Continuous Conditional Random Fields
for Drug Target Interaction Prediction

by

Marten Heidemeyer

B.Sc., Universität zu Lübeck, 2013

Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of
Master of Science

in the
School of Computing Science
Faculty of Applied Sciences

© Marten Heidemeyer 2016
SIMON FRASER UNIVERSITY
Fall 2016

All rights reserved.
However, in accordance with the Copyright Act of Canada, this work may be
reproduced without authorization under the conditions for “Fair Dealing.”
Therefore, limited reproduction of this work for the purposes of private study,
research, education, satire, parody, criticism, review and news reporting is likely
to be in accordance with the law, particularly if cited appropriately.
Approval

Name: Marten Heidemeyer
Degree: Master of Science (Computing Science)
Title: Continuous Conditional Random Fields for Drug Target Interaction Prediction

Examining Committee: Chair: Faraz Hach
Research Associate

Martin Ester
Senior Supervisor
Professor

Artem Cherkasov
Senior Research Scientist, Vancouver
Prostate Center (Former Adjunct Professor at Simon Fraser University)

Leonid Chindelevitch
Assistant/Associate/Professor
School/Department or Faculty

Date Defended: October 3rd, 2016
Abstract

Knowledge about the interaction between drugs and target proteins is essential in the drug discovery process. Understanding the relationship between compounds and proteins through wetlab experiments alone is time-consuming and costly. With the motivation to support the experimental work by systematically prioritizing the most potent compounds for a target, numerous methods for the in-silico prediction of drug-target interaction have been proposed recently and high performance on binary datasets, where drug-target pairs are classified as either binding or non-binding have been reported. A possible drawback of binary datasets is that missing values and non-interacting drug-target pairs are not differentiated. In this thesis, a model is developed that predicts the binding strengths of drug-target pairs as continuous values and thus incorporates the whole interaction spectrum from true negative to true positive interactions in the learning phase. The developed model combines two previously used approaches for the drug-target problem, which are Matrix Factorization and Conditional Random Fields. The model is evaluated in terms of the metrics $AUC$, $AUPR$ and $CI$ on three datasets and a slight improvement in performance is observed when compared to the state of the art method.

**Keywords:** Drug-target interaction, predictive modeling, machine learning
Dedication

I dedicate this work to my mother, my father and all my friends
Acknowledgements

I would like to express my gratitude to my senior supervisor Prof. Martin Ester for his guidance, patience and support throughout the process of my master studies. I feel specially grateful for having worked under Prof. Ester’s supervision not only for the invaluable advice on this thesis but also for the encouragement and inspiration to appreciate the world and life outside of the lab. Additionally, I would like to thank Dr. Artem Cherkasov for being my supervisor, for the helpful discussions and for the support. Further, I have to thank Prof. Barbara Hammer who gave me valuable input during my research exchange with the University of Bielefeld. I also thank Dr. Leonid Chindelevitch for serving as the internal examiner of this thesis and Faraz Harach for serving as a graduate chair. Special thanks go to my lab mates for their companionship: Tong He, Dr. Yao Wu, Xin Wang, Beidou Wang, Mehrdad Mansouri, Sahand Khakabimamaghani, Boshra Nabaei, Jiaxi Tang, Weipeng Lin, Hongwei Liang, Zhilin Zhang and Yue Wang. Finally, my sincerest gratitude goes to my parents for their endless love and support through all the years.
# Table of Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Approval</td>
<td>ii</td>
</tr>
<tr>
<td>Abstract</td>
<td>iii</td>
</tr>
<tr>
<td>Dedication</td>
<td>iv</td>
</tr>
<tr>
<td>Acknowledgements</td>
<td>v</td>
</tr>
<tr>
<td>Table of Contents</td>
<td>vi</td>
</tr>
<tr>
<td>List of Tables</td>
<td>ix</td>
</tr>
<tr>
<td>List of Figures</td>
<td>x</td>
</tr>
<tr>
<td>1 Introduction</td>
<td>1</td>
</tr>
<tr>
<td>1.1 Terms and Definitions</td>
<td>2</td>
</tr>
<tr>
<td>2 Review of Literature</td>
<td>4</td>
</tr>
<tr>
<td>2.1 Yoshihiro Yamanishi, &quot;Prediction of drug-target interaction networks from the integration of chemical and genomic spaces.&quot;, 2008</td>
<td>5</td>
</tr>
<tr>
<td>2.2 Twan van Laarhoven, &quot;Gaussian interaction profile kernels for predicting drug-target interaction.&quot;, 2011</td>
<td>8</td>
</tr>
<tr>
<td>2.3 Fan Yang, &quot;Drug-target interaction prediction by integrating chemical, genomic, functional and pharmacological data.&quot;, 2014</td>
<td>9</td>
</tr>
<tr>
<td>2.4 Yong Liu, &quot;Neighborhood Regularized Logistic Matrix Factorization for Drug-Target Interaction Prediction&quot;, 2016</td>
<td>12</td>
</tr>
<tr>
<td>Section</td>
<td>Title</td>
</tr>
<tr>
<td>---------</td>
<td>------------------------------------------------------------------------</td>
</tr>
<tr>
<td>2.5</td>
<td>Tapio Pahikkala, &quot;Toward more realistic drug-target interaction predictions.&quot;, 2014</td>
</tr>
<tr>
<td>3</td>
<td>Method</td>
</tr>
<tr>
<td>3.1</td>
<td>Model Intuition</td>
</tr>
<tr>
<td>3.2</td>
<td>Notation</td>
</tr>
<tr>
<td>3.3</td>
<td>Matrix Factorization</td>
</tr>
<tr>
<td>3.3.1</td>
<td>Parameter Learning</td>
</tr>
<tr>
<td>3.3.2</td>
<td>Inference</td>
</tr>
<tr>
<td>3.4</td>
<td>Continuous Conditional Random Fields</td>
</tr>
<tr>
<td>3.4.1</td>
<td>Formal Definition</td>
</tr>
<tr>
<td>3.4.2</td>
<td>Parameter Learning</td>
</tr>
<tr>
<td>3.4.3</td>
<td>Inference</td>
</tr>
<tr>
<td>3.5</td>
<td>MF+CCRF</td>
</tr>
<tr>
<td>3.5.1</td>
<td>Integrating the similarity matrices</td>
</tr>
<tr>
<td>3.5.2</td>
<td>Graphical structure of the CCRF</td>
</tr>
<tr>
<td>3.6</td>
<td>Proof of Concept</td>
</tr>
<tr>
<td>3.6.1</td>
<td>Data Simulation</td>
</tr>
<tr>
<td>3.6.2</td>
<td>Experiments on Simulated Data</td>
</tr>
<tr>
<td>3.6.3</td>
<td>Summary</td>
</tr>
<tr>
<td>4</td>
<td>Experiments</td>
</tr>
<tr>
<td>4.1</td>
<td>Datasets</td>
</tr>
<tr>
<td>4.1.1</td>
<td>The Davis Dataset</td>
</tr>
<tr>
<td>4.1.2</td>
<td>The Metz Dataset</td>
</tr>
<tr>
<td>4.1.3</td>
<td>The KIBA Dataset</td>
</tr>
<tr>
<td>4.1.4</td>
<td>Similarity Metrics of Drugs and Targets</td>
</tr>
<tr>
<td>4.2</td>
<td>Experimental Settings</td>
</tr>
<tr>
<td>4.2.1</td>
<td>Binary vs. Continuous Prediction</td>
</tr>
<tr>
<td>4.2.2</td>
<td>Performance Evaluation</td>
</tr>
</tbody>
</table>
4.3 Results ................................................................. 56

5 Discussion and Future Work ...................................... 60
  5.1 Discussion ......................................................... 60
  5.2 Future Work ....................................................... 61

Bibliography ............................................................ 63
List of Tables

Table 3.1 $RMSE$ of $MF$, $CCRF$ and $MF + CCRF$ on simulated dataset. 41
Table 4.1 Statistics of the used evaluation datasets. 44
Table 4.2 $CI$ of $KronRLS$ and $MF + CCRF$ on the three evaluation datasets. 58
Table 4.3 $AUC$ of $KronRLS$ and $MF + CCRF$ on the three evaluation datasets. For a description of the table entries see Figure 4.12 . 58
Table 4.4 $AUPR$ of $KronRLS$ and $MF + CCRF$ on the three evaluation datasets. 59
### List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Illustration of Supervised Bipartite Graph Inference Method</td>
<td>7</td>
</tr>
<tr>
<td>2.2</td>
<td>Illustration of Gaussian Interaction Profile Kernel Method</td>
<td>10</td>
</tr>
<tr>
<td>3.1</td>
<td>Illustration of steps to get vector $X$</td>
<td>32</td>
</tr>
<tr>
<td>3.2</td>
<td>Illustration of Parameter Learning</td>
<td>33</td>
</tr>
<tr>
<td>3.3</td>
<td>Integrating similarity matrices</td>
<td>35</td>
</tr>
<tr>
<td>3.4</td>
<td>Davis drugs clustering</td>
<td>39</td>
</tr>
<tr>
<td>3.5</td>
<td>Sampling CCRF input</td>
<td>39</td>
</tr>
<tr>
<td>3.6</td>
<td>Defining the structure of the CRF</td>
<td>40</td>
</tr>
<tr>
<td>3.7</td>
<td>Example of CCRF sample</td>
<td>40</td>
</tr>
<tr>
<td>3.8</td>
<td>Generated dataset.</td>
<td>41</td>
</tr>
<tr>
<td>3.9</td>
<td>Training dataset.</td>
<td>41</td>
</tr>
<tr>
<td>3.10</td>
<td>$MF$ prediction</td>
<td>41</td>
</tr>
<tr>
<td>3.11</td>
<td>$MF + CCRF$ prediction</td>
<td>41</td>
</tr>
<tr>
<td>4.1</td>
<td>Distribution of $pK_d$ values given in the <em>Davis</em> dataset</td>
<td>45</td>
</tr>
<tr>
<td>4.2</td>
<td>Distribution of $K_i$ values given in the <em>Metz</em> dataset</td>
<td>46</td>
</tr>
<tr>
<td>4.3</td>
<td>Distribution of KIBA scores given in the <em>KIBA</em> dataset</td>
<td>47</td>
</tr>
<tr>
<td>4.4</td>
<td>Correlation of drug-similarity and binding behaviour, similar drug pairs, <em>Metz</em> dataset</td>
<td>50</td>
</tr>
<tr>
<td>4.5</td>
<td>Correlation of drug-similarity and binding behaviour, un-similar drug pairs, <em>Metz</em> dataset</td>
<td>50</td>
</tr>
<tr>
<td>4.6</td>
<td>Correlation of drug-similarity and binding behaviour, similar drug pairs, <em>KIBA</em> dataset</td>
<td>51</td>
</tr>
</tbody>
</table>
Figure 4.7 Correlation of drug-similarity and binding behaviour, un-similar drug pairs, KIBA dataset ........................................... 51
Figure 4.8 Correlation of drug-similarity and binding behaviour, similar drug pairs, Davis dataset ..................................... 52
Figure 4.9 Correlation of drug-similarity and binding behaviour, un-similar drug pairs Davis dataset ................................. 52
Figure 4.10 Correlation of target-similarity and binding behaviour, similar target pairs, Davis dataset ............................. 53
Figure 4.11 Correlation of target-similarity and binding behaviour, un-similar target pairs, Davis dataset ............................ 53
Figure 4.12 Description of evaluation tables ................................. 55
Chapter 1

Introduction

In the field of pharmacology, drug discovery is the process through which new medications are brought to the market. To date, drug discovery is a highly expensive and difficult process with a low success rate, despite advances in technology and understanding of biological systems. Although the investment in drug research and development has increased substantially in recent years, the number of newly approved drugs by the US Food and Drug Administration (FDA) per year has not [19]. Since the sequencing of the human genome, the drug development process is mainly focused on proteins that are hypothesized to play a key role in diseases for which a drug is developed. Identifying chemical structures that modify the activity of these target proteins is a fundamental challenge in drug discovery. The most frequently used approach today is known as reverse pharmacology which involves large libraries of chemical compounds which need to be screened against the target proteins which are thought to be linked to the disease [19]. The identification of compounds which physically bind to the disease target is the key challenge here and thus the knowledge about the interaction strength between chemical structures and proteins is an important topic in drug development. The safest and most accurate method to gain knowledge about the interaction strength of drug candidates and target proteins is through wetlab experiments. On the other hand, wetlab experiments are costly in terms of time and money, as there are thousands of potential drug candidates. The
failure of a new ligand in toxicity tests is higher than 90% which is the most significant reason for the high cost of the drug development process.

In drug development, *drug repositioning* is a technique in which known drugs and drug candidates are used to treat new diseases. Existing drugs may bind to the target protein of a disease other than the disease that the drug was originally developed for. Using an existing drug as a basis for the development of a new drug is far more likely to succeed, because the existing drug has already passed toxicity tests and its safety is known. Numerous openly accessible databases exist, listing the interaction of known compounds, which can be either already approved drugs or experimental drug candidates, against known target proteins (ChEMBL [5], DrugBank [23], KEGG[9], SuperTarget [7], BindingDB [14]). The high cost of in vitro methods for the testing of drug-target binding behaviour and the availability of experimental results in public databases give strong incentives to develop in silico methods for the prediction of new drug-target interactions. The development of a new model for this task is the topic of this thesis.

### 1.1 Terms and Definitions

Throughout this thesis the terms *drugs* and *targets* are used frequently. A simple introduction to these terms is given in this section.

- **Drugs**: In this thesis, the terms *compounds* and *drugs* are often used interchangeably. In drug discovery, a lead compound is a chemical structure that showed pharmacological activity that is likely to be therapeutically useful. The lead compound is used as a starting point which might be modified to increase the binding strength to the disease's target or to increase the selectivity with which it binds to the target. A high binding selectivity is important to reduce the potential side effects that the compound causes in the organism. The terms *compounds* and *drugs* are used interchangeably because the used datasets for
the in silico prediction of drug-target interaction often contain both approved drugs and chemical structures that might be potential lead compounds.

- **Targets:** The terms *target* and *proteins* are often used interchangeably in this thesis. The identification of the biological origin of a disease is the first step in the discovery of a medicine. In a simplified formulation, certain proteins are hypothesized to be targets of intervention for the disease for which a drug is developed which changes the behavior or function of the respective proteins by physically binding to them. Here, the terms describe the proteins in the body whose activity can be modified by a drug through physical binding which causes a certain effect. The effect can be therapeutic if the drug binds to the targets of a disease or an unwanted side effect if the drug binds to other proteins.
Chapter 2

Review of Literature

Machine Learning and Data Mining techniques for drug development is a hot topic. In most existing methods the problem is formulated as a binary classification problem, where the drug-target pairs are treated as instances and the chemical structures of drugs and the amino acid subsequences of the targets are used as features, describing the instances. The goal in the binary formulation is to classify a given drug-target pair into binding and non binding. This approach stands in contrast with the developed method in this thesis, which predicts continuous drug-target binding affinities, where the binding affinity describes the strength of the binding between the drug and the target. The related methods which classify drug-target pairs will still be introduced here because the majority of existing work formulates the problem as such. Additionally, the only existing method in the literature, KronRLS, which predicts continuous binding affinities is introduced here. This method will be used as a comparison model to evaluate the developed model in this thesis. The following sections describe the state of the art Machine Learning based methods for drug-target interaction prediction, including the comparison model KronRLS. As described below, Machine Learning based methods for drug-target interaction prediction typically utilize similarity matrices of the drugs and targets to predict missing interactions. The method that is developed in this thesis follows this approach and the details of the types of similarity metrics for the drugs and targets are described in section 4.1. To
understand the related work, it is sufficient to know that a similarity value for each
drug-drug pair and each target-target pair is given.

Two existing methods for drug-target interaction prediction that are not based
on Machine Learning techniques are docking simulation and ligand-based approaches
[10], [13], [22]. In docking simulation the interaction strength of ligands and proteins
is estimated based on the structure of the target protein. This process is extremely
time-consuming and the structural information of a protein is not always available
[15]. In ligand based approaches, the interaction strength of a candidate ligand to a
target protein is obtained by comparing the candidate ligand to ligands for which the
interaction strength to the target is known. This approach is not applicable, when
information of candidate-similar ligands is not available for the target protein. Both
approaches will not be examined further here.

The following sections are titled by the first author of the respective method,
followed by the title of the paper and the year that the method was published.

2.1 Yoshihiro Yamanishi, "Prediction of drug-target interac-
tion networks from the integration of chemical and ge-
nomic spaces.", 2008

One of the first proposed models for the task [25], is a supervised bipartite graph
learning method. The motivation of the model is to reveal the correlations between
drug similarity, target similarity and the drug-target interaction network. The authors
define the chemical space for drugs, the genomic space for targets and the pharmaco-
cological space for drug-target pairs and propose a method to embed compounds and
proteins from the chemical and genomic spaces respectively into the unified pharma-
cological space. New drug-target interactions are then predicted by connecting drugs
and targets which are closer than a threshold in the pharmacological space.
In the authors' model, the drug-target interaction network is represented by a bipartite graph $G = (V_1 + V_2, E)$, where $V_1$ is a set of drugs, $V_2$ is a set of target proteins and $E$ is a set of interactions between the drugs and targets. A graph-based similarity matrix $K = \begin{pmatrix} K_{cc} & K_{cg} \\ K_{cg}^T & K_{gg} \end{pmatrix}$ is constructed, where the elements of $K_{cc}$, $K_{gg}$, $K_{cg}$ are computed by using Gaussian functions: $(K_{cc})_{ij} = \exp(-d^2_{c_ic_j}/h^2)$ for $i, j = 1, \ldots, n_c$, $(K_{gg})_{ij} = \exp(-d^2_{g_ig_j}/h^2)$ for $i, j = 1, \ldots, n_g$ and $(K_{cg})_{ij} = \exp(-d^2_{c_ig_j}/h^2)$ for $i = 1, \ldots, n_c$, $j = 1, \ldots, n_g$. Here $d$ stands for the shortest distance between two objects, $n_c$ and $n_g$ stand for the number of known drugs and targets respectively and $h$ is a width parameter. To compute the vectors that span the pharmacological space, the eigenvalue decomposition of $K$ is computed as:

$$K = \Gamma \Lambda \frac{1}{2} \Lambda \frac{1}{2} \Gamma^T = U U^T$$

and all drugs and targets are represented by using the row vectors of the matrix $U = (u_{c_1}, \ldots, u_{c_{n_c}}, u_{g_1}, \ldots, u_{g_{n_g}})^T$.

Now two models are learned to map new compounds and targets from the chemical and genomic spaces respectively into the pharmacological space. A kernel regression model that learns the feature vectors of new compounds and targets is proposed for this task:

$$u = \sum_{i=1}^{n} s(x, x_i) w_i + \epsilon$$

for the mapping of the compounds, $s(x, x_i)$ represents the compound similarity and for the mapping of the targets $s(x, x_i)$ represents the target similarity (and $n$ represents the number of drugs/targets in the training set). $\epsilon$ is a noise vector and $w_i$ is a weight vector that is learned by minimizing the loss function:

$$L = ||UU^T - SWW^T S^T||_F^2$$

where $S$ represents the respective similarity matrix, $W$ represents the matrix of weight vectors and $||.||_F$ represents the Frobenius norm. Two such models, meaning two sets of weight vectors are learned to map new compounds and new targets onto the pharmacological space. Finally, based on the feature vectors $u$ in the pharmacological
Figure 2.1: Illustration of the Supervised Bipartite Graph Inference Model, [25]. All drugs and targets are mapped into the unified pharmacological space. Similarity scores in the pharmacological space for the drug-target pairs are then calculated as the inner product of the feature vectors.

space, feature-based similarity scores for three types of drug-target pairs are computed as the inner product as follows.

- $corr(c_{\text{new}}, g_j) = u_{c_{\text{new}}}u_{g_j}$
- $corr(c_i, g_{\text{new}}) = u_{c_i}u_{g_{\text{new}}}$
- $corr(c_{\text{new}}, g_{\text{new}}) = u_{c_{\text{new}}}u_{g_{\text{new}}}$

The three types of drug-target pairs, correspond to new drugs $c_{\text{new}}$ and known targets $g_j$, known drugs $c_i$ and new targets $g_{\text{new}}$ and new drugs and new targets $c_{\text{new}}, g_{\text{new}}$. High-scoring compound-protein pairs of any of the three above types are predicted to interact with each other. Figure 2.1 illustrates the Supervised Bipartite Graph Inference Model.
2.2 Twan van Laarhoven, "Gaussian interaction profile kernels for predicting drug-target interaction.", 2011

The authors of this method first build an interaction profile for each drug and for each target [21]. The interaction profile of each compound is a binary vector describing the presence or absence of interaction with every target in the network. The interaction profile for each target is defined analogously. Figure 2.2 illustrates the interaction profiles, constructed from the drug-target interaction network. The interaction profiles of the drugs and targets are used as feature vectors and a Gaussian kernel is constructed for the drugs and targets respectively. Let $y_{d_i}$ be the interaction profile of drug $d_i$, then the Gaussian kernel for the drugs is defined as

$$K_{GIP}(d_i, d_j) = \exp(-\gamma_d ||y_{d_i} - y_{d_j}||^2)$$

and the Gaussian kernel for the targets can be defined analogously. Here, $GIP$ stands for Gaussian Interaction Profile and the parameter $\gamma_d$ controls the kernels bandwidth. The authors set $\gamma_d$ to:

$$\gamma_d = \frac{1}{\frac{1}{n_d} \sum_{i=1}^{n_d} |y_{d_i}|^2}$$

and $\gamma_t$ is defined analogously. The similarity information of the drugs and targets is integrated by defining two new kernels $K_{chemical}$ and $K_{genomic}$ for the drugs and targets respectively which are defined as:

$$K_{chemical} = (S_D + S_D^T)/2$$

$$K_{genomic} = (S_T + S_T^T)/2$$

(2.1)

where $S_D$ and $S_T$ stand for the given similarity matrices of the drugs and targets. Finally, combined kernels of the two kernels $K_{GIP}$ and $K_{chemical}/K_{genomic}$ are defined as a weighted average:

$$K_d = \alpha_d K_{chemical} + (1 - \alpha_d) K_{GIP}$$

$$K_t = \alpha_t K_{genomic} + (1 - \alpha_t) K_{GIP}$$
In the authors implementation, the parameters $\alpha_d, \alpha_t$ were chosen as $\alpha_d = \alpha_t = 0.5$.

A Regularized Least Squares (RLS) classifier is chosen to generalize from the training data. The predicted values $\hat{y}$ of the RLS classifier for a given kernel $K$ and training vector $y$ are defined as:

$$\hat{y} = K(K + \sigma I)^{-1}y$$

Here $\sigma$ is a regularization parameter. Higher values of $\sigma$ give a smoother result, while for $\sigma = 0$, we get $\hat{y} = y$. With the matrix $Y \in n_d \times n_t$ being the binary matrix of training values of $n_d$ drugs and $n_t$ targets and the kernels for the drugs and targets as defined above, the authors of the model propose two ways to predict the interaction of all drug-target pairs in the matrix. The first type of prediction $RLS$-avg is defined as:

$$\hat{Y} = \frac{1}{2}(K_d(K_d + \sigma I)^{-1}Y) + \frac{1}{2}(K_t(K_t + \sigma I)^{-1}Y^T)^T$$

For the second type of prediction, the authors propose to compute yet a fourth kernel, defined by the Kronecker product $K = K_d \otimes K_t$, which gives a similarity for all drug-target pairs. The model named $RLS$-$Kron$ predicts $\hat{Y}$ as:

$$vec(\hat{Y}^T) = K(K + \sigma I)^{-1}vec(Y^T)$$

### 2.3 Fan Yang, "Drug-target interaction prediction by integrating chemical, genomic, functional and pharmacological data.", 2014

This method is introduced here because it is the only method in the literature that also utilizes Conditional Random Fields to predict drug-target interaction. The authors propose the following model to classify drug-target pairs into binding or non-binding: Let $\{d_i\}$ be the set of known drugs and $\{t_j\}$ be the set of known targets. $X$ is used to denote the known drug-target interactions and the similarity information among the
drugs and targets. For each drug $d_i$ an undirected graph $G = (V_t, E_t)$ is defined, where $V_t = \{t_j\}$ is the set of targets and each edge in $E_t$ represents the similarity between a pair of targets. Let $Y = (y_1, y_2, \ldots, y_{n_t})$, where $n_t$ is the total number of known targets, denote the prediction, where each $y_j$ is a binary random variable representing the prediction on target $t_j$, that is, $y_j = 1$ if the predicted interaction between $d_i$ and target $t_j$ is true, and $y_j = 0$ otherwise. These graphs will in the following be referred to as the target-based CRFs. Similarly, an undirected graph $G = (V_d, E_d)$, where $V_d$ is the set of known drugs and $E_d$ represents the similarity between a pair of drugs is constructed for each target. These graphs will in the following be referred to as the drug-based CRFs.

In order to predict the interaction of all drugs, using the target-based CRFs, a joint probability distribution which is conditioned on the observation $X$ is defined for each target-based CRF, meaning a separate CRF is constructed for each drug. In the underlying graph, each node represents a target $t_i$ and its associated binary random variable $y_i$, and each edge connecting two nodes represents the dependency
between these two nodes. The energy of a joint configuration $Y$, given $X$ for any of the target-based CRFs is defined as follows:

$$E(Y|X) = \sum_i a_i f(y_i|X) + \sum_{i,j} b_{ij} g(y_i,y_j|X) \quad (2.2)$$

where $f(y_i|X)$ is a local node feature function defined based on the state of $y_i$, $g(y_i,y_j|X)$ is a relational edge feature function defined based on states of both $y_i$ and $y_j$ and $a_i \geq 0$ and $b_{ij} \geq 0$ are weight parameters that are learned from training data.

The functions $f$ and $g$ are defined as follows:

$$f(y_i|X) = -(y_i - H_{x_i}(y_i))^2 \quad (2.3)$$

$$g(y_i,y_j|X) = -H_{x_i,x_j}(y_i - y_j)^2 \quad (2.4)$$

where $H_{x_i}(y_i)$ represents the observed feature of target $t_i$. In the authors implementation, $H_{x_i}(y_i)$ is the average number of observed drug interactions for target $t_i$. $H_{x_i,x_j}(y_i - y_j)$ denotes the difference between binary variables $y_i$ and $y_j$. Intuitively, this formulation adds a penalization when predictions for two similar nodes are different and when the prediction of a given node deviates from its average state. The authors of the method let all target-based CRFs share the same parameters $a_i$ and $b_{ij}$ and similarly for all drug-based CRFs. The joint probability density function of $Y$ given $X$ is then defined as:

$$P(Y|X) = \frac{1}{Z(X)} \exp(-E(Y|X)) \quad (2.5)$$

where $Z(X) = \sum_Y \exp(-E(Y|X))$ is the partition function of the CRF. As described in [27], the parameters can be learned by optimizing the log-likelihood of the training data by stochastic gradient ascent. To predict unknown drug-target interactions for a query drug and target $t_k$, the conditional probability $p(y_k|y_{-k},X)$ can be computed for each target $t_k$, where $y_{-k}$ denotes all other targets except $t_k$. The
conditional expectation of $y_k$ is calculated as the prediction score of the interaction between target $t_k$ and the query drug. Similarly, a prediction score can be computed for a query target and each drug. A weighted average of the drug-based and the target-based CRFs is taken as the final prediction for each drug-target pair in the authors implementation.

2.4 Yong Liu, "Neighborhood Regularized Logistic Matrix Factorization for Drug-Target Interaction Prediction", 2016

The study [15] proposes a method named Neighborhood Regularized Logistic Matrix Factorization (NRLMF) for the task of drug-target interaction prediction which predicts the probability that a drug would interact with a target. In this method, the properties of a drug and a target are represented by two latent vectors in a shared low dimensional latent space. The interaction probability of a drug-target pair is modeled by a logistic function of the drug-specific and the target-specific latent vectors. In the learning step, regularization constraints between the latent representations of similar drugs and targets are applied, such that similar latent vectors are learned for similar drugs and targets. Formally, the problem is described as follows: The set of known drugs is denoted by $D = \{d_i\}$, with $i = 1, \ldots, m$ and the set of known targets is denoted by $T = \{t_j\}$, $j = 1, \ldots, n$, where $m$ and $n$ are the number of drugs and targets, respectively. In addition, the drug similarities are represented by $S^d \in \mathbb{R}^{m \times m}$ and the target similarities by $S^t \in \mathbb{R}^{n \times n}$. A binary matrix $Y \in \mathbb{R}^{m \times n}$, where each element $y_{ij} \in \{0, 1\}$ is given as training data. If drug $d_i$ has been experimentally verified to interact with target $t_j$, $y_{ij}$ is set to 1 and otherwise $y_{ij}$ is set to 0. Both, the drugs and the targets are mapped into a shared latent space, with a low dimensionality $r$. The properties of a drug $d_i$ and a target $t_j$ are described by two latent vectors $u_i \in \mathbb{R}^{1 \times r}$ and $v_j \in \mathbb{R}^{1 \times r}$, respectively. The interaction probability $p_{ij}$ of a drug-target
pair \((d_i, t_j)\) is then modeled by the logistic function:

\[
p_{ij} = \frac{\exp(u_i v_j^T)}{1 + \exp(u_i v_j^T)}
\]

(2.6)

In the following, let \(U \in \mathbb{R}^{m \times r}\) and \(V \in \mathbb{R}^{n \times r}\) denote the latent vectors of all drugs and targets respectively, such that \(u_i\) is the \(i^{th}\) row in \(U\) and \(v_j\) is the \(j^{th}\) row in \(V\). Under the assumption that all training examples are independent, the probability of the observations in the training matrix \(Y\) is:

\[
p(Y|U, V) = \prod_{1 \leq i \leq m, 1 \leq j \leq n, y_{ij} = 1} [p_{ij}^{y_{ij}} (1 - p_{ij})^{(1 - y_{ij})}]^c \times \prod_{1 \leq i \leq m, 1 \leq j \leq n, y_{ij} = 0} p_{ij}^{y_{ij}} (1 - p_{ij})^{(1 - y_{ij})}
\]

(2.7)

where \(c\) is a constant used to control the importance levels of observed interactions. This importance weighting strategy is used to assign higher importance to positive interactions, because the positive interactions are biologically validated and thus more trustworthy. The negative interactions in contrast could contain potential drug-target interactions and are thus unreliable.

As derived in [15], the model parameters \(U\) and \(V\) can be learned by maximizing the posterior distribution, which is equivalent to minimizing the following objective function:

\[
\min_{U, V} = \sum_{i=1}^{m} \sum_{j=1}^{n} (1 + c y_{ij} - y_{ij}) \log[1 + \exp(u_i v_j^T)] - c y_{ij} u_i v_j^T + \frac{\lambda_d}{2} ||U||_F^2 + \frac{\lambda_t}{2} ||V||_F^2
\]

(2.8)

where \(\lambda_d = \frac{1}{\sigma_d^2}\) and \(\lambda_t = \frac{1}{\sigma_t^2}\) are parameters controlling the variances of Gaussian priors on the latent vectors of the drugs and targets and \(||.||_F\) denotes the Frobenius norm.

Next, the drug and target similarities are integrated as follows: For drug \(d_i\), let \(N(d_i)\) denote the \(K_1\) most similar drugs to \(d_i\) and analogously let \(N(t_j)\) denote the \(K_2\) most similar targets to \(t_j\). The drug neighborhood information is represented using
an adjacency matrix $A$, where the $(i, y)$ element $a_{iy}$ is defined as:

$$a_{iy} = \begin{cases} S_{iy}^d & \text{if } d_y \in N(d_i) \\ 0 & \text{otherwise} \end{cases}$$  \hspace{1cm} (2.9)$$

and similarly the target neighborhood information is denoted by $B$, where its $(j, x)$ element $b_{jx}$ is defined as:

$$b_{jx} = \begin{cases} S_{jx}^t & \text{if } t_x \in N(t_j) \\ 0 & \text{otherwise} \end{cases}$$  \hspace{1cm} (2.10)$$

Now, intuitively the drug and target similarities are integrated by adding constraints such that the distance of the latent vectors between each drug/target $d_i/t_j$ and its nearest neighbors $N(d_i)/N(t_j)$ is minimized. This can be achieved by adding constraining terms to the objective function. For the drugs, the constraining term has the following form:

$$\frac{\alpha}{2} \sum_{i=1}^{m} \sum_{y=1}^{m} a_{iy} ||u_i - u_y||_F^2$$  \hspace{1cm} (2.11)$$

A similar term is added for the latent vectors of the targets. These constraining terms are added to the objective function 2.8 to integrate the similarity metrics. The full derivation and the resulting objective function which minimizes the posterior distribution including the constraining terms is shown in [15]. Once the latent vectors $U$ and $V$ have been learned, the probability associated with any unknown drug-target pair $(d_i, t_j)$ can be predicted by equation 2.6. In [15] the drugs and targets are further distinguished into drugs/targets with known interactions and drugs/targets without any known interaction. These details are left out here.
2.5 Tapio Pahikkala, "Toward more realistic drug-target interaction predictions.", 2014

*KronRLS* is a model for continuous drug-target interaction prediction that was presented in [17]. Note that *RLS – Kron*, which is introduced in section 2.2 and *KronRLS*, which is introduced here are two different methods. The model that is developed in this thesis will later be compared to the performance of *KronRLS*. *KronRLS* is a kernel methods which formulates the problem as follows: for a set $D$ of drugs and a set $T$ of targets, we have a set $X = x_1, \ldots, x_m$ of drug-target pairs ($X$ is a subset of $D \times T$) with an associated vector $y = y_1, \ldots, y_m$ of continuous binding affinities as training data. To find a prediction function $f(x)$ for all possible drug-target pairs $x \in D \times T$, we learn $f$ from the training data by minimizing the objective:

$$J(f) = \sum_{i=1}^{m} (y_i - f(x_i))^2 + \lambda \|f\|_k^2$$

(2.12)

In the objective function, $\|f\|_k^2$ is the norm of $f$, which is associated to a kernel function $k$ (which is described below), and $\lambda > 0$ is a user defined regularization parameter. A minimizer of the above objective can be expressed as

$$f(x) = \sum_{i=1}^{m} a_i k(x, x_i)$$

(2.13)

The kernel function $k$ is a symmetric similarity measure between each of the $m$ drug-target pairs, which can be represented by a $m \times m$ matrix $K$. When we have two individual similarity matrices $K_d$ and $K_t$ for the drugs and targets respectively, a similarity matrix for each drug-target pair can be computed as $K_d \otimes K_t$, where $\otimes$ stands for the Kronecker product. We can see that intuitively, the parameters $a_i$ represent how much can be interpolated from each known drug-target pair.

Now we first assume that the training set $X$ contains every possible pair of drugs and targets (which is not the case in the real application of the method). With this assumption, we can compute $K$ as $K = K_d \otimes K_t$ and we can learn the parameters $a_i$
which define the minimizer $f$ by solving the system of linear equations:

$$ (K + \lambda I)a = y $$

(2.14)

where $I$ is the $D \times T$ identity matrix. For the real world scenario, where only a subset of $D \times T$ is given as training data, the vector $y$ has missing values in the above formulation. To find the parameters $a$, [17] suggests to formulate the problem as above and use approaches such as conjugate gradient, combined with Kronecker algebraic optimization to solve the system of linear equations.

Finally, predictions for all drug-target pairs of all drugs and targets that appear in the training data can be made through equation 2.13.
Chapter 3

Method

3.1 Model Intuition

The model that is developed in this thesis consists of two parts. As a first step, Matrix Factorization (MF) is applied to make a first prediction on the missing drug-target pairs. MF predicts the missing values based only on the observations in the training dataset and does not take additional features of the drugs and targets into account. This is where the second part of the model, Continuous Conditional Random Fields (CCRF) comes into play. The CCRF gets the similarity matrices of the drugs and targets as input, together with the training dataset, the predictions that were made by MF on the missing values and the predictions that were made by MF on the training data (MF is applied twice, to make predictions on missing values and the training data itself as described below). The intuition behind the CCRF is to re-predict the missing values based on three parameters $\alpha$, $\beta_1$, $\beta_2$ which represent respectively:

- $\alpha$: the *trustworthiness* of the MF prediction.

- $\beta_1$: how much can be interpolated between drug-drug pairs with high similarity scores in the drug similarity matrix.

- $\beta_2$: how much can be interpolated between target-target pairs with high similarity scores in the target similarity matrix.
The trustworthiness of the \( MF \) prediction is learned from the training data and the predictions that were made by \( MF \) on these data points. The dependance of drug-drug and target-target pairs with high similarity is captured by learning how similar the binding values of these pairs are in the training data. The idea of the combined \( MF+CCRF \) model is that the \( CCRF \) can improve the initial \( MF \) prediction of a drug-target pair by interpolating from given observations of a similar drug on the same target or by interpolating from a given observation of a similar target on the same drug. When it is not possible to interpolate from a given observation to make a prediction for a drug-target pair, the combined model will just fall back onto the initial \( MF \) prediction.

The two parts of the model, \( MF \) and \( CCRF \) are described first separately and then in combination in the following sections. Matrix Factorization is described in section 3.3. \( CCRF \) is described in more detail in section 3.4. Both, \( MF \) and \( CCRF \) are explained by first giving a short overview, then explaining the parameter learning step and finally describing how missing values can be predicted (the inference step). The combination of the two models (\( MF+CCRF \)) is described in section 3.5. This thesis focuses mainly on the \( CCRF \) part of the \( MF+CCRF \) model. Before going into the model details, section 3.2 gives an overview of the used notation.

\section{Notation}

Let \( D \) and \( T \) denote sets of drugs and targets respectively. Let \( R \in \mathbb{R}^{\mid D \mid \times \mid T \mid} \) denote a drug-target interaction matrix, where the rows represent the drugs in \( D \) and the columns represent the targets in \( T \). \( R \) contains missing values for drug-target pairs for which the interaction strength is unknown and real values representing the interaction strength for drug-target pairs whose binding affinity was measured in wetlab experiments. The known values in \( R \) serve as training data for the developed model and the missing values are those that we wish to predict. Further, let \( S_D \in \mathbb{R}^{\mid D \mid \times \mid D \mid} \) and \( S_T \in \mathbb{R}^{\mid T \mid \times \mid T \mid} \) denote matrices containing similarity scores for the drugs \( D \) and
targets $T$ respectively. The interaction data and similarity scores can be of different kind as described in section 4.1. In the drug-target interaction prediction task, the number of missing values in $R$ usually outweights the number of known interactions and $R$ is only sparsely populated.

3.3 Matrix Factorization

The Matrix Factorization technique has been demonstrated to be effective specially for personalized recommendation tasks [11] and it has been previously applied for drug-target interaction prediction [15], [4], [6]. Here, the MF technique is utilized in its simplest form, without incorporating the similarity matrices directly as it is done in [15] and [6]. MF is a process in which the sparsely filled training matrix $R$ is approximated by the product of two factor matrices $P \in \mathbb{R}^{k \times |D|}$ and $Q \in \mathbb{R}^{k \times |T|}$ as described in the following two sections.

3.3.1 Parameter Learning

The factor matrices $P$ and $Q$ are learned by minimizing the regularized squared error on the set of known affinities $\kappa$:

$$\min_{Q,P} \sum_{(d_i,t_j) \in \kappa} (a_{i,j} - q_i^T p_j)^2 + \lambda(||p||^2 + ||q||^2)$$  \hspace{1cm} (3.1)

The term $(a_{i,j} - q_i^T p_j)^2$ represents the fit of the learned parameters to the previously observed binding affinities. The term $\lambda(||p||^2 + ||q||^2)$ penalizes the magnitudes of the learned parameters to prevent overfitting and the constant $\lambda$ controls the extend of overfitting. The above optimization problem can be solved by stochastic gradient descent as described in [11]: The algorithm loops through all pairs $(d_i, t_j)$ in $R$ for which the binding affinity is known, computes the current prediction for $(d_i, t_j)$ and
computes the associated prediction error as:

\[ e_{i,j} = r_{i,j} - p_i^T q_j \]  

where \( p_i \) denotes the \( i \)th column of \( P \) and \( q_j \) denotes the \( j \)th column of \( Q \). The columns of \( P \) and \( Q \) are then modified in the opposite direction of the gradient:

\[ p_i := q_i + \gamma (e_{i,j} q_j - \lambda q_i) \]  
\[ q_j := p_i + \gamma (e_{i,j} p_i - \lambda q_j) \]

where \( \gamma \) is a constant specifying the magnitude of the update.

### 3.3.2 Inference

With learned matrices \( P \) and \( Q \), a missing affinity of drug-target pair \((d_i, t_j)\) in \( R \) can be predicted by the inner product of the \( i \)th column of \( P \) and the \( j \)th column of \( Q \). A matrix \( R' \) with predictions for all drug-target pairs can be computed as:

\[ R' = P^T Q \]

### 3.4 Continuous Conditional Random Fields

Conditional Random Fields were originally developed for the task of segmenting and labeling sequence data [12]. In contrast to predicting continuous values as for drug-target binding affinities, the original formulation predicts a label vector \( Y \), where all components of \( Y_i \) are assumed to range over a finite label alphabet \( \mathcal{Y} \). For the task of classifying whether or not a drug-target pair is interacting, CRFs were applied previously [27].

To the best of my knowledge, the continuous variant of Conditional Random fields were first introduced by [18] for ranking tasks in document retrieval systems. An other
application of CCRFs for recommender systems can be found in [24]. The problem formulation in [24] is very similar to the problem of drug-target interaction prediction. However, the formulation in [24] requires taking a Gibbs-sample of the distribution defined through the CCRF in each update step during parameter learning. Although with the drawback of reducing the feasible size of the graphical model, [2] presents a closed form solution for the parameter learning and inference step. This closed formulation of the CCRF is applied for the model that is used in this thesis. The following sections formally describe the CCRF and explain the parameter learning and the inference step.

3.4.1 Formal Definition

Here I will use the same notation as it is presented in [2]. A CCRF is an undirected graphical model which models a set of output variables $Y = \{y_1, \ldots, y_n\}, y_i \in \mathbb{R}$ that we wish to predict as $P(Y|X)$ where $X = \{x_1, \ldots, x_n\}, x_i \in \mathbb{R}^m$ is a set of observed input variables. The CCRF defines a conditional probability distribution with the density function:

$$P(Y|X) = \frac{\exp(\Psi)}{\int_{-\infty}^{\infty} \exp(\Psi)dy}$$

$$\Psi = \sum_i \sum_{k=1}^{K_1} \alpha_k f_k(y_i, X) + \sum_{i,j} \sum_{k=1}^{K_2} \beta_k g_k(y_i, y_j, X)$$

The term $\int_{-\infty}^{\infty} \exp(\Psi)dy$ is a normalization term which makes the probability distribution valid. The $f_k$ terms will be called vertex features and the $g_k$ terms edge features. The feature functions are defined as:

$$f_k(y_i, X) = -(y_i - X_{i,k})^2$$

$$g_k(y_i, y_j, X) = -\frac{1}{2}S_{i,j}^k (y_i - y_j)^2$$

Intuitively, the weights $\alpha_k$ on the feature functions $f_k$ represent the reliability of observation $X_{i,k}$ in regard to the true label $y_i$. Here, the observations $X_{i,k}$ can for example represent the initial predictions of a set of $K_1$ regressors. In the CCRF that is used for the experiments in this thesis there is always only one regressor (and thus
$K_1 = 1$) which is the Matrix Factorization part of the model and $X_{i,1}$ represents the initial prediction coming from Matrix Factorization. Edge features $g_k$ represent the dependency between labels $y_i$ and $y_j$ with similarity $S_{i,j}^k$, $k \in 1, \ldots, K_2$. With given drug-drug similarity matrix $S_D$ and target-target similarity matrix $S_T$, $S_{i,j}^1$ and $S_{i,j}^2$ can for example be defined as:

\[
S_{i,j}^1 = \begin{cases} 
1, & S_D(i,j) > \text{thresh}_d \\
0, & S_D(i,j) < \text{thresh}_d 
\end{cases}
\]  

(3.7)

\[
S_{i,j}^2 = \begin{cases} 
1, & S_T(i,j) > \text{thresh}_t \\
0, & S_T(i,j) < \text{thresh}_t 
\end{cases}
\]  

(3.8)

where $\text{thresh}_d$ and $\text{thresh}_t$ are user defined thresholds. Intuitively, the weights $\beta_1$ and $\beta_2$ on $S_{i,j}^1$ and $S_{i,j}^2$ respectively, now represent the importance of assigning similar prediction values to drug pairs with similarity larger than $\text{thresh}_d$ and of assigning similar prediction values to target pairs with similarity larger than $\text{thresh}_t$. Note, that the dimensions of $S^k$ and the similarity matrices $S_D$ and $S_T$ are usually different. The matrices $S^k$ contain as many rows and columns as there are nodes in the graphical structure of the CCRF. Here $S_D(i,j)/S_T(i,j)$ refers to the similarity of the drugs/targets that correspond to the nodes $i$ and $j$ in the graphical structure.

As explained in the following sections, the conditional probability distribution that is defined through the CCRF can be transformed into a multivariate Gaussian distribution in closed form, which is a useful property for both learning the weights on the feature functions as well as making predictions for missing values.

### 3.4.2 Parameter Learning

This section describes the parameter learning procedure of the CCRF for quadratic vertex and edge functions as described above. Assume, we are given training values $Y = \{y_1, \ldots, y_n\}$. Additionally, for each $y_i$ we have given the prediction $X_{i,k}$ that
regressor $k$ would have predicted. As explained above, for the model in this thesis we have $K_1 = 1$ and thus we have only one prediction $X_{i,1}$ for each training data point, which is the Matrix Factorization prediction. Further we have $K_2$ similarity matrices given, which in the scope of this thesis can for example be the matrices $S_{i,j}^1$ and $S_{i,j}^2$ as defined in equation 3.7 and equation 3.8 respectively. In the learning step, we want to find the $\alpha$ and $\beta_k$ weights, that optimize the conditional log-likelihood of the CCRF on the training data:

$$L(\alpha, \beta) = \log P(Y|X) \quad (3.9)$$

$$(\alpha_{opt}, \beta_{opt}) = \arg\max_{\alpha,\beta} L(\alpha, \beta) \quad (3.10)$$

Because this problem is convex [18], standard techniques such as stochastic gradient descent can be utilized to determine the optimal parameters $\alpha$ and $\beta_k$. For the computation of the derivatives of $\log P(Y|X)$ it helps to convert equation 3.6 into the form of a multivariate Gaussian distribution. The conversion into the Gaussian formulation and the computation of the derivatives are described in detail in the following two sections. The log-likelihood is optimized with respect to $\log \alpha$ and $\log \beta_k$ in order to guarantee that $\alpha > 0$ and $\beta_k > 0$, which in turn guarantees that the normalization function is integrable. Using the partial derivatives, the CCRF learning algorithm is as follows (here it is assumed that we are learning one parameter $\alpha$ and two $\beta$ parameters ($\beta_1, \beta_2$). The two $\beta$ parameters describe the weight on the drug-similarity and the target-similarity matrix respectively):

**Gaussian Formulation of the CCRF**

This subsection describes the conversion of the probability distribution defined through equation 3.6 into the formulation of a Gaussian distribution, as derived in [2]:

First, the feature functions are plugged into the energy function $\Psi$: 
**Algorithm**  CCRF learning algorithm

**Require:** \{\(X_{\text{train}}, Y_{\text{train}}, S_D = S^1, S_T = S^2\)\}

Params: number of iterations \(T\), learning rate \(\eta\)

Initialise Parameters \(\alpha\) and \(\beta\)

\[
\text{for } r = 1 \text{ to } T \quad \text{do}
\]

Compute current gradients with respect to \(\log \alpha\), \(\log \beta_1\), \(\log \beta_2\)

\[
\begin{align*}
\log \alpha &= \log \alpha + \eta \frac{\delta \log(P(Y|X))}{\delta \log \alpha} \\
\log \beta_1 &= \log \beta_1 + \eta \frac{\delta \log(P(Y|X))}{\delta \log \beta_1} \\
\log \beta_2 &= \log \beta_2 + \eta \frac{\delta \log(P(Y|X))}{\delta \log \beta_2}
\end{align*}
\]

Update \(\alpha\), \(\beta_1\), \(\beta_2\)

\[
\text{end for}
\]

Return \(\alpha\), \(\beta_1\), \(\beta_2\)

\[
\Psi = \sum_{k=1}^{K_1} \alpha_k f_k(y_i, X) + \sum_{i,j} \sum_{k=1}^{K_2} \beta_k g_k(y_i, y_j, X)
\]

\[
= - \sum_{k=1}^{K_1} \alpha_k (y_i - X_{i,k})^2 - \frac{1}{2} \sum_{i,j} \sum_{k=1}^{K_2} \beta_k S_{i,j}^k (y_i - y_j)^2
\]  \hspace{1cm} (3.11)

For the next step, we first define a diagonal matrix \(A\) and a symmetric matrix \(B\):

\[
A_{i,j} = \begin{cases} 
\sum_{k=1}^{K_1} \alpha_k, & i = j \\
0, & i \neq j
\end{cases}
\]  \hspace{1cm} (3.12)

\[
B_{i,j} = \begin{cases} 
(\sum_{k=1}^{K_2} \beta_k \sum_{r=1}^{n} S_{i,r}^k) - (\sum_{k=1}^{K_2} \beta_k S_{i,j}^k), & i = j \\
- \sum_{k=1}^{K_2} \beta_k S_{i,j}^k, & i \neq j
\end{cases}
\]  \hspace{1cm} (3.13)

as well as

\[
\Sigma^{-1} = 2(A + B)
\]  \hspace{1cm} (3.14)

and a vector \(b\):

\[
b_i = 2 \sum_{k=1}^{K_1} \alpha_k X_{i,k}
\]  \hspace{1cm} (3.15)

\[
b = 2X\alpha
\]  \hspace{1cm} (3.16)
The factor \( \Psi \) in equation 3.11 can then be expressed in terms of \( A, B \) and \( b \): First, terms of \( \Psi \) containing \( \alpha \) parameters are collected:

\[
- \sum_{i}^{K_1} \sum_{k=1}^{\alpha} \alpha_k (y_i - X_{i,k})^2 \\
= - \sum_{i}^{K_1} \sum_{k=1}^{\alpha} \alpha_k (y_i^2 - 2y_i X_{i,k} + X_{i,k}^2) \\
= - \sum_{i}^{K_1} \sum_{k=1}^{\alpha} \alpha_k y_i^2 + \sum_{i}^{K_1} \sum_{k=1}^{\alpha} \alpha_k 2y_i X_{i,k} - \sum_{i}^{K_1} \sum_{k=1}^{\alpha} \alpha_k X_{i,k}^2 \\
= - y^T A y + y^T b - \sum_{i}^{K_1} \sum_{k=1}^{\alpha} \alpha_k X_{i,k}^2 
\tag{3.17}
\]

Then, terms of \( \Psi \) containing \( \beta \) parameters are collected:

\[
- \frac{1}{2} \sum_{i,j}^{K_2} \sum_{k=1}^{\beta} \beta k S_{i,j}^k (y_i - y_j)^2 \\
= - \frac{1}{2} \sum_{i,j}^{K_2} \sum_{k=1}^{\beta} \beta k S_{i,j}^k (y_i^2 - 2y_i y_j + y_j^2) \\
= - \frac{1}{2} \sum_{i,j}^{K_2} \sum_{k=1}^{\beta} \beta k S_{i,j}^k (y_i^2 + y_j^2) + \sum_{i,j}^{K_2} \sum_{k=1}^{\beta} \beta k S_{i,j}^k y_i y_j \\
= - \frac{1}{2} \sum_{k=1}^{K_2} \beta_k \sum_{i,j}^{S_{i,j}^k} y_i^2 + \sum_{k=1}^{K_2} \beta_k S_{i,j}^k \sum_{i,j} y_i y_j \\
= - y^T B y 
\tag{3.18}
\]

Combining these we can write \( \Psi \) as:

\[
\Psi = - y^T A y + y^T b - y^T B y - d = - \frac{1}{2} (y^T \Sigma^{-1} y) + y \Sigma^{-1} \mu - d 
\tag{3.19}
\]

where \( d = \sum_{i}^{K_1} \sum_{k=1}^{\alpha} \alpha_k X_{i,k}^2 \) and \( \mu = \Sigma b \).

The \(-d\) term can be cancelled out in the probability distribution:
\[ P(y|X) = \frac{\exp(\Psi)}{\int_{-\infty}^{\infty} \exp(\Psi) dy} \]
\[ = \frac{\exp\left(-\frac{1}{2}(y^T \Sigma^{-1} y) + y \Sigma^{-1} \mu\right)\exp(-d)}{\int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}(y^T \Sigma^{-1} y) + y \Sigma^{-1} \mu\right)\exp(-d) dy} \]
\[ = \frac{\exp\left(-\frac{1}{2}(y^T \Sigma^{-1} y) + y \Sigma^{-1} \mu\right)}{\int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}(y^T \Sigma^{-1} y) + y \Sigma^{-1} \mu\right) dy} \]

The definite integral of \(\exp\left(-\frac{1}{2}(y^T \Sigma^{-1} y) + y \Sigma^{-1} \mu\right)\) with respect to \(y\) can be found by using the formula for an integral with square and linear terms:

\[ \exp\left(-\frac{1}{2}(y^T \Sigma^{-1} y) + y \Sigma^{-1} \mu\right) = \frac{(2\pi)^{\frac{n}{2}}}{|\Sigma^{-1}|^{\frac{1}{2}}} \exp\left(\frac{1}{2} \mu \Sigma^{-1} \mu\right) \]

Finally, we obtain the Gaussian formulation of the probability distribution:

\[ P(y|X) = \frac{\exp\left(-\frac{1}{2}(y^T \Sigma^{-1} y) + y \Sigma^{-1} \mu\right)}{(2\pi)^{\frac{n}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left(\frac{1}{2} \mu \Sigma^{-1} \mu\right) \]

**Partial Derivatives of the CCRF log-likelihood**

This section describes the calculation of the partial derivatives of the CCRF log-likelihood with respect to the parameters \(\alpha\) and \(\beta\) as it is described in [2]. The log-likelihood expressed as the Gaussian distribution as derived above is:
\[
\log(P(y|X)) = -\frac{1}{2}(y - \mu)^T \Sigma^{-1}(y - \mu) - \log((2\pi)^{\frac{n}{2}}|\Sigma|^\frac{1}{2}) \\
= -\frac{1}{2}(y - \mu)^T \Sigma^{-1}(y - \mu) - \left(\frac{n}{2}\log(2\pi) + \frac{1}{2}\log|\Sigma|\right) \\
= -\frac{1}{2}(y - \mu)^T \Sigma^{-1}(y - \mu) + \frac{1}{2}\log|\Sigma^{-1}| - \frac{n}{2}\log(2\pi) \\
= -\frac{1}{2}y^T \Sigma^{-1}(y - \mu) + \frac{1}{2}\mu^T \Sigma^{-1}\mu + \frac{1}{2}\log|\Sigma^{-1}| - \frac{n}{2}\log(2\pi) \\
= -\frac{1}{2}y^T \Sigma^{-1}y + \frac{1}{2}\mu^T \Sigma^{-1}\mu + \frac{1}{2}\log|\Sigma^{-1}| - \frac{n}{2}\log(2\pi) \\
= -\frac{1}{2}y^T \Sigma^{-1}y + \frac{1}{2}\mu^T \Sigma^{-1}\mu + \frac{1}{2}\log|\Sigma^{-1}| - \frac{n}{2}\log(2\pi) + \frac{1}{2}\log|\Sigma| \\
= -\frac{1}{2}y^T \Sigma^{-1}y + \frac{1}{2}\mu^T \Sigma^{-1}\mu + \frac{1}{2}\log(|\Sigma|) \\
\]

(3.23)

Now we can note all the necessary partial derivatives. We start with the derivatives of the likelihood for the alphas.

\[
\frac{\delta \Sigma^{-1}}{\delta \alpha} = \frac{\delta 2A + 2B}{\delta \alpha} = 2\delta A \frac{\delta A}{\delta \alpha} = 2I \\
\frac{\delta b}{\delta \alpha} = \frac{\delta 2X\alpha}{\delta \alpha} = 2X
\]

(3.24)

(3.25)

where \(I\) is the identity matrix of size \(n \times n\) and \(n\) denotes the number of nodes in the CCRF. To calculate the derivative of the term \(\frac{\delta b^T \Sigma b}{\delta \alpha}\) a trick of using the partial derivative of a matrix inverse is used \((\frac{\delta M^{-1}}{\delta \alpha} = -M^{-1}\frac{\delta M}{\delta \alpha}M^{-1})\) to get the partial derivative of \(\Sigma:\)

\[
\frac{\delta b^T \Sigma b}{\delta \alpha} = \frac{\delta b^T \Sigma b}{\delta \alpha} \Sigma b + b^T \frac{\delta \Sigma b}{\delta \alpha} = 2X\mu + b^T \left(\frac{\delta \Sigma}{\delta \alpha} b + \Sigma \frac{\delta b}{\delta \alpha}\right) \\
= 2X\mu + b^T \Sigma \frac{\delta b}{\delta \alpha} \Sigma b + b^T \Sigma 2(X)^T = 4X\mu + b^T \Sigma \frac{\delta b}{\delta \alpha} b \\
= 4X\mu + b^T \left(-\Sigma \frac{\delta \Sigma^{-1}}{\delta \alpha} \Sigma\right) b = 4X\mu - 2b^T \Sigma \Sigma b \\
= 4X\mu - 2\mu^T \mu
\]

(3.26)

What is left is the derivative of the term coming from the normalization function:
\[
\frac{\delta |\Sigma^{-1}|}{\delta \alpha} = \frac{1}{|\Sigma^{-1}|} \frac{\delta |\Sigma^{-1}|}{\delta \alpha} = \frac{1}{|\Sigma^{-1}|} |\Sigma^{-1}| \times \text{trace}(\Sigma \frac{\delta \Sigma^{-1}}{\alpha}) \\
= 2 \times \text{trace}(\Sigma I) = 2 \times \text{trace}(\Sigma)
\]

(3.27)

And the full derivative of \( \log(P(y|X)) \) with respect to \( \alpha \) is:

\[
\frac{\delta \log(P(y|X))}{\delta \alpha} = -y^T y + 2y^T X^T - 2X\mu + \mu^T \mu + \text{trace}(\Sigma)
\]

(3.28)

Next, we note the partial derivatives of \( \log(P(y|X)) \) with respect to \( \beta \):

\[
\frac{\delta \Sigma^{-1}}{\delta \beta_k} = 2B^{(k)}
\]

(3.29)

with

\[
B^{(k)} = \begin{cases}
(\sum_{r=1}^n S^{(k)}_{i,r}) - S^{(k)}_{i,i}, & i = j \\
-S^{(k)}_{i,k}, & i \neq j
\end{cases}
\]

(3.30)

\[
\frac{\delta b}{\delta \beta_k} = 0
\]

(3.31)

\[
b^T \Sigma b = -b^T (\Sigma \frac{\delta \Sigma^{-1}}{\delta \beta}) \Sigma b = -2b^T \Sigma B^{(k)} \Sigma b = -2\mu^T B^{(k)} \mu
\]

(3.32)

\[
\frac{\delta \log|\Sigma^{-1}|}{\delta \beta_k} = \frac{1}{|\Sigma^{-1}|} \frac{\delta |\Sigma^{-1}|}{\delta \beta_k} = \frac{1}{|\Sigma^{-1}|} |\Sigma^{-1}| \times \text{trace}(\Sigma \frac{\delta \Sigma^{-1}}{\beta_k}) \\
= 2 \times \text{trace}(\Sigma B^{(k)}) = 2 \times \text{vec}(\Sigma)^T \text{vec}(B^{(k)})
\]

(3.33)

Now the above derivations can be combined to obtain the partial derivative of \( \log(P(y|X)) \) with respect to \( \beta \):

\[
\frac{\delta \log(P(y|X))}{\beta_k} = -y^T B^{(k)} y + \mu^T B^{(k)} \mu + \text{vec}(\Sigma)^T \text{vec}(B^{(k)})
\]

(3.34)
where \( \text{Vec} \) stands for the matrix vectorization operation which stacks up the columns of a matrix to form a single column matrix.

As stated above, the log-likelihood is actually optimized by using the partial derivatives with respect to \( \log \alpha \) and \( \log \beta_k \) instead of using just \( \alpha \) and \( \beta_k \), in order to guarantee that \( \alpha > 0 \) and \( \beta_k > 0 \). These derivatives are given by:

\[
\frac{\delta \log(P(y|X))}{\delta \log \alpha} = \alpha \left( \frac{\delta \log(P(y|X))}{\delta \alpha} \right) \tag{3.35}
\]

\[
\frac{\delta \log(P(y|X))}{\delta \log \beta_k} = \beta_k \left( \frac{\delta \log(P(y|X))}{\delta \beta_k} \right) \tag{3.36}
\]

### 3.4.3 Inference

As described above, the \( \text{CCRF} \) can be viewed as a multivariate Gaussian. Therefore, the prediction \( Y' \) that maximizes \( P(Y'|X) \) is the mean value of the distribution:

\[
Y' = \arg \max_{Y} P(Y'|X) = \mu = \Sigma b \tag{3.37}
\]

This way, the \( \text{CCRF} \) predicts the missing values \textit{as well as} the values already in the training set. In order not to predict the known values anew, the resulting multivariate Gaussian can be conditioned on the known training values [8]:

Let \( Y' \sim \mathcal{N}(\mu, \Sigma) \) denote the normal distribution that is defined by \( \Sigma \) and \( \mu \) as derived above. \( Y' \) can be partitioned into values that we wish to predict \( y_1 \) and known values \( y_2 \) as follows:

\[
Y' = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}
\]

Let the length of \( y_1 \) be \( q \) and the total number of values to predict be \( N \). \( \mu \) and \( \Sigma \) can be partitioned accordingly as follows:

\[
\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}
\]
\[
\Sigma = \begin{pmatrix}
\Sigma_{11} & \Sigma_{12} \\
\Sigma_{21} & \Sigma_{22}
\end{pmatrix}
\]

where \( \mu_1 \) is of length \( q \), \( \mu_2 \) is of length \( N - q \), \( \Sigma_{11} \in q \times q \), \( \Sigma_{12} \in q \times (N - q) \), \( \Sigma_{21} \in (N - q) \times q \) and \( \Sigma_{22} \in (N - q) \times (N - q) \). The conditional of \( y_1 \) on \( y_2 = a \) is again a multivariate normal Gaussian \( (y_1|y_2 = a) \sim N(\hat{\mu}, \hat{\Sigma}) \), with

\[
\hat{\mu} = \mu_1 + \Sigma_{12} \Sigma_{22}^{-1} (a - \mu_2)
\]

\[
\hat{\Sigma} = \Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21}
\]

### 3.5 MF+CCRF

This section gives an overview of the combined \( MF+CCRF \) model. Assume we have a matrix \( R \in \mathbb{R}^{[D] \times [T]} \) given and corresponding similarity matrices \( S_D \in \mathbb{R}^{[D] \times [D]} \) and \( S_T \in \mathbb{R}^{[T] \times [T]} \) for the drugs \( D \) and targets \( T \) as described in sec 3.2. It is our goal to predict the binding affinities of those drug-target pairs in \( R \) for which no binding value is given. The given binding affinities serve as training data.

The main part of the \( MF+CCRF \) model is on the \( CCRF \) part. The purpose of \( MF \) is to give a first prediction on the missing drug-target pairs which is then refined by the \( CCRF \). The \( CCRF \) interpolates between similar drugs and targets, where similarity is defined through the given similarity matrices of the drugs and targets. This interpolation can happen either between two drug-target pairs, where a prediction is made for both pairs or from a known affinity value of one drug-target pair to a pair that we wish to predict. The \( CCRF \) learns the importance of the \( MF \) prediction and the importance of the interpolation between drug-target pairs from the training data. Assume that \( Y_{train} = \{y_1, \ldots, y_n\} \) is the set of known binding values in \( R \) and \( X_{train} = x_1, \ldots, x_n \) represents the \( MF \) prediction on those training values, meaning that \( x_1 \) is the \( MF \) prediction on training value \( y_1 \), \( x_2 \) is the \( MF \) prediction on training value \( y_2 \) and so on. The purpose of the vector \( X_{train} \) is to represent the performance of \( MF \) on the missing values. In order not to overfit the \( MF \) model on
the training data, when obtaining \( X_{\text{train}} \) (as this would result in an over optimistic vector \( X_{\text{train}} \)), the training data is partitioned into 5 folds of which 4 folds are used as training data to predict the fold that was left out. The procedure is repeated for each fold to get the complete vector \( X_{\text{train}} \) as illustrated in figure 3.1. When cross-validation is used to evaluate the models performance, the above procedure of obtaining \( X_{\text{train}} \) can be seen as an inner cross-prediction step on the training data itself. Each element in \( Y_{\text{train}} \) and each element in \( X_{\text{train}} \) correspond to a node of the graphical structure of the CCRF as illustrated in Figure 3.2. The edges of the graphical model are defined through the similarity matrices of the drugs and targets. One possible way of connecting the nodes is by connecting each drug to its \( k \) most similar drugs and similarly connect each target to its \( k \) most similar targets. Another way of choosing the edges is by connecting all drugs and targets respectively whose similarity is above a user defined threshold. Following the notation in 3.4, we have one initial prediction \( X_{\text{train}} \) (thus \( K_1 = 1 \) as mentioned in 3.4). Looking back at equation 3.6, we see that \( K_2 \) different similarity matrices \( S_{i,j}^k, k \in \{1, \ldots, K_2\} \) can be given for the nodes. We can for example choose to learn one parameter \( \beta \) for all edges but we can also split the edges into edges that are defined through the drug-drug similarity and edges that are defined through the target-target similarity. In the second case we would learn two edge parameters \( \beta_1 \) and \( \beta_2 \). Other variations are of course possible, such as learning separate edge parameters for each drug and target. Figure 3.2 shows an example, where all nodes are connected to those nodes where the drug or target similarity is above a threshold of 0.9. Note that the \( S_{i,j}^k \) matrices are of dimension \( n \times n \) where \( n \) is the number of nodes in the CCRF. For the training step, the number of nodes in the CCRF is \( n = Y_{\text{train}} \) and thus \( S_{i,j} \in \mathbb{R}^{Y_{\text{train}} \times Y_{\text{train}}} \). For the prediction step, the number of nodes in the CCRF is equal to the number of cells in the matrix \( R \) and thus \( S_{i,j} \in \mathbb{R}^{|D| \times |T| \times |D| \times |T|} \).
Figure 3.1: Illustration of steps to get vector $X$: The training data is partitioned into $k$ folds. For each fold, the training data is masked and $MF$ predicts all remaining values in the matrix including the masked fold. Vector $X$ is obtained, by combining the predictions of each fold. Note that the vector $X$ represents the predictions of the given matrix values as a vector in row-first order and that the length of the vector $X$ is equal to the number of training observations. In this example, a matrix of 4 drugs and 4 targets with 9 training points is given, thus the length of $X$ is 9.
Figure 3.2: Illustration of Parameter Learning: Input for the learning step of the CCRF are the given observations in the drug-target matrix, the vector $X$ that was predicted by $MF$, and the similarity matrices of the drugs and targets. Here, the given observations are transformed into a vector $Y$ in row-first order (corresponding to vector $X$). The graphical structure of the CCRF is defined by the similarity matrices of the drugs and targets. A threshold $t$ can be applied to connect only those nodes that have a similarity $> t$. In the above example, drugs $(d_1, d_3)$ and $(d_3, d_4)$ and targets $(t_2, t_4)$ have similarity scores above the threshold. Thus the nodes $(1,5)$ and $(2,6)$ are connected (because they belong to drugs $(d_1, d_3)$). Further the nodes $(5,7)$ and $(6,8)$ are connected (because they belong to drugs $(d_3, d_4)$) and the nodes $(4,8)$ are connected because they belong to targets $(t_2, t_4)$. One parameter $\alpha$ is learned for the dependance between $Y$ and $X$. Another parameter $\beta_1$ is learned for the dependance between similar drugs (green edges) and a third parameter $\beta_2$ is learned for the dependance between similar targets (orange edges).
3.5.1 Integrating the similarity matrices

For the $MF+CCRF$ method that is developed in this thesis it is possible to integrate:

- none of the similarity matrices
- only the drug-drug similarity matrix
- only the target-target similarity matrix
- both similarity matrices

Here, using no similarity metric means to use just the $MF$ prediction which in regard to the developed model is the least interesting use case. The results of using only the $MF$ prediction are still listed in section 4.3 to show the improvement of integrating the similarity matrices. When using only the drug similarity, we can build separate graphical models for each target because no dependency between the targets is introduced. Thus, in this case, a separate set of parameters $\alpha$ and $\beta$ can be learned for each target and a separate prediction can be made for each target (for each column of the drug-target matrix). Analogously, when using only the target similarity, we can build separate graphical models for each drug, learn a separate set of parameters $\alpha$ and $\beta$ for each drug and make a separate prediction for each drug (each row of the drug-target matrix). The graphs of the underlying $CCRF$ for the cases when only one of the similarity matrices is used are rather small: when the drug-similarity is used, each graph has $n_d$ nodes, where $n_d$ stands for the number of drugs in the dataset. Analogously, when the target-similarity is used, each graph has $n_t$ nodes, where $n_t$ stands for the number of targets in the dataset. In the first case, we need to compute the inverse of a matrix of size $n_d \times n_d$ and in the second case we have to compute the inverse of a matrix of size $n_t \times n_t$ in the inference step.

When both similarity matrices are introduced the case is more complicated. While for the cases of introducing only one of the similarity matrix, the underlying graphs of the $CCRF$ are rather small, this is not the case when both similarity matrices are used simultaneously. Following the strategy of connecting similar drugs and targets,
we would construct one large graph, where each drug-target pair is one node in the CCRF. The datasets that were used for the experiments in this thesis are too large to create a single graphical model for all drugs and targets because in the inference step it is necessary to compute the inverse of a matrix of size $n_d n_t \times n_d n_t$. In order to overcome this difficulty the drugs and targets were clustered to generate smaller drug-target submatrices of feasible size as explained in section 3.5.2 and a separate CCRF was utilized for each smaller matrix to make a prediction.

The integration of the similarity matrices and the resulting underlying CCRF-graphs are illustrated in figure 3.3.

3.5.2 Graphical structure of the CCRF

As mentioned above, different strategies can be applied to chose the edges of the graphical model. Which of the CCRF nodes are connected is defined through the similarity matrices $S^k$ in the formulation of the CCRF. Theoretically, one can connect all nodes to all other nodes and weight the edges with the similarity between the
corresponding drugs or targets. However, the time of computing the covariance matrix of the Gaussian distribution, which is defined through the CCRF depends on the sparsity of the matrices $S^k$. It is necessary to compute this inverse in each step of the parameter learning as well as in the inference step. In order to get CCRFs with a feasible structure, the edges for the experiments in this thesis where chosen by connecting all drugs and targets to its $k_{\text{neighbors}} = 4$ nearest neighbors. An other strategy to connect the nodes, would be to chose a similarity-threshold and connect all drugs and targets with a similarity above the threshold. However, this can cause that the parameter training and the inference step become expensive in terms of computation time, when too many edges are introduced. The reason for this, is that in each parameter update step, we need to compute the inverse of the matrix $\Sigma^{-1}$ to get the covariance matrix $\Sigma$ (see 3.4). The computation time of this step depends on the sparsity of the matrix $\Sigma^{-1}$, which again depends on the graphical structure of the CCRF.

**Clustering of the Drug-Target Matrix**

Hierarchical clustering was utilized on the drug and target similarity matrices respectively to first cluster the drugs and targets into smaller subsets. For this task the R functions `hclust` and `cutree` were used. When clustering the drugs, based on the drug similarity matrix, `hclust` initially assigns each drug to its own cluster and then proceeds iteratively, at each stage joining the two most similar clusters, until there is only one single cluster left, resulting in a tree as illustrated in figure 3.4. The `cutree` function with parameter $k_d$ then groups the clusters into $k_d$ disjunct sets of drugs. The same procedure is applied on the target similarity matrix resulting in $k_t$ disjunct sets of targets. Finally, we obtain $k_d \times k_t$ separate drug-target sets. For all $k_d$ drug sets we can now go over all $k_t$ target sets and construct a CCRF for the submatrix of the respective drugs and targets.
3.6 Proof of Concept

This section serves as a proof of concept for the model that is developed in this thesis. In a first step, a matrix of \( n \) rows and \( m \) columns is generated which is then partitioned into training and test data (\( n \) simulated drugs and \( m \) simulated targets). The matrix values are generated, such that the following assumptions are fulfilled:

- the dataset has underlying latent factors
- there are certain pairs of columns in the generated dataset, s.th. the pairs of columns have similar values (simulating similar targets).

In a second step, a fraction of observations in this generated dataset is masked as test data. The remaining values are used as training data for the MF+CCRF model. The model parameters are trained based on the training data and in the next step the values that were previously masked as test data are predicted. Finally, the performance of this model is evaluated in terms of Root Mean Square Error (\( RMSE \)) and compared to the performance of using only MF or only CCRF. In the following, each of these steps is documented, starting with the simulation of the data.

3.6.1 Data Simulation

A dataset of dimension \( m \times n \) (\( m \) drugs and \( n \) targets) that has underlying latent factors and column-pairs with similar values can be simulated by first constructing a CCRF and then taking a sample from it. At first, a matrix \( X \) of size \( m \times n \) is generated which has only underlying latent factors as shown in figure 3.5. Next, the graphical structure of the CCRF is defined by arbitrarily choosing pairs of columns for which the CCRF distribution should produce similar values. In the CCRF formulation each matrix cell is a node and each node can be adjacent to any other node in the graphical model. Therefore the graphical structure of the CCRF is defined by an adjacency matrix \( B \) of size \( mn \times mn \). Figure 3.6 shows an example of a choice of structure for the CCRF. The values in matrix \( X \) are used as node input for the CCRF. Next
\(\alpha\) (importance of the node values) and \(\beta\) (importance of the adjacency matrix) are chosen arbitrarily and a sample from the multivariate Gaussian distribution that can be derived from the CCRF formulation as described in section 3.4 is taken. This sample is used as the dataset for the numerical experiments. It has both underlying latent factors (because the input node values of matrix \(X\) has underlying latent factors) and the assumption that similar targets (which were chosen arbitrarily by defining the matrix \(B\)) have similar values is fulfilled. Figure 3.7 shows an example of a sample, taken from the multivariate Gaussian that is defined by the CCRF where the structure of the model was chosen as described in figure 3.6.

### 3.6.2 Experiments on Simulated Data

For the numerical experiments a dataset of \(n = 40\) rows and \(m = 40\) columns was generated. As described above, for the dataset generation, a first dataset \(X\) was generated which has only underlying latent factors. Next the CCRF parameters were chosen arbitrarily as \(\alpha = 1\) and \(\beta = 2\). The graphical structure of the CCRF was chosen as illustrated in figure 3.6. Next, the corresponding multivariate Gaussian was obtained, by first calculating the covariance matrix \(\Sigma\) and mean vector \(\mu\) as described in section 3.4. A sample from this distribution was taken as the simulated dataset which is visualized in figure 3.8. Next, around 120 training values were sampled from the complete dataset. An example of a sampled training dataset is shown in figure 3.9. First, Matrix Factorization was applied to predict the missing values and the \(RMSE\) of the prediction was computed. Next, the CCRF+MF model was applied to predict the missing values and the \(RMSE\) was computed again for this prediction. The predictions of the MF and MF+CCRF model are visualized in figures 3.10 and 3.11. The prediction of the CCRF model without MF as an initial prediction was also tested. As it is necessary to have some node input \(X\), the mean value of the training observations was used instead of the MF prediction for the CCRF model.
Figure 3.4: Clustering the drugs in the *Davis* dataset, which is introduced in section 4.1. The dendogram illustrates the similarities of the drugs in the *Davis* dataset. The numbers at each leaf represent the drug indices.

Figure 3.5: Sampling matrix $X$ which is used for the node input of the *CCRF* distribution. First two latent-factor-matrices $LF_d$ and $LF_t$ of dimension $n \times k$ and $m \times k$, where $n$ is the number of drugs, $m$ is the number of targets and $k$ is the number of latent factors (here $k = 10$ is chosen arbitrarily) are sampled randomly. $X$ is defined as the product $X = LF_d LF_t^T$ and is used as the values for the node input of the *CCRF*.
Figure 3.6: Defining the structure of the CRF. Assume each matrix cell is a node of the CCRF. For the numerical experiments, the matrix columns were divided into groups of 10 columns. Each cell in each group was connected to all other cells in the same group, which is illustrated by the archs.

Figure 3.7: Example of a dataset that is sampled from the CCRF when the structure is chosen, such as described in figure 3.6. As can be seen, this dataset has the property that the nodes which are interconnected by the structure of the graphical model have similar values. Additionally this matrix has underlying latent factors because the node input values were generated by multiplying randomly sampled latent factor matrices.
by itself. Table 3.1 lists the RMSE of the methods and shows that the MF+CCRF model outperforms the MF and CCRF models when used singularly.

\begin{tabular}{|c|c|c|}
\hline
RMSE & MF  & CCRF & MF+CCRF \\
\hline
11.493 & 11.909 & 7.514 \\
\hline
\end{tabular}

Table 3.1: RMSE of MF, CCRF and MF + CCRF on simulated dataset.

3.6.3 Summary

As pointed out in 3.1, the idea of the combined MF+CCRF model is that the CCRF can improve the initial MF prediction of a drug-target pair by interpolating from given observations of a similar drug on the same target or by interpolating from a given observation of a similar target on the same drug. For the numerical experiments, the
first type of interpolation proved to be useful. The interpolation among the columns through the CCRF improved the performance of using only MF significantly. The MF method alone does not capture the high correlation of the values inside the target groups as well as the combined method, which can be seen in figure 3.10 and figure 3.11. The combined method also improves the performance of using the CCRF alone.

The reason for this is simple: some rows of the training dataset will not contain any observations for a complete target group (in figure 3.9 for example the first row does not contain any observations for the second group of targets $t_{11}, \ldots, t_{20}$). When using only the CCRF without MF as an initial prediction, the prediction in these groups will be the mean value of the training data. When using MF as an initial prediction instead, the MF will have captured some information for these groups because of the underlying latent factors, which can be learned from the remaining target groups.
Chapter 4

Experiments

4.1 Datasets

In the majority of studies that can be found in the literature, the presented models for drug-target interaction prediction are trained and evaluated on binary datasets. Typically, the existing models are evaluated on the four binary datasets that were first presented in [26]. In these datasets a label of $y_{d_i,t_j} = 1$ is given for a drug-target pair $(d_i,t_j)$ which is known to interact and a label of $y_{d_i,t_j} = 0$ is given when either the drug-target pair is known not to interact or when it is unknown whether the pair interacts. In contrast to a model that classifies if a drug-target pair interacts or not, the model that is developed in this thesis learns and predicts the continuous binding affinities of drug-target pairs. To the best of my knowledge, only one existing study can be found in the literature which presents a model for the prediction of the continuous binding affinity of drugs and targets, which is the model *KronRLS* [17] which was introduced in section 2. In the original paper, *KronRLS* is evaluated on two continuous datasets (*Metz* and *Davis*) that are also used in this thesis for the evaluation of the presented model. A third dataset *KIBA* is obtained by preprocessing the drug-target dataset that is presented in [20]. The three datasets named *Metz*, *Davis* and *KIBA* respectively that are used for evaluating the developed model are described in the following chapters. Additionally, the corresponding drug-drug and
target-target similarity matrices that were used to construct the graphical model of the CCRF are described in a following section. Table 4.1 lists the sizes and densities of the datasets. Here, density means the percentage of drug-target pairs in the dataset for which an observation is given.

4.1.1 The Davis Dataset

The continuous dataset Davis was used for the evaluation of the drug-target interaction prediction model presented in [17]. The dataset itself was published in the study [3]. For this dataset, the interaction of 68 kinase inhibitors with 442 kinases was tested and measured as the $K_d$ value. The kinase inhibitors are the drugs and the kinases are the targets in the more general formulation of drugs and targets. The Davis dataset contains the full information of binding affinities for all drug-target pairs in the dataset, and thus contains no missing values. A lower $K_d$ value indicates a higher binding affinity between the drug and the target. As described in [3], the binding affinity is not reported if it was measured to be $> 10000$. For these drug-target pairs, a $K_d$ value of 10000 was used for the experiments in this thesis. The $K_d$ values in the Davis dataset were log transformed, according to the formula:

$$pK_d := -\log_{10}(\frac{K_d}{1e9})$$

(4.1)

and thus after the log-transformation a higher $pK_d$ value represents a higher binding affinity. The drug-drug and target-target similarity matrices for this dataset can be downloaded from the website of the author of [17]. The distribution of $pK_d$ values in this dataset is illustrated in figure 4.1.
4.1.2 The Metz Dataset

Just as the Davis dataset, the continuous dataset Metz was used for the evaluation of the drug-target interaction prediction model presented in [17]. The dataset was published in the study [16]. The Metz dataset consists of 1421 drugs and 156 targets. The binding affinity is given as log transformed $K_i$ values (called $pK_i$ values) for 42% of the drug-target pairs. As drug-drug and target-target similarities for this dataset the matrices were used that can be downloaded from the website of the author of [17]. The distribution of $pK_i$ values in this dataset is illustrated in figure 4.2.

4.1.3 The KIBA Dataset

The Davis and Metz datasets are suitable for the evaluation of predictive models for drug target interaction because data heterogeneity is not an issue. We can assume that the experimental settings for the measured drug target pairs in each dataset were the same and the binding affinities are comparable. When working with experimental results that come from multiple sources the data might be heterogeneous: In one case the binding affinity might be measured by $K_i$, in another
case by $K_d$ and in a third case by $IC_{50}$ value. Another source of data heterogeneity are different experimental settings. An approach to integrate the observations from different sources, named KIBA (short for Kinase Inhibitor BioActivity) and a corresponding dataset is presented in [20]. With their method, the authors of [20] integrated the experimental results from multiple databases into a bioactivity matrix of 52498 compounds and 467 targets, including 246088 observations. The binding affinities in this matrix are given as KIBA-scores. This dataset was used to obtain a third evaluation dataset, which is called the KIBA dataset, by removing all drugs and targets with less than 10 observations from the original dataset that was downloaded from the supplementary material of [20], resulting in a dataset of 2116 drugs and 229 targets with a density of 24%. For this dataset the drug-drug similarity matrix was computed through the PubChem structure clustering tool (https://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?p=clustering). The target target similarity matrix was obtained by computing the normalized Smith Waterman Score [26] for each pair of targets. The distribution of KIBA scores in this dataset is illustrated in figure 4.3.
4.1.4 Similarity Metrics of Drugs and Targets

As drug-drug and target-target similarity matrices for the Davis and Metz dataset the precomputed matrices that are provided on the website of [17] were used. Here, the drug-drug similarity was computed based on the 2D chemical structure of the compounds: The compounds are first represented by the graph of their 2D chemical structure and the similarity score is computed based on the size of the common substructures [1]. The target-target similarity was computed based on the protein sequences, using the normalized Smith-Waterman score [25]. For the KIBA dataset, the drug-drug similarity matrix was obtained through the compound clustering tool of PubChem. The given ChEMBL IDs of the compounds were first matched to their PubChem CIDs which were then used as input to the PubChem web interface (The PubChem website offers a tool that allows to match ChEMBL IDs to PubChem CIDs). The clustering tool allows to download a similarity matrix for the compounds which is computed based on the compound structure (similarly as for the drug-drug similarity of the Metz and David datasets). For the KIBA dataset, the protein sequences
were downloaded from NCBI and the normalized Smith Waterman similarity was computed for each pair by aligning the sequences using the Biostrings R package.

**Correlation of Drug/Target Similarity and Binding Behaviour**

In order to examine the correlation between the similarity matrices and the binding behaviour (meaning to find out if similar drugs or targets actually show similar binding behaviour) the following data analysis was performed. First, the analysis regarding the drug-similarity: all pairs of drugs were selected, for which five or more targets can be found, such that both drugs were tested against those five or more targets. The pairs of drugs were subset into two disjunct sets. The first subset contains only pairs of drugs for which the similarity is below a threshold of 0.7, meaning that this subset contains drug pairs of low similarity. The second subset contains all the drug pairs for which the similarity is above the threshold, meaning this subset contains drug pairs of high similarity. The following steps were then performed for both subsets: For each pair of drugs \((d_a, d_b)\) in the set, all targets that both drugs were tested against were selected. This set of targets was subset into \(binding_a\), representing all targets to which \(d_a\) binds with a high affinity (the thresholds to define high affinity for the datasets are described in the next section) and \(binding_b\), representing all targets to which \(d_b\) binds with a high affinity. For each pair of drugs, the size of the union of \(binding_a\) and \(binding_b\) (x-Axis in the following plots) and the size of the intersection of \(binding_a\) and \(binding_b\) (y-Axis in the following plot) was computed. Finally, the number of times each combination of union and intersection was obtained is counted (the count is represented by the bubble size and color in the following plots). The same analysis can be repeated analogously for the similar and un-similar target pairs of the datasets.

Figures 4.4 and 4.5 show on the x-Axis the size of the union and on the y-Axis the size of the intersection of the drug-drug pairs for the Metz dataset. The bubble-size represents the number of drug-drug pairs that were found with the corresponding
sizes of the union and intersection. Comparing figures 4.4 and 4.5, we can see that drug-drug pairs with a similarity above 0.7 tend to share more targets to which both drugs bind with high affinities than drug-drug pairs with a similarity below 0.7. The same can be observed for drug-drug pairs in the KIBA dataset illustrated in figures 4.6 and 4.7. For the Davis dataset it seems like no such correlation can be observed (see figures 4.8 and 4.9) which goes hand in hand with the observation in the results section, that the drug similarity does not improve the prediction performance. On the other hand as illustrated in figures 4.10 and 4.11, the target similarity correlates with the binding behaviour of the targets for the Davis datasets.

4.2 Experimental Settings

4.2.1 Binary vs. Continuous Prediction

For the evaluation of the model, a performance evaluation for the classification of drug-target pairs into binding or non-binding was also included, using the metrics AUC and AUPR. Therefore, the datasets were binarized by applying thresholds as it was done in [17]. The MF + CCRF model can only predict continuous values, therefore the threshold was applied on the true values after the prediction step, in order to compute AUC and AUPR. In the Davis and Metz datasets, the higher the $pK_d$ or $pK_i$ value, the higher the binding affinity between a drug and a target. For the Metz dataset, the same threshold of $pK_i \geq 7.6$ was used as suggested in [17] to assign a label of 1 meaning binding and 0 meaning non-binding. For the Davis dataset, a threshold of $pK_d \geq 7.0$ which is a bit less stringent than the threshold that is suggested in [17] is used. In the original KIBA dataset, the lower the KIBA-score, the higher the binding affinity and [20] suggests a threshold of KIBA-score $\leq 3$ to binarize the dataset. In an additional preprocessing step, the KIBA-scores were transformed by taking the negative of each value and adding the minimum to all values in order to obtain a threshold where all values above the threshold are classified as binding.
Figure 4.4: Illustration of binding behaviour of drug-drug pairs in *Metz* dataset with similarity above 0.7.

Figure 4.5: Illustration of binding behaviour of drug-drug pairs in *Metz* dataset with similarity below 0.7.
Figure 4.6: Illustration of binding behaviour of drug-drug pairs in KIBA dataset with similarity above 0.7.

Figure 4.7: Illustration of binding behaviour of drug-drug pairs in KIBA dataset with similarity below 0.7.
Figure 4.8: Illustration of binding behaviour of drug-drug pairs in *Davis* dataset with similarity above 0.7.

Figure 4.9: Illustration of binding behaviour of drug-drug pairs in *Davis* dataset with similarity below 0.7.
Figure 4.10: Illustration of binding behaviour of target-target pairs in *Davis* dataset with similarity above 0.7.

Figure 4.11: Illustration of binding behaviour of target-target pairs in *Davis* dataset with similarity below 0.7.
The $KIBA$ threshold of 3 in the un-transformed dataset then becomes 12.1 in the transformed $KIBA$ dataset.

It is noteworthy, that when the classification metrics $AUC$ and $AUPR$ are applied, $KronRLS$ learns and predicts binary labels, meaning that the datasets are binarized according to the cutoff thresholds before the training step. The $MF+CCRF$ method in contrast, only predicts continuous values and the cutoff threshold is applied after the prediction step to calculate the $AUC$ and $AUPR$. One can argue, that given two models $A$ and $B$, where $A$ learns to predict continuous values and model $B$ learns to predict binary values, and the performance of model $A$ in terms of $AUC$ and $AUPR$ is as good as the performance of model $B$, then model $A$ is advantageous because it does not need to be retrained (as model $B$) when the cutoff threshold for the dataset is changed.

4.2.2 Performance Evaluation

The model is evaluated using 5 fold cross-validation. In this procedure, the given values in the datasets are randomly partitioned into 5 subsets of equal size. Each subset is in turn used as validation data to test the method that was trained on the remaining 4 subsets. Both, the $MF+CCRF$ model and the comparison model $KronRLS$ can use as input either only one of the similarity metrics or both (in addition to the training matrix with the observed binding affinities). When $KronRLS$ gets as input only one of the similarity matrices for example only the drug similarity, it automatically generates a similarity kernel for the targets, where each target is only similar to itself, meaning the identity matrix is imputed for the targets (and vice-versa, when only the target-similarity is given, the identity matrix is used for the drugs). For each dataset, for each performance metric and for each method ($KronRLS$ and $MF+CCRF$), we get 3 evaluation scores that can be compared among the two methods as illustrated in. For the $MF+CCRF$ model a fourth evaluation score is listed in the results section, which is the performance of using no similarity
Figure 4.12: Description of the evaluation tables. 2D refers to the drug similarity (the drug similarity is based on the 2-dimensional structure of the compounds as described above). SW refers to the target similarity and stands for Smith-Waterman as described in the Dataset section. \( \delta \) is used in the row/column where the drug/target similarity is not used.

information, resulting in the prediction of only \( MF \). The used performance metrics are described in the following sections. Figure 4.12 explains how to read the tables, listing the performance scores

**Evaluation Metrics**

As suggested in [17], the concordance index \( (CI) \) can be used as an evaluation metric for the prediction accuracy as it takes into account that the interaction affinities behind drug-target interactions are continuous values rather than binary ones. The intuition behind the \( CI \) is as follows: the \( CI \) over a set of paired data is the probability that the predictions for two randomly drawn drug-target pairs with different label values are in the correct order, meaning that the prediction \( f_i \) for the larger affinity \( y_i \) is larger than the prediction \( f_j \) for the smaller affinity value \( y_j \):

\[
CI = \frac{1}{Z} \sum_{y_i > y_j} h(f_i - f_j)
\]  

(4.2)
where $Z$ is a normalization constant that equals the number of data pairs with different label values, and $h(u)$ is the step function returning 1.0, 0.5 and 0.0 for $u > 0$, $u = 0$ and $u < 0$ respectively. The $CI$ ranges between 0.5 and 1.0, where 0.5 corresponds to a random predictor and 1.0 corresponds to perfect prediction accuracy [17].

As mentioned above, the datasets were also binarized according to cut-off thresholds to evaluate the performance on the classification task. In case of binary interaction labels, the $CI$ becomes equal to the commonly used Area Under the Receiver Operating Characteristic Curve (AUC) metric [17]:

$$AUC = \frac{1}{m_+ m_-} \sum_{y_i=1, y_j=-1} h(f_i - f_j)$$ (4.3)

where $m_+$ and $m_-$ are the numbers of drug-target pairs belonging to the positive and negative classes. For the binary classification task, the performance was also evaluated using the Area Under the Precision-Recall curve (AUPR), which has been used in most of the previous studies on drug-target interaction prediction.

### 4.3 Results

This chapter lists the results of the experiments as described in the previous section for the comparison model KronRLS and the MF+CCRF model. Tables 4.2, 4.3 and 4.4 list the performances in terms of $CI$, $AUC$ and $AUPR$ respectively for both methods and all three datasets. The cases of integrating either only one similarity matrix or both similarity matrix are listed separately as discussed above. For the MF+CCRF model the performance of using only MF is listed additionally. For the KronRLS model the table entry where none of the similarities is used, is left empty. We observe the following:

- Regarding the performance of the MF+CCRF model, we observe that the integration of the similarity matrices through the CCRF brings a significant im-
Improvement in performance when compared to the performance of using only MF. The improvement is most evident for the classification metrics AUC and AUPR. For the Davis dataset, the integration of the similarity matrices CCRF raises the AUC from 0.86 (MF) to 0.95 (MF+CCRF, both similarity matrices). For the Metz dataset, the AUC is raised from 0.88 (MF) to 0.94 (MF+CCRF, both similarity matrices). In terms of AUPR, the CCRF rises the performance from 0.49 (MF) to 0.67 (MF+CCRF, both similarity matrices) on the Davis dataset, from 0.33 (MF) to 0.58 (MF+CCRF, both similarity matrices) on the Metz dataset and from 0.63 (MF) to 0.72 (MF+CCRF, both similarity matrices) on the KIBA dataset.

- When comparing the performance of MF+CCRF to KronRLS in the settings where only one of the similarity matrices is used, we observe that MF+CCRF significantly outperforms KronRLS on most of the datasets. In particular for the KIBA dataset, MF+CCRF raises the CI from 0.67 to 0.80 when only the drug similarity is used and from 0.65 to 0.80 when only the target similarity is used. On the Metz dataset, MF+CCRF raises the CI from 0.74 to 0.81 when only the drug similarity is used and from 0.69 to 0.78 when only the target similarity is used. Further, MF+CCRF raises the AUC on the Metz dataset from 0.86 to 0.91, when using only the drug similarity and from 0.84 to 0.90, when using only the target similarity. On the Davis dataset, MF+CCRF raises the AUC from 0.75 to 0.81 when using only the drug similarity.

- When integrating both similarity metrics, we observe almost equal performances for KronRLS and MF+CCRF for across all datasets and metrics: MF+CCRF marginally improves the performance of KronRLS in regard of CI for the Metz dataset, where it raises the CI of 0.78 to 0.82. In all other cases the performances of KronRLS and MF+CCRF are quite similar and do not differ by more than 3%.
Table 4.2: CI of KronRLS and MF + CCRF on the three evaluation datasets.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>CI</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>KronRLS</td>
</tr>
<tr>
<td>Davis</td>
<td></td>
</tr>
<tr>
<td>2D</td>
<td>0.85</td>
</tr>
<tr>
<td>δ</td>
<td>0.84</td>
</tr>
<tr>
<td>SW</td>
<td>δ</td>
</tr>
<tr>
<td>KronRLS</td>
<td>CCRF</td>
</tr>
<tr>
<td>Metz</td>
<td></td>
</tr>
<tr>
<td>2D</td>
<td>0.78</td>
</tr>
<tr>
<td>δ</td>
<td>0.69</td>
</tr>
<tr>
<td>SW</td>
<td>δ</td>
</tr>
<tr>
<td>KronRLS</td>
<td>CCRF</td>
</tr>
<tr>
<td>Kiba</td>
<td></td>
</tr>
<tr>
<td>2D</td>
<td>0.79</td>
</tr>
<tr>
<td>δ</td>
<td>0.65</td>
</tr>
<tr>
<td>SW</td>
<td>δ</td>
</tr>
</tbody>
</table>

Table 4.3: AUC of KronRLS and MF + CCRF on the three evaluation datasets. For a description of the table entries see Figure 4.12

<table>
<thead>
<tr>
<th>Dataset</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>KronRLS</td>
</tr>
<tr>
<td>Davis</td>
<td></td>
</tr>
<tr>
<td>2D</td>
<td>0.93</td>
</tr>
<tr>
<td>δ</td>
<td>0.91</td>
</tr>
<tr>
<td>SW</td>
<td>δ</td>
</tr>
<tr>
<td>KronRLS</td>
<td>CCRF</td>
</tr>
<tr>
<td>Metz</td>
<td></td>
</tr>
<tr>
<td>2D</td>
<td>0.93</td>
</tr>
<tr>
<td>δ</td>
<td>0.84</td>
</tr>
<tr>
<td>SW</td>
<td>δ</td>
</tr>
<tr>
<td>KronRLS</td>
<td>CCRF</td>
</tr>
<tr>
<td>Kiba</td>
<td></td>
</tr>
<tr>
<td>2D</td>
<td>0.88</td>
</tr>
<tr>
<td>δ</td>
<td>0.84</td>
</tr>
<tr>
<td>SW</td>
<td>δ</td>
</tr>
</tbody>
</table>
Table 4.4: AUPR of KronRLS and MF + CCRF on the three evaluation datasets.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>AUPR</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>KronRLS</td>
</tr>
<tr>
<td>Davis</td>
<td></td>
</tr>
<tr>
<td>2D</td>
<td>0.68</td>
</tr>
<tr>
<td>δ</td>
<td>0.64</td>
</tr>
<tr>
<td>SW</td>
<td>δ</td>
</tr>
<tr>
<td>δ</td>
<td>0.60</td>
</tr>
<tr>
<td>Metz</td>
<td></td>
</tr>
<tr>
<td>2D</td>
<td>0.57</td>
</tr>
<tr>
<td>δ</td>
<td>0.28</td>
</tr>
<tr>
<td>SW</td>
<td>δ</td>
</tr>
<tr>
<td>δ</td>
<td>0.43</td>
</tr>
<tr>
<td>Kiba</td>
<td></td>
</tr>
<tr>
<td>2D</td>
<td>0.75</td>
</tr>
<tr>
<td>δ</td>
<td>0.66</td>
</tr>
<tr>
<td>SW</td>
<td>δ</td>
</tr>
<tr>
<td>δ</td>
<td>0.69</td>
</tr>
</tbody>
</table>
Chapter 5

Discussion and Future Work

5.1 Discussion

In this work, a new method, $MF+CCRF$, for the prediction of drug target interaction, which combines Matrix Factorization and Continuous Conditional Random Fields is proposed. In contrast to the vast majority of previous work on machine-learning based methods for drug-target interaction prediction that classify drug-target pairs as either binding or non-binding the proposed model predicts the binding affinity as continuous values, which better reflects the true complexity of the drug-target prediction problem. The model is evaluated on three datasets which are arguably of especially high quality in terms of data homogeneity as the first two data sets, Davis and Metz, originate from individual wetlabs, where the measurements were taken under unified experimental conditions and for the third dataset, KIBA, observations from multiple sources where carefully integrated. Rather than containing only true positive interactions as the previously used binary datasets, the datasets that were used here contain standardized mappings of the $K_i$, $K_d$ and KIBA-scores which provide broader insights into the interaction patterns of the drugs and their potential protein targets. The obtained results in terms of the ranking metric $CI$ and the classification metrics $AUC$ and $AUPR$ show that the proposed methods performs as well as the state of the art method $KronRLS$. As described in section 4.2.1, $KronRLS$ predicts binary values
when the classification metrics $AUC$ and $AUPR$ are applied, while $MF + CCRF$ predicts continuous values on which the binarization threshold is applied after the prediction step. $MF + CCRF$ still performs as good as $KronRLS$ in terms of $AUC$ and $AUPR$ and therefore one can argue that $MF + CCRF$ has the advantage that for different binarization thresholds, $MF + CCRF$ does not need to be retrained, in contrast to $KronRLS$. Especially for the cases when only one similarity matrix of either the drugs or the targets is given, $MF + CCRF$ outperforms $KronRLS$ significantly across almost all datasets and metrics.

Further, we observe that the integration of the similarity matrices through the $CCRF$ significantly improves the performance of Matrix Factorization alone. This suggests that the model can be used to improve the $MF$ prediction in similar settings, for example in user-item recommendation tasks, when suitable similarity matrices for the users/items are given.

### 5.2 Future Work

One issue that is not addressed in this work is the problem of the biased nature of drug-target datasets. The used evaluation datasets are usually highly biased, containing only a small number of drugs and targets with many observations and a large number of drugs and targets with only a few observations. In the cross validation setting this leads to overoptimistic results because the model is mainly evaluated on the few drugs and targets with many observations and the reported evaluation metrics can not be generalized to the complete set of drugs and targets. One direction of future work would be to provide researchers a measure of the confidence of the prediction, so that the most confident predictions can be validated in wetlab experiments.

An other direction of future work would be to find a better strategy for the parameter tuning of the presented model. $KronRLS$ uses an inner cross validation step to find the optimal regularization parameter $\lambda$. A similar strategy could be applied here, where the $MF + CCRF$ model could automatically search for drug-similarity
and target-similarity thresholds, that define which nodes of the CCRF to connect, instead of using the fixed parameter setting of connecting all $k$-nearest neighbors.
Bibliography


