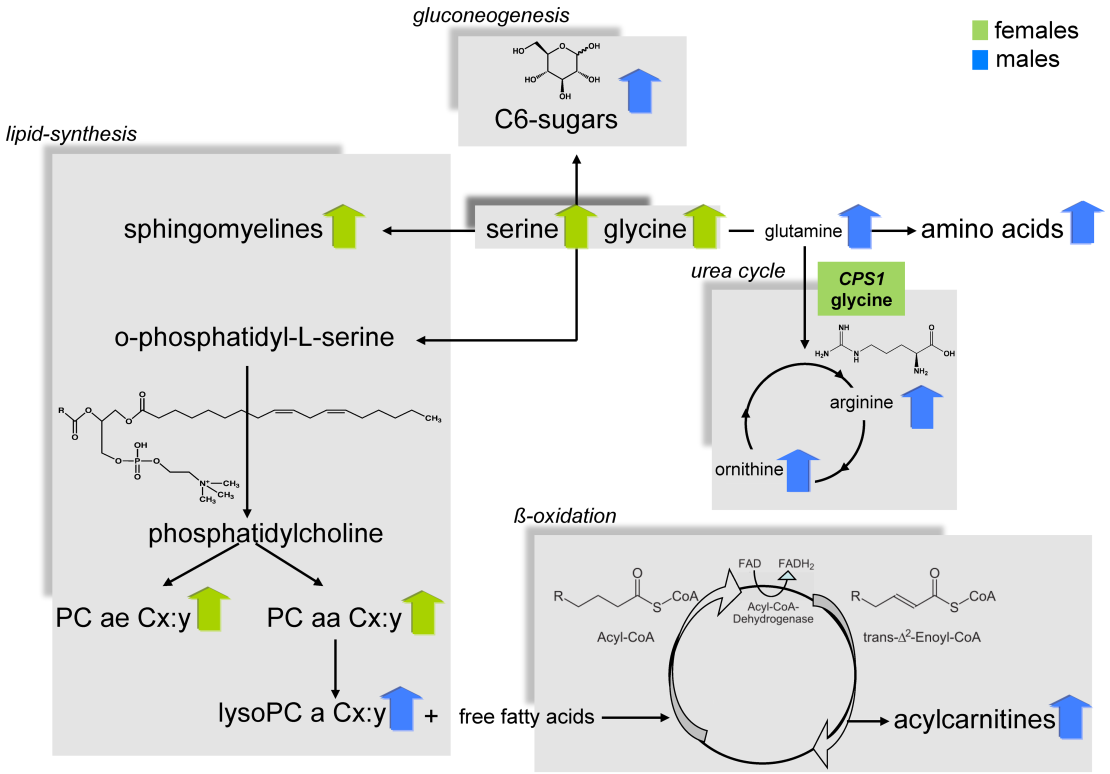


Figure 7.2: Metabolic differences between males and females. Green arrows indicate metabolite classes upregulated in females, whereas blue arrows represent higher concentrations in males. Several sphingolipids, phosphatidylcholines and serine and glycine are higher in females. In contrast, C6 sugars (primarily glucose), lyso-phosphatidylcholines, acylcarnitines (representing β-oxidation intermediates) and other amino acids are higher in males. Abbreviations: PC=phosphatidylcholine. Reprinted from Mittelstrass et al. [25].

where m_i represents the concentration vector of the i-th metabolite, gender, age and BMI are the respective phenotype vectors (one value for each sample), the $\beta_{\cdot i}$ represent the fitted coefficients, ϵ_i is a normally distributed error term and p is the number of metabolites (131 in this case). Gender is coded with discrete values of zero and one. The linear model is then fitted according to equation (3.8) in Chapter 3.

For a detailed list of results we refer the reader to the original publication. Briefly, phosphatidylcholines, sphingomyelins, as well as serine and glycine were generally higher in females compared to males (Figure 7.2). In contrast, lysophosphatidylcholine levels, acylcarnitines, C6-sugars (which primarily represents glucose), and the remaining amino acids were higher in males than in females. In total, 102 out of 131 metabolites were reported to be significantly different after Bonferroni correction between genders. In addition to the metabolomics analysis alone, we reported a strong gender difference for the genetic association between the CPS1 locus (carbamoyl-phosphate synthase 1) and glycine. In this thesis, however, we will focus on the metabolic changes only.

In addition to the rather knowledge-driven result analysis displayed in Figure 7.2, we applied our GGM methodology to get a more comprehensive picture of gender differences. Projecting the β_{1i} values from equation (7.1) onto the GGM allows to actually follow the propagation of gender difference effects through the metabolic network (Figure 7.3). Importantly, for this analysis the GGM was corrected for gender-specific effects (by adding gender to the data matrix), in order to specifically avoid re-using the same data. Since in a Gaussian graphical model each pairwise correlation is corrected for all remaining variables, adding a column to the data matrix will automatically remove confounding effects of this variable on the GGM. We applied a high cut-off of ζ=0.3 to emphasize strong inter-metabolite effects for this analysis. With this cutoff, the GGM contains a total of 1.28% out of all possible edges (Figure 7.4A). The distribution of partial correlation coefficients resembles the distribution shown in Figure 5.1, thus further emphasizing the stability of GGM estimations for the given sample size. Note that in contrast to the GGM calculated in Chapter 5, we here omitted negative edges and focused only on positive partial correlations. As discussed in that chapter, negative correlations generally represent a special case of partial correlations which can be excluded for this analysis.

In order to further investigate topological properties of the GGM, we plotted the number of clustered groups in the GGM as a function of the absolute partial correlation cutoff (Figure 7.4B). For this analysis, we excluded singleton metabolites without any partial correlation above the threshold. Most non-singleton groups emerge in the cutoff range between 0.3 and 0.7, which corresponds to the cutoffs from Figure 7.3. For our lower cutoff of 0.3, we obtain 14 groups, which can here be regarded as independent phenotypes in the metabolite pool.

Strikingly, sex-specific effects appear localized with respect to both the measured metabolic classes and the GGM structure. For instance, while most sphingomyelin concentrations were shown to be higher in females (Figure 7.2), they also represent a connected component in the GGM (Figure 7.3). Similarly, acylcarnitines are higher in males and mostly share partial correlation edges with other acylcarnitines. The analysis suggests that sex-specific concentration differences affect whole metabolic pathways rather than being randomly spread over the different metabolites.

One specific result in this combined analysis is of particular interest. Three metabolite pairs display strong edges in the GGM, but substantially different gender-specific regulation (yellow ellipses in Figure 7.3). Specifically, lysoPC a C20:3, lysoPC a C20:4 and lysoPC a 18:2 concen-

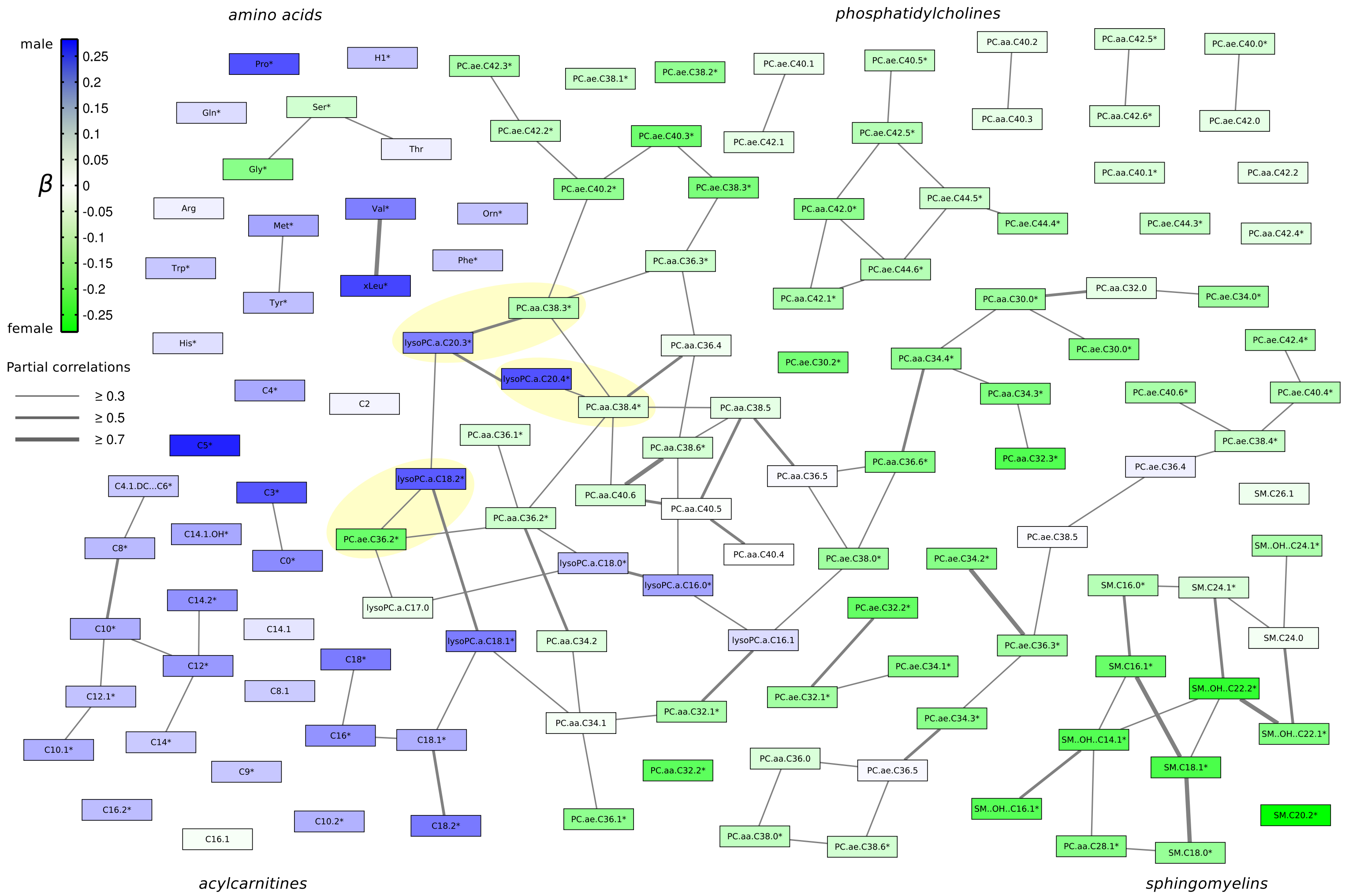


Figure 7.3: Gender effect network. Gaussian graphical model illustrating the propagation of gender-specific effects through the metabolic network. Each node represents one metabolite whereas edge weights correspond to partial correlation strengths. The diagram only shows partial correlations ≥ 0.3. This networks is essentially based on the same dataset as the GGM shown in Figure 5.1, only on a slightly smaller set of metabolites and a higher partial correlation cutoff. Node coloring represents the strength of association – measured using β_{i1} from equation (7.1) – towards either males or females. Asterisks indicate significantly different metabolites between genders. Yellow highlighted metabolite pairs differ by a C18:0 fatty acid residue. Reprinted from Mittelstrass et al. [25].

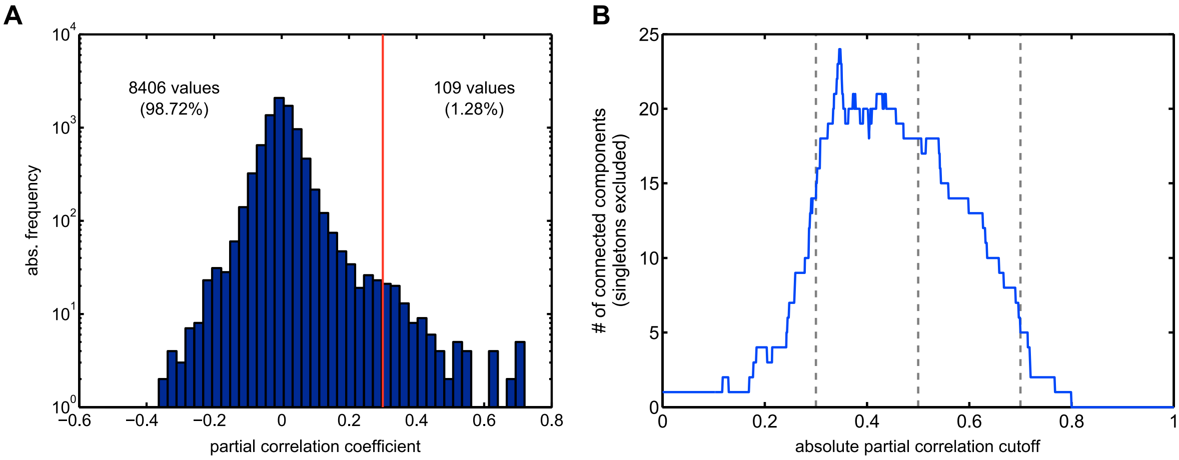


Figure 7.4: Topological properties of the GGM used in the gender study. **A:** Histogram of partial correlation coefficients. For the chosen cutoff ζ=0.3, the network contains 1.28% of all possible edges. **B:** Number of connected components in the GGM as a function of the absolute partial correlation cutoff. For ζ=0.3, we obtain 14 non-singleton groups. The vertical dashed lines at ζ=0.5 and ζ=0.7 correspond to the two additional partial correlation cutoffs annotated in Figure 7.3.

trations are significantly higher in males, whereas their respective GGM neighbors PC aa C38:3, PC aa C38:4 and PC ae C36:2 are higher in females. Interestingly, each pair of lyso PC and diacyl PC shares a carbon atom / double bond difference of exactly C18:0. Thus, the regulation of C18:0 (stearic acid) might represent a key metabolic difference between males and females. From the network point-of-view, C18:0 differences may be regarded as an *entry point* of gender differences into the metabolic network. Differences between males and females primarily arise from these localized positions in the metabolic pathway and further observed differences might rather be caused by propagation of the signal through the network.

In summary, this study presented a possibility to use GGMs for the extended analysis of statistical associations in metabolomics data. Our results assign a key role to stearic acid as a possible entry point of gender-specific metabolic differences. Importantly, the approach is not limited to allegedly biased metabolic pathway databases, since only measured metabolomics data with a phenotypic trait are required. The study is currently being repeated with data from the Metabolon platform, which provides are broader panel of measured metabolites in contrast to the lipid focused Biocrates kit used here. Furthermore, the analysis will be extended by a specifically designed enrichment algorithm which directly points out areas in the metabolic network carrying a localized signal.

7.1.2 Metabolome associations with fat free mass

In this study, we investigated associations between the blood metabolome and features of body composition. In particular, we analyzed the fat-free body mass (FFM) which mainly represents skeletal muscle mass [168]. FFM has been used in various contexts, for instance to investigate associations of early nutritional programming and metabolic disease risk [169], weight loss due to lung diseases [170, 171] or the effects of anabolic steroids on body mass [172]. Since especially changes in lipid metabolism can be expected for varying fractions of fat-free mass, the Biocrates metabolites here allow for an explorative investigation of possible impacts of muscle content on the human metabolome. As a proxy of skeletal muscle mass, we used the fat-free mass index (FFMI), which is a height-independent measure of fat-free mass based on the body fat percentage. To the best of our knowledge, it represented the first systematic study comparing FFMI with high-throughput metabolomics data. The analysis was again performed on data from the KORA F4 study (3061 probands) for discovery, and complemented by samples from KORA S4 for replication. This dataset is identical to the one used in Section 7.1.1. A co-authored manuscript for the project has been prepared:

⋆ Jourdan, C., Petersen, A.K., Gieger, C., Döring, A., Illig, T., Wang-Sattler, R., Meisinger, C., Peters, A., Adamski, J., Prehn, C., Suhre, K., Altmaier, E., Kastenmüller, G., Römisch-Margl, W., Theis, F.J., **Krumsiek, J.**, Wichmann, H.E., and Linseisen, J. Association between Fat Free Mass and Serum Metabolite Profile in a Population-Based Study at Two Points in Time. *PLoS ONE*, 7(6):e40009, 2012.

A linear regression analysis with metabolites as dependent variables, FFMI as the explanatory variable, and age and gender as covariates for correction was performed analogously to equation (7.1) in section 7.1.1. The analysis revealed significant associations with the FFMI for various metabolites from different metabolic classes. For instance, the branched-chain amino acids (BCAAs), tyrosine and phenylalanine were found to be positively associated with FFMI. Furthermore, a positive association of the ratio of branched-chain amino acids to glucogenic amino acids was discovered. This indicates increased BCAAs concentrations in relationship to glucogenic amino acids in subjects with higher FFMI. For the carnitine class, an increase in short odd-chained carnitines (especially C3 and C5) with FFMI as well as a decrease of long chain C18 carnitine was detected. This combination of associated metabolites suggests an increased β-oxidation rate for higher FFMI values.

To further elucidate the biochemical context of the recovered FFMI effects, we again colored the metabolomics GGM with β-values from linear regression analysis (Figure 7.5). Two clusters of particular interest will be discussed in the following. First, ‘cluster 1’ in this figure shows diverse, rather unlocalized effects. Interestingly, however, PC aa 38:3 appears in the center of this cluster and represents the only phospholipid species with a positive association to the FFMI. The GGM neighbors of PC aa 38:3 do not display any significant impacts of the FFMI. Hence, this phospholipid might represent a specific point in the metabolome association with fat-free body mass, whereas surrounding metabolic interaction partners are decoupled in terms of effect propagation. A bulk fatty acid side chain composition of 38:3 most probably either represents a combination of the ω-6 unsaturated 20:3 with 18:0, or of the ω-3 or ω-6 18:3 with 20:0. It might thus be worthwhile to further investigate the fatty acid biosynthesis or degradation pathways specifically involving these fatty acids species. 'Cluster 2' contains a set of coordinately downregulated phosphatidylcholines with very long fatty acid side chains. In line with our findings from the carnitine class, the GGM results further strengthen the hypothesis of an increased fatty acid oxidation in subjects with a higher relative amount of fat-free mass.

Taken together, in this project we repeated the effect network approach with FFMI as the analyzed phenotype. We found indications for increased oxidation of fatty acids with higher FFMI as well as a rather isolated signal for the diacyl phosphatidylcholine C38:3, which requires further in-depth investigations.

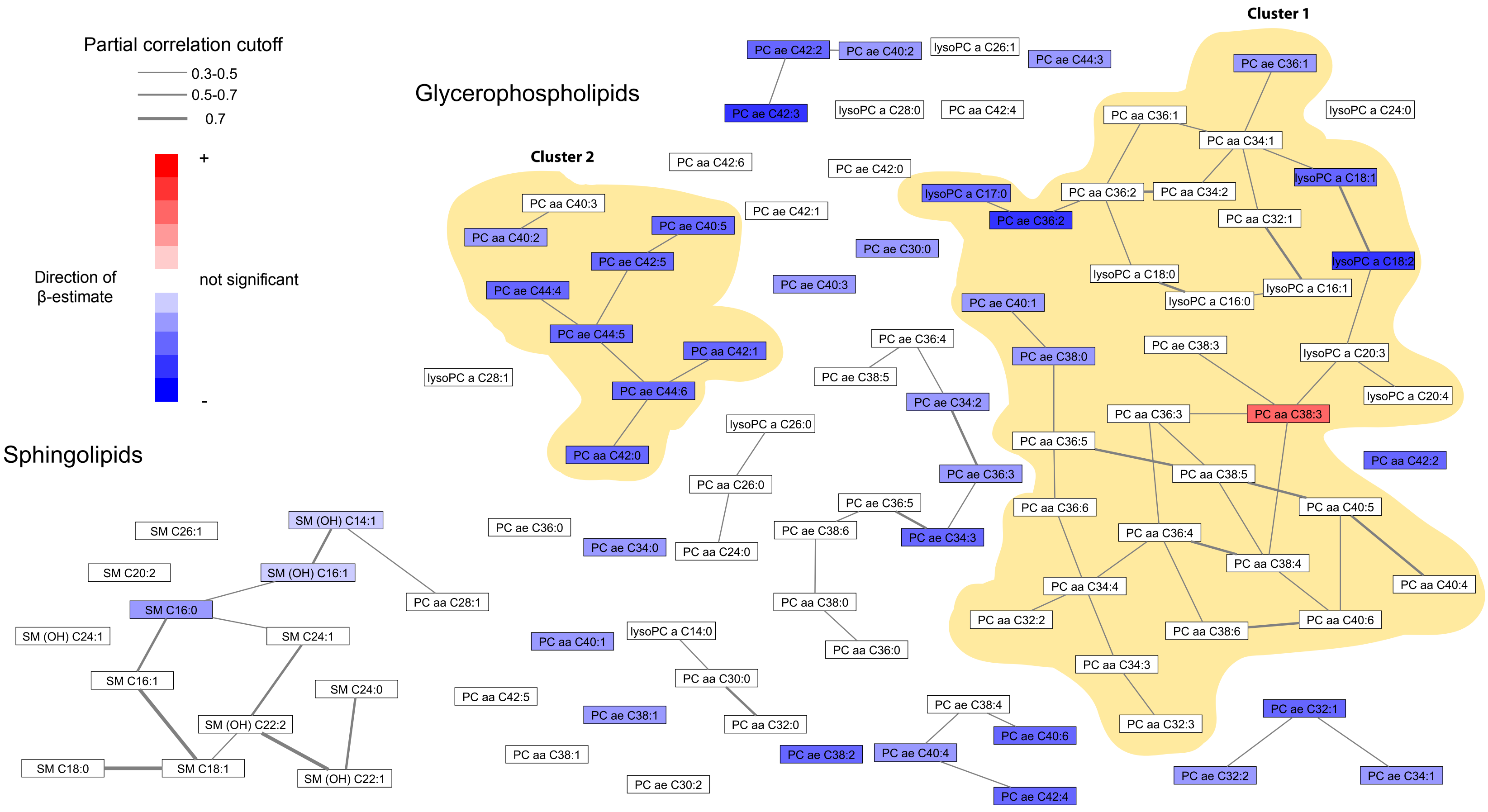


Figure 7.5: Effect network with FFMI (fat-free mass index) associations. Cluster 1 shows rather diverse associations with localized, positive effect of PC aa C38:3. This might represent an isolated spot in the metabolic pathway which associates with fat-free body mass. Cluster 2 displays coordinately downregulated PCs with very long chain fatty acid residues. PC=phosphatidylcholine. Figure adapted from Jourdan et al. [173].

7.1.3 Effects of Type D personality on the metabolome

We applied effect network approach in an additional project, where metabolomics differences of study participants diagnosed with a *Type-D personality* were investigated. Since the statistical backgrounds of graph coloring were already thoroughly introduced, we will only briefly describe the results here. A Type-D personality refers to a mental characteristics, where patients show a general liability to psychological distress, e.g. social inhibition and negative affect [61]. Moreover, Type-D has been associated with an increased risk of cardiovascular disease [174]. Again, we statistically assessed the differences between healthy and diseased subjects for all metabolites. We used data from the KORA F4 cohort with Metabolon measurements, including unknown metabolites (cf. Chapters 2 and 6).

The study is currently being published:

⋆ Altmaier, E., Emeny, R., **Krumsiek, J.**, Lacruz, E., Lukaschek, K., Haefner, S., Kastenmüller, G., Römisch-Margl, W., Prehn, C., Mohney, R.P., Milburn, M.V., Illig, T., Adamski, J., Theis, F.J., Suhre, K., and Ladwig, K.H. Metabolomic profiles in individuals with negative affectivity and social inhibition: a population-based study of Type D personality. *Psychoneuroendocrinology*, in press.

For 1,509 out the 1,768 KORA F4 participants we had questionnaire information on mental health, and a total 387 participants were subsequently diagnosed with a Type-D personality. A linear regression analysis with metabolites as dependent variables, the Type-D state as the explanatory variable, and age, gender, HDL, LDL, cholesterol, triglycerides, hypertension, BMI, diabetes and the intake of antidepressive medications as covariates for correction was performed (analogously to equation (7.1) in section 7.1.1). The statistical analysis of metabolite concentrations alone only revealed a single metabolite, kynurenine, which was significantly associated with Type-D after multiple testing correction. Kynurenine represents a trypthophan metabolite and was decreased in Type-D individuals. Interestingly, though distinct with respect to their pathogenesis, depression and schizophrenia have previously been reported to associated with kynurenine as well [175].

Using Gaussian graphical models, we could furthermore detect clusters of borderline-significant metabolite sets which, in the network context, still contained interesting signals. In particular, four metabolite clusters showing a Type-D impact could be identified: (1) several steroid sulfates and X-18601, which most likely also represents a steroid (cf. Figure 6.3) are connected in the GGM and display a positive association with Type-D. Interestingly, steroid hormones have previously been linked with schizophrenia [176]. (2) Tyrosine, gamma-glutamyltyrosine and gamma-glutamylphenylalanine were down-regulated in Type-D individuals. This association suggests a relationship of Type-D with the tyrosine-dopamine neurotransmitter pathway. (3) Caffeine, paraxanthine, piperine and X-11485 displayed lower concentrations in affected individuals. This signal most likely reflects dietary differences between affected and healthy individuals (suggesting that Type-D personalities drink less coffee). (4) Finally, 3-indoxylsulfate and X-12405 also displayed a negative association with Type-D.

In summary, we demonstrated another application of the effect network approach. Using the network context between metabolites, we uncovered statistical associations that would have been considered insignificant otherwise. Methodologically, we will attempt to develop network-based clustering algorithms which assign actual p-values instead of manually deriving groups from the networks.

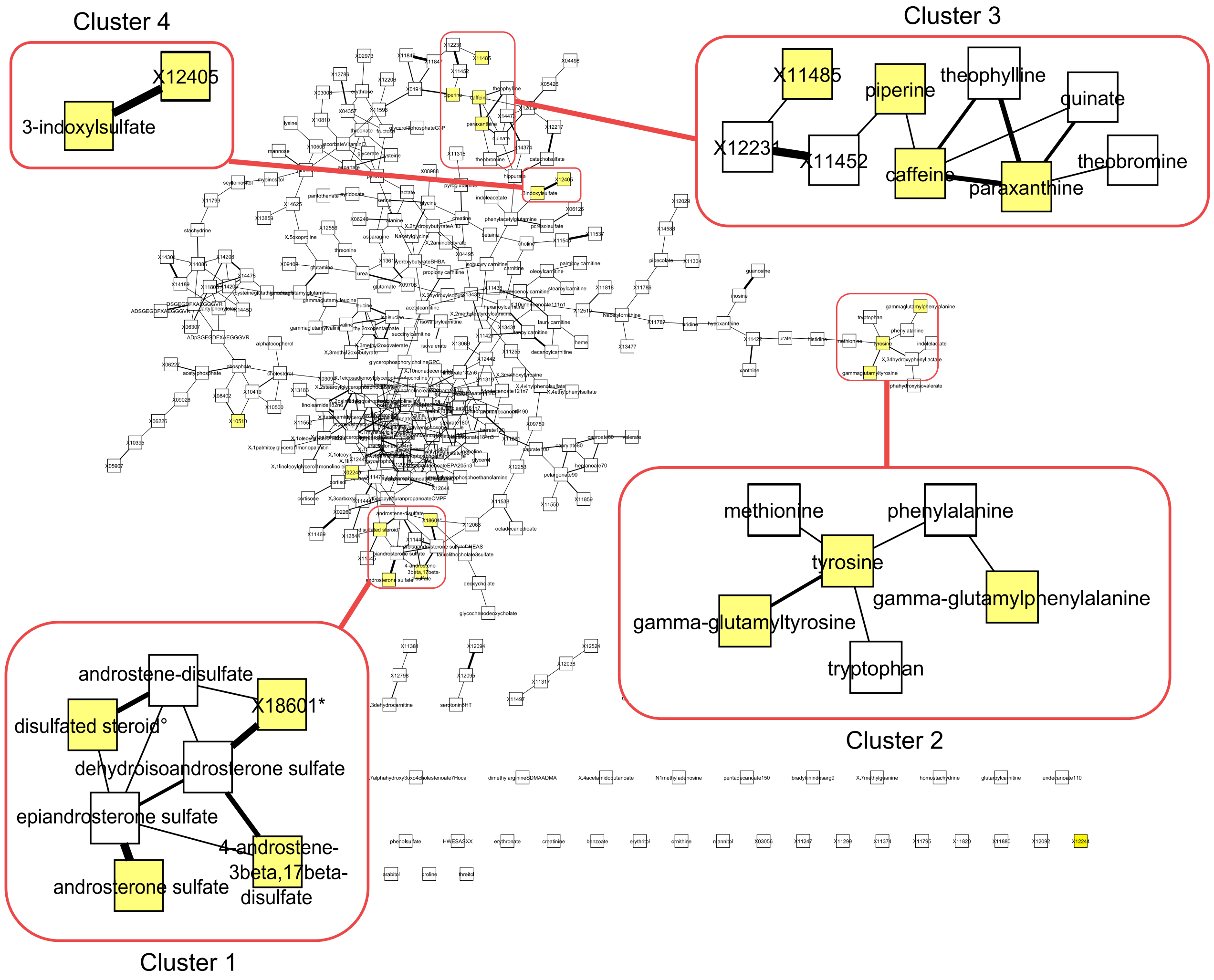


Figure 7.6: Effect network with Type-D personality associations. Each yellow node corresponds to a metabolite constituting a p-value below 0.01 for the association with the Type-D phenotype. We identified four metabolite clusters in the GGM which contain localized phenotype effects. These include clusters related to steroid hormones (Cluster 1), amino acids and dipeptides (Cluster 2), caffeine derivatives and piperine (Cluster 3), and 3-indoxylsulfate (Cluster 4). Adapted from Altmaier et al. [177].

7.2 Differentially regulated metabolism in glioblastoma cells

Malignant primary brain tumors, such as glioblastoma, are nearly always fatal despite considerable progress in clinical cancer therapy [178]. Glioblastoma are characterized by a resistance to apoptosis stimuli and the invasion of surrounding normal tissue. Experimental access to glioblastoma cells for *in vitro* experiments is given through the U87 cell line, which was derived from a human grade IV glioma in 1968 [179], and has been used in numerous publications since its generation [180, 181]. Moreover, U87 was the first fully sequenced cancer cell line genome [182]. A decade ago, Lang et al. [183] reported apoptosis and G2 arrest in U87 cells transfected with the tumor suppressor p53, followed by treatment with the chemotherapeutic agent Irinotecan or its active metabolite SN-38.

We investigated lipidomics data from U87 cells under seven treatment conditions – out of which only one constitutes a relevant apoptotic effect on the immortal brain tumor cells – and one control condition without treatment. Using a specialized Fourier-Transform Ion-Cyclotron-Resonance (FT-ICR) MS/MS technique [184], 167 polar lipids were measured across six lipid classes. In contrast to the phosphatidylcholine-centered metabolite panel from the Biocrates platform (cf. Chapter 2), the U87 experiments comprise phospholipids with additional head groups, including phosphatidylinositols (PI), phosphatidylserines (PS), phosphatidylethanolamines (PE), phosphatidic acid and sphingomyelins. Furthermore, a series of gangliosides were measured, a glycosylated lipid class specific to the nervous system [185]. The measured lipid panel also displays the previously-mentioned side chain ambiguity problem. That is, we again only get the sum of carbon atoms and double bonds for lipids with two fatty acyl side chains.

Cells were grown under eight different medium conditions (Table 7.1) in three biological replicates with three technical replicates each. All possible combinations of 24h treatment of SN-38 (chemotherapeutic agent), p53 (tumor suppressor viral transfection), and DI312 (control adenovirus vector transfection) were applied. Interestingly, the variant were p53 is applied first and SN-38 afterwards induces modest apoptosis and cell cycle arrest in G$_2$, whereas the reverse treatment induces almost almost complete arrest in G_2 and apoptosis of the majority of cells. Since the latter effect does not allow a proper analysis of the lipidome, we consider the p53/24hr + SN-38/24hr treatment to be the relevant experimental condition for our analyses. The data set was originally published by He et al. [178].

We applied a differential Gaussian graphical modeling approach in order to elucidate specific metabolic changes introduced by the apoptosis-inducing treatment variant. The study was per-

Condition	**Effect**
DI312/24hr + SN-38/24hr	—
p53/24hr + SN-38/24hr	modest apoptosis and cell cycle arrest in G_2 (relevant)
SN-38/24hr + DI312/24hr	almost complete G_2 arrest and apoptosis of 90% of the cells
SN-38/24hr + p53/24hr	—
DI312	—
p53/24hr	—
SN-38/24hr	—
no treatment	—

Table 7.1: Experimental conditions for U87MG glioblastoma cells. Out of the eight possible combinations, one induces almost full apoptosis, and one induced modest apoptosis with cell cycle arrest. The latter is considered relevant or 'active' for our study. SN-38: treatment with chemotherapeutic agent. p53: tumor suppressor viral transfection. DI312: control adenovirus vector transfection.

formed in close collaboration with Nikola Müller and Anke Meyer-Baese and was published in:

⋆ Mueller, N.S., **Krumsiek, J.**, Theis, F.J., Böhm, C., and Meyer-Baese, A. Gaussian graphical modeling reveals specific lipid correlations in glioblastoma cells. volume 8058, page 805819. SPIE, 2011.

From the methodological point-of-view, this project was particularly challenging due to the very small number of samples. In the following, we briefly summarize the introduced concepts and subsequent findings.

A differential Gaussian graphical model of glioblastoma metabolomics data

The experimental design for the glioblastoma study is substantially different to the epidemiological analyses discussed in the previous sections. Specifically, we here have eight different experimental conditions, out of which only one actually induces the desired apoptosis and cell cycle arrest effects in U87 cells (Table 7.1). Investigating pairs of measured metabolites, there are generally three scenarios of how the data point of this condition could influence their correlations (Figure 7.7): (1) The correlation could be *unspecific*, that is leaving in or out the specific data point would not significantly change the respective correlation. (2) In a *treatment-induced* scenario, the metabolites are actually uncorrelated, but coordinately react in the same direction upon treatment. We then observe a correlation which would not be present if the data point was

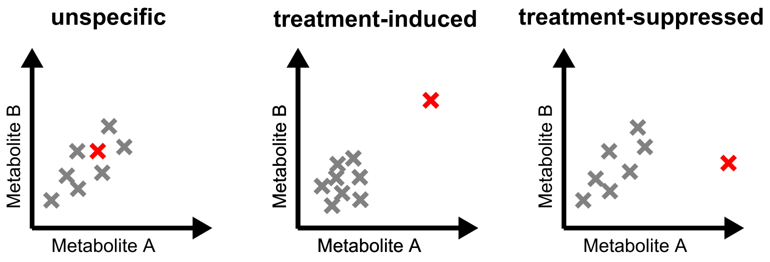


Figure 7.7: Influence of data points from the relevant treatment on pairwise correlations. See main text for a detailed description of the three scenarios. Note that replicates have been omitted in this diagram for simplicity.

left out. (3) Finally, there could be a *treatment-suppressed* situation, where the metabolites are correlated but only one of them changes upon the specific treatment. This would diminish the otherwise present correlation if the data point was used in the correlation analysis.

The GGM-based method to generate a treatment-specific metabolome network works as follows: First, the GGM based on all eight conditions is estimated, followed by eight specific GGMs where each experimental condition is left out once. Since we had considerably less samples than variables, and the covariance matrix of metabolites can thus not be inverted, we here employed a shrinkage-based GGM approach developed by Schäfer and Strimmer [83] (see also Chapter 3.5). GGMs were again constructed by means of statistical significance of the respective partial correlations, here assessed by a false-discovery rate [186] approach with q=0.01. An edge between two metabolites is then included in the differential GGM if it fulfills either one out of two criteria: (1) The edge is not present in the GGM where the active experimental condition was left out, but present in all other GGMs including the full GGM based on all samples (*treatment-induced* edge). (2) Vice versa, the edge is present in the GGM where the active condition was left out, but absent in all other GGMs (*treatment-suppressed* edge). As a result, we obtain a differential GGM for the glioblastoma lipidome, containing treatment-specific metabolite-metabolite associations resulting from the combination of p53 transfection prior to SN-38 chemotherapy. The network contains 45 out of the original 167 lipids, and 33 edges out of which 25 are *treatment-induced* and 8 are *treatment-suppressed* (Figure 7.8).

The results point out several positions in the metabolic network where specific changes due to the treatment might have occurred. For instance, three out of five measured sulfatides (a specialized class of ceramides) are present in the network. In particular, the oxidized sulfatide (34:2)+O plays a prominent role in the differential GGM with five suppressed edges. Furthermore, 17 out of 32 measured gangliosides (a lipid class primarily present in the nervous system) occur differentially regulated in the GGM. Finally, 14 out of 55 phosphatidylinositols (a com-

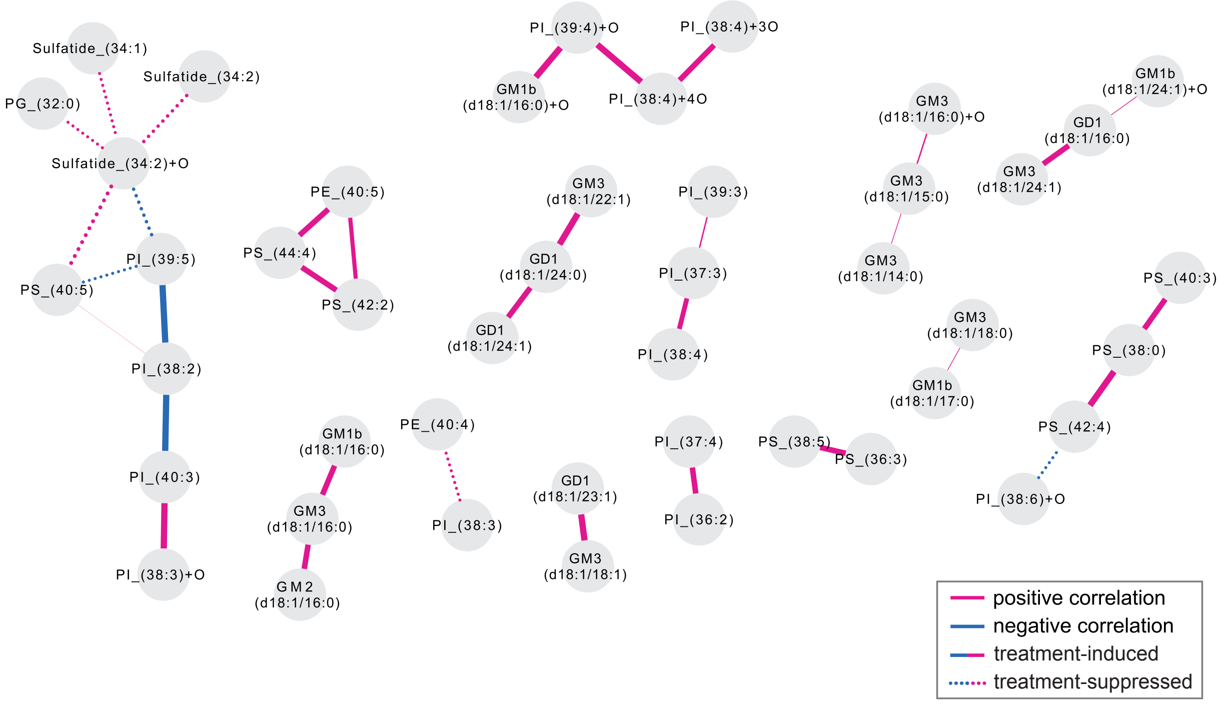


Figure 7.8: Differential GGM of glioblastoma lipidomics data. Colors indicate positive or negative correlations. Solid and dashed lines represent *treatment-induced* or *treatment-suppressed* edges with respect to the 'p53/24hr + SN-38/24hr' treatment. We observe an overrepresentation of sulfatides, gangliosides, and phosphatidylinositols in the network.

mon membrane lipid class) appear in the differential network. An in-depth biological analysis of the obtained results, possibly with experimental support from the collaboration partners, was left for future projects.

In summary, this project introduced an alternative approach of analyzing sample groups using GGMs. In contrast to the effect networks from the previous section, we here directly took into account the experimental design of the study. Certainly, the statistical power for GGM calculation with only eight (or even less for the sub-GGMs) samples should be considered rather limited. On the other hand, in the original shrinkage-based GGM paper (Schäfer and Strimmer [83]), the authors also worked on a set with only 8 experimental conditions and $p = 102$ measured variables. The differential GGM approach introduced here represents a pilot study which can be extended to other projects with more samples and by going deeper into the biological interpretation of the obtained results.

7.3 Phenotype set enrichment analysis

In this project, the GGM was used to define biologically meaningful metabolite groups. These groups were used as substrates for a previously published enrichment algorithm which incorporates multiple phenotypic traits at once in a genome-wide association study (GWAS). The method is called *phenotype set enrichment analysis* (PSEA) and has been published by Ried et al. [187]. Usually, GWAS primarily either focus on single phenotypic traits, like like diabetes type II [188] or coronary heart disease [189], or carry out the analyses for multiple phenotypes independently, as for instance in the metabolomics GWAS published recently [27, 29]. The PSEA algorithm, in contrast, performs an integrated statistical test of a gene[1] against a whole set of phenotypes. The rationale behind this approach are supposedly stronger associations of these phenotype sets with genetic variation in a locus than with the single phenotypes alone. Briefly, the study showed that using this method, new gene-phenotype associations can be revealed as well existing ones confirmed.

The statistical test behind the method is based on an aggregate of the single phenotype statistics followed by a permutation step in order to derive an empirical p-value. Note that this approach is very similar to the weighted enrichment algorithm we developed in Chapter 8, only that in our case we are working with contributions from the source matrix of an independent component analysis instead of statistical test statistics from a GWAS.

The following section describes how our metabolomics Gaussian graphical modeling methodology helped to defined meaningful phenotype sets in a data-driven manner, irrespective of the genetic associations. In contrast to the previous sections, rather than following the spread of a statistical signal through the network, we were interested in metabolite groups naturally arising from the partial correlation network.

Metabolomics GGMs and PSEA

Metabolite concentrations were used as phenotypic traits, and GGMs were employed to define the respective phenotype sets. For this project, partial correlations based on both the Biocrates dataset (Chapter 5) and on the Metabolon dataset (Chapter 6, but only calculated on the known metabolites) were used. The partial correlation matrices were cut at different absolute partial correlation values (0.3 and 0.45). This yields two granularities of connected components in the

[1] This actually refers to a gene and not a SNP, since the authors first combine all SNP hits of a given gene into a single statistic.

graph, and thus overlapping clusters of metabolites. As already seen in the previous chapters, these clusters are homogeneous with respect to the annotated metabolic classes or major pathway assignments (e.g. carnitines, phosphatidylcholines, amino acids, etc.). Thus, independent of the partitioning of metabolites provided by, for instance, putatively biased pathway databases, we here generated biologically feasible sets from dependency structures in the data only. The PSEA was then carried out both on 1,809 genotyped individuals from the KORA cohort for discovery, as well as on data from the Twins UK study [190] for replication. Interestingly, several loci could be detected using the GGM-based metabolite sets that would otherwise only be detectable by the usage of metabolite ratios. While the ratios represented a (rather successful but) simple approach to capture metabolic relationships, GGMs specifically describe the metabolites' biochemical relations. Furthermore, using metabolite sets defined by the GGM instead of ratios tremendously reduces the amount of statistical tests that have to be performed.

In summary, this application represents another case study utilizing the dependency structures behind metabolomics data in the sense of a data-driven metabolic network. Here, the GGM provide substrate sets for the evaluation of a weighted enrichment algorithm developed by our collaboration partners. From a conceptual point-of-view, using biochemically related groups of metabolites instead of single metabolites in a GWAS might improve the power of the analysis. Note that this idea resembles the effect networks from Section 7.1, where we also combine common statistical analyses with the network structure in order to improve the sensitivity and biological relevance.

7.4 Metabolite ratios, genetic networks and GGMs

Recent metabolomics studies demonstrated metabolite ratios, that is the concentration of one metabolite divided by the concentration of another, as valuable markers for statistical analyses. For instance, metabolite ratios have been used to detect metabolic relationships with medication [120], smoking [23] and genetic variation [27, 29]. This effect can most probably be attributed to the reduction of biological variation for correlated metabolites. When two metabolites display similar concentration patterns, then taking the ratio of both concentrations cancels out the biological variation and produces a more accurate readout of the current relationship between the two compounds. In particular, metabolite ratios turned out to substantially improve p-values and increase the explained variance in statistical analyses compared to the metabolite concentrations alone [27, 29]. Historically, metabolite ratios have long been known to be valuable biomarkers, e.g. the phenylalanine-tyrosine ratio for the diagnosis of phenylketonuria [191], or the lactate-pyruvate ratio for the detection of deficiencies in energy metabolism [192].

The above-mentioned studies introduced a specific measure which captures the improvement in statistical association due to using the metabolite ratio: the *p-gain*. It is defined as the ratio of the smaller of the two single metabolite p-values divided by the p-value of the ratio:

$$\text{p-gain}\,(M_1, M_2, \text{SNP}) = \frac{\min\left(\text{p}(M_1|\,\text{SNP}), \text{p}(M_2|\,\text{SNP})\right)}{\text{p}(M_1/M_2|\,\text{SNP})},$$

where M_1 and M_2 represent metabolites, SNP is a specific SNP under investigation, and $\text{p}(x|\,\text{SNP})$ represents the p-value obtained when regressing x against the SNP in a model with additive genetic effects. In other words, it reflects the factor of p-value decrease achieved by taking the ratio. A major drawback of the p-gain application, however, was (a) the usage of vague 'rule-of-thumb' criteria to determine whether a p-gain itself be considered significant or not, and (b) the lack of a systematic evaluation of whether metabolite pairs with high p-gain values indeed represent biologically meaningful connections. Therefore, in a collaboration with Ann-Kristin Petersen, we sought to investigate both statistical as well as biological properties of p-gains:

⋆ Petersen, A.K., **Krumsiek, J.**, Wägele, B., Theis, F.J., Wichmann, H.E., Gieger, C., and Suhre, K. On the hypothesis-free testing of metabolite ratios in genome-wide and metabolome-wide association studies. *BMC Bioinformatics*, 13:120, 2012.

A specific derivation of the cumulative distribution function of p-gains is provided in this work, which then allows to construct statistical tests for the p-gain measure. In the following, we will summarize the findings of this paper and then provide a direct comparison of GGMs with genetically-determined metabolic networks resulting from GWAS with metabolite ratios. From the statistical side, the main finding of this work was a connection between p-gain significance and the metabolite correlation structure: If two metabolites are uncorrelated, then their ratio will display a high variation. Thus, a high p-gain is required in order to be considered significant in this case. On the other hand, if metabolites are strongly correlated, then their ratio will display low variation, and already modest p-gain values may become significant.

The p-gain significance calculation method was then used on n=1,814 samples from the KORA cohort with metabolomics measurements from the Metabolon platform and the large-scale genotyping data (the same dataset as in Chapter 6, without unknown metabolites). P-gains were then compared to several predefined pathway-based metabolite sets in order to evaluate whether pairs with a high p-gain are 'biologically related' (i.e. participating in the same pathway). The analysis showed that even down to a p-gain value of 10, 13.97% of all metabolite pairs were biologically related for at least one of the metabolite sets.

The p-gain values were systematically compared to the data-driven metabolite pairs defined by the metabolomics GGM. Note that this analysis is not part of the original publication. The number of metabolites pairs that display a p-gain above a given threshold rapidly decreases for larger p-gain cutoffs (Figure 7.9A). For instance, while there are 2396 metabolite pairs with a p-gain above 10^3, only 65 pairs show a p-gain above 10^{10}. Interestingly, higher p-gain values for metabolite pairs from the genetics analysis coincide with lower p-values for the respective GGM edge (Figure 7.9B). This indicates an interesting relationship between associations with genetic variation in the large population cohort and intrinsic dependencies between the metabolites as determined by the GGM. Both GGM and GWAS appear to recover similar pairs of metabolites from the data independently. In other words, there is not only a biochemical footprint of metabolic pathways in the blood serum data, but also a genetically-determined one.

Lists of metabolite pairs with a high p-gain can be regarded as a genetically-determined metabolite association network extending and complementing the structures detected by the GGM or the integrated GGM/genetics networks discussed in Chapter 6. Importantly, these networks contain ternary edges between two metabolites and one SNP (represented by its respective gene) each, thus adding an additional layer of functional information. A small example network is shown in Figure 7.9B. It displays several metabolite-gene associations for carnitine metabolism and ACAD β-oxidation enzymes as well as several fatty acids and the SCD desaturase enzymes. Note that MSH4 represents a poorly characterized homolog from E.coli, which possibly represents a meiosis-specific protein. Given the associations shown here, the functional annotation might be reconsidered to represent a fatty-acid related enzyme. Importantly, in this association network, the ternary metabolite-metabolite-gene associations provide more complex relationships than the simple genetic associations investigated in Chapter 6.

Taken together, we performed a rigorous analysis of a simple statistical enhancement of genome-wide association studies. The results will further aid in the analysis of GWAS and a more meaningful interpretation of the p-gains derived from metabolite ratio analysis. Compared with GGM edges, metabolite pairs with higher p-gain values tend to have smaller p-values in the GGM, providing a direct link between genetically-determined metabotypes and intrinsic metabolite dependencies. Finally, we outlined an extended genetic network approach which incorporates both metabolite ratio information as well as genetic information.

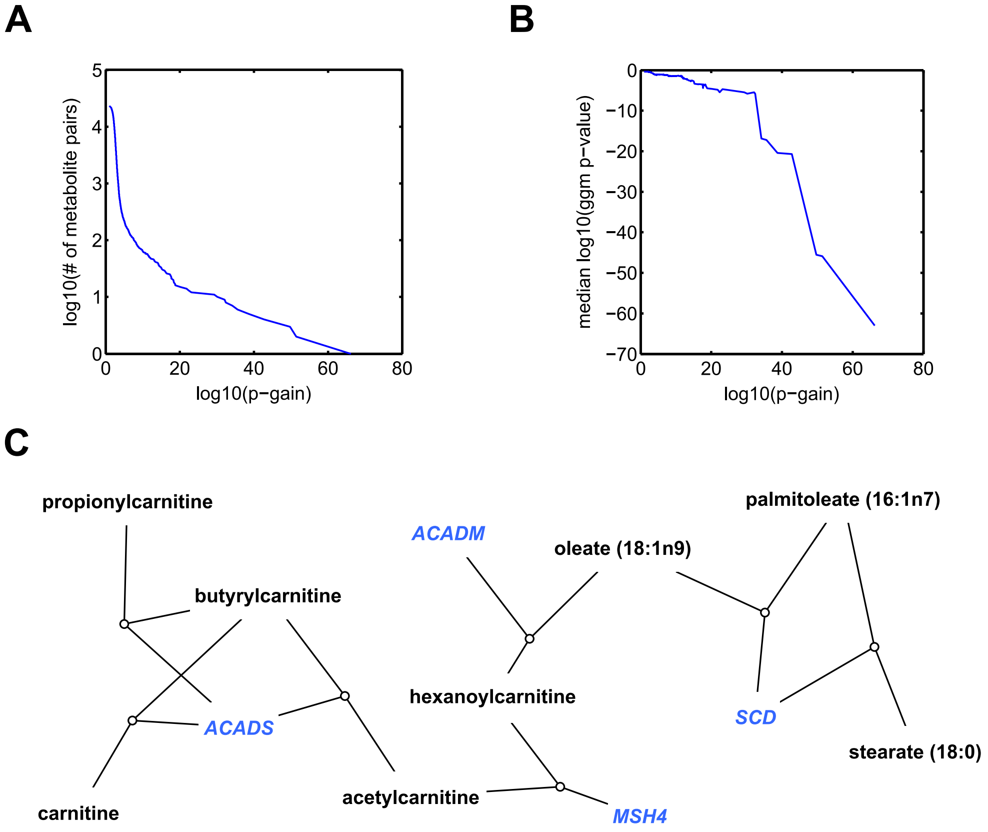


Figure 7.9: Network analysis for metabolite ratios, the p-gain and GGMs. **A:** The number of metabolite pairs above a given threshold rapidly decreases for increasing p-gain values. **B:** Relationship between p-gain and GGM edge p-values. For each p-gain (x-axis) we collected all metabolite pairs that constitute a p-gain of at least this value and subsequently calculated the median GGM edge p-value of the corresponding metabolite pair list (y-axis). We observe a clear tendency of lower GGM edge p-values for higher p-gain values, providing a direct link between genetic associations and intrinsic metabolite dependencies. **B:** Genetically-determined metabolic network derived from genotyping data combined with metabolomics. We show an exemplary subnetwork of lipid-associated processes including ratio information. Each small circle represents one ratio, connecting two metabolites and one gene locus each. The original SNPs are not provided in this diagram; genes were assigned via linkage disequilibrium as before (cf. Chapter 6, Table 6.1).

Chapter 8

Beyond covariance: Higher-order dependencies in metabolomics data

In the past chapters, we discussed the inference and functional analysis of Gaussian graphical models (GGMs) on metabolomics data. While we have seen that GGMs recover biologically related pairs of metabolites, covariance-based approaches only exploit second-order dependencies in the data (recall that variance is the second central moment of a distribution). However, in practice we frequently observe higher-order dependencies, which may yield additional information that is otherwise neglected. Metabolomics data, for instance, will never perfectly follow a Gaussian distribution even after logarithmizing[1], thus leaving multivariate dependencies which cannot be captured. Another prominent example of a second-order statistics-based analysis, principal component analysis (PCA), searches for mutually decorrelated directions in the data matrix, which then explain maximal variance [193]. PCA is commonly used as a tool for the initial analysis of high-dimensional data sets, especially in the metabolomics field [194]. In this study, we aimed at using the full-order multivariate statistics in an explorative analysis of metabolomics data; hence we proposed the use of independent component analysis (ICA) as a statistically motivated extension of PCA for metabolomics data [195]. The introduction of statistical independence here naturally generalizes the concept of decorrelation for non-normal data.

[1] QQ plots against a normal distribution can be downloaded from `http://helmholtz-muenchen.de/cmb/ggm`

For ICA we assume metabolite profiles to be composed of statistically independent components (ICs), whose mixture makes up the measured metabolomics profile. Let $\mathbf{X} = (x_{ij}) \in \mathbb{R}_+^{n \times p}$ be the pre-processed data matrix, where each of the n rows corresponds to one measured study proband, and each of the p columns represents one metabolite. For a given number of components k, independent component analysis attempts to find a factorization of the data matrix

$$x_{ij} = \sum_{l=1}^{k} a_{il} \cdot s_{lj} + \epsilon_{ij}, \tag{8.1}$$

where the *mixing matrix* $\mathbf{A} = (a_{il})$ is of dimension $n \times k$, the *source matrix* $\mathbf{S} = (s_{lj})$ is $k \times p$, and ϵ_{ij} represents independent, normally distributed noise (Figure 8.1A). The particularity of ICA is the requirement of all rows $s_{l\cdot}$ in $\mathbf{S}$ (which we will refer to as IC$_j$) to be samples of a statistically independent random vector. Interpreted biologically, each row in $\mathbf{S}$ represents a distinct metabolic process, which contributes to the overall concentration profile. The matrix $\mathbf{A}$, on the other hand, reflects how strong each of these processes is *active* in a given sample (study proband in our case). In other words, instead of describing the metabolome of each proband by p numeric values, after ICA we can equivalently represent the metabolome using only $k << p$ values. It can be shown that the decomposition into $\mathbf{A}$ and $\mathbf{S}$ is unique given sufficiently many samples [196, 197].

In biomedical research, ICA is commonly used as a method for high-dimensional data reduction and analysis. Early applications from the neuroscience field include the analysis of electroencephalographic measurements [198] and fMRI data [199–201]. For molecular biology, ICA has frequently been used to analyze transcriptomics data, e.g. for cancer classification [202–204] or the investigation of cell differentiation [205, 206]. Moreover, several studies already applied ICA in the context of metabolomics data, for instance for the analysis of plant parasites [207] and toxins [208], and for metabolite fingerprinting [209]. While certainly interesting for their respective biological questions, these metabolomics studies merely used ICA as a data compression and visualization method rather than functionally investigating the reconstructed independent components in detail. The only studies which, to the best of our knowledge, performed a functional analysis of $\mathbf{A}$ and $\mathbf{S}$ are (i) Wienkoop et al. [210], who did a joint ICA of metabolomics and proteomics data in starch metabolism and (ii) Martin et al. [211], who investigated the development of colitis in mice using NMR metabolomics.

For this study we employed a Bayesian independent component analysis approach. The key idea of Bayesian inference is to interpret each parameter as a random distribution. These distributions are then estimated using Bayes rule, for example by Markov chain Monte Carlo methods, or simply by maximum a posteriori estimation. With an inferred parameter distribution at hand,

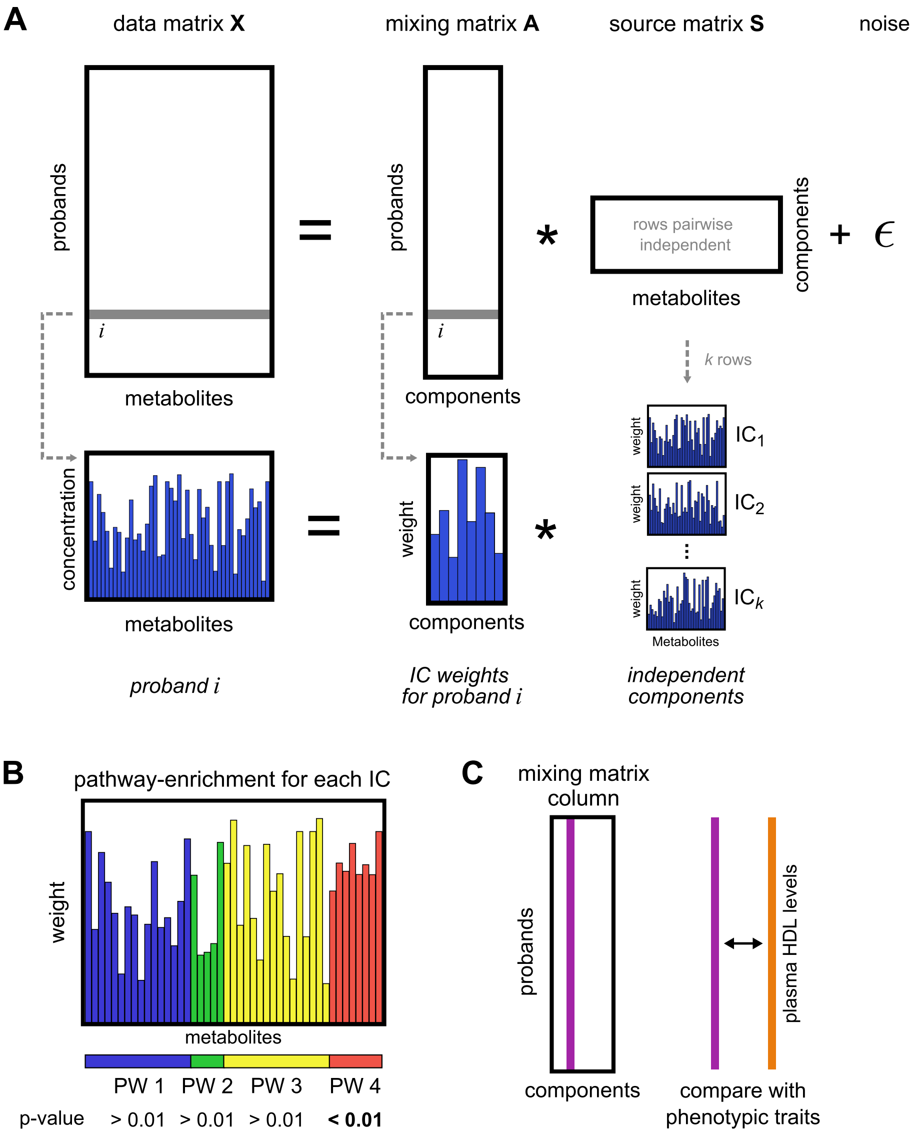


Figure 8.1: **A:** Independent component analysis model applied to metabolomics data. The data matrix **X** is decomposed into the product of a mixing matrix **A** and a source matrix **S**, cf. equation (8.1) in the text. The source matrix contains statistically independent profiles of metabolites ($s_{l\cdot}$, termed 'IC' = independent component throughout the chapter), whereas the mixing matrix represents the contribution strengths of each component to the respective metabolomics sample. **B:** Concept of pathway enrichment performed for each independent component. We statistically assess whether the IC contributions for the metabolites from a specific pathway are higher than expected by chance. **C:** Each column in the mixing matrix represents a newly derived variable in the dataset which can be correlated with other proband-specific traits.

we can obtain both conventional point estimates, but also parameter error estimates as provided by the respective variance. Moreover, by choosing adequate priors, we can include known information beforehand. In our case, we require nonnegative values of both the source and the mixing matrix. We argue that such nonnegativity better represents biological processes than arbitrarily negative matrix entries. In classical ICA, the choice of model parameters such as the number of components k to be reconstructed is a non-trivial problem. Usually, an ad-hoc number of components is chosen, thereby accepting possible fusions of components (if too few are selected) or generation of information-free noise components [195]. A series of tools for identifying the correct model have been developed in the ICA community, mostly using heuristics e.g. based on clustering similar components [212, 213]. We here evaluate the Bayesian Information Criterion (BIC) for each ICA calculation to get a trade-off between model accuracy (how close the matrix product gets to the original data matrix) and the number of parameters in the model. Finally, we select the number of components for which we obtained the highest BIC value. Methodologically, we applied a Bayesian mean-field ICA method [214], which uses an EM-like parameter estimation scheme.

The novelty of our approach is the application of a parameter-free, Bayesian, noisy ICA approach to metabolomics data, followed by a functional analysis of both independent metabolite processes in **S** as well as proband-specific signals in **A**. *Parameter-free, noisy, Bayesian* here refers to, (i) avoiding a manual selection of the number of components k, (ii) obtaining an actual distribution for **S**, thus providing confidence intervals for the reconstructed values, and (iii) allowing for an independently estimated noise term ϵ_{ij}.

The chapter is organized as follows: First, we apply ICA to a large dataset of human blood serum metabolomics samples of 1764 probands and 218 measured metabolites (Figure 8.1A), and estimate the number of components k using the above-mentioned Bayesian mean-field ICA approach. Next, we investigate the source matrix **S**, first by manual investigation and then by calculating the statistical enrichment of known metabolic pathways in each component (Figure 8.1B). We demonstrate that the approach outperforms PCA, k-means clustering as well as fuzzy c-means with respect to biological pathway enrichment. In the final results part, we correlate the columns of the mixing matrix **A** to HDL (high-density lipoprotein) concentrations in blood plasma (Figure 8.1C). One independent component correlates stronger with HDL concentrations than all metabolites in the dataset alone. We thereby establish a novel connection between blood plasma HDL and branched-chain amino acids, and discuss potential biological implications.

All results reported in this chapter are part of the following publication:

$\star$ **Krumsiek, J.**, Suhre, K., Illig, T., Adamski, J., and Theis, F.J. Bayesian Independent Component Analysis recovers pathway signatures from blood metabolomics data. *Journal of Proteome Research*, 11(8):41204131, 2012.

We published a preliminary extension to independent *subspace* analysis on a similar dataset in Gutch et al. [215].

8.1 Methods

Bayesian ICA model & component selection

In this study we used the Metabolon dataset without unknown metabolites (cf. Chapters 2 and 6). For preprocessing, the data matrix **X** was column-normalized to unit variance and subsequently scaled between 0 and 1.

We solved the described noisy source separation problem by probabilistic independent component analysis [216, 217]. Assuming normally distributed white noise with covariance matrix Σ, the mixing model results in the model likelihood

$$P(\mathbf{X}|\mathbf{A},\mathbf{S},\Sigma) = (\det 2\pi\Sigma)^{-N/2} \exp\left(-\frac{1}{2} tr(\mathbf{X}-\mathbf{AS})^T \Sigma^{-1} (\mathbf{X}-\mathbf{AS})\right),$$

which describes the probability of observing data **X** given mixing matrix **A**, sources **S** and noise with covariance Σ. Instead of maximizing this likelihood, we follow a Bayesian approach and consider the model posterior $P(\mathbf{A},\mathbf{S},\Sigma|\mathbf{X}) \propto P(\mathbf{X}|\mathbf{A},\mathbf{S},\Sigma)P(\mathbf{A})P(\mathbf{S})P(\Sigma)$ with (independent) priors $P(\mathbf{A}), P(\mathbf{S})$ and $P(\Sigma)$. Full sampling of this posterior is too time consuming and requires more elaborate Markov Chain Monte Carlo sampling. We decided to follow a simpler two-step EM-type algorithm by iteratively estimating first source posterior $P(\mathbf{S}|\mathbf{X},\mathbf{A},\Sigma)$ and then point estimates of **A** and Σ using a MAP (maximum-a-posteriori) estimator. We used a mean-field based algorithm proposed by Højen-Sørensen et al. [214], since it allows flexible choice of source priors. We assumed nonnegative mixing matrix and exponentially distributed source weights. We then analyzed the resulting point estimates for mixing matrix and noise covariance as well as the source distributions, which are shown componentwise as mean and standard deviation.

The model assumes a fixed number k of source components. We determined the optimal number of components using the Bayesian information criterion (BIC) [218]. It is here defined as

BIC $= -pL + \frac{1}{2}(nk+1)\log(p)$, where L represents the log-likelihood of the fitted ICA model. We chose the k for which BIC gets minimal.

The information content of each independent component was assessed by means of kurtosis, i.e. the fourth standardized moment. The kurtosis β_i of each IC$_i$ is defined as

$$\beta_i = \frac{\frac{1}{p}\sum_{i=j}^{p}\left(\mathbf{S}_{ij} - \overline{\mathbf{S}_{i\cdot}}\right)^4}{\left(\frac{1}{p}\sum_{i=j}^{p}\left(\mathbf{S}_{ij} - \overline{\mathbf{S}_{i\cdot}}\right)^2\right)^2},$$

where p is the number of metabolites (i.e. the number of columns in $\mathbf{S}$) and $\overline{\mathbf{S}_{i\cdot}}$ denotes the average value of independent component i.

Weighted enrichment analysis

Let p again be the number of metabolites in our dataset and c be the number of distinct class annotations. We investigate the class enrichment in a vector $\mathbf{w}$ of non-negative weights: $w_i \in \mathbb{R}_+,$ for each metabolite $i = 1, ..., p$. Class assignments are specified in the Boolean matrix $\mathbf{B} = (b_{ij})$ of dimension $p \times c$ by

$$b_{ij} = \begin{cases} 1, & \text{if metabolite } i \text{ belongs to class } j \\ 0, & \text{else} \end{cases} .$$

We now compute the class enrichment vector $\mathbf{e}$ of dimension c as $\mathbf{e} = \mathbf{B} \cdot \mathbf{w} \in \mathbb{R}^c$, i.e. for each class we simply sum up the contributions of all metabolites that belong to that specific class.

The values in $\mathbf{e}$ have no properly defined scale and can thus not be directly interpreted. Instead, we randomly shuffle the metabolite-class associations $r = 10^7$ times and recalculate a randomized vector $\mathbf{e}_r$. Let $\mathbf{f}$ contain the number of randomized values among all sampled $\mathbf{e}_r$ that are larger than the respective elements in $\mathbf{e}$. We compute the empirical p-value vector of length c as $\mathbf{p} := \frac{\mathbf{f}}{r}$. The result vector $\mathbf{p}$ thus contains one empirical p-value for the enrichment of each class in $\mathbf{w}$.

PCA, k-means and fuzzy c-means clustering

Principal component analysis (PCA) represents a standard multivariate data analysis procedure reviewed, for instance, in Shlens [193]. Briefly, similar to ICA, PCA represents a mixture model, where the data matrix $\mathbf{X}$ is split into two matrices $\mathbf{A}$ and $\mathbf{S}$ such that $\mathbf{X} = \mathbf{A} \cdot \mathbf{S}$. In contrast to ICA, $\mathbf{S}$ is here chosen such that all components are decorrelated, i.e. cov $\left(\mathbf{S}^T\right) = \mathbf{0}$. For k-means and fuzzy c-means clustering, we used the MATLAB-integrated functions `kmeans` and `fcm`, respectively. As a second variant of the fuzzy c-means approach, we only set the highest value of each metabolite in the fuzzy clustering matrix to 1 and the rest to 0 (thus again creating a hard clustering as produced by k-means). For all methods but ICA we logarithmized and subsequently column-normalized the data matrix.

Regression analysis

Associations between HDL values and the component strength vectors (columns) of the mixing matrix as well all metabolites were estimated using linear regression analysis. Before performing the actual analysis we removed from the data (i) age effects by only taking the residuals from a linear regression of the mixing matrix and the metabolite matrix columns on age, and (ii) gender-specific effects by subtracting the group-wise medians from each column in the data. We then regressed the HDL values on both the mixing matrix columns and each metabolite using the MATLAB `regress` function. P-values were obtained from the t-distribution with studentized residuals, the explained variance is determined by the coefficient of determination R^2. For the linear model forward feature selection algorithm based on AIC (Akaike information criterion), we used the `R platform` function `step` with setting `direction='forward'`.

8.2 Bayesian noisy ICA on metabolomics data

For data preprocessing, we normalized each column in the data matrix (1764 probands, 218 metabolites) to a standard deviation of 1 and subsequently scaled the values between 0 and 1. The following ICA calculations are based on the Bayesian mean-field ICA approach described in Højen-Sørensen et al. [214]. We assumed a nonnegativity prior for $\mathbf{A}$, an exponential distribution (and thus positive values) for $\mathbf{S}$, and an isotropic noise model for ϵ_{ij}. In order to determine the number of components k to be used, we calculated the Bayesian Information Criterion (BIC) for $k = 2$ up to $k = 30$ components, with 100 random initial conditions (Figure 8.2A). The diagram demonstrates (i) proper convergence of the algorithm due to similar BIC values in multiple runs for each k, and (ii) a clear BIC peak around 7 to 10 components. The highest score in the analysis was achieved for one run at $k = 8$, so we chose this number of components for all subsequent analysis steps. For higher numbers of k, the increase in reconstruction quality is

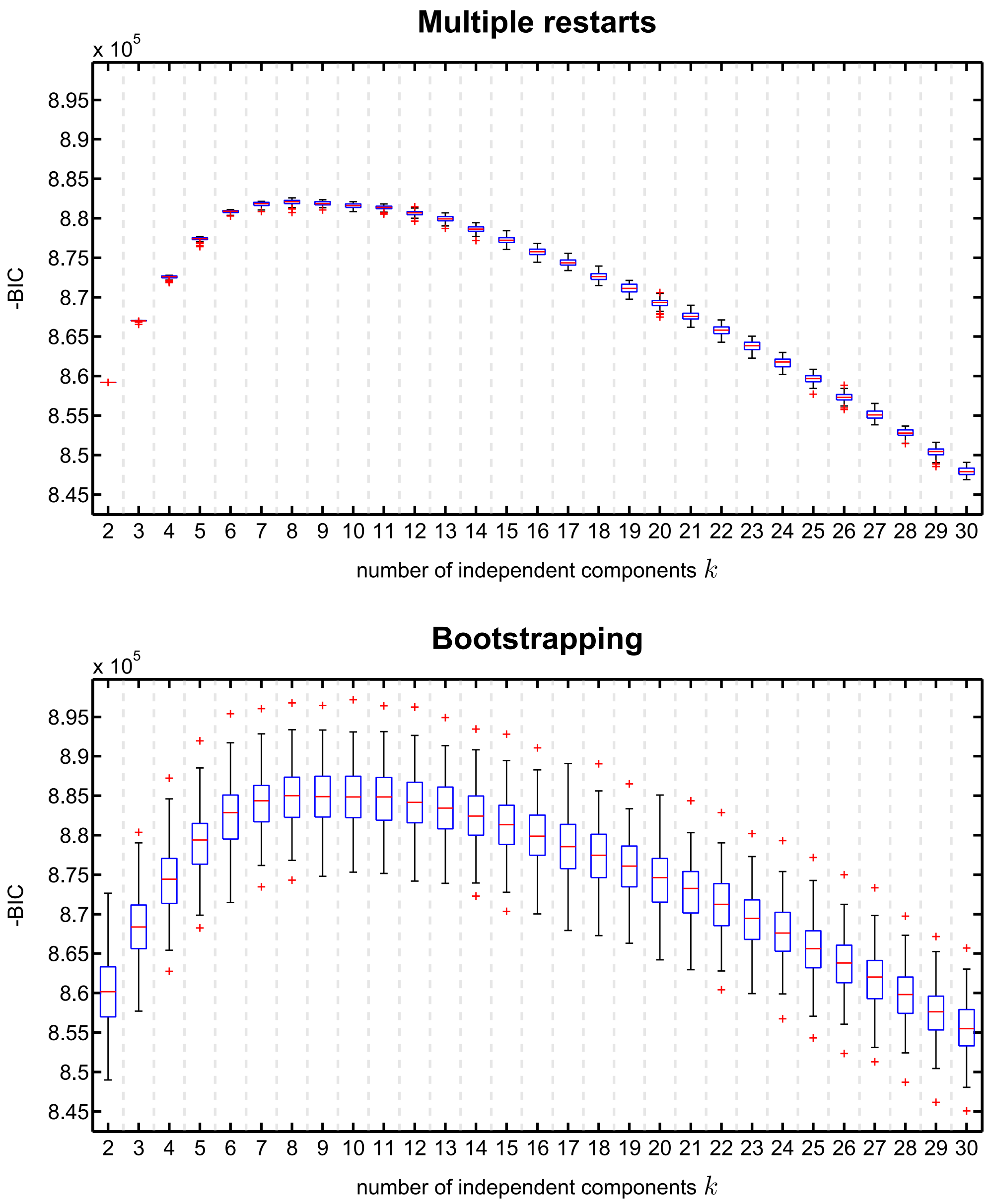


Figure 8.2: Selection of the number of components. **A:** The Bayesian information criterion (BIC) of the ICA model was estimated according to Højen-Sørensen et al. [214] for a range of k values, with 100 random initial value conditions for each k. We observe a clear peak around 7 to 10 components and choose $k = 8$ for all subsequent analyses. **B:** Stability analysis. The estimation variance is higher when performing ICA on bootstrap samples, but the position of the minimum BIC peak remains stable.

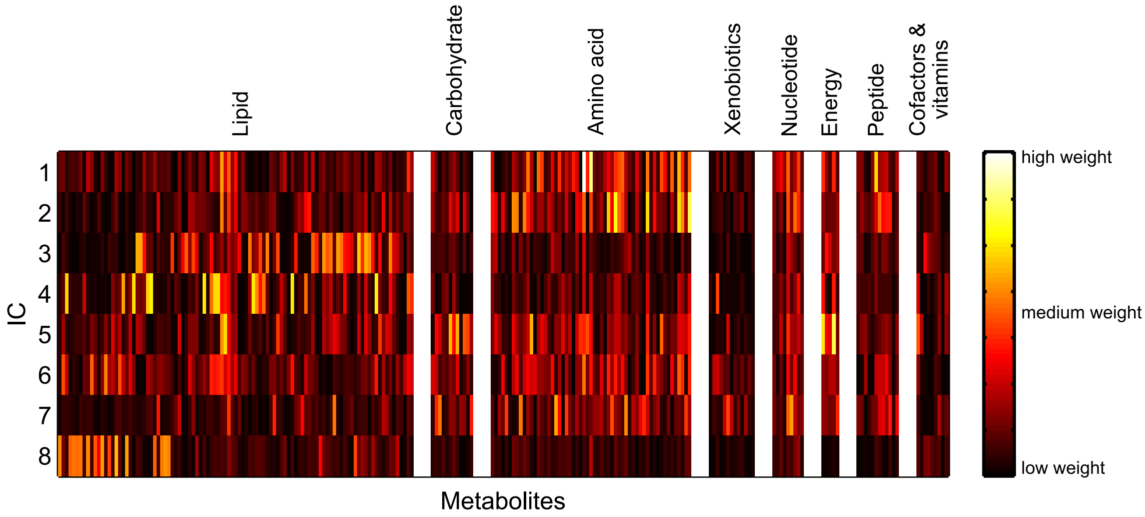


Figure 8.3: The source matrix **S**, grouped by the 8 metabolic *super-pathways* in our dataset. Rows are pairwise statistically independent and contain the contributions of all metabolites to the respective component. Already from this visual inspection we can see enrichments for specific pathways in each component, e.g. *Amino acid* in IC_1 and IC_2 and *Lipid* in IC_4 and IC_8.

not sufficient to compensate for the penalty imposed due to more parameters in the model. In order to verify the stability of the choice of k with respect to changes in the underlying dataset, we employed a sample bootstrapping approach (8.2B). This robustness analysis did not reveal significant differences to the full dataset run.

The resulting matrices **S** (with estimated parameter variance) and **A** are visualized in Figures 8.3/8.4 and 8.7, respectively, and will be subject to detailed functional analyses in the following sections.

8.3 Manual investigation of independent components in S

While the separation of the metabolomics dataset into 8 independent components might be sound from a statistical point-of-view, we have to ask whether we can gain insights into metabolic processes underneath giving rise to the data. Each component consists of a vector $s_{l\cdot}$ of non-negative contribution strengths, that is one value for each metabolite (Figure 8.3). In order to get an overview of the metabolic functions the components might be involved in, we manually investigated the 15 strongest contributions for each component (Figure 8.4). Estimation certainty is generally high, as indicated by small error bars resulting from the probabilistic ICA

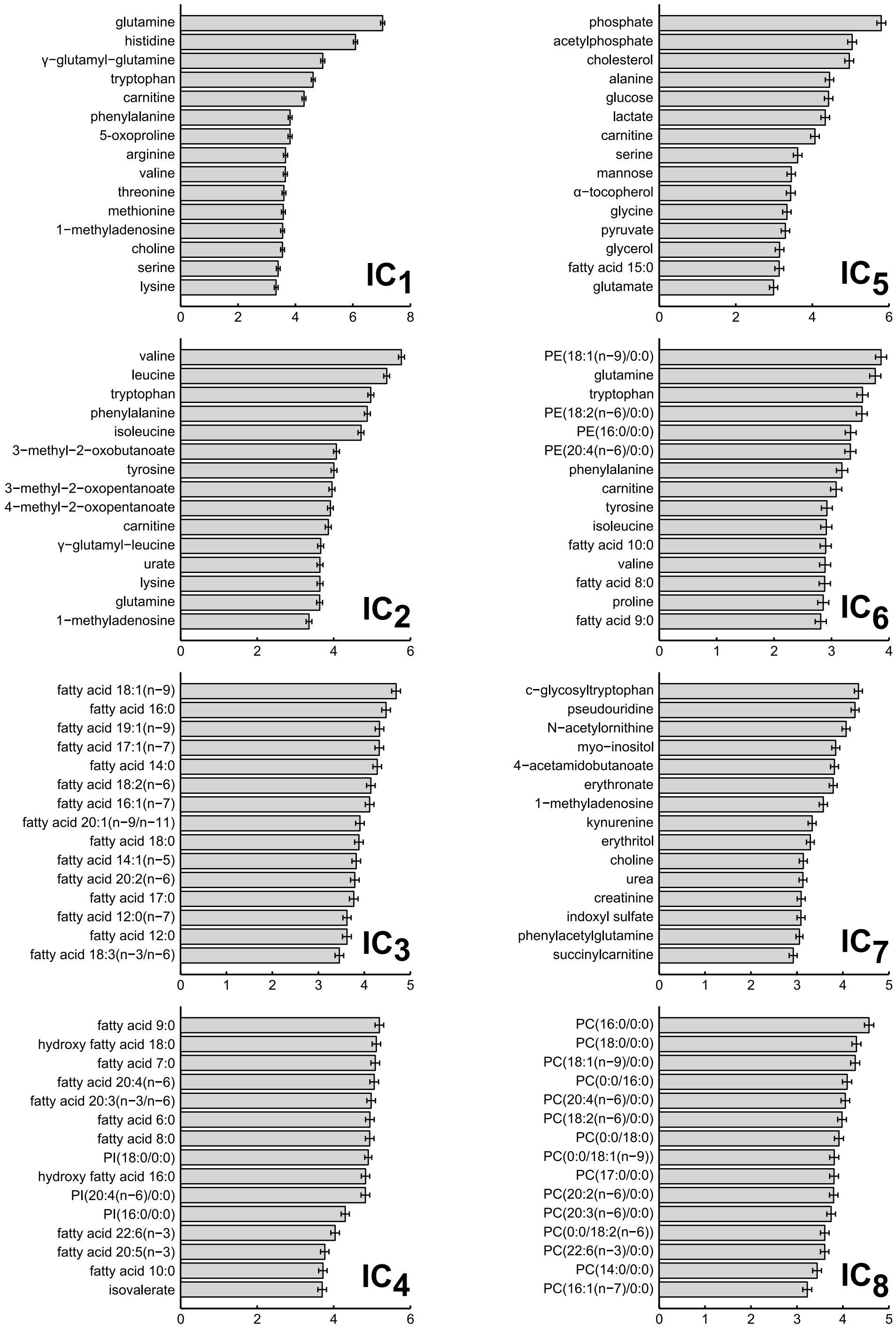


Figure 8.4: Top 15 metabolite contributions for each independent component in **S**. For most components, we observe strong tendencies towards specific parts of cellular metabolism. For instance, IC_2 contains branched-chain amino acids and their degradation product among its highest contributing metabolites. IC_8 contains phosphatidylcholines for various chain lengths and desaturation grades, and so on. Error bars indicate standard deviations from the estimation algorithm. Abbreviations: PC=phosphatidylcholine, PI=phosphatidylinositol, PE=phosphatidylethanolamine.

approach. Functionally, we observe prominent metabolites from each independent component to be biologically related. The following paragraph briefly describes each of the eight reconstructed independent components with respect to biochemical characteristics of the top-scoring metabolites.

IC_1 primarily contains amino acids and related substances. Among the top-scoring metabolites in this component are amino acids containing functional amine groups, like glutamine, histidine, arginine and carnitine, as well as several aromatic compounds, including tryptophan and phenylalanine. The strongest metabolites in IC_2 are again primarily amino acids. We observe phenylalanine and tryptophan in the top-scoring compound list, and in particular various branched-chain amino acids. Valine, leucine and isoleucine constitute high contributions, but also their direct degradation products 3-methyl-2-oxobutyrate, 4-methyl-2-oxopentanoate and 3-methyl-2-oxovalerate, respectively. Independent component IC_3 exclusively contains long chain fatty acids comprising 12 to 20 carbon atoms among its 15 strongest metabolites. This includes fatty acids with both even numbers of carbon atoms as well as a few odd numbered fatty acids, and various levels of desaturation (i.e. number of double bonds). IC_4 represents a rather heterogeneous set of fatty acid-based lipids. These include short and medium chain fatty acids, hydroxy fatty acids, two polyunsaturated fatty acids (arachidonate and dihomolineolate), and several phospatidylinositols. The fifth independent component IC_5 contains as its strongest entries several metabolites involved in energy homoeostatic processes. This includes phosphate and acetylphosphate, lactate, pyruvate, but also carbohydrates like glucose and mannose. IC_6 contains both signals from amino acids (including glutamine, tryptophan, phenylalanine, isoleucine, valine and proline), and from lipid metabolism including phosphatidylethanolamines and medium chain fatty acids. IC_7 also constitutes a rather mixed component with metabolites from tryptophan metabolism (glycosyltryptophane, kynurenin, 3-indoxylsulfate), nucleotide-related substances (pseudouridine, N1-methyladenosine), carbohydrates (myo-inositol, erythronate, erythritol) and others. Finally, IC_8 primarily represents the phosphatidylcholine (PC) lipid class, particularly lyso-PCs with a single fatty acid residue bound to either the sn-1 or sn-2 position of the glycerol backbone. Fatty acid side chains vary from medium chain saturated 14:0 up to poly-unsaturated fatty acid residues 20:4.

Taken together, these results suggest that each metabolomics profile represents a mixture of statistically independent signals, each of which corresponds to a distinct part in cellular metabolism.

8.4 Systematic analysis and statistical enrichment

Motivated by the findings of our manual investigation, we next asked the question whether this signal can be systematically verified. More specifically, we evaluated whether the reconstructed independent components indeed represent distinct subparts of cellular metabolism. For this purpose, we designed a weighted class enrichment algorithm. Regular hypergeometric enrichment tests like *gene set enrichment analysis* (GSEA) [141] and *metabolite set enrichment analysis* (MSEA) [219] analyze discrete yes/no assignments of each analyzed item (metabolite in our case) to one or more classes. Our approach, in contrast, takes into account the weight of each item in the group (in our case the contribution of each metabolite to each IC) in order to calculate the corresponding enrichment.

For each metabolite, one of the following eight *super-pathway* annotations was provided: 'Lipid', 'Carbohydrate', 'Amino acid', 'Xenobiotics', 'Nucleotide', 'Energy', 'Peptide', 'Cofactors and vitamins'. Furthermore, there are a 61 *sub-pathway* annotations like 'Oxidative phosphorylation', 'Carnitine metabolism' or 'Valine, leucine and isoleucine metabolism'. In the following analysis we first determined whether each independent component significantly enriches metabolites from one of the super-pathways ($p \leq 0.01$). For each enriched super-pathway, we then investigated whether the component also enriches one of the sub-pathways (Table 8.1). Further confirming the manual analysis, we observe strong enrichments for amino acids, lipids and energy metabolism. In particular, independent components separate histidine, branched-chain amino acid (valine, leucine, isoleucine) and tryptophan-related processes in the amino acid super-pathway class. For the lipid class, we observe two mixed components involving various types of fatty acids as well as a third, glycerolipid-centered component. The energy-related component splits into oxidative phosphorylation and central carbon metabolism (glycolysis, gluconeogenesis and pyruvate metabolism).

We compared the weighted enrichment algorithm with hypergeometric enrichment as used in GSEA and MSEA. The weighted approach displays a slightly higher sensitivity for the detection of enriched pathways, but the results of weighted and hypergeometric enrichment are generally comparable (results not shown). Importantly, however, hypergeometric enrichment requires a hard yes/no assignment of metabolites to each component, i.e. whether it can be considered 'present' in the component or not. This introduces an additional cutoff parameter that needs to be defined before the analysis. Weighted enrichment, on the other hand, works parameter-free and directly uses the actual strength of each metabolite in the components.

	Super pathway	**p**	**Sub-pathway**	**p**
IC$_1$	Amino acid	$3.0 \cdot 10^{-7}$	Histidine metabolism	$4.6 \cdot 10^{-3}$
IC$_2$	Amino acid	$< 1.0 \cdot 10^{-7}$	Valine, leucine and isoleucine metabolism	$8.0 \cdot 10^{-7}$
IC$_6$	Amino acid	$4.0 \cdot 10^{-3}$	Valine, leucine and isoleucine metabolism	$3.5 \cdot 10^{-3}$
IC$_7$	Amino acid	$5.4 \cdot 10^{-4}$	Tryptophan metabolism	$4.0 \cdot 10^{-3}$
IC$_3$	Lipid	$< 1.0 \cdot 10^{-7}$	Fatty acid, saturated, even Fatty acid, monoene Fatty acid, monoene, odd Fatty acid, polyene Carnitine metabolism	$2.3 \cdot 10^{-4}$ $4,0 \cdot 10^{-7}$ $4.3 \cdot 10^{-4}$ $6.6 \cdot 10^{-4}$ $7.1 \cdot 10^{-3}$
IC$_4$	Lipid	$3.9 \cdot 10^{-5}$	Fatty acid, saturated, even Fatty acid, saturated, odd Fatty acid, polyene Fatty acid, saturated, monohydroxy	$2.0 \cdot 10^{-3}$ $7.2 \cdot 10^{-5}$ $1.2 \cdot 10^{-4}$ $1.0 \cdot 10^{-3}$
IC$_8$	Lipid	$< 1.0 \cdot 10^{-7}$	Glycerolipid metabolism	$< 1.0 \cdot 10^{-7}$
IC$_5$	Energy Carbohydrate	$2.0 \cdot 10^{-4}$ $2.4 \cdot 10^{-3}$	Oxidative phosphorylation Glycolysis, gluconeogenesis, pyruvate metabolism	$< 1.0 \cdot 10^{-7}$ $1.5 \cdot 10^{-3}$

Table 8.1: Statistical enrichment of metabolic pathways in the independent components. We employed a weighted enrichment test which makes use of the actual contributions of each metabolite in the ICs (see main text). As suggested by our manual investigation, we find strong enrichment for different parts of metabolism, e.g. amino acid pathways, lipid-specific pathways, and energy-related processes. Interestingly, except for a few overlaps, each IC specifically enriches a distinct major pathway.

We furthermore complemented the functional enrichment analysis from an information theoretical point-of-view, by inspecting the information content in each independent component. ICA seeks for maximal non-Gaussianity, a feature commonly measured by the fourth central distribution moment (*kurtosis*). Decreasingly ordered kurtosis values for all eight components are displayed in Figure 8.5. Interestingly, the two components containing the least amount of information, namely IC$_6$ and IC$_3$ are those that displayed a significant overlap in functional enrichment with other components (IC$_2$ and IC$_4$, respectively). This indicates that kurtosis can be used to sort out components containing rather little biological information; an approach that has been employed in previous studies already [209, 210]. On the other hand, components displaying significant, distinct associations with biological processes also contain a high amount of information (e.g. IC$_8$ and IC$_1$). This finding establishes an appealing bridge between the statistical information content in the reconstructed components, and the biological information content encoded therein.

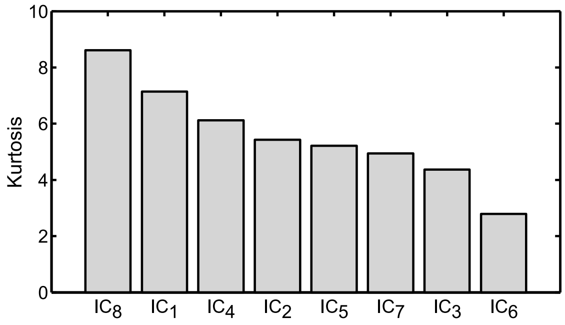


Figure 8.5: Kurtosis as a measure of information content for each independent component. Remarkably, those components with a high information content also tend to display strong functional enrichment of a metabolic pathway.

8.5 Comparison with PCA and k-means clustering

To get an objective view of the quality of our ICA approach, we compared the weighted enrichment results obtained using Bayesian ICA with commonly used data analysis techniques. We ran the enrichment calculations on the results of principal component analysis (PCA) and k-means clustering with the same number of components (or clusters), see Figure 8.6. Furthermore, we introduce the concept of *consistent* and *inconsistent* sub-pathway enrichments. The enrichment of a sub-pathway is considered inconsistent, if the super-pathway this sub-pathway belongs to is not enriched in the same component. For ICA, we detect one inconsistent enrichment of the gamma-glutamyl peptide pathway for IC_2, which enriches the amino acid super-pathway.

PCA yields seven out of eight enriched components, with a total of three distinct enriched super-pathways. For the sub-pathway enrichment, six enrichments can be considered inconsistent since the respective super-pathways are not enriched in the same component. Several components display similar enrichments as independent components from the ICA. Specifically, IC_2/PC_5 as well as IC_6/PC_2 enrich branched-chain amino acids, IC_3/PC_1 as well as IC_4/PC_4 show specific fatty acid pathway enrichments, IC_5/PC_6 enrich the glycolysis pathway, and finally IC_8/PC_3 enrich the glycerolipids. PCA does not detect enrichments of histidine metabolism (IC_1), oxidative phosphorylation (IC_5) and tryptophane metabolism (IC_7). Furthermore, p-values for PCA enrichment are generally higher in comparison to ICA (colors in Figure 8.6), e.g. with three out of seven enriched super-pathways which are only borderline significant. K-means clustering produces a substantial number of enrichments for sub-pathways which are mostly inconsistent. In other words, k-means recovers parts of the metabolism, which however do not belong to the same super-pathway and cannot be considered as specific metabolic signals.

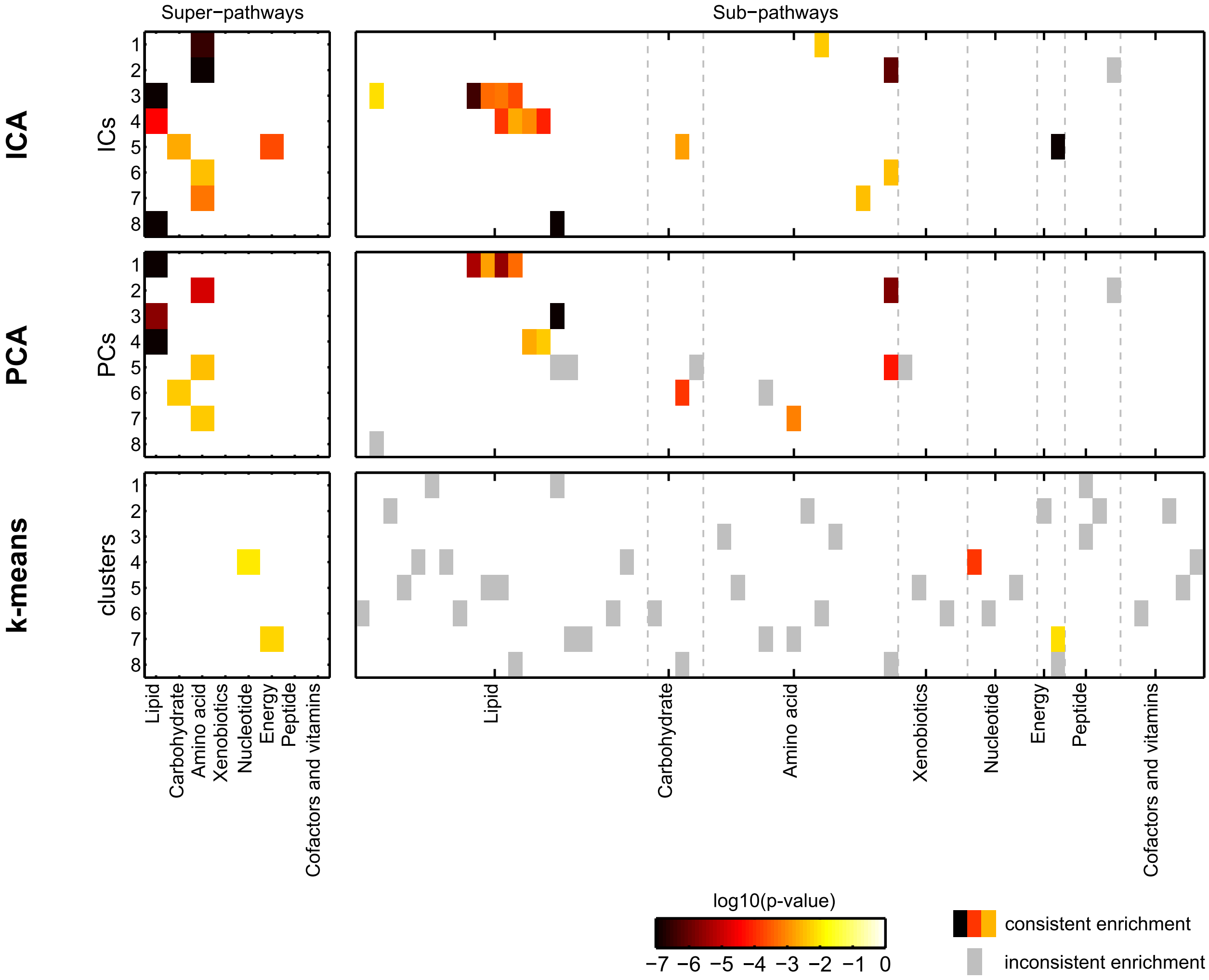


Figure 8.6: Comparison of pathway enrichment for ICA, PCA and k-means clustering. ICA and PCA produce generally comparable results, but ICA appears more sensitive (enriches more super-pathways), more specific (less inconsistent enrichments) and displays lower association p-values. Note that the components are not comparable in order, e.g. IC_1 does not correspond to PC_1.

To further compare ICA with a regular clustering algorithm that supports weighted cluster assignments, we applied fuzzy c-means clustering. The analysis produced no significantly enriched clusters with respect to the super-pathways, and only few enriched sub-pathways. Finally, c-means clustering with subsequent selection of the clusters displaying the highest contribution for each metabolite (see Methods) yields similar results as the k-means approach. Detailed enrichment results of Bayesian ICA, PCA, k-means, and the two variant of c-means clustering are collected in the supplementary material of the original publication.

8.6 Analyzing the mixing matrix A — associations with HDL

Up to this point we have demonstrated that, to a certain extent, metabolomics profiles may be interpreted as a mixture of independent processes from different parts of the metabolic pathways. We next sought to investigate whether the mixing matrix **A** contains biologically interesting information as well. Recall that **A** gives us another 8 variables for each sample (proband in the study cohort) in addition to the metabolite concentrations. These 8 variables encode how strong each IC, i.e. each recovered biological process, contributes to the respective metabolite profile. As can be seen in the clustering displayed in Figure 8.7, the IC weights certainly contain proband-specific information suitable for further analysis. The question now is how to determine whether these weights represent biologically meaningful descriptors. A straightforward approach is to correlate the columns of **A** with other, sample-specific parameters and measurements (Figure 8.1C). One such example is provided in a transcriptomics ICA study by Schachtner et al. [206], where the mixing matrix columns were compared with so-called *design vectors* – which essentially encode the different conditions cells in that particular study were cultured in.

We here chose blood plasma high-density lipoprotein (HDL) levels, which represent a complex quantitative trait influenced by a variety of metabolic and physiological parameters [220]. HDL belongs to the class of lipoproteins, small particles circulating in the blood responsible for the transport of insoluble lipids through the body. We conducted a linear regression analysis of both metabolites and IC strengths against HDL levels, corrected for gender and age effects (Figure 8.8A). Associations with HDL are generally high throughout the dataset, with 88 out of 218 metabolites and 5 out of 8 ICs displaying statistically significant associations ($\alpha = 0.05$ after Bonferroni correction). Two independent components, IC_2 and IC_1, show profound signals with p-values below 10^{-17}. Remarkably, IC_2 even constitutes the strongest association throughout all

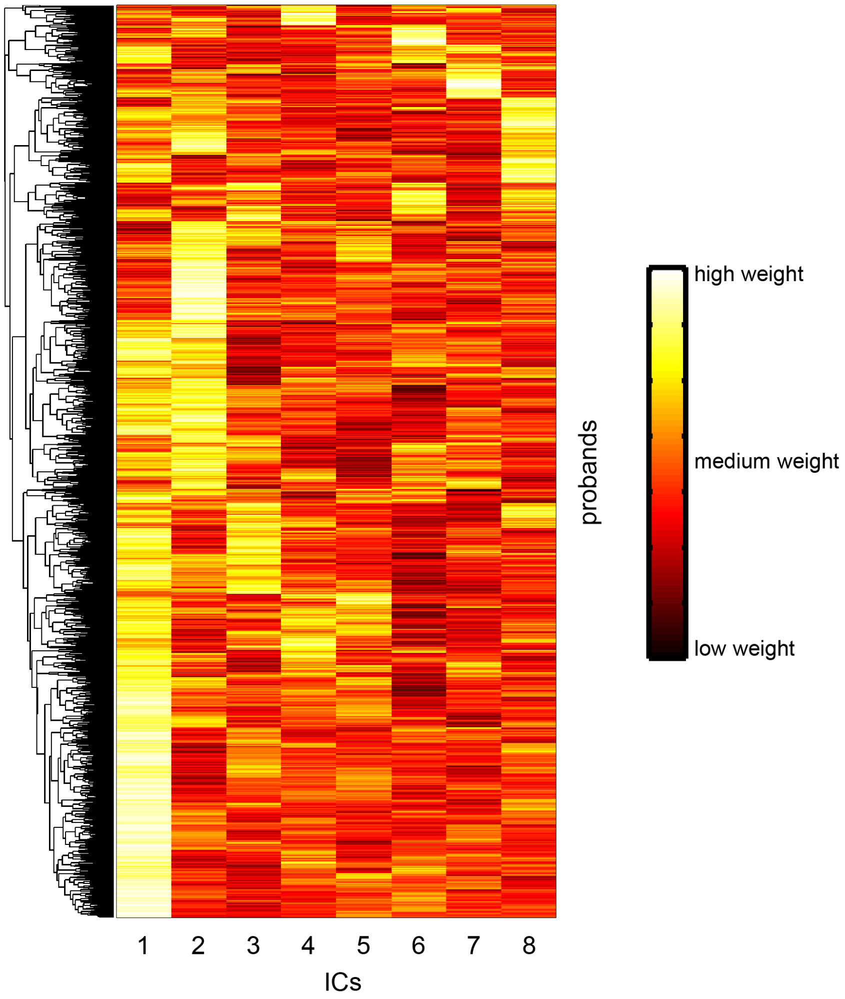
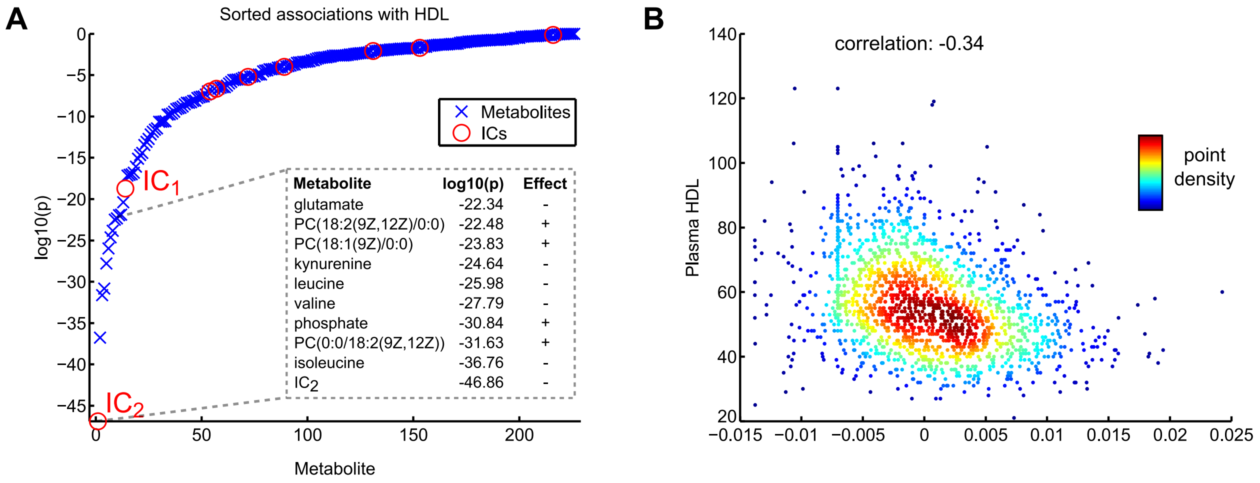


Figure 8.7: The mixing matrix **A**. Rows represent the strengths of each independent component's contribution to the respective proband metabolome. The hierarchical clustering in proband direction demonstrates the presence of clear-cut groups reconstructed from the ICA. Each column in the matrix is then subject to correlation with plasma HDL levels in the next step.



Metabolite	log10(p)	Effect
glutamate	-22.34	-
PC(18:2(9Z,12Z)/0:0)	-22.48	+
PC(18:1(9Z)/0:0)	-23.83	+
kynurenine	-24.64	-
leucine	-25.98	-
valine	-27.79	-
phosphate	-30.84	+
PC(0:0/18:2(9Z,12Z))	-31.63	+
isoleucine	-36.76	-
IC_2	-46.86	-

Figure 8.8: Linear regression of plasma HDL levels on metabolite levels and independent component contributions, corrected for gender and age effects. **A:** The strongest association of all variables is constituted by IC_2, followed by the branched-chain amino acids, other amino acids and several phosphatidylcholines. **B:** Negative correlation between plasma HDL and the contribution strength of IC_2 (which primarily contains contributions from branched-chain amino acids). Note that negative values for the IC occur due to the correction for gender and age.

analyzed variables. As described above, IC_2 primarily contains signatures of the three branched-chain amino acids valine, leucine and isoleucine as well as their respective degradation products.

We detect a *negative* effects on plasma HDL levels for both branched-chain amino acids alone, and for the IC_2 contribution strength (‘Effect’ column in Figure 8.8A, and Figure 8.8B). This means, a stronger contribution of this component, and thus higher values of the involved metabolites, coincides with lower values of HDL. This finding represents a novel connection between branched-chain amino acids and blood plasma HDL levels. For comparison, we performed the HDL comparison with loadings from PCA instead of ICA. The branched-chain amino acid principal component displays a profoundly weaker association with HDL than IC_2 ($p = 3.28 \cdot 10^{-5}$). The strongest association of a principal component with HDL ranks number 20 in the sorted association list.

In order to get an additional comparison with common regression-based approaches, we generated a linear model with multiple metabolite predictor variables. To this extent, we ran a forward feature selection approach based on AIC (Akaike information criterion, see Methods). Interestingly, when ordering the metabolites by their importance for the overall model performance, isoleucine is the only branched-chain amino acid-related metabolite appearing among the top hits (results not shown). This is an effect of high correlations between metabolites:

Once isoleucine is added to the model, the other branched-chain amino acid compounds cannot improve model performance any further. Hence, while such a multipredictor linear regression model might produce a reasonably good description of HDL levels, the interpretation of metabolites with high weights in this model might be misleading.

8.7 Conclusion

In this study, we evaluated a Bayesian independent component analysis (ICA) approach as a tool for the investigation of a population-based metabolomics dataset containing 1764 probands and 218 metabolites. The Bayesian framework provides several advantages over a regular ICA: (1) We can implement distribution priors (a nonnegativity constraint in our case) to construct a biologically meaningful factorization of the data matrix. (2) Since we get distributions of fitted parameters, we obtain information on the estimation certainty for each entry in **S**. (3) Using a Bayesian Information Criterion-based model selection approach, we can automatically determine the number of components to be reconstructed from the data.

We evaluated the source matrix **S** of statistically independent metabolite profiles from a biological point-of-view and demonstrated strong enrichment of distinct metabolic pathways in the reconstructed components. This implies that the human blood metabolome represents a mixture of overlaying, statistically independent signals, each of which can be attributed to a specific set of metabolic pathways. While this concept is quite similar to the idea of *eigengenes* and *eigenmetabolites* [221], our approach extends the standard ICA approach by a Bayesian, noisy framework which allows for the estimation of confidence intervals for the reconstructed values.

The results obtained from the investigation of **S** are in general accordance with our findings from Gaussian graphical models (GGMs) of metabolomics data, as described in the previous chapters. While GGMs only evaluate pairwise associations instead of whole groups as in the ICA approach, the recovery of functionally related metabolites from blood plasma metabolomics samples is similar for both approaches. This fosters the idea of an actual *snapshot* of an organism's metabolism in the blood, rather than mere signatures of transportation and disposal processes in this biofluid.

Correlating the columns of the mixing matrix **A** with plasma HDL levels, we detected a possibly novel association between branched-chain amino acids and HDL blood plasma levels. HDL represents a complex, heterogeneous phenotype which is still poorly understood and associated with a variety of biological processes [222, 223]. The metabolic process encoded by independent component 2 in our study now adds an additional piece of functional information for the interpretation of plasma HDL. Interestingly, both HDL levels and branched-chain amino acids are well-known to be strongly connected with obesity, insulin resistance and diabetes type II. On the one hand, branched-chain amino acid levels are altered as a direct consequence of changed insulin sensitivity, and have been shown to be markers for the prediction of future diabetes type II [224, 225]. Furthermore, leucine is known to directly interact on a cellular level with the insulin signaling cascade [226]. On the other hand, the pathological phenotype is known to lower HDL blood plasma levels, a condition that severely increases the risk for cardiovascular disease [227]. Using cross-sectional metabolomics data from a population cohort, we could now establish the additional association between branched-chain amino acids and HDL, irrespective of a diabetic phenotype. Interestingly, we could recover this association despite the unsupervised approach taken by ICA. In other words, independent component 2 has not been specifically tailored to explain HDL levels, but rather seems to reflect an intrinsic metabolic process around branched-chain amino acids that strongly associates with HDL. The only (biologically motivatable) assumption going into the ICA model is the independence of metabolite profiles to hold throughout all samples in the data.

We systematically compared the ICA results with commonly used multivariate data analysis methods like PCA and k-means clustering. The comparison with PCA was of particular interest here, since it is widely used for metabolomics data and, similar to ICA, also represent a linear mixture model separating the data matrix into a source and a mixing matrix. While PCA produced a series of enriched components with direct IC counterparts, ICA appeared to be more sensitive. Specifically, ICA enrichments were generally stronger in comparison to PCA and detected several pathway enrichments that could not observed for PCA. Moreover, our findings from the HDL analysis could not be reproduced in the PCA approach. These results could be due to the rather arbitrary constraint of orthogonal basis vectors in PCA, which can hardly be biologically motivated. The notion of statistically independent processes acting in the system, as recovered by the ICA, can directly be interpreted in the context of a metabolic system.

Taken together, Bayesian ICA on metabolomics data can be used both to reconstruct meaningful metabolic profiles which underly the measured concentrations, and to detect novel relationships with complex phenotypic traits like plasma HDL levels.

Chapter 9

Summary & Outlook

The field of metabolomics has tremendously advanced in the past few years, with discoveries in epidemiology [228, 229], nutritional challenging [17, 230] and molecular cell biology mechanisms [16, 231]. Metabolite profiles are frequently used for both biomarker discovery of phenotypic states (like a disease state), but also to elucidate general metabolic mechanisms for fundamental research. Understanding the functional relationships between metabolite concentrations and physiological traits, however, remains a challenging task. Especially from a statistical or bioinformatical point-of-view, dealing with high-dimensional data matrices produced by metabolomics measurements holds numerous problems and pitfalls. In this thesis, we laid a particular focus on exploring statistical metabolite dependencies, which arise due to naturally occurring biological variation in large datasets. For the first time in a systematic fashion, we applied Gaussian graphical models, which estimate conditional dependencies between variables, to metabolomics data in order to tackle the problem of indirect effects and spurious correlations. Furthermore, an independent component analysis model was applied to metabolomics data, which is capable of detecting higher-order statistical relationships beyond pairwise covariance. In the following, we will summarize the scientific contributions developed in this thesis, and discuss possible extensions and future directions.

Scientific achievements

The following novel scientific contributions and insights were obtained throughout the work of this thesis:

- Gaussian graphical models are generally capable of reconstructing the structure of computer-simulated reaction systems when applying a log-normal noise model to the reaction parameters (Chapter 4). We furthermore demonstrated a few exceptions where reconstruction was impaired, like feedback mechanisms, which need to be kept in mind when working with real data. In general, the forward simulation of reaction systems is a valuable tool to determine beforehand what we can expect to recover using GGMs, and what might remain hidden.

- A statistically significant fraction of metabolite pairs which share an edge in the Gaussian graphical model are also directly connected in the metabolic pathway (Chapter 5). This indicates that the simple correlation structures between metabolites carry a strong, systematic signal of the underlying pathways, which are detectable in high-throughput data when accounting for indirect effects. Furthermore, those metabolite pairs which we considered 'wrong' in our analysis might be worthwhile for further analysis. False positives, i.e. pairs with significant partial correlations but no known pathway connection, might represent previously unknown pathway reactions or specific co-regulatory mechanisms. False negatives, i.e. pairs with insignificant partial correlations but a known biochemical interaction, might point towards the specificity of a metabolic reaction. The reaction is present in the organism, but no signal is detectable in the blood system.

- GGMs can be exploited to derive functional classifications of unidentified metabolites from untargeted metabolomics experiments (Chapter 6). In combination with large-scale genotyping data and pathway database information, systematic classifications can readily be obtained. For some cases, the metabolic context provided by this integration approach is even precise enough to derive a testable chemical identity prediction. Several of our newly assigned metabolite identities shed new light on existing biomarker studies on liver detoxification, hypertension and insulin resistance.

- In addition to the unknown classification analysis in Chapter 6, we detected seven new loci of metabolic individuality: SLC22A2, COMT, CYP3A5, CYP2C18, GBA3, UGT3A1, and rs12413935 (for the last locus, no known gene has been annotated yet). To the best of our knowledge, no previous studies associated variations in these SNP loci with changes in blood metabolite concentrations.

- Applying the GGM methodology to various biological questions in Chapter 7, we were able to generate specific insights for the respective biological systems under investigation: (1) Gender-specific metabolome differences might originate from a particular change in stearic acid (C18:0) metabolism. (2) Investigating the association between fat-free body mass and metabolome changes, we detected a specific signal for the phosphatidylcholine C38:3, which does not appear to propagate through the metabolic network. Furthermore, we detected a coordinated downregulation of phosphatidylcholines with very long fatty acid side chains. (3) A differential GGM approach elucidated specific metabolic changes in a glioblastoma cell line upon chemotherapeutic treatment and gene therapy. Specifically, the partial correlations between oxidized sulfatides, gangliosides and phosphatidylinositols were affected by the treatment. (4) GGMs can be used to define biologically meaningful metabolite groups in the sense of a graph clustering. In the simplest approach, a high partial correlation cutoff will yield a graph with multiple connected components, which then represent groups of biochemically related metabolites. In summary, all of these examples represent studies where the systematic metabolic picture provided by a GGM aided the biological interpretation of results.

- Investigating higher-order statistical associations beyond covariance in an independent component analysis, we detected profound pathway footprints for entire groups of metabolites in the data (Chapter 8). Furthermore, these pathway signatures displayed a stronger correlation with blood HDL levels than any metabolite alone. ICA can thus be seen as a promising alternative to GGM analysis, which investigates group-wise signals in addition to solely pairwise associations as estimated by correlation measures.

- In contrast to most previous systems biological frameworks for metabolism, our modeling and network inference approach specifically works with metabolomics data. For example, constraint-based modeling [38], which was tremendously successful for almost 20 years, has never been properly adapted for metabolomics data. In combination with the integrated pathway models we derived in Chapter 5 and especially Chapter 6, we developed a generic approach for the functional analysis of cross-sectional metabolomics data.

- Methodologically, we introduced several approaches to verify the stability of GGM estimation, both using varying sample sizes and sample bootstrapping (Chapter 5). Furthermore, we introduced a preliminary differential GGM algorithm, which detects group-specific changes in metabolite associations (Chapter 7).

- A particularly important feature of a *data-driven* metabolic reconstruction approach is the conceptual independence from prior knowledge, which is still far from complete for human metabolism. For example, numerous measured substances cannot be found in public reaction databases and subsequently not be analyzed in a knowledge-driven fashion. Moreover, we have seen in Chapter 6 that unknown metabolites constitute a plethora of both biochemical and genetic interactions. With the GGM approach, these unknowns may be kept in the dataset and subjected to follow-up analyses like the network-based biomarker discovery approaches from Chapter 7.1.

- All results could be obtained from metabolomics data of human blood in a large population cohort. This provides important insights into the nature of metabolites that can be found in the blood. Inspecting the GGMs shown in Figures 5.1 and 6.3, the majority of all metabolites is connected to other compounds in the network. Furthermore, throughout the work of this thesis, we found only very few examples of GGM edges that appear to be biologically unreasonable. Our results thus suggest that metabolites present in the blood are not only products of unspecific leaking from larger metabolically active organs into the vascular system, but also carry a full footprint of the metabolic pathways. Where these signal actually originate from, i.e. liver, muscle or other tissues, is still to be determined.

Taken together, human blood metabolomics data contain strong footprints of biochemical pathways, which can be reconstructed using statistical methods like GGMs and ICA. Furthermore, reconstructed metabolic pathways can be used to address biological questions like group-specific metabolome differences on a systematic level.

Since our metabolomics GGMs were estimated from a very large number of samples, they can be used as a ‘ground truth’ for other metabolomics projects with smaller sample sizes. If we assumed the metabolite-metabolite interactions in fasting state to be conserved throughout all humans, we could use the KORA population GGM to functionally analyze statistical results from a different study cohort (e.g. a challenging study with only few participants). The validity of this conservation assumption will soon be evaluated on independent population cohorts measured using the same metabolomics platforms (e.g. the TwinsUK study [232]).

Extensions and future directions

There are a variety of possible extensions to both the GGM calculation as well as the evaluation methods. These will be discussed in the following.

First, we will include further layers of molecular information in addition to mass-flow metabolic reaction networks. This primarily includes gene regulatory processes and the corresponding proteomics and transcriptomics data. Second, the simulated reaction systems should also be extended by a regulatory layer, again to check what we can expect from a GGM reconstruction and what might not be revealed. In Chapter 6 we also included SNP genotyping data in a rather pragmatic and functionally-oriented fashion (through a GWAS). Such variation might also be subject to more specific modeling, however only for genetic variants where the causal effect on the respective gene products is known.

Methodologically, there are several ways to improve and extend the GGM estimation procedure. First, an important issue is the presence of outliers in the data, which might substantially falsify correlation estimators. In principle, there are two possible ways of dealing with data outliers. On the one hand, one might attempt to detect outliers, e.g. using the coefficient of variation, and subsequently filter them out from the data as a preprocessing step before the actual analysis. Examples for this simple approach can be found in virtually any present metabolomics study (e.g. [18, 29]). A more involved approach is to incorporate robust estimation into the GGM calculation process. An example can be found in Miyamura and Kano [90], who introduced a robust maximum likelihood method of covariance estimation. Furthermore, rank-based correlation approaches like Spearman correlation [72] could be used. This approach has the additional advantage of circumventing the need for Gaussianity of the measured compounds and linearity of associations. Spearman correlations will detect arbitrary monotonic relationships. Any method correcting for data outliers could then simply replace the standard GGM estimation procedure in the data analysis workflow.

In this thesis, we mainly calculated GGMs by simple inversion of the covariance matrix. For most metabolomics studies today, however, we cannot expect the number of samples to be larger than the number of measured metabolites, as for the KORA data. Therefore, the 'small n, large p' approaches introduced in Chapter 3.5 will play an important role in future applications of the metabolomics GGM approach. Methods that allow for an inversion of the covariance matrix despite small sample sizes are suitable tools for the quick generation of GGMs for any type of dataset. Moreover, methods that directly work with the Markov properties from graphical modeling theory might be even more promising. For example, we have seen that two uncorrelated variables might become strongly negatively correlated when conditioning against further variables. Specialized algorithms specifically reconstructing the graph neighborhood (the *Markov boundary*) of a given variable might be more suitable for such scenarios [55, 79]. Furthermore, such correlation scenarios can be used to introduce directionality into the GGM calculation, see below. A particularly important pre-analysis for the application of 'small n, large p' approaches will be a systematic assessment of their reconstruction capabilities (similar to Chapters 4 and 5) as well as a direct comparison between the different GGM approaches.

GGM per se only reconstruct *undirected* effects from the data. Real biochemical networks, however, often contain a directionality of the effect (e.g. in an irreversible reaction). In future projects, we will employ algorithms that recover directed edges in a network. Our toy model framework (Chapter 4) is ideally fit to generate evaluation data for such algorithms, since we can specifically control the directionality of information flow in the underlying network. A popular example for a directed graphical model used in biomedical research are Bayesian networks [74], which encode a (simplified) factorization of the joint probability of all variables. For partial correlations, several approaches have been proposed to include directionality of the edges. For example, Freudenberg et al. [59] used the concept of *d-separation* to rule out directed connectivities which are not supported by the data. Again, this approach elucidates the conditional and marginal independence relations between variables. Opgen-Rhein and Strimmer [233] focused on the connection between linear regression models, partial correlation and partial variances (the variance left after regression). Briefly, if two variable are (undirectedly) connected in the GGM, but the variables show a significantly different reduction in variance due to the regression against all other variables, then the edge points from the variable with the higher partial variance to the variable with the lower partial variance. The authors demonstrate that in such a case the regression coefficient (after scale normalization) will be asymmetric, i.e. the mutual effect of the variable onto each other is not equal. In a third study, Yuan et al. [234] investigated the change of partial correlations of a variable X_i when including or not including a specific effector variable X_j. The authors argue that if the partial correlations with all other variables change significantly, the effector variable X_j does have a directed influence on the prediction of X_i. Importantly, however, fitting a model with unidirectional influences must still not be confused with causality (although many authors claim so). The direction of an edge in the model tells us that the dataset can be fitted best using this edge in the model; there is no guarantee that this direction also holds true in biological reality.

Furthermore, the inclusion of higher-order interactions into the graphical modeling context might be extended in future projects. In Chapter 7.4 we have introduced a variant of metabolite-metabolite networks that include genotyping data. The ternary edges induced by the ratio of two metabolites with a genetic locus provide more biological information than interactions between metabolites alone. From a methodological point-of-view, higher-order statistical dependencies should be included into our models. The independent component analysis in Chapter 8 demonstrated a first attempt to introduce such dependencies in the analysis of metabolomics data. Moreover, there are specific graphical modeling approach that encode information beyond the conditional independence between variables. For example, 'Vines' [235] include specific conditional dependence information (e.g. non-zero partial correlations) between two variables given a set of conditioning variables.

A particular problem that should be addressed in future applications of correlation coefficients is the reasonability of significance cutoffs. Throughout all analyses, we followed a straightforward, statistically sound approach: We used established statistical models which test for non-zero correlation coefficients, corrected for multiple hypothesis testing and applied a standard significance level α (of 0.05, 0.01 or 0.001). However, it is doubtful whether checking for a non-zero correlation coefficient always represents the biologically relevant question. For example, assume we obtained a correlation of 0.1 between two metabolites in a sample of n=1000 study participants. This correlation yields a p-value of p=$1.53 \cdot 10^{-3}$. When increasing the number of samples to n=2000 participants, the correlation coefficient will most likely be the same (we get a reasonably good estimate with 1000 samples already), but the p-value drops to p=$7.33 \cdot 10^{-6}$. The situation also holds true for partial correlation coefficients, where some correlation might remain due to dependencies that cannot be entirely partialized out. Whether or not an edge will be present in the GGM is thus not only determined by the actual association between compounds, but also by the number of samples in the dataset. This obviously should not be the case and must be considered in future analyses. One approach to solve this issue would be to apply a constant correlation cutoff (cf. Chapter 7.1). Although the problem is then shifted to determining a proper value for this correlation cutoff, this approach would not dependent on the actual sample size for stably estimated correlation coefficients.

The above-mentioned extensions concerned graphical model estimation as such. There are numerous possibilities to extend the biological evaluation of our statistical results. As discussed above, the *false positive* GGM edges identified in Chapter 5 cannot be considered *false* in biological sense, but rather represent findings that are not in accordance with current pathway knowledge. The list of GGM edges that have no evidence in the public reaction databases represent suitable candidates for subsequent deeper analyses and possibly experimental testing.

Moreover, several approaches to improve the analysis of a certain phenotype in the light of GGMs can be imagined. (1) Our differential analysis presented in Chapter 7.2 merely represented an ad-hoc approach to find changed partial correlations between different conditions. A recent statistics master thesis conducted in our group investigated hypothesis testing for differential correlation values both in computer-simulated systems and metabolomics data. This approach will be used to soundly assess the significance of changed correlations between two groups in the data, rather than comparing ‘present’ and ‘absent’ edges in two distinct GGMs. (2) The analysis of colored GGMs (‘effect networks’, Chapter 7.1) can be complemented by the algorithmic detection of regions in the graph with a strong signal. Computationally, this requires algorithms similar to clustering, which find regions of enriched signals in a weighted graph. Rather than manually finding highly colored regions by visual inspection, we could then recover regions with strong signals automatically. Such approaches will be particularly useful for analyses where only very few significant metabolite associations are detected. (3) In addition to projecting statistical results to the data-driven GGMs, real metabolic pathways from public databases should also be used. An example for this idea was published by Chuang et al. [236], who projected the results of a metastatic/non-metastatic differential proteomics analysis to publically available protein-protein interaction networks. This data combination then allowed the identification of specific, metastasis-related protein subnetworks. Interestingly, the concept of *differential network biology*, i.e. differential calculation of networks or coloring of networks with statistical results, has recently been announced to become a major tool for future 'omics' analyses (Ideker and Krogan [161]).

Addressing the above-mentioned issues will be a challenge for future GGM-based projects, but will allow us to gain further insights into the biochemical interplay of metabolites.

Conclusion

In this thesis, we presented Gaussian graphical models as a valuable tool for the recovery of biochemical reactions from high-throughput targeted metabolomics data. Using techniques from mathematical modeling and bioinformatics, we could proof the validity of the approach by computer simulations and systematic comparisons against public databases. Furthermore, several approaches of how to investigate specific phenotypic groups in the study samples have been proposed. Independent component analysis (ICA) was introduced as an extension of correlation-based approaches, which provides particular insights into the interplay of metabolite groups. Concluding, we suggest to use GGMs and ICA as standard tools of investigation in future metabolomics studies, utilizing the upcoming wealth of metabolic profiling data to form a more comprehensive picture of metabolic pathways.

Bibliography

[1] Poretsky, L. *Principles of Diabetes Mellitus*. Springer, 2010. ISBN 9780387098401.

[2] Buchner, E. Alkoholische Gärung ohne Hefezellen (Vorläufige Mitteilung). *Berichte der Deutschen Chemischen Gesellschaft*, 30:117–124, 1897.

[3] Gonzalez, J. and Willis, M.S. Ivar Asbjörn Følling Discovered Phenylketonuria (PKU). *Lab Medicine*, 41(2):118–119, 2010.

[4] Blow, N. Metabolomics: Biochemistry's new look. *Nature*, 455(7213):697–700, 2008.

[5] Kaddurah-Daouk, R., Kristal, B.S., and Weinshilboum, R.M. Metabolomics: a global biochemical approach to drug response and disease. *Annu Rev Pharmacol Toxicol*, 48:653–683, 2008.

[6] Tweeddale, H., Notley-McRobb, L., and Ferenci, T. Effect of slow growth on metabolism of Escherichia coli, as revealed by global metabolite pool (”metabolome”) analysis. *J Bacteriol*, 180(19):5109–5116, 1998.

[7] Oliver, S.G., Winson, M.K., Kell, D.B., and Baganz, F. Systematic functional analysis of the yeast genome. *Trends in Biotechnology*, 16(9):373 – 378, 1998. ISSN 0167-7799.

[8] Griffin, J.L. The Cinderella story of metabolic profiling: does metabolomics get to go to the functional genomics ball? *Philos Trans R Soc Lond B Biol Sci*, 361(1465):147–161, 2006.

[9] Ludwig, C. and Viant, M.R. Two-dimensional J-resolved NMR spectroscopy: review of a key methodology in the metabolomics toolbox. *Phytochem Anal*, 21(1):22–32, 2010.

[10] Roux, A., Lison, D., Junot, C., and Heilier, J.F. Applications of liquid chromatography coupled to mass spectrometry-based metabolomics in clinical chemistry and toxicology: A review. *Clin Biochem*, 44(1):119–135, 2011.

[11] Fiehn, O. Metabolomics–the link between genotypes and phenotypes. *Plant Mol Biol*, 48(1-2):155–171, 2002.

[12] Keurentjes, J.J.B., Fu, J., de Vos, C.H.R., Lommen, A., Hall, R.D., Bino, R.J., van der Plas, L.H.W., Jansen, R.C., Vreugdenhil, D., and Koornneef, M. The genetics of plant metabolism. *Nat Genet*, 38(7):842–849, 2006.

[13] Morgenthal, K., Weckwerth, W., and Steuer, R. Metabolomic networks in plants: Transitions from pattern recognition to biological interpretation. *Biosystems*, 83(2-3):108–117, 2006.

[14] Scherling, C., Ulrich, K., Ewald, D., and Weckwerth, W. A metabolic signature of the beneficial interaction of the endophyte paenibacillus sp. isolate and in vitro-grown poplar plants revealed by metabolomics. *Mol Plant Microbe Interact*, 22(8):1032–1037, 2009.

[15] Cuadros-Inostroza, A., Giavalisco, P., Hummel, J., Eckardt, A., Willmitzer, L., and Pena-Cortes, H. Discrimination of Wine Attributes by Metabolome Analysis. *Analytical Chemistry*, 82(9):3573–3580, 2010.

[16] Fendt, S.M., Buescher, J.M., Rudroff, F., Picotti, P., Zamboni, N., and Sauer, U. Tradeoff between enzyme and metabolite efficiency maintains metabolic homeostasis upon perturbations in enzyme capacity. *Mol Syst Biol*, 6:356, 2010.

[17] Fav, G., Beckmann, M.E., Draper, J.H., and Mathers, J.C. Measurement of dietary exposure: a challenging problem which may be overcome thanks to metabolomics? *Genes Nutr*, 4(2):135–141, 2009.

[18] Krug, S., Kastenmüller, G., Stückler, F., Rist, M.J., Skurk, T., Sailer, M., Raffler, J., Römisch-Margl, W., Adamski, J., Prehn, C., Frank, T., Engel, K.H., Hofmann, T., Luy, B., Zimmermann, R., Moritz, F., Schmitt-Kopplin, P., Krumsiek, J., Kremer, W., Huber, F., Oeh, U., Theis, F.J., Szymczak, W., Hauner, H., Suhre, K., and Daniel, H. The dynamic range of the human metabolome revealed by challenges. *FASEB Journal*, 2012.

[19] Konttinen, H., Männistö, S., Sarlio-Lähteenkorva, S., Silventoinen, K., and Haukkala, A. Emotional eating, depressive symptoms and self-reported food consumption. A population-based study. *Appetite*, 54(3):473–479, 2010.

[20] Zhai, G., Wang-Sattler, R., Hart, D.J., Arden, N.K., Hakim, A.J., Illig, T., and Spector, T.D. Serum branched-chain amino acid to histidine ratio: a novel metabolomic biomarker of knee osteoarthritis. *Ann Rheum Dis*, 69(6):1227–1231, 2010.

[21] Fiehn, O., Garvey, W.T., Newman, J.W., Lok, K.H., Hoppel, C.L., and Adams, S.H. Plasma metabolomic profiles reflective of glucose homeostasis in non-diabetic and type 2 diabetic obese African-American women. *PLoS One*, 5(12):e15234, 2010.

[22] Gall, W.E., Beebe, K., Lawton, K.A., Adam, K.P., Mitchell, M.W., Nakhle, P.J., Ryals, J.A., Milburn, M.V., Nannipieri, M., Camastra, S., Natali, A., Ferrannini, E., and Group, R.I.S.C.S. alpha-hydroxybutyrate is an early biomarker of insulin resistance and glucose intolerance in a nondiabetic population. *PLoS One*, 5(5):e10883, 2010.

[23] Wang-Sattler, R., Yu, Y., Mittelstrass, K., Lattka, E., Altmaier, E., Gieger, C., Ladwig, K.H., Dahmen, N., Weinberger, K.M., Hao, P., Liu, L., Li, Y., Wichmann, H.E., Adamski, J., Suhre, K., and Illig, T. Metabolic profiling reveals distinct variations linked to nicotine consumption in humans–first results from the KORA study. *PLoS One*, 3(12):e3863, 2008.

[24] Altmaier, E., Kastenmüller, G., Römisch-Margl, W., Thorand, B., Weinberger, K.M., Adamski, J., Illig, T., Döring, A., and Suhre, K. Variation in the human lipidome associated with coffee consumption as revealed by quantitative targeted metabolomics. *Mol Nutr Food Res*, 53(11):1357–1365, 2009.

[25] Mittelstrass, K., Ried, J.S., Yu, Z., Krumsiek, J., Gieger, C., Prehn, C., Roemisch-Margl, W., Polonikov, A., Peters, A., Theis, F.J., Meitinger, T., Kronenberg, F., Weidinger, S., Wichmann, H.E., Suhre, K., Wang-Sattler, R., Adamski, J., and Illig, T. Discovery of Sexual Dimorphisms in Metabolic and Genetic Biomarkers. *PLoS Genetics*, 7(8):e1002215, 2011.

[26] Gieger, C., Geistlinger, L., Altmaier, E., de Angelis, M.H., Kronenberg, F., Meitinger, T., Mewes, H.W., Wichmann, H.E., Weinberger, K.M., Adamski, J., Illig, T., and Suhre, K. Genetics meets metabolomics: a genome-wide association study of metabolite profiles in human serum. *PLoS Genet*, 4(11):e1000282, 2008.

[27] Illig, T., Gieger, C., Zhai, G., Römisch-Margl, W., Wang-Sattler, R., Prehn, C., Altmaier, E., Kastenmüller, G., Kato, B.S., Mewes, H.W., Meitinger, T., de Angelis, M.H., Kronenberg, F., Soranzo, N., Wichmann, H.E., Spector, T.D., Adamski, J., and Suhre, K. A genome-wide perspective of genetic variation in human metabolism. *Nat Genet*, 42(2):137–141, 2010.

[28] Suhre, K., Wallaschofski, H., Raffler, J., Friedrich, N., Haring, R., Michael, K., Wasner, C., Krebs, A., Kronenberg, F., Chang, D., Meisinger, C., Wichmann, H.E., Hoffmann, W., Völzke, H., Völker, U., Teumer, A., Biffar, R., Kocher, T., Felix, S.B., Illig, T., Kroemer,

H.K., Gieger, C., Römisch-Margl, W., and Nauck, M. A genome-wide association study of metabolic traits in human urine. *Nat Genet*, 43(6):565–569, 2011.

[29] Suhre, K., Shin, S.Y., Petersen, A.K., Mohney, R.P., Meredith, D., Wägele, B., Altmaier, E., CARDIoGRAM, Deloukas, P., Erdmann, J., Grundberg, E., Hammond, C.J., de Angelis, M.H., Kastenmüller, G., Köttgen, A., Kronenberg, F., Mangino, M., Meisinger, C., Meitinger, T., Mewes, H.W., Milburn, M.V., Prehn, C., Raffler, J., Ried, J.S., Römisch-Margl, W., Samani, N.J., Small, K.S., Wichmann, H.E., Zhai, G., Illig, T., Spector, T.D., Adamski, J., Soranzo, N., and Gieger, C. Human metabolic individuality in biomedical and pharmaceutical research. *Nature*, 477(7362):54–60, 2011.

[30] Hirschhorn, J.N. and Daly, M.J. Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet*, 6(2):95–108, 2005.

[31] Morton, N.E. Significance levels in complex inheritance. *Am J Hum Genet*, 62(3):690–697, 1998.

[32] 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–1073, 2010.

[33] Holle, R., Happich, M., Löwel, H., Wichmann, H.E., and MONICA/KORA Study Group. KORA–a research platform for population based health research. *Gesundheitswesen*, 67 Suppl 1:S19–S25, 2005.

[34] Kacser, H. and Burns, J.A. The control of flux. *Symposia of the Society for Experimental Biology*, 27:65–104, 1973. ISSN 0081-1386.

[35] Cascante, M., Boros, L.G., Comin-Anduix, B., de Atauri, P., Centelles, J.J., and Lee, P.W.N. Metabolic control analysis in drug discovery and disease. *Nat Biotechnol*, 20(3):243–249, 2002.

[36] Rees, T. and Hill, S. Metabolic control analysis of plant metabolism. *Plant, Cell & Environment*, 17(5):587–599, 1994. ISSN 1365-3040.

[37] Fell, D.A. Increasing the flux in metabolic pathways: A metabolic control analysis perspective. *Biotechnology and Bioengineering*, 58(2-3):121–124, 1998. ISSN 1097-0290.

[38] Varma, A. and Palsson, B.O. Metabolic Capabilities of Escherichia coli: I. Synthesis of Biosynthetic Precursors and Cofactors. *Journal of Theoretical Biology*, 165(4):477–502, 1993.

[39] Papin, J.A., Price, N.D., Wiback, S.J., Fell, D.A., and Palsson, B.O. Metabolic pathways in the post-genome era. *Trends Biochem Sci*, 28(5):250–258, 2003.

[40] Gupta, S., Maurya, M.R., Merrill, A.H., Glass, C.K., and Subramaniam, S. Integration of lipidomics and transcriptomics data towards a systems biology model of sphingolipid metabolism. *BMC Syst Biol*, 5:26, 2011.

[41] Hirai, M.Y., Yano, M., Goodenowe, D.B., Kanaya, S., Kimura, T., Awazuhara, M., Arita, M., Fujiwara, T., and Saito, K. Integration of transcriptomics and metabolomics for understanding of global responses to nutritional stresses in Arabidopsis thaliana. *Proc Natl Acad Sci U S A*, 101(27):10205–10210, 2004.

[42] Xiao, X., Dawson, N., Macintyre, L., Morris, B.J., Pratt, J.A., Watson, D.G., and Higham, D.J. Exploring metabolic pathway disruption in the subchronic phencyclidine model of schizophrenia with the Generalized Singular Value Decomposition. *BMC Syst Biol*, 5:72, 2011.

[43] Kanehisa, M. and Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*, 28(1):27–30, 2000.

[44] Szklarczyk, D., Franceschini, A., Kuhn, M., Simonovic, M., Roth, A., Minguez, P., Doerks, T., Stark, M., Muller, J., Bork, P., Jensen, L.J., and von Mering, C. The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res*, 39(Database issue):D561–D568, 2011.

[45] Kamburov, A., Pentchev, K., Galicka, H., Wierling, C., Lehrach, H., and Herwig, R. ConsensusPathDB: toward a more complete picture of cell biology. *Nucleic Acids Res*, 39(Database issue):D712–D717, 2011.

[46] Kauffman, K.J., Pajerowski, J.D., Jamshidi, N., Palsson, B.O., and Edwards, J.S. Description and analysis of metabolic connectivity and dynamics in the human red blood cell. *Biophys J*, 83(2):646–662, 2002.

[47] Gille, C., Bölling, C., Hoppe, A., Bulik, S., Hoffmann, S., Hübner, K., Karlstädt, A., Ganeshan, R., König, M., Rother, K., Weidlich, M., Behre, J., and Holzhütter, H.G. HepatoNet1: a comprehensive metabolic reconstruction of the human hepatocyte for the analysis of liver physiology. *Mol Syst Biol*, 6:411, 2010.

[48] Steuer, R., Kurths, J., Fiehn, O., and Weckwerth, W. Observing and interpreting correlations in metabolomic networks. *Bioinformatics*, 19(8):1019–1026, 2003.

[49] Camacho, D., de la Fuente, A., and Mendes, P. The origin of correlations in metabolomics data. *Metabolomics*, 1(1):53–63, 2005.

[50] Krumsiek, J., Stückler, F., Kastenmüller, G., and Theis, F.J. *Systems Biology meets Metabolism*. Springer, 2012.

[51] Schäfer, J. and Strimmer, K. Learning Large-Scale Graphical Gaussian Models from Genomic Data. volume 776, pages 263–276. AIP, 2005.

[52] Lee, J.M., Lee, J.M., Gianchandani, E.P., Eddy, J.A., and Papin, J.A. Dynamic analysis of integrated signaling, metabolic, and regulatory networks. *PLoS Comput Biol*, 4(5):e1000086, 2008.

[53] Rothman, K., Greenland, S., and Lash, T. *Modern Epidemiology*. Wolters Kluwer Health/Lippincott Williams & Wilkins, 2008. ISBN 9780781755641.

[54] Burns, W.C. Spurious Correlations [online paper]. *http://www.burns.com/wcbspurcorl.htm*, 1996.

[55] de la Fuente, A., Bing, N., Hoeschele, I., and Mendes, P. Discovery of meaningful associations in genomic data using partial correlation coefficients. *Bioinformatics*, 20(18):3565–3574, 2004.

[56] Magwene, P.M. and Kim, J. Estimating genomic coexpression networks using first-order conditional independence. *Genome Biol*, 5(12):R100, 2004.

[57] Schäfer, J. and Strimmer, K. An empirical Bayes approach to inferring large-scale gene association networks. *Bioinformatics*, 21(6):754–764, 2005.

[58] Wille, A., Zimmermann, P., Vranov, E., Fürholz, A., Laule, O., Bleuler, S., Hennig, L., Prelic, A., von Rohr, P., Thiele, L., Zitzler, E., Gruissem, W., and Bühlmann, P. Sparse graphical Gaussian modeling of the isoprenoid gene network in Arabidopsis thaliana. *Genome Biol*, 5(11):R92, 2004.

[59] Freudenberg, J., Wang, M., Yang, Y., and Li, W. Partial correlation analysis indicates causal relationships between GC-content, exon density and recombination rate in the human genome. *BMC Bioinformatics*, 10 Suppl 1:S66, 2009.

[60] Ball, P. Tangled relationships unpicked. *Nature News (online)*, 2011.

[61] Denollet, J. and Conraads, V.M. Type D personality and vulnerability to adverse outcomes in heart disease. *Cleve Clin J Med*, 78 Suppl 1:S13–S19, 2011.

[62] van Buuren, S. and Groothuis-Oudshoorn, K. MICE: Multivariate Imputation by Chained Equations in R. *Journal of statistical software*, in press:1–68, 2010.

[63] Grimmett, G. and Stirzaker, D. *Probability and Random Processes*. Oxford University Press, Oxford, 3rd edition, 2001.

[64] Pearson, K. Regression, heredity, and panmixia. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 187:253–318, 1896.

[65] Muirhead, R.J. *Aspects of Multivariate Statistical Theory*. Wiley-Interscience, 2005.

[66] Fahrmeir, L., Künstler, R., Pigeot, I., and Tutz, G. *Statistik*. Springer-Lehrbuch Series. Springer, 5. edition, 2004. ISBN 9783540212324.

[67] Melnick, E. and Tenenbein, A. *Misspecifications of the normal distribution*. Working paper series. New York University, Graduate School of Business Administration, 1978.

[68] Fisher, R.A. Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika*, 10(4):507–521, 1915.

[69] Weisberg, S. *Applied linear regression*. Wiley series in probability and statistics. Wiley-Interscience, 2005. ISBN 9780471663799.

[70] Lauritzen, S.L. *Graphical Models (Oxford Statistical Science Series)*. Oxford University Press, USA, 1996.

[71] Dempster, A. Covariance selection. *Biometrics*, 28:157–75, 1972.

[72] Spearman, C. The Proof and Measurement of Association Between Two Things. *American Journal of Psychology*, 15:88–103, 1904.

[73] Whittaker, J. *Graphical models in applied multivariate statistics*. Wiley series in probability and mathematical statistics: Probability and mathematical statistics. Wiley, 1990. ISBN 9780471917502.

[74] Neapolitan, R. *Learning Bayesian networks*. Prentice Hall series in artificial intelligence. Pearson Prentice Hall, 2004. ISBN 9780130125347.

[75] Bruex, A., Kainkaryam, R.M., Wieckowski, Y., Kang, Y.H., Bernhardt, C., Xia, Y., Zheng, X., Wang, J.Y., Lee, M.M., Benfey, P., Woolf, P.J., and Schiefelbein, J. A gene regulatory network for root epidermis cell differentiation in Arabidopsis. *PLoS Genet*, 8(1):e1002446, 2012.

[76] Kindermann, R., Snell, J., and Society, A.M. *Markov random fields and their applications*. Contemporary mathematics. American Mathematical Society, 1980. ISBN 9780821850015.

[77] Bondy, J. and Murty, U. *Graph theory*. Graduate texts in mathematics. Springer, 2007. ISBN 9781846289699.

[78] Chandrasekaran, V., Johnson, J., and Willsky, A. Estimation in Gaussian Graphical Models Using Tractable Subgraphs: A Walk-Sum Analysis. *Signal Processing, IEEE Transactions on*, 56(5):1916 –1930, 2008. ISSN 1053-587X.

[79] Peña, J.M. Learning Gaussian graphical models of gene networks with false discovery rate control. In *Proceedings of the 6th European conference on Evolutionary computation, machine learning and data mining in bioinformatics*, EvoBIO'08, pages 165–176. Springer-Verlag, Berlin, Heidelberg, 2008. ISBN 3-540-78756-9, 978-3-540-78756-3.

[80] Castelo, R., Roverato, A., and Chickering, M. A robust procedure for gaussian graphical model search from microarray data with p larger than n. *Journal of Machine Learning Research*, 7:2006, 2006.

[81] Whittaker, J. *Graphical Models in Applied Multivariate Statistics*. Wiley Publishing, 2009. ISBN 0470743662, 9780470743669.

[82] Daniels, M.J. and Kass, R.E. Shrinkage estimators for covariance matrices. *Biometrics*, 57(4):1173–1184, 2001.

[83] Schäfer, J. and Strimmer, K. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical applications in genetics and molecular biology*, 4, 2005. ISSN 1544-6115.

[84] Castelo, R. and Roverato, A. Reverse engineering molecular regulatory networks from microarray data with qp-graphs. *J Comput Biol*, 16(2):213–227, 2009.

[85] Penrose, R. A generalized inverse for matrices. *Mathematical Proceedings of the Cambridge Philosophical Society*, 51(03):406–413, 1955.

[86] Meinshausen, N. and Bühlmann, P. High dimensional graphs and variable selection with the Lasso. *Annals of Statistics*, 34:1436–1462, 2006.

[87] Yuan, M. and Lin, Y. Model selection and estimation in the Gaussian graphical model. *Biometrika*, 94(1):19–35, 2007.

[88] Friedman, J., Hastie, T., and Tibshirani, R. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

[89] Wong, F., Carter, C.K., and Kohn, R. Efficient estimation of covariance selection models. *Biometrika*, 90(4):809–830, 2003.

[90] Miyamura, M. and Kano, Y. Robust Gaussian graphical modeling. *Journal of Multivariate Analysis*, 97(7):1525 – 1550, 2006. ISSN 0047-259X.

[91] Liebermeister, W. and Klipp, E. Bringing metabolic networks to life: integration of kinetic, metabolic, and proteomic data. *Theor Biol Med Model*, 3:42, 2006.

[92] Palsson, B.O. *Systems Biology: Properties of Reconstructed Networks*. Cambridge University Press, 1 edition, 2006. ISBN 0521859034.

[93] Famili, I., Mahadevan, R., and Palsson, B.O. k-Cone analysis: determining all candidate values for kinetic parameters on a network scale. *Biophys J*, 88(3):1616–1625, 2005.

[94] Michaelis, L. and Menten, M.L. Die Kinetik der Invertinwirkung. *Biochem. Z*, 49(333-369):352, 1913.

[95] Dräger, A., Kronfeld, M., Ziller, M.J., Supper, J., Planatscher, H., Magnus, J.B., Oldiges, M., Kohlbacher, O., and Zell, A. Modeling metabolic networks in C. glutamicum: a comparison of rate laws in combination with various parameter optimization strategies. *BMC Syst Biol*, 3:5, 2009.

[96] Shampine, L.F. and Reichelt, M.W. The MATLAB ODE Suite. *SIAM Journal on Scientific Computing*, 18(1):1–22, 1997.

[97] Berg, J.M., Tymoczko, J.L., and Stryer, L. *Biochemistry*. W. H. Freeman, sixth edition edition, 2006. ISBN 0716787245.

[98] Hynne, F., Dan, S., and Srensen, P.G. Full-scale model of glycolysis in Saccharomyces cerevisiae. *Biophys Chem*, 94(1-2):121–163, 2001.

[99] Altman, D.G. and Bland, J.M. Diagnostic tests. 1: Sensitivity and specificity. *BMJ*, 308(6943):1552, 1994.

[100] Van Rijsbergen, C.J. *Information Retrieval, 2nd edition*. Dept. of Computer Science, University of Glasgow, 1979.

[101] Winicov, I. and Pizer, L.I. The mechanism of end product inhibition of serine biosynthesis. IV. Subunit structure of phosphoglycerate dehydrogenase and steady state kinetic studies of phosphoglycerate oxidation. *J Biol Chem*, 249(5):1348–1355, 1974.

[102] Tyson, J.J., Csikasz-Nagy, A., and Novak, B. The dynamics of cell cycle regulation. *Bioessays*, 24(12):1095–1109, 2002.

[103] Craciun, G., Tang, Y., and Feinberg, M. Understanding bistability in complex enzyme-driven reaction networks. *Proc Natl Acad Sci U S A*, 103(23):8697–8702, 2006.

[104] Huang, S., Guo, Y.P., May, G., and Enver, T. Bifurcation dynamics in lineage-commitment in bipotent progenitor cells. *Dev Biol*, 305(2):695–713, 2007.

[105] Soranzo, N. and Altafini, C. ERNEST: a toolbox for chemical reaction network theory. *Bioinformatics*, 25(21):2853–2854, 2009.

[106] Feinberg, M. Chemical reaction network structure and the stability of complex isothermal reactors - I. The deficiency zero and deficiency one theorems. *Chemical Engr. Sci.*, 42:2229–2268, 1987.

[107] Banaji, M., Donnell, P., and Baigent, S. P Matrix Properties, Injectivity, and Stability in Chemical Reaction Systems. *SIAM Journal of Applied Mathematics*, 67(6):1523–1547, 2007.

[108] White, S. and Smyth, P. A Spectral Clustering Approach To Finding Communities in Graphs. In *SIAM International Conference on Data Mining*. 2005.

[109] Maslov, S. and Sneppen, K. Specificity and stability in topology of protein networks. *Science*, 296(5569):910–913, 2002.

[110] Wong, P., Althammer, S., Hildebrand, A., Kirschner, A., Pagel, P., Geissler, B., Smialowski, P., Blöchl, F., Oesterheld, M., Schmidt, T., Strack, N., Theis, F.J., Ruepp, A., and Frishman, D. An evolutionary and structural characterization of mammalian protein complex organization. *BMC Genomics*, 9:629, 2008.

[111] Hartsperger, M.L., Blöchl, F., Stümpflen, V., and Theis, F.J. Structuring heterogeneous biological information using fuzzy clustering of k-partite graphs. *BMC Bioinformatics*, 11:522, 2010.

[112] Newman, M.E.J. and Girvan, M. Finding and evaluating community structure in networks. *Phys. Rev. E*, 69(2):026113, 2004.

[113] Matsuzaka, T., Shimano, H., Yahagi, N., Kato, T., Atsumi, A., Yamamoto, T., Inoue, N., Ishikawa, M., Okada, S., Ishigaki, N., Iwasaki, H., Iwasaki, Y., Karasawa, T., Kumadaki, S., Matsui, T., Sekiya, M., Ohashi, K., Hasty, A.H., Nakagawa, Y., Takahashi, A., Suzuki, H., Yatoh, S., Sone, H., Toyoshima, H., ichi Osuga, J., and Yamada, N. Crucial role of a long-chain fatty acid elongase, Elovl6, in obesity-induced insulin resistance. *Nat Med*, 13(10):1193–1202, 2007.

[114] Eaton, S., Bartlett, K., and Pourfarzam, M. Mammalian mitochondrial beta-oxidation. *Biochem J*, 320 (Pt 2):345–357, 1996.

[115] Spector, A. Essentiality of fatty acids. *Lipids*, 34(0):S1–S3, 1999.

[116] Duarte, N.C., Becker, S.A., Jamshidi, N., Thiele, I., Mo, M.L., Vo, T.D., Srivas, R., and Palsson, B.O. Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proc Natl Acad Sci U S A*, 104(6):1777–1782, 2007.

[117] Ma, H., Sorokin, A., Mazein, A., Selkov, A., Selkov, E., Demin, O., and Goryanin, I. The Edinburgh human metabolic network reconstruction and its functional analysis. *Mol Syst Biol*, 3:135, 2007.

[118] Arkin, A., Shen, P., and Ross, J. A Test Case of Correlation Metric Construction of a Reaction Pathway from Measurements. *Science*, 277(5330):1275–1279, 1997.

[119] Steuer, R. Review: on the analysis and interpretation of correlations in metabolomic data. *Brief Bioinform*, 7(2):151–158, 2006.

[120] Altmaier, E., Ramsay, S.L., Graber, A., Mewes, H.W., Weinberger, K.M., and Suhre, K. Bioinformatics analysis of targeted metabolomics–uncovering old and new tales of diabetic mice under medication. *Endocrinology*, 149(7):3478–3489, 2008.

[121] Nicholson, G., Rantalainen, M., Li, J.V., Maher, A.D., Malmodin, D., Ahmadi, K.R., Faber, J.H., Barrett, A., Min, J.L., Rayner, N.W., Toft, H., Krestyaninova, M., Viksna, J., Neogi, S.G., Dumas, M.E., Sarkans, U., Consortium, M.A.G.E., Donnelly, P., Illig, T., Adamski, J., Suhre, K., Allen, M., Zondervan, K.T., Spector, T.D., Nicholson, J.K., Lindon, J.C., Baunsgaard, D., Holmes, E., McCarthy, M.I., and Holmes, C.C. A genome-wide metabolic QTL analysis in Europeans implicates two loci shaped by recent positive selection. *PLoS Genet*, 7(9):e1002270, 2011.

[122] Horai, H., Arita, M., Kanaya, S., Nihei, Y., Ikeda, T., Suwa, K., Ojima, Y., Tanaka, K., Tanaka, S., Aoshima, K., Oda, Y., Kakazu, Y., Kusano, M., Tohge, T., Matsuda, F., Sawada, Y., Hirai, M.Y., Nakanishi, H., Ikeda, K., Akimoto, N., Maoka, T., Takahashi,

H., Ara, T., Sakurai, N., Suzuki, H., Shibata, D., Neumann, S., Iida, T., Tanaka, K., Funatsu, K., Matsuura, F., Soga, T., Taguchi, R., Saito, K., and Nishioka, T. MassBank: a public repository for sharing mass spectral data for life sciences. *J Mass Spectrom*, 45(7):703–714, 2010.

[123] Afeefy, H., Liebman, J., and Stein, S. *NIST Chemistry WebBook, NIST Standard Reference Database Number 69*, chapter Neutral Thermochemical Data, http://webbook.nist.gov. 2011.

[124] Wishart, D.S., Knox, C., Guo, A.C., Eisner, R., Young, N., Gautam, B., Hau, D.D., Psychogios, N., Dong, E., Bouatra, S., Mandal, R., Sinelnikov, I., Xia, J., Jia, L., Cruz, J.A., Lim, E., Sobsey, C.A., Shrivastava, S., Huang, P., Liu, P., Fang, L., Peng, J., Fradette, R., Cheng, D., Tzur, D., Clements, M., Lewis, A., Souza, A.D., Zuniga, A., Dawe, M., Xiong, Y., Clive, D., Greiner, R., Nazyrova, A., Shaykhutdinov, R., Li, L., Vogel, H.J., and Forsythe, I. HMDB: a knowledgebase for the human metabolome. *Nucleic Acids Res*, 37(Database issue):D603–D610, 2009.

[125] Steffens, D.C., Jiang, W., Krishnan, K.R.R., Karoly, E.D., Mitchell, M.W., O'Connor, C.M., and Kaddurah-Daouk, R. Metabolomic differences in heart failure patients with and without major depression. *J Geriatr Psychiatry Neurol*, 23(2):138–146, 2010.

[126] Kind, T. and Fiehn, O. Seven Golden Rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. *BMC Bioinformatics*, 8:105, 2007.

[127] Bowen, B.P. and Northen, T.R. Dealing with the unknown: metabolomics and metabolite atlases. *J Am Soc Mass Spectrom*, 21(9):1471–1476, 2010.

[128] Wishart, D.S. Advances in metabolite identification. *Bioanalysis*, 3(15):1769–1782, 2011.

[129] Rasche, F., Svatoš, A., Maddula, R.K., Böttcher, C., and Böcker, S. Computing Fragmentation Trees from Tandem Mass Spectrometry Data. *Analytical Chemistry*, 83(4):1243–1251, 2011.

[130] Mihaleva, V.V., Verhoeven, H.A., de Vos, R.C.H., Hall, R.D., and van Ham, R.C.H.J. Automated procedure for candidate compound selection in GC-MS metabolomics based on prediction of Kovats retention index. *Bioinformatics*, 25(6):787–794, 2009.

[131] Creek, D.J., Jankevics, A., Breitling, R., Watson, D.G., Barrett, M.P., and Burgess, K.E.V. Towards Global Metabolomics Analysis with Liquid Chromatography-Mass Spectrometry: Improved Metabolite Identification by Retention Time Prediction. *Anal Chem*, 2011.

[132] Böcker, S., Letzel, M.C., Liptk, Z., and Pervukhin, A. SIRIUS: decomposing isotope patterns for metabolite identification. *Bioinformatics*, 25(2):218–224, 2009.

[133] Gipson, G., Tatsuoka, K., Sokhansanj, B., Ball, R., and Connor, S. Assignment of MS-based metabolomic datasets via compound interaction pair mapping. *Metabolomics*, 4:94–103, 2008. ISSN 1573-3882. 10.1007/s11306-007-0096-9.

[134] Weber, R.J. and Viant, M.R. MI-Pack: Increased confidence of metabolite identification in mass spectra by integrating accurate masses and metabolic pathways. *Chemometrics and Intelligent Laboratory Systems*, 104(1):75 – 82, 2010. ISSN 0169-7439.

[135] Nayak, R.R., Kearns, M., Spielman, R.S., and Cheung, V.G. Coexpression network based on natural variation in human gene expression reveals gene interactions and functions. *Genome Res*, 19(11):1953–1962, 2009.

[136] Sharan, R., Ulitsky, I., and Shamir, R. Network-based prediction of protein function. *Mol Syst Biol*, 3:88, 2007.

[137] Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., and Sham, P.C. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*, 81(3):559–575, 2007.

[138] Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., and Sherlock, G. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, 25(1):25–29, 2000.

[139] Zhai, G., Teumer, A., Stolk, L., Perry, J.R.B., Vandenput, L., Coviello, A.D., Koster, A., Bell, J.T., Bhasin, S., Eriksson, J., Eriksson, A., Ernst, F., Ferrucci, L., Frayling, T.M., Glass, D., Grundberg, E., Haring, R., Hedman, A.K., Hofman, A., Kiel, D.P., Kroemer, H.K., Liu, Y., Lunetta, K.L., Maggio, M., Lorentzon, M., Mangino, M., Melzer, D., Miljkovic, I., Consortium, M.H.E.R., Nica, A., Penninx, B.W.J.H., Vasan, R.S., Rivadeneira, F., Small, K.S., Soranzo, N., Uitterlinden, A.G., Völzke, H., Wilson, S.G., Xi, L., Zhuang, W.V., Harris, T.B., Murabito, J.M., Ohlsson, C., Murray, A., de Jong, F.H., Spector, T.D., and Wallaschofski, H. Eight common genetic variants associated with serum DHEAS levels suggest a key role in ageing mechanisms. *PLoS Genet*, 7(4):e1002025, 2011.

[140] Otterness, D.M., Wieben, E.D., Wood, T.C., Watson, W.G., Madden, B.J., McCormick, D.J., and Weinshilboum, R.M. Human liver dehydroepiandrosterone sulfotransferase: molecular cloning and expression of cDNA. *Mol Pharmacol*, 41(5):865–872, 1992.

[141] Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., and Mesirov, J.P. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*, 102(43):15545–15550, 2005.

[142] Tate, S.S. and Meister, A. gamma-Glutamyl transpeptidase from kidney. *Methods Enzymol*, 113:400–419, 1985.

[143] Kováts, E. Gas-chromatographische Charakterisierung organischer Verbindungen. Teil 1: Retentionsindices aliphatischer Halogenide, Alkohole, Aldehyde und Ketone. *Helvetica Chimica Acta*, 41(7):1915–1932, 1958. ISSN 1522-2675.

[144] Eaton, S. Control of mitochondrial beta-oxidation flux. *Prog Lipid Res*, 41(3):197–239, 2002.

[145] Boden, G., Scapa, E.F., Kanno, K., Cohen, D.E., Brosnan, M.E., Brosnan, J.T., Pessayre, D., Roy-Chowdhury, N., Lu, Y., Roy-Chowdhury, J., Jansen, P.L., Faber, K.N., Hussinger, D., Lingappa, V.R., Fernndez-Checa, J.C., Garca-Ruiz, C., Puy, H., Deybach, J.C., Okuno, M., Matsushima-Nishiwaki, R., Kojima, S., Brown, K.E., Brewer, G.J., Harris, E.D., Askari, F.K., Neuschwander-Tetri, B.A., Liddle, C., and Stedman, C.A. *Metabolism*, pages 129–249. Blackwell Publishing Ltd, 2008. ISBN 9780470691861.

[146] Tukey, R.H. and Strassburg, C.P. Human UDP-glucuronosyltransferases: metabolism, expression, and disease. *Annu Rev Pharmacol Toxicol*, 40:581–616, 2000.

[147] Bosma, P.J. Inherited disorders of bilirubin metabolism. *J Hepatol*, 38(1):107–117, 2003.

[148] Abu-Bakar, A., Moore, M.R., and Lang, M.A. Evidence for induced microsomal bilirubin degradation by cytochrome P450 2A5. *Biochem Pharmacol*, 70(10):1527–1535, 2005.

[149] Kibriya, M.G., Jasmine, F., Argos, M., Andrulis, I.L., John, E.M., Chang-Claude, J., and Ahsan, H. A pilot genome-wide association study of early-onset breast cancer. *Breast Cancer Res Treat*, 114(3):463–477, 2009.

[150] Djouss, L., Levy, D., Cupples, L.A., Evans, J.C., D'Agostino, R.B., and Ellison, R.C. Total serum bilirubin and risk of cardiovascular disease in the Framingham offspring study. *Am J Cardiol*, 87(10):1196–200; A4, 7, 2001.

[151] Bosma, P.J., van der Meer, I.M., Bakker, C.T., Hofman, A., Paul-Abrahamse, M., and Witteman, J.C. UGT1A1*28 allele and coronary heart disease: the Rotterdam Study. *Clin Chem*, 49(7):1180–1181, 2003.

[152] Baranano, D.E., Rao, M., Ferris, C.D., and Snyder, S.H. Biliverdin reductase: a major physiologic cytoprotectant. *Proc Natl Acad Sci U S A*, 99(25):16093–16098, 2002.

[153] Bowers-Komro, D.M., McCormick, D.B., King, G.A., Sweeny, J.G., and Iacobucci, G.A. Confirmation of 2-O-methyl ascorbic acid as the product from the enzymatic methylation of L-ascorbic acid by catechol-O-methyltransferase. *Int J Vitam Nutr Res*, 52(2):186–193, 1982.

[154] Milburn, M., Guo, L., Wulff, J.E., and Lawton, K.A. DETERMINATION OF THE LIVER TOXICITY OF AN AGENT. 2010.

[155] Männistö, P.T. and Kaakkola, S. Catechol-O-methyltransferase (COMT): biochemistry, molecular biology, pharmacology, and clinical efficacy of the new selective COMT inhibitors. *Pharmacol Rev*, 51(4):593–628, 1999.

[156] Butterworth, M., Lau, S.S., and Monks, T.J. 17 beta-Estradiol metabolism by hamster hepatic microsomes. Implications for the catechol-O-methyl transferase-mediated detoxication of catechol estrogens. *Drug Metab Dispos*, 24(5):588–594, 1996.

[157] Imig, J.D. ACE Inhibition and Bradykinin-Mediated Renal Vascular Responses: EDHF Involvement. *Hypertension*, 43(3):533–535, 2004.

[158] Acharya, K.R., Sturrock, E.D., Riordan, J.F., and Ehlers, M.R.W. Ace revisited: a new target for structure-based drug design. *Nat Rev Drug Discov*, 2(11):891–902, 2003.

[159] Wilson, P.W., D'Agostino, R.B., Levy, D., Belanger, A.M., Silbershatz, H., and Kannel, W.B. Prediction of coronary heart disease using risk factor categories. *Circulation*, 97(18):1837–1847, 1998.

[160] Adams, S.H., Hoppel, C.L., Lok, K.H., Zhao, L., Wong, S.W., Minkler, P.E., Hwang, D.H., Newman, J.W., and Garvey, W.T. Plasma acylcarnitine profiles suggest incomplete long-chain fatty acid beta-oxidation and altered tricarboxylic acid cycle activity in type 2 diabetic African-American women. *J Nutr*, 139(6):1073–1081, 2009.

[161] Ideker, T. and Krogan, N.J. Differential network biology. *Mol Syst Biol*, 8:565, 2012.

[162] Collaer, M.L. and Hines, M. Human behavioral sex differences: a role for gonadal hormones during early development? *Psychol Bull*, 118(1):55–107, 1995.

[163] Murphy, D.G., DeCarli, C., McIntosh, A.R., Daly, E., Mentis, M.J., Pietrini, P., Szczepanik, J., Schapiro, M.B., Grady, C.L., Horwitz, B., and Rapoport, S.I. Sex differences in human brain morphometry and metabolism: an in vivo quantitative magnetic resonance imaging and positron emission tomography study on the effect of aging. *Arch Gen Psychiatry*, 53(7):585–594, 1996.

[164] Hankin, B.L. and Abramson, L.Y. Development of gender differences in depression: an elaborated cognitive vulnerability-transactional stress theory. *Psychol Bull*, 127(6):773–796, 2001.

[165] Blaak, E. Gender differences in fat metabolism. *Curr Opin Clin Nutr Metab Care*, 4(6):499–502, 2001.

[166] Kim, A.M., Tingen, C.M., and Woodruff, T.K. Sex bias in trials and treatment must end. *Nature*, 465(7299):688–689, 2010.

[167] Wizemann, T.M., Mary-Lou Pardue, Editors, C.o.U.t.B.o.S., and Gender Differences, B.o.H.S.P. *Exploring the Biological Contributions to Human Health: Does Sex Matter?* The National Academies Press, 2001. ISBN 9780309072816.

[168] Heymsfield, S.B., Smith, R., Aulet, M., Bensen, B., Lichtman, S., Wang, J., and Pierson, Jr, R. Appendicular skeletal muscle mass: measurement by dual-photon absorptiometry. *Am J Clin Nutr*, 52(2):214–218, 1990.

[169] Wells, J.C.K., Chomtho, S., and Fewtrell, M.S. Programming of body composition by early growth and nutrition. *Proc Nutr Soc*, 66(3):423–434, 2007.

[170] Schols, A.M.W.J., Broekhuizen, R., Weling-Scheepers, C.A., and Wouters, E.F. Body composition and mortality in chronic obstructive pulmonary disease. *Am J Clin Nutr*, 82(1):53–59, 2005.

[171] King, S.J., Nyulasi, I.B., Strauss, B.J.G., Kotsimbos, T., Bailey, M., and Wilson, J.W. Fat-free mass depletion in cystic fibrosis: associated with lung disease severity but poorly detected by body mass index. *Nutrition*, 26(7-8):753–759, 2010.

[172] Kouri, E.M., Pope, Jr, H., Katz, D.L., and Oliva, P. Fat-free mass index in users and nonusers of anabolic-androgenic steroids. *Clin J Sport Med*, 5(4):223–228, 1995.

[173] Jourdan, C., Petersen, A.K., Gieger, C., Döring, A., Illig, T., Wang-Sattler, R., Meisinger, C., Peters, A., Adamski, J., Prehn, C., Suhre, K., Altmaier, E., Kastenmüller, G., Römisch-Margl, W., Theis, F.J., Krumsiek, J., Wichmann, H.E., and Linseisen, J. Body

fat free mass is associated with the serum metabolite profile in a population-based study. *PLoS One*, 7(6):e40009, 2012.

[174] Kupper, N. and Denollet, J. Type D personality as a prognostic factor in heart disease: assessment and mediating mechanisms. *J Pers Assess*, 89(3):265–276, 2007.

[175] Myint, A.M., Kim, Y.K., Verkerk, R., Scharp, S., Steinbusch, H., and Leonard, B. Kynurenine pathway in major depression: evidence of impaired neuroprotection. *J Affect Disord*, 98(1-2):143–151, 2007.

[176] Ritsner, M., Gibel, A., Maayan, R., Ratner, Y., Ram, E., Modai, I., and Weizman, A. State and trait related predictors of serum cortisol to DHEA(S) molar ratios and hormone concentrations in schizophrenia patients. *European Neuropsychopharmacology*, 17(4):257 – 264, 2007. ISSN 0924-977X.

[177] Altmaier, E., Emeny, R., Krumsiek, J., Lacruz, E., Lukaschek, K., Haefner, S., Kastenmüller, G., Römisch-Margl, W., Prehn, C., Mohney, R.P., Milburn, M.V., Illig, T., Adamski, J., Theis, F.J., Suhre, K., and Ladwig, K.H. Metabolomic profiles in individuals with negative affectivity and social inhibition: a population-based study of Type D personality. *Psychoneuroendocrinology*, in press.

[178] He, H., Nilsson, C.L., Emmett, M.R., Marshall, A.G., Kroes, R.A., Moskal, J.R., Ji, Y., Colman, H., Priebe, W., Lang, F.F., and Conrad, C.A. Glycomic and transcriptomic response of GSC11 glioblastoma stem cells to STAT3 phosphorylation inhibition and serum-induced differentiation. *J Proteome Res*, 9(5):2098–2108, 2010.

[179] Pontn, J. and Macintyre, E.H. Long term culture of normal and neoplastic human glia. *Acta Pathol Microbiol Scand*, 74(4):465–486, 1968.

[180] Seton-Rogers, S. Hypoxia: HIF switch. *Nat Rev Cancer*, 11(6):391–391, 2011. ISSN 1474-175X.

[181] Luqiu, Z., Yiquan, K., Gengqiang, L., Yijing, L., Xiaodan, J., and Yingqian, C. A new design immunotoxin for killing high-grade glioma U87 cells: From in vitro to in vivo. *J Immunotoxicol*, 2012.

[182] Clark, M.J., Homer, N., O'Connor, B.D., Chen, Z., Eskin, A., Lee, H., Merriman, B., and Nelson, S.F. U87MG Decoded: The Genomic Sequence of a Cytogenetically Aberrant Human Cancer Cell Line. *PLoS Genet*, 6(1):e1000832, 2010.

[183] Lang, F., Shono, T., and Gilbert, M. Ad-p53 sensitizes wild-type p53 gliomas to the topoisomerase I inhibitor SN-38. *Neuro-Oncol*, 4:323–324, 2002.

[184] He, H., Conrad, C.A., Nilsson, C.L., Ji, Y., Schaub, T.M., Marshall, A.G., and Emmett, M.R. Method for lipidomic analysis: p53 expression modulation of sulfatide, ganglioside, and phospholipid composition of U87 MG glioblastoma cells. *Anal Chem*, 79(22):8423–8430, 2007.

[185] Sonnino, S., Mauri, L., Chigorno, V., and Prinetti, A. Gangliosides as components of lipid membrane domains. *Glycobiology*, 17(1):1R–13R, 2007.

[186] Benjamini, Y. and Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995. ISSN 00359246.

[187] Ried, J.S., Döring, A., Oexle, K., Meisinger, C., Winkelmann, J., Klopp, N., Meitinger, T., Peters, A., Suhre, K., Wichmann, H.E., and Gieger, C. PSEA: Phenotype Set Enrichment AnalysisA New Method for Analysis of Multiple Phenotypes. *Genetic Epidemiology*, 36(3):244–252, 2012. ISSN 1098-2272.

[188] Zeggini, E., Scott, L.J., Saxena, R., Voight, B.F., Marchini, J.L., Hu, T., de Bakker, P.I.W., Abecasis, G.R., Almgren, P., Andersen, G., Ardlie, K., Boström, K.B., Bergman, R.N., Bonnycastle, L.L., Borch-Johnsen, K., Burtt, N.P., Chen, H., Chines, P.S., Daly, M.J., Deodhar, P., Ding, C.J., Doney, A.S.F., Duren, W.L., Elliott, K.S., Erdos, M.R., Frayling, T.M., Freathy, R.M., Gianniny, L., Grallert, H., Grarup, N., Groves, C.J., Guiducci, C., Hansen, T., Herder, C., Hitman, G.A., Hughes, T.E., Isomaa, B., Jackson, A.U., Jrgensen, T., Kong, A., Kubalanza, K., Kuruvilla, F.G., Kuusisto, J., Langenberg, C., Lango, H., Lauritzen, T., Li, Y., Lindgren, C.M., Lyssenko, V., Marvelle, A.F., Meisinger, C., Midthjell, K., Mohlke, K.L., Morken, M.A., Morris, A.D., Narisu, N., Nilsson, P., Owen, K.R., Palmer, C.N.A., Payne, F., Perry, J.R.B., Pettersen, E., Platou, C., Prokopenko, I., Qi, L., Qin, L., Rayner, N.W., Rees, M., Roix, J.J., Sandbaek, A., Shields, B., Sjögren, M., Steinthorsdottir, V., Stringham, H.M., Swift, A.J., Thorleifsson, G., Thorsteinsdottir, U., Timpson, N.J., Tuomi, T., Tuomilehto, J., Walker, M., Watanabe, R.M., Weedon, M.N., Willer, C.J., , W.T.C.C.C., Illig, T., Hveem, K., Hu, F.B., Laakso, M., Stefansson, K., Pedersen, O., Wareham, N.J., Barroso, I., Hattersley, A.T., Collins, F.S., Groop, L., McCarthy, M.I., Boehnke, M., and Altshuler, D. Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nat Genet*, 40(5):638–645, 2008.

[189] McCarthy, M.I., Abecasis, G.R., Cardon, L.R., Goldstein, D.B., Little, J., Ioannidis, J.P.A., and Hirschhorn, J.N. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat Rev Genet*, 9(5):356–369, 2008. ISSN 1471-0056.

[190] Andrew, T., Hart, D.J., Snieder, H., de Lange, M., Spector, T.D., and MacGregor, A.J. Are twins and singletons comparable? A study of disease-related and lifestyle characteristics in adult women. *Twin Res*, 4(6):464–477, 2001.

[191] Eastman, J.W., Sherwin, J.E., Wong, R., Liao, C.L., Currier, R.J., Lorey, F., and Cunningham, G. Use of the phenylalanine:tyrosine ratio to test newborns for phenylketonuria in a large public health screening programme. *J Med Screen*, 7(3):131–135, 2000.

[192] Burton, B.K. Inborn errors of metabolism in infancy: a guide to diagnosis. *Pediatrics*, 102(6):E69, 1998.

[193] Shlens, J. A tutorial on Principal Component Analysis. In *Systems Neurobiology Laboratory, Salk Institute for Biological Studies*. 2005.

[194] Werth, M.T., Halouska, S., Shortridge, M.D., Zhang, B., and Powers, R. Analysis of metabolomic PCA data using tree diagrams. *Anal Biochem*, 399(1):58–63, 2010.

[195] Hyvärinen, A., Karhunen, J., and Oja, E. *Independent component analysis*. Adaptive and learning systems for signal processing, communications, and control. J. Wiley, 2001. ISBN 9780471405405.

[196] Comon, P. Independent component analysis, a new concept? *Signal Process.*, 36:287–314, 1994. ISSN 0165-1684.

[197] Theis, F. Uniqueness of real and complex linear independent component analysis revisited. In *Proc. EUSIPCO 2004*, pages 1705–1708. Vienna, Austria, 2004.

[198] Makeig, S., Bell, A.J., Jung, T.P., and Sejnowski, T.J. Independent Component Analysis of Electroencephalographic Data. In D.S. Touretzky, M.C. Mozer, and M.E. Hasselmo, editors, *Advances in Neural Information Processing Systems*, volume 8, pages 145–151. The MIT Press, 1996.

[199] Mckeown, M.J., Makeig, S., Brown, G.G., Jung, T.P., Kindermann, S.S., Kindermann, R.S., Bell, A.J., and Sejnowski, T.J. Analysis of fMRI Data by Blind Separation Into Independent Spatial Components. *Human Brain Mapping*, 6:160–188, 1998.

[200] Karvanen, J. and Theis, F.J. Spatial ICA of fMRI data in time windows. In *Proceedings: Bayesian inference and maximum entropy methods in science and engineering: 24th internat. workshop, Garching, Germany, 25 - 30 July 2004*, volume 735 of *AIP conference proceedings*, pages 312–319. American Institute of Physics, Melville, NY, 2004.

[201] Keck, I.R., Theis, F.J., Gruber, P., Lang, E., Specht, K., and Puntonet, C.G. 3D spatial analysis of fMRI data on a word perception task. In C.G. Puntonet, editor, *Independent component analysis and blind signal separation: fifth international conference, ICA 2004, Granada, Spain, September 22 - 24, 2004; proceedings*, volume 3195 of *Lecture Notes in Computer Science*, pages 977–984. Springer, Berlin, 2004.

[202] Zhang, X.W., Yap, Y.L., Wei, D., Chen, F., and Danchin, A. Molecular diagnosis of human cancer type by gene expression profiles and independent component analysis. *Eur J Hum Genet*, 13(12):1303–1311, 2005.

[203] Huang, D.S. and Zheng, C.H. Independent component analysis-based penalized discriminant method for tumor classification using gene expression data. *Bioinformatics*, 22(15):1855–1862, 2006.

[204] Teschendorff, A.E., Journe, M., Absil, P.A., Sepulchre, R., and Caldas, C. Elucidating the altered transcriptional programs in breast cancer using independent component analysis. *PLoS Comput Biol*, 3(8):e161, 2007.

[205] Lutter, D., Ugocsai, P., Grandl, M., Orso, E., Theis, F., Lang, E.W., and Schmitz, G. Analyzing M-CSF dependent monocyte/macrophage differentiation: expression modes and meta-modes derived from an independent component analysis. *BMC Bioinformatics*, 9:100, 2008.

[206] Schachtner, R., Lutter, D., Knollmüller, P., Tom, A.M., Theis, F.J., Schmitz, G., Stetter, M., Vilda, P.G., and Lang, E.W. Knowledge-based gene expression classification via matrix factorization. *Bioinformatics*, 24(15):1688–1697, 2008.

[207] Hofmann, J., Ashry, A.E.N.E., Anwar, S., Erban, A., Kopka, J., and Grundler, F. Metabolic profiling reveals local and systemic responses of host plants to nematode parasitism. *Plant J*, 62(6):1058–1071, 2010.

[208] Führs, H., Götze, S., Specht, A., Erban, A., Gallien, S., Heintz, D., Dorsselaer, A.V., Kopka, J., Braun, H.P., and Horst, W.J. Characterization of leaf apoplastic peroxidases and metabolites in Vigna unguiculata in response to toxic manganese supply and silicon. *J Exp Bot*, 60(6):1663–1678, 2009.

[209] Scholz, M., Gatzek, S., Sterling, A., Fiehn, O., and Selbig, J. Metabolite fingerprinting: detecting biological features by independent component analysis. *Bioinformatics*, 20(15):2447–2454, 2004.

[210] Wienkoop, S., Morgenthal, K., Wolschin, F., Scholz, M., Selbig, J., and Weckwerth, W. Integration of metabolomic and proteomic phenotypes: analysis of data covariance dissects starch and RFO metabolism from low and high temperature compensation response in Arabidopsis thaliana. *Mol Cell Proteomics*, 7(9):1725–1736, 2008.

[211] Martin, F.P.J., Rezzi, S., , I.M., Philippe, D., Tornier, L., Messlik, A., Holzlwimmer, G., Baur, P., Quintanilla-Fend, L., Loh, G., Blaut, M., Blum, S., Kochhar, S., and Haller, D. Metabolic Assessment of Gradual Development of Moderate Experimental Colitis in IL-10 Deficient Mice. *Journal of Proteome Research*, 8(5):2376–2387, 2009.

[212] Himberg, J., Hyvärinen, A., and Esposito, F. Validating the independent components of neuroimaging time series via clustering and visualization. *NeuroImage*, 22(3):1214 – 1222, 2004. ISSN 1053-8119.

[213] Keck, I., Theis, F., Gruber, P., Lang, E., Specht, K., Fink, G., Tomé, A., and Puntonet, C. Automated clustering of ICA results for fMRI data analysis. In *Proc. CIMED 2005*, pages 211–216. Lisbon, Portugal, 2005.

[214] Højen-Sørensen, P.A.R., Winther, O., and Hansen, L.K. Mean-field approaches to independent component analysis. *Neural Comput*, 14(4):889–918, 2002.

[215] Gutch, H., Krumsiek, J., and Theis, F. An ISA Algorithm With Unknown Group Sizes Identifies Meaningful Clusters in Metabolomics Data. EURASIP, 2011.

[216] Belouchran, A. and Cardoso, J.F. Maximum Likelihood Source Separation By the Expectation-Maximization Technique: Deterministic and Stochastic Implementation. In *in Proc. NOLTA*, pages 49–53. 1995.

[217] Hansen, L.K. *Advances in independent components analysis*, chapter Blind separation of noisy image mixtures., pages 165–187. Springer-Verlag, 2000.

[218] Fahrmeir, L., Kneib, T., and Lang, S. *Regression. Modelle, Methoden und Anwendungen*. Springer, Heidelberg, 2nd edition, 2009.

[219] Xia, J. and Wishart, D.S. Web-based inference of biological patterns, functions and pathways from metabolomic data using MetaboAnalyst. *Nat Protoc*, 6(6):743–760, 2011.

[220] Lusis, A.J. and Pajukanta, P. A treasure trove for lipoprotein biology. *Nat Genet*, 40(2):129–130, 2008.

[221] DiLeo, M.V., Strahan, G.D., den Bakker, M., and Hoekenga, O.A. Weighted Correlation Network Analysis (WGCNA) Applied to the Tomato Fruit Metabolome. *PLoS ONE*, 6(10):e26683, 2011.

[222] Camont, L., Chapman, M.J., and Kontush, A. Biological activities of HDL subpopulations and their relevance to cardiovascular disease. *Trends Mol Med*, 17(10):594–603, 2011.

[223] Petersen, A.K., Stark, K., Musameh, M.D., Nelson, C.P., Römisch-Margl, W., Kremer, W., Raffler, J., Krug, S., Skurk, T., Rist, M.J., Daniel, H., Hauner, H., Adamski, J., Tomaszewski, M., Döring, A., Peters, A., Wichmann, H.E., Kaess, B.M., Kalbitzer, H.R., Huber, F., Pfahlert, V., Samani, N.J., Kronenberg, F., Dieplinger, H., Illig, T., Hengstenberg, C., Suhre, K., Gieger, C., and Kastenmüller, G. Genetic associations with lipoprotein subfractions provide information on their biological nature. *Hum Mol Genet*, 2012.

[224] Felig, P., Marliss, E., and Cahill, G.F. Plasma amino acid levels and insulin secretion in obesity. *N Engl J Med*, 281(15):811–816, 1969.

[225] Wang, T.J., Larson, M.G., Vasan, R.S., Cheng, S., Rhee, E.P., McCabe, E., Lewis, G.D., Fox, C.S., Jacques, P.F., Fernandez, C., O'Donnell, C.J., Carr, S.A., Mootha, V.K., Florez, J.C., Souza, A., Melander, O., Clish, C.B., and Gerszten, R.E. Metabolite profiles and the risk of developing diabetes. *Nat Med*, 17(4):448–453, 2011.

[226] Layman, D.K. and Walker, D.A. Potential importance of leucine in treatment of obesity and the metabolic syndrome. *J Nutr*, 136(1 Suppl):319S–323S, 2006.

[227] Betteridge, D.J. Lipid control in patients with diabetes mellitus. *Nat Rev Cardiol*, 8(5):278–290, 2011.

[228] Suhre, K., Meisinger, C., Döring, A., Altmaier, E., Belcredi, P., Gieger, C., Chang, D., Milburn, M.V., Gall, W.E., Weinberger, K.M., Mewes, H.W., de Angelis, M.H., Wichmann, H.E., Kronenberg, F., Adamski, J., and Illig, T. Metabolic footprint of diabetes: a multiplatform metabolomics study in an epidemiological setting. *PLoS One*, 5(11):e13953, 2010.

[229] Hu, F.B. Metabolic profiling of diabetes: from black-box epidemiology to systems epidemiology. *Clin Chem*, 57(9):1224–1226, 2011.

[230] Bondia-Pons, I., Nordlund, E., Mattila, I., Katina, K., Aura, A.M., Kolehmainen, M., Oresic, M., Mykkanen, H., and Poutanen, K. Postprandial differences in the plasma metabolome of healthy Finnish subjects after intake of a sourdough fermented endosperm rye bread versus white wheat bread. *Nutr J*, 10(1):116, 2011.

[231] Heiden, M.G.V. Targeting cancer metabolism: a therapeutic window opens. *Nat Rev Drug Discov*, 10(9):671–684, 2011.

[232] Spector, T.D. and Williams, F.M. The UK Adult Twin Registry (TwinsUK). *Twin research and human genetics : the official journal of the International Society for Twin Studies*, 9(6):899–906, 2006. ISSN 1832-4274.

[233] Opgen-Rhein, R. and Strimmer, K. From correlation to causation networks: a simple approximate learning algorithm and its application to high-dimensional plant gene expression data. *BMC Syst Biol*, 1:37, 2007.

[234] Yuan, Y., Li, C.T., and Windram, O. Directed Partial Correlation: Inferring Large-Scale Gene Regulatory Network through Induced Topology Disruptions. *PLoS One*, 6(4):e16835, 2011.

[235] Bedford, T. and Cooke, R. Vines - A new graphical model for dependent random variables. *Annals of Statistics*, 30(4):1031–1068, 2002. ISSN 0090-5364.

[236] Chuang, H.Y., Lee, E., Liu, Y.T., Lee, D., and Ideker, T. Network-based classification of breast cancer metastasis. *Mol Syst Biol*, 3:140, 2007.