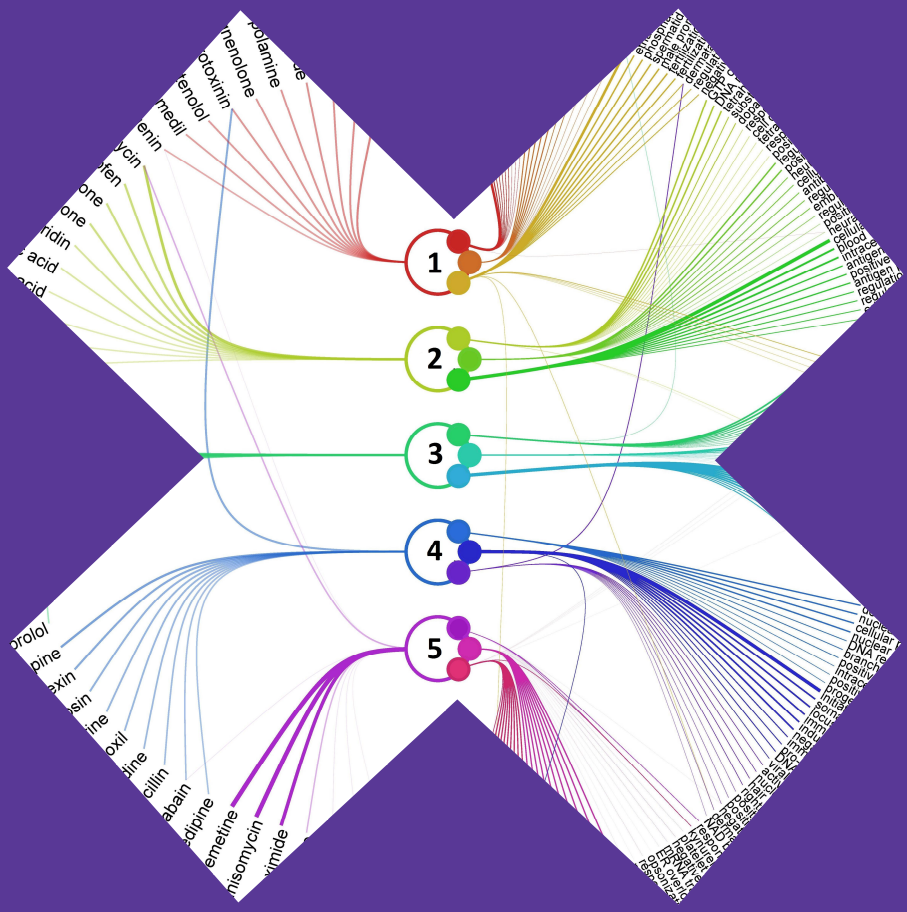
Department of Computer Science

Bayesian multi-view models for data-driven drug response analysis

Suleiman Ali Khan



Aalto University

DOCTORAL DISSERTATIONS

Aalto University publication series
DOCTORAL DISSERTATIONS 105/2015

Bayesian multi-view models for data-driven drug response analysis

Suleiman Ali Khan

A doctoral dissertation completed for the degree of Doctor of Science (Technology) to be defended, with the permission of the Aalto University School of Science, at a public examination held at the Auditorium U4 of the school on 7th September 2015 at 12 noon.

Aalto University
School of Science
Department of Computer Science
Statistical Machine Learning and Bioinformatics Group

Supervising professor
Prof. Samuel Kaski

Preliminary examiners
Prof. Sampsa Hautaniemi, University of Helsinki, Finland
Prof. Manfred Claassen, Institute of Molecular Systems Biology,
ETH Zurich, Switzerland

Opponent
Asst. Prof. Sara Mostafavi, University of British Columbia, Canada

Aalto University publication series
DOCTORAL DISSERTATIONS 105/2015

ISBN 978-952-60-6309-6 (printed)
ISBN 978-952-60-6310-2 (pdf)
ISSN-L 1799-4934
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)
http://urn.fi/URN:ISBN:978-952-60-6310-2

Unigrafia Oy
Helsinki 2015

Finland

Publication orders (printed book):
khan.suleiman@gmail.com

Aalto University

Abstract

Aalto University, P.O. Box 11000, FI-00076 Aalto www.aalto.fi

Author
Suleiman Ali Khan

Name of the doctoral dissertation
Bayesian multi-view models for data-driven drug response analysis

Publisher School of Science

Unit Department of Computer Science

Series Aalto University publication series DOCTORAL DISSERTATIONS 105/2015

Field of research Information and Computer Science

Manuscript submitted 29 April 2015 **Date of the defence** 7 September 2015

Permission to publish granted (date) 26 June 2015 **Language** English

☐ **Monograph** ☒ **Article dissertation (summary + original articles)**

Abstract

A central challenge faced by biological and medical research is to understand the impact of chemical entities on living cells. Identifying the relationships between the chemical structures and their cellular responses is valuable for improving drug design and targeted therapies. The chemical structures and their detailed molecular responses need to be combined through a systematic analysis to learn the complex dependencies, which can then assist in improving understanding of the molecular mechanisms of drugs as well as predictions on the effects of unknown molecules. Moreover, with emerging drug-response data sets being profiled over several disease types and phenotypic details, it is pertinent to develop advanced computational methods that can be used to study multiple sets of data together.

In this thesis, a novel multi-disciplinary challenge is undertaken for computationally analyzing interactions between multiple biological responses and chemical properties of drugs, while simultaneously advancing the computational methods to better learn these interactions. Specifically, multi-view dependency modeling of paired data sets is formulated as a means of systematically studying the drug-response relationships. First, the systematic analysis of drug structures and their genome-wide responses is presented as a multi-set dependency modeling problem and established methods are adopted to test the novel hypothesis.

Several novel extensions of the drug-response analysis are then presented that explore responses measured over multiple disease types and multiple levels of phenotypic detail, uncovering novel biological insights of potential impact. These analyses are made possible by novel advancements in multi-view methods. Specifically, the first Bayesian tensor canonical correlation analysis and its extensions are introduced to capture the underlying multi-way structure and applied in analyzing novel toxicogenomic interactions. The results illustrate that modeling the precise multi-view and multi-way formulation of the data is valuable for discovering interpretable latent components as well as for the prediction of unseen responses of drugs.

Therefore, the original contribution to knowledge in this dissertation is two-fold: first, the data-driven identification of relationships between structural properties of drugs and their genome-wide responses in cells and, second, novel advancements of multi-view methods that find dependencies between paired data sets. Open source implementations of the new methods have been released to facilitate further research.

Keywords Bayesian modeling, Machine learning, Multi-view learning, Computational biology, Bioinformatics, Toxicogenomics, Latent variable models, Bayesian tensor CCA

ISBN (printed) 978-952-60-6309-6 **ISBN (pdf)** 978-952-60-6310-2

ISSN-L 1799-4934 **ISSN (printed)** 1799-4934 **ISSN (pdf)** 1799-4942

Location of publisher Helsinki **Location of printing** Helsinki **Year** 2015

Pages 178 **urn** http://urn.fi/URN:ISBN:978-952-60-6310-2

Preface

This work has been carried out in the Statistical Machine Learning and Bioinformatics (MI) group in the Department of Computer Science (formerly known as Department of Information and Computer Science), Aalto University School of Science, Finland. I have had the privilege of being a member of the Finnish Center of Excellence in Computational Inference Research (COIN) as well as the Helsinki Institute of Information Technology (HIIT), both of which have provided a vast exposure and excellent networking with top researchers in the field. This research and thesis has been funded by the Academy of Finland project Computational modeling of the biological effects of chemicals (grant no. 140057), the Finnish Doctoral Programme in Computational Sciences (FICS), and the Finnish Foundation for Technology Promotion.

I am grateful to my supervisor Prof. Samuel Kaski for teaching me the principles of science and for giving me the opportunity to work in a fully interdisciplinary field with top-notch collaborations. I feel deeply privileged to have learned so much from a top researcher! I wish to especially thank Prof. Olli Kallioniemi, Director of Institute of Molecular Medicine Finland (FIMM), Dr. Krister Wennerberg from Cancer Chemical Systems Biology group at FIMM, and Prof. Antti Poso from Drug Design Laboratory at University of Eastern Finland, both for their guidance in the collaborative research, and for teaching me how to envision a multitude of outcomes from such a multidisciplinary setup. Your contributions extend way beyond what we have written together! This outstanding multidisciplinary environment has enabled me to acquire skills in both statistical machine learning and computational systems biology.

My sincere compliments belong to all co-authors, especially Ali Faisal, Juuso Parkkinen, Seppo Virtanen, Mehmet Gönen and Eemeli Leppäaho. It has been a pleasure working with you all! During these years, I have got

to know many wonderful colleagues at the MI research group. Therefore I would like to thank all the current and former members of the group, especially Muhammad Ammad-ud-din, Elisabeth Georgii, Leo Lahti, Sohan Seth, Arto Klami, Jaakko Peltonen, Sami Remes, Antti Honkela, Tommi Suvitaival and Jussi Gillberg for fruitful discussions on science as well as life in general.

I would like to thank the pre-examiners Prof. Sampsa Hautaniemi and Prof. Manfred Claassen for their valuable comments. A special thanks is also attributed to Dr. Tero Aittokallio and his group at Institute of Molecular Medicine Finland for their interest in my work. I would also like to express my gratitude towards all the friends for the shared moments especially those at weekly tea, Aqdas Malik, Hussnain Ahmed, Rao Anwer and Adnan Ghani.

Finally, I am indebted to my parents for their consistent support and prayers throughout my doctoral studies. I would like to especially thank my mother, who along with so many other things also made my foundations in mathematics, which made related subjects easier to me throughout my career; and my father, for inculcating in me the freedom of choice. I am also extremely grateful to my brothers and sisters for the continuous support they provided, not only through the doctoral studies, but all throughout. Lastly, I wish to thank my wife and our children Mansur and Maimunah for absolutely everything.

Espoo, July 29, 2015,

Suleiman Ali Khan

Contents

List of Publications

This thesis consists of an overview and of the following publications which are referred to in the text by their Roman numerals.

I Suleiman A Khan, Ali Faisal, John P Mpindi, Juuso A Parkkinen, Tuomo Kalliokoski, Antti Poso, Olli P Kallioniemi, Krister Wennerberg and Samuel Kaski. Comprehensive data-driven analysis of the impact of chemoinformatic structure on the genome-wide biological response profiles of cancer cells to 1159 drugs. *BMC Bioinformatics*, 13:112, 2012.

II Seppo Virtanen, Arto Klami, Suleiman A Khan and Samuel Kaski. Bayesian Group Factor Analysis. In *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics AISTATS, JMLR W&CP*, 22:1269–1277, 2012.

III Suleiman A Khan, Seppo Virtanen, Olli P Kallioniemi, Krister Wennerberg, Antti Poso and Samuel Kaski. Identification of structural features in chemicals associated with cancer drug response: A systematic data-driven analysis. In *Proceedings of the Thirteenth European Conference on Computational Biology ECCB, Bioinformatics*, 30:i497–i504, 2014.

IV Mehmet Gönen, Suleiman A Khan and Samuel Kaski. Kernelized Bayesian Matrix Factorization. In *Proceedings of the Twenty-Ninth International Conference on Machine Learning ICML, JMLR W&CP*, 28: 864–872, 2012.

V Suleiman A Khan and Samuel Kaski. Bayesian Multi-View Tensor Factorization. In *Proceedings of the Seventh European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases ECML PKDD, editors T. Calders et al., Springer-Verlag Berlin Heidelberg*, 8724:656-671, 2014.

VI Suleiman A Khan, Eemeli Leppäaho and Samuel Kaski. Multi-Tensor Factorization. *Submitted to a journal*, 23 pages, 2015.

Author's Contribution

Publication I: "Comprehensive data-driven analysis of the impact of chemoinformatic structure on the genome-wide biological response profiles of cancer cells to 1159 drugs"

The ideas and experiments presented in this article were designed jointly. In particular, the author implemented the data fusion with canonical correlations and the analysis and visualization pipeline. The writing of the article was a combined effort.

Publication II: "Bayesian Group Factor Analysis"

The author designed and carried out the drug-response case study and participated in writing the article. Dr. Virtanen had the main responsibility of the model and other experiments.

Publication III: "Identification of structural features in chemicals associated with cancer drug response: A systematic data-driven analysis"

The article was jointly designed and written. The author participated in the model implementation, and implemented and carried out all the experiments and visualizations.

Publication IV: "Kernelized Bayesian Matrix Factorization"

The author designed and carried out the second drug protein interaction case study and participated in writing the article. Dr. Gönen had the

main responsibility of the model, experiments and writing of the article.

Publication V: "Bayesian Multi-View Tensor Factorization"

The author had the main responsibility for the design and implementation of the model, experiments, and for the preparation of the article. The main modelling idea was developed jointly.

Publication VI: "Multi-Tensor Factorization"

The author had the main responsibility of the initial idea, design and implementation of the MTF model, and conducted the molecular biology experiments. Mr. Leppäaho designed and implemented the relaxed formulation and conducted the neuroimaging experiments. The manuscript was written jointly.

List of Abbreviations and Symbols

Abbreviations

ARD	automatic relevance determination
ATC	anatomical therapeutic chemical classification
CCA	canonical correlation analysis
CP	canonical decomposition / parallel factor analysis
DNA	deoxyribonucleic acid
FCFP	functional connectivity fingerprints
GO	gene ontology
GSEA	gene set enrichment analysis
MCMC	Markov chain Monte Carlo
RNA	ribonucleic acid
siRNA	silencing RNA
NeRV	neighbor retrieval visualizer

Symbols

In this thesis caligraphic symbols (for example $\mathcal{X}$) are used to denote tensors, bold uppercase to signify matrices ($\mathbf{X}$), and bold lowercase column vectors ($\mathbf{x}$). Normal lowercase symbols (x) indicate scalar variables.

$\mathbb{R}$	real domain
x, y	scalar data point
$\mathbf{x}$, $\mathbf{y}$	data vectors
$\mathbf{X}$, $\mathbf{Y}$	data matrix
$\mathbf{X}^T$	transpose of matrix $\mathbf{X}$
$\mathcal{X}$, $\mathcal{Y}$	data tensor
$\mathcal{N}(\mu, \sigma^2)$	Gaussian distribution with mean μ and variance σ^2
$\mathit{Gamma}(\alpha, \beta)$	gamma distribution with shape α and rate β
$\mathit{Beta}(\alpha, \beta)$	beta distribution with shape parameters α and β
$\mathit{Bernoulli}(\pi)$	Bernoulli distribution with probability π
$p(X\|Y)$	conditional probability distribution of X given Y
$\mathbf{I}$	identity matrix
λ	scalar, regularization parameter
$\log(.)$	logarithmic function
$\circ$	outer product of two vectors
$\odot$	Khatri-Rao product of two matrices, reshaped into a tensor
$\times_k$	mode k product of a tensor and a matrix
$*$	Hadamard or element-wise product of two matrices, vectors or scalars

1. Introduction

One of the fundamental challenges for life sciences in general, and medicine in particular, is to understand how chemical entities in food, medicines, and environment affect living cells. This is difficult because the chemicals are expected to produce multiple and overwhelmingly diverse responses. Genome-wide measurements of these responses such as the gene expression, can provide insights into the system-level understanding and action mechanisms [1, 2]. With millions of available compounds and the multitude of responses of cells, understanding the effects of drugs is a challenging yet significantly important goal. A key challenge here is to learn the complex dependencies between the structures and their genome-wide responses. These dependencies can assist in better understanding the effects of drugs, as well as to make predictions of effects for previously unseen molecules. Therefore, if uncovered, these dependencies may assist in significantly enhancing optimization efforts on design, re-purposing and personalized applications of the drugs.

Traditionally, cellular response data has been scarce and has hosted only a few samples. These limited measurements have been only partially relevant as they present a view of the response to specific conditions only. With advancements in high-throughput response measurement techniques, data sets are being profiled that measure genome-wide effects of chemical perturbations on multiple specialized cells [3]. This knowledge offers the opportunity to study the diverse responses, over several different cell-types in a systematic fashion.

One plausible direction is to harness the existing measurement data and apply computational approaches to predict the relationships between structural aspects of the drugs and the observed systems-wide biological responses. On the computational front, machine learning is playing an important role in developing methods to analyze data. Lately, Bayesian

machine learning has become increasingly valuable allowing principled means of handling uncertainty in the model parameters and incorporation of the prior probabilities, which make it possible to study data even with small sample sizes. In this work, generative latent variable models enable building representations that explain the processes that have generated the observed data [4, 5]. By discovering the descriptions of these underlying processes, one can potentially understand them or produce the required predictions.

When the task is to search for relationships between two or more variables, machine learning approaches that model dependencies among variables fit the goal directly. These methods seek co-occurring patterns that relate the variables of interest. Component-based dependency models are additionally able to segregate multiple distinct dependencies existing in the data. For example, factor analysis is a commonly used method that captures dependencies between variables into distinct factors, with each factor capturing different dependencies [6, 7]. Methods such as the canonical correlation analysis [CCA; 8, 9] identify dependencies between two data sets, decomposing them into distinct components. Recent advances in Bayesian factor analysis and canonical correlation analysis have made them practically applicable in many real-world scenarios [10, 11].

The central hypothesis of this dissertation is that component-based dependency modeling can be used to systematically study the systems-wide responses of cells to drugs and their structures, uncovering the links between them in a data-driven fashion. Here the basic assumption is that the common statistical patterns existing between the datasets represent underlying mechanistic processes. These processes are hypothesized to be informative for the action mechanisms of the drugs, i.e. the biochemical interactions through which a drug produces its pharmacological effect. Analyzing the responses of multiple types of cells brings up the requirement for new multi-source methods. Moreover, even more structured methods are needed for studying hypothesis that emerge when responses are measured at several levels of phenotypic details. This thesis starts by adapting canonical correlations for the structure-response modeling to test the basic premise. It then presents several novel multi-source models that remove some of the limitations of existing methods and applies them to novel drug response analysis problems.

Typically, the action mechanisms of drugs have been studied computationally using chemical properties of the drugs or biological properties

of the specific targets [12]. Approaches such as the quantitative structure-activity analysis (QSAR) commonly use structural properties of the drugs to predict their biological activity. For example, the classical QSAR work by Cramer et al. [13] predicted a single biological activity from structural features of the drugs. However, recent evidence suggests the importance of tailoring the drugs for multiple targets to enhance their efficacy [14]. Moreover, drugs interacting with several targets may also produce toxicity and other side effects. Therefore, balancing the efficacy vs. toxicity is a crucially important problem. On the other hand, genome-wide responses to drugs have also been successfully used in explaining the mechanisms of drug actions [15, 16]. However, to the best of our knowledge, the systematic integration of these complementary approaches, to study structure-response relationships with component-based dependency models, has not been explored earlier.

Therefore, this research is positioned at the intersection of three different fields: machine learning, bioinformatics, and chemoinformatics.

1.1 Contributions and Organization of the thesis

This thesis presents several novel multi-view dependency modeling methods and explores their applicability for the data-driven decomposition of drug structure-response links. The applications utilize the drug response interactions for both understanding the underlying processes and the prediction of unseen responses.

Publication I adapts the existing canonical correlation model family, hypothesizing the latent components as descriptions of underlying drug response mechanisms. This formulation extends the drug response analysis from standard QSAR, which relates drug properties with their univariate responses, to finding relationships between structural descriptors of the drugs and their genome-wide responses. The paper demonstrates that component-based dependency modeling can successfully capture structure-related drug response patterns in a data-driven fashion.

Publications II and III generalize the analysis to multiple diseases, with novel multi-view model formulations. Specifically, the group factor analysis (GFA) model carries out a data-driven search for relationships between the chemical properties of the drug molecules and their disease-specific biological response profiles. The relationships governed by distinct underlying biological processes are segregated automatically by the

sparse factorization. This ability of GFA makes it possible to distinguish between responses that are disease-specific from those common to all the disease types analyzed. A focused experiment was carried out to study this in Publication II. When exposed to genome-wide data sets, a critical challenge for algorithms is the high dimensionality and small sample size of the data. The small number of samples leave considerable uncertainty in the data, and prioritizing the interacting features becomes a major concern. In Publication III, feature-level sparsity is applied in GFA to cater for these concerns, and a detailed structure-response analysis is performed. The study demonstrated that data-driven modeling of drug responses in multiple diseases can be informative for identifying established and novel structure-response links, as well as for exploring disease-specific action mechanisms. Such systematic large-scale studies can also identify new repositioning opportunities for existing drugs. If validated with wet lab experiments, the approach may also open up the opportunity for drug designers to tailor drug molecules based on the genome-wide responses of existing drugs.

Publication IV addresses the analogous problem of predicting drug targets from a multi-source formulation. The paper advances the state of the art by concluding that protein targets of the drugs can be better predicted when multiple types of descriptions of the drugs are used as side information sources.

Publication V introduces a new problem formulation and presents a model called Bayesian multi-view tensor factorization (BMTF) for solving it. BMTF learns dependencies between multiple co-occurring tensors comprehensively, decomposing them into a set of underlying factors that can be shared between some or all of the tensors. This paper advances the state of the art by making it possible to jointly factorize the multi-view tensor data sets. In short, there is now a model similar to Bayesian CCA but for tensors. The method is applied to decompose drug responses of multiple diseases at different levels of phenotypic detail, helping application experts to construct targeted hypotheses for the underlying processes generating the data.

Publication VI presents a new model for joint factorization of multiple matrices and tensors, coined multi-tensor factorization (MTF). The method generalizes the matrix-tensor factorization to novel Bayesian multi-view settings as well as factorization of arbitrary sets of tensors. The model factorizes multiple matrices and tensors collectively, allowing in-

vestigation of relationships that may exist between some or all of them. The new model also addresses several key practical issues in a principled fashion. The Bayesian formulation provides improved performance in real-world applications with small samples and large dimensions of the matrix-tensor data sets. The paper finally presents the decomposition of structurally driven responses of multiple diseases when integrating responses from various levels of phenotypic detail. Thereby, demonstrating that the method can assist application scientists to explore and predict drug response mechanisms, by integrating the data sets of choice.

Chapter 2 discusses the matrix and tensor factorization methods that form the basis of this thesis. In particular, it presents the relevant assumptions of these approaches that are harnessed for both the applications and machine learning advancements. Chapter 3 discusses the biological responses to drugs, the measurement data sets, and the structural properties of chemical compounds. Chapters 4 and 5 present the main contributions of the thesis, describing the different multi-source models while increasing the amount of structure they capture as the thesis progresses. The chapters simultaneously discuss the corresponding novel drug response analysis that is made possible by the advanced methods. Chapter 6 concludes the thesis and suggests directions for further research.

2. Bayesian machine learning and factorizations

Machine learning focuses on algorithms that search for patterns in the data and extract useful information [5]. The algorithms learn a model from existing data samples and utilize it for various tasks. In several applications, the data measurements can be hypothesized to have been generated from an underlying process that may not be measurable itself [17]. Moreover, the measurements are typically high-dimensional and may contain correlated variables which are additionally corrupted with noise, making the direct analysis complicated. One of the key features of machine learning is to learn low-dimensional latent representations that summarize the data. The learned data summaries can then be hypothesized as the descriptions of the phenomenon that have generated the data. These summaries are then commonly used to understand the mechanisms of the data generation process, or then to predict the unseen outcomes. A salient latent summarization techniques is data factorization that forms the core of all the methods developed in this thesis. This chapter describes data factorization approaches, identifying the key assumptions and the latest developments.

2.1 Latent variable models

Latent variable models are a powerful approach to machine learning [4, 18]. They provide a flexible way of describing dependencies between the data variables by assuming that the observed data was generated through the interactions of a few unobserved (latent) variables. These latent variables present a low-dimensional summary of the observed data and can be considered as concise and denoised descriptions of the underlying processes that have generated the data. The representations can then be used to understand the data generation processes or predict the unob-

served data entities.

Formally, the higher-dimensional data variables $\mathbf{x} \in \mathbb{R}^D$ can be represented using lower-dimensional latent variables $\mathbf{z} = \{z_1, z_2, \dots z_K\}$ where $K < D$, providing a flexible way to represent the dependencies between the observed variables. The number of latent variables is typically much smaller than the data dimensionality and each z_k usually follows a simple distribution.

The matrix and tensor factorization methods discussed in the following sections are key examples of latent variable models that find low-dimensional factors (latent variables) to represent the high-dimensional observed data.

2.2 Matrix factorization and co-occurring patterns

Matrix factorization (MF) is a well-established approach having vast scientific applicability including missing value prediction, dimensionality reduction, and data visualization [19–21]. Among other motivations, matrix factorization can be seen as a means of identifying underlying processes that produced the data. In such a scheme, measurements are thought to have been generated from a combination of multiple latent processes, each generating some parts of the data. For matrix factorization, the task is to decompose a matrix into several *factors* or *components* that describe the underlying (hopefully meaningful) processes. The terms factors and components are used interchangeably in this thesis, as both have been commonly used in the literature.

A broad set of approaches has been studied to factorize matrices, optimizing on different criteria [22, 23]. For factorization of a single matrix *factor analysis*, and for joint factorization of two matrices *canonical correlation analysis*, are well-established methods. Both factor analysis and canonical correlation analysis form the basis of all the methods developed and presented in this thesis, and are described next.

Factor Analysis (FA) [6] is an unsupervised technique for low-dimensional factorization of a single matrix. FA assumes that the data matrix $\mathbf{X} \in \mathcal{R}^{\mathcal{N}\times\mathcal{D}}$ can be modeled by a latent factor representation such that the factors capture dependencies between the variables. For N samples, each

described by a D-dimensional data vector $\mathbf{x}_n$, FA can be represented as

$$\begin{aligned} \mathbf{x}_n &\sim N(\mathbf{W}\mathbf{z}_n, \Sigma) \\ \mathbf{z}_n &\sim N(0, \mathbf{I}) \ , \end{aligned} \tag{2.1}$$

where $\mathbf{W}\mathbf{z}_n$ is a low-dimensional latent representation of $\mathbf{x}_n$, while Σ is a diagonal noise covariance matrix, with a separate term $\sigma_1, \ldots, \sigma_D$ for each of the D variables. The noise model captures the individual variation of each variable, and the latent representation $\mathbf{W}\mathbf{z}_n$ models the covariance patterns *between* the variables. This assumption is of key importance as it allows FA to capture patterns common between two or more variables. In contrast, principal component analysis [PCA; 20] assumes the more restrictive isotropic noise model with all the variables having a single variance parameter.

Bayesian factor analysis has been used successfully in modeling the factors from real data applications. For example, in modeling genomic data sets, the low-dimensional factors are used to represent the biological processes driving the mechanisms. Recently, several studies have been conducted using microarray gene expression data sets to create hypotheses for the cellular response patterns [7, 10, 24]. Similarly, learning the denoised low-rank structure with factorization has also been proved useful for the prediction of missing values [19, 25].

Canonical Correlation Analysis (CCA) [8] is an unsupervised method that decomposes two paired matrices into a shared low-dimensional representation. The paired matrices are characterized by having a common identity of the samples. Therefore, unlike FA that finds dependencies between two or more variables, CCA aims to capture the correlated patterns between two matrices. CCA linearly transforms the matrices into a maximally correlated subspace of components, such that any two components are uncorrelated with each other. This way, it can find distinct components that are common to both of the matrices. For two data vectors $\mathbf{x}_n^{(1)}$ and $\mathbf{x}_n^{(2)}$, CCA can be represented as a generative process [11, 26]:

$$\begin{aligned} \mathbf{x}_n^{(m)} &\sim N(\mathbf{W}^{(m)}\mathbf{z}_n, \Psi^{(m)}) \qquad m = 1, 2 \\ \mathbf{z}_n &\sim N(0, \mathbf{I}) \ , \end{aligned} \tag{2.2}$$

where $\mathbf{z}_n$ is the latent vector common to both matrices, $\mathbf{W}^{(m)}$ are loadings for each matrix, and $\Psi^{(m)}$ the corresponding noise covariance matrix. The shared latent representation $\mathbf{z}_n$ of CCA models the covariation patterns between the two matrices, while $\Psi^{(m)}$ models the variation specific

to each matrix. This division implies that $\mathbf{z}_n$ can capture the dependencies between the two matrices. An efficient CCA solution using group-sparse priors was recently presented by [11]. Their formulation uses the latent variables to represent both the correlated patterns between the matrices as well as the matrix specific variation, while the noise covariance is assumed isotropic for each of the data sets.

For a comprehensive review of Bayesian canonical correlation analysis see [11], while [9] for classical CCA. CCA has been successfully used for modeling dependencies between data sets. For example, in genomics it has been used to identify chromosomal regions showing dependencies in copy number and gene expression of a set of samples [27, 28].

2.3 Tensor factorizations

When data sets have more dimensions than just samples and features, they may present tensorial relationships. While matrix factorization methods are optimized to decompose matrices, the structure in more than two modes can be handled appropriately by tensor factorizations. In order to capture the more structured patterns of such data sets and avoid overfitting, tensor methods use more constrained formulations that have fewer parameters than their matrix counterparts.

Analogous to MF, tensor factorizations perform decompositions of the tensors into their constituent parts. However, the decompositions of multimode data sets allow factorizations with several different interaction assumptions and present additional modeling issues. Consequently, a wide range of low-dimensional representations of tensors have been proposed in the literature [29]. The most well-studied models include the CANDECOMP/PARAFAC [30, 31] and the Tucker model family [32]. For a comprehensive review of tensor factorizations and their properties, see [29].

Tensor factorizations have obtained significant success in chemometrics and psychometrics in the last decade. Recently, they have also been adopted in bioinformatics and related application fields. For example, tensor methods have been used in exploring factors of gene expression patterns over replicates of several stimuli [33], as well as integrating responses from different studies [34, 35]. The rapid accumulation of biological measurement data is leading to new and extended hypotheses, which may be investigated with tensor formulations. Very recently, gene

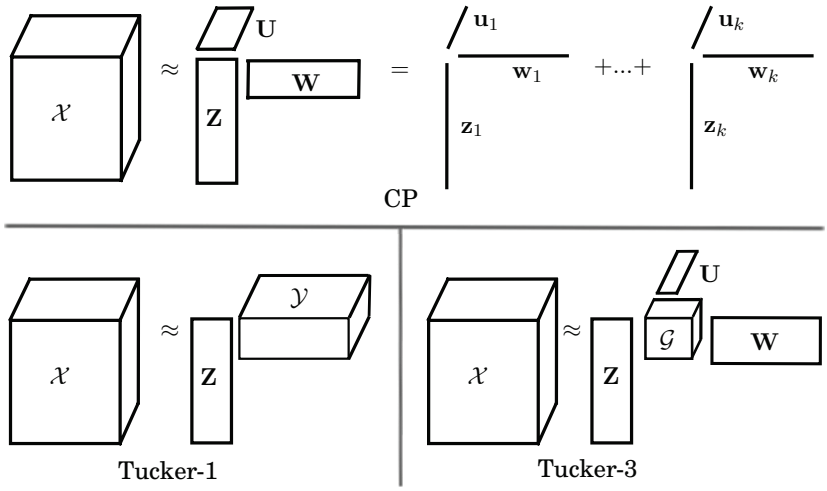


Figure 2.1. Top: CANDECOMP/PARAFAC (CP) factorizes a tensor $\mathcal{X}$ into K components in each mode and is equivalent to sum of rank one component tensors. **Bottom:** Tucker-1 and Tucker-3 decompositions that factorize the tensors in one and three modes, respectively. Tucker-3 models the interactions between components via a core tensor $\mathcal{G}$.

expression time series of several samples [36] and integrative analyses of metabolic and gene expression networks [37] have also been explored with tensor factorizations.

2.3.1 CANDECOMP/PARAFAC (CP)

The CP decomposition was presented independently as canonical decomposition (CANDECOMP by [30]) and parallel factor analysis (PARAFAC [31, 38]). CP is a direct extension of matrix factorization to higher order data sets. The decomposition originates from Cattell's theory of parallel proportional profiles [39]. Cattell stated that two independent realizations of similar data sets could be decomposed jointly with a simultaneous factor analysis. This process learns a common projection matrix that differs only in the scale of factors for the two data sets and captures the intrinsic axis of the underlying factors. Parallel factor analysis [31] extends this conceptualization to a tensor (a set of matrices placed together), allowing the projections of FA to differ only in their scales in the third mode. CP can therefore be seen as multiple factor analysis of a single phenomenon being performed simultaneously, to infer the empirically meaningful factors.

The CP decomposition is defined in a symmetric way to factorize a tensor into a sum of rank-one tensors, where each rank one tensor is the

outer product of vector loadings in all modes (Figure 2.1-top). For a third-order tensor $\mathcal{X} \in \mathbb{R}^{N \times D \times L}$, a rank-$K$ CP is represented as:

$$\mathcal{X} = \sum_{k=1}^{K} \mathbf{z}_k \circ \mathbf{w}_k \circ \mathbf{u}_k + \boldsymbol{\epsilon} \,, \tag{2.3}$$

where $\mathbf{Z} \in \mathbb{R}^{N \times K}$ and $\mathbf{U} \in \mathbb{R}^{L \times K}$ and $\mathbf{W} \in \mathbb{R}^{D \times K}$ are the latent variables corresponding to the three modes.

This component decomposition is valuable for many applications, as the different rank-1 terms can be related to separate mechanisms that may have contributed to the higher order tensor, making the model readily interpretable. Furthermore, the CP factorization solutions are known to be unique up to a given permutation and scaling [40]. These characteristics of CP, make it a suitable choice for the underlying tensor factorization in the methods presented in this thesis.

Commonly, CP is solved for a given K via Alternating Least Squares (ALS) algorithm [41, 42]. ALS finds a locally optimal CP solution by updating one loading matrix at a time while keeping others fixed [29]. It estimates the parameters by a least squares approach that is equivalent to maximum likelihood estimate under the assumption of isotropic Gaussian noise. Recently, several authors have also solved CP in a Bayesian setting demonstrating its advantage [43–45].

Rank. Unlike matrix factorizations, in CP the exact rank determination can be crucially important as both under and over-estimation may result in invalid solutions [29]. The primary reason is that a rank-k CP solution is not guaranteed to be the best rank-k approximation if the true number of factors is larger than k [46]. On the other hand, the factorization could produce artificial splits or noise components if k is set larger than the actual value. Therefore, the best rank-k solution can not be computed sequentially, rather all the factors must be found simultaneously [29]. Determining the CP rank is a challenging problem and solutions based on cross-validation tend to be computationally expensive [47]. Recently, a Bayesian solution for automatic rank identification has been proposed [44], which is also computationally fast, however, its robustness could be further studied.

Degeneracy. Practical application of CP can occasionally suffer from degenerate solutions, in which two or more components become highly correlated in all the modes and some of the loading values becoming arbitrarily large. The degenerate solutions are not interpretable and compli-

cate the use of CP [48]. These degeneracies have primarily been observed when data sets having a non-trilinear structure are decomposed with CP [49, 50]. Nevertheless, the strong interpretative power of CP has brought up degeneracies as an important research topic [51–53]. As high negative correlations are a primary characteristic of degeneracies, several researchers have recently studied solutions with strict constraints that force uncorrelatedness, such as orthogonality and non-negativity [52, 53].

2.3.2 Tucker

The Tucker model family [32] defines several levels of factorization, and has three main forms, Tucker-1, Tucker-2, and Tucker-3. The 1-mode factorization of Tucker, is the most relaxed formulation that decomposes only one of the modes while Tucker-2 and Tucker-3 factorize two and all three of the modes, respectively. Tucker-1 being the most flexible is also equivalent to matrix factorization of a matricized tensor (Figure 2.1, bottom-left). On the other hand, Tucker-3 enforces more structure and is characterized by interactions between a different set of factors in each mode (Figure 2.1, bottom-right). Unlike in CP, in the Tucker-3 model, a factor does not represent an additive source of information; rather, it represents a pattern of variation in a given mode only, and the factors are thought to have generated the data by interacting with several other patterns of variation (factors),

$$\mathcal{X} \approx \sum_{p=1}^{P} \sum_{q=1}^{Q} \sum_{r=1}^{R} g_{p,q,r} \mathbf{z}_p \circ \mathbf{w}_q \circ \mathbf{u}_r \ .$$

The loadings $\mathbf{Z} \in \mathbb{R}^{N \times P}$, $\mathbf{W} \in \mathbb{R}^{D \times Q}$ and $\mathbf{U} \in \mathbb{R}^{L \times R}$ are accompanied with a core tensor $\mathcal{G} \in \mathbb{R}^{P \times Q \times R}$ that captures the interactions between factors of different modes (Figure 2.1, bottom-right). The complex interaction of factors via $\mathcal{G}$ makes interpretation of factors difficult. Moreover, in contrast to the CP factorization, the Tucker model is not guaranteed to provide unique solutions. The Tucker component matrices can be rotated, and the core tensor $\mathcal{G}$ counter-rotated to obtain infinite number of models with an equal fit. This rotational ambiguity requires additional constraints to make the solutions interpretable [54, 55]. However, CP can be seen as a restricted version of Tucker-3 with the core tensor $\mathcal{G}$ constrained to be a superdiagonal and of size $K \times K \times K$. While Tucker better fits complex structures, the CP outperforms it when the data contains trilinear relationships [56].

Table 2.1. The novel methods applied or developed in this thesis are based on matrix or tensor factorizations as summarized in this table. The matrix factorization methods of Publications I-IV are discussed in Chapter 4 while tensor factorization methods of Publications V-VI are described in Chapter 5.

Pub.	Matrix Factorization	Tensor Factorization
I	Canonical Correlation Analysis	
II	Group Factor Analysis	
III	Sparse Group Factor Analysis	
IV	Kernelized Bayesian Matrix Factorization	
V		Bayesian Multi-view Tensor Factorization
VI	Multi-tensor Factorization	Multi-tensor Factorization

This thesis utilizes and develops novel extensions of matrix and tensor factorization methods, when multiple data sets are factorized together. The methods are detailed in Chapters 4 and 5, however, a summary of their matrix or tensor nature is presented in Table 2.1.

2.4 Bayesian modeling

Bayesian modeling is being increasingly used in modern machine learning research. The modeling task is to define a model and learn the unknown parameters (or latent variables) using the observed data. The Bayesian data analysis models the uncertainty observed in the true value of an unknown parameter using a probability distribution. This principled means of representing uncertainty is especially advantageous, when the data samples are few and noisy. Therefore, the salient feature of the Bayesian analysis is that when uncertainty exists, it learns a posterior distribution specifying a range of values for the parameter, while taking into account the prior information.

The posterior distribution of the parameter is defined by the likelihood and the prior. The likelihood function determines the impact of data on the parameter while the prior probability distribution encodes any prior information concerning the parameter. The Bayes theorem gives this for parameter θ and data $\mathbf{X}$ as

$$P(\theta|\mathbf{X}) = \frac{P(\mathbf{X}|\theta)P(\theta)}{P(\mathbf{X})},$$

where the marginal $P(\mathbf{X})$ is the normalization constant integrating the posterior to one. The posterior probabilities $P(\theta|\mathbf{X})$ can be seen as likelihood $P(\mathbf{X}|\theta)$ weighted by the prior $P(\theta)$, where the prior increases or decreases the impact of likelihood in the posterior. In other words, the modeling favors the solutions matching the prior. The prior may be used to encode evidence or expert knowledge into the model. Moreover, it can even assume a 'non-informative' setting, which enforces minimal assumptions so that the data can guide the posterior.

In this thesis, priors are used for inducing structure-revealing patterns and regularizing the solutions by giving higher probability to simpler representations. The regularization is especially beneficial when the sample count is low, and the level of noise is high.

2.5 Inference and Gibbs sampling

Once the model is specified, the learning task is to compute the posterior distribution of the latent variables $\mathbf{z}$ and the parameters θ. However, except for only the simplest models, there may not exist closed-form analytical solutions for the posterior evaluation [5]. In such cases, inference of the parameters is performed via approximate schemes.

The approximate methods fall into two broad categories, deterministic and sampling approaches [5]. The deterministic methods use analytical approximations of the posterior distribution. For example, they may assume that the posterior factorizes into simpler distributions. Quite contrary, the sampling methods work by sampling from the joint posterior and have the property that given infinite computational resources they can generate the exact result. Sampling methods are considered approximate because only a finite number of samples can be obtained practically. The deterministic approaches such as the variational inference are usually suitable for simple and smooth distributions, and are faster than sampling methods. On the other hand, as sampling methods obtain the samples from the actual posterior distribution, they are usually more suitable for complex distributions. Gibbs sampling is one of the most widely used sampling algorithm and is described in the rest of this section. It was used in Publications III, V and VI of this thesis.

Gibbs sampling is a widely applicable approximation algorithm [5]. It simulates observations as approximations of the joint posterior distribution without directly sampling from the joint distribution itself. The

method iterates the samples from the conditional posterior of each variable while conditioned on the existing estimates of the remaining variables and the data,

$$z_i^t \sim p(z_i|\mathbf{z}_{-i}^{t-1}, \mathbf{X}, \theta).$$

Here the latent variables $\mathbf{z}$ are split into k sub-parts $\mathbf{z} = (z_1, z_2, ...z_k)$, and t indicates the current iteration number of the sampler. In every iteration, each of the k sub-parts is sampled sequentially based on the latest updates of all other parts, improving the simulation at each step. The algorithm is a Markov chain Monte Carlo (MCMC) method, where successively sampling from the conditionals converges to the stationary distribution, which, in this case, is the joint posterior distribution [57]. The samples from the stationary distribution can then be used to approximate the joint posterior distribution, the marginal distribution of any of the variables or for computing the expected values of the variables.

In practice, as the chain iterates towards convergence, the initial set of samples prior to the stationary distribution are discarded as the *burn in*. The samples from successive draws of the posterior may be strongly correlated, in which case *thinning* may be used to break the dependence by keeping only the every l^{th} sample [57]. The $\mathbf{z}^{t=0}$ can be initialized in several ways with sampling from the priors being a common choice. The effect of this random initialization and the stochastic nature of the chains can be partly mitigated by using multiple chains for averaging or selecting a reliable one. Therefore, with its ease of formulation, Gibbs sampling is a simple, reliable and a well studied standard choice.

3. Biological responses to drugs

Cellular response to drugs depends on several factors including the multitude of targets the drug can bind and the resulting pathways perturbed. It is also well understood that there are potentially several pathways connected with each cellular phenotype. This presents a many-to-many relationship from drugs to targets and from pathways to responses, with the majority of this information being unknown. Given this incomplete knowledge, comprehensively modeling the drug responses over a diverse drug library is a goal yet to be achieved. However, as drugs bind to the targets yielding the effects, the *binding* is of central importance to the response. This binding of the drug-target pair depends primarily on the structural correspondence of the drug molecule and the binding cavity of the target. Therefore, in principle, the responses to drugs can be modeled comprehensively by learning a mapping between the structural properties of the drugs and their responses, if done on a genome-wide scale.

In this chapter, the response measurements and structural properties of the drugs used in the thesis are presented briefly. Interested readers are referred to more details and the Publications in each section.

3.1 Gene expression measurements

Cellular proteins are the functional blocks executing the mechanisms in every cell. In order to learn the actions or reactions of the cells, it would be ideal to measure the complete protein expression of the cells. However, with the current state of the art, reliable and cost effective techniques are not available and, therefore, approximate alternatives are commonly used. The most widely used means of estimating the type of activity in a cell is the expression of all the genes in a particular condition. These patterns, referred to as *gene expression*, are well known for their ability

to differentiate between the different types [58] and the behavior of cells [59]. Studying the changes in gene expression has been valuable in understanding different medical conditions [60], disease processes [61], and otherwise to explore the therapeutic applications of drugs [62]. For example in cancers, it is valuable to identify the survival-related genes whose expressions are altered by a drug [63].

Microarrays [64] are extensively used for measuring genome-wide gene expressions. They measure the expression of thousands of short predefined sequences, which are then preprocessed to obtain gene-level activity. These measurements can be significantly noisy [65], and require preprocessing and corrections prior to their practical use [66]. More recently, RNA sequencing has emerged as a technique with wider coverage for measuring gene expression responses [67, 68]. However, this thesis focuses on microarray gene expression data, due to the availability of large-scale drug profiling data sets [3].

A common way of analyzing the processed gene expression data is *differential expression*, where the expression of interest is compared with a control signal. This procedure builds the notion of directional changes in expression. For example, comparing post and pre-treatment expression of genes can identify which genes have been *up-regulated* (increased in expression) and which *down-regulated*, as a result of a treatment. Differential expression is usually computed as fold-change, i.e., a $\log_2$-ratio between treatment and control. In order to determine which genes have been differentially expressed, standard statistical tests such as the t-test are commonly used to test significances [69]. A large number of genes is usually tested for the significance, and multiple hypothesis correction needs to be carried out. For corrections, the standard methods are the Bonferroni correction and the false discovery rate [70, 71]. With advancements in machine learning, it has also become common to study differential gene expression values of several case-control samples directly with computational methods, to test and identify different hypotheses [15, 72]. It is also informative to analyze expression changes in light of prior biological knowledge. In this line, Gene Set Enrichment Analysis [GSEA; 73] and Gene Ontology enrichments [GO; 74] of known biological pathways and processes have gained significant success. In short, gene expression responses form a valuable genome-wide resource to study cellular responses, in particular, drug action mechanisms.

This thesis utilizes the drug-treatment responses from the Connec-

Table 3.1. Data set usage in the publications of this thesis. The data sets are obtained from the CMap and the NCI60 databases, and also the structural descriptors of the drugs. The GSEA and Gene expression present two different representations of the CMap data. The toxicity data set from NCI60 and different types of structural descriptors are described in Sections 3.2 and 3.3, respectively.

	Data sets		
Publication	CMap	NCI60	Structural descriptors
I	GSEA		VolSurf
II	GSEA		VolSurf
III	Gene expression		FCFP + Pentacle
IV			VolSurf + Pentacle
V	Gene expression	Toxicity	
VI	Gene expression	Toxicity	FCFP

tivity Map database (CMap, [3]) that hosts microarray gene expression responses of several cancer cell lines to over 1300 drugs. The CMap database has been successfully used by the scientific community, for example, to discover the mode of action of drugs [16] as well as new biological links between multiple drugs [15]. The three major cell lines used in CMap, HL60, MCF7, and PC3 come from three different cancers, namely blood, breast and prostate cancers. Therefore, this data presents a unique view of the genome-wide responses of cancers to drugs, over a set of 11,000 genes that are measured for each sample. The data set contains $\sim$7000 gene expression measurement samples including replicates. The measurements come from three different microarray platforms, with over 85% of the samples profiled with one of them, the Affymetrix HT-HGU133A chip, and was therefore used in this thesis. The data set contains both post and pre-treatment cellular responses of genes and is preprocessed to obtain a treatment vs. control differential gene expression response for each drug-cell pair (further details are given in Publication I). A positive value in differential expression corresponds to up-regulation of the genes while negative corresponds to down-regulation. Gene set summaries were also computed as a dimensionality reduced representation of each sample using GSEA (Publications I and II, Section 4.1). The CMap data set was used in all publications of this thesis except Publication IV as enlisted in Table 3.1.

3.2 Drug sensitivity measurements

The drug sensitivity analysis aims to identify the concentration threshold of a drug required for a particular pharmacological action. The process is usually conducted by administering multiple different doses of a drug to a cell and measuring the cell viability or inhibition, resulting in a dose-response curve. The dose-response curves can then be used directly in the analysis or summarized in one or more standardized values such as the GI50 (growth inhibition of 50%). Drug profiling presents a phenotypic view of a cell's response to a drug, and can also be used to describe toxic outcomes depending upon the application [75, 76]. The drug sensitivity measurements have gained significant attention in studying responses of cancer cells, where a set of drugs are tested over a set of cell lines to identify sensitive and resistance behaviors [77, 78].

Large scale drug profiling has led to the possibility of not only identifying sensitivity or toxicity patterns of the drugs [79], but to also search for genomic markers that may be indicative of the responses [80, 81]. Several studies have recently been conducted to measure drug sensitivity of human cancer cell lines against large collections of drugs, along with genomic profiling of the cells to identify potential biomarkers [78, 82]. On a complementary front, drug treatment gene expression data has also recently been shown as a better classifier of drug toxicity than chemical descriptors of the drugs [83].

The NCI60 is the largest panel of drug sensitivity data measured with over 40,000 compounds tested on 60 different cancer cell lines [77]. The data set has been used in several studies, for example to establish the mechanisms of action of the compounds [84, 85] and to predict their cytotoxicity [75, 86]. In this thesis, dose-response summaries are obtained from NCI60 for all the common drugs and cell lines that have corresponding drug treatment gene expression measurements in CMap [3]. The data set presents three response measures, the GI50 (growth inhibition of 50%), LC50 (50% lethal concentration) and TGI (total growth inhibition). The measurements are then preprocessed to obtain a positive value if the concentration of the drug used in the Connectivity Map was toxic (higher than the dose-response values), and negative otherwise. The transformed data set therefore represents concentration-dependent toxicity values and has been used in Publications V and VI of the thesis (Table 3.1).

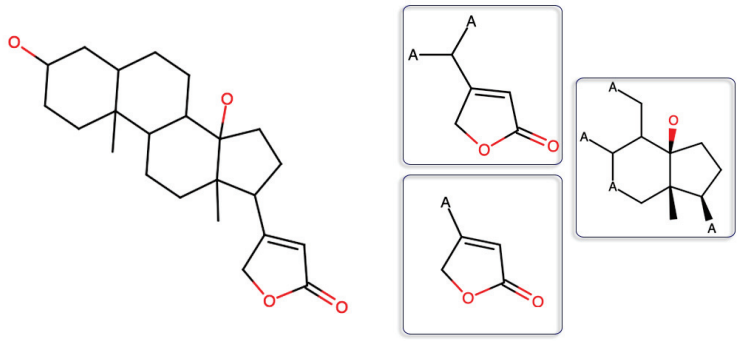


Figure 3.1. 2D Fingerprints of an example compound (digitoxigenin). **Left**: The 2D structure of digitoxigenin is shown. **Right**: The 2D fingerprints represent the compound as structural fragments. In this example, Functional Connectivity Fingerprints (FCFP4, Section 3.3.1) are used to compute the structural fragments of digitoxigenin. A total of 34 fragments are found in digitoxigenin, 3 of which are shown here.

3.3 Structural descriptors

Structural descriptors of drugs represent each compound with a set of features of its structure and function. Different types of 2D and 3D descriptors exist, each optimized for different criteria [87]. The selection of appropriate types of descriptors is important, especially in this work as the aim is to represent drug molecules in a way that assists capturing the biological functions. Some of the key descriptor types are elaborated and reasoned in the rest of this section. One or more of these descriptor sets are used in each publication of this thesis, except Publication V as enumerated in Table 3.1.

3.3.1 2D fingerprints

The 2-dimensional fingerprints represent the structural and functional properties of drugs by formulating the presence or absence of the fragments, allowing easy visual inspection (Figure 3.1). MACCS fingerprints are one of the traditional examples of 2D fingerprints that represent drugs with a predefined set of 166 structural fragments [88]. The Functional Connectivity Fingerprints [FCFP; 89]) are an advanced formulation of the 2D circular topological fingerprints. They have been designed specifically for modeling of structure-activity relationships and similarities between the drug, and, therefore, have been a descriptor of choice in such

studies [90–92]. FCFP can represent variation in novel structures as they compute the fragments dynamically, instead of using a predefined set. Therefore, FCFP can be used for identifying 2D substructures that make novel compounds structurally similar and are responsible for biological activity.

3.3.2 3D descriptors

Amongst a host of 3D descriptors, those based on molecular interaction fields (MIFs) are well suited for structure-activity modeling and drug discovery [93]. Instead of describing the molecules structural skeleton, these MIFs take a complementary approach and sketch out the interaction potential of a compound with a set of different probes. For a compound, this could potentially be used to characterize interaction sites around the molecule or summarize the different physicochemical, pharmacophoric, and shape-based properties that are relevant for its pharmaceutical application. When analyzing multiple compounds with MIFs, extracting the relevant information, and the alignment of molecules have traditionally been the major bottlenecks. However, recent advancements such as the VolSurf and Pentacle have substantially overcome these.

VolSurf

The 3-dimensional VolSurf descriptors [94] are based on molecular interaction fields and are specifically designed for optimization of pharmacokinetic properties. Unlike the 2D fingerprints, which represent the structural fragments, VolSurf describes the generic physicochemical properties of the molecules. They are, therefore, capable of grouping together compounds that have different chemical structure yet possess the same type of chemical properties. VolSurf summarizes the interaction contours from MIF by computing shape and volume-related statistics into a set of predefined descriptors. These include interactions such as lipophilic, lipophobic, hydrophilic, and hydrophobic, resulting in a set of 76 structural features.

Pentacle

The Pentacle descriptors are advanced 3-dimensional field distance descriptors [95] that also use molecular interaction fields to capture the functional properties. They encode the detailed interaction potential of the compounds with the chemical probes, at dynamically computed markers. These interaction potentials represent the molecule (its charge and

shape) as the target molecule would see it. Therefore, Pentacle descriptors can group together compounds with different chemical structures, but having similar interaction potential with a receptor. This characteristic is important in the applications of this thesis, as it allows grouping together compounds that bind into the same binding pocket despite structural differences, which traditional fingerprints are unable to recognize. In comparison to VolSurf, Pentacle descriptors are more detailed and describe pharmacophore features extensively. The dimensionality of the resulting Pentacle descriptors depends on the size of the molecules and is usually much higher than VolSurf. Very recently, both VolSurf and Pentacle have shown success in structure-activity analysis [96–98].

4. Multi-view models for drug responses

In modern chemical systems biology, the high-throughput profiling of a compendium of compounds against genomically heterogeneous cancer cell lines has unveiled diverse and interesting relationships between the responses. Several computational approaches have used massively high-dimensional genome-wide responses to investigate the drug action mechanisms [12, 15, 16, 99]. On the other hand, various studies use the chemical properties of the drugs to predict their univariate effects [97, 100]. To combine these complementary approaches, this research formulates the hypothesis that drug structure-response relationships can be studied on the systems-wide level in a data-driven manner. Such systematic studies are valuable as they may ultimately assist in improving drug design and personalizing treatments. Specifically, this chapter pursues machine learning methods to model the relationships between structural descriptors of drugs and their corresponding genome-wide responses, over multiple types of cancer.

The first aim here is to study the effects of chemical structure in the context of *multidimensional* biological readouts, rather than to use a single target 'activity' as in the conventional QSAR studies [13]. When the goal is to search for relationships between two multivariate data sets in an unbiased fashion, dependency modeling based approaches such as canonical correlation analysis (CCA) match the objective directly. Section 4.1 uses canonical correlation analysis (CCA) on the largest genome-wide drug profiling data resource (CMap, section 3.1) and a set of chemical features, in search of correlated patterns (Publication I). The study formulates the structure-response problem in a dependency modeling framework and presents a systematic pipeline for analysis of the genome-wide responses.

The results demonstrate that systematic dependency modeling suc-

cessfully identifies both known and novel insights; and also opens up newer horizons to be explored. The study also suggested that multiple diseases may produce response patterns that may be partly specific to only one or few of the diseases (cancer types). Such response patterns, if existing, may be discovered through multi-source modeling. Section 4.2 presents a novel multi-view model, and applies in collaboration to explore patterns between drug structures and their responses over multiple diseases (Publications II and III). The results illustrate the advantages of the improved methodology. Finally, multi-source modeling is also evaluated for prediction of drug targets in section 4.3 (Publication IV).

4.1 Canonical correlation analysis for drug responses

The structure-response relationships can be learned comprehensively by *searching for links* between drug structures and their genome-wide responses. This translates to the machine learning task of identifying latent interactions between multivariate structural descriptors of drugs and their corresponding multivariate (genome-wide) responses. Publication I proposes a data-driven solution to model the hidden interactions in a systematic manner. The interactions are hypothesized to describe the underlying biological processes, and are identified by discovering the joint low-dimensional subspace that relates drug structures with cellular responses.

The analysis is based on the differential gene expression response of 1159 drugs originating from the Connectivity Map database [3] (section 3.1). In order to reduce the dimensionality of this gene expression data and to bring in prior knowledge of known biological responses, Gene Set Enrichment Analysis (GSEA) [73] was performed. For GSEA, the curated gene sets (C2) from the Molecular Signatures Database[1] [73] were selected for enrichment, as they are both carefully curated and cover over 90% of the genes in this data. These include gene sets from online pathway databases and knowledge of domain experts. A few example gene sets are the *HSA05221_acute_myeloid_leukemia (AML)* that lists known genes involved in AML, *ET743_Sarcoma_up* enumerates genes up-regulated in sarcoma's as a result of treatment with Trabectedin (an anti-cancer drug), and the *Oxiadative_Phosphorylation* which is a metabolic pathway. The GSEA summarizes the gene-level expression into a gene set activation

[1] http://www.broadinstitute.org/gsea/msigdb/collections.jsp

profile. As a result, the 11,327 genes from CMap were summarized into 1321 gene sets. On the other hand, the chemical space of the 1159 drugs was represented by the 3D VolSurf descriptors (section 3.3.2). VolSurf comprises of 76 descriptors that are capable of grouping together compounds having the same type of chemical properties despite having different chemical structures.

4.1.1 Dependency modeling via canonical correlations

For discovering relationships in a data-driven way, the computational task is to model the dependencies between the two data sets (structural descriptors and gene set responses) in a comprehensive fashion. CCA, fitting the goal directly, discovers a joint low-dimensional representation that splits the two input spaces (chemical and biological) into distinct components. Each of the components statistically correlates the patterns of one space (chemical descriptors) with those of the other space (biological response). As drugs generate diverse effects, component-based models are suitable for segregating the multiple different responses. Therefore, with CCA, the joint decomposition of the structure-response data sets results into correlated components, which are hypothesized to describe the underlying biological phenomena driving cancer drug response. The key assumption here is that if there is any statistical dependency between the patterns in the chemical space and the biological responses, these patterns are then informative for developing hypotheses about the mechanisms of drug actions.

Methodologically, CCA is an established data integration approach that maximally explains the dependency between two data sets [8]. The method linearly projects the data sets to obtain a maximally correlated low-dimensional representation. This low-dimensional representation aka the *shared space* or the *components* capture the statistically shared patterns between the two data sets, whereas patterns specific to any one of the data set are considered noise and ignored. This split matches exactly to the assumption that shared structure-response patterns are of prime interest.

Analogous to most genomics data sets, the CMap gene set profiles have more dimensions than the number of samples with several highly correlated variables. In such a situation, there is a potential for the CCA covariance matrices to become ill-conditioned, which could result in numerical inaccuracies while computing the inverse. This is a classical prob-

lem when $D > N$, and regularized solutions are available [101–103].

Given two data sets $\mathbf{X} \in \mathcal{R}^{N \times D_1}$ and $\mathbf{Y} \in \mathcal{R}^{N \times D_2}$, with N paired occurrences of samples in the two views, regularized CCA finds K linear projections[2] of the data sets, $\mathbf{X}\mathbf{w}_k$ and $\mathbf{Y}\mathbf{v}_k$, such that their correlation P_k is maximized as,

$$\begin{aligned} P_k &= \underset{\mathbf{w}_k, \mathbf{v}_k}{\arg\max}\, \mathbf{cor}(\mathbf{X}\mathbf{w}_k, \mathbf{Y}\mathbf{v}_k) \\ &= \underset{\mathbf{w}_k, \mathbf{v}_k}{\arg\max}\, \frac{\mathbf{w}_k^T \mathbf{C_{xy}} \mathbf{v}_k}{\sqrt{\mathbf{w}_k^T \mathbf{C_{xx}} \mathbf{w}_k + L_1 \|w_k\|^2} . \sqrt{\mathbf{v}_k^T \mathbf{C_{yy}} \mathbf{v}_k + L_2 \|v_k\|^2}}. \end{aligned} \tag{4.1}$$

The vectors $\mathbf{w}_k$ and $\mathbf{v}_k$ are the *projection weights*, or *loadings* when normalized, while the projected space of data sets $\mathbf{X}\mathbf{w}_k$ and $\mathbf{Y}\mathbf{v}_k$ constitutes the *CCA components* or *canonical covariates*, capturing the shared patterns between the two data sets. The first component $k = 1$ is found such that the correlation P_k is the largest possible. All other components $k = 2, 3...$ are computed analogously, but with the additional constraint that they are uncorrelated with the previously obtained components. These constraints ensure that the first K components capture the strongest shared effects, which are also distinct from each other. The regularization replaces the empirical covariances $\mathbf{C_{xx}}$ and $\mathbf{C_{yy}}$ by their regularized estimates $\mathbf{C_{xx}} + L_1\mathbf{I}$ and $\mathbf{C_{yy}} + L_2\mathbf{I}$, and also acts as a penalizer on the projection weights $L_1\|w_k\|^2$, $L_2\|v_k\|^2$, preferring a simpler solution.

Once found, the CCA components decomposing the data sets (structural descriptors and gene set responses) into components are represented by the sum of the projections $\mathbf{Z} = \mathbf{X}\mathbf{w}_k + \mathbf{Y}\mathbf{v}_k$. The most prominent samples (drugs) that represent the component are those having the largest values in $|\mathbf{Z}|$. The positive and negative scores can be thought to represent two distinct sets of samples (drugs) that present opposing behaviors and should be analyzed separately. The canonical loadings with high correlation values identify the most important structural features in a component that share the patterns with corresponding gene set responses. Therefore, a CCA component k can be represented by a set of samples (drugs) that have specific features (structural properties) activating the biological response in particular gene sets.

[2]In Publication I the symbol S was used to denote components, while K in all the rest. In the interest of consistency, K is used throughout the thesis.

4.1.2 Drug structure-response relationships

To validate if the structure-response relationships could be modeled with CCA, the discovered components were subjected to quantitative and qualitative validations. Quantitative validation was done by evaluating the model's ability to retrieve similar drugs when queried with a single drug. The retrieval performance was measured as the mean average precision of retrieving similar compounds, with similarity measured over an independent drug-protein target data set compiled from several publicly available databases. To test the model for its ability to extract relevant information, the retrieval experiment was performed on both the integrated data space and the individual biological or chemical data spaces separately. The results confirmed that the integrated space is more informative of drug response than any of the data sets separately, verifying that the CCA components extracted biologically meaningful signals from the data. In this respect, the conclusions of this work can also be useful as a means of better characterizing the drug molecules based on joint modeling of their structure and response profiles.

For qualitative validation and exploration of the novel links, the learned model parameters were studied with novel visualization schemes and gene ontology enrichments for interpretations. The resulting interpretations and visualizations brought out new links of interest to biologists, as briefly summarized next.

The top 10 components having significant correlations were examined for detailed interpretation and are presented in Publication I, while an overview and most significant findings are elaborated here. Here each component represented the relationships between structural features and biological responses that were triggered by a set of drugs (as described in sec 4.1.1). The key drugs in each component were analyzed for the common mechanisms and the drug classes to which they belong, revealing several known drug groups such as HDAC inhibitors, cardiac glycosides, and protein synthesis inhibitors, being enriched in the components. The biological action captured by each component was identified using the most strongly correlated gene sets. In several components, the biological response was recognized as the known effects of the drug groups confirming them to be in line with established biological knowledge. The strongly correlated chemical features in each component were found to be representing particular VolSurf properties that were linked, by the mo-

del, to the biological response. In summary, the components were verified to represent well-established findings while revealing novel relationships between the structural properties and the biological responses.

Several of these 10 components also discovered novel and interesting responses. For example, components 2B and 10A both were found to represent two separate aspects of DNA damage response, connected to two separate molecular features, potentially indicative of different mechanisms. The key drugs of these two components were then examined for toxic indications using the NCI60 database [77], confirming that most of them had been administered at the toxic doses in CMap.

Component 3A also revealed a potentially interesting response, relating to mitochondrial and metabolic stress related processes. The component was therefore further investigated for functional transitions caused by drug responses. To this end, a non-linear dimensionality reduction tool called Neighbor Retrieval Visualizer [NeRV; 104, 105] was used. The visualization using NeRV mapped the component non-linearly onto a 2-dimensional display of the reference states, such that similarities were preserved as faithfully as possible. The reference states were represented using 30 independent and untreated breast cancer cell lines denoting different molecular subtypes. Interestingly, the NeRV-visualization of the component (3A) revealed that the captured response of DNA-damaging drugs made the cells resemble therapy-resistant cancer cell lines. This transition indicates that the regulation of these metabolic genes was potentially a protective or resistance-mediating response.

Since there is emerging interest in the involvement of metabolism in cancer and drug resistance [106–108], the expression of component 3A genes was further explored for its therapeutic relevance. To this end, initial wet lab experiments were followed up in collaboration, which indicated that siRNA-mediated silencing of 14 of these genes (including metabolic genes such as *HADHA*, *IDH3B*, *ME2*, *PDAP1* and *UQCRC1*) did result in chemosensitization of cells to the used DNA damaging agents [109], giving confidence to the discovered hypothesis.

To sum up, decomposing structure-response interactions on a genome-wide scale can be solved via dependency modeling methods such as CCA, generating hypothesis for unexplored polypharmacology. The approach is flexible and can be extended to model dependencies from other types of chemical descriptors and genome-wide biological responses. However, there were limitations of the study. The study was constrained to a limited

set of descriptors and the dimensionality-reduced GSEA profiles. Though the analysis indicated that the use of GSEA was not detrimental to the performance of the method, gene interpretability would become easier by directly using gene-level data in the modeling process. The more detailed gene-level data and chemical descriptors were incorporated in section 4.2.2, by using a Bayesian solution that can better handle uncertainty when the data dimensionality is high.

4.2 Group factor analysis for drug responses

The cellular responses to a drug are complex in nature and depend on a multitude of characteristics. One of the most important aspects is the type of the cancer. Specially, different cancers are heterogeneous in nature and respond selectively to the drugs. This selectiveness makes it valuable to learn which of the responses are specific to a cancer type and which common across several of them. This section extends the challenge of structure-response modeling to multiple different diseases (cancer types) and addresses the question: *Can systematic modeling distinguish between structurally driven responses common to multiple cancers from those specific to one or few*. First, in section 4.2.1 a targeted study designed for structure-response relationships over multiple cancers is presented, with a new multi-source dependency modeling method called *Group Factor Analysis* (Publication II). The study confirms the hypothesis that modeling dependencies between multiple-cancer responses from more than two paired data sets (also known as '*views*') via multi-view models is feasible and promising. An extension of the method for high-dimensional data is then presented, and a detailed analysis of the multi-structure multi-cancer response is followed (section 4.2.2, Publication III).

4.2.1 Dependency modeling via group factor analysis

The dependency modeling task now becomes the identification of common patterns from multiple paired data sets, as the problem extends to responses of multiple types of cancers. Existing multi-set extensions of CCA formulate the task for more than two data sets [110, 111]; however, they only model components common to all views, and hence are unable to identify patterns shared by only a subset of the views. Publication II presents a novel latent component model *Group Factor Analysis* (GFA)

that not only extracts the statistical dependencies between all the data sets but also identifies dependencies between any subset of them.

Group Factor Analysis (GFA) is a model designed to capture relationships (statistical dependencies) by reducing the collection of data sets (views) into a combined set of low-dimensional factors (components). A component can be active in one or more of the data views, representing that it captures the underlying relationships between the corresponding views only. For example, a component active in all the views captures the shared dependency structure between all the views while one active in a single view identifies variation and features specific to that particular view only. GFA learns the activity of the components in a data-driven manner making it possible to identify dependencies that exist between a subset of the views. In the structure-response decomposition problem, each of the factors can be thought to represent a distinct underlying biological process that has generated parts of one or more of the data sets. Therefore, the task of GFA here is to factorize the collection of structures and multiple responses while separating the components capturing the structure-biology relationships from the rest.

Formally, given a collection of M data sets $\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, ...\mathbf{X}^{(M)} \in \mathcal{R}^{N \times D_m}$, having N paired samples (drugs), and a separate set of dimensions D_m in each view, GFA searches for a K-dimensional matrix factorization for the entire collection. The model is formulated in a Bayesian setting as a product of a Gaussian latent component matrix $\mathbf{Z} \in \mathcal{R}^{N \times K}$ containing the K components, and a projection weight matrix $\mathbf{W}^{(m)} \in \mathcal{R}^{D_m \times K}$ for each data set m. The model is represented as

$$
\begin{aligned}
\mathbf{x}_n^{(m)} &\sim \mathcal{N}\left(\mathbf{W}^{(m)}\mathbf{z}_n, \mathbf{I}(\tau^{(m)})^{-1}\right) \\
\mathbf{z}_n &\sim \mathcal{N}(0, I) \\
\mathbf{w}_{:,k}^{(m)} &\sim \mathcal{N}\left(0, (\alpha_k^{(m)})^{-1}\right) \\
\alpha_k^{(m)} &\sim Gamma(a^\alpha, b^\alpha) \\
\tau^{(m)} &\sim Gamma(a^\tau, b^\tau) \ ,
\end{aligned}
\tag{4.2}
$$

where $\tau^{(m)}$ denotes the view-specific noise precision. The latent variables $\mathbf{z}_n$ are common between all the views, representing the response patterns. GFA solves the joint decomposition problem using the *group-wise sparse* matrix factorization of all data sets, where each data set is considered a group. The group-wise projections $\mathbf{W}^{(m)}$ capture both group-specific variation (activity seen exclusively in one view), and dependencies between the groups (activity in more than one views). This is achieved by model-

ing the total variation in the data while constrained by the group-sparse prior. The group-sparse prior (via $\alpha_k^{(m)}$) controls the scale of the projection weights $\mathbf{w}_{:,k}^{(m)}$ for each of the component-view pairs. Higher values of $\alpha_k^{(m)}$ shrink the corresponding $\mathbf{w}_{:,k}^{(m)}$ towards zero switching the component off, while smaller values of $\alpha_k^{(m)}$ increase the scale of $\mathbf{w}_{:,k}^{(m)}$, making the component active in the view. GFA learns the $\alpha_k^{(m)}$ in a data-driven way, yielding dependency patterns between the views. As a practical step, $\alpha_k^{(m)}$ is thresholded with respect to the captured variance to obtain active and non-active status of the components.

GFA is applied on the four co-occurring data sets of the chemical systems biology problem. The chemical space is represented by the 3D VolSurf descriptors as in section 4.1; however, the biological space is represented by three views, formed by gene set activation profiles of the three different cancer types [CMap; 3]. The application goal here is to discover factors that describe interesting statistical dependencies between the views. The exact question is: *Does there exist any relationship between drug structures and their biological responses, and can we differentiate between responses that are disease-specific from those which are generic to all three subtypes of cancer*.

To validate if multi-source structure existed and was captured by the model the component activity of GFA was examined. The activities revealed that indeed the model discovered several components shared between descriptors and biological responses. Components of both types were found: (i) components shared by chemical structures and a few of the cell lines giving hypotheses for specific cancer variants, and (ii) components shared by chemical structures and all of the cell lines constituting responses common to all cancer subtypes. The first components of both types were examined by collaborators and found in accordance with established biological knowledge.

To validate the model quantitatively, for the task of discovering biologically meaningful components, its ability to retrieve similar drugs was evaluated on an independent drug-target data set. The mean average precision of retrieving the drugs having similar known targets was found to be significantly higher for the integrated component space of the model, than the chemical descriptors and the three gene set profile spaces separately. The result confirms that the model extracted biologically meaningful information.

Given the heterogeneity of available high-throughput biological data,

the application though limited in its interpretation did demonstrate the model to be useful in identifying relevant and biologically meaningful components. The subsequent section extends the model with sparsity priors and presents a detailed structure-response analysis using genome-wide *gene expression profiles* of the three cancer types, when coupled with *multiple* detailed and extensive sets of structural descriptors.

4.2.2 Bayesian sparse group factor analysis

With plausible results in section 4.2.1, an extended model is now proposed for the massively high-dimensional data while simultaneously increasing the interpretability of the components by enforcing sparse priors. The extended model is applied to the genome-wide screening assay data of three cancers from CMap and two sets of advanced chemical descriptors, to investigate the cancer-specific and across-cancer effects of drug structures. The gene-level expression data of 11,327 genes is incorporated directly into the model, instead of the 1321 gene set responses (used in section 4.1 and 4.2.1). This detailed information makes it possible to link the structural features directly to the activity of the genes and is expected to make components easier to interpret. Additionally, the advanced structural features may extract novel as well as known links of interest.

The formulation of GFA in eqn. 4.2 uses a group-wise sparse prior for determining the view activity. The specification is sparse at the level of data sets but dense at the level of individual features. Feature-level sparsity can better regularize the solution while also aiding easier interpretability. The introduction of feature-level sparsity in the model is non-trivial since the model is already group-wise sparse. This is overcome by introducing layered sparsity: group-level sparsity followed by a sub-layer of feature-level sparsity. The group sparsity (view activation) is now encoded by formulating a group spike and slab prior [112], and is represented by binary variables that control the activity of the k^{th} component for group m. The second layer of sparsity is encoded on the features within a group via an element-wise automatic relevance determination prior [ARD; 113]. The ARD enforced on the projection weight matrices pushes individual weight values of irrelevant variables towards zero, inducing each of the active components to become feature-wise sparse. For-

mally, the sparse GFA model is represented as

$$\begin{aligned}
\mathbf{x}_n^{(m)} &\sim \mathcal{N}\left(\mathbf{W}^{(m)}\mathbf{z}_n, \mathbf{I}(\tau^{(m)})^{-1}\right) \\
\mathbf{z}_n &\sim \mathcal{N}(0, I) \\
w_{d,k}^{(m)} &\sim h_k^{(m)}\mathcal{N}\left(0, (\alpha_{d,k}^{(m)})^{-1}\right) + (1 - h_k^{(m)})\delta_0 \\
h_k^{(m)} &\sim Bernoulli(\pi_k) \\
\pi_k &\sim Beta(a^\pi, b^\pi) \\
\alpha_{d,k}^{(m)} &\sim Gamma(a^\alpha, b^\alpha) \\
\tau^{(m)} &\sim Gamma(a^\tau, b^\tau)\ ,
\end{aligned} \tag{4.3}$$

where the $h_k^{(m)}$ are binary variables controlling group sparsity. For the view-component pair shut down by the binary variables ($h_k^{(m)}$=0), the delta function δ_0 forces zero on all the corresponding weights $\mathbf{w}_{:,k}^{(m)}$. Whereas for the active view-component pairs ($h_k^{(m)}$=1), the weights are sampled from the element-wise ARD $\alpha_{d,k}^{(m)}$ that brings in the second layer of sparsity, but for each of the d features. This formulation allows both the entire components as well as features within active components, to be switched off. A Gibbs sampler was implemented to perform model inference, and an open source implementation of sparse GFA is provided at `http://research.ics.aalto.fi/mi/software/GFAsparse`.

GFA decomposes the data sets into components that are shared between one, more, or all of the views. This decomposition results in several exciting possibilities of view-shared components. First, the components shared between chemical and all biological views hypothesize for the structurally driven responses that are common across all of the analyzed cancer-types. Second, the components that are shared between chemical and one (or few) of the biological views present cancer-specific hypotheses. Finally, the remaining components that are specific to the biological views or the chemistry views do not represent structure-response relationships.

A component k active in at least one chemical and one biological view encodes the cross-space relationships and can be analyzed as follows. The weight vector $\mathbf{w}_{:,k}^{(m)}$ represents the contribution of each feature into the k^{th} component for identifying the linked descriptors and genes. The descriptor-gene relationships are observed prominently in the most significant drugs of the component characterized by the highest magnitude scores of $\mathbf{z}_{:,k}$.

4.2.3 Multi-response relationships

The multi-response relationships between responses of multiple cancer types and multiple structural descriptors is now studied with sparse group factor analysis. Specifically, the multi-cancer relationships are explored between the gene expression responses of three different cancers and two sets of advanced structural descriptors. The two types of advanced structural descriptors are the 3D *Pentacle* and the detailed though more traditional functional connectivity fingerprints *FCFP*. The Pentacle can capture molecular field similarities between drugs corresponding to potential biological functional similarity despite the drugs being structurally different (section 3.3.2). The FCFP are more traditionally used fingerprints for structure-response analysis and capture similarity between drugs that can be directly related to presence or absence of structural fragments (section 3.3.1). While FCFP provide a direct mechanism of identifying links between structural fragments and response, the Pentacle adds in the potential for novel drug associations. For the biological response, the 11,327 treatment vs. control differential gene expression responses of the three different cancer cell lines are used as three separate data views. The gene expression data coming from the Connectivity Map is obtained over the common set of 682 drugs that were profiled on all the three cancer cell lines. GFA run on the $M = 5$ data sets with $K = 80$, discovers 11 components of interest, i.e. shared between one or more structural descriptors and one or more biological responses. These components form hypotheses for structure-driven responses of drugs indicative for both the cancer-specific variants and those common across all subtypes of cancer.

To validate the model's informativeness, it was quantitatively evaluated for discovering biologically and chemically meaningful components. This was done by studying the components against the Chemical Entities of Biological Interest Ontology [ChEBI; 114]. The ChEBI is currently the largest curated ontology of small molecule drugs that classifies the compounds with respect to their chemical structure, biological roles they are known to play, and their applications.

The GFA components shared between chemical and biological views were, therefore, hypothesized to be in line with the known relationships encoded in ChEBI. To test this, the similarity of top compounds from the components was computed, as indicated by ChEBI regarding their structural and functional relationships (data external to the model). This is

compared to the corresponding similarities from (i) random choice of compounds, (ii) drug clusters learned for elicitation of action mechanisms using the CMap biological response data by [15], and (iii) structure-response components of Publication I.

The result clearly showed that the shared GFA components significantly outperformed the random baseline, concluding that the components were strongly in line with the known structure-role relationships of ChEBI. Additionally, the GFA based components fared much better than both of the comparison approaches endorsing the advantage of the method.

All the 11 shared components were also qualitatively explored seeking structural and functional similarities. Each component was hypothesized to describe an underlying process that explained the relationships between chemical descriptors and biological responses. The analysis identified three major results, as described next.

First, the components shared between the FCFP4 descriptors and one or more cell lines identified the core 2D structural groups that were linked with the response, creating a structure-response map. As key examples, these included (i) the common Steroid Backbone of all the Cardenolides in Component 1 that is linked to the DNA Damage Response of the drugs, at the high concentrations at which the drugs were administered; and (ii) the well known aromatic ring of the HDAC inhibitors in Component 2 was linked with the corresponding inhibition response of the drugs. These structure-response relationships demonstrate that the methodology correctly identifies the structural groups responsible for gene-level response and, therefore, can be used as a principled tool for unsupervised exploration and mining of structure-response relationships.

Second, 3 out of the 11 components showed cancer-specific responses, entirely missed by the key earlier studies Publication I and [15]. The most salient one was the leukemia-specific response of corticosteroids, which indicated that the drug targets may be selectively active in leukemia cells. The same component also revealed that simvastatin and repaglinide, which are two structurally very dissimilar drugs, shared the same response pattern of corticosteroids. This is an important finding with potential therapeutic implications in the light that recently lovastatin, which is a close structural analog of simvastatin, was shown to have anti-cancer activity in leukemic stem cells [115].

Third, the use of advanced Pentacle descriptors with the structure-

response linking capability of GFA, allowed discovery of novel drug associations. Though several examples are presented in Publication III, the most prominent one is that of the 15 Delta Prostaglandin J2's (PGJ2) role as an HSP90 inhibitor drug. HSP90 inhibitors are well known for anti-cancer activity, and their mechanism of action is also well-understood [116]. PGJ2 and its analogs have also been studied for their anti-cancer activity though their mechanisms of action have not been revealed earlier. The analysis suggests that PGJ2 and HSP90 inhibitors despite being structurally very dissimilar, share specific Pentacle field properties that are linked in HSP90 inhibitor Geldanamycin to its binding with the HSP90 protein. This chemical similarity coupled with a similar response to the drugs on HSP related genes gives an indication that the PGJ2 may be targeting HSP90. PubChem drug-target data reveals that PGJ2-HSP pair falls in untested/unknown category as of yet, though the novel analysis is indicating a partly common mechanism of action.

The results demonstrate that GFA can discover meaningful relationships between compounds and cell line expression that were not found using simpler methods. Specifically, with multi-source modeling, it was successfully possible to separate responses common to cancer subtypes from those specific to a cancer variant. The model found both known and novel structure-response links, leading to potentially valuable drug indications. The quantitative validation confirmed that the analysis is well in line with the established structure-biology links. In summary, the practical applicability of GFA is demonstrated to analyze the relationships between differential gene expression data and structural properties of drugs. This is an important problem because a successful method of understanding the dependencies between the structures and responses of specific diseases would enable additional optimization efforts for drug designs and precision medicine.

4.3 Multi-source prediction of targets

The previous two sections discussed modeling the relationships between structural properties of drugs and their biological function represented by gene expression responses. The drug-target binding was not explicitly modeled due to massively missing amounts of information. To explore the potential of multi-source modeling in predicting the targets of drugs, a study is carried out and presented next (Publication IV). The task is mod-

eled as factorization of the drug-target matrix aided with side information of the drugs (multiple types of structural descriptors) to better predict the interactions. The analysis demonstrates that integrating multiple data sets (as side information sources) improves the predictions.

4.3.1 Kernelized Bayesian matrix factorization

Publication IV presents Kernelized Bayesian Matrix Factorization (KBMF) algorithm for factorizing a matrix by leveraging additional information from side data-sources. KBMF performs the low-dimensional factorization, such that the factors are learned using *multiple* side data-sources in a non-linear fashion. To capture the non-linear effects, the side data sets are encoded using Kernels, and their low-dimensional projections are combined via Multiple Kernel Learning (MKL). Therefore, KBMF learns the factorization using a weighted combination of the multiple side data sets while learning the weights in a data-driven fashion.

4.3.2 Multi-structure prediction

The drug-target predictions were studied using multiple structural descriptors of the drugs and the drug-protein interaction network data from [117]. The data set comprises target information of 800 proteins over 855 drugs, with 4659 experimentally validated interactions. The side information for drugs is represented by two types of structural descriptors, the VolSurf and the Pentacle (also called Amanda).

The task of KBMF then is to use the multi-view structural properties of the drugs and predict the missing interactions in the drug-target matrix. To model the non-linear relationships, Gaussian Kernels were computed on both data sets representing them as side information Kernels. KBMF employs MKL to utilize the two descriptor data sets simultaneously, for learning the drug-target interactions. The model is first validated for predicting the interactions in comparison to existing non-Bayesian alternatives as measured by the AUC (area under the ROC curve). The experiment is performed in a cross-validation setting while varying the number of components and confirms that KBMF outperforms the comparison methods.

Having confidence that the model outperforms the existing state of the art, the particular multi-source hypothesis was tested. Therefore, to determine the benefit gained by the weighted combination of multi-

ple side data sets, the model was compared to the cases, (i) when a single side data set is used, or (ii) both the side data sets are concatenated in a single view. The results confirm that the principled handling of multiple views via weighted combination of MKL results in the best prediction performance (AUC), validating the hypothesis that multi-source models can be used to predict drug-target interactions better. Very recently, a closely related multi-source multi-task variant of KBMF has been used to predict drug responses, outperforming 40 other methods in the crowdsourced NCI-Dream drug sensitivity challenge [118].

The model's ability to retrieve similar drugs is also explored. Specifically, the use of multiple side data sets is evaluated for its prospects in retrieving similar drugs. Here the hypothesis is that the subspace useful for prediction of drug-target interactions could be valuable for drug classifications as well. The performance is measured as the mean average precision of retrieving similar drugs when evaluated over an independent validation data set. Specifically, the therapeutic classification of drugs are used as an independent validation data set, to measure the drug similarity. The retrieval performance of the model with multiple descriptors combined via MKL is compared to the performance of each descriptor individually as well as when concatenated in a single view. The results validate that the drug-target prediction space supplemented with the multi-view descriptors outperforms the alternatives. The finding implies that KBMF can also be used as a metric learning method for drug similarities.

5. Multi-tensor factorizations for drug responses

When data sets are measured over multiple variables of interest, the underlying structure may present multi-way relationships. In such cases, tensor methods can preserve the natural structure in the data, and obtain more compact and accurate representations in an efficient way. Tensor factorizations have recently gained much interest in the machine learning community to address some of the challenges posed by multi-way data sets [43, 44, 119]. With heterogeneous and partially paired multi-way data sets emerging, it is of prime importance to have integrative tensor methods that can jointly factorize several tensorial data sets. As a key application problem, different types of biological response measurements under multiple experimental conditions require multi-way methods to study the underlying structure appropriately. This chapter aims to contribute to this growing area of research by proposing the first Bayesian multi-source tensor factorization methods.

In the following, novel tensor integration methods are introduced. Section 5.1 presents a Bayesian multi-view tensor factorization method that finds a combined low-dimensional representation of multiple paired tensors (Publication V). The method is applied as the first Bayesian tensor canonical correlation analysis to a novel toxicogenomics application of integrating drug-disease responses. Section 5.2 extends the formulation to allow coupled matrices and tensors, coining *Multi-tensor factorization* (Publication VI). Finally, the methods are demonstrated in use for both interpretability of underlying processes and predictions, on an extended formulation of the drug-response decomposition problem (section 5.3). The novel application sits at the juncture of toxicity, chemistry and bioinformatics; and serves as an example of the novel formulations that can be solved with the new multi-source multi-mode factorization methods. Nevertheless, the new methods developed in this chapter are

generic and applicable in any field of science. For notational simplicity, the case of third-order tensors is presented.

5.1 Bayesian multi-view tensor factorization

A novel model coined *Bayesian multi-view tensor factorization* (BMTF) is presented that jointly factorizes multiple co-occurring tensors into a low-dimensional and interpretable representation. Specifically, a novel multi-view tensor factorization problem is formulated, and a Bayesian model is presented as a solution. The multi-view tensor factorization problem is introduced as joint factorization of paired tensors to learn a concise set of multi-way factors determining the dependencies between the data sets, in a data-driven fashion. This unique formulation poses three essential modeling choices to be made, as discussed next.

First, the multi-way nature of tensors allows them to be factorized in several different forms. The most widely used factorization approaches such as the CANDECOMP/PARAFAC (CP) and the Tucker model family were discussed in section 2.3. The CP factorization decomposes a tensor as a sum of rank-1 tensors and has the advantage of being readily interpretable similar to matrix factor analysis. It is based on the principle of parallel proportional profiles [39], which allows the intrinsic axes to be discovered automatically, providing unique solutions under mild assumptions [31, 40]. These properties make CP a valuable choice for BMTF. The CP decomposes a tensor $\mathcal{X} \in \mathcal{R}^{N \times D \times L}$ into three constituent loading matrices $\mathbf{Z} \in \mathcal{R}^{N \times K}$, $\mathbf{W} \in \mathcal{R}^{D \times K}$, $\mathbf{U} \in \mathcal{R}^{L \times K}$. Therefore, the joint CP factorization of multiple paired tensors can be formulated by assuming at least one of the loading matrices (for example $\mathbf{Z}$) to be common across all the tensors. The joint loading matrix makes it possible to capture dependencies between the data sets.

Second, the joint factorization can be formulated considering the tensor structure from two different perspectives. First, a tensor can consist of a vector of samples, where each sample has two set of dimensions (i.e. a matrix), forming a three-way tensor. This construction is analogous to a set of slabs placed on top of each other where each slab is a single sample. In a multi-view case, as both dimensions of each view could be different, this setting corresponds to pairing between the views in the single sample mode as done by [120] for a non-probabilistic version of tensor CCA. Alternatively, a tensor may be made up of a matrix of samples where each

sample is a vector, forming a three-way tensor. This case is more similar to the traditional matrix case, where the samples are vectorial in nature (analogous to a fiber), with the difference that samples now come from multiple covariates. In multi-view case, this corresponds to pairing in the two sample modes of the tensor. While none of the existing tensor CCA methods formulate the problem in the later fashion, it corresponds well to the natural extension of most problems (including biological) that are currently being analyzed via matrix methods. In BMTF, the pairing in two modes is achieved by assuming the loadings in $\mathbf{U}$ and $\mathbf{Z}$ to be common across all the tensors.

Third, the factorization should contain components that can be shared by any subset of the tensors, i.e. one, some, or all the tensors. This generalization is necessary to learn all the possible dependencies between the views, as well as to differentiate amongst dependencies that exist between two or more views from those specific to only one. This flexibility can be achieved by modeling the total variation in the tensors, such that a set of view-specific loadings $\mathbf{W}^{(m)}$ control which of the patterns from $\mathbf{Z}$ and $\mathbf{U}$ are active in each view. Patterns active in more than one view are shared between the views while those active in only one are specific to it. BMTF enforces group-sparsity on $\mathbf{W}^{(m)}$ making it possible to learn the exact nature and sharing of the factors automatically from the data. In addition, BMTF can determine the total number of components that represent each view, as well as the entire collection.

Formally, the BMTF model for multiple ($m = 1 : M$) paired tensors $\mathcal{X}^{(m)} \in \mathcal{R}^{N \times L \times D^{(m)}}$ is formulated, to learn a low-dimensional (K) representation explaining the entire collection. Assuming normal distributions

and conjugate priors, the Bayesian model is constructed[1] as

$$
\begin{aligned}
\mathbf{x}_{n,l}^{(m)} &\sim \mathcal{N}\left(\mathbf{W}^{(m)}(\mathbf{z}_n * \mathbf{u}_l), \mathbf{I}(\tau^{(m)})^{-1}\right) \\
\mathbf{z}_n &\sim \mathcal{N}(0, I) \\
u_{l,k} &\sim \mathcal{N}\left(0, (\beta_{l,k})^{-1}\right) \\
w_{d,k}^{(m)} &\sim h_k^{(m)} \mathcal{N}\left(0, (\alpha_{d,k}^{(m)})^{-1}\right) + (1 - h_k^{(m)})\delta_0 \\
h_k^{(m)} &\sim Bernoulli(\pi_k) \\
\pi_k &\sim Beta(a^\pi, b^\pi) \\
\beta_{l,k} &\sim Gamma(a^\beta, b^\beta) \\
\alpha_{d,k}^{(m)} &\sim Gamma(a^\alpha, b^\alpha) \\
\tau^{(m)} &\sim Gamma(a^\tau, b^\tau) \ .
\end{aligned}
$$

The latent variables $\mathbf{Z}$ and $\mathbf{U}$ are common to all the tensors and capture the underlying patterns while the $\mathbf{W}^{(m)}$ translate the patterns for each tensor. A two layered sparsity formulation is used to learn the required structure. The first, *group-sparsity* controls the component-view activation via binary variables $h_k^{(m)}$ inducing a spike and slab prior [112]. The $h_k^{(m)}$ is automatically learned from the data, such that a value of $h_k^{(m)} = 1$ corresponds to the component being active ($\mathbf{w}_{:,k}^{(m)} \in \mathcal{R}^{1 \times D_m}$) while $h_k^{(m)} = 0$ means not-active ($\mathbf{w}_{:,k}^{(m)} \in 0^{1 \times D_m}$). As an example, a component active in two views has $h_k^{(m)} = 1$ for the corresponding two views, while $h_k^{(m)} = 0$ for the rest. Therefore, $h_k^{(m)}$ learns the sharing structure between the tensors. A component active in two or more views captures a shared response pattern while those active in only one represent variation specific to that tensor. The model also learns the total number of active components in the data collection by switching all the extra $h_k^{(1:M)} = 0$, given K is initialized to a large enough value.

The *feature-level* sparsity is induced via the automatic relevance determination prior [ARD; 113]. The ARD shrinks the individual loadings of $\mathbf{W}^{(m)}$ and $\mathbf{U}$ via $\alpha_{d,k}^{(m)}$ and $\beta_{l,k}$. This formulation regularizes the solution as well as assists in overcoming degenerate solutions. A Gibbs sampler was implemented to perform the model inference, and an open source implementation of BMTF is provided at `http://research.ics.aalto.fi/mi/software/BMTF`.

The approach is general enough to include sparse Bayesian CP (CAN-

[1] Publication V uses slightly different notations, with the vector, matrix or tensor nature of a symbol being defined entirely by its subscripts. Here we use the given notation in the interest of consistency throughout the thesis.

DECOMP/PARAFAC) factorization as a special case when only one tensor is factorized. In this case, the formulation performs comparable to the state of the art CP solutions and outperforms them when data has degenerate components. The model is simultaneously also the first Bayesian tensor canonical correlation analysis (CCA) method, when the data set contains two tensors; and is validated on both simulated and real data sets to capture the multi-view trilinear tensorial structure. In short, the methodology allows solutions for new problems. As a key example, it is successfully demonstrated in a novel problem setting, of collectively decomposing toxic and gene expression responses of multiple cancers to multiple drugs. The problem setting is a novel drug-response formulation, allowing identification of toxic drug responses that are shared by all subtypes, as well as those specific to some cancers. The application is described in Publication V, and an even extended formulation is presented in section 5.3.

5.2 Bayesian coupled matrix-tensor factorization

An even more interesting problem is when multi-source data collections include both matrices and tensors, coupled on a set of common samples. While BMTF (section 5.1) is already equipped to solve the case when data sets are of the same order, the case of differing orders (matrices and tensors) in multi-view data sets is presented in this section. Coupled matrix-tensor factorization methods have recently been introduced [121], however, have limited applicability. Here the first Bayesian formulation is presented that simultaneously extends to form a generic *Multi-tensor factorization* (MTF) while learning the nature and cardinality of factors automatically, in a data-driven fashion.

The multi-view coupled matrix-tensor factorization problem is studied here for multiple coupled matrices and tensors, such that samples are paired in two modes within the tensors while one mode of the tensors is common with the matrices. The goal now is to perform a collective decomposition of the matrices and tensors while segregating between the shared and private components, irrespective of the matrix or tensor nature of the view. This is achieved in an unsupervised way via a joint Factor Analysis and CP-type decomposition of matrices and tensors, having three key characteristics. First, the joint decomposition is characterized by a common set of latent variables $\mathbf{Z}$ between all the views (tensor and matrices).

This allows the formulation to capture common patterns between tensors and matrices, irrespective of their nature. Second, the view-specific loadings $\mathbf{W}^{(m)}$ control which of the patterns in $\mathbf{Z}$ are active in each of the views, enabling the model to find dependencies between any subset of the matrices or tensors. Third, only the tensor views are additionally modeled with loadings for the third mode $\mathbf{U}$ allowing both matrices and tensor to be factorized together.

Formally, given a collection of paired matrices and tensor views $\mathcal{X}^{(1)}$, $\ldots \mathcal{X}^{(T)}$, with an indicator variable β_t identifying the tensors ($\beta_t = 1$) and the matrices ($\beta_t = 2$), the MTF model is

$$
\begin{aligned}
\mathbf{x}^{(t)}_{n,:,l} &\sim \mathcal{N}\left(\mathbf{W}^{(t)}(\mathbf{z}_n * \mathbf{u}_l^{\beta_t}), \mathbf{I}(\tau^{(t)})^{-1}\right) \\
\mathbf{Z}, \mathbf{U}^{(1)} &\sim \mathcal{N}(0, \mathbf{I}) \\
\mathbf{U}^{(2)} &\sim \mathbf{1}_{1\times K} \\
w^{(t)}_{d,k} &\sim h_{t,k}\, \mathcal{N}\left(0, (\alpha^{(t)}_{d,k})^{-1}\right) + (1 - h_{t,k})\, \delta_0 \\
h_{t,k} &\sim Bernoulli(\pi_k) \\
\pi_k &\sim Beta(a^\pi, b^\pi) \\
\alpha^{(t)}_{d,k} &\sim Gamma(a^\alpha, b^\alpha) \\
\tau^{(t)} &\sim Gamma(a^\tau, b^\tau)\ .
\end{aligned}
$$

Here the joint matrix and tensor decomposition is governed by the common $\mathbf{Z}$ latent variables that capture the activity patterns of all matrix and tensor views. The $\mathbf{W}$ loadings (via the spike and slab prior, $h_{t,k}$) control the possibly different number and activity of components across all the views, irrespective of their matrix or tensor formulation. The $\mathbf{U}$ captures the patterns across the tensor mode of the tensor views. The remaining distributional assumptions are as defined in section 5.1.

Generalized MTF. The coupling of matrices and tensors can be formulated for all tensors, making it possible to decompose arbitrarily coupled data sets. Noting that matrices are tensors of order two, a fully extendable formulation of MTF is proposed that decomposes arbitrarily coupled tensors (including matrices) for investigating components shared and specific to each. The task is conceived as factorization of a large tensor $\widehat{\mathcal{X}} \in \mathbb{R}^{\sum D_i \times \sum D_i \times \sum L_i}$ that is made up by a block structured formulation of all the tensors and matrices. Here a key assumption is that the distinction between samples and dimensions is removed, and each mode of a tensor or matrix can be represented as a group or block of variables. The factorization can then be computed using a group-sparse prior that

controls the activity of a component in each block. A component is active in a view only if it is active in all the corresponding blocks (i.e. modes) of that view. Analogous to MTF a component active in two or more views is said to be shared between them. Formally, for the combined latent block factor loadings $\mathcal{W} \in \mathbb{R}^{\sum D_i \times K \times \sum L_i}$, the model is defined as

$$\begin{aligned}
\hat{\mathbf{X}}_{:,:,l} &\sim \mathbf{W}_{:,:,l}\mathbf{W}_{:,:,l}^{\top} \\
w_{d,k,l} &\sim h_{b_d,k,l}\mathcal{N}(v_{d,k}u_{l,k}, \lambda_{\beta_l}^{-1}) + (1 - h_{b_d,k,l})\delta_0 \\
\mathbf{U} &\sim \mathcal{N}(0,1) \\
v_{d,k} &\sim \mathcal{N}\left(0, (\alpha_{d,k})^{-1}\right) \ .
\end{aligned}$$

Here b_d denotes the block to which feature d belongs while β_l denotes whether slab l belongs to a tensor ($\beta_l = 1$) or a matrix ($\beta_l = 2$). The binary variable $h_{b_d,k,l}$ imposes the block structure for $b_d = 1 \ldots B$ blocks via a spike and slab prior, automatically learning the exact activation profiles. The $h_{b_d,k,l}$ activity patterns for each block-factor pair can then be used to infer the shared and specific factors for each data set. The $\mathbf{V}$ and $\mathbf{U}$ capture the trilinear CP factors, while λ_{β_l} tunes the factorization structure.

MTF performs well in simulated illustrations and its multi-view application on novel structural toxicogenomics is discussed in section 5.3. An open source implementation of a Gibbs sampler for MTF is provided at `http://research.ics.aalto.fi/mi/software/MTF`.

5.3 Toxicogenomics dependencies

Drug responses are measured at several levels of detail and type of activity. The gene expression responses of drugs present a systems-level view, while the toxicity summarizes the toxic behavior. Toxicogenomics aims to identify the links between genomic measurements of the cells and the toxicological profiles of drugs. These links, if uncovered can be illustrative of the molecular mechanisms of toxicity [122]. Interestingly, the toxicogenomic responses can also be analyzed together over a series of cell lines. Such an analysis poses the hypothesis that common patterns in the activity of genes and toxicity profiles of drugs can identify cellular response mechanisms, as well as be useful in predicting the toxicity outcome of a drug-cell treatment. The novel multi-tensor factorization methods of sections 5.1 and 5.2 model the multi-way structure of such coupled responses for identifying the underlying relationships.

A novel extension of the structure-response application is formulated next. To recap, in chapter 4, the cellular responses to drugs that are related to their structural descriptors were analyzed. The formulations also explored which of the responses were specific to a particular cancer type and which shared across all sub-types. In this section, the novel hypothesis studied is that *multi-source tensor formulations can identify and predict responses that are additionally related to drugs toxicity, while simultaneously discovering their cancer specificity and correspondence to structural properties of the drugs*.

The toxicogenomics data set contains three drug views, specifically, (i) the differential gene expression of several drugs as measured over three different cancers, (ii) the toxicity profiles of the same set of cancers and drugs; and (iii) the structural properties that describe the drugs. The gene expression is, therefore, a tensor of $drugs \times cancers \times genes$, while toxicity a tensor of $drugs \times cancers \times toxicity\_measures$, and structural properties a matrix of $drugs \times descriptors$. The expression and toxicity data sets come from CMap and NCI60 respectively and are processed as described in Chapter 3. For gene expression, the most varying genes are selected for a targeted analysis while the toxicity data comprises of three different toxicity measurements for each of the corresponding drug-cell pairs. The FCFP4 (section 3.3.1) are used for representing the structural properties of the drugs. The three data sets are paired on a common identity of 73 drugs.

MTF run with $K = 30$ successfully identified three drug-response components shared between all the views, demonstrating that the hypothesized structure exists and can be modeled with principled solutions. Several empty components were found indicating that the model can effectively learn the cardinality of the data. The model demonstrates its interpretive power that comes both from the component-view activities, which enable easier component identification then matrix methods, and by using the latent variables $\mathbf{Z}, \mathbf{U}$ and $\mathbf{W}^{(m)}$ for collective component visualization schemes.

To validate the model the joint components were interpreted, revealing both recently discovered findings, as well as those with potential for new biological discoveries of impact. The strongest component shows a response that is primarily driven by three structurally analogous drugs (geldanamycin, tanespimycin, and alvespimycin), all of which belong to the same class of HSP90 inhibitors. The component indicates that the

drugs are inducing an HSP response of the cells (by up-regulation of HSP genes) in all the three cancers, and is linked to the toxicity indicator GI50 (Growth Inhibition of 50%). It is well-known that the HSP90 protein is a molecular chaperone that stabilizes a variety of other proteins, including those that are crucial for the survival of cancer cells [123]. The HSP90 inhibitors bind to the protein resulting in a loss of function, because of which they have also been evaluated for their therapeutic efficacy in cancers [116, 124]. The component, therefore, presents a well known HSP90 response of cancer cells. The remaining components detailed in Publication VI also find well-formed findings though with potential novel implications. Publication V studied a subset of the problem using BMTF with gene expression, and toxicity data sets only, finding meaningful results.

The models were also validated quantitatively to assess if principled handling of multi-view multi-way structure in the data impacts the predictive performance. The performance was measured as RMSE (root mean squared error) of predicting missing data points and compared with state or the art multi-view matrix, as well as coupled matrix-tensor methods. The experiments verified that MTF and BMTF were able to predict the toxic response to drugs over multiple cancers significantly better than the existing alternatives. In summary, the principled models can decompose the novel drug-disease-response relationships more accurately and enable solutions for new applications. For example, they open up a new direction for QSAR modelers to integrate high-dimensional multi-way data sources for predictive analysis. For medicinal chemists, it makes it possible to explore the functional mechanisms and polypharmacology for a particular disease and response type, in a data-driven fashion.

6. Discussion and conclusions

Returning to the hypothesis posed at the beginning of this thesis, it is now possible to state that component-based dependency modeling methods can be used to uncover informative structure-response links and novel mechanisms of drug action, by taking advantage of the structure in all the data sets. The proposed methods were able to extract interesting relationships and segregate dependencies between multiple data sets from those specific to only one data set (noise, in this case).

This thesis has presented several novel multi-source methods each for different but increasingly complex and novel formulations of drug response analyses, demonstrating that it is important to take the appropriate structure of the data sets into account. Specifically, the multi-view models make it possible to discover interactions and mechanisms that can not be learned from individual measurement sources. Consequently, this research has shown that the multi-view approach not only discovers potentially novel biological findings but also identifies dependencies that are more informative quantitatively as well as qualitatively.

The second major finding of this thesis was that when the measurements come from multiple variables of interest as well as multiple paired response types, models that handle the paired multi-way structure work the best. This dissertation presented new problem formulations of multi-view tensor factorization and multi-tensor factorization along with the models for solving them. These studies generate several contributions to the current literature. First, the BMTF work contributed to the existing domain of tensor factorizations by providing a Bayesian method similar to multi-set CCA, but for tensors. Second, MTF extends the Bayesian tensor literature by allowing tensors of different orders and arbitrary coupling to be factorized for links shared and specific to each. These methods have validated that principled handling of the multi-mode multi-source struct-

ure is more accurate than the simpler alternatives. Finally, some of the limitations of single tensor factorizations were also removed, improving the current state of the art.

One of the more significant and novel findings to emerge from this thesis is that structure-response relationships of drugs can be studied systematically on a genome-wide level, using component-based approaches. To this end, data-driven models were learned from global profiling datasets, to identify the relationships between chemical features and biological responses. The studies presented in this dissertation have shown that analysis of drug compounds against a set of chemical descriptors and genome-wide measurements can be done by extracting a set of latent representations, which highlight features in the chemical descriptor set that account for differential biological changes in responses. Given the data-driven nature of the presented studies, the components capture drug-response links that are not constrained to known target information; therefore, many components capture potentially novel but still unidentified mechanisms of action. The proposed methods provide a systematic way to identify the links between the structural properties of drugs and the cell-specific gene expression and toxicity responses. Consequently, a potential implication of this thesis could be the possibility to assist drug designers to tailor compounds for matching the desired response patterns. It also opens up the opportunity for medicinal chemists to better understand the functional mechanisms of drug structures. In summary, this research moves a step further for assisting the targeted interventions of drugs.

Therefore, this thesis produces progress in both machine learning and molecular biology, emphasizing that by bridging fragmented disciplines interdisciplinary research plays an important role, not only in achieving the joint goals, but also in generating ideas for scientific advances within each.

This dissertation has opened up several possibilities for further investigation both in applications and methods development. The novel multi-view and multi-mode methods developed in this thesis may be applied to any other field, where systematic exploration of dependencies in partly related data sets is of interest. For example, they can be used in computational neuroimaging where it could be interesting to learn the relationships between a stimulus, its annotations, and the corresponding brain measurements. Such studies could help to understand the functional regions of the brain better. Another example could be computational social

sciences where modeling of user interactions can be studied using data from social apparatus such as Facebook, Twitter, and SMS. Such explorations could seek insights into a host of issues such as behaviors, norms, privacy concerns, and their inter-relationships.

Given the success of these first genome-wide and multi-disease structure-response decompositions, further research in this field would greatly assist in generalizing the patterns and exploring them for potential drug efficacies. One line of action for the generalizations could be an expansion to large-scale analyses with the data resources similar to that of LINCS, as they become fully available. For efficacies, it would be interesting for biological chemists to explore the plausible findings in the lab that the methods hypothesize for novel indications. As the results also demonstrated that integrating multiple data sources improves drug characterization, a recommended direction is to adapt and explore multi-source drug characterization further in practice.

The thesis also opens up new solutions for the existing questions in molecular biology. For example, it would be interesting to explore if the multi-tensor assumptions are useful in studying responses of several organisms. This can potentially be used to study which of the responses translate across organisms while segregating them from those which do not. Such findings can allow systematic quantification of the effects that can be translated from model organisms.

On the machine learning front, a natural direction is to postulate the new multi-view matrix and tensor models for exponential family distributions, increasing the scope of the methods. The choice of underlying factorization assumptions can be worthwhile to study. For example, it can be interesting to explore the new model formulations for incorporating biological pathway and target knowledge. A plausible way might be to develop informative priors to incorporate the knowledge. This approach could potentially support in building informative hypothesis for the data-driven analysis. Also, the technical choices and sparsity assumptions of the model's may be improved further, as needed by an application.

In summary, the thesis contributes data-driven methods to analyze the relationships between differential genome-wide responses of compounds and their structural properties while demonstrating their practical potential. This is a valuable scientific advancement because understanding the correlation between the structure of compounds and the differential responses of a particular disease can assist lead optimization efforts.

Bibliography

[1] E. C. Butcher, E. L. Berg, and E. J. Kunkel, "Systems biology in drug discovery," *Nature biotechnology*, vol. 22, no. 10, pp. 1253–1259, 2004.

[2] M. Schenone, V. Dančík, B. K. Wagner, and P. A. Clemons, "Target identification and mechanism of action in chemical biology and drug discovery," *Nature chemical biology*, vol. 9, no. 4, pp. 232–240, 2013.

[3] J. Lamb *et al.*, "The connectivity map: Using gene-expression signatures to connect small molecules, genes, and disease," *Science*, vol. 313, no. 5795, pp. 1929–1935, 2006.

[4] C. M. Bishop, "Latent variable models," in *Learning in graphical models*, pp. 371–403, Springer, 1998.

[5] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, New York, USA, 2006.

[6] C. Spearman, "General intelligence, objectively determined and measured," *The American Journal of Psychology*, vol. 15, no. 2, pp. 201–292, 1904.

[7] M. West, "Bayesian factor regression models in the large p, small n paradigm," *Bayesian statistics*, vol. 7, pp. 733–742, 2003.

[8] H. Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, pp. 321–377, 1936.

[9] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural Computation*, vol. 16, no. 12, pp. 2639–2664, 2004.

[10] C. Carvalho, J. Chang, J. Lucas, J. Nevins, Q. Wang, and M. West, "High-dimensional sparse factor modeling: Applications in gene expression genomics," *Journal of the American Statistical Association*, vol. 103, no. 484, pp. 1438–1456, 2008.

[11] A. Klami, S. Virtanen, and S. Kaski, "Bayesian canonical correlation analysis," *Journal of Machine Learning Research*, vol. 14, pp. 965–1003, 2013.

[12] M. J. Keiser, V. Setola, J. J. Irwin, C. Laggner, A. I. Abbas, S. J. Hufeisen, N. H. Jensen, M. B. Kuijer, R. C. Matos, T. B. Tran, *et al.*, "Predicting new molecular targets for known drugs," *Nature*, vol. 462, no. 7270, pp. 175–181, 2009.

[13] R. D. Cramer, D. E. Patterson, and J. D. Bunce, "Comparative molecular field analysis (comfa). 1. effect of shape on binding of steroids to carrier proteins," *Journal of the American Chemical Society*, vol. 110, no. 18, pp. 5959–5967, 1988.

[14] A. L. Hopkins, "Network pharmacology: the next paradigm in drug discovery," *Nature chemical biology*, vol. 4, no. 11, pp. 682–690, 2008.

[15] F. Iorio *et al.*, "Discovery of drug mode of action and drug repositioning from transcriptional responses," *Proceedings of the National Academy of Sciences, USA*, vol. 107, pp. 14621–14626, 2010.

[16] M. Iskar, M. Campillos, M. Kuhn, L. J. Jensen, V. Van Noort, and P. Bork, "Drug-induced regulation of target expression," *PLoS computational biology*, vol. 6, no. 9, p. e1000925, 2010.

[17] D. J. Bartholomew, M. Knott, and I. Moustaki, *Latent variable models and factor analysis: A unified approach*, vol. 904. John Wiley & Sons, 2011.

[18] A. Skrondal and S. RABE-HESKETH, "Latent variable modelling: A survey*," *Scandinavian Journal of Statistics*, vol. 34, no. 4, pp. 712–745, 2007.

[19] R. Salakhutdinov and A. Mnih, "Bayesian probabilistic matrix factorization using markov chain monte carlo," in *Proceedings of the 25th international conference on Machine learning*, pp. 880–887, ACM, 2008.

[20] M. E. Tipping and C. M. Bishop, "Probabilistic principal component analysis," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 61, no. 3, pp. 611–622, 1999.

[21] C. M. Bishop and M. E. Tipping, "A hierarchical latent variable model for data visualization," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 20, no. 3, pp. 281–293, 1998.

[22] D. M. Witten, R. Tibshirani, and T. Hastie, "A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis," *Biostatistics*, p. kxp008, 2009.

[23] A. V. Kossenkov and M. F. Ochs, "Matrix factorization for recovery of biological processes from microarray data," *Methods in enzymology*, vol. 467, pp. 59–77, 2009.

[24] D. Knowles and Z. Ghahramani, "Nonparametric bayesian sparse factor models with application to gene expression modeling," *Ann. Appl. Stat.*, vol. 5, no. 2, pp. 1534–1552, 2011.

[25] M. Gönen, "Predicting drug–target interactions from chemical and genomic kernels using bayesian matrix factorization," *Bioinformatics*, vol. 28, no. 18, pp. 2304–2310, 2012.

[26] F. R. Bach and M. I. Jordan, "A probabilistic interpretation of canonical correlation analysis," Tech. Rep. 688, Department of Statistics, University of California, Berkeley, 2005.

[27] D. M. Witten and R. J. Tibshirani, "Extensions of sparse canonical correlation analysis with applications to genomic data," *Statistical applications in genetics and molecular biology*, vol. 8, no. 1, pp. 1–27.

[28] L. Lahti, S. Myllykangas, S. Knuutila, and S. Kaski, "Dependency detection with similarity constraints," in *Machine Learning for Signal Processing, 2009. MLSP 2009. IEEE International Workshop on*, pp. 1–6, IEEE, 2009.

[29] T. Kolda and B. Bader, "Tensor decompositions and applications," *SIAM Review*, vol. 51, no. 3, pp. 455–500, 2009.

[30] J. D. Carroll and J.-J. Chang, "Analysis of individual differences in multidimensional scaling via an n-way generalization of Eckart-Young decomposition," *Psychometrika*, vol. 35, no. 3, pp. 283–319, 1970.

[31] R. A. Harshman, "Foundations of the parafac procedure: models and conditions for an explanatory multimodal factor analysis," *UCLA Working Papers in Phonetics*, vol. 16, pp. 1–84, 1970.

[32] L. R. Tucker, "Some mathematical notes on three-mode factor analysis," *Psychometrika*, vol. 31, no. 3, pp. 279–311, 1966.

[33] B. Yener, E. Acar, P. Aguis, K. Bennett, S. L. Vandenberg, and G. E. Plopper, "Multiway modeling and analysis in stem cell systems biology," *BMC systems biology*, vol. 2, no. 1, p. 63, 2008.

[34] L. Omberg, G. H. Golub, and O. Alter, "A tensor higher-order singular value decomposition for integrative analysis of dna microarray data from different studies," *Proceedings of the National Academy of Sciences*, vol. 104, no. 47, pp. 18371–18376, 2007.

[35] W. Li, C.-C. Liu, T. Zhang, H. Li, M. S. Waterman, and X. J. Zhou, "Integrative analysis of many weighted co-expression networks using tensor computation," *PLoS computational biology*, vol. 7, no. 6, p. e1001106, 2011.

[36] Y. Li and A. Ngom, "Non-negative matrix and tensor factorization based classification of clinical microarray gene expression data.," in *BIBM*, pp. 438–443, 2010.

[37] K. Brink-Jensen, S. Bak, K. Jørgensen, and C. T. Ekstrøm, "Integrative analysis of metabolomics and transcriptomics data: a unified model framework to identify underlying system pathways," *PloS one*, vol. 8, no. 9, p. e72116, 2013.

[38] R. A. Harshman and M. E. Lundy, "Parafac: Parallel factor analysis," *Computational Statistics & Data Analysis*, vol. 18, no. 1, pp. 39–72, 1994.

[39] R. B. Cattell, "Parallel proportional profiles and other principles for determining the choice of factors by rotation," *Psychometrika*, vol. 9, no. 4, pp. 267–283, 1944.

[40] J. B. Kruskal, "Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics," *Linear Algebra and its Applications*, vol. 18, no. 2, pp. 95 – 138, 1977.

[41] R. Bro, *Multi-way analysis in the food industry: models, algorithms, and applications*. PhD thesis, Københavns Universitet, Department of Food Science, 1998.

[42] C. A. Andersson and R. Bro, "The N-way toolbox for MATLAB," *Chemometrics and Intelligent Laboratory Systems*, vol. 52, no. 1, pp. 1 – 4, 2000.

[43] L. Xiong, X. Chen, T.-K. Huang, J. Schneider, and J. G. Carbonell, "Temporal collaborative filtering with bayesian probabilistic tensor factorization," in *Proceedings of SIAM Data Mining*, vol. 10, pp. 211–222, 2010.

[44] M. Mørup and L. K. Hansen, "Automatic relevance determination for multiway models," *Journal of Chemometrics*, vol. 23, no. 7-8, pp. 352 – 363, 2009.

[45] P. D. Hoff, "Hierarchical multilinear models for multiway data," *Computational Statistics & Data Analysis*, vol. 55, no. 1, pp. 530 – 543, 2011.

[46] A. Smilde, R. Bro, and P. Geladi, *Multi-way analysis: applications in the chemical sciences*. John Wiley & Sons, 2005.

[47] J. Håstad, "Tensor rank is np-complete," *Journal of Algorithms*, vol. 11, no. 4, pp. 644–654, 1990.

[48] R. A. Harshman and M. E. Lundy, "Data preprocessing and the extended parafac model," in *Research Methods for Multi-mode Data Analysis*, pp. 216–284, Praeger Publishers, 1984.

[49] J. Kruskal, R. Harshman, and M. Lundy, "How 3-mfa data can cause degenerate parafac solutions, among other relationships," *In Coppi R., Bolasco S., editors, Multiway data analysis*, pp. 115–121, 1989.

[50] M. E. Lundy, R. A. Harshman, and J. B. Kruskal, "A two-stage procedure incorporating good features of both trilinear and quadrilinear models," *Multiway data analysis*, pp. 123–130, 1989.

[51] V. de Silva and L. Lim, "Tensor rank and the ill-posedness of the best low-rank approximation problem," *SIAM Journal on Matrix Analysis and Applications*, vol. 30, no. 3, pp. 1084–1127, 2008.

[52] W. Krijnen, T. Dijkstra, and A. Stegeman, "On the non-existence of optimal solutions and the occurrence of "degeneracy" in the candecomp/parafac model," *Psychometrika*, vol. 73, no. 3, pp. 431–439, 2008.

[53] M. Sørensen, L. Lathauwer, P. Comon, S. Icart, and L. Deneire, "Canonical polyadic decomposition with a columnwise orthonormal factor matrix," *SIAM Journal on Matrix Analysis and Applications*, vol. 33, no. 4, pp. 1190–1213, 2012.

[54] H. A. Kiers, "Tuckals core rotations and constrained tuckals modelling," *Statistica Applicata*, vol. 4, no. 4, pp. 659–667, 1992.

[55] R. Henrion and C. A. Andersson, "A new criterion for simple-structure transformations of core arrays in n-way principal components analysis," *Chemometrics and Intelligent Laboratory Systems*, vol. 47, no. 2, pp. 189–204, 1999.

[56] L. D. Lathauwer, B. D. Moor, and J. Vandewalle, "A multilinear singular value decomposition," *SIAM J. Matrix Anal. Appl*, vol. 21, pp. 1253–1278, 2000.

[57] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin, *Bayesian Data Analysis*. Chapman & Hall/CRC Texts in Statistical Science, Boca Raton, Florida, USA, 2nd ed., 2003.

[58] L. Zhang, W. Zhou, V. E. Velculescu, S. E. Kern, R. H. Hruban, S. R. Hamilton, B. Vogelstein, and K. W. Kinzler, "Gene expression profiles in normal and cancer cells," *Science*, vol. 276, no. 5316, pp. 1268–1272, 1997.

[59] D. T. Ross, U. Scherf, M. B. Eisen, C. M. Perou, C. Rees, P. Spellman, V. Iyer, S. S. Jeffrey, M. Van de Rijn, M. Waltham, *et al.*, "Systematic variation in gene expression patterns in human cancer cell lines," *Nature genetics*, vol. 24, no. 3, pp. 227–235, 2000.

[60] J. Lu, G. Getz, E. A. Miska, E. Alvarez-Saavedra, J. Lamb, D. Peck, A. Sweet-Cordero, B. L. Ebert, R. H. Mak, A. A. Ferrando, *et al.*, "Microrna expression profiles classify human cancers," *nature*, vol. 435, no. 7043, pp. 834–838, 2005.

[61] U. R. Chandran, C. Ma, R. Dhir, M. Bisceglia, M. Lyons-Weiler, W. Liang, G. Michalopoulos, M. Becich, and F. A. Monzon, "Gene expression profiles of prostate cancer reveal involvement of multiple molecular pathways in the metastatic process," *BMC cancer*, vol. 7, no. 1, p. 64, 2007.

[62] L. J. van't Veer, H. Dai, M. J. Van De Vijver, Y. D. He, A. A. Hart, M. Mao, H. L. Peterse, K. van der Kooy, M. J. Marton, A. T. Witteveen, *et al.*, "Gene expression profiling predicts clinical outcome of breast cancer," *nature*, vol. 415, no. 6871, pp. 530–536, 2002.

[63] M. H. Cheok, W. Yang, C.-H. Pui, J. R. Downing, C. Cheng, C. W. Naeve, M. V. Relling, and W. E. Evans, "Treatment-specific changes in gene expression discriminate in vivo drug response in human leukemia cells," *Nature genetics*, vol. 34, no. 1, pp. 85–90, 2003.

[64] J. Sobek *et al.*, "Microarray technology as a universal tool for high-throughput analysis of biological systems," *Combinatorial Chemistry & High-throughput Screening*, vol. 9, pp. 365–380, 2006.

[65] J. M. Raser and E. K. OŚhea, "Noise in gene expression: origins, consequences, and control," *Science*, vol. 309, no. 5743, pp. 2010–2013, 2005.

[66] R. A. Irizarry, B. Hobbs, F. Collin, Y. D. Beazer-Barclay, K. J. Antonellis, U. Scherf, and T. P. Speed, "Exploration, normalization, and summaries of high density oligonucleotide array probe level data," *Biostatistics*, vol. 4, no. 2, pp. 249–264, 2003.

[67] E. R. Mardis, "Next-generation dna sequencing methods," *Annu. Rev. Genomics Hum. Genet.*, vol. 9, pp. 387–402, 2008.

[68] Z. Wang, M. Gerstein, and M. Snyder, "Rna-seq: a revolutionary tool for transcriptomics," *Nature Reviews Genetics*, vol. 10, no. 1, pp. 57–63, 2009.

[69] X. Cui and G. A. Churchill, "Statistical tests for differential expression in cdna microarray experiments," *Genome Biol*, vol. 4, no. 4, p. 210, 2003.

[70] J. D. Storey and R. Tibshirani, "Statistical significance for genomewide studies," *Proceedings of the National Academy of Sciences*, vol. 100, no. 16, pp. 9440–9445, 2003.

[71] S. Dudoit, J. P. Shaffer, and J. C. Boldrick, "Multiple hypothesis testing in microarray experiments," *Statistical Science*, pp. 71–103, 2003.

[72] M. Iskar, G. Zeller, P. Blattmann, M. Campillos, M. Kuhn, K. H. Kaminska, H. Runz, A.-C. Gavin, R. Pepperkok, V. van Noort, *et al.*, "Characterization of drug-induced transcriptional modules: towards drug repositioning and functional understanding," *Molecular systems biology*, vol. 9, no. 1, 2013.

[73] A. Subramanian *et al.*, "Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles," *Proceedings of the National Academy of Sciences, USA*, vol. 102, pp. 15545–15550, 2005.

[74] M. Ashburner *et al.*, "Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium," *Nature Genetics*, vol. 25, pp. 25–9, 2000.

[75] R. Huang, A. Wallqvist, and D. G. Covell, "Assessment of in vitro and in vivo activities in the national cancer institute's anticancer screen with respect to chemical structure, target specificity, and mechanism of action," *Journal of medicinal chemistry*, vol. 49, no. 6, pp. 1964–1979, 2006.

[76] U. Schmidt, S. Struck, B. Gruening, J. Hossbach, I. S. Jaeger, R. Parol, U. Lindequist, E. Teuscher, and R. Preissner, "Supertoxic: a comprehensive database of toxic compounds," *Nucleic acids research*, vol. 37, no. suppl 1, pp. D295–D299, 2009.

[77] R. H. Shoemaker, "The nci60 human tumour cell line anticancer drug screen," *Nature Reviews Cancer*, vol. 6, no. 10, pp. 813–823, 2006.

[78] J. Barretina, G. Caponigro, N. Stransky, K. Venkatesan, A. A. Margolin, S. Kim, C. J. Wilson, J. Lehár, G. V. Kryukov, D. Sonkin, *et al.*, "The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity," *Nature*, vol. 483, no. 7391, pp. 603–607, 2012.

[79] A. C. Lee, K. Shedden, G. R. Rosania, and G. M. Crippen, "Data mining the nci60 to predict generalized cytotoxicity," *Journal of chemical information and modeling*, vol. 48, no. 7, pp. 1379–1388, 2008.

[80] Z. Kutalik, J. S. Beckmann, and S. Bergmann, "A modular approach for integrative analysis of large-scale gene-expression and drug-response data," *Nature biotechnology*, vol. 26, no. 5, pp. 531–539, 2008.

[81] Q. Mo, S. Wang, V. E. Seshan, A. B. Olshen, N. Schultz, C. Sander, R. S. Powers, M. Ladanyi, and R. Shen, "Pattern discovery and cancer gene identification in integrated cancer genomic data," *Proceedings of the National Academy of Sciences*, vol. 110, no. 11, pp. 4245–4250, 2013.

[82] M. J. Garnett, E. J. Edelman, S. J. Heidorn, C. D. Greenman, A. Dastur, K. W. Lau, P. Greninger, I. R. Thompson, X. Luo, J. Soares, *et al.*, "Systematic identification of genomic markers of drug sensitivity in cancer cells," *Nature*, vol. 483, no. 7391, pp. 570–575, 2012.

[83] Y. Low, T. Uehara, Y. Minowa, H. Yamada, Y. Ohno, T. Urushidani, A. Sedykh, E. Muratov, V. Kuz?min, D. Fourches, *et al.*, "Predicting drug-induced hepatotoxicity using qsar and toxicogenomics approaches," *Chemical research in toxicology*, vol. 24, no. 8, pp. 1251–1262, 2011.

[84] R. Huang, A. Wallqvist, and D. G. Covell, "Anticancer metal compounds in nci's tumor-screening database: putative mode of action," *Biochemical pharmacology*, vol. 69, no. 7, pp. 1009–1039, 2005.

[85] D. G. Covell, A. Wallqvist, R. Huang, N. Thanki, A. A. Rabow, and X.-J. Lu, "Linking tumor cell cytotoxicity to mechanism of drug action: An integrated analysis of gene expression, small-molecule screening and structural databases," *Proteins: Structure, Function, and Bioinformatics*, vol. 59, no. 3, pp. 403–433, 2005.

[86] A. C. Lee, K. Shedden, G. R. Rosania, and G. M. Crippen, "Data mining the nci60 to predict generalized cytotoxicity," *Journal of chemical information and modeling*, vol. 48, no. 7, pp. 1379–1388, 2008.

[87] A. R. Leach and V. J. Gillet, "An introduction to chemoinformatics," 2007.

[88] "Chemical fragment generation and clustering software §,"

[89] R. C. Glem, A. Bender, C. H. Arnby, L. Carlsson, S. Boyer, and J. Smith, "Circular fingerprints: Flexible molecular descriptors with applications from physical chemistry to adme.," *IDrugs: the investigational drugs journal*, vol. 9, no. 3, pp. 199–204, 2006.

[90] M. Hassan, R. D. Brown, S. Varma-O'Brien, and D. Rogers, "Cheminformatics analysis and learning in a data pipelining environment," *Molecular diversity*, vol. 10, no. 3, pp. 283–299, 2006.

[91] D. Rogers and M. Hahn, "Extended-connectivity fingerprints," *Journal of chemical information and modeling*, vol. 50, no. 5, pp. 742–754, 2010.

[92] S. Ekins, R. C. Reynolds, H. Kim, M.-S. Koo, M. Ekonomidis, M. Talaue, S. D. Paget, L. K. Woolhiser, A. J. Lenaerts, B. A. Bunin, *et al.*, "Bayesian models leveraging bioactivity and cytotoxicity information for drug discovery," *Chemistry & biology*, vol. 20, no. 3, pp. 370–378, 2013.

[93] R. Mannhold, H. Kubinyi, G. Folkers, and G. Cruciani, *Molecular interaction fields: applications in drug discovery and ADME prediction*. John Wiley & Sons, 2006.

[94] G. Cruciani, M. Pastor, and W. Guba, "VolSurf: a new tool for the pharmacokinetic optimization of lead compounds," *European Journal of Pharmaceutical Sciences*, vol. 11, pp. S29–S39, 2000.

[95] A. Duran, G. C. Martinez, and M. Pastor, "Development and validation of amanda, a new algorithm for selecting highly relevant regions in molecular interaction fields," *Journal of chemical information and modeling*, vol. 48, no. 9, pp. 1813–1823, 2008.

[96] G. Ermondi, G. Caron, I. G. Pintos, M. Gerbaldo, M. Pérez, D. I. Pérez, Z. Gándara, A. Martínez, G. Gómez, and Y. Fall, "An application of two MIFs-based tools (Volsurf+ and Pentacle) to binary QSAR: The case of a palinurin-related data set of non-ATP competitive Glycogen Synthase Kinase 3β (gsk-3β) inhibitors," *European journal of medicinal chemistry*, vol. 46, no. 3, pp. 860–869, 2011.

[97] B. J. Drakulić, T. P. Stanojković, Z. S. Zižak, and M. M. Dabović, "Antiproliferative activity of aroylacrylic acids. structure-activity study based on molecular interaction fields," *European journal of medicinal chemistry*, vol. 46, no. 8, pp. 3265–3273, 2011.

[98] D. Dong and B. Wu, "In silico modeling of udp-glucuronosyltransferase 1a10 substrates using the volsurf approach," *Journal of pharmaceutical sciences*, vol. 101, no. 9, pp. 3531–3539, 2012.

[99] M. Sirota, J. T. Dudley, J. Kim, A. P. Chiang, A. A. Morgan, A. Sweet-Cordero, J. Sage, and A. J. Butte, "Discovery and preclinical validation of drug indications using compendia of public gene expression data," *Science translational medicine*, vol. 3, no. 96, pp. 96ra77–96ra77, 2011.

[100] R. Perkins, H. Fang, W. Tong, and W. J. Welsh, "Quantitative structure-activity relationship methods: Perspectives on drug discovery and toxicology," *Environmental Toxicology and Chemistry*, vol. 22, no. 8, pp. 1666–1679, 2003.

[101] H. Vinod, "Canonical ridge and econometrics of joint production," *Journal of Econometrics*, vol. 4, no. 2, pp. 147 – 166, 1976.

[102] S. E. Leurgans, R. A. Moyeed, and B. W. Silverman, "Canonical correlation analysis when the data are curves," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 55, pp. 725–740, 1993.

[103] I. González, S. Déjean, P. G. Martin, A. Baccini, *et al.*, "Cca: An R package to extend Canonical correlation analysis," *Journal of Statistical Software*, vol. 23, pp. 1–14, 2008.

[104] J. Venna and S. Kaski, "Nonlinear dimensionality reduction as information retrieval," in *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 2007.

[105] J. Venna, J. Peltonen, K. Nybo, H. Aidos, and S. Kaski, "Information retrieval perspective to nonlinear dimensionality reduction for data visualization," *The Journal of Machine Learning Research*, vol. 11, pp. 451–490, 2010.

[106] R. A. Cairns, I. S. Harris, and T. W. Mak, "Regulation of cancer cell metabolism," *Nature Reviews Cancer*, vol. 11, no. 2, pp. 85–95, 2011.

[107] T. N. Seyfried and L. M. Shelton, "Cancer as a metabolic disease," *Nutr Metab (Lond)*, vol. 7, no. 7, pp. 269–70, 2010.

[108] M. M. Gottesman, "Mechanisms of cancer drug resistance," *Annual review of medicine*, vol. 53, no. 1, pp. 615–627, 2002.

[109] M. Majumder, S. Khan, A. Lehto, L. Turunen, S. Kaski, and K. Wennerberg, "A chemical systems biological approach to understand the association of metabolic stress and resistance to dna damaging drugs in breast cancer," *European Journal of Cancer*, vol. 48, p. 170 (abstract), 2012.

[110] C. Archambeau and F. R. Bach, "Sparse probabilistic projections," in *Advances in neural information processing systems*, pp. 73–80, 2009.

[111] F. Deleus and M. M. Van Hulle, "Functional connectivity analysis of fmri data based on regularized multiset canonical correlation analysis," *Journal of Neuroscience methods*, vol. 197, no. 1, pp. 143–157, 2011.

[112] T. J. Mitchell and J. J. Beauchamp, "Bayesian variable selection in linear regression," *Journal of the American Statistical Association*, vol. 83, no. 404, pp. 1023–1032, 1988.

[113] R. M. Neal, "Bayesian learning for neural networks," *Springer-Verlag*, 1996.

[114] K. Degtyarenko, P. De Matos, M. Ennis, J. Hastings, M. Zbinden, A. Mcnaught, R. Alcantara, M. Darsow, M. Guedj, and M. Ashburner, "Chebi: a database and ontology for chemical entities of biological interest," *Nucleic acids research*, vol. 36, no. suppl 1, pp. D344–D350, 2008.

[115] K. A. Hartwell, P. G. Miller, S. Mukherjee, A. R. Kahn, A. L. Stewart, D. J. Logan, J. M. Negri, M. Duvet, M. Järås, R. Puram, *et al.*, "Niche-based screening identifies small-molecule inhibitors of leukemia stem cells," *Nature chemical biology*, 2013.

[116] A. Kamal *et al.*, "A high-affinity conformation of HSP90 confers tumour selectivity on HSP90 inhibitors," *Nature*, vol. 425, no. 6956, pp. 407–410, 2003.

[117] S. A. Khan, A. Faisal, J. P. Mpindi, J. A. Parkkinen, T. Kalliokoski, A. Poso, O. P. Kallioniemi, K. Wennerberg, and S. Kaski, "Comprehensive data-driven analysis of the impact of chemoinformatic structure on the genome-wide biological response profiles of cancer cells to 1159 drugs," *BMC bioinformatics*, vol. 13, no. 1, p. 112, 2012.

[118] J. C. Costello, L. M. Heiser, E. Georgii, M. Gönen, M. P. Menden, N. J. Wang, M. Bansal, P. Hintsanen, S. A. Khan, J.-P. Mpindi, *et al.*, "A community effort to assess and improve drug sensitivity prediction algorithms," *Nature Biotechnology*, 2014.

[119] W. Chu and Z. Ghahramani, "Probabilistic models for incomplete multi-dimensional arrays," in *Proceedings of AISTATS, JMLR W&CP*, vol. 5, pp. 89 – 96, 2009.

[120] H. Lu, "Learning canonical correlations of paired tensor sets via tensor-to-vector projection," in *Proceedings of IJCAI*, pp. 1516–1522, 2013.

[121] E. Acar, M. A. Rasmussen, F. Savorani, T. Naes, and R. Bro, "Understanding data fusion within the framework of coupled matrix and tensor factorizations," *Chemometrics and Intelligent Laboratory Systems*, vol. 129, pp. 53 – 63, 2013.

[122] T. Hartung, E. V. Vliet, J. Jaworska, L. Bonilla, N. Skinner, and R. Thomas, "Food for thought ... systems toxicology," *ALTEX*, vol. 29, no. 2, pp. 119–128, 2012.

[123] J. S. Isaacs, W. Xu, and L. Neckers, "Heat shock protein 90 as a molecular target for cancer therapeutics," *Cancer Cell*, vol. 3, no. 3, pp. 213 – 217, 2003.

[124] L. Neckers and P. Workman, "Hsp90 molecular chaperone inhibitors: are we there yet?," *Clinical Cancer Research*, vol. 18, no. 1, pp. 64–76, 2012.

DISSERTATIONS IN INFORMATION AND COMPUTER SCIENCE

Aalto-DD177/2014 Laitinen, Tero
Extending SAT Solver with Parity Reasoning. 2014.

Aalto-DD178/2014 Gonçalves, Nicolau
Advances in Analysis and Exploration in Medical Imaging. 2014.

Aalto-DD191/2014 Kindermann, Roland
SMT-based Verification of Timed Systems and Software. 2014.

Aalto-DD207/2014 Chen, Xi
Real-time Action Recognition for RGB-D and Motion Capture Data. 2014.

Aalto-DD211/2014 Soleimany, Hadi
Studies in Lightweight Cryptography. 2014.

Aalto-DD28/2015 Su, Hongyu
Multilabel Classification through Structured Output Learning – Methods and Applications. 2015.

Aalto-DD31/2015 Talonen, Jaakko
Advances in Methods of Anomaly Detection and Visualization of Multivariate Data. 2015.

Aalto-DD43/2015 van Heeswijk, Mark
Advances in Extreme Learning Machines. 2015.

Aalto-DD62/2015 Luttinen, Jaakko
Bayesian Latent Gaussian Spatio-Temporal Models. 2015.

Aalto-DD91/2015 Kohonen, Oskar
Advances in Weakly Supervised Learning of Morphology. 2015.

Aalto-DD 105/2015

ISBN 978-952-60-6309-6 (printed)
ISBN 978-952-60-6310-2 (pdf)
ISSN-L 1799-4934
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)

Aalto University
School of Science
Department of Computer Science
www.aalto.fi

BUSINESS +
ECONOMY

ART +
DESIGN +
ARCHITECTURE

SCIENCE +
TECHNOLOGY

CROSSOVER

**DOCTORAL
DISSERTATIONS**