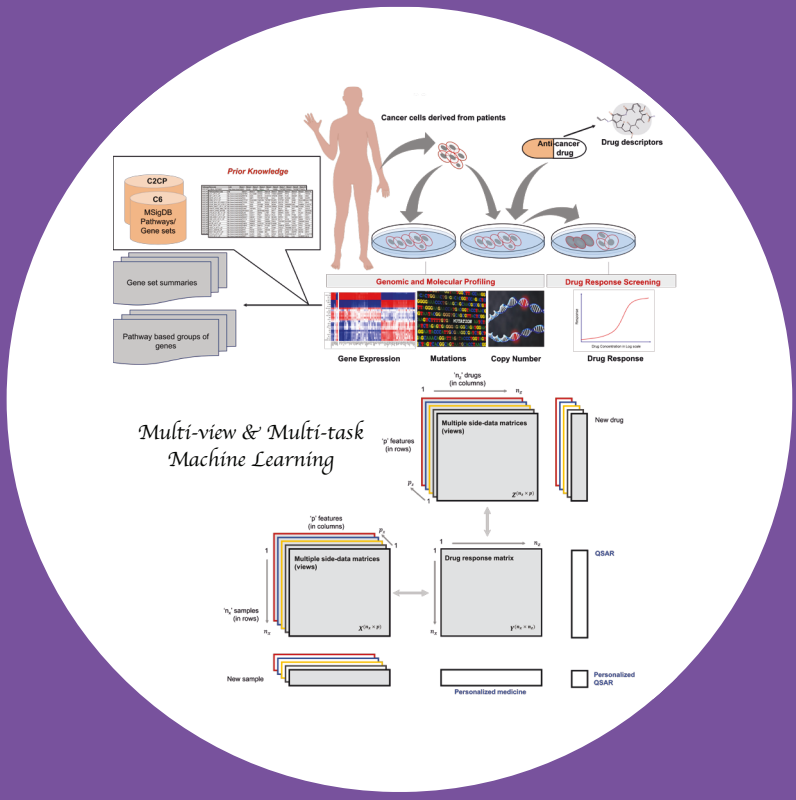
Department of Computer Science

Machine learning methods for improving drug response prediction in cancer

Muhammad Ammad-ud-din



Aalto University

DOCTORAL DISSERTATIONS

Aalto University publication series
DOCTORAL DISSERTATIONS 127/2017

Machine learning methods for improving drug response prediction in cancer

Muhammad Ammad-ud-din

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 T2 of the school on 27th July 2017 at 12 noon.

Aalto University
School of Science
Department of Computer Science
Probabilistic Machine Learning Group

Supervising professors
Samuel Kaski, Ph.D.,
Professor Academy of Finland,
Professor of Computer Science, Aalto University,
Director, Finnish Centre of Excellence in Computational Inference Research COIN.

Preliminary examiners
Janne Lehtio, Ph.D.,
Professor in Medical Proteomics,
Platform Director, Mass spectrometry, Science for Life Laboratory,
Group leader, Cancer proteomics,
Department of Oncology-Pathology,
Karolinska Institutet, Stockholm,
Sweden.

Tapio Pahikkala, Ph.D.,
Assistant Professor,
Department of Future Technologies,
University of Turku,
Finland.

Opponents
Anil Korkut, Ph.D.,
Assistant Professor
Department of Bioinformatics and Computational Biology
The UT MD Anderson Cancer Center,
United States of America.

Aalto University publication series
DOCTORAL DISSERTATIONS 127/2017

ISBN 978-952-60-7514-3 (printed)
ISBN 978-952-60-7513-6 (pdf)
ISSN-L 1799-4934
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)
http://urn.fi/URN:ISBN:978-952-60-7513-6

Unigrafia Oy
Helsinki 2017

Finland

441 697
Printed matter

Aalto University

Abstract

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

Author
Muhammad Ammad-ud-din

Name of the doctoral dissertation
Machine learning methods for improving drug response prediction in cancer

Publisher School of Science

Unit Department of Computer Science

Series Aalto University publication series DOCTORAL DISSERTATIONS 127/2017

Field of research Machine Learning and Bioinformatics

Manuscript submitted 14 March 2017 **Date of the defence** 27 July 2017

Permission to publish granted (date) 9 May 2017 **Language** English

☐ **Monograph** ☒ **Article dissertation** ☐ **Essay dissertation**

Abstract

Personalizing medicine, by choosing therapies that maximize effectiveness and minimize side effects for individual patients, is one of the prime challenges in cancer treatment. At the core of personalized medicine is a machine learning problem: Given a set of patients whose response to some drugs has been observed, predict the response of a new patient or to a new drug. Computationally predicted responses can then be used to generate hypotheses for selecting therapies tailored to individual patients. However, the prediction task is exceedingly challenging, raising the need for the development of new machine learning methods.

This thesis undertakes a unique multi-disciplinary approach to predict drug responses by utilizing multiple data sources in cancer, while simultaneously advancing the computational methods to improve accuracy. Specifically, the thesis presents a new Bayesian multi-view multi-task method that outperformed existing computational models in an international crowdsourcing challenge to predict drug responses. The method is further extended to solve the more challenging task of predicting drug responses in multiple cancer types. Notably, the thesis extends the kernelized Bayesian matrix factorization method with component-wise multiple kernel learning for effectively inferring associations between a large number of biologically motivated data sources and the latent factors. The results demonstrate that the new formulation of the method, supplemented with prior biological knowledge, is helpful for discovering interpretable associations as well as for predicting the drug responses of new cancer cells.

The original contribution of this thesis is two-fold: First, the thesis proposes novel multi-view and multi-task methods to predict drug responses in cancer cells with increased accuracy. Second, new ways of incorporating prior biological knowledge are explored to further improve drug response predictions. Open source implementations of the new methods have been released to facilitate further research.

Keywords Data integration, Multi-view Multi-task Machine Learning, Personalized Medicine

ISBN (printed) 978-952-60-7514-3 **ISBN (pdf)** 978-952-60-7513-6

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

Location of publisher Helsinki **Location of printing** Helsinki **Year** 2017

Pages 133 **urn** http://urn.fi/URN:ISBN:978-952-60-7513-6

Preface

This work has been carried out in the Probabilistic Machine Learning (PML) group in the Department of 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 broad exposure and excellent networking with top researchers in the field. This research and thesis have been funded by the Academy of Finland (Finnish Center of Excellence in Computational Inference Research COIN, grant nos. 251170, 140057 and 295503).

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 unique collaborations. I feel deeply privileged to have learned so much from a top researcher!

I wish to especially thank Prof. Tero Aitokallio from Computational Systems Medicine group at Institute for Molecular Medicine Finland (FIMM), Dr. Krister Wennerberg from Cancer Chemical Systems Biology team at FIMM.

I am also thankful to Dr. Elisabeth Georgii from PML group (now at Helmholtz Zentrum München, Germany), Dr. Mehmet Gönen from PML group (now at Koç University, Turkey), Dr. Pekka Marttinen from PML group and Dr. Suleiman Ali Khan from PML group (now at Computational Systems Medicine group, FIMM), both for their guidance in the collaborative research, and for teaching me how to envisage multiple opportunities from such a multidisciplinary setup. Your contributions extend way beyond what we have written together! This outstanding interdisciplinary environment has enabled me to acquire skills in probabilistic machine learning, bioinformatics, and personalized medicine. My sincere compliments belong to all co-authors, especially Eemeli Leppäaho, Marta Soare,

Tuomo Laitinen, Prof. Antti Poso and Prof. Olli Kallioniemi. It has been a pleasure working with you all!

I have got to know many wonderful colleagues, while working on other related projects. In particular Luana Micallef, Iiris Sundin, Homayun Afrabandpey, Pedram Daee and Tomi Peltola from the PML group, Baris Serim and Prof. Giulio Jacucci from HIIT. Thank you all!

I would also like to express my gratitude to all the current and former members of the PML group especially Jussi Gillberg, Ziyuan Lin, Sami Remes, Jonathan Strahl, Xiangju Qin, Juuso Parkkinen, Tommi Suvitaival and Manuel Eugster for fruitful discussions on science as well as life in general.

I would like to thank the pre-examiners Prof. Janne Lehtiö from Karolinska Institutet, Sweden and Asst Prof. Tapio Pahikkala form University of Turku, Finland for their valuable comments on the thesis. Their feedback and suggestions helped me to improve the thesis. I would also like to express my gratitude towards all the friends for the shared moments at Aalto University Adnan Ghani, Ali Faisal, Aqdas Malik, Hussnain Ahmed, Muhammad Irfan Khan and Rao Anwer.

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 started teaching me how to read and write; and my father, for always standing beside me in every difficult moment of life. I am also incredibly grateful to my brothers for the continuous support they provided, not only during the doctoral studies but all throughout the life. Lastly, I wish to express my most sincere gratitude and appreciation to my wife and our son Wahaaj for absolutely everything.

Espoo, June 20, 2017,

Muhammad Ammad-ud-din

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 James C. Costello, Laura M. Heiser, Elisabeth Georgii, Mehmet Gönen, Michael P. Menden, Nicholas J. Wang, Mukesh Bansal, Muhammad Ammad-ud-din, Petteri Hintsanen, Suleiman A. Khan, John-Patrick Mpindi, Olli Kallioniemi, Antti Honkela, Tero Aittokallio, Krister Wennerberg, NCI DREAM Community, James J. Collins, Dan Gallahan, Dinah Singe, Julio Saez-Rodriguez, Samuel Kaski, Joe W. Gray and Gustavo Stolovitzky. A Community Effort to Assess and Improve Drug Sensitivity Prediction Algorithms. *Nature Biotechnology*, 32, 12, 1202-1212, 2014.

II Muhammad Ammad-ud-din, Elisabeth Georgii, Mehmet Gönen, Tuomo Laitinen, Olli Kallioniemi, Krister Wennerberg, Antti Poso, and Samuel Kaski. Integrative and Personalized QSAR Analysis in Cancer by Kernelized Bayesian Matrix Factorization. *Journal of Chemical Information and Modeling*, 54, 8, 2347-2359, 2014.

III Muhammad Ammad-ud-din, Suleiman A.Khan, Disha Malani, Astrid Murumägi, Olli Kallioniemi, Tero Aittokallio and Samuel Kaski. Drug response prediction by inferring pathway-response associations with Kernelized Bayesian Matrix Factorization. *Bioinformatics*, 32, 17, i455-i463, 2016.

IV Solveig Sieberts, Fan Zhu, Javier García-García, Eli Stahl, Abhishek

Pratap, Gaurav Pandey, Dimitrios Pappas, Daniel Aguilar, Bernat Anton, Jaume Bonet, Ridvan Eksi, Oriol Fornés, Emre Guney, Hongdong Li, Manuel Marín, Bharat Panwar, Joan Planas-Iglesias, Daniel Poglayen, Jing Cui, Andre Falcao, Christine Suver, Bruce Hoff, Venkat Balagurusamy, Donna Dillenberger, Elias Chaibub Neto, Thea Norman, Tero Aittokallio, Muhammad Ammad-ud-din, Chloe-Agathe Azencott, Víctor Bellón, Valentina Boeva, Kerstin Bunte, Himanshu Chheda, Lu Cheng, Jukka Corander, Michel Dumontier, Anna Goldenberg, Peddinti Gopalacharyulu, Mohsen Hajiloo, Daniel Hidru, Alok Jaiswal, Samuel Kaski, Beyrem Khalfaoui, Suleiman Khan, Eric Kramer, Pekka Marttinen, Aziz Mezlini, Bhuvan Molparia,Matti Pirinen, Janna Saarela, Matthias Samwald, Veronique Stoven, Hao Tang, Jing Tang, Ali Torkamani, Jean-Philippe Vert, Bo Wang, Tao Wang, Krister Wennerberg, Nathan Wineinger, Guanghua Xiao, Yang Xie, Rae Yeung, Xiaowei Zhan, Cheng Zhao, Jeff Greenberg, Joel Kremer, Kaleb Michaud, Anne Barton, Marieke Coenen, Xavier Mariette, Corinne Miceli, Nancy Shadick, Michael Weinblatt, Niek de Vries, Paul Tak, Danielle Gerlag, Tom W. J. Huizinga, Fina Kurreeman, Cornelia Allaart, Stanley Bridges, Lindsey Criswell, Larry Moreland, Lars Klareskog, Saedis Saevarsdottir, Leonid Padyukov, Peter Gregersen, Stephen Friend, Robert Plenge, Gustavo Stolovitzky, Baldomero Oliva, Yuanfang Guan, and Lara Mangravite. Crowdsourced assessment of common genetic contribution to predicting anti-TNF treatment response in rheumatoid arthritis. *Nature Communication*, 7, EP:12460, 2016.

V Eemeli Leppäaho and Muhammad Ammad-ud-din and Samuel Kaski. GFA: Exploratory Analysis of Multiple Data Sources with Group Factor Analysis. *Journal of Machine Learning Research*, 18, 39, 1-5, 2017.

VI Marta Soare, Muhammad Ammad-ud-din and Samuel Kaski. Regression with n$\rightarrow$1 by Expert Knowledge Elicitation. In *15th IEEE International Conference on Machine Learning and Applications (ICMLA)*, Anaheim, USA, 734-739, Dec 2016.

Author's Contribution

Publication I: "A Community Effort to Assess and Improve Drug Sensitivity Prediction Algorithms"

The author took part in conducting the data analysis for the top-performing approach in the drug sensitivity prediction challenge. In particular, the author processed the data, conducted experiments to evaluate the best modeling choices for the top-performing approach, generated the final predictions and participated in writing the article.

Publication II: "Integrative and Personalized QSAR Analysis in Cancer by Kernelized Bayesian Matrix Factorization"

The author had the main responsibility for the design, experiments and the preparation of the article. The main modeling idea was developed jointly. Dr. Gönen implemented the method. The manuscript was written jointly.

Publication III: "Drug response prediction by inferring pathway-response associations with Kernelized Bayesian Matrix Factorization"

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

Publication IV: "Crowdsourced assessment of common genetic contribution to predicting anti-TNF treatment response in rheumatoid arthritis"

The author took part in conducting the data analysis for the Bayesian multi-task multiple kernel learning method and participated in writing the article. In particular, the author conducted experiments to choose the best modeling choices for the model and generated the final predictions.

Publication V: "GFA: Exploratory Analysis of Multiple Data Sources with Group Factor Analysis"

The author designed and carried out the model complexity selection on the genomic case study and participated in writing the article. Mr. Leppäaho had the main responsibility of the model, experiments, and writing of the article.

Publication VI: "Regression with n$\rightarrow$1 by Expert Knowledge Elicitation"

The author conducted the experiments, evaluated the results and participated in writing the article. Ms. Soare had the main responsibility of the methodology, algorithm, and preparation of the article.

List of Abbreviations and Symbols

Abbreviations

AUC	Area Under the dose response Curve
CCLE	Cancer Cell line Encylopedia
CTRP	Cancer Therapeutic Response Portal
DNA	Deoxyribonucleic Acid
DREAM	Dialogue on Reverse Engineering Assessment and Methods
GDSC	Genomics of Drug Sensitivity in Cancer
MIF	Molecular Interaction Fields
MKL	Multiple Kernel Learning
MSigDB	Molecular Signature Database
MT	Multi-Task
NCI	National Cancer Institute
PCR	Principal Component Regression
PDF	Probability Density Function
PLS	Partial Least Squares
QSAR	Quantitative Structure Activity Relationship
RNA	Ribonucleic Acid

Symbols

In this thesis calligraphic symbols, for example bold uppercase to denote matrices ($\mathbf{X}$), and bold lowercase column vectors ($\mathbf{x}$). Normal lowercase symbols (x) represent 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 $\mathbf{X}$
$p(y)$	probability distribution of y
$p(\theta, y)$	joint probability distribution of θ and y
$p(\theta \vert y)$	conditional probability distribution of θ given y
λ	scalar-valued, regularization parameter

1. Introduction

1.1 Computational personalized medicine

The fundamental goal of personalized medicine is to identify individualized therapies that maximize effectiveness with minimal side-effects. The effectiveness, however, depends on a variety of factors such as genetic, molecular, and environmental, and much of this information remains unknown. A promising approach is then to learn computational models using the available genomic and molecular profiles of the patient samples and the responses they elicit when exposed to a spectrum of drugs. The models serve in two ways: (1) predicting responses for a new sample and (2) identifying features predictive of drug responses. These predictions can then be used to generate hypotheses on choosing therapies that are potentially effective for the individual patient.

Drug response prediction is valuable in treating many diseases and is especially plausible for cancer, where the genetic heterogeneity of the cells has a significant impact on the response. The development of computational models has been made possible through recent large-scale high-throughput screening studies [1, 2], providing genomic, transcriptomic (or molecular) as well as drug response measurements on preclinical human models of cancers (commonly known as "cell lines"). Importantly, these benchmark studies showed that computational models could be learned to gain mechanistic and therapeutic insights.

However, the modeling task is difficult and poses computational challenges that raise the need for developing new and robust models. A key challenge underlying drug response modeling is the "small n, large p" problem. Compared to the very high dimensionality of the '*-omics*' features, which are often highly correlated as well, small sample sizes offer limited

statistical power, leading to highly uncertain predictions. Moreover, the cancer types are inherently heterogeneous, thus making the robust inference even harder. This thesis addresses a fundamental research question in personalized medicine: how to improve the drug response prediction in cancer given the aforementioned computational challenges. The ultimate goal would be to develop models that provide better predictions, a step towards finding a personalized cure in cancer. Here, a key assumption is that genome affects the response, and we want to learn that.

The main idea in computational personalized medicine is very simple: given the genome-wide features of the cell lines as input (also known as independent variables or covariates) and drug responses as the target (output or dependent variable), learn a model of the drug response. The model could predict responses to previously unseen cell lines and could help interpret features relevant to the prediction task. Alternatively, given the chemical and structural data of the drugs as input and their responses on a single cell line (output or dependent variable), learn a predictive model of the drug response. The model could predict responses to untested drugs and compounds. It is commonly known as a quantitative structure-activity relationship (QSAR) task in pharmaceuticals and has wide applicability in drug design and discovery [3].

Machine learning models, commonly used to solve these tasks, could broadly be grouped into two categories: linear and nonlinear. Multivariate regression, partial least squares (PLS), and principal component regression (PCR) are some of the frequently used linear models, while kernels methods (such as kernelized regression), neural networks, and random forest are well-known examples of nonlinear counterparts (Chapter 2 presents these models briefly).

While there are several advantages and disadvantages, linear models neglect the relevant nonlinear structure in the data and only consider the linear relationships between the variables. Commonly, these models are used to predict for a target variable only [1, 2]. In practice, for each drug, a separate model is learned. When the data are scarce, a natural motive would be to gain statistical strength by gathering evidence from multiple sources, for instance, the use of response measurements of multiple drugs and the integration of multiple '-*omics*' data sources. However, the features encoded in the genomic and transcriptomic data may provide varying levels of information about the functional activities in the cell. For example, a mutated gene present in the genomic data may show up or

down-regulation of its expression in the transcriptomic data. Many existing models do not offer any approach to exploit this information among the features of different data sources systematically, other than treating them as independent covariates in a predictive model. There is a need to develop new methods that can integrate these various types of features in a principled way, effectively learning the relationship between them and all the target variables simultaneously. The thesis adopts a multi-disciplinary strategy and combines three different research fields: machine learning, bioinformatics, and personalized medicine.

1.2 Contribution

This thesis proposes novel machine learning methods for improving the accuracy of predicting drug responses in cancer cells[1]. Specifically, the contributions of the thesis are two-fold:

1. The thesis presents multi-view and multi-task methods to improve the drug response predictions by efficiently addressing the underlying computational challenges. Computational methods, commonly referred to as ***multi-view***, aim to effectively integrate multiple data sources, yielding an increased signal-to-noise ratio in the parameter space. Here, a key assumption is that a joint modeling of the features from multiple data sources reveal hidden statistical relationships, which may not be obvious from the data itself and are relevant for the drug response prediction. Another closely related class of methods, commonly known as ***multi-task***, allows for the learning of a task from other related tasks. For example, predicting one drug response alone can be considered as an individual task, whereas two drugs whose responses are highly correlated can provide the statistical boost when learned together. These models are especially beneficial when the number of samples is small, or when the samples come from a diverse collection of cancers.

2. The thesis introduces novel ways of incorporating prior biological knowledge. Due to the small sample sizes, the data-driven task requires additional information to improve drug response predictions. A valuable source is prior knowledge in public databases (or from biomedical

[1]This thesis uses both these terms ”cancer cells” or “cancer cell lines” interchangeably.

experts), readily available to be utilized in a biologically meaningful way.

Publication I presents a novel multi-view multi-task kernelized probabilistic regression method to predict drug responses on previously unseen cancer cell lines. Unlike the classical personalized medicine task of predicting a drug response for the new cell line, the proposed method predicts responses to multiple drugs simultaneously. The results demonstrate that nonlinear modeling, multi-task and multi-view learning supplemented with the prior biological knowledge significantly improve the drug response predictions. Publication IV extends the scope of the model in a disease other than cancer. The prediction task is to predict the drug responses in cells derived from rheumatoid arthritis patients.

Publication II introduces a new multi-view multi-task kernelized probabilistic matrix factorization method for drug response prediction. The factorization is needed to automatically differentiate the underlying distinct drug responses across and within different cancer types. The publication studies novel applications of drug response prediction, made possible by the proposed model. The multi-view matrix factorization allows for the inclusion of input data sources for both cell lines and drugs simultaneously, making it possible to solve the difficult task of predicting response to an entirely new drug on a previously unseen cell line.

Publication III proposes a flexible formulation for the multi-view part of the model introduced in Publication II, additionally proposing a new way to incorporate prior biological knowledge. The results show that the combination of new model formulation and prior knowledge improves drug response analysis in comparison to existing approaches (including its predecessor proposed in Publication II). Also, Publication III proposes interpretable relationships between the groups of molecular features and drug responses.

Publication V presents an R package that implements Group Factor Analysis (GFA), a linear multi-view factor analysis model. Unlike kernel methods, GFA searches for linear relationships between the genome-wide features of the cell lines and their drug response profiles. Interpreting these feature-response relationships could be useful in several clinical applications. The R package provides a complete data analysis pipeline to study drug response predictions in future studies.

Publication VI introduces an approach to extracting and using prior knowledge from an expert in drug response modeling. Publication I, II

and III incorporate prior knowledge from public databases; however, another useful source of knowledge is an expert (a clinician or biomedical researcher) who could provide information on the importance of the features. Specifically, Publication VI presents an approach to extracting prior knowledge from an expert efficiently, and the results demonstrate that this approach significantly improves drug response predictions in "small n, large p" problem setting.

1.3 Organization of the thesis

Chapter 2 presents computational models that are commonly used to predict drug responses. Chapter 3 discusses the drug response data, the genomic, transcriptomic, and other genome-wide features of the cell lines, and the chemical as well as structural properties of drugs in addition to the notion of prior biological knowledge. Chapter 4 describes the multi-view and multi-task learning that form the basis of this thesis. In particular, the chapter presents the relevant theoretical knowledge for "learning from multiple data sources". Chapter 5 presents the main contributions of the thesis, describing the new multi-view and multi-task methods for drug response prediction. Chapter 6 concludes the thesis and suggests directions for further research.

2. Computational models for drug response prediction

In drug response analysis, two complementary approaches exist that exploit high-throughput drug screening data from cancer cells. First, personalized medicine approach: given genome-wide features of a cell, predict the cell's responses to *a priori* known set of drugs. Second, QSAR approach: given chemical features of a drug, predict the response of this drug in *apriori* known set of cells. The idea of computational modeling for these approaches consists of two steps. The first step is to learn a model that describes the relationship between drug responses and genomic (or chemical) features. Next, the model predicts for new data with unknown response values. This chapter briefly describes the commonly used computational models for both approaches.

2.1 Linear regression

Consider $\mathbf{X} \in \mathbb{R}^{n \times p}$, the matrix of '*-omics*' data and $\mathbf{y} \in \mathbb{R}^{n \times 1}$, the vector of drug responses. Where n denotes the number of samples (here cell lines or patients) and p represents the number of features.
Linear regression assumes that the drug responses $\mathbf{y}$ have been generated from a linear combination of unknown weight vector $\mathbf{w} \in \mathbb{R}^{1 \times p}$ and the features in $\mathbf{X}$, corrupted with a noise ϵ (usually known as an error term). Mathematically, the model can be expressed as

$$\mathbf{y} = \mathbf{X}\mathbf{w}^T + \epsilon. \tag{2.1}$$

The machine learning goal is then to learn $\mathbf{w}$ such that $\mathbf{y} - \mathbf{X}\mathbf{w}^T$ is minimized, or more intuitively the sum of squared errors $\sum_{i=1}^{n} \epsilon_i^2$ is minimized. Once the $\mathbf{w}$ are learned, they can be used for two purposes: (1) to gain biological insights from important features and (2) to predict drug

response for a new sample x_*, by

$$y_* = \mathbf{x}_* \mathbf{w}^T. \tag{2.2}$$

The classical ordinary least squares approach infers w such that $\sum_{i=1}^{n} C(\mathbf{w}, x_i, y_i) = \sum_{i=1}^{n} \epsilon_i^2 = (\mathbf{y} - \mathbf{X}\mathbf{w})^T(\mathbf{y} - \mathbf{X}\mathbf{w})$ is minimized. In genomic and molecular data, since the number of features is often much higher than the number of samples, the inference becomes ill-posed and suffers from over-fitting. A frequent solution is to introduce a regularization term in the cost function. The regularization penalizes the complexity of the regression model by adding penalty terms, $||\mathbf{w}||_1$ and $||\mathbf{w}||_2$ norms to the cost function. Recently, Zou et al [4] have proposed the elastic net regularization, as follows:

$$\min_{\mathbf{w}} \sum_{i=1}^{n} C(\mathbf{w}, x_i, y_i) + \lambda\big(\alpha||\mathbf{w}||_1 + \frac{(1-\alpha)}{2}||\mathbf{w}||_2^2\big). \tag{2.3}$$

Here $\lambda > 0$ is the penalty parameter that controls the amount of regularization and shrinking of the weight vector w. The penalty reduces to the ridge regression [5] when $\alpha = 0$ and, the lasso regression [6] when $\alpha = 1$, while for all $\alpha \in (0, 1)$ is the combination of the ridge and lasso regression.

A striking characteristic of the elastic net regularization is that it can select groups of correlated features while favoring automatic feature selection and continuous shrinkage. The "grouped selection" property of the elastic net penalty gives an advantage over the *lasso* penalty for selecting the features (in this case genes) whose expressions are predictive of drug responses. Typically, for those genes sharing the same biological process, the correlations among them can be very high. The *lasso* regression only selects at most n genes out of p candidates (while $p >> n$). Whereas, with the elastic net regression, when a gene is selected, the whole group of correlating genes is included into the model.

Owing to these nice properties of feature selection, elastic net regression has become a popular computational model in drug response analysis and, has been applied in numerous benchmark studies to identify genes predictive of drug responses in cancer cells [1, 2, 7, 8].

Principal Component Regression (PCR) performs a two-step approach. In the first step, top principal components Z (also known as latent factors) of the genomic data X are obtained. In the second step, a linear model between the y and Z is learned using ordinary least squares. The inferred weight vector w in the transformed space, is then mapped back into the original space to obtain biological interpretations.

Partial Least Squares (PLS) Regression. While the main idea is similar to PCR, PLS [9] regression additionally exploits the output $\mathbf{y}$ to construct the latent factors $\mathbf{Z}$. A key motivation is that these factors capture the underlying noise-free true signal, which may be substantial for the model. It is commonly assumed that using the subset of latent factors for learning the regression model behaves like a regularization and avoids over-fitting [10]. Multiple studies have used both PCR and PLS for drug response analysis [11, 12, 13].

2.2 Kernel methods

In contrast to the linear counterparts, kernel methods can model the non-linear relationships in the data by choosing a suitable similarity measure between the samples. The main idea of a kernel-based formulation is to learn a decision function in the feature space where data points are implicitly mapped to using a kernel function [14].

Consider $\mathbf{y} \in \mathbb{R}^{n \times 1}$, the vector of drug responses and $\mathbf{X} \in \mathbb{R}^{n \times p}$, the matrix of genomic data containing n independent and identically distributed samples $\{x_i \in \boldsymbol{\mathcal{X}}\}_{i=1}^{n}$. The decision function that is used to predict the drug response of a new sample x_* can be written as

$$f(x_*) = \mathbf{w}^T \boldsymbol{k_*} + b \tag{2.4}$$

where $\mathbf{w}$ denotes the unknown weight vector for the samples and b represents the error term (also called as bias). And $\boldsymbol{k_*} = [k(\mathbf{x}_1, \mathbf{x}_*).....k(\mathbf{x}_n, \mathbf{x}_*)]^T$ where $k : \boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{X}} \rightarrow \mathbb{R}$ is the kernel function (usually known as kernels in machine learning community) that gives similarities between the samples. Intuitively, a regression framework that uses a kernel as input can be termed as kernelized regression.

Several drug response studies have used kernelized regression, for example to predict drug responses in cancer cells [2, 11, 15] as well as to predict drug side effects from chemical structures and target information [16]. A closely related concept in the Bayesian domain is known as Gaussian Process (GP) regression [17]. Likewise, GP regression uses a kernel to define the covariance of a prior distribution over the decision function.

Kernel function

In machine learning, kernels are used to change the input space of the data using a mapping function. In other words, any method whose formulation has the input data always in dot-product form, can be kernelized by replacing these dot-product terms by a mapping function, called a *kernel*; however, that should satisfies certain mathematical properties [18]. Kernelizing a method has several advantages, such as obtaining a richer feature representation, which yields better learning abilities.

When kernelizing a method, a central question is how to assure whether the chosen kernel corresponds to a dot product on a higher-dimensional vector space, in other words, whether a kernel is *valid*. Two concepts are essential for assuring the validity of a kernel: Gram matrix and positive semi-definiteness (see [18] for a more detailed explanation and mathematical proofs). According to the Mercer's Theorem [19] and Reproducing Kernel Hilbert Space (RKHS) [20, 18], any positive semi-definite matrix that is represented as a dot product of two functions on the RKHS, spanned by the kernel. In this way, a kernel can be treated as a similarity measure for pairs of samples.

Gaussian kernel. The Gaussian kernel (also known as Radial Basis Function RBF) between two samples $\mathbf{x}$ and $\mathbf{x}_*$ is defined as

$$k(\mathbf{x}, \mathbf{x}_*) = \exp\Big(\frac{\|\mathbf{x} - \mathbf{x}_*\|^2}{2\sigma^2}\Big). \tag{2.5}$$

Where σ is a hyperparameter referred to as the *length scale*. It determines the smoothness of the decision boundary imposed by the kernel. The larger σ is, the smoother the decision surface is. The kernel is called *Gaussian* since its formula is proportional to the PDF of the normal distribution. Gaussian is one of the most widely used kernels due to its simplicity, interpretability, and ability to capture nonlinear relationships.

2.3 Random forest and ensemble methods

The random forest method [21] works on two principles: (1) It generates several regression trees based on random selection of samples and random selection of features. (2) It provides the prediction of the new sample by averaging the predictions of several regression trees.

For each tree, a random subset of samples is selected using a bootstrap sampling of the observed samples, and the features to be selected at each

branch are a random subset of the full set of features. Essentially, a random subset of samples, as well as a random subset of features, is selected to generate each random tree, thus forming a random forest. Because of this two-step selection at random, the generated regression trees are decorrelated, and thus, averaging their prediction responses reduces the variance of the error. As the random forest regression is built on the ensemble approach, it is expected to provide high prediction accuracy, but the biological interpretability is limited.

Though the method can handle a large number of features, the number of regression trees needed to model the data would also be very high, hence raising potential scalability and complexity challenges. Nonetheless, random forest regression has frequently been used for predicting drug responses in multiple studies [15, 22, 23, 24].

2.4 Deep learning and neural networks

Representation learning takes to input the raw data and automatically learns the representations needed for robust prediction. Deep learning, in particular, deep neural networks, are representation learning methods with multiple layers of representation, obtained by defining simple but nonlinear functions that each transform the representation at one layer (beginning from the raw input data features) into a representation at a higher, slightly more abstract layer. Eventually, defining enough such transformations supports learning complex functions. Deep learning models do not require designing these layers of representations manually. Instead, these are learned from data using a general-purpose learning algorithm [25].

Usually, neural networks (consisting of one layer) have been applied to solve classical QSAR problems [26], as well as to predict drug responses in cancer cells [27]. However, due to high-dimensional data, these shallow neural networks are limited in their applicability and depends on appropriate preselection of features. However, more recently, deep learning methods have been proposed that utilize high-dimensional chemical and structural features to predict drug responses [28].

A beneficial aspect of the linear models is that they are easier to interpret and provide a straightforward analysis of the relationship between the genomic and molecular features and drug responses. However,

the linear models neglect the relevant nonlinear structure in the data which may result in a bad prediction accuracy. On the contrary, nonlinear method provide improved prediction performance, however compromise the interpretability.

3. Multiple data sources

Thanks to recent large-scale high-throughput screening efforts that have generated, not only the genomic and molecular data but also the response measurements of hundreds of anti-cancer drugs against several hundreds of human cancer cell lines [1, 2]. Figure 3.1 explains the various types of data sources that are potentially available in drug response analysis (not always in practice) and, this chapter presents them briefly.

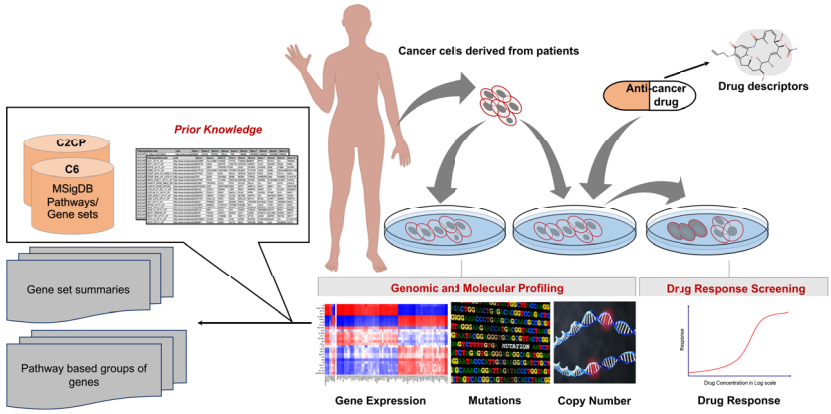


Figure 3.1. There are available multiple types of high-throughput data sources for drug response analysis. Cells are derived from cancer patients and cultured in a laboratory. On the one hand, the cells are subjected to genomic and transcriptomic profiling to measure their mutational statuses, copy number variants, gene expression levels and similar other molecular features. On the other hand, the same cells are treated with anti-cancer drugs and their growth in response to the given concentrations is summarized with a drug response curve. Moreover, a drug's activity can be characterized by chemical and structural features (known as descriptors) using chemo-informatics tools. Additionally, much prior knowledge about the features is readily available, for instance, MSigDB contains a collection of gene sets and pathways which can be used to extract pathway-based groups of features (here genes).

3.1 Genomic and transcriptomic profiles

Gene expression is the process of transcribing a gene (essentially a DNA sequence) into RNA. The expression level of a particular gene indicates the approximate number of RNA copies present in a cell and generally correlates with the abundance of the corresponding proteins produced in the cell. Thus, the expression of a gene characterizes its activity under certain normal or disease conditions in cells. Recent technologies such as microarray [29] and RNA sequencing [30], have enabled measuring the activities of several thousand of genes simultaneously (hence named *high-throughput*). Mainly, the gene expression measurements characterize the genome-wide transcriptomic profiles that may be indicative of the particular response patterns and helps in identification of molecular targets predictive of drug responses in cancer cells.

A simple **mutation** defines a change in a genome on a small scale affecting one or a few nucleotides. Mutations can arise from a change of a single nucleotide (i.e., point mutation), insertion or deletion of one or more extra nucleotides. Such simple mutations (referred to as genomic features) occurring in the coding regions of the genome alter the protein product, resulting in the change of functionality of that particular protein. These mutations can contribute to cancer development. The widely known p53, is a tumor suppressor gene and is critical for protection against cancer; mutation in the p53 gene alters the normal functioning of the gene and can lead to the development of cancer. Hence, the analysis of cancer cells to identify the status of genomic features as mutated or not mutated is essential.

Copy Number Variations (CNVs) are a form of structural variation in the genome that produces an unusual number of copies of one or more genomic regions. In contrast to point mutations described above, CNVs affect larger genome regions. Deletions of such larger genome regions reduce the copy number whereas duplications increase the copy number compared to the normal level. Studies have shown that CNVs can greatly affect cancer development. For example, the EFGR gene amplification can contribute to the development of certain types of non-small lung cancer [31].

Another type of genome rearrangements is translocations. They also play a role in cancer, for instance, the translocation of EWS-FLI1 gene is

the cause of Ewing's cancer [1].

Exome sequencing [32] is a new technique that aims to capture sequence variation in the form of mutations and indels (insertion/deletions) in all coding regions of the genome. This approach requires a hybridization step to capture all exons and a sequencing step to identify sequence variation in the exons.

In addition to these transcriptomic and genomic profiles, other high-throughput technologies measure the epigenome and proteome level changes as well. A technology by Bibikova et al [33] monitors DNA methylation process at the epigenetic level by counting the number of methyl-groups attached to the DNA. At the proteome level, reverse phase protein array (RPPA) technique measures the abundance of proteins in the cell [34].

3.2 Drug response measurements

In cancer studies, a drug response measurement is a measure of the effectiveness of a drug inhibiting cell viability. Growing cell cultures are treated with different concentrations of a particular drug, and after a fixed time period the number of cells in the culture is measured by a fluorescent signal. These measurements result in a drug dose-response curve, where the x-axis denotes the drug concentrations and the y-axis the corresponding cell viability (Figure 3.1 shows an example curve). Typically the curve is fitted using a small number of measurement points. It is a common practice to summarize the curve by the IC_{50} value, which is defined as a quantitative measure indicating the dose of a particular drug needed to inhibit the biological activity by half. Other characteristic values reported in the literature are GI_{50} (concentration required to inhibit 50% of maximal cell growth) [35, 36], IC_{25} (inhibit the biological activity by 25%), IC_{75} (inhibit the biological activity by 75%) and AUC [1, 2, 7].

3.3 Drug descriptors and targets

Chemo-informatics tools link the chemical and structural properties of drugs to their biological efficacy. The most traditional chemical properties are 2D structural descriptors considering the number of atoms & bonds (such as the number of non-H atoms & rotatable bonds), number of each functional group, the number of carbon chains, the presence of rings and

ring sizes. These descriptors are well suited for investigating a variety of absorption, distribution, metabolism, and excretion (i.e., ADME) properties and to generate predictive models of bio-activity of the drugs.

Beyond that, 3D MIFs have been studied extensively. They describe the interaction energies between the drug molecules and standardized probes (representing functional groups) and have been shown to be useful in pharmacokinetics, drug discovery, and drug design [37, 38]. Modern software tools extract relevant numerical descriptors such as Volsurf and Pentacle using the MIFs. Volsurf calculates the volume and the surface of the interaction contours at predefined energy values [39], whereas Pentacles selects informative points from MIFs based on energy value distributions. The resulting descriptors are represented independently of the original 3D coordinates by an auto-correlation transform. Hence, these are called GRIND, GRID Independent Descriptors [40, 41].

3.4 Prior biological knowledge

To address “small n, large p” problem, a perspective direction is to incorporate informative prior biological knowledge. The underlying assumption is that prior biological knowledge introduces additional structure and information that is valuable for the modeling task. There are several ways of including additional biological knowledge.

For instance, much prior knowledge about the molecular and genomic features is readily available in public databases. For example, Molecular Signature Database MSigDB [42] contains a collection of gene sets and pathways which can be used to derive pathway-based groups of features. Other the other hand, COSMIC database [43] maintains a list of features frequently implicated in cancer, and these features can potentially be used to give emphasis during the modeling task. In many cases, primary therapeutic targets of the drugs are also known, that can be used to extract potential relevant data from the public databases, to be used as complementary information in the model. Moreover, new data sources denoting the gene set summaries and activities of pathways can also be computed using the tool PARADIGM [44] and biological knowledge stored in the public databases. This thesis makes use of prior knowledge for the modeling task and also explores various ways of incorporating meaningful biological information in the form of prior knowledge.

Table 3.1 indicates (with a tick mark) the various types of inputs

such as gene expression, mutations, copy number variations, drug descriptors and prior knowledge-based data sources that have been used in the publications reported in this thesis.

Table 3.1. Summary of the genomic, transcriptomic, drug descriptors and prior knowledge based data sources used in the publications presented in this thesis

Pub.	Gene Expression	Copy Number Variations	Mutations	Prior knowledge	Drug Descriptors
I	✓	✓	✓	✓	
II	✓	✓	✓	✓	✓
III	✓			✓	
IV			✓	✓	
V	✓				
VI	✓			✓	

4. Learning from multiple data sources

One of the prime goals of learning from multiple data sources in machine learning is to integrate multiple views systematically, yielding increased prediction accuracy. A key assumption is that the integration of several data sources extracts more relevant information and structure from the data that is valuable for the prediction task. Specifically, in “small n,

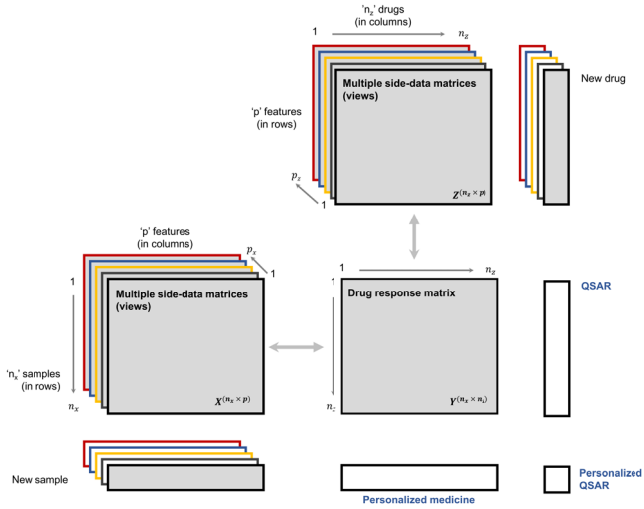


Figure 4.1. The figure conceptualizes the learning from multiple data sources for the drug response prediction. Computationally, the data sources are encoded as data matrices of size n-by-p, where n is the number of samples and p is the number of features. The thesis abbreviates the data matrix with a keyword ‘view’ and, as there are multiple views, multi-view learning aims to effectively integrate the views, yielding an increased signal-to-noise ratio in the parameter space. The drug response matrix consists of measurements from multiple drugs. Hence multi-task learning allows joint modeling of all the drugs, making it possible to gather statistical evidence across multiple drugs. Essentially, multi-view multi-task learning can learn an integrated model using a set of samples whose responses to drugs have been observed previously. The model can then predict the response of the drugs on a new sample or predict the response of the samples to a new drug or predict the response of a new drug on a new sample (personalized QSAR task).

large p" problem setting, combining multiple data sources can be beneficial for two reasons. First, the modeling exploits the additional information that naturally arises from the integration. Second, relevant information present across multiple sources obviously provides statistical strength when learned jointly. For example, the gene expression patterns of multiple pathways may be linked to drug responses, or a mutated gene present in one view may show up- or down-regulation of its expression in the other views.

A fundamental methodological challenge involves determining what is the 'relevant information' to extract and developing models for predicting drug responses given the other '*-omics*' data sources. A naive approach would integrate different sources into one data source and learn a set of potentially predictive features by explicitly optimizing a cost function. However, it may result in too simple approach requiring strong regularization to eliminate the false positives, and it may be difficult to incorporate prior knowledge about the relevant information adequately.

A likely solution addressing these limitations is a generative probabilistic approach, that models each data source by a set of latent variables. Multi-view learning can then extends the latent variable modeling to integrate multiple data sources and jointly learn the hidden statistical relationships within and between data sources, which may not be obvious from the data itself and are relevant for the prediction task.

In particular, Figure 4.1 conceptualizes the learning from multiple data sources for drug response analysis. With the blend of these powerful machine learning approaches, it is possible to address several challenging and novel prediction tasks such as personalized medicine and QSAR.

4.1 Probabilistic machine learning

In drug response analysis, gene expression and similar other '*-omics*' data sources are highly noisy due to the measurement techniques, and additionally contain irrelevant information due to the complexity of the biological system. Moreover, the small sample sizes offer limited statistical power leading to high uncertainty in the predictions. Use of probabilistic machine learning is beneficial, since it provides principled ways to handle uncertainties by assuming suitable probability distributions for the unobserved data (in the form of priors) [45].

Given the observed data, the posterior of the model parameter (unob-

served data) can be obtained using the Bayes theorem:

$$p(\theta|y) = \frac{p(\theta, y)}{p(y)} = \frac{p(\theta)p(y|\theta)}{p(y)}$$

The probabilistic model can be defined by specifying the *joint probability distribution* $p(\theta, y)$ for the model *parameters* θ and the observed data y. This can be written as the product of the *prior distribution* $p(\theta)$ and the *likelihood function* $p(y|\theta)$: $p(\theta, y) = p(\theta)p(y|\theta)$. Bayes' rule then provides the conditional probability of θ given y, called the *posterior* distribution. The marginal $p(y)$ is a constant which normalizes the posterior such that it integrates to one.

In other words, these simple mathematical expressions express the essential core of probabilistic machine learning: define the model $p(\theta, y)$ and perform the necessary computations, known as *inference*, to summarise $p(\theta|y)$ in appropriate ways (or the inference of the unobserved data) [45]. The posterior probabilities $p(\theta|y)$ can be seen as likelihood $p(y|\theta)$ weighted by the prior $p(\theta)$, where the prior increases or decreases the impact of likelihood in the posterior. Hence, the modeling favors the solutions matching the prior. The prior may be used to encode our belief 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.

Given the model, the inference task requires computing the posterior distribution of the parameters θ. However, except for only the simplest models, the exact computation of the posterior is intractable and requires approximate methods, which fall into two broad categories, deterministic and sampling [46].

Sampling draws values of θ from some approximate distributions, and then correct those draws to better approximate the true posterior distribution $p(\theta|y)$. The samples are drawn sequentially, with a distribution of the samples drawn previously. Given infinite computational resources, sampling methods can infer the true posterior distribution. However, only a finite number of samples can be drawn practically.

On the other hand, *Variational Bayesian (VB)* approximation, is a widely used deterministic approach for inferring the posterior which is computationally too intensive to sample [46]. In VB, the posterior distribution $p(\theta|y)$ is approximated by a variational distribution $q(\theta)$: $p(\theta|y) \approx q(\theta)$, where q is chosen as a simpler distribution than the original posterior. The goal is then to make q as close to p as possible, using the Kullback-Leibler divergence [45].

Latent variable models

Latent variable modeling has emerged as a power tool for data analysis in Bayesian machine learning [47, 48]. A latent variable model provides a flexible way of modeling dependencies between the high-dimensional data variables $\mathbf{x} \in \mathbb{R}^p$ by assuming that the data is generated by a set of underlying low-dimensional variables $\mathbf{z} = \{z_1, z_2, \dots z_R\}$ (also termed as factors or components). Since these components are not observed, they are called *latent*. The latent components present a concise and denoised summary of the underlying processes that have generated the data and can be used to create hypotheses about the dependencies between the variables or to predict the unobserved variables. The number of latent components is typically much smaller than the observed data dimensionality $R << p$, while each component follows a simpler distribution.

Matrix factorization

Matrix factorization has evolved as a fundamental approach to machine learning with many applications such as missing value prediction, dimensionality reduction, and data visualization [49, 50, 51]. Among other motivations, matrix factorization is commonly used to decompose the high-dimensional observed data into multiple low-dimensional latent factors. The underlying assumption is that the combination of multiple latent factors has generated the observed data, however, each combination has generated some parts of the data. A wide variety of approaches has been studied to factorize matrices, while optimizing different criteria [52, 53].

For factorization of a single matrix, *Factor Analysis (FA)* [54] is a well-studied method. FA assumes that the data matrix $\mathbf{Y} \in \mathbb{R}^{n \times p}$ can be modeled by a latent factor formulation such that the factor capture dependencies between the variables. Mathematically, FA is expressed as

$$\mathbf{Y} = \mathbf{Z}\mathbf{W}^T + \mathbf{E}. \tag{4.1}$$

Where the columns of $\mathbf{Z}$ are the R latent factors, $\mathbf{W} \in \mathbb{R}^{n \times R}$ contains their loadings, and $\mathbf{E}$ is residual noise, respectively.

In Bayesian domain, the factor analysis model can be defined by choosing suitable probability distributions (as priors) for $\mathbf{Z}$, $\mathbf{W}$ and $\mathbf{E}$ [55].

4.2 Multi-view learning

In several applications, data from multiple, related sources are available providing complementary information about the problem under study. Especially in drug response analysis, examples mentioned in Chapter 3 include various types of genomic and transcriptomic data, drug response, and the descriptor data. Thus, it is practical to define a joint probabilistic model for all available data sources.

In the machine learning domain, learning from multiple data sources with paired samples is called *multi-view* learning, where a '*view*' refers to a single data source (or data matrix). A classical two-view method for modeling linear dependencies between two data sources is *canonical correlation analysis* (CCA) [56, 57].

Multiple kernel learning

When the task is to learn a composite representation of the available data, *Multiple Kernel Learning (MKL)* [58] provides a principled solution. MKL can integrate multiple data sources, effectively yielding an increased signal-to-noise ratio in the parameter space. The data sources are encoded in the form of kernels, hence model nonlinear relationships as well.

Mercer's Theorem [19] assures that combining valid kernels using simple operations, such as addition and element-wise multiplication produces a valid kernel. This theorem provides the opportunity to obtain composite kernels from simple ones, facilitating to capture, more complex properties of data, in addition to integrating data sources with incompatible representations. These benefits result in improved model performance.

MKL algorithms basically replace the kernel in Equation 2.4 with a combined kernel calculated as a function of the input kernels (also termed as base kernels). The most common combination is to use a weighted sum of P kernels
$\{k_m : \boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{X}} \to \mathbb{R}\}_{m=1}^{P}$:

$$f(x_*) = \mathbf{w}^T \underbrace{\left(\sum_{m=1}^{P_{\mathrm{x}}} \boldsymbol{e}_m \mathbf{K}_{\mathbf{x}_*,m} \right)}_{\text{composite kernel}} + \quad b. \tag{4.2}$$

where the vector of kernel weights is denoted by $\boldsymbol{e}$, kernels by $\mathbf{K}_{\mathbf{x}_*}$, the unknown weight vector of samples by $\mathbf{w}$ and, bias term by b respectively. The methods can formulate different MKL algorithms by assuming constraints on the kernel weights e.g., arbitrary weights, nonnegative weights

or weights on a simplex (see [58, 59], for a more detailed explanation on MKL methods).

4.3 Multi-task learning

Multi-task learning is an approach to machine learning that learns a task together with other related tasks at the same time using a shared representation [60]. MTL often leads to a better model for the all the tasks, as it allows the model to use the commonalities among the tasks [61, 62].

In drug response analysis, predicting one drug response alone can be considered as an individual task, whereas two drugs whose responses are highly correlated can provide a statistical boost when learned together. It is especially beneficial when the number of samples is small, or when the samples come from a diverse collection such as in the pan-cancer scenario. Additionally, multi-task learning may yield better predictive accuracy by diminishing the impact of 'off-target effects' and drug-specific experimental noise. Moreover, methods that jointly model response profiles across multiple drugs may yield insights into groups of drugs targeting similar pathways or having similar mechanisms of action.

Mathematically, the multi-task learning can be expressed for joint modeling of multiple drug responses (or relevant tasks) as a linear combinations of features and the unknown weights as:

$$\mathbf{Y} = \mathbf{X}\mathbf{W}^T + \epsilon. \tag{4.3}$$

Here, in contrast to Equation 2.1, $\mathbf{Y} \in \mathbb{R}^{n \times t}$ is the matrix with the columns representing t drug responses (or learning tasks) and the rows n denoting the number of samples (cell lines or patients). Also, $\mathbf{W} \in \mathbb{R}^{t \times p}$ is the matrix of unknown weights (or co-efficients), where $\mathbf{w}^t \in \mathbb{R}^p$ is the unknown weight vector for each task. And, $\mathbf{X} \in \mathbb{R}^{n \times p}$ denotes the matrix of '*-omics*' data consisting of p features.

Several approaches have been proposed to infer w jointly while learning multiple tasks [63, 64, 65]. Inferring $\mathbf{w}^t$ independently with l_1 regularization would yield the following cost function [64]:

$$\min_{\mathbf{w}^t} \frac{1}{n_t} \sum_{i=1}^{n_t} C^t(\mathbf{w}^t, x_i^t, y_i^t) + \lambda ||\mathbf{w}^t||_1. \tag{4.4}$$

Here $\lambda > 0$ is the penalty parameter that controls the amount of regularization and shrinking of the weight vector w. Assuming λ is the same across the tasks, inferring each of the $\mathbf{w}^t$ is equivalent to inferring $\mathbf{W}$ globally

obtained by summing the cost functions [64]:

$$\min_{\mathbf{W}} \sum_{t=1}^{T} \frac{1}{n_t} \sum_{i=1}^{n_t} C^t(\mathbf{w}^t, x_i^t, y_i^t) + \lambda \sum_{t=1}^{T} ||\mathbf{w}^t||_1. \tag{4.5}$$

Essentially, Equation 4.5 implies task-specific selection of the features. In other words, features are selected predictive of each task individually.

However, the ultimate goal of multi-task learning in drug response prediction could be to prefer regrularization scheme that favors feature selection with shared pattern of response predictiveness. For example, when learning personalized models for predicting responses of multiple drugs, even though individual responses can vary. There could be a common subset of genomic features shared across drugs. Considering the entire block of coefficients associated with a feature across tasks as a unit, encouraging regrularization at the block level, leads to infer weights with several blocks of coefficients are set to zero.

Thus, in order to select features common across tasks, an alternatively regularization scheme would be [64, 65]:

$$\min_{\mathbf{W}} \sum_{t=1}^{T} \frac{1}{n_t} \sum_{i=1}^{n_t} C^t(\mathbf{w}^t, x_i^t, y_i^t) + \lambda \sum_{j=1}^{p} ||w_j||_2, \tag{4.6}$$

where

$$\sum_{j=1}^{p} ||w_j||_2 = \sum_{j=1}^{p} \Big(\sum_{t=1}^{T} |w_{tj}|^2 \Big)^{\frac{1}{2}}. \tag{4.7}$$

Notably, compared to Equation 4.5, the regularization in Equation 4.6 is a two-step approach. In the first step, the l_2- norms of the feature-specific coefficient vectors are penalized followed by the l_1- norm of the vector. In other words, non-relevant features that are common across tasks are set to zeros collectively. In a Bayesian approach, assuming a appropriate prior on the features would help regularize the model and learn predictive features common across the tasks.

Likewise, multi-task MKL have shown to be useful in many kernel based formulations [66, 67, 68, 69], here a obvious goal is to predict the output variables (e.g., the drug responses) rather than the feature selection. The key idea is to formulate multi-task learning as that of learning a composite kernel, obtained by combining a given set of input (or base) kernels, in order to improve predictions for all the tasks simultaneously. In other words, multi-task MKL can also be seen as an extension of the MKL framework, to the case of multiple tasks. However, the primary objective in MKL is indeed to learn a composite kernel best suited for a given task by combining the individual kernels optimally. Since MKL

learns the weights of the input kernels in a data-driven way, there are two alternatives proposed for the multi-task MKL approach in the literature. The first approach assumes a separate set of kernel weights for each task and regularizes them globally [69]. The second approach assumes that tasks share a common set of kernel weights that jointly optimizes all the tasks [68].

In this thesis, Publication I and Publication II assume a common set of kernel weights across tasks. In this way, the weights would then reflect the individual contribution of the kernels (or data sources) in predicting the responses of all the drugs jointly. Whereas, Publication III assumes a seperate set of kernel weights for each task, in other words the multi-task MKL model learns the task-specific kernel weights. Resultantly, the weights would then give the selective contribution of the individual kernels (or data sources) in predicting the responses for particular groups of drugs (or tasks).

Table 4.1. The combinations of approaches used in the publications included in this thesis, to apply or develop novel machine learning methods in drug response prediction. The methods are described in Chapter 5

Pub.	Kernelized regression	Matrix factorization	Multi-view learning	Multi-task learning
I	✓		✓	✓
II		✓	✓	✓
III		✓	✓	✓
IV	✓		✓	✓
V		✓	✓	✓

5. Multi-view and multi-task methods for drug response prediction

Given multiple types of data sources (described in Chapter 3) and the technical machinery (discussed in Chapter 4), it can now be possible to build novel computational models for predicting drug responses in cancer cells. The modeling task is extremely challenging due to the "small n, large p" problem. Specifically, the thesis advances in two complementary directions.

1. Methodological: combine multi-view and multi-task approach to learn an integrative model, that can naturally exploit the benefit of more data.

2. Biological: incorporate additional prior knowledge in a biologically meaningful way to solve the modeling task better.

The aim here is then to show that multi-view multi-task learning supplemented with prior biological knowledge provides better drug response predictions, compared to existing computational models (explained in Chapter 2).

5.1 Bayesian multi-task multiple kernel learning

Publication I proposes a novel Bayesian multi-task multiple kernel learning (or MT MKL) method to predict drug responses in cancer cells.
Bayesian MT MKL (illustrated in Figure 5.1) combines four modeling principles: kernelized regression, multi-view learning, multi-task learning and Bayesian inference. While kernels denote the input data sources and tasks represent the columns of the observed output matrix (shown as $\mathbf{Y}$ in Figure 5.1), the main idea is a two-step algorithm. In the first step, for each task, estimate intermediate variables from the kernels using a

fixed set of weight parameters of samples. In the second step, combine the intermediate variables using the kernel weights to estimate the output matrix. Similar to classical regression models, Bayesian MT MKL uses the input data corresponding to the rows of the output matrix. Whereas, the proposed method goes beyond commonly used computational methods in drug response analysis in several aspects.

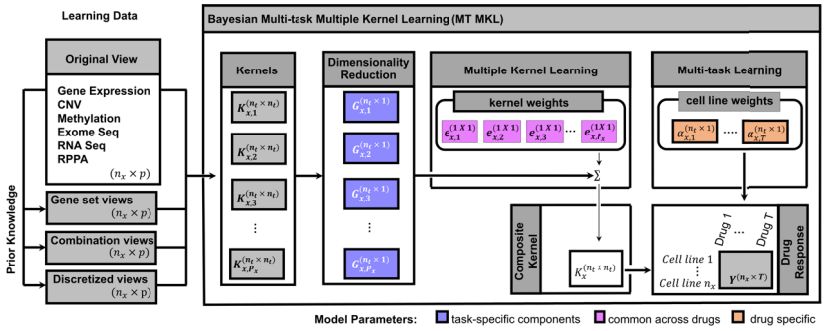


Figure 5.1. Flow diagram of the Bayesian MT MKL method. Each of the input data sources is transformed into a kernel matrix that denotes the pairwise similarities between the cells in the training data $\mathbf{K}_{x,1}...\mathbf{K}_{x,P_x}$). For each task t (here a drug), the model assumes intermediate variables $\mathbf{G}_{t,1}...\mathbf{G}_{t,P_x}$ obtained from each kernel. A weighted combination of the matrices $\mathbf{G}_{t,1}...\mathbf{G}_{t,P_x}$ parameterized by the weight vector $\boldsymbol{e}_k$ (one weight coefficient per kernel) yields the composite kernel matrix $\mathbf{K}_*$. Cell line weight vector $\boldsymbol{a}_t$ is specific to each drug, while $\boldsymbol{e}_k$ shared across all drugs.

The kernelized regression predicts drug responses from similarities between cell lines. In contrast, other methods such as linear regression, random forests, and neural networks use input features to predict drug responses. The kernel representation provides two benefits: First, the number of model parameters reduces to match the number of samples and not the number of features. It helps to control over-fitting especially in the case of "small n, large p" setup. Secondly, kernels can model nonlinear relationships between genomic and molecular features and, the drug responses.

Multi-view learning enables integration of heterogeneous data sources into a single model. Essentially, multi-view not only helps to integrate genomic and transcriptomic data sources but also allows to include various representations of the same data sources. Many computational models do not facilitate such systematic integration. For example, in linear regression, one way to use multiple sources would be to concatenate them in one data matrix (which is also called *early fusion* in data science). Moreover, the multi-view approach capitalizes prior biological knowledge in the form of

additional data sources (examples shown in Figure 5.1:“Learning Data”). In contrast to classical kernelized regression model where the predictive performance depends on the choice of a suitable kernel (i.e., choosing the functional form and its parameters), MKL learns a combination of different kernel weights to obtain a similarity measure that best matches the underlying problem. The kernel weights reflect the importance of each data source for predicting the drug responses.

Multi-task learning allows simultaneous modeling of responses and sharing of information across all the drugs. Specifically, kernel weights are shared between all the drugs, while the model learns the parameters related to the kernelized regression, specific to each drug (box “Multi- Task Learning” in Figure 5.1). Conversely, existing computational methods study drug-specific models only. Notably, the Bayesian inference and regularizations are used to handle uncertainty in the model parameters that resulted due to the small sample sizes.

Importantly, Bayesian MT MKL won the NCI-DREAM drug sensitivity prediction challenge organized by NCI and the DREAM project. The main goal of the challenge was to identify and benchmark the top performing methods in predicting drug responses from genomic, transcriptomic, epigenomic, and proteomic data sources in breast cancer cell lines. In the challenge, the Bayesian MT MKL method provided better drug response predictions, outperforming 43 sets of predictions from state-of-the-art predictive models including kernel methods, nonlinear regression (such as random forests), sparse linear regression (for instance lasso and elastic net) and, PLS or PCR.

In addition to six ‘*omics*’ data sources, Bayesian MT MKL integrated three types of additional data based on prior biological knowledge into the model (as shown in Figure 5.1: “Learning Data”), while predicting responses to all drugs simultaneously. First, gene set views aggregated summaries across functionally related genes as defined in C2 and CP collections from MSigDB. Second, combination views merged gene expression with CNV and DNA methylation to compute pathway activity scores from the PARADIGM tool and gene-wise product of the individual sources. Third, discretized views transformed real valued data sources (for instance gene expression) into binary valued sources, denoting either the gene is expressed or not expressed.

The results demonstrate that modeling nonlinearity in the data essentially provided improved predictions, and incorporating biologically

relevant additional data further increases the prediction performance. Since, MKL learned the importance of each source in a data-driven way, the gene expression and prior knowledge-based data derived from gene expression were found to be the most predictive, nevertheless the integration of all the views provided the best performance.

Publication IV applied Bayesian MT MKL in the context of personalized medicine, however in a disease different than cancer. The prediction task was to predict the response to anti-TNF (tumor necrosis factor) treatment in Rheumatoid Arthritis (RA) patients. Specifically, the study intended to assess the utility of SNP[1] data in predicting the anti-TNF treatment efficacy. For this purpose, the biologically relevant SNPs have been extracted in various ways such as association tests using $\triangle DAS28$[2] and EULAR non-response[3], the earlier published literature [72], eQTL and differential gene expression analysis. The findings demonstrate that no significant genetic contribution to prediction accuracy was found, despite there was evidence of a significant genetic heritability estimate of treatment non-response trait. The results also confirm the expectations of the RA community that standard clinical features were more predictive as compared to the SNPs, hence emphasizing to use other genome-wide data in future studies.

5.2 Kernelized Bayesian matrix factorization

Bayesian MT MKL takes input the multiple data sources for rows only. Next, it is extended to incorporate data sources, also for columns of the drug response matrix.

Publication II proposes a kernelized Bayesian matrix factorization (KBMF) method to predict drug responses on human cancer cells of multiple types. KBMF simultaneously utilizes both '*-omics*' and descriptor data as input to learn a joint model of drug responses. Essentially, KBMF factorizes a matrix by leveraging additional information from multiple data sources, as a weighted combination of these sources while learning the weights in a data-driven way. In other words, the method combines

[1] a type of mutations that denote the variation in a single nucleotide occuring at a particular position in genome

[2] absolute change in disease activity score in 28 joints following 3–6 months of anti-TNF treatment [70]

[3] EULAR response is calculated based on the pre- and post-treatment disease activity score and is widely used in clinical research and practice [71]

Bayesian matrix factorization with MKL to solve the drug response prediction problem (as shown in Figure 5.2).

KBMF, similar to Bayesian MT MKL, uses kernels to denote the input data sources. Whereas, the probabilistic generative model specifies a low dimensional factorization of the observed output matrix, in three steps. In the first step, kernel-based nonlinear dimensionality reduction projects each kernel onto the low-dimensional factors (or components) to obtain kernel-specific latent factors. In the second step, MKL integrates the kernel-specific factors with the weights to get composite latent factors ($\mathbf{H_x}$ or $\mathbf{H_z}$). In the third step, the product of the composite latent factor generates the observed matrix (matrix factorization).

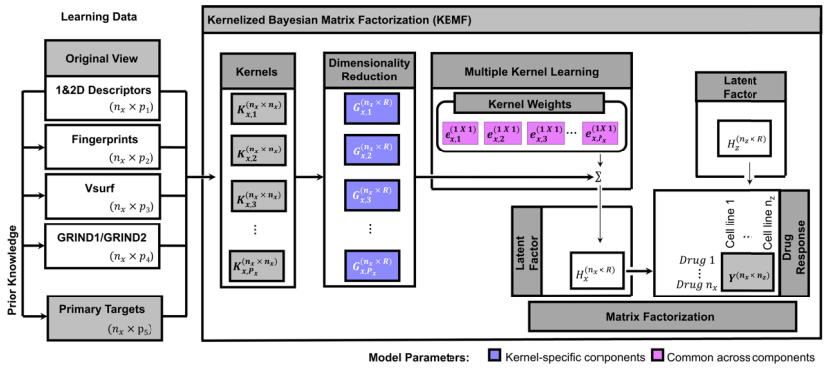


Figure 5.2. Flow diagram of kernelized Bayesian matrix factorization with multiple kernel learning. Each of the input data sources is converted into a kernel matrix $\mathbf{K}_{x,1}...\mathbf{K}_{x,P_x}$). The model assumes a low-dimensional representation $\mathbf{G}_{x,1}...\mathbf{G}_{x,P_x}$ obtained from each kernel (kernel-specific components via nonlinear dimensionality reduction). A weighted combination of the matrices $\mathbf{G}_{x,1}...\mathbf{G}_{x,P_x}$ parameterized by the weight vector e_x (one weight coefficient per kernel) yields the latent factors $\mathbf{H}_x$ (MKL). Similarly, the latent factors $\mathbf{H_z}$ for the columns can be obtained (not shown). The output matrix Y (here, denoting drug responses) is calculated as a product of latent factors $\mathbf{H}_x$ and $\mathbf{H}_z$ (matrix factorization).

The Bayesian MT MKL method uses a kernelized regression approach that combines multi-task and multi-view learning and can be viewed as an extension of single-view single-task kernelized regression. The KBMF method can then be seen, extending Bayesian MT MKL from kernelized regression to kernelized matrix factorization. All the benefits of the Bayesian MT MKL method are naturally inherited in KBMF while an additional and intriguing benefit is the assumption of multiple latent components (or factors). In contrast to Bayesian MT MKL, KBMF components take the role of tasks such that similar tasks (here drugs) are modeled with one component. Primarily, Bayesian MT MKL is a flexible method modeling

each task with a separate set of latent factors (the intermediate variables in Figure 5.1). And, KBMF further adopts the modeling assumptions, where two or more similar tasks can be modeled with one component, assuming fewer parameters, without compromising the generalizability and accuracy.

The factorization of the drug response matrix (where columns of the matrix, i.e., drugs denote separate tasks) into latent components, while integrating multiple data sources for both the rows and columns provides unprecedented flexibility in modeling. On the one hand, KBMF can capture distinct response patterns decomposed into components. This is especially advantageous to model responses to multiple drugs over multiple cells having diverse cancer types. Many other simpler computational models may not model the diverse and distinct responses with improved accuracy. On the other hand, KBMF can exploit several relevant prediction tasks. For instance, (1) predicting responses of known drugs to new cells, (2) predicting responses of new drugs on existing cells and (3) predicting responses of new drugs on new cells (visualized in Figure 4.1).

In other words, the blend of these powerful machine learning tools makes it possible to combine personalized medicine with QSAR in Publication II and address two novel drug response prediction tasks. First, an integrative QSAR task to predict the response of a new drug on multiple cells rather than to predict the response to a single cell as done in the classical studies. Second, a personalized QSAR task to predict the response of an entirely new drug on a previously unseen cell. KBMF showed increased predictive performance compared to the existing methods, for instance, neural networks, PLS, and ensembles on five out of eight benchmark QSAR data sets.

KBMF solved the new prediction tasks (stated above), on the GDSC data set comprising of 116 anti-cancer drugs and 650 cancer cell lines of diverse types. As prior knowledge, the method exploited known primary targets of the drugs in a biologically meaningful approach, in addition to multiple '-*omics*' data sources, chemical, and structural descriptors collectively, to learn a model of the drug responses.

The results substantiate that simultaneous use of both '-*omics*' and drug data sources improved the prediction performance. In particular, integrating all the data yielded better drug response predictions than descriptors or targets alone. Furthermore, the high performance in predicting missing values by the KBMF method empowered the reconstruction of

a global map of complete drug responses. The map was then explored to assess the treatment potential of therapeutically interesting anti-cancer drugs.

Kernelized Bayesian Matrix Factorization with component-wise multiple kernel learning (cwKBMF)

Publication III extends KBMF to model the complex relationships between a large number of multiple data sources (given as inputs) and the latent component space of the output matrix. This new formulation of KBMF allows component-wise multiple kernel learning (MKL), referred to as *cw*KBMF for brevity and is illustrated in Figure 5.3. The new method is similar to KBMF in two aspects: the kernel-based dimensionality reduction and matrix factorization, whereas *cw*KBMF proposes a novel formulation of multiple kernel learning.

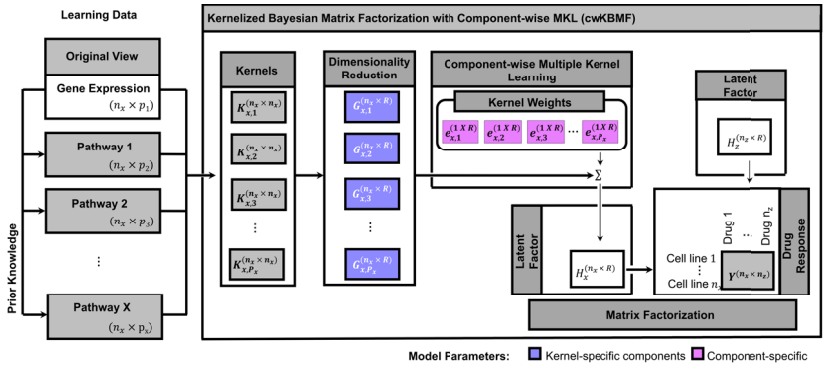


Figure 5.3. Flow diagram of kernelized matrix factorization with component-wise multiple kernel learning. This model assumes a weighted combination of the matrices $\mathbf{G}_{x,1}...\mathbf{G}_{x,P_x}$ parameterized by the weight vector $\boldsymbol{e}_{x,\cdot}^{1\times R}$ spanning the number of components R, yielding the latent factors $\mathbf{H}_x$ (component-wise multiple kernel learning). Likewise, the latent factors $\mathbf{H}_z$ for the columns can be obtained (not shown). The output matrix Y (here, containing drug responses) is calculated as a product of the latent factors $\mathbf{H}_x$ and $\mathbf{H}_z$ (matrix factorization).

The component-wise MKL benefits in learning the latent components $\mathbf{H}_x$ as a combination of kernel-specific components $\{\mathbf{G}_{x,m} \in \mathbb{R}^{n_x \times R}\}_{m=1}^{P_x}$ while segregating between kernels that are component-specific and those which are shared across all components. The new model formulation introduces component-specific kernel weights $\{\boldsymbol{e}_{x,m} \in \mathbb{R}^{1\times R}\}_{m=1}^{P_x}$ that control the activity of each kernel in each component. This extension allows the method to effectively learn the underlying structure for identifying the relationship between kernels and components.

KBMF, the predecessor method (shown in Figure 5.2), integrates multiple data sources assuming that a source is either relevant for all drugs or none and lacks to identify the component-specific relationships between the data sources and the drugs. On the contrary, *cw*KBMF solves the prediction task by gathering evidence from multiple sources, selectively, for each of the drug groups. The extension can be perceived as multi-task learning by matrix factorization, however with selective data integration. A key assumption is that component-wise MKL allows the method to exploit the data sources better. Hence, the model advances in two ways: i) improve the predictive power, and ii) identify the component-specific latent relationships between the data sources and the drug responses.

The pathway information from MSigDB can be utilized in several biologically relevant ways. For example, one way is to split the '*-omics*' data into groups of genes where each group denotes one pathway. Intuitively, the groups would then represent the multiple data sources and *cw*KBMF would model the nonlinear interactions between genes of each pathway. Since, the pathways linked to the known primary targets of the drugs are biologically meaningful, these are used to learn the drug response relationships with the *cw*KBMF method in Publication III.

The results summarized several significant findings. First, the *cw*KBMF method quantitatively outperformed the state-of-art methods (including KBMF and Bayesian MT MKL) on predicting drug responses in two publicly available drug response data sets (i.e., GDSC and CTRP). Second, the results generalized previous findings (for instance, reported in Publication I), that nonlinear models provide better drug response predictions and incorporating prior biological knowledge improves the prediction performance. Third, systematically modeling the relationships between pathway-based data sources and drug responses with *cw*KBMF, provided significantly better predictions compared to existing methods. Fourth, the use of prior knowledge not only improved the prediction performance but also helped to infer pathway-drug response relationships by *cw*KBMF, becoming the first kernelized method that made it possible to study such relationships. As a proof-of-concept, positive controls, pathway-response associations learned by the model for the well known EGFR and MEK inhibitors provided biologically meaningful interpretations. Finally, the predictive power of the *cw*KBMF method was confirmed on additional data set (fully-blinded) using the experimental validation, performed independently in the lab.

5.3 Multi-view factor analysis

The machine learning method presented in this thesis so far have demonstrated improved prediction accuracy, while modeling nonlinear relationships in the data. However, the biological interpretability of the methods is limited for several translational applications, where the goal is to identify features that are predictive of the drug responses. In this realm, linear models are easier to interpret and a natural choice. The thesis next discusses a linear multi-view model for drug response modeling.

Group Factor Analysis (GFA) is a generalization of the factor analysis model from a single data source to multiple sources [73, 74]. Specifically, GFA is a model designed to capture relationships by reducing the collection of data sources into a combined set of low-dimensional factors (or latent components). A component can be active in one or more of the sources, capturing the hidden relationships between the variables of the corresponding data sources only. For example, a component active in all the sources captures the shared relationship structure between all the data sources while a component active in a single source identifies features specific to that particular source only. In drug response prediction, the component active in the genomic, transcriptomic and drug response data sources, represents relationships between the genome-wide features and drug responses, yielding predictability from genes to drugs.

Publication V presents an R package that implements a full Bayesian formulation of the GFA model with a Gibbs sampling approach. Specifically, the package defines sparse priors for the latent factors to tackle "small n, large p" problem as well as to improve the biological interpretations. Another related challenge is to infer the number of components (model complexity), needed to explain the relationships in the data. The GFA priors can detect the true model complexity automatically. Publication V demonstrates the model complexity selection mechanism using the GDSC data set. The package supports optimization of the model complexity selection by assuming *apriori* particular signal-to-noise ratio. The package also facilitates to obtain a set of robust components that occur across multiple factorizations, whose interpretations are likely to be more biologically plausible and statistically significant. Lastly, the GFA package provides functionality to explore and visualize the component activities that define the factorization of the data sources. In summary, the package provides a complete data analysis pipeline to support drug response analysis in

future studies.

5.4 Incorporating prior knowledge from experts

In Sections 5.1 and 5.2, the thesis proposes methods that incorporate prior knowledge extracted from public databases, such as known cancer genes and pathways or gene sets. This use of prior knowledge has shown to provide better drug response predictions in the "small n, large p" setting. However, when the sample sizes are remarkably small, the data-driven prediction task becomes progressively harder and requires more information. A valuable source could be an expert (in this case a clinician or a researcher), who could provide useful information on the relevance of the features for the prediction task. However, extracting relevance information for thousands of features (here genes) from the expert is not practically feasible.

Publication VI presents an approach to efficiently extract prior knowledge from the expert by asking feedback on a limited set of genomic features. The study compared three strategies to identify the most important features on which to ask expert feedback. The first strategy selected features at random, while the second strategy chose the features based on their largest absolute values. Whereas, the third strategy selected the top features to ask expert feedback, whose ranks were determined based on the largest point-wise product of the features and their estimated regression coefficients. Experiments performed with simulated experts, and drug response dataset from GDSC demonstrate that the third strategy provided the best performance, while the differences in performances increases as the expert gave more feedback.

The results signify that the prior knowledge, efficiently extracted from the experts, provides improved accuracy for predicting drug responses. The proposed approach intends to be a simple proof-of-concept study and a starting point for developing advanced approaches to extract and incorporate prior knowledge from experts.

6. Discussion and conclusion

Returning to the research question raised at the beginning of this thesis: how to improve drug response predictions, given the computational challenge of "small n, large p" in personalized medicine. It is now possible to answer that drug response predictions can be improved using multi-view multi-task methods supplemented with prior knowledge. The methods proposed in this thesis can effectively integrate multiple data sources and apply a joint model of all the drugs, yielding better predictions as compared to commonly used methods. Specifically, this thesis presents three novel, multi-view multi-task methods, each for different but increasingly complex and novel formulations of the drug response prediction problem. It can be claimed explicitly that the multi-view models can extract useful, shared information from multiple sources, which individual data source alone cannot provide.

The thesis demonstrates that the proposed methods provide better drug response predictions while they progress in the hierarchy. For instance, *cw*KBMF outperforms both KBMF and Bayesian multi-task MKL in predicting drug responses of new cells. Another significant finding of the thesis is that modeling nonlinear relationships in the form of kernels provides improved drug response predictions. In particular, kernelized regression, combined with multi-task and multiple-kernel learning, showed increased predictive performance as compared to the other widely used methods. This finding opens doors to a new line of research in the personalized medicine domain. Notably, the fully-blinded experimental validation presented in Publication III confirms the predictive power of *cw*KBMF. The experimental validation builds confidence that in-silico predictions can be reasonably robust and may be used to explore the spectrum of therapeutic choices in future studies.

One of the more significant and novel findings to emerge from this

thesis is the use of prior biological knowledge to improve drug response predictions. The thesis shows several data-driven ways of incorporating prior biological knowledge into computational models extracted from public databases. On the machine learning front, a simple approach has also been presented that explores ways to extract prior knowledge from experts. The results support the need for incorporating experts' knowledge in the modeling loop (Publication VI). Incorporating prior knowledge not only gives better predictions but also allows identifying potential relationships between pathways and drugs, which is predictive of responses in cancer cells. Essentially, the use of appropriate prior biological knowledge helps to tackle the "small n, large p" problem. In addition to other related findings, gene expressions and prior knowledge-based data sources derived from gene expressions provide the best predictive accuracy.

Consequently, the methods presented in this thesis could possibly be adopted to assist clinicians in choosing effective therapies for individual patients. The drug response predictions generated by the proposed methods could be useful to medical researchers, both to pre-select potential drugs for further screening and to enhance their understanding of the functional mechanisms of the drugs. In summary, this research takes us a step closer to achieving the personalized and targeted interventions of drugs. This thesis demonstrates progress in both machine learning and personalized medicine, emphasizing that interdisciplinary research plays an important role, not only in achieving mutual goals, but also in generating ideas for scientific advances within each discipline. In particular, the release of source codes and an R-package is beneficial to both communities. On the medical side, researchers can use the package to identify potential biomarkers when analyzing relationships between genomic and molecular features and the drug responses. On the other hand, machine learners can utilize the freely available code to make advancements on the modeling side.

However, the thesis also identifies areas of application where the proposed methods do not improve the drug response predictions. In particular, those applications that use SNP-based genotype data (Publication IV and [75]). It can be argued that either the data is not suitable for studying such problems, or there is a need to develop a more appropriate methodology to study the problem correctly. The SNP data is extremely high-dimensional (~ 2 billion features) and essentially discrete in nature (0, 1, and 2 count data). With the kernelized methods, the employed kernel functions may simply not be enough to correctly model the relationships between the

samples.

This thesis reveals several possibilities for future research, both in applications and methods development. Any other domain, in which systematic integration of multiple, related data sources is of interest, can benefit from the proposed multi-view and multi-task methods. For example, recommender engines where multiple data sources are available nowadays. As the methods can integrate multiple data sources, challenging prediction tasks such as out-of-matrix prediction can effectively be addressed. Practically, the methods can predict ratings for a new movie or can recommend new items to a customer, in addition to predicting the missing ratings or recommendations. In comparison to many conventional rating prediction algorithms, the proposed methods may provide improved predictions, however may compromise the scalability.

While the thesis proposes kernel-based methods, the limited interpretability in the feature space is a key factor hindering their usabilities in translational studies. The thesis creates new opportunities in drug response prediction, for developing methods that are interpretable in the original feature spaces; finding a suitable trade-off between linear and nonlinear models. Moreover, certain topics in drug response prediction have not been explored in this thesis and could be investigated in future studies. For example, the effect of the differences in the drug response measurements arising from the different experimental protocols [76, 77, 78] and the effect of '-*omics*'-based quantitative analysis using the cell line data on predictions. Though the use of cancer cell lines primarily serves as good preclinical models, the patterns of sensitivity in cell lines may be different from a drug's response in vivo, offering a limited understanding of the human pharmacology [79]. Since increasing amounts of data are becoming available especially emerging from ex-vivo patient samples, the proposed methods could be used to predict drug responses; a step closer to clinical applications in future.

This thesis also creates new directions for the current research questions in personalized medicine. For instance, it would be interesting to explore the abilities to predict drug combinations by the proposed methods. Alternatively, when the sample size approaches to n = 1 of a completely personalized scenario (e.g., predicting treatment outcome of an individual patient), use of prior knowledge becomes inevitable. Assuming all the prior knowledge available in public databases has already been incorporated, a remaining source of information is an expert; however, the knowledge

is often tacit and uncertain for any specific patient. There is a need to develop new interactive machine learning methods that combine principles of human-computer interaction and multi-view learning. The interactive methods effectively elicit tacit knowledge from the expert as well as simultaneously improve model predictions while incorporating the elicited prior knowledge [80]. Furthermore, it can be interesting to explore the new design formulations for incorporating pathway and target knowledge. A plausible way may be to develop informative prior to integrating the knowledge. This approach could potentially support generating the biological hypotheses for the data-driven analysis. Also, the technical choices and sparsity assumptions of the model may be improved further, as required by an application.

In summary, the thesis contributes Bayesian multi-view multi-task methods for predicting drug response in cancer cells. The thesis is a valuable scientific advancement because predicting such responses can significantly enhance our understanding of the action mechanism of anti-cancer drugs and may ultimately assist in personalizing treatments for individual patients.

Bibliography

[1] 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.

[2] 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.

[3] 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.

[4] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 2, pp. 301–320, 2005.

[5] A. Hoerl and R. Kennard, "Ridge regression, in 'Encyclopedia of Statistical Sciences', vol. 8," 1988.

[6] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

[7] A. Basu, N. E. Bodycombe, J. H. Cheah, E. V. Price, K. Liu, G. I. Schaefer, R. Y. Ebright, M. L. Stewart, D. Ito, S. Wang, *et al.*, "An interactive resource to identify cancer genetic and lineage dependencies targeted by small molecules," *Cell*, vol. 154, no. 5, pp. 1151–1161, 2013.

[8] F. Iorio, T. A. Knijnenburg, D. J. Vis, G. R. Bignell, M. P. Menden, M. Schubert, N. Aben, E. Gonçalves, S. Barthorpe, H. Lightfoot, *et al.*, "A landscape of pharmacogenomic interactions in cancer," *Cell*, vol. 166, no. 3, pp. 740–754, 2016.

[9] P. Geladi and B. R. Kowalski, "Partial least-squares regression: a tutorial," *Analytica Chimica Acta*, vol. 185, pp. 1–17, 1986.

[10] T. Dijkstra, "Some comments on maximum likelihood and partial least squares methods," *Journal of Econometrics*, vol. 22, no. 1, pp. 67–90, 1983.

[11] I. S. Jang, E. C. Neto, J. Guinney, S. H. Friend, and A. A. Margolin, "Systematic assessment of analytical methods for drug sensitivity prediction

from cancer cell line data," in *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, p. 63, NIH Public Access, 2014.

[12] P. Geeleher, N. J. Cox, and R. S. Huang, "Clinical drug response can be predicted using baseline gene expression levels and in vitro drug sensitivity in cell lines," *Genome Biology*, vol. 15, no. 3, p. 1, 2014.

[13] Z. Dong, N. Zhang, C. Li, H. Wang, Y. Fang, J. Wang, and X. Zheng, "Anticancer drug sensitivity prediction in cell lines from baseline gene expression through recursive feature selection," *BMC cancer*, vol. 15, no. 1, p. 489, 2015.

[14] V. N. Vapnik and V. Vapnik, *Statistical learning theory*, vol. 1. Wiley New York, 1998.

[15] H. Hejase and C. Chan, "Improving drug sensitivity prediction using different types of data," *CPT: Pharmacometrics and Systems Pharmacology*, vol. 4, no. 2, pp. 98–105, 2015.

[16] Y. Yamanishi, E. Pauwels, and M. Kotera, "Drug side-effect prediction based on the integration of chemical and biological spaces," *Journal of Chemical Information and Modeling*, vol. 52, no. 12, pp. 3284–3292, 2012.

[17] C. E. Rasmussen, "Gaussian processes in machine learning," in *Advanced lectures on machine learning*, pp. 63–71, Springer, 2004.

[18] B. Scholkopf and A. Smola, *Learning with kernels: Support vector machines, regularization, optimization, and beyond*. MIT Press, 2001.

[19] J. Mercer, "Functions of positive and negative type and their connection with the theory of integral equations," *Philosophical Transactions of the Royal Society of London. Series A)*, vol. 83, no. 559, pp. 69–70, 1909.

[20] H. Fröhlich, "Kernel methods in chemo- and bioinformatics," *Ph.D Thesis, University of Tübingen*, 2006.

[21] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[22] G. Riddick, H. Song, S. Ahn, J. Walling, D. Borges-Rivera, W. Zhang, and H. A. Fine, "Predicting in vitro drug sensitivity using random forests," *Bioinformatics*, vol. 27, no. 2, pp. 220–224, 2011.

[23] Q. Wan and R. Pal, "An ensemble based top performing approach for nci-dream drug sensitivity prediction challenge," *PloS One*, vol. 9, no. 6, p. e101183, 2014.

[24] J. D. Ospina, J. Zhu, C. Chira, A. Bossi, J. B. Delobel, V. Beckendorf, B. Dubray, J.-L. Lagrange, J. C. Correa, A. Simon, *et al.*, "Random forests to predict rectal toxicity following prostate cancer radiation therapy," *International Journal of Radiation Oncology* Biology* Physics*, vol. 89, no. 5, pp. 1024–1031, 2014.

[25] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[26] J. J. Sutherland, L. A. O'Brien, and D. F. Weaver, "A comparison of methods for modeling quantitative structure-activity relationships," *Journal of Medicinal Chemistry*, vol. 47, no. 22, pp. 5541–5554, 2004.

[27] M. P. Menden, F. Iorio, M. Garnett, U. McDermott, C. Benes, P. J. Ballester, and J. Saez-Rodriguez, "Machine learning prediction of cancer cell sensitivity to drugs based on genomic and chemical properties," *PloS One*, vol. 8, no. 4, p. e61318, 2013.

[28] J. Ma, R. P. Sheridan, A. Liaw, G. E. Dahl, and V. Svetnik, "Deep neural nets as a method for quantitative structure–activity relationships," *Journal of Chemical Information and Modeling*, vol. 55, no. 2, pp. 263–274, 2015.

[29] M. J. Heller, "Dna microarray technology: devices, systems, and applications," *Annual Review of Biomedical Engineering*, vol. 4, no. 1, pp. 129–153, 2002.

[30] 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.

[31] F. Cappuzzo, F. R. Hirsch, E. Rossi, S. Bartolini, G. L. Ceresoli, L. Bemis, J. Haney, S. Witta, K. Danenberg, I. Domenichini, *et al.*, "Epidermal growth factor receptor gene and protein and gefitinib sensitivity in non–small-cell lung cancer," *Journal of the National Cancer Institute*, vol. 97, no. 9, pp. 643–655, 2005.

[32] S. B. Ng, E. H. Turner, P. D. Robertson, S. D. Flygare, A. W. Bigham, C. Lee, T. Shaffer, M. Wong, A. Bhattacharjee, E. E. Eichler, *et al.*, "Targeted capture and massively parallel sequencing of 12 human exomes," *Nature*, vol. 461, no. 7261, pp. 272–276, 2009.

[33] M. Bibikova, B. Barnes, C. Tsan, V. Ho, B. Klotzle, J. M. Le, D. Delano, L. Zhang, G. P. Schroth, K. L. Gunderson, *et al.*, "High density dna methylation array with single cpg site resolution," *Genomics*, vol. 98, no. 4, pp. 288–295, 2011.

[34] B. Spurrier, S. Ramalingam, and S. Nishizuka, "Reverse-phase protein lysate microarrays for cell signaling analysis," *Nature protocols*, vol. 3, no. 11, pp. 1796–1808, 2008.

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

[36] S. L. Holbeck, J. M. Collins, and J. H. Doroshow, "Analysis of food and drug administration–approved anticancer agents in the nci60 panel of human tumor cell lines," *Molecular Cancer Therapeutics*, vol. 9, no. 5, pp. 1451–1460, 2010.

[37] R. Mannhold, H. Kubinyi, G. Folkers, and G. Cruciani, *Molecular interaction fields: applications in drug discovery and ADME prediction*, vol. 27. John Wiley & Sons, 2006.

[38] G. Cruciani, P. Crivori, P.-A. Carrupt, and B. Testa, "Molecular fields in quantitative structure–permeation relationships: the volsurf approach," *Journal of Molecular Structure: THEOCHEM*, vol. 503, no. 1, pp. 17–30, 2000.

[39] 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.

[40] M. Pastor, G. Cruciani, I. McLay, S. Pickett, and S. Clementi, "Grid-independent descriptors (grind): a novel class of alignment-independent three-dimensional molecular descriptors," *Journal of Medicinal Chemistry*, vol. 43, no. 17, pp. 3233–3243, 2000.

[41] Á. Durán, G. C. Martínez, 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.

[42] A. Liberzon, A. Subramanian, R. Pinchback, H. Thorvaldsdóttir, P. Tamayo, and J. P. Mesirov, "Molecular signatures database (msigdb) 3.0," *Bioinformatics*, vol. 27, no. 12, pp. 1739–1740, 2011.

[43] S. A. Forbes, D. Beare, P. Gunasekaran, K. Leung, N. Bindal, H. Boutselakis, M. Ding, S. Bamford, C. Cole, S. Ward, C. Y. Kok, M. Jia, T. De, J. W. Teague, M. R. Stratton, U. McDermott, and P. J. Campbell, "Cosmic: exploring the world's knowledge of somatic mutations in human cancer," *Nucleic Acids Research*, vol. 43, no. D1, pp. D805–D811, 2015.

[44] C. J. Vaske, S. C. Benz, J. Z. Sanborn, D. Earl, C. Szeto, J. Zhu, D. Haussler, and J. M. Stuart, "Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using paradigm," *Bioinformatics*, vol. 26, no. 12, pp. i237–i245, 2010.

[45] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin, *Bayesian data analysis*, vol. 2. Chapman & Hall/CRC Boca Raton, FL, USA, 2014.

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

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

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

[49] 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.

[50] 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.

[51] 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.

[52] 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.

[53] 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.

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

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

[56] 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.

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

[58] M. Gönen and E. Alpaydın, "Multiple kernel learning algorithms," *Journal of Machine Learning Research*, vol. 12, pp. 2211–2268, 2011.

[59] M. Gönen, "Bayesian efficient multiple kernel learning," in *Proceedings of the 29th International Conference on Machine Learning*, pp. 1–8, 2012.

[60] R. Caruana, "Multitask learning," *Machine Learning*, vol. 28, no. 1, pp. 41–75, 1997.

[61] J. Baxter, "A bayesian/information theoretic model of learning to learn via multiple task sampling," *Machine Learning*, vol. 28, no. 1, pp. 7–39, 1997.

[62] J. Baxter, "A model of inductive bias learning," *Journal of Artificial Intellengence Research (JAIR)*, vol. 12, no. 149-198, p. 3, 2000.

[63] B. A. Turlach, W. N. Venables, and S. J. Wright, "Simultaneous variable selection," *Technometrics*, vol. 47, no. 3, pp. 349–363, 2005.

[64] G. Obozinski, B. Taskar, and M. Jordan, "Multi-task feature selection," *Statistics Department, UC Berkeley, Tech. Rep*, vol. 2, pp. 1–15, 2006.

[65] Y. Zhang, D.-Y. Yeung, and Q. Xu, "Probabilistic multi-task feature selection," in *Advances in neural information processing systems*, pp. 2559–2567, 2010.

[66] A. Rakotomamonjy, F. R. Bach, S. Canu, and Y. Grandvalet, "Simplemkl," *Journal of Machine Learning Research*, vol. 9, no. Nov, pp. 2491–2521, 2008.

[67] P. Jawanpuria and J. S. Nath, "Multi-task multiple kernel learning," in *Proceedings of the 2011 SIAM International Conference on Data Mining*, pp. 828–838, SIAM, 2011.

[68] M. Gönen, "A bayesian multiple kernel learning framework for single and multiple output regression," in *Proceedings of the 20th European Conference on Artificial Intelligence*, pp. 354–359, IOS Press, 2012.

[69] M. Kandemir, A. Vetek, M. Gönen, A. Klami, and S. Kaski, "Multi-task and multi-view learning of user state," *Neurocomputing*, vol. 139, pp. 97–106, 2014.

[70] M. Prevoo, M. Van't Hof, H. Kuper, M. Van Leeuwen, L. Van De Putte, and P. Van Riel, "Modified disease activity scores that include twenty-eight-joint counts development and validation in a prospective longitudinal study of patients with rheumatoid arthritis," *Arthritis & Rheumatism*, vol. 38, no. 1, pp. 44–48, 1995.

[71] A. Van Gestel, M. Prevoo, M. Van't Hof, M. Van Rijswijk, L. Van de Putte, and P. Van Riel, "Development and validation of the european league against rheumatism response criteria for rheumatoid arthritis: comparison with the preliminary american college of rheumatology and the world health organization/international league against rheumatism criteria," *Arthritis and Rheumatism*, vol. 39, no. 1, pp. 34–40, 1996.

[72] Y. Okada, D. Wu, G. Trynka, T. Raj, C. Terao, K. Ikari, Y. Kochi, K. Ohmura, A. Suzuki, S. Yoshida, *et al.*, "Genetics of rheumatoid arthritis contributes to biology and drug discovery," *Nature*, vol. 506, no. 7488, pp. 376–381, 2014.

[73] S. A. Khan, S. Virtanen, O. P. Kallioniemi, K. Wennerberg, A. Poso, and S. Kaski, "Identification of structural features in chemicals associated with cancer drug response: a systematic data-driven analysis," *Bioinformatics*, vol. 30, no. 17, pp. i497–i504, 2014.

[74] K. Bunte, E. Leppäaho, I. Saarinen, and S. Kaski, "Sparse group factor analysis for biclustering of multiple data sources," *Bioinformatics*, vol. 32, no. 16, pp. 2457–2463, 2016.

[75] F. Eduati, L. M. Mangravite, T. Wang, H. Tang, J. C. Bare, R. Huang, T. Norman, M. Kellen, M. P. Menden, J. Yang, *et al.*, "Prediction of human population responses to toxic compounds by a collaborative competition," *Nature biotechnology*, vol. 33, no. 9, pp. 933–940, 2015.

[76] B. Haibe-Kains, N. El-Hachem, N. J. Birkbak, A. C. Jin, A. H. Beck, H. J. Aerts, and J. Quackenbush, "Inconsistency in large pharmacogenomic studies," *Nature*, vol. 504, no. 7480, pp. 389–393, 2013.

[77] P. Geeleher, E. R. Gamazon, C. Seoighe, N. J. Cox, and R. S. Huang, "Consistency in large pharmacogenomic studies," *Nature*, vol. 540, no. 7631, pp. E1–E2, 2016.

[78] J. P. Mpindi, B. Yadav, P. Östling, P. Gautam, D. Malani, A. Murumägi, A. Hirasawa, S. Kangaspeska, K. Wennerberg, O. Kallioniemi, *et al.*, "Consistency in drug response profiling," *Nature*, vol. 540, no. 7631, pp. E5–E6, 2016.

[79] J. N. Weinstein, "Drug discovery: Cell lines battle cancer," *Nature*, vol. 483, no. 7391, pp. 544–545, 2012.

[80] L. Micallef, I. Sundin, P. Marttinen, M. Ammad-ud din, T. Peltola, M. Soare, G. Jacucci, and S. Kaski, "Interactive elicitation of knowledge on feature relevance improves predictions in small data sets," in *Proceedings of the 22nd International Conference on Intelligent User Interfaces*, pp. 2–6, ACM, 2017.

Personalizing medicine is one of the prime challenges in cancer treatment. Computationally, **machine learning models** can predict drug responses in cancer cells utilizing the high-throughput data available now. The drug response predictions can then be analyzed to generate hypotheses for selecting therapies tailored to individual patients. However, the prediction task is exceedingly challenging, due to the small sample sizes and large dimensionality of the data. To improve drug response prediction in cancer, this thesis presents novel **Bayesian multi-task multi-view machine learning methods.** The key assumptions are that the integration of several data sources extracts more relevant information and structure from the data ***(multi-view)*** and the joint modeling of multiple drugs provides a statistical boost when learned together ***(multi-task)***, are valuable for the prediction task. Moreover, new ways of incorporating prior biological knowledge are explored to further improve the predictions. Open source implementations of the new methods have been released to facilitate further research.

ISBN 978-952-60-7514-3 (printed)
ISBN 978-952-60-7513-6 (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

Aalto-DD 127/2017