BIOLOGICAL DATA INTEGRATION AND RELATION PREDICTION BY
MATRIX FACTORIZATION

A THESIS SUBMITTED TO

THE GRADUATE SCHOOL OF INFORMATICS OF

THE MIDDLE EAST TECHNICAL UNIVERSITY

BY

GÖKÇE ABAY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE
OF

MASTER OF SCIENCE

IN

BIOINFORMATICS

JANUARY 2020

Approval of the thesis:

BIOLOGICAL DATA INTEGRATION AND RELATION PREDICTION BY MATRIX FACTORIZATION

Submitted by Gökçe Abay in partial fulfillment of the requirements for the degree of **Master of Science in Health Informatics Department, Middle East Technical University** by,

Prof. Dr. Deniz Zeyrek Bozşahin
Dean, **Graduate School of Informatics**

Assoc. Prof Dr. Yeşim Aydın Son
Head of Department, **Health Informatics**

Assist. Prof Dr. Aybar Can Acar
Supervisor, **Health Informatics, METU**

Assoc. Prof. Dr. Tunca Doğan
Co-Supervisor, **Computer Engineering Dept., Hacettepe University**

Examining Committee Members:

Assoc. Prof Dr. Yeşim Aydın Son
Health Informatics Dept., METU

Assist. Prof Dr. Aybar Can Acar
Health Informatics Dept., METU

Assoc. Prof Dr. Tunca Doğan
Computer Engineering Dept., Hacettepe University

Assoc. Prof. Dr. Özlen Konu
Molecular Biology and Genetics Dept., Bilkent University

Assoc. Prof. Dr. Nurcan Tunçbağ
Health Informatics Dept., METU

Date: **30.01.2020**

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last name :   GÖKÇE ABAY

Signature   :   _____

ABSTRACT

BIOLOGICAL DATA INTEGRATION AND RELATION PREDICTION BY
MATRIX FACTORIZATION

Abay, Gökçe

MSc, Department of Bioinformatics

Supervisor: Assist. Prof Dr. Aybar Can Acar

Co-Supervisor: Assoc. Prof Dr. Tunca Doğan

The available molecular sequence data has increased greatly in the last decades, thanks to the new technological developments in the field of life-sciences. In order for this data to be useful to the scientific community, it should be characterized. Traditionally, this characterization is done manually, where the experimentally produced molecular data is curated and stored in the biological databases. The huge volume of the currently available data summons the need for the automatic and systematic analysis. A crucial part of this systematic analysis is data integration with the identification of the relationships between the elements from different biological data types. In this study, we propose to integrate large-scale gene/protein annotation data using non-negative matrix factorization (NMF), which is a frequently used method for recommender systems with successful real-world applications. NMF has also been employed for uniting multi-relational data in many different fields including bioinformatics and cheminformatics. Within the purposes of this study, we first collected protein annotations such as molecular functions, biological processes, sub-cellular localizations and disease relations from different resources such as UniProt-GOA and DisGeNET, and organized them as binary relation matrices. We then applied various NMF-based algorithms to this multi-dimensional relational biomolecular sequence annotation data (i.e. genes/proteins vs. functions, genes/proteins vs. diseases, diseases vs. functions) and evaluated the results of each model in terms of their capacity to learn the intrinsic structure in relational data, via cross-validation. The results indicated that NMF has the capacity to retrieve most of the known protein annotations without using any sequence or structure-based protein features (AUROC: $0.80 - 0.94$, accuracy: $0.53 - 0.64$, F1-score: $0.06 - 0.40$, MCC: $0.13 - 0.38$). Using NMF, the ultimate aim here is to predict the unknown binary relationships between these biological entities; and to

represent these entities (i.e., proteins, functions and disease entries) as informative and non-redundant quantitative feature vectors (using the low-rank feature matrices generated by the factorization process), which can be used in diverse data mining and machine learning tasks in the future, such as the automated annotation of proteins or the construction of biological knowledge graphs.

ÖZ

MATRİS FACTORİZASYONU YÖNTEMİ İLE BİYOLOJİK VERİ ENTEGRASYONU VE İLİŞKİ TAHMİNİ

Abay, Gökçe

Yüksek Lisans, Biyoenformatik Bölümü

Tez Yöneticisi: Dr. Öğr. Üyesi Aybar Can Acar

Ortak Tez Yöneticisi: Doç. Dr. Tunca Doğan

Ocak 2020, 101 sayfa

Yaşam bilimleri alanındaki yeni teknolojik gelişmeler sayesinde, üretilen moleküler sekans verisi miktarında son yıllarda büyük bir artış olmuştur. Bu verinin bilimsel literatüre fayda sağlayabilmesi için anlamlandırılması gerekmektedir. Geleneksel olarak bu anlamlandırma işlemi, deneyler ile üretilen moleküler verinin elle işlenmesi ve biyolojik veri tabanlarında saklanması suretiyle yapılır. Ancak bu verinin muazzam büyüklüklere ulaşması, otomatik ve sistematik analiz ihtiyacını doğurmuştur. Bu sistematik analizin önemli bir kısmını, farklı veri tabanlarından elde edilen öğelerin arasındaki ilişkilerin tanımlanması ile verinin entegre edilmesi oluşturmaktadır. Bu çalışmada negatif olmayan matris faktorizasyonu (non-negative matrix factorization – NMF) yöntemi ile büyük çaplı gen/protein verisini entegre edecek bir yaklaşım önerilmektedir. NMF ürün tavsiye sistemlerinde sıklıkla kullanılan ve başarılı uygulamaları olan bir yöntemdir. NMF ayrıca biyoenformatik ve kemoenformatik gibi alanlarında çoklu-ilişkili verinin birleştirilmesi için de uygulanmıştır. Bu çalışmanın amacı doğrultusunda, öncelikle moleküler işlev, biyolojik süreç, hücre-içi lokasyon ve hastalık ilişkileri gibi protein anotasyonları UniProt-GOA, DisGeNET gibi farklı kaynaklardan toplanmıştır ve bu veri ikili ilişki matrisleri olarak düzenlenmiştir. Sonrasında bu çok-boyutlu ilişkili biyomoleküler sekans anotasyon verisine (genler/proteinler ve işlevler, genler/proteinler ve hastalıklar, hastalıklar ve işlevler) çeşitli NMF tabanlı algoritmaları uygulanmıştır; ardından her modelin sonuçları ilişkili verideki esas yapıyı öğrenme yeteneği üzerinden çapraz doğrulama aracılığıyla değerlendirilmiştir. Sonuçlar, NMF'in bilinen protein anotasyonlarının çoğunu herhangi bir sekans veya yapı tabanlı protein özelliği kullanmadan ortaya çıkarma yeteneğine sahip olduğunu göstermiştir (AUROC: 0.80 – 0.94, doğruluk: 0.53 – 0.64,

F1-skoru: 0.06 – 0.40, MCC: 0.13 – 0.38). Bu çalışmanın nihai amacı, NMF'i kullanarak bu biyolojik varlıklar arasındaki bilinmeyen ikili ilişkileri tahmin etmektir. Devamında ise, bu varlıkları (proteinler, işlevler ve hastalık girdileri) faktorizasyon işlemiyle üretilmiş düşük boyutlu matrislerini kullanarak bilgilendirici ve artıksız niceliksel öznitelik vektörleri olarak ifade etmektir. Bu öznitelik vektörlerinin gelecekte proteinlerin otomatik anotasyonu veya biyolojik ağların oluşturulması gibi çeşitli veri madenciliği ve makine öğrenmesi uygulamalarında kullanılması hedeflenmektedir.

Anahtar kelimeler: negatif olmayan matris faktorizasyonu, çoklu ilişkili veri, biyolojik veri birleştirmesi, makine öğrenmesi, protein anotasyonları.

To My Family…

ACKNOWLEDGEMENTS

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

BP	Biological Process
CC	Cellular Component
FN	False Negative
FP	False Positive
FPR	False Positive Rate
GO	Gene Ontology
HNMF	Hybrid Nonnegative Matrix Factorization
ID	Identifier
KL	Kullback-Leibler
MCC	Matthews Correlation Coefficient
MF	Molecular Function
NMF	Nonnegative Matrix Factorization
NMTF	Nonnegative Matrix Tri-Factorization
PPI	Protein-Protein Interaction
TN	True Negative
TP	True Positive
TPR	True Positive Rate

CHAPTER 1

1. INTRODUCTION

1.1. Motivation

Biological data production has greatly increased thanks to the technological developments in recent years. Due to this increase, data characterization and analysis (traditionally, achieved by manual curation and the storage of experimentally produced data in biological databases) needs to be automated and systematized. A crucial part of this systematic analysis is data integration with the identification of relationships between the elements from different biological data types, such as functions of biomolecules, biological processes and their relationship with disease mechanisms. Automatic identification of these relations can be achieved via several computational approaches, such as network-based or module-assisted methods (Sharan et al., 2007). These automatically identified relationships may be a decent starting point to plan and to conduct targeted experiments, rather than trying to find a relationship by trial and error. Possessing prior information in this regard can be a huge advantage since an experimental procedures may long durations and high costs. In this thesis, we aimed to address this problem and suggested a computational approach for automated integration of biomolecular data.

1.2.   Biological Definitions

1.2.1. Genes, Proteins and Functions

Genes can be identified as information storage units of the organisms, holding all the data of the organism (Lewis, 2005). Genes are responsible for various activities; some are involved in protein encoding while some others take role as regulatory units and so on. Their building blocks are DNA molecules that form the nucleotide sequences. Genes that are responsible for protein encoding does not directly produce proteins. Instead, there exists a flow called central dogma, which contains the procedures replication, transcription, and translation. Replication is the creation of a copy of a gene of interest, while transcription is the production of mRNA molecules from that gene. The mRNA molecules produced by the transcription process is involved in production of amino acid sequences (proteins). The process of protein production from the mRNA is called translation. Amino acid sequences produced in translation fold in favorable positions to have a certain 3-dimensional shape. This shape, the structure of the protein, determines its function (Doolittle, 1985). Proteins are essential molecules

in biological systems, and they not only contribute to the structure of the cell, but also involve in almost all of cell's, and though, the organism's dynamic processes. Some of these molecular processes include enzymic activity, transporting of molecules, message carrying as well as acting as antibodies, hormones, toxins and many more (Alberts et al., 2008).

1.2.2. *Diseases*

A disease is the abnormality in the structural and functional condition of an organism (William Burrows & Dante G. Scarpelli, 2019). Diseases can be caused by several factors. Genetic diseases are caused by a change in DNA sequence from its normal state (*Genetic Disorders | NHGRI*, n.d.). This type of diseases may be resulted from mutations in one gene or in combination of genes. Mutations are the changes in the nucleotide sequence of the genes that can or cannot be resulted in functional changes (Ripley, 2013). The mutations can affect a part of the genome or can occur in only one nucleotide, which is then called point mutations. Depending on where in the gene the mutation occurred, and what kind of change it procures, mutations are named differently. If the point mutation substitutes an amino acid in the original protein sequence, it is called a missense mutation. While a mutation which is not changing the wild-type protein sequence is called a silent mutation. A point mutation can sometimes lead to premature termination of the production of amino acid sequence. This type of mutation is named as a nonsense mutation (William S. Klug, Michael R. Cummings, 2006). Mutations can result in a functional change or the entire loss of function in the protein depending on where in the gene sequence it occurs. These functional changes in the protein may result in diseases and disorders varying in type and severity.

1.3. Biological Databases

The biological entities of discovered proteins, genes, functions, diseases etc. are stored in biological databases as entries. A biological ontology is a way to store this information using standard vocabulary. An annotation is the storage of relations between biomolecules and relevant function defining ontological terms, in biological databases. This does not only provide a more systematic approach, but also provides organization for the data. There are several biological ontologies available for different biological data types, such as gene ontology, disease ontology and human phenotype ontology. Furthermore, these ontologies are connected to each other via cross-ontology mappings.

In the following sub-sections of this section, the data used in the study are explained along with the databases that provide this data.

1.3.1. Protein Databases

The most well-known resource for proteins is the UniProt (The Universal Protein Resource) (UniProt Consortium, 2018), in which, protein entries include information related to amino acid sequences, functions, locations, the genes they are encoded from, interaction with other proteins and so on. UniProt is composed of two main databases. The first one includes manually curated and reviewed data, named UniProtKB/Swiss-Prot, and the second one includes data obtained via electronical annotations that are yet to be reviewed, which is named UniProtKB/TrEMBL. Another database storing protein related information is PDB (Protein Data Bank), which is the archive of 3-D structures of these biomolecules (Berman et al., 2002). In PDB, the structure information is stored with the related sequence and ligand information, which are small molecules that bind to the protein of interest (Gordon & Perugini, 2016).

1.3.2. Gene Ontology

Functions of genes/proteins are stored as Gene Ontology based annotations (Ashburner et al., 2000). The functions are categorized in three main branches; cellular component (CC), molecular function (MF) and biological process (BP). Cellular component indicates where in the cell the protein performs its mission, molecular function is the information about what specific molecular job the protein does in the cell, while biological process defines large scale mechanisms such as the oxygenation of tissues.

There exists a hierarchical system among the GO terms, where higher (more generic) terms are called parent terms and the lower (more specific) ones are called child terms (Hennig et al., 2003). For example, the GO term "molecular function" is one of the three most generic terms (a parent term), while "protein binding" is one of the child of terms of "molecular function". The term "protein binding" has its own children terms as well.

1.3.3. Disease Ontologies

Disease databases hold information regarding which genes the genetic disorders are related to each other, where and on which function on the body it is observed, and so forth. There are several disease databases, such as OMIM, Orphanet and DECIPHER. OMIM (Online Mendelian Inheritance in Man) is a catalog of human genetic disorders, focusing on the molecular relationship between genetic variation and phenotypic expression (Hamosh et al., 2005). Orphanet, on the other hand, is the portal that collects and stores the information about rare diseases (Montani et al., 2013). There also exist databases that collect information from multiple disease resources, such as DisGeNET. DisGeNET is a platform containing publicly available collections of genes and human diseases, obtained from various databases (Piñero et al., 2019, 2017, 2015).

This database also integrates data from other resources, such as the expert curated repositories, animal models, scientific literature and so on. DisGeNET includes gene-disease associations from UniProt, ClinGen, CTD and such, as well as variant-disease associations obtained from ClinVar, GWAS Catalog, GWAS database and so on.

1.4. Nonnegative Matrix Factorization

Matrix factorization is the decomposition of a two-dimensional matrix using the properties of matrix algebra, where the matrix contains different entity types at its columns and rows, and the relationship between these entities are expressed with values inside the corresponding cells (Cai et al., 2008). Matrix factorization is a frequently used method for recommender systems, which learns from the users' previous interactions to recommend new items that the user might be interested (Koren et al., 2009). The method is successfully applied in real-world cases. It has also been employed for uniting multi-relational data in various fields.

1.4.1. Matrix

A matrix is a two-dimensional array with size *m x n*, where m represents the quantity of rows of the matrix, while n represents the quantity of columns. Matrices can be fully or partially filled with real or complex numbers, and they are used to solve various mathematical problems (Cherney et al., 2013). Algebraic operations can be applied on matrices, provided that they follow particular rules. For instance, if the matrices are the same size they can be added or subtracted; or if the inner dimensions are the same size they can be multiplied.

1.4.2. NMF Algorithm

Nonnegative matrix factorization (NMF) is based on matrix multiplication. The principle is to obtain two low-rank matrices by factorizing the main matrix, where the large matrix is recovered when the low-rank matrices are multiplied (Lee & Seung, 2001). This method saves computational space in storing information since matrices takes up exponentially more space as they get bigger, as well as the computation time, while operating with the data stored in the matrix as smaller matrices are easier to work with. As for the recommendation systems' aspect, the low-rank matrices are constructed by discovering the similarities between entities of the original matrix (called latent factors) by the factorization algorithm, and thus learning the intrinsic properties, which is then used for predicting the unknown values in the matrix. In **Figure 1.1.** a simple representation of latent factors is visualized through movie genres and user gender. Here, the latent factor number is arbitrarily selected as two, at the input level; in other words, the algorithm is instructed to distribute movies and users along 2 axis, according to their similarities. When the results are examined, it is

observed that the algorithm had implicitly grouped the movies into two main genres of serious and escapist; and the users according to their genders (Koren et al., 2009). As the latent factor values get higher, the model becomes more complex and what the groups are constructed upon becomes more and more incomprehensible by examining alone. On the other hand, very low values of latent factors are usually not sufficient to successfully express the data, and thus, produce random results. Therefore, there is a trade off in between and the correct selection of the number of latent factors is critical.
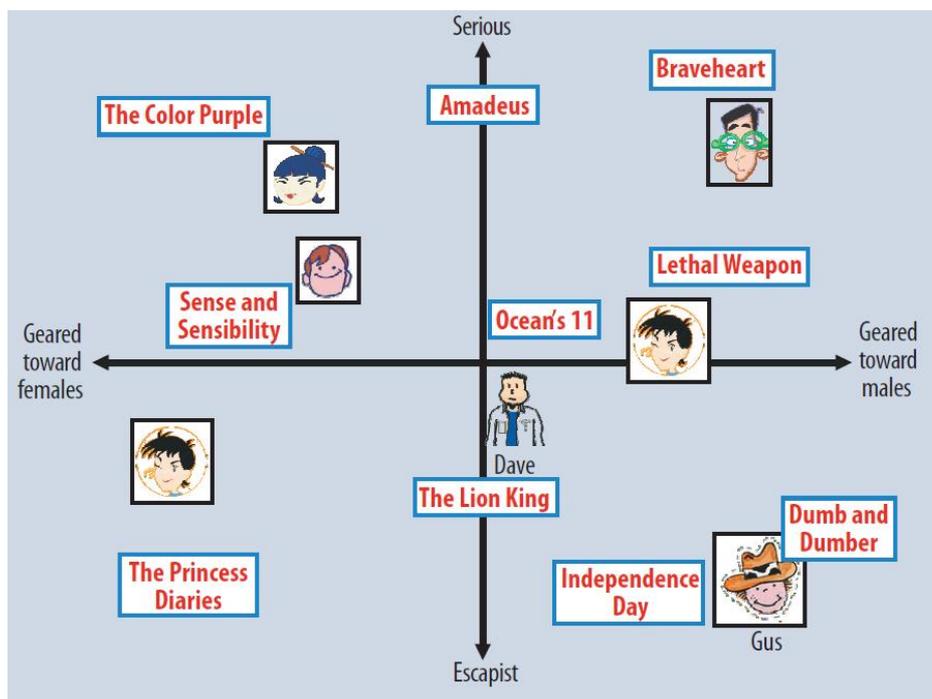
Figure 1.1. An example of simple latent factor inductions of the algorithm. (Koren et al., 2009).

1.5. Aim of the Thesis

The aim of this thesis is, first, to gather and to integrate the available large-scale biomolecular relation data. Second, to predict the relationships of genes/proteins with biological processes, molecular functions, cellular locations, and with genetic diseases, using matrix factorization.

Our hypothesis in this study was that, it can be possible to estimate the unknown relations between proteins, functional aspects and diseases, using only their previously known relations. In literature, protein function annotation studies mostly employ sequence information while constructing machine learning-based prediction tools. Here, our expectation was that, the model would learn the intrinsic structure of the data not only via known explicit protein to function/disease relations, but also using the

implicit relations between functions or between diseases. We expect that, the relation predictions produced by our approach will be complementary to the conventional sequence/structure feature based protein function annotation methods.

1.6. Outline of the Thesis

In chapter 2, the literature review is presented. At first, previous studies on recommender systems are explained, and then the articles that used the matrix factorization approach are reviewed, as well as the studies focusing on the integration of multiple data types.

In chapter 3, the methodology is described in detail. First, the algorithm and the mathematical definitions behind the algorithm is explained, followed by the data types and their retrieval from different databases. Lastly, we get down to the details of how the chosen algorithms are applied to our data.

In chapter 4, the results and performance values that were retrieved from different algorithmic applications are presented. It displays the change in performance values, as the applied method is upgraded from the baseline NMF to more complex algorithms.

In chapter 5, the discussion and the conclusion of study is given. Finally, potential future works are explained.

CHAPTER 2

2. LITERATURE REVIEW

2.1. Recommender Systems and Matrix Factorization

Nonnegative matrix factorization is firstly applied in 1999 to learn different parts and features of faces and to detect semantic features of the texts. The study had brought new aspect to the problem since other algorithms like principal components and vector quantization had been learning the data not via parts of the objects (Lee & Seung, 1999). The method then became a popular approach for recommender systems, starting from the Netflix Prize in 2006. Netflix is a streaming service for watching movies, TV shows, documentaries and other visual production (Adhikari et al., 2012). Netflix Prize was a competition that the company itself started by sharing a sample data of user-product rating as training set to the participants and expecting them to come up with expected ratings the users would give to other movies etc., i.e. rating predictions. The company calculated the performance improvements via not-shared data, and promised to give the big prize to the first participating team whose work would do at least 10% better than the company's existing systems, and a smaller money prize for the leading team that would not be able to reach the threshold (Koren et al., 2009). The winning team was announced as BellKor's Pragmatic Chaos with 10.06% improvement to the company's own algorithm in 2009. After the popularity the method had gained from the competition, it has been applied to other fields as well, like life-sciences.

2.2. Applications of Matrix Factorization

As matrix factorization has started to be used in more and more fields, the researchers wanted to produce more accurate and abundant predictions via adding more data to the algorithm so that it could learn relations from other data types to predict binary relations. Lippert et al.(2008) has studied on such approach to predict movie ratings and gene functions of yeast. In movie-rating prediction application, they added user's rating to the movies as well as information about users (gender, age and occupation) and movie entries (the movies are separated to 20 different genres). For the gene function prediction, different types of data such as gene, function, chromosome, phenotype, motif are inserted to the algorithm to obtain predictions. The algorithm they proposed is then compared to the regular matrix factorization and singular value decomposition algorithms.

Another study focuses on nonnegative matrix factorization with multiple data types for movie rating prediction (Singh & Gordon, 2008). In this study of collective matrix

factorization, they factorized all data matrices, and used the shared entities between matrices to predict new relationships between the elements of matrices. They also introduced weight matrix for the user-movie rating matrix, to indicate whether the rating was present in the initial data matrix. If the rating was present, the value in the weight matrix was 1, and 0 if the rating was not available in initial data. By using of the weight matrix, they made sure that the initially unavailable ratings had not been used while predicting new ratings, and in general, predicting new relationships between the entities of different data types. The relations used in the study (other than movie ratings of users and the weight matrix) were genres for each movie, and a list of actors in each movie. They showed that using relations improved predictions compared to the matrix factorization methods that use a single matrix.

There exists another study that collects all matrices in a one large matrix (collective matrix factorization) instead of using all the data matrices separately via Singular Value Thresholding algorithm (Bouchard et al., 2013). They put the matrices together after computing their nuclear norms, which can also be represented by decomposition norm, based on same mentality of matrix factorization. An example for the collection of all matrices as one can be seen in **Figure 2.1.** In the figure, the relation diagram of three data types are given, in addition to the placement of their norm matrices. The matrices are placed in a way so that none of the relations represented in the diagram is missed. They applied this method in two types of real-life data: MovieLens and Flickr, the first being a movie rating database containing users' ratings, their demographic information and the movies descriptors; and the other being a social photo bookmarking site, of which the data having user-user interaction, user-tags, item-item feature etc.
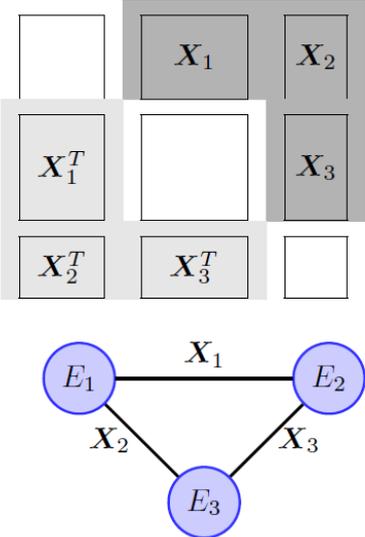
Figure 2.1. A symmetric block matrix representation of collective matrix factorization and the entity relations (Bouchard et al., 2013).

Hybrid nonnegative matrix factorization (HNMF) algorithm was presented by Luo et al. with the purpose of integrating phenotypic and genotypic features of hypertension patients for clustering and therefore future predictions (Luo et al., 2019). Their difference from the previous methods is that they use specialized loss functions for different data types, with purpose of obtaining better results in the end. This algorithm is applied for three data types (patients, genotype and phenotype features) with two relations, these being patient-genotype and patient-phenotype. The low-rank matrices of the data types are constructed and then improved by benefitting from binary relations between other data entities. This algorithm is used in our study as well, so its mathematical approach and the application with the data will be explained in Chapter 3 in more detail.

Another differentiation of the NMF method is an algorithm based on Nonnegative Matrix Tri-Factorization (NMTF). Their purpose of using this method is drug repositioning by using different features of drugs and targets as data types (Ceddia et al., 2019; Dissez et al., 2019). The difference of this method is that they can determine different numbers of latent factors for each of the data types, instead of the traditional approach of using same number of latent factors for both data types in a particular relation matrix. To achieve this the algorithm adds a third matrix that connects the low-rank matrices while factorizing. The study uses intra-type relations for some of the data types as well with the relations with different entries. Another contribution they added to the method is how to start the latent factor matrices of the data types. Traditionally, and in most of the cases mentioned before, the low-rank matrices are started with random values and then are updated in each iteration according to the update rules. On the other hand, they compare four different starting methods and use the one with the most efficient result based on the claim that random starting makes the results inconsistent at each application. To further improve the results of the algorithm, intra-type relations are also added for some of the data types, claiming that the proteins, pathways that interact with each other etc. tend to be classified together and show similar biological characteristics. This algorithm is also used in this thesis as the last upgrade to the baseline nonnegative matrix factorization algorithm, so its principles etc. will be explained again in Chapter 3 in more detail.

2.3. Matrix Factorization Applications on Biological Data

Matrix factorization and its derivatives are used in multiple fields of sciences, including computational biology. The scientific problems in this field include molecular pattern discovery, as in protein and gene microarray studies and expression profiles, cross-platform and cross-species analyses, function-gene relations, drug-target interactions and so on (Devarajan, 2008). Pehkonen et al. used nonnegative matrix factorization to identify and visualize the clusters of genes via their functional classes (Pehkonen et al., 2005). They obtained various grouping results for different numbers of clusters, in other words, different numbers of latent factors. They shared these various clusters they obtained with their developed tool called GENERATOR and the differentiation of clusters as the clustering number changes. Additionally, they reported the comparison of their tool and other computational tools to show the performance of their algorithm.

Another popular objective is drug-target interactions. Since the drug prediction experiments take years, cost huge amounts of money and have no guarantee to be successful in the end of the experiment, the knowledge in the field is limited. There are databases for drug molecules that are experimentally proven such as ChEMBL (Gaulton et al., 2012), DrugBank (Wishart, 2012) and so on. The scarcity of information due to previously explained reasons encourages the scientists of the field to find methods for automatic and systematic predictions. One example that matrix factorization is used for drug-target prediction is performed by Cobanoglu et al. (2013). They used Probabilistic Matrix Factorization to find additional interactions between drugs and their target molecules (proteins etc.) given that they give all the drugs and target molecules of interest and some of the binary interactions. Another necessity of the method is that the number of initially known interactions should be high enough for the method to outperform the ones available in literature. The visualization of the method for drug-target prediction is shown in **Figure 2.2.** In the figure, the initially available interactions between drugs and target molecules are represented by black lines, while new predictions of interactions are shown with red lines, along with the probability of that association. Initial interactions are used to construct the latent vectors of different entities, and then the vectors of drug and target entities are multiplied to obtain the likelihood of that entity pair in association with each other.
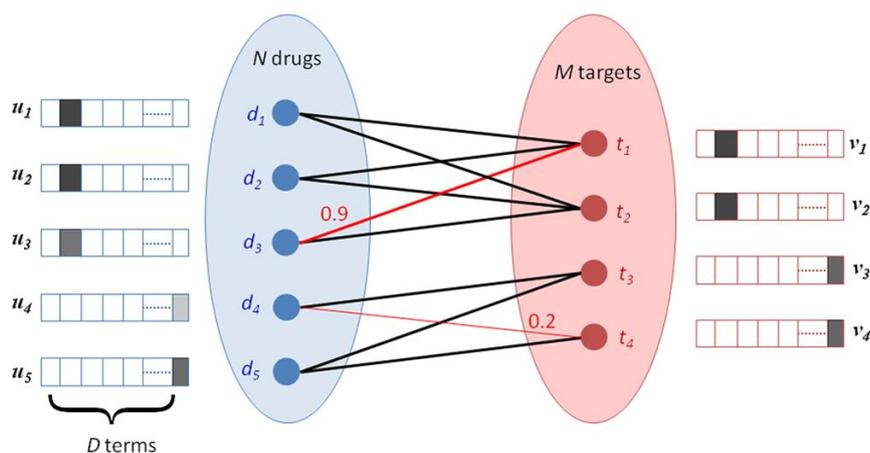
Figure 2.2. Probabilistic matrix factorization for drug-target interaction. (Cobanoglu et al., 2013).

Another study that focuses on integration of multiple datasets introduces intra-type relations as well to the algorithm so that the similarities between entities of same data type can also be considered for relation predictions (Žitnik & Zupan, 2015). They use this algorithm on two experiments: gene function prediction and drug-target prediction. For gene function prediction, they used genes, ontology terms (function annotations), experimental conditions, publications, descriptors, and pathways as data types; and also added scores of interaction and similarity for data types such as genes and functions. The other case was for pharmacological (drug-target) prediction, in which they used chemicals, pharmacologic actions, publications, depositors and their categorization, and substructure fingerprints, with chemicals having similarity scores as intra-type information.

Drug-target interaction prediction studies are very popular due to above-mentioned reasons. Each study tries to achieve more accurate results by adding different types of extra information to the algorithm. One such study uses an algorithm called Collaborative Matrix Factorization (Zheng et al., 2013). The method is based on the approach of adding similarity matrices representing different properties of main data types (drugs and targets) to improve the drug-target prediction. They add chemical structure and ATC (Anatomical Therapeutic Chemical) (Skrbo et al., 2004) similarities for chemical drugs; and genomic sequence, gene ontology (GO) and protein-protein interaction (PPI) network similarity matrices for targets in addition to the relation matrix of drugs and target molecules.

Last but not least, there is another study which is similar to the one above by the data they used, but separated in some aspects by the factorization method. They study used Bayesian matrix factorization to predict drug-target interactions from chemical and genomic kernels (Gönen, 2012). They separated the drug-target interactions for humans into four groups, which were enzymes, ion channels, G-protein-coupled receptors and nuclear receptors. Then they inserted these groups of interactions to Bayesian matrix factorization for predictions. They used only the chemical similarities between drugs and the genomic similarities of proteins (targets). Their method of Bayesian matrix factorization was claimed to be combining kernel-based dimensionality reduction, matrix factorization, and binary classification (for prediction).

All the methods mentioned above either focus on the binary relations completely independent from each other (i.e., no multi-type data integration), or they use biological data types and relations that are different than ours, while integrating multiple relations. The main contribution of our study is integrating the protein-function relations (with functions of all three main GO categories) and the protein(gene)-disease relations, together with intra-type relationships between protein entities (protein–protein interactions -PPI-), to predict the unknown relations between these input entities. Another important contribution of this study is testing the idea that inserting more relational data to the model would improve the prediction performance.

CHAPTER 3

3. MATERIALS AND METHODS

3.1. Overview of the Method

The pipeline of the study can be observed in **Figure 3.1**. At the beginning of this work, the binary relation datasets protein and functional aspects were downloaded from the UniProt–GOA database, while the relationships between proteins and diseases were retrieved from the DisGeNET database. After the conversion of the data into relation matrices of Protein x CC, Protein x MF, Protein x BP and Protein x Disease, the matrices were inserted to the chosen algorithms. First, the simple nonnegative matrix factorization algorithm was used to obtain relation predictions, without any integration. Then, to observe the improvements of the multiple relation data integration, the HNMF method was used, in which, pairs of relation matrices are inserted as input to the algorithm. Lastly, the NMTF algorithm was used, where it was possible to insert all of the data matrices at once, that being the main purpose of this study. Finally, the performance evaluations were performed for each algorithm and compared to each other for a thorough discussion. Every methodological step mentioned above is explained in detail in following sub-sections of this chapter.
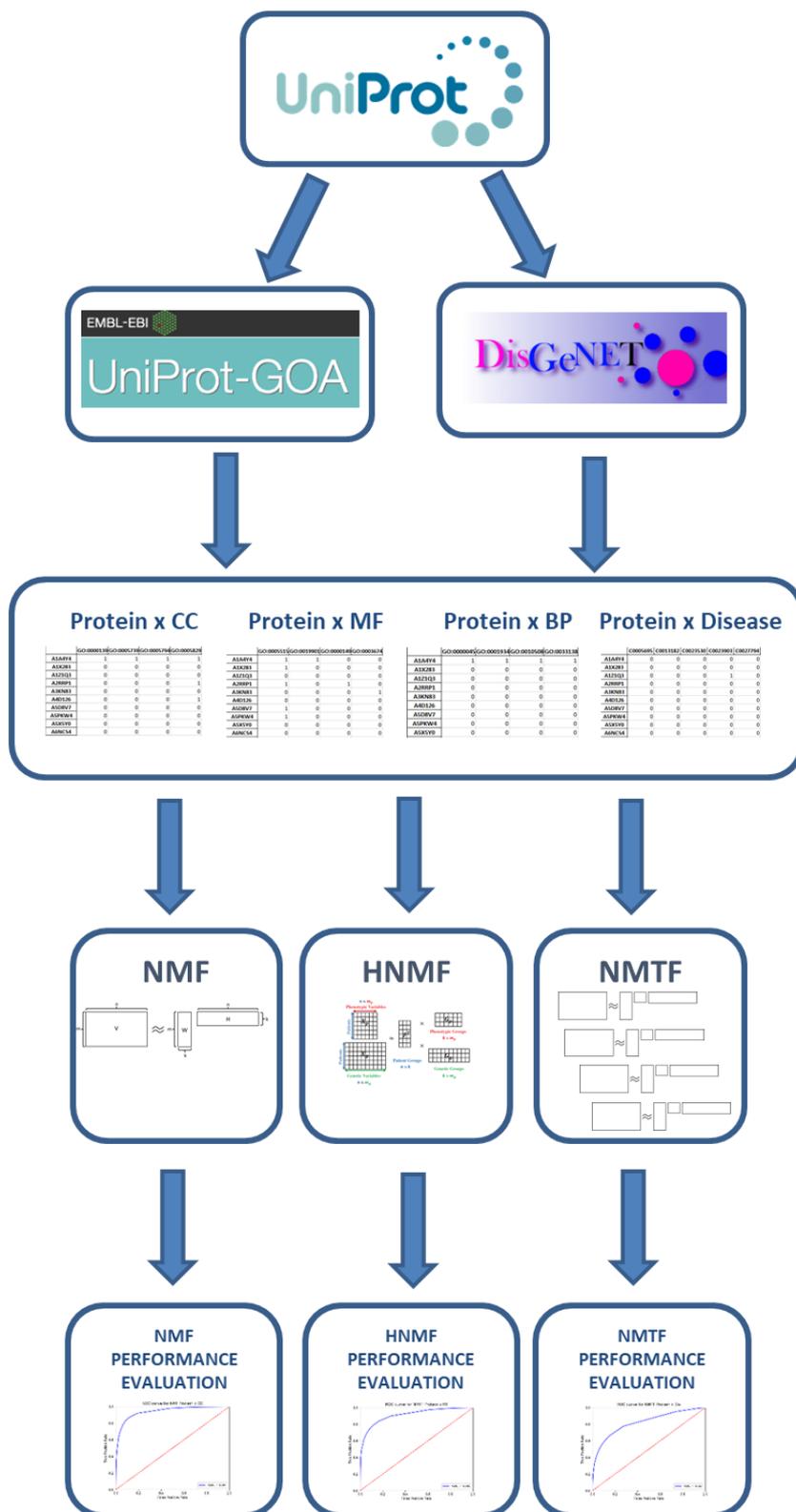
Figure 3.1. The pipeline of the methodology.

3.2. Datasets

3.2.1. Retrieval of Protein GO Annotation Datasets

As explained earlier in this thesis, GO (Gene Ontology) annotations are consisted of three main branches; cellular component (CC), molecular functions (MF) and biological process (BP). All protein – GO term binary relationships were exported from GO Annotation Database of EMBL-EBI (Huntley et al., 2015). The fields included in the data file can be seen in **Table 3.1**.

Table 3.1. The columns in GO annotation data file.

Column	Content	Example
1	DB	UniProtKB
2	DB Object ID	P12345
3	DB Object Symbol	PHO3
4	Qualifier	NOT
5	GO ID	GO:0003993
6	DB:Reference (|DB:Reference)	PMID:2676709
7	Evidence Code	IMP
8	With (or) From	GO:0000346
9	Aspect	F
10	DB Object Name	Toll-like receptor 4
11	DB Object Synonym (|Synonym)	hToll
12	DB Object Type	protein
13	Taxon(|taxon)	taxon:9606
14	Date	20090118
15	Assigned By	SGD
16	Annotation Extension	part_of(CL:0000576)
17	Gene Product Form ID	UniProtKB:P12345-2

First filter applied to the data was for taxonomy; the taxon of interest was human in the study. The next filter was exclusion of electronically curated relations by other prediction algorithms, to work on experimentally proven and curated annotations only. For this, the annotations coded with IEA (which indicates that the curation is done via prediction algorithms) as evidence code are excluded from the data. Afterwards, the relations are separated according to their GO term types, from which respective data matrices are constructed in the later steps. Then the unnecessary columns in the data were deleted, as in removing all columns except protein and GO IDs. From the remaining list of binary relationships, the repeating rows were removed in order to have only the unique pairs in the end.

As the next step, the parent GO terms were added to the relationship list, to improve the performance of algorithm; since it was important for the algorithm to catch the common features between entries. The data file downloaded from the UniProt - GOA database of UniProt (Binns et al., 2009) had been containing only the most specific terms, meaning the terms in the downloaded data had no child terms.

15

3.2.2.   Retrieval of Protein – Disease Relation Dataset

The protein – disease relation data was exported from DisGeNET website. The data downloaded from this database was already filtered as experimentally curated. Unlike the GO Annotation database, the data in this platform contains gene IDs as one of the columns instead of proteins. So, they were converted to protein UniProtKB ID for consistency among relation data. Like in protein – GO terms data, the unnecessary columns and then the repeating rows were removed.

The next procedure was to take the protein IDs that were present in binary relationships with all three GO categories and disease entries, as the number of the entities of the same group used in the algorithms should be the same. Taking the common proteins was necessary for not only a meaningful comparison among results of different applications, but also for the NMF applications using multiple relation matrices to work. As clarification; the HNMF and NMTF algorithms would not work unless the number of proteins in the inserted data matrices is the same. Afterwards, again for performance-related reasons we applied a filter such that only the GO terms that are associated with at least 50 proteins and disease terms that form binary relationship at least with 30 proteins were remained. Lastly, the data was converted to matrix form, wherein the cell that corresponds to that particular entity binary was filled as 1 if the protein – GO (or disease) pair is present in the data (i.e. the binary relationship between that protein and GO/disease term), and 0 otherwise. In the end, four different relation data matrices were constructed, which are Protein x CC, Protein x MF, Protein x BP and Protein x Disease relations. The size of the data matrices before and after the common proteins are taken is shown in **Table 3.2.** and **Table 3.3**, respectively.

Table 3.2. The number of each entries and the binary relation number before common proteins in all matrices were taken.

	Binary relation number	Protein number	GO/disease number
Protein x CC	117,163	16,345	364
Protein x MF	72,425	15,583	446
Protein x BP	137,444	15,043	1679
Protein x Disease	83,446	7227	3670

Table 3.3. The size of the data matrices after all filters were applied and common proteins were taken.

	Binary relation number	Protein number	GO/disease number
Protein x CC	52,203	5424	280
Protein x MF	34,270	5424	274
Protein x BP	75,880	5424	987
Protein x Disease	71,376	5424	930

The histogram graph of proteins for each matrix can be observed in **Figure 3.2**. The graphs show the occurrence numbers of the proteins; for example, how many of the

proteins has a binary relationship with a GO term only once, how many of them occurs two-times and so on.

a.
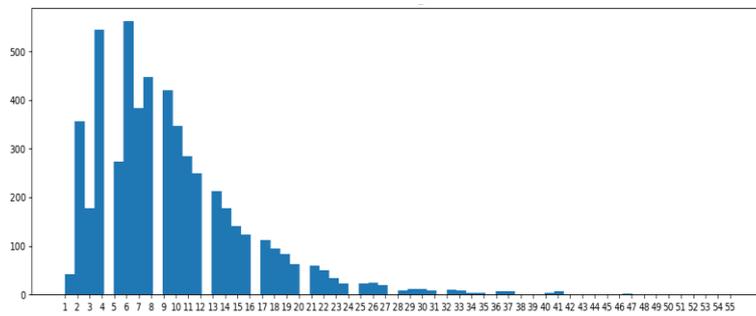
b.
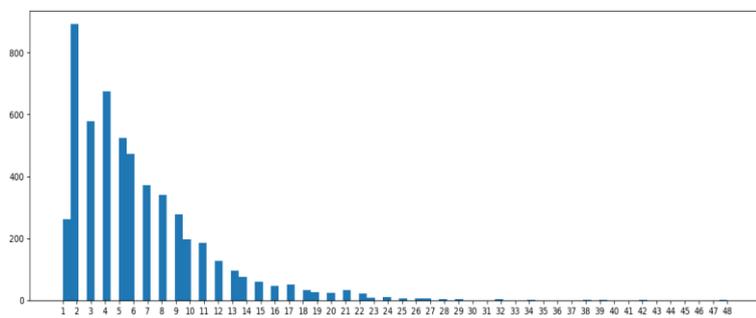
c.
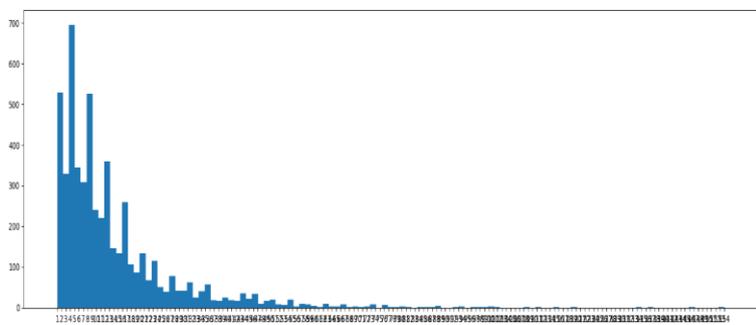
d.
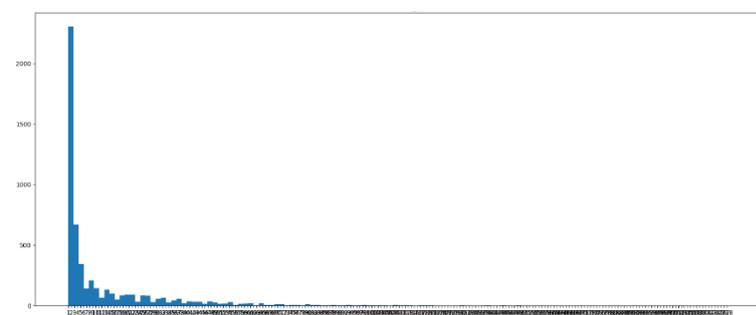
Figure 3.2. The histogram of matrices with respect to proteins. a. Protein x CC, b. Protein x MF, c. Protein x BP, and d. Protein x Disease matrices.

As can be seen in the histogram figures, there were proteins present only once in the binary relationship list, so they had not shared any common relations with other entities. To explain in more detail, these proteins that are present only once would not contribute to the algorithm since it learns from the shared relationships between entities. So, as an additional NMF analysis to compare the effect of data structure, these proteins were also excluded from the data to observe whether it would boost the performance of the algorithms. Afterwards, again the common proteins (proteins present in all 4 matrices) were taken. The final size of the matrices after this procedure can be seen in **Table 3.4.**

Table 3.4. The size of the matrices after the proteins with one binary relationship were removed.

	Binary relation number	Protein number	GO/disease number
Protein x CC	37,300	3575	279
Protein x MF	25,444	3575	273
Protein x BP	59,221	3575	987
Protein x Disease	65,785	3575	930

A representation of the constructed matrices can be seen below in **Table 3.5**. In the matrices, the row elements are protein entries, and the columns are GO / disease terms. The cells were filled as 1 if there is a binary relation between the protein and GO/disease terms, and filled with 0 if there was no relation between the entities.

Table 3.5. Representation of an example data matrix for Protein x MF.

	GO:0005515	GO:0003674	GO:0140096	GO:0016787	GO:0035091	…
A1A4Y4	1	0	0	0	0	
A1X283	1	0	0	0	1	
A3KN83	0	1	0	0	0	
A6NNW6	0	1	0	0	0	
P00797	1	0	1	1	0	
…						

3.3. Mathematical Definitions

In this section of the thesis, some of the mathematical terms are explained as an introduction before the baseline algorithm and the upgraded versions are explained in detail.

3.3.1. Matrix and Basic Operations

A matrix is a rectangular array with size $r \times k$ (i.e. with r number of rows and k number of columns), containing real or complex numbers, and each number on the matrix is called entries (Cherney et al., 2013). Various operations can be applied to a matrix as well as among multiple matrices. The transpose of a matrix of r x k size is a matrix of

k x r size, in which the rows and columns of the matrix are replaced with each other. Matrices can be added and subtracted if their sizes are equal. Matrix multiplication is an operation where one matrix is obtained from two matrices. For two matrices to be multiplied, the inner sizes of the matrices when put side by side should be equal. For clarification, imagine two matrices, first with size of *m x j* and the second with size of *k x n*. For these two matrices to be multiplied, *j* should be equal to *k* (*j = k*). The result of this multiplication is an *m x n* matrix.

$$
\begin{bmatrix} a_{11} & \cdots & a_{1j} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mj} \end{bmatrix} \cdot \begin{bmatrix} b_{11} & \cdots & b_{1n} \\ \vdots & \ddots & \vdots \\ b_{k1} & \cdots & b_{kn} \end{bmatrix} = \begin{bmatrix} c_{11} & \cdots & c_{1n} \\ \vdots & \ddots & \vdots \\ c_{m1} & \cdots & c_{mn} \end{bmatrix} \tag{1}
$$

where a_{11}…are elements of matrix A of size *m x j*, b_{11}… are elements of matrix B of size *k x n*, and c_{11}… are elements of matrix C of size *m x n*.

3.3.2. Likelihood Function

The definition of likelihood function is the joint probability of observing $x_1...x_n$ given that the parameter is Θ (Liu & Jiang, 2013). The mathematical notation of the function is:

$$
L\left(\Theta \mid x_1, x_2, \ldots, x_n\right) = p(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} p(x_i \mid \Theta) \tag{2}
$$

where L is the likelihood and p( ) is the probability.

Because the likelihood function gets harder to solve as n becomes larger, a workaround is used as:

$$
\log L\left(\Theta \mid x_1, x_2, \ldots, x_n\right) = \log \sum_{i=1}^{n} p(x_i \mid \Theta) \tag{3}
$$

where log is the logarithmic function.

The purpose of the function is to find the most likely parameters given the observations, thus making this an optimization problem. Since we try to find the maximum similarity, there comes the maximum likelihood estimation (MLE), whose mathematical formulation is:

$$
\theta_{MLE} = \arg\max_{\Theta} \log L\left(\Theta \mid x_1, x_2, \ldots, x_n\right) \tag{4}
$$

3.4. Nonnegative Matrix Factorization Algorithm

Nonnegative matrix factorization (NMF) is an algorithm based on the idea that a matrix V with size *m x n* is factorized into two low-rank feature matrices W (*m x k*) and H (*k x n*) such a way that when they are multiplied the original matrix is regained (Lee & Seung, 1999).

$$V \approx W \, x \, H \tag{5}$$

Below in **Figure 3.3.** a schematic representation of matrix factorization is also available.
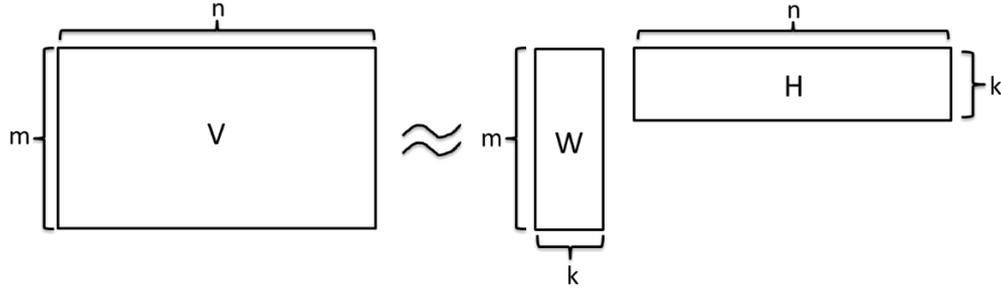
Figure 3.3. Schematic figure of matrix factorization.

k represents the latent factor (the cluster value of the data types) that is determined in the beginning, and the low rank matrices are computed accordingly. Finding the best k value is important since there is a trade-off between approximation and complexity for the model, where greater k values provide better approximations (because the generalization and thus data loss degree is smaller) while smaller k gives less complex model for computation.

The algorithm tries to find the optimum low-rank matrices that would give the closest result to the original data matrix; in other words, it tries to make the difference between original data matrix V and the multiplication of the low-rank matrices W and H minimum. So;

$$\arg\min_{A,B} = \, || \, V - W \cdot H \, ||_F^2 \tag{6}$$

is taken, which is called loss function, where the loss between the data matrix and the predicted matrix is calculated. In the equation $|| \; ||_F^2$ notation is the Frobenius norm. Frobenius norm of a matrix is the square root of the sum of the absolute squares of its elements (entries, cells) (Golub & Loan, 1996):

$$||A||_F = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{m} |a_{ij}|^2} \tag{7}$$

where A is a m x n sized matrix, and a_{ij} is the element of A in i-th row and j-th column.

In order to find the optimum solution, the low rank matrices are updated and compared to the original matrix. Traditionally, they are started with random values and then updated in each iteration until the best matrices to minimize the difference is reached. The stopping criterion can be either a very small value of loss that will be the result of loss function, or a maximum number of iterations to be reached. The update methods (i.e. the learning method of the algorithm) are generally improved from two basic algorithm; stochastic gradient descent and alternating least squares (Koren et al., 2009). In stochastic gradient descent every element of predicted matrix is compared to the data matrix and updated by a given learning rate in every iteration, (Bottou, 2012). Alternating least squares method for the update of low-rank matrices is preferred for some since stochastic gradient descent may cost too much computational time. The update method simplifies the optimization problem by taking each low-rank matrix as constant to optimize the other one, and continues the procedure until convergence (Stanford, 2015).

In this study, first of all, the baseline NMF algorithm was applied, where only one matrix was introduced to the method and the factorization results for each matrix was independent from the others. For this, the nnmf function of the MATLAB (R2018b) software was used. Since there were multiple parameters for the algorithm that can change the result, the algorithm was first run for a small toy matrix of 15x10 to find the best parameter combinations for the iteration number, update rule and so on. Afterwards, different latent factor (k) values were tested to find out the best option for each relation data matrix. At last, the algorithm's performance was evaluated with 10-fold cross-validation. The results obtained from this algorithm were to be compared to the results from algorithms using multiple-relation data matrices, to observe the performance improvement from using the multiple relation matrices simultaneously.

3.4.1. HNMF Algorithm

HNMF algorithm is an upgraded version of the baseline NMF algorithm., where the approach in the study was to maximize the addition of likelihoods (joint likelihood) of two approximations to achieve introducing two relations to the base model (Luo et al., 2019). They used discrete values for genotypic data (counts of genetic variants) and continuous values for phenotypic data. For this reason, they used Kullback-Leibler (KL) divergence for the genotype part of the likelihood function, and Frobenius norm for phenotype part of the function. They have also added a parameter lambda to the joint likelihood function so that the trade-off between the loss functions can be entered. As the update rule of low-rank matrices (of patients, phenotype values and genotype values) they adapted the alternating projected gradient descent method, and the stopping criteria as either small enough joint loss or convergence in difference between loss of last two iterations is reached.

Below in **Figure 3.4** the HNMF model in the original study can be observed according to their own data types. X_p and X_g represents their relation matrices of Patients x Phenotypic variables and Patients x Genetic Variables, respectively. F is the low-rank latent groups matrix for patients while G_p and G_g are the low rank matrices for phenotypic and genotypic groups. In this study the model was adapted to the data here as the X_g matrix is the Protein x Disease matrix and the X_p matrix is either Protein x CC, Protein x MF or Protein x BP in each application.
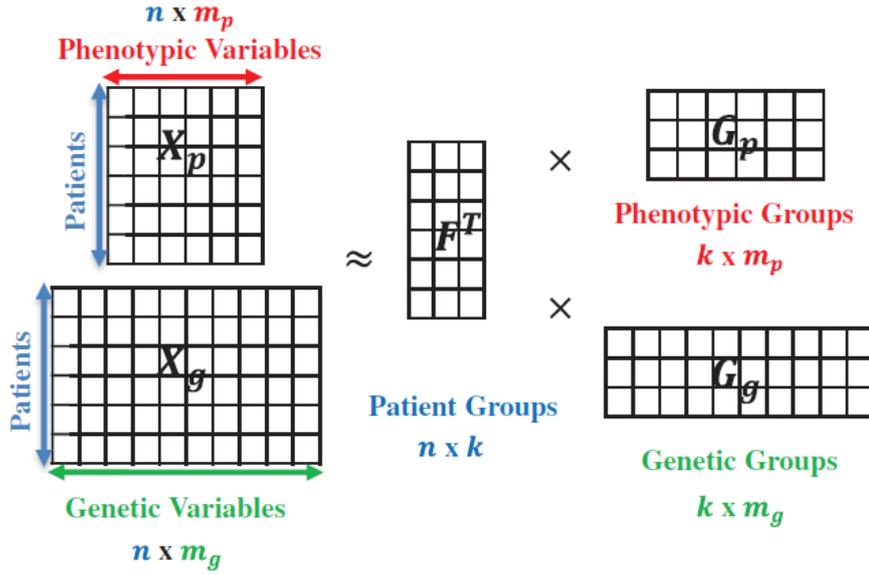
Figure 3.4. A hybrid nonnegative matrix factorization model (Luo et al., 2019).

For adaptation to data in this study, the joint likelihood function is converted so that both parts are suitable for KL-loss since all data is of categorical values (either 1 or 0); likewise, the update rules are converted accordingly. The study has shared their MATLAB code as an open source, thus giving opportunity to update the equations to fit this study.

Parameters, functions etc. in the equations given below are explained where they are firstly encountered and are not defined again as they reappear.

The minimized negative log likelihood function in the matrix form of original study:

$$L\,(F,\,G_p, G_g) \;=\; \sum_{ij} [\frac{\lambda}{2}\,||X_p \,-\, \hat{X}_p||_F^2 + \hat{X}_g \,-\, X_g\,log\,(\hat{X}_g) \tag{8}$$

where λ is the weight of the function, $\hat{X}_p$ is $F.G_p$ and $\hat{X}_g$ is $F.G_g$

The updated version of the function to fit the case in this study:

$$L\ (F,\ G_p, G_g)\ =\ \sum_{ij}\ [\frac{\lambda}{2}\ (X_p\ -\ \hat{X}_p\ log\ (\hat{X}_p)) + \hat{X}_g\ -\ X_g\ log\ (\hat{X}_g)] \qquad (9)$$

Here we converted the first part of the equation, where X_p matrix and its derivatives are involved.

The update equations for each of the factor matrices for original study:

$$\nabla G_p\ L\ (F, G_p, G_g)\ =\ \lambda\ (\ F\ F^T G_p\ -\ F\ X_p) \qquad (10)$$

$$\nabla G_g\ L\ (F, G_p, G_g)\ =\ F\ (E_G\ -\ \tilde{X}_g) \qquad (11)$$

$$\nabla\ F\ L\ (F, G_p, G_g)\ =\ \lambda\ (-\ G_p\ X_p^T\ +\ G_p\ G_p^T\ F)\ +\ G_g\ (E_F\ -\ \tilde{X}_g^T\ ) \qquad (12)$$

where E_G and E_F are all-one matrix and $\tilde{X}_g = X_g\ /\ \tilde{X}_g$.

The update equations of low-rank matrices are converted for this study as:

$$\nabla G_p\ L\ (F, G_p, G_g)\ =\ F\ (E_p\ -\ \tilde{X}_p) \qquad (13)$$

$$\nabla G_g\ L\ (F, G_p, G_g)\ =\ F\ (E_G\ -\ \tilde{X}_g) \qquad (14)$$

$$\nabla\ F\ L\ (F, G_p, G_g)\ =\ \lambda\ (G_p\ (E_q\ -\ \tilde{X}_p^T\ )]\ +\ G_g\ (E_F\ -\ \tilde{X}_g^T\ ) \qquad (15)$$

where λ is taken as 1 in this study since their weight are equal. We converted the equation (13) according to our data, and also the first part of equation (15), since (15) is the sum of (13) and (14).

As the next step, as in basic NMF procedure, firstly the algorithm is run for toy matrices (of 15x10 and 15x8, respectively) to determine the parameters like iterations, tolerance etc. Then the algorithm of both versions (before and after conversion) is run with the determined parameters to find the most befitting latent values (k) for relation matrix pairs to be inserted. Another reason of this part was to compare the performances of both versions, to see whether the conversion improved the performance as presumed. One of the constant matrix for taken the pairs was Protein x Disease matrix as we wanted to use one relation matrix with GO term and the other with disease terms so that the algorithm would be able to learn the correlation between diseases and functions. After determining the optimum k values for every relation pair, 10-fold cross validation is applied for evaluation, and then performance scores are obtained.

3.4.2.   NMTF Algorithm

NMTF algorithm is another model based on NMF approach, where the improvement of the algorithm to the baseline is the ability to use all the relation matrices at one run. In this method different k values for each data type in a relation can be given, with the purpose of obtaining better results due to data types having the opportunity of being clustered via their respective optimum group numbers (Ceddia et al., 2019; Dissez et al., 2019). As mentioned in Section 2.3.2., they also compared different starting methods for low-rank matrices in order to achieve more consistent results. They compared the results of random uniform, random ACOL, k-means clustering and spherical k-means clustering methods for starting of low-rank matrices and discovered that spherical k-means clustering gives the best result. We used their spherical k-means clustering method as well for this part of the study.
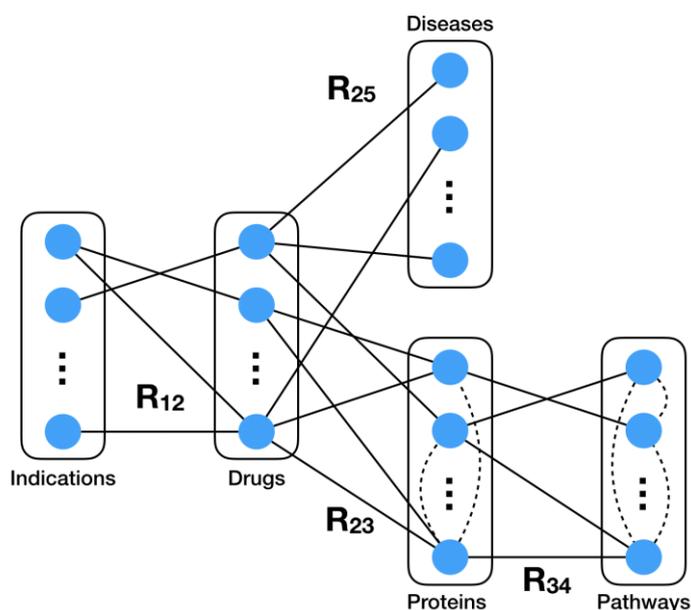
Figure 3.5. An example schematic representation of relation diagram. (Dissez et al., 2019).

An example representing the diagram of interactions among data types in the original study is given in **Figure 3.5.** In their case, drugs were related with indications, diseases and proteins; therefore, the relation matrices involving drugs were indications x drugs (R12), drugs x proteins (R23) and drugs x diseases (R25). The relation diagram of their data was a little different from this study, so the formulas in update rules etc. are converted accordingly. Unlike in the source study, the data types in this study were proteins (1), CC (2), MF (3), BP (4) GO annotation terms and diseases (5). Thus, the relation matrices here were R12 (Protein x CC), R13 (Protein x MF), R14 (Protein x BP) and R15 (Protein x Disease). The diagram applied to our study based on the above one is available in **Figure 3.6**.

25
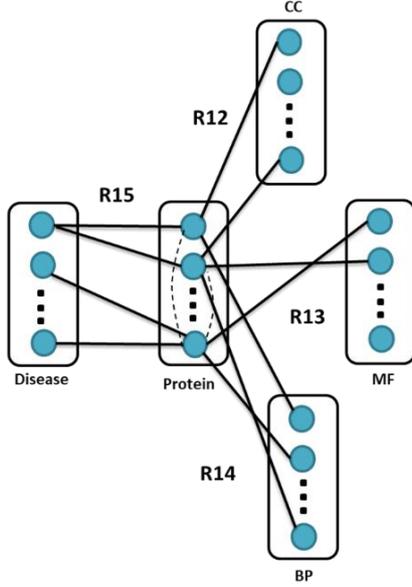
Figure 3.6. The schematic representation of the relation diagram of this study.

The NMTF objective function in the original study according to their data was:

$$J(G,S) = \left\lVert R_{12} - G_1 S_{12} G_2^T \right\rVert^2 + \left\lVert R_{23} - G_2 S_{23} G_3^T \right\rVert^2 + \left\lVert R_{34} - G_3 S_{34} G_4^T \right\rVert^2 \quad (16)$$
$$+ \left\lVert R_{25} - G_2 S_{25} G_5^T \right\rVert^2 + tr(G_3^T L_3 G_3) + tr(G_4^T L_4 G_4)$$

where G_x's are the data types' respective low-rank matrices, S_x's are the intermediate matrices of latent factor (k) values, L_x's are the intra type relation matrices of their respective data types, and tr is trace function. Trace of a square matrix A is the sum of the diagonal elements (Lang, 2013):

$$Tr\,(A) = \sum_{i=1}^{n} a_{ii} \quad (17)$$

The NMTF objective function for our data is:

$$J(G,S) = \left\lVert R_{12} - G_1 S_{12} G_2^T \right\rVert^2 + \left\lVert R_{13} - G_1 S_{13} G_3^T \right\rVert^2 + \left\lVert R_{14} - G_1 S_{14} G_4^T \right\rVert^2 \quad (18)$$
$$+ \left\lVert R_{15} - G_1 S_{15} G_5^T \right\rVert^2 + tr(G_1^T L_1 G_1)$$

26

where 1 is the representation of protein, 2 of CC, 3 of MF, 4 of BP and 5 of disease terms. The trace function is for protein-protein interaction matrix. The equation was updated according to the relation diagram of this study.

The update rule equations are converted to fit the data of this study as well. The original form of update rule equations was:

$$G_1 \leftarrow G_1 \sqrt{\frac{R_{12}G_2S_{12}^T}{G_1G_1^TR_{12}G_2S_{12}^T}} \tag{19}$$

$$G_2 \leftarrow G_2 \sqrt{\frac{R_{12}^TG_1S_{12} + R_{23}G_3S_{23}^T + R_{25}G_5S_{25}^T}{G_2G_2^TR_{12}^TG_1S_{12} + G_2G_2^TR_{23}G_3S_{23}^T + G_2G_2^TR_{25}G_5S_{25}^T}} \tag{20}$$

$$G_3 \leftarrow G_3 \sqrt{\frac{R_{23}^TG_2S_{23} + R_{34}G_4S_{34}^T}{G_3G_3^TR_{23}^TG_2S_{23} + G_3G_3^TR_{34}G_4S_{34}^T}} \tag{21}$$

$$G_4 \leftarrow G_4 \sqrt{\frac{R_{34}^TG_3S_{34}}{G_4G_4^TR_{34}^TG_4S_{34}}} \tag{22}$$

$$G_5 \leftarrow G_5 \sqrt{\frac{R_{25}^TG_2S_{25}}{G_5G_5^TR_{25}^TG_2S_{25}}} \tag{23}$$

$$S_{12} \leftarrow S_{12} \sqrt{\frac{G_1^TR_{12}G_2}{G_1^TG_1S_{12}G_2^TG_2}} \tag{24}$$

$$S_{23} \leftarrow S_{23} \sqrt{\frac{G_2^TR_{23}G_3}{G_2^TG_2S_{23}G_3^TG_3}} \tag{25}$$

$$S_{34} \leftarrow S_{34} \sqrt{\frac{G_3^TR_{34}G_4}{G_3^TG_3S_{34}G_4^TG_4}} \tag{26}$$

$$S_{25} \leftarrow S_{25} \sqrt{\frac{G_2^T R_{25} G_5}{G_2^T G_4 S_{25} G_5^T G_5}} \tag{27}$$

The above equations are converted for this study as shown in below:

$$G_1 \leftarrow G_1 \sqrt{\frac{R_{12} G_2 S_{12}^T + R_{13} G_3 S_{13}^T + R_{14} G_4 S_{14}^T + R_{15} G_5 S_{15}^T}{G_1 G_1^T R_{12} G_2 S_{12}^T + G_1 G_1^T R_{13} G_3 S_{13}^T + G_1 G_1^T R_{14} G_4 S_{14}^T + G_1 G_1^T R_{15} G_5 S_{15}^T}}$$

$$\tag{28}$$

$$G_2 \leftarrow G_2 \sqrt{\frac{R_{12}^T G_1 S_{12}}{G_2 G_2^T R_{12}^T G_1 S_{12}}} \tag{29}$$

$$G_3 \leftarrow G_3 \sqrt{\frac{R_{13}^T G_1 S_{13}}{G_3 G_3^T R_{13}^T G_1 S_{13}}} \tag{30}$$

$$G_4 \leftarrow G_4 \sqrt{\frac{R_{14}^T G_1 S_{14}}{G_4 G_4^T R_{14}^T G_1 S_{14}}} \tag{31}$$

$$G_5 \leftarrow G_5 \sqrt{\frac{R_{15}^T G_1 S_{15}}{G_5 G_5^T R_{15}^T G_1 S_{15}}} \tag{32}$$

$$S_{12} \leftarrow S_{12} \sqrt{\frac{G_1^T R_{12} G_2}{G_1^T G_1 S_{12} G_2^T G_2}} \tag{33}$$

$$S_{13} \leftarrow S_{13} \sqrt{\frac{G_1^T R_{13} G_3}{G_1^T G_1 S_{13} G_3^T G_3}} \tag{34}$$

$$S_{14} \leftarrow S_{14} \sqrt{\frac{G_1^T R_{14} G_4}{G_1^T G_1 S_{14} G_4^T G_4}} \qquad (35)$$

$$S_{15} \leftarrow S_{15} \sqrt{\frac{G_1^T R_{15} G_5}{G_1^T G_1 S_{15} G_5^T G_5}} \qquad (36)$$

In the original study, the drug and protein data types were in relation with more than one other data type, so the update rules regarding these data types contained all the relation and low-rank matrices they were involved in. For example, the protein data type was present in both drug x protein and protein x pathway relations (please refer to **Figure 3.4** for more clear representation), so the update rule of the low-rank matrix of protein contains matrices from both relations' approximations. In the converted version of the equations for this study, protein was the data type involved in all the relation matrices, so the update rule of it contains all the relation and low-rank matrices.

Their code for the application was also available as an open source, but the programming platform was Python in this case. Python v.3.7 was used to run this section of the experiment. After adaptation to the data of this study, firstly run the data matrices and compared their error rates to determine the k values to be used, and then applied 10-fold cross-validation and obtained the performance scores.

3.5. Performance Evaluation

For every method used in this study, at first the latent factor (k) values are determined via error comparison. The algorithms are run with training data to detect the best k values for each matrix. For comparison, the error rates are used, of which the formula is;

$$error = \frac{\sum_{ij} |data\ matrix_{ij} - prediction\ matrix_{ij}|}{size\ of\ the\ matrix\ (m\ x\ n)} \qquad (37)$$

where i and j are the index numbers of each element of matrices.

The result is then compared with random error rate, which is defined as;

$$random\ error = \frac{number\ of\ positive\ points}{size\ of\ the\ matrix\ (m\ x\ n)} \qquad (38)$$

Random error is a special case of error when a model predicts all points as zero without any prediction (machine learning). Logically, it is needed to take k values that give lower error values than of random error value, so that it can be said that the model performs better than random predictor for that k value.

After the determination of k values to be used with the training data, 10-fold cross validation is applied to test the performance of the method, where the available data in hand (the annotations) are divided into ten parts of equal size, and each time one of the parts are excluded from the training data to be fitted to the model. After obtaining the results for 10-fold cross-validation, as an additional analysis we performed 3-fold cross validation. By further evaluating the algorithms with 3-fold cross validation we aimed to prevent overfitting of the models; in other words, to memorize the data instead of learning, and thus failing to model future data (Oxford, n.d.). Another aim was to create more challenging test for the models to compare the performance of the models by the training data sparsity. While applying 3-fold cross validation, we took the latent factor (k) values for a matrix same in each application, to compare the performance changes as the model becomes more complex.

Later, according to the results depending on different threshold values, where the predicted point is considered positive if equal or above the threshold and negative otherwise, the confusion matrices are constructed, and then these points are compared to the data points. The confusion matrix contains the four main classes identifying each result point; true positive (TP – both the prediction and the given point is positive), true negative (TN – both the prediction and the given data is negative), false positive (FP – where the data point is negative but the prediction for it is positive), and false negative (FN – where the real data point is positive but the prediction is negative). From these, performance scores are calculated to better understand the model, such as recall, precision, false positive rate (FPR), accuracy and so on.

Recall (also called sensitivity or true positive rate -TPR-) shows how many of the positive values in the data are correctly predicted as positive.

$$\text{Recall} = \frac{TP}{TP + FN} \tag{39}$$

Precision shows the ratio of how many of the predicted positives are true.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{40}$$

False positive rate (FPR) is the ratio of how many of the negative points are predicted as positive. It is used with TPR for ROC curve plotting.

$$\text{False Positive Rate (FPR)} = \frac{FP}{FP + TN} \tag{41}$$

30

Accuracy is the ratio indicating how many of the predictions are correct, positive or negative.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FN} + \text{FP}} \qquad (42)$$

F1 score helps the data to be better understood when the precision and recall are misleading due to imbalance of the data.

$$\text{F1 score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \qquad (43)$$

Another score to describe the imbalanced data in a better way is Matthews correlation coefficient (MCC).

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}} \qquad (44)$$

With TPR and FPR scores for each threshold, the receiver operating characteristic (ROC) curve is plotted, which shows the performance of the prediction model. The diagonal line in the graph shows the random (uninformative) guessing, and the models having ROC curves that are above this diagonal line are said to be better predictor than random while the models with curves below the line are said to predict worse than random classifiers. An example of ROC curve can be seen in **Figure 3.7.** The green curve in the figure, which is also pointed by black arrows, is the ROC curve, and the red diagonal line is the random classifier. For comparison of different models, the Area Under the Curve (AUC) is calculated, which is the area between ROC curve and the x axis, where the area is bounded by $x=1$. It is interpreted as the higher the AUC, the better the predictive model.
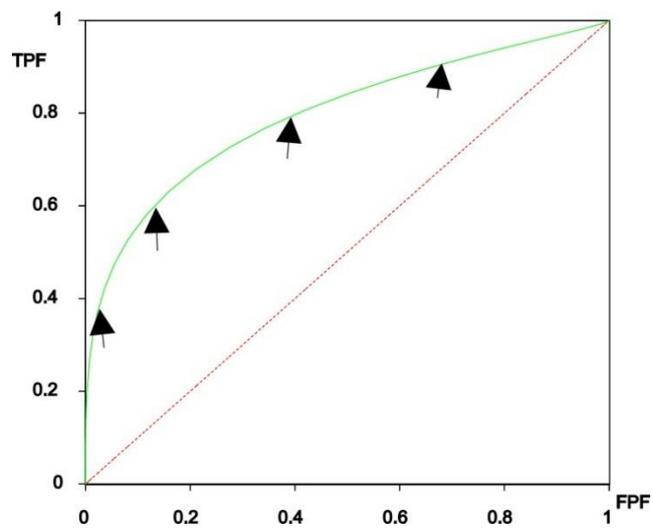
Figure 3.7. An example of ROC curve. The area under the curve gives the AUC value. (Hajian-Tilaki, 2013).

CHAPTER 4

4. RESULTS

4.1. Application of the Baseline NMF Algorithm

Following the algorithm run with toy matrices (with the purpose of determining the parameters), it is observed that there was no significant performance difference for iteration number and tolerance value selections, so the default parameters were employed. On the other hand, "als" update method was found to perform better than "mul" method, so the experiments in this field from this point were continued with the "als" update method.

Afterwards, the algorithm was run with the actual training data to detect the best k values for each matrix. The graphical plots of this analysis for each relation matrix can be seen in **Figure 4.1**. The random prediction model error was 0.034 for Protein x CC, 0.023 for Protein x MF, and 0.014 for Protein x BP and Protein x Disease matrices. One of the criteria while selecting the number of latent factors was that the value should have lower error rate than the random predictor. The other criterium was the value being not too high, so that, there would still be a classification among data type entities. If the k values were too close to the number of entities of the original data matrices, there would be no grouping at all. As a result of this experiment, k values of 50, 50, 100 and 100, were selected as the number of latent factors for CC, MF, BP and diseases, respectively.
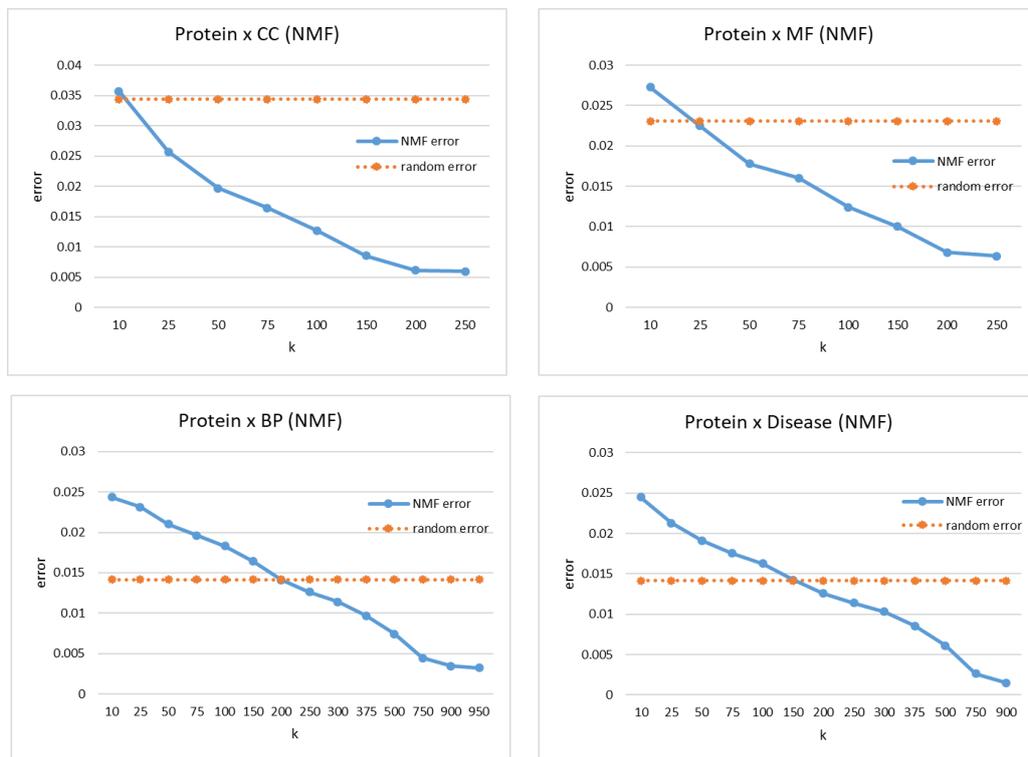
Figure 4.1. The comparison of error values of NMF algorithm and uniform predictor for respective relation matrices.

After determining the latent factor (k) values, the model was evaluated with 10-fold cross validation, and then the performance scores were calculated. Below are the scores for the threshold 0.5 in **Table 4.1.** The threshold was chosen as 0.5 while giving the scores since it is the middle point of 0 and 1. Furthermore, the threshold is generally taken as 0.5 in literature as well. The MCC scores for matrices are 0.26, 0.29, 0.32, and 0.40 for Protein x CC, Protein x MF, Protein x BP and Protein x Disease matrices, respectively. On the other hand, the given scores for the threshold of 0.5 are not the optimal performance scores of the models. However, the commonly chosen threshold of 0.5 was not the one giving the best performance results, the models gave the best performance scores when the threshold was taken as 0.02. Since this threshold is not close to the universal threshold of 0.5, the scores for this threshold were presented in separate tables as well. The scores according to the threshold of 0.02 can also be seen in **Table 4.2**. The scores for all threshold values are available in **Appendix A.1.**

Table 4.1. The scores at threshold 0.5 for each relation matrix in the baseline NMF algorithm for 10-fold cross-validation.

	Thr.	TP	TN	FP	FN	Rec.	Prec.	FPR	Acc.	F-sc	MCC
PxCC	0.5	6558	52170	30	45642	0.13	1.00	0.00	0.56	0.22	**0.26**
PxMF	0.5	5305	34250	20	28965	0.15	1.00	0.00	0.58	0.27	**0.29**
PxBP	0.5	14065	75853	27	61815	0.19	1.00	0.00	0.59	0.31	**0.32**
PxD	0.5	19389	71356	14	51981	0.27	1.00	0.00	0.64	0.43	**0.40**

PxCC: Protein x CC, PxMF: Protein x MF, PxBP: Protein x BP, PxD: Protein x Disease, Thr.: Threshold, Rec.: Recall, Prec.: Precision, Acc: Accuracy, F-sc.: F-score

Table 4.2. The scores at threshold=0.02 for each relation matrix in the baseline NMF algorithm for 10-fold cross-validation.

	Thr.	TP	TN	FP	FN	Rec.	Prec.	FPR	Acc.	F-sc	MCC
PxCC	0.02	45403	46065	6135	6797	0.87	0.88	0.12	0.88	0.88	**0.75**
PxMF	0.02	29022	30890	3380	5248	0.85	0.90	0.10	0.87	0.87	**0.75**
PxBP	0.02	64165	66833	9047	11715	0.85	0.88	0.12	0.86	0.86	**0.73**
PxD	0.02	59794	63476	7894	11576	0.83	0.88	0.11	0.86	0.85	**0.72**

PxCC: Protein x CC, PxMF: Protein x MF, PxBP: Protein x BP, PxD: Protein x Disease, Thr.: Threshold, Rec.: Recall, Prec.: Precision, Acc: Accuracy, F-sc.: F-score

The ROC curves and the AUC values are given in **Figure 4.2.** It is seen that the AUC values for the algorithm run with each matrix are 0.94, 0.92, 0.93 and 0.92 for Protein x CC, Protein x MF, Protein x BP and Protein x Disease matrices, respectively.
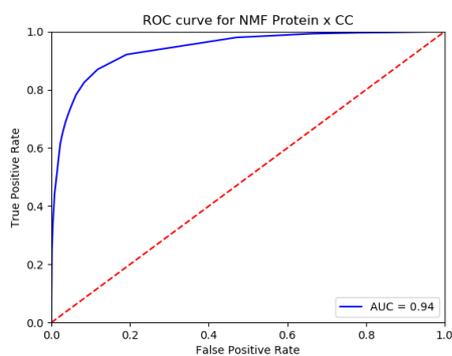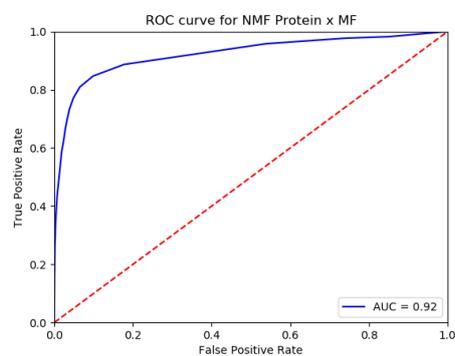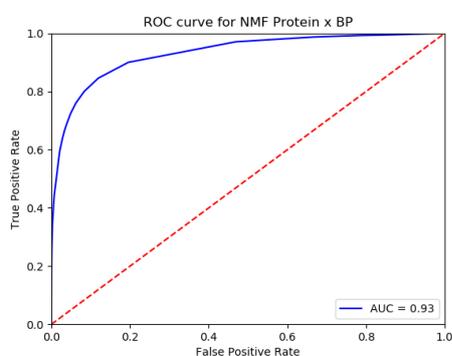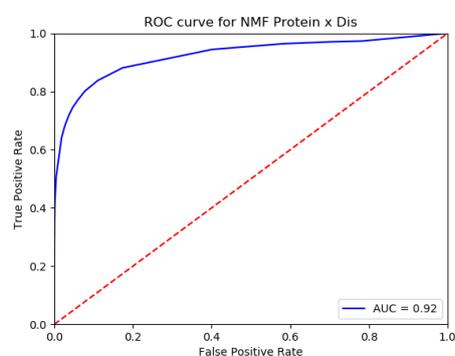
Figure 4.2. The ROC curves and AUC values for NMF cross-validation scores. **a.** Protein x CC, **b.** Protein x MF, **c.** Protein x BP, **d.** Protein x Disease

As mentioned in Chapter 3, the tests were re-run with further filtered datasets. The performance evaluation of these analyses was done via 3-fold cross-validation, in order to test the performance of models when they were run in more challenging conditions. The latent factor values were 150 for Protein x CC and Protein x MF matrices, and 200 for Protein x BP and Protein x Disease matrices. The results of these analyses for baseline NMF algorithm can be examined in **Table 4.3** and **Table 4.4** for thresholds of 0.5 and the threshold giving the best performance scores for each matrix model, respectively. The scores for all matrices are around 0.65 when the best threshold (score-wise) was taken. On the other hand, the scores when the threshold was taken as 0.5 were 0.13 for Protein x CC, 0.19 for Protein x MF, 0.21 for Protein x BP and 0.27 for Protein x Disease matrices.

Table 4.3. The scores at threshold 0.5 for each relation matrix in the baseline NMF algorithm for 3-fold cross-validation.

	Thr.	TP	TN	FP	FN	Rec.	Prec.	FPR	Acc.	F-sc	MCC
PxCC	0.5	1120	37297	2	33020	0.03	1.00	0.00	0.54	0.06	0.13
PxMF	0.5	1575	25439	4	21034	0.07	1.00	0.00	0.56	0.13	0.19
PxBP	0.5	4786	59216	4	51173	0.09	1.00	7E-05	0.56	0.16	0.21
PxD	0.5	8209	65778	6	54912	0.13	1.00	9E-05	0.57	0.23	0.27

PxCC: Protein x CC, PxMF: Protein x MF, PxBP: Protein x BP, PxD: Protein x Disease, Thr.: Threshold, Rec.: Recall, Prec.: Precision, Acc: Accuracy, F-sc.: F-score

Table 4.4. The scores at threshold giving the best scores for each relation matrix in the baseline NMF algorithm for 3-fold cross-validation.

	Thr.	TP	TN	FP	FN	Rec.	Prec.	FPR	Acc.	F-sc	MCC
PxCC	**0.01**	24626	33443	3856	9514	0.72	0.86	0.10	0.81	0.79	0.63
PxMF	**0.01**	16560	22774	2669	6049	0.73	0.86	0.10	0.82	0.79	0.64
PxBP	**0.02**	42754	53021	6199	13205	0.76	0.87	0.10	0.83	0.82	0.67
PxD	**0.02**	48616	57686	8098	14505	0.77	0.86	0.12	0.82	0.81	0.65

PxCC: Protein x CC, PxMF: Protein x MF, PxBP: Protein x BP, PxD: Protein x Disease, Thr.: Threshold, Rec.: Recall, Prec.: Precision, Acc: Accuracy, F-sc.: F-score

4.2. Application of the HNMF Algorithm

In HNMF method, the parameters such as tolerance, "max_iter", "max_sub_iter", "timelimit" etc. were determined by running the algorithm with the handmade toy data matrices. There was not any significant difference observed as the parameters change. So, the iteration number was taken as high as 500 since small iteration number may cause lower performance; the tolerance as 1e-6 (the default tolerance in the code), and the "timelimit" was kept high so that the algorithm was not terminated prematurely. In addition, the lambda value was kept as 1, meaning no weight is given to any of the matrices. Later, different latent factor (k) values were observed with training data to find the optimum ones for each pair. In this stage, the error values of the algorithms for before and after the equation conversions were compared. The comparison of error rates for both versions can be seen in **Table 4.5.** Since even one run could took days in some cases, the tested latent factor values were limited. It is observed that the results were improved as expected in the converted part of the algorithm. For example, the error rate for Protein x CC matrix was 0.0454 before conversion while it improved to 0.0180 after the algorithm was tinkered with. As a result, the latent factor (k) values were taken as 150 for the pairs of Protein x CC / Protein x MF and Protein x Disease matrices, while as 200 for Protein x BP and Protein x Disease pair since the size of Protein x BP matrix was bigger than the other two GO term matrices.

Table 4.5. The error rates for the HNMF algorithm before and after the loss functions and the update equations were converted.

	model	k	Error (P x __)	Error (P x D)
HNMF before conversion	PxMF and PxD	10	0.0313	0.0218
		20	0.0321	0.0204
		50	0.0338	0.0174
		100	0.0343	0.0142
		150	0.0333	0.0133
		200	**0.0314**	**0.0102**
	PxCC and PxD	50	0.0486	0.0175
		150	**0.0454**	**0.0120**
	PxBP and PxD	50	0.0247	0.0177
		150	0.0248	0.0125
		200	**0.0246**	**0.0110**
HNMF after conversion	PxMF and PxD	50	0.0250	0.0186
		150	**0.0168**	**0.0132**
	PxCC and PxD	50	0.0281	0.0187
		150	**0.0180**	**0.0137**
	PxBP and PxD	50	0.0212	0.0200
		150	0.0171	0.0159
		200	**0.0154**	**0.0144**

(Please refer to Methods 3.4.1). The lowest error rates for each model is in bold. Protein x (__) represents the other pair used in the model with Protein x Disease matrix, and it is indicated in the model column. PxCC: Protein x CC, PxMF: Protein x MF, PxBP: Protein x BP, PxD: Protein x Disease

Afterwards, the cross validation was performed with determined k values for each model using the updated algorithm. Below in **Table 4.6.** and **Table 4.7** the performance scores for the model with Protein x CC and Protein x Disease matrices can be examined for thresholds of 0.5 and 0.02, respectively. As in the case of baseline NMF algorithm, the optimal scores were calculated when the threshold was taken as 0.02, but results for threshold of 0.5 were also given for clearer comparison with the studies in literature. Likewise, in **Table 4.8.** the performance scores for threshold of 0.5, and in **Table 4.9** the scores for the threshold of 0.02 for the HNMF model with Protein x MF and Protein x Disease model is showed. Lastly, in **Table 4.10.** and in **Table 4.11** the scores for the model run with Protein x BP and Protein x Disease matrices can be seen, for the thresholds of 0.5 and 0.02 respectively. The MCC scores for Protein x Disease is around 0.40 for all three models at threshold=0.5. MCC score is 0.45 for Protein x CC matrix, 0.37 for Protein x MF matrix, and 0.32 for Protein x BP matrix. The complete versions of the tables are available in **Appendix A.2.** In **Figure 4.3.** the ROC curves and AUC scores can be seen for each model. The AUC values are around 0.87 for every matrix in this experiment.

Table 4.6. The scores for threshold of 0.5 for Protein x CC and Protein x Disease model of HNMF application for 10-fold cross-validation.

	Thr.	TP	TN	FP	FN	Rec.	Prec.	FPR	Acc.	F-sc	MCC
PxCC	0.5	17811	52112	88	34389	0.34	1.00	0.00	0.67	0.51	**0.45**
PxD	0.5	19665	71299	71	51705	0.28	1.00	0.00	0.64	0.43	**0.40**

PxCC: Protein x CC, PxD: Protein x Disease, Thr.: Threshold, Rec.: Recall, Prec.: Precision, Acc: Accuracy, F-sc.: F-score

Table 4.7. The scores for threshold of 0.02 for Protein x CC and Protein x Disease model of HNMF application for 10-fold cross-validation.

	Thr.	TP	TN	FP	FN	Rec.	Prec.	FPR	Acc.	F-sc	MCC
PxCC	0.02	40620	49881	2319	11580	0.78	0.95	0.04	0.87	0.85	**0.75**
PxD	0.02	53149	68116	3254	18221	0.74	0.94	0.05	0.85	0.83	**0.72**

PxCC: Protein x CC, PxD: Protein x Disease, Thr.: Threshold, Rec.: Recall, Prec.: Precision, Acc: Accuracy, F-sc.: F-score

Table 4.8. The scores for threshold of 0.5 for Protein x MF and Protein x Disease model of HNMF application for 10-fold cross-validation.

	Thr.	TP	TN	FP	FN	Rec.	Prec.	FPR	Acc.	F-sc	MCC
PxMF	0.5	8556	34223	47	25714	0.25	0.99	0.00	0.62	0.40	**0.37**
PxD	0.5	20324	71280	90	51046	0.28	1.00	0.00	0.64	0.44	**0.40**

Protein x MF, PxD: Protein x Disease, Thr.: Threshold, Rec.: Recall, Prec.: Precision, Acc: Accuracy, F-sc.: F-score

Table 4.9. The scores for threshold of 0.02 for Protein x MF and Protein x Disease model of HNMF application for 10-fold cross-validation.

	Thr.	TP	TN	FP	FN	Rec.	Prec.	FPR	Acc.	F-sc	MCC
PxMF	0.02	25503	32661	1609	8767	0.74	0.94	0.05	0.85	0.83	**0.71**
PxD	0.02	52468	68233	3137	18902	0.74	0.94	0.04	0.85	0.83	**0.71**

Protein x MF, PxD: Protein x Disease, Thr.: Threshold, Rec.: Recall, Prec.: Precision, Acc: Accuracy, F-sc.: F-score

Table 4.10. The scores in threshold of 0.5 for Protein x BP and Protein x Disease model of HNMF application for 10-fold cross-validation.

	Thr.	TP	TN	FP	FN	Rec.	Prec.	FPR	Acc.	F-sc	MCC
PxBP	0.5	13934	75808	72	61946	0.18	0.99	0.00	0.59	0.31	**0.32**
PxD	0.5	18975	71282	88	52395	0.27	1.00	0.00	0.63	0.42	**0.39**

PxBP: Protein x BP, PxD: Protein x Disease, Thr.: Threshold, Rec.: Recall, Prec.: Precision, Acc: Accuracy, F-sc.: F-score

Table 4.11. The scores in threshold of 0.02 for Protein x BP and Protein x Disease model of HNMF application for 10-fold cross-validation.

	Thr.	TP	TN	FP	FN	Rec.	Prec.	FPR	Acc.	F-sc	MCC
PxBP	0.02	58258	72158	3722	17622	0.77	0.94	0.05	0.86	0.85	**0.73**
PxD	0.02	53643	67838	3532	17727	0.75	0.94	0.05	0.8	0.83	**0.72**

PxBP: Protein x BP, PxD: Protein x Disease, Thr.: Threshold, Rec.: Recall, Prec.: Precision, Acc: Accuracy, F-sc.: F-score
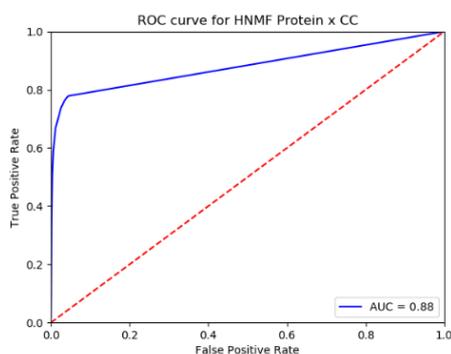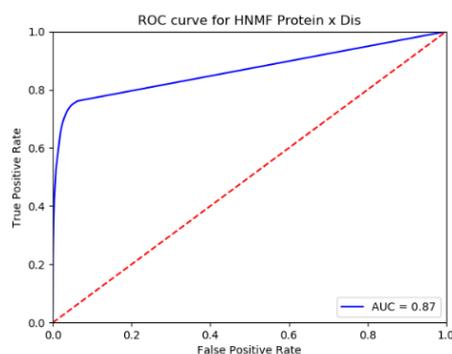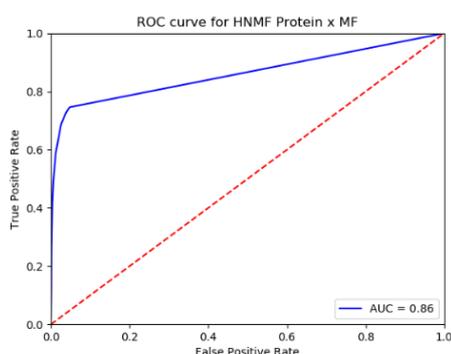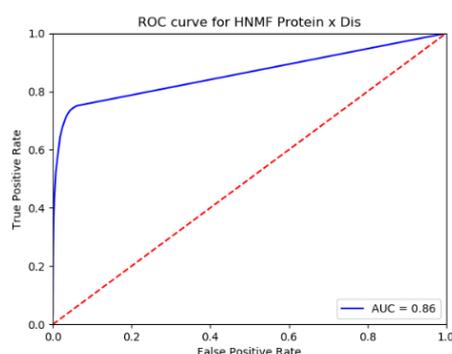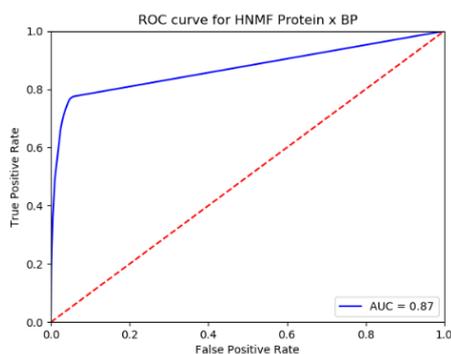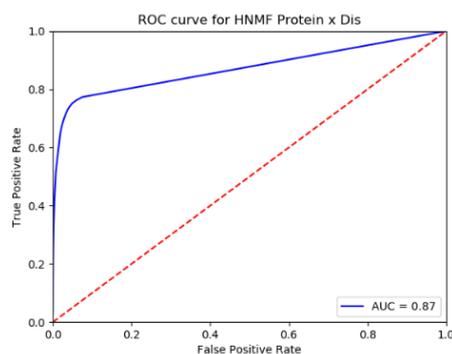
Figure 4.3. The ROC curves and AUC values for HNMF cross-validation scores. a. Protein x CC and Protein x Disease model (a.1. Protein x CC, a.2. Protein x Disease), b. Protein x MF and Protein x Disease model (b.1. Protein x MF, b.2. Protein x Disease), c. Protein x BP and Protein x Disease model (c.1. Protein x BP, c.2. Protein x Disease).

To investigate the unexpected appearance of straight line in the ROC curves of HNMF models above, the graphs were re-drawn to discover where in the curve each threshold point was. The new ROC curves can be seen in **Figure 4.4**. As can be seen in the

figures, the straight line was drawn between the closest threshold points (0 and 1e-6) while connecting the scatter points because there was no point in between.
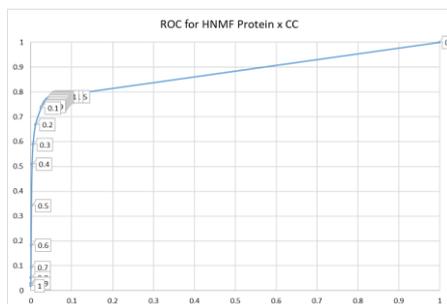
a.1.   **a.2.**
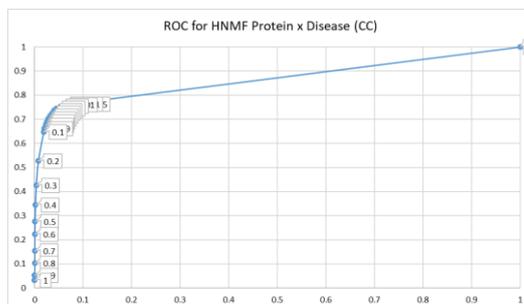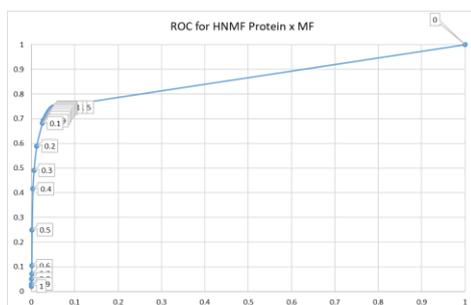
b.1.   **b.2.**
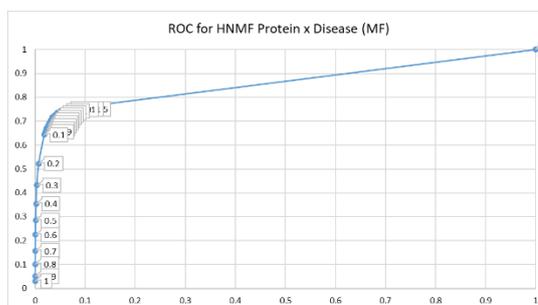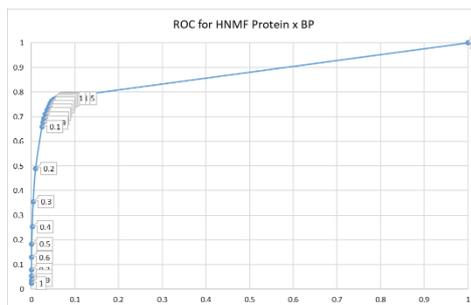
c.1.   **c.2.**
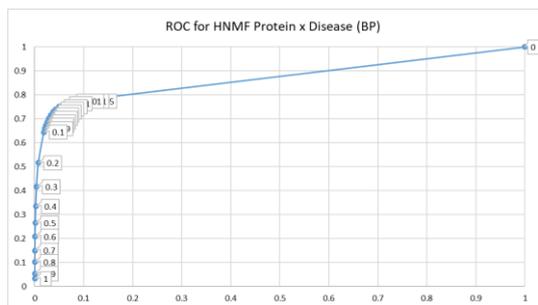
Figure 4.4. The ROC curves for HNMF models, along with threshold points. a. Protein x CC and Protein x Disease model (a.1. Protein x CC, a.2. Protein x Disease), b. Protein x MF and Protein x Disease model (b.1. Protein x MF, b.2. Protein x Disease), c. Protein x BP and Protein x Disease model (c.1. Protein x BP, c.2. Protein x Disease).

As in baseline NMF model, the HNMF models were also run with new dataset, and evaluated by 3-fold cross-validation. The results of these analyses can be seen in **Table 4.12, Table 4.13, Table 4.14, Table 4.15, Table 4.16** and **Table 4.17** for each matrix pair (the pairs were same as in the 10-fold cross-validation), at the thresholds of 0.5 and the one giving the best scores.

Table 4.12. The scores for threshold of 0.5 for Protein x CC and Protein x Disease model of HNMF application for 3-fold cross-validation.

	Thr.	TP	TN	FP	FN	Rec.	Prec.	FPR	Acc.	F-sc	MCC
PxCC	0.5	8539	37235	64	25601	0.25	0.99	0.00	0.64	0.40	0.38
PxD	0.5	11155	65722	62	51966	0.18	0.99	0.00	0.60	0.30	0.31

PxCC: Protein x CC, PxD: Protein x Disease, Thr.: Threshold, Rec.: Recall, Prec.: Precision, Acc: Accuracy, F-sc.: F-score

Table 4.13. The scores for threshold of 0.02 for Protein x CC and Protein x Disease model of HNMF application for 3-fold cross-validation.

	Thr.	TP	TN	FP	FN	Rec.	Prec.	FPR	Acc.	F-sc	MCC
PxCC	0.03	23480	35463	1836	10660	0.69	0.93	0.05	0.83	0.79	0.67
PxD	0.01	43562	61723	4061	19559	0.69	0.91	0.06	0.82	0.79	0.65

PxCC: Protein x CC, PxD: Protein x Disease, Thr.: Threshold, Rec.: Recall, Prec.: Precision, Acc: Accuracy, F-sc.: F-score

Table 4.14. The scores for threshold of 0.5 for Protein x MF and Protein x Disease model of HNMF application for 3-fold cross-validation.

	Thr.	TP	TN	FP	FN	Rec.	Prec.	FPR	Acc.	F-sc	MCC
PxMF	0.5	4404	25414	29	18205	0.19	0.99	0.00	0.62	0.33	0.33
PxD	0.5	11468	65734	50	51653	0.18	1.00	0.00	0.60	0.31	0.32

Protein x MF, PxD: Protein x Disease, Thr.: Threshold, Rec.: Recall, Prec.: Precision, Acc: Accuracy, F-sc.: F-score

Table 4.15. The scores for threshold of 0.02 for Protein x MF and Protein x Disease model of HNMF application for 3-fold cross-validation.

	Thr.	TP	TN	FP	FN	Rec.	Prec.	FPR	Acc.	F-sc	MCC
PxMF	0.02	15348	24127	1316	7261	0.68	0.92	0.05	0.82	0.78	0.66
PxD	0.02	42038	62410	3374	21083	0.67	0.93	0.05	0.81	0.77	0.64

Protein x MF, PxD: Protein x Disease, Thr.: Threshold, Rec.: Recall, Prec.: Precision, Acc: Accuracy, F-sc.: F-score

Table 4.16. The scores in threshold of 0.5 for Protein x BP and Protein x Disease model of HNMF application for 3-fold cross-validation.

	Thr.	TP	TN	FP	FN	Rec.	Prec.	FPR	Acc.	F-sc	MCC
PxBP	0.5	6733	59179	41	49226	0.12	0.99	0.00	0.57	0.21	0.25
PxD	0.5	10977	65717	67	52144	0.17	0.99	0.00	0.59	0.30	0.31

PxBP: Protein x BP, PxD: Protein x Disease, Thr.: Threshold, Rec.: Recall, Prec.: Precision, Acc: Accuracy, F-sc.: F-score

Table 4.17. The scores in threshold of 0.02 for Protein x BP and Protein x Disease model of HNMF application for 3-fold cross-validation.

	Thr.	TP	TN	FP	FN	Rec.	Prec.	FPR	Acc.	F-sc	MCC
PxBP	0.02	37770	56124	3096	18189	0.67	0.92	0.05	0.82	0.78	0.65
PxD	0.02	43069	62107	3677	20052	0.68	0.92	0.06	0.82	0.78	0.65

PxBP: Protein x BP, PxD: Protein x Disease, Thr.: Threshold, Rec.: Recall, Prec.: Precision, Acc: Accuracy, F-sc.: F-score

4.3.Application of NMTF Algorithm

In the last method of the study, NMTF algorithm, as first step the latent factor (k) value combination was determined. All the results were obtained for spherical k-means as initialization method for low-rank matrices. The k values and the error rates before protein - protein interaction were added to the algorithm are given in **Table 4.18**, where R12, R13, R14 and R15 matrices are Protein x CC, Protein x MF, Protein x BP and Protein x Disease matrices, respectively, together with averages of error rates for every k set. To determine the best latent factor numbers, the average of errors for each matrix was taken. The latent factor numbers were determined as 150 for CC, 2500 for proteins, 150 for MF, 150 for BP and 150 for disease data types. Full version of the table is available in **Appendix A**.**3**.

Table 4.18. The best latent factor (k) value set and the resulting lowest error rates (without PPI matrix).

	k1 (CC)	k2 (Prot)	k3 (MF)	k4 (BP)	k5 (Dis)
k10	150	2500	150	150	150
	R12	**R13**	**R14**	**R15**	**Avg error**
error	0.027782	0.03032	0.024335	0.024669	0.026777

CC: Cellular Component, Prot: Protein, MF: Molecular Function, BP: Biological Process, Dis: Disease
Avg: Average

The optimum latent factor (k) values and the error rates obtained from prediction matrices when run with these k values with the PPI matrix added are given in **Table 4.19.** The k values were determined as 50 for CC and MF, 150 for BP and disease, and 500 for protein data types. The average of error rates obtained by running the algorithm with these latent factor values is calculated as 0.037.

Table 4.19. The best latent factor (k) value set and the resulting lowest error rates (with PPI matrix).

	k1 (CC)	k2 (Prot)	k3 (MF)	k4 (BP)	k5 (Dis)
k3	50	500	50	150	150
	R12	**R13**	**R14**	**R15**	**Avg error**
error	0.055046	0.037071	0.028604	0.028466	0.037297

CC: Cellular Component, Prot: Protein, MF: Molecular Function, BP: Biological Process, Dis: Disease
Avg: Average

After the determination of latent factor (k) values, 10-fold cross-validation was applied and the performance scores were obtained. The performance scores of models without and with the PPI matrix added to the algorithm at threshold=0.5 are available in **Table 4.20** and **Table 4.22**. The best thresholds and their scores for the NMTF algorithms without and with PPI matrix added to the algorithm are given in **Table 4.21.** and **Table 4.23**, respectively, for each relation matrix.

For both experiments (i.e. without and with PPI matrix) the thresholds were taken as 0.5 when showing the results of the algorithm. The MCC scores for Protein x CC matrix are close for without and with the addition of PPI matrix to the algorithm, but it is observed a little lower for the model with PPI matrix (0.44) than the model without it (0.47). The MCC scores of Protein x MF are 0.39 in the model without PPI matrix, and 0.35 with PPI matrix ; 0.23 and 0.18 for Protein x BP matrix, and 0.20 and 0.17 for Protein x Disease matrix, with respective to PPI matrix addition. The full version of scores are available in **Appendix A.3** and **Appendix A.4.**

The ROC curves and the AUC values of NMTF algorithm run without PPI matrix can be examined in **Figure 4.5.** The AUC scores for Protein x CC, Protein x MF, Protein x BP and Protein x Disease matrices are 0.96, 0.91, 0.85 and 0.82, respectively. Likewise, the ROC curves and the AUC scores for the algorithm run with PPI matrix are available in **Figure 4.6.** The AUC score for Protein x CC matrix is 0.95 while it is 0.90 for Protein x MF matrix, 0.84 for Protein x BP matrix and 0.79 for Protein x Disease matrix.

Table 4.20. The performance at threshold=0.5 for NMTF algorithm without PPI matrix for 10-fold cross-validation.

	Thr.	TP	TN	FP	FN	Rec.	Prec.	FPR	Acc.	F-sc	MCC
PxCC	0.5	19383	52117	83	32817	0.37	1.00	0.00	0.68	0.54	**0.47**
PxMF	0.5	9323	34230	40	24947	0.27	1.00	0.00	0.64	0.43	**0.39**
PxBP	0.5	7660	75862	18	68220	0.10	1.00	0.00	0.55	0.18	**0.23**
PxD	0.5	5627	71358	12	65743	0.08	1.00	0.00	0.54	0.15	**0.20**

PxCC: Protein x CC, PxMF: Protein x MF, PxBP: Protein x BP, PxD: Protein x Disease, Thr.: Threshold, Rec.: Recall, Prec.: Precision, Acc: Accuracy, F-sc.: F-score

Table 4.21. The thresholds of best performance for NMTF algorithm without PPI matrix for 10-fold cross-validation.

	Thr.	TP	TN	FP	FN	Rec.	Prec.	FPR	Acc.	F-sc	MCC
PxCC	**0.03**	46171	47772	4428	6029	0.88	0.91	0.08	0.90	**0.90**	**0.80**
PxMF	**0.03**	27061	30346	3924	7209	0.79	0.87	0.11	0.83	0.83	**0.68**
PxBP	**0.02**	52717	64945	10935	23163	0.69	0.83	0.14	0.78	0.76	**0.56**
PxD	**0.02**	46929	60903	10467	24441	0.66	0.82	0.15	0.76	0.73	**0.52**

PxCC: Protein x CC, PxMF: Protein x MF, PxBP: Protein x BP, PxD: Protein x Disease, Thr.: Threshold, Rec.: Recall, Prec.: Precision, Acc: Accuracy, F-sc.: F-score

Table 4.22. The performance at threshold=0.5 for NMTF algorithm with PPI matrix for 10-fold cross-validation.

	Thr.	TP	TN	FP	FN	Rec.	Prec.	FPR	Acc.	F-sc	MCC
PxCC	0.5	17287	52078	122	34913	0.33	0.99	0.00	0.66	0.50	**0.44**
PxMF	0.5	7745	34237	33	26525	0.23	1.00	0.00	0.61	0.37	**0.35**
PxBP	0.5	4717	75867	13	71163	0.06	1.00	0.00	0.53	0.12	**0.18**
PxD	0.5	3900	71358	12	67470	0.05	1.00	0.00	0.53	0.10	**0.17**

PxCC: Protein x CC, PxMF: Protein x MF, PxBP: Protein x BP, PxD: Protein x Disease, Thr.: Threshold, Rec.: Recall, Prec.: Precision, Acc: Accuracy, F-sc.: F-score

Table 4.23. The thresholds of best performance for NMTF algorithm with PPI matrix for 10-fold cross-validation.

	Thr.	TP	TN	FP	FN	Rec.	Prec.	FPR	Acc.	F-sc	MCC
PxCC	**0.03**	46497	46821	5379	5703	0.89	0.90	0.10	0.89	**0.89**	**0.79**
PxMF	**0.03**	26092	30181	4089	8178	0.76	0.86	0.12	0.82	0.81	**0.65**
PxBP	**0.02**	51318	64365	11515	24562	0.68	0.82	0.15	0.76	0.74	**0.53**
PxD	**0.02**	43079	61017	10353	28291	0.60	0.81	0.15	0.73	0.69	**0.47**

PxCC: Protein x CC, PxMF: Protein x MF, PxBP: Protein x BP, PxD: Protein x Disease, Thr.: Threshold, Rec.: Recall, Prec.: Precision, Acc: Accuracy, F-sc.: F-score

a.
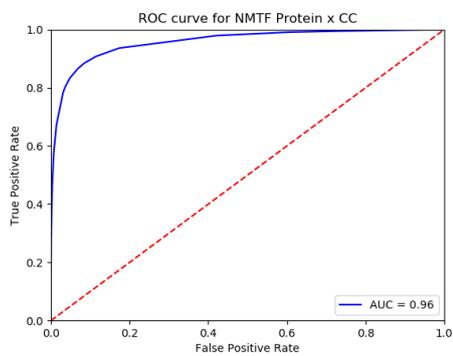
b.
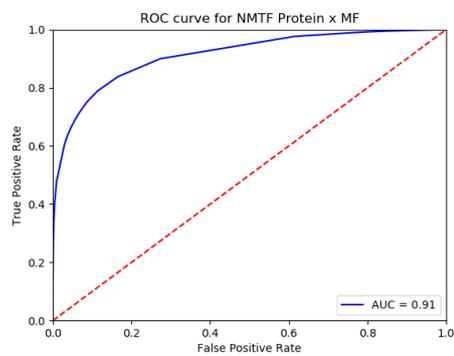
c.
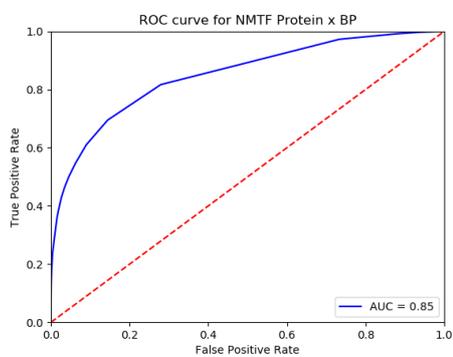
d.
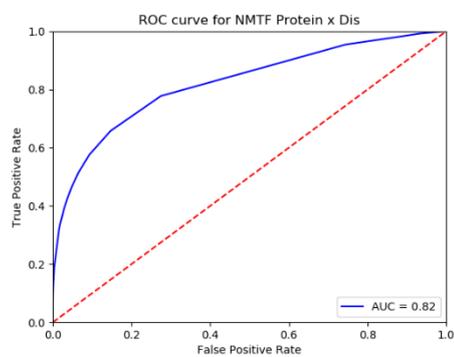
Figure 4.5. The ROC curves and AUC scores of matrices for the NMTF model without PPI matrix.

a.

b.

c.

d.
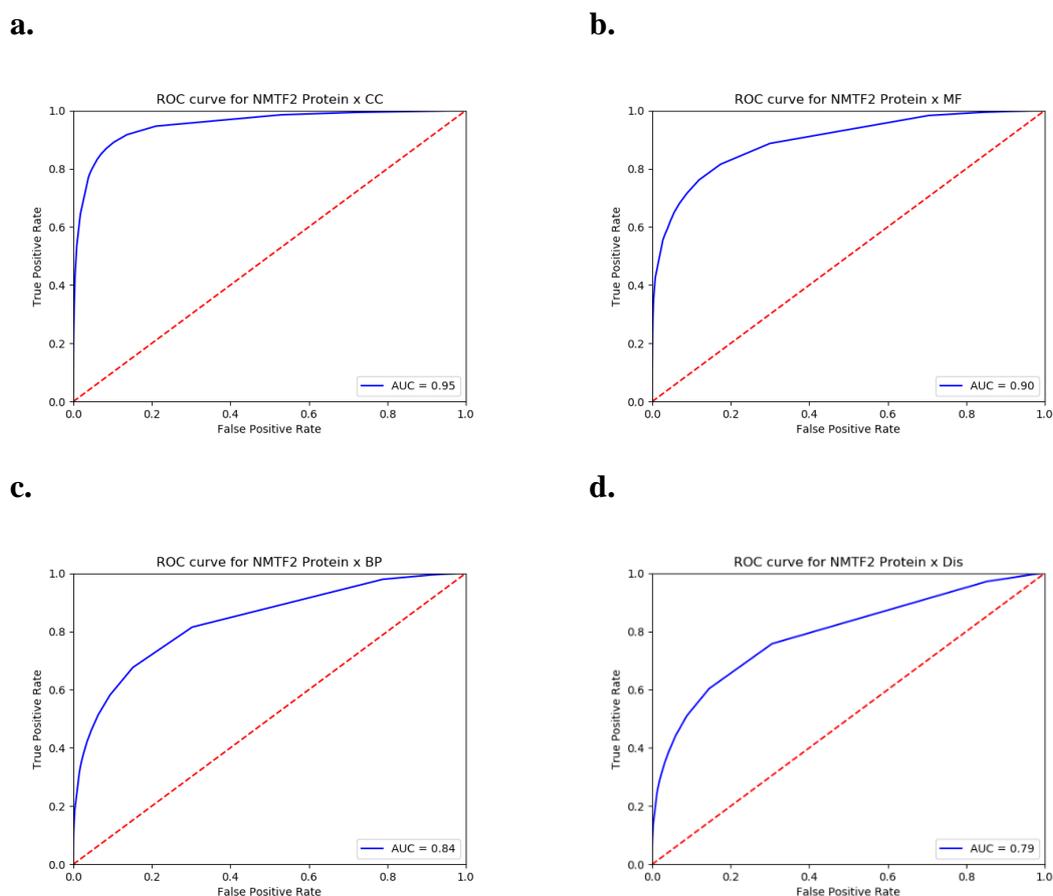
Figure 4.6. The ROC curves and AUC scores of matrices for the NMTF model with PPI matrix.

NMTF algorithm was also re-applied to the new dataset, and evaluated by 3-fold cross-validation for aforementioned reasons. The results of NMTF models with and without the PPI matrix introduced to the algorithm is available in **Table 4.24, Table 4.25, Table 4.26** and **Table 4.27**, for thresholds of 0.5 and the ones giving the best scores. The k values in this experiment were 150 for CC and MF, 200 for BP and Disease, and 2000 for Protein data types.

Table 4.24. The performance at threshold=0.5 for NMTF algorithm without PPI matrix for 3-fold cross-validation.

	Thr.	TP	TN	FP	FN	Rec.	Prec.	FPR	Acc.	F-sc	MCC
PxCC	0.5	7956	37279	20	26184	0.23	1.00	0.00	0.63	0.38	0.37
PxMF	0.5	4084	25432	11	18525	0.18	1.00	0.00	0.61	0.31	0.32
PxBP	0.5	3034	59215	5	52925	0.05	1.00	8E-5	0.54	0.10	0.17
PxD	0.5	3082	65778	6	60039	0.05	1.00	9E-5	0.53	0.09	0.16

PxCC: Protein x CC, PxMF: Protein x MF, PxBP: Protein x BP, PxD: Protein x Disease, Thr.: Threshold, Rec.: Recall, Prec.: Precision, Acc: Accuracy, F-sc.: F-score

Table 4.25. The thresholds of best performance for NMTF algorithm without PPI matrix for 3-fold cross-validation.

	Thr.	TP	TN	FP	FN	Rec.	Prec.	FPR	Acc.	F-sc	MCC
PxCC	**0.03**	28348	34510	2789	5792	0.83	0.91	0.07	0.88	0.87	0.76
PxMF	**0.02**	18048	21620	3823	4561	0.80	0.83	0.15	0.83	0.81	0.65
PxBP	**0.02**	35191	51781	7439	20768	0.63	0.83	0.13	0.76	0.71	0.52
PxD	**0.02**	39638	56024	9760	23483	0.63	0.80	0.15	0.74	0.70	0.49

PxCC: Protein x CC, PxMF: Protein x MF, PxBP: Protein x BP, PxD: Protein x Disease, Thr.: Threshold, Rec.: Recall, Prec.: Precision, Acc: Accuracy, F-sc.: F-score

Table 4.26. The performance at threshold=0.5 for NMTF algorithm with PPI matrix for 3-fold cross-validation.

|	Thr.	TP	TN	FP	FN	Rec.	Prec.	FPR	Acc.	F-sc	MCC
PxCC	0.5	8090	37279	20	26050	0.24	1.00	0.00	0.64	0.38	0.37
PxMF	0.5	4681	25429	14	17928	0.21	1.00	0.00	0.63	0.34	0.35
PxBP	0.5	3127	59217	3	52832	0.06	1.00	5E-05	0.54	0.11	0.17
PxD	0.5	3057	65781	3	60064	0.05	1.00	5E-05	0.53	0.09	0.16

PxCC: Protein x CC, PxMF: Protein x MF, PxBP: Protein x BP, PxD: Protein x Disease, Thr.: Threshold, Rec.: Recall, Prec.: Precision, Acc: Accuracy, F-sc.: F-score

Table 4.27. The thresholds of best performance for NMTF algorithm with PPI matrix for 3-fold cross-validation.

	Thr.	TP	TN	FP	FN	Rec.	Prec.	FPR	Acc.	F-sc	MCC
PxCC	**0.03**	28646	34424	2875	5494	0.84	0.91	0.08	0.88	0.87	0.77
PxMF	**0.02**	18820	21518	3925	3789	0.83	0.83	0.15	0.84	0.83	0.68
PxBP	**0.02**	35771	51678	7542	20188	0.64	0.83	0.13	0.76	0.72	0.53
PxD	**0.02**	39036	55891	9893	24085	0.62	0.80	0.15	0.74	0.70	0.48

PxCC: Protein x CC, PxMF: Protein x MF, PxBP: Protein x BP, PxD: Protein x Disease, Thr.: Threshold, Rec.: Recall, Prec.: Precision, Acc: Accuracy, F-sc.: F-score

4.4. Performance Comparison Between Different Algorithms

For more clear comparison, the AUC values of all models according to the relation matrices are given in **Table 4.28.** Likewise, the comparison of accuracy, F and MCC scores are given in **Table 4.29**, **4.30** and **4.31**, respectively for the threshold of 0.5. There exists more than one score for Protein x Disease matrix run with HNMF model for each table; however, the best of the three is given in the table of scores below. The accuracy, F-score and MCC scores at threshold giving the best performance scores are given in **Table 4.32**, **Table 4.33** and **Table 4.34**.

Table 4.28. AUC values of all the models performed in this study.

	NMF	HNMF	NMTF (without PPI matrix)	NMTF (with PPI matrix)
Protein x CC	0.94	0.88	0.96	0.95
Protein x MF	0.92	0.86	0.91	0.90
Protein x BP	0.93	0.87	0.85	0.84
Protein x Dis	0.92	0.87	0.82	0.79

Table 4.29. Accuracy scores of all models performed in this study (at threshold=0.5).

	NMF	HNMF	NMTF (without PPI matrix)	NMTF (with PPI matrix)
Protein x CC	0.56	0.67	0.68	0.66
Protein x MF	0.58	0.62	0.64	0.61
Protein x BP	0.59	0.59	0.55	0.53
Protein x Dis	0.64	0.64	0.54	0.53

Table 4.30. F-scores of all models performed in this study (at threshold=0.5).

	NMF	HNMF	NMTF (without PPI matrix)	NMTF (with PPI matrix)
Protein x CC	0.22	0.51	0.54	0.50
Protein x MF	0.27	0.40	0.43	0.37
Protein x BP	0.31	0.31	0.18	0.12
Protein x Dis	0.43	0.44	0.15	0.10

Table 4.31. MCC scores of all models performed in this study (at threshold=0.5).

	NMF	HNMF	NMTF (without PPI matrix)	NMTF (with PPI matrix)
Protein x CC	0.26	0.45	0.47	0.44
Protein x MF	0.29	0.37	0.39	0.35
Protein x BP	0.32	0.32	0.23	0.18
Protein x Dis	0.40	0.40	0.20	0.17

Table 4.32. Best Accuracy scores of all models performed in this study.

	Thr.	NMF	Thr.	HNMF	Thr.	NMTF (without PPI matrix)	Thr.	NMTF (with PPI matrix)
Protein x CC	0.02	0.88	0.02	0.87	0.03	0.90	0.03	0.89
Protein x MF	0.02	0.87	0.02	0.85	0.03	0.84	0.03	0.82
Protein x BP	0.02	0.86	0.02	0.86	0.02	0.78	0.02	0.76
Protein x Dis	0.02	0.86	0.02	0.85	0.02	0.76	0.02	0.73

Thr.:Threshold

Table 4.33. Best F-scores of all models performed in this study.

	Thr.	NMF	Thr.	HNMF	Thr.	NMTF (without PPI matrix)	Thr.	NMTF (with PPI matrix)
Protein x CC	0.02	0.88	0.02	0.85	0.03	0.90	0.03	0.89
Protein x MF	0.02	0.87	0.02	0.83	0.03	0.84	0.03	0.82
Protein x BP	0.02	0.86	0.02	0.85	0.02	0.78	0.02	0.77
Protein x Dis	0.02	0.86	0.02	0.84	0.02	0.76	0.02	0.73

Thr.:Threshold

Table 4.34. Best MCC scores of all models performed in this study.

	Thr.	NMF	Thr.	HNMF	Thr.	NMTF (without PPI matrix)	Thr.	NMTF (with PPI matrix)
Protein x CC	0.02	0.75	0.02	0.75	0.03	0.80	0.03	0.79
Protein x MF	0.02	0.75	0.02	0.71	0.03	0.68	0.03	0.65
Protein x BP	0.02	0.73	0.02	0.73	0.02	0.56	0.02	0.53
Protein x Dis	0.02	0.73	0.02	0.72	0.02	0.52	0.02	0.47

Thr.:Threshold

4.5. Computation Time Comparison Between Different Algorithms

For the comparison, the running-time of each algorithm for 3-fold cross-validation is given in **Table 4.35**. The time given is in hours. For the baseline NMF algorithm, the times are given for each matrix. For HNMF algorithm the times are given for Protein x CC and Protein x Disease, Protein x MF and Protein x Disease and Protein x BP and Protein x Disease models. Their running time is given in the row for the matrix inserted to the model with Protein x Disease matrix. For example, the running time of the Protein x CC and Protein x Disease model of HNMF algorithm is given in Protein x CC row. For this reason, the Protein x Disease row is empty for HNMF algorithm. Since NMTF algorithm allows all relation matrices to be inserted simultaneously, there is no different run times for each matrix. Instead, one run-time is given for each NMTF application with and without PPI matrix. The baseline NMF algorithm only took seconds, while each NMTF algorithm run in around 4 hours. On the other hand, the HNMF applications lasted the longest to finished. The running time of the HNMF algorithm was about 8 hours for Protein x CC and Protein x Disease model, and close to 12 hours for Protein x MF and Protein x Disease model. On the other hand, the Protein x BP and Protein x Disease model took days to finish. The model took more than 109 hours to finish. All algorithms were run in a computer with Intel® Core™ i7-8750H CPU @ 2.20 GHz, 16 GB RAM and in Windows 10 64-bit operating system.

Table 4.35. Running time of each algorithm.

	NMF	HNMF	NMTF (without PPI matrix)	NMTF (with PPI matrix)
Protein x CC	0.0021	8.1	4.21	4.64
Protein x MF	0.0022	11.65		
Protein x BP	0.0027	109.2		
Protein x Dis	0.0031			

CHAPTER 5

5. DISCUSSION AND CONCLUSION

Recent technological developments in many different scientific fields resulted in a dramatical increase in the size of the produced biological data, making the manual review and curation processes almost an impossible task. Automatic annotation systems have become crucial in order to ease the workload of manual curators. In this study, we investigated different nonnegative matrix factorization (NMF) algorithms for large-scale biological data integration and relation prediction, where new relation predictions were obtained by only using the existing relations between protein/gene vs. functions and proteins/genes vs. diseases.

The hypothesis of this study was that the performance of the system could be improved by inserting more relational data to the system, for it to better learn the similarities between entities. For this, first of all, each relational data matrix was inserted to the baseline NMF algorithm separately, then stepwise integration of other relation data was achieved with HNMF and NMTF algorithms. HNMF algorithm was able to use only two of relation data at the same time, while NMTF allowed all relation data to be added simultaneously. NMTF also provided the opportunity to insert PPI information to the algorithm, with the aim of improving the quality and quantity of the predictions. We compared the results of the baseline NMF, HNMF and NMTF (with and without PPI information). While comparing the performances, the scores were considered for the binary classification threshold of 0.5. Actually, as can be seen in confusion matrices given in **Appendix A.1** and other sections of **Appendix A,** the optimal thresholds providing the best performance scores change for each application. However, for the comparison to be fair, a standard threshold should have been selected, and it was chosen as 0.5 (the midpoint between 0 and 1, which are the minimum and maximum values), as in most of the studies in the literature. Even though the models seemed to be performing lower for the threshold of 0.5, we were mostly interested in their comparison with each other. The algorithms were firstly evaluated by 10-fold cross-validation. We then evaluated the algorithms by 3-fold cross-validation, since we suspected of overfitting of the models in some cases (e.g. BP and disease matrix factorizations), when we examined the results. The trend in results were similar when the scores for thresholds of 0.5 and 0.02 (the mostly optimal threshold) are examined. Another purpose of 3-fold cross-validation was to perform a more challenging test for the models to be able to clearly observe the separation in performance.

When we compared the NMF algorithms, generally it can be said that the baseline NMF method displayed the worst performance, while the more complicated NMF

methods such as HNMF and NMTF performed significantly better. It is considered that this outcome was obtained dues to the use of multiple relations between several data types during factorization, unlike the baseline NMF algorithm, which uses the relation between only two different data types. If the results of HNMF and NMTF are to be compared, it was observed that the performance of HNMF and NMTF algorithms are close to each other; although the HNMF algorithm performed slightly better in general. It was expected for the NMTF algorithm to perform better since it integrates all relation matrices simultaneously. It is believed that, the opposite was occurred because the processed data was quite incomplete, noisy and heterogeneous, resulting in the learning process not to be as good as expected. When the contribution of PPI matrix was examined, it was observed that the addition of PPI information resulted in slightly better scores. This was not observed when the results of 10-fold cross-validation were examined because there was a technical error while constructing the PPI information matrices in that analysis. This error has been eliminated when the datasets were reconstructed to perform the 3-fold cross-validation. As a result, 3-fold cross-validation results should be taken into account while evaluating the contribution of PPI information. The increase in the performance with the inclusion of PPI information is thought to be the contribution of an extra source of data.

The HNMF algorithm was observed to performed slightly better when Protein x Disease was factorized together with Protein x Cellular Component and Protein x Molecular Function, compared to the model with Protein x Biological Process. However, we expected the model with BP data type to produce better scores since BP is more closely related to the occurrence of diseases. Since, the occurrence of diseases is directly related to the disruptions in biological processes. We believe that the reason is again the heterogeneity and noisiness of the BP data, resulting in bias while the model learns the features from the data during the factorization process.

For the relation matrices with CC and MF, the best performances were obtained by the NMTF algorithm, with a slight improvement with the addition of PPI information. This is believed to be the outcome of NMTF using all relations simultaneously, providing better a extraction of the latent features of proteins, cellular components and molecular functions. However, this was not the case when the scores of biological process and disease relation models were examined. This may be resulted from overfitting of the models, since biological process and disease relations are much noisier and more heterogeneous.

Protein function prediction problem has been a struggle for some time for the scientific community. One of the projects that address this problem is the CAFA challenge. CAFA (abbr. for Critical Assessment of Function Annotation) is an ongoing experiment for assessing protein function prediction methods via computational algorithms. In CAFA publications, the F-scores for protein - molecular function relation prediction is around 0.7, and even lower for other relations such as cellular components and biological processes (Dessimoz et al., 2013; Jiang et al., 2016; Radivojac et al., 2013; Zhou et al., 2019). The results we obtained for these relations

are similar to the ones found in CAFA, especially for MF and CC prediction. Considering, Protein x BP and Protein x Disease relations, our performance was significantly lower. One of the underlying reasons for this can be that, the available Protein x Disease relation data is not clean, since it includes many false positives.

In general, it is possible to say that, NMF is not an optimal approach for multiple relation prediction of biological data, especially when the data is highly incomplete (i.e., low number of known relations), since the only available information for the algorithm to learn the features are these relations. There could also be some additional reasons for the observed low performance in some cases. One of the reasons might be the algorithms relying on random factors. In NMF algorithms, as in some other machine learning approaches, optimization is to be achieved via starting from a random point, followed by an approximation to reach the optimum point. In cases where starting point is not efficient, lower performance values can be observed by getting stuck into local minima. This problem might be further promoted by the addition of more types of relational data to the system, where the search space for the optimum point is expanded.

Another disadvantage in this study was the size of the data. Even with the data filtering operations, complex algorithms took extremely long times to run. When we compared the run times of 3 different algorithms for 3-fold cross-validation, we observed that the baseline NMF algorithm took only seconds to finish, regardless of the matrix type. Both NMTF algorithms (with and without PPI matrix) took a little more than 4 hours. However, the HNMF algorithm took the most time to finish for all 3 models. Especially, the Protein x BP and Protein x Disease model of HNMF algorithm took close to a week to finalize, since it was the most complex pair of the matrices. In general, the running time increased as the complexity of the model increased. Although, the HNMF algorithm took much more time to finish compared to NMTF algorithm, which is more complex than HNMF algorithm. This might be resulted from the selection of stopping criteria of the algorithms. While the NMTF algorithm used maximum number of iterations to finish the run, HNMF used a convergence threshold to finalize. The running time of the NMTF algorithm may increase if a constant convergence value is selected as the stopping criteria, or even run longer than HNMF algorithm. In the end, running of the algorithms may be impossible to run as the size of the data increase and as more relations are added to the algorithms.

For future work, we plan to construct the Protein x Disease relation data categorically by separating it into multiple matrices according to disease types, to make it more homogeneous. Then these different groups of disease matrices may be inserted to the algorithms as separate relation matrices. In addition, we plan to insert additional relational data to the algorithm, such as GO semantic similarities and disease – disease interactions as new intra-type relations. We also intend to improve the PPI information used in the factorization by including second, third and fourth degree neighborhoods in the PPI matrix using different scoring schemes reflecting the interaction proximity on the network (i.e. 1 for first, 0.75 for second, 0.5 for third, 0.25 for fourth neighbors).

Another type of intra-type gene/protein information can be the gene co-expression profiles. Similar genes (proteins) tend to be expressed or silenced together in a biological processes (Stuart et al., 2003). As a result, co-expressed genes can be denoted by higher values in the gene-gene/protein-protein matrices. Another possible addition to this study as future work would be to embed the low-rank latent features of biological entities by dimensionality reduction, for visualization and data exploration. These features can be embedded using Principal Component Analysis (PCA) or with t-stochastic neighbor embedding (tSNE) methods to visualize the features' characterizations on a 2 or 3 dimensional plane (Abdi & Williams, 2010). This way, the nature and the biological relevance of the discovered latent vectors can be assessed from a general perspective (i.e. similar proteins from the same protein families should be embedded close to each other, or semantically similar GO terms should appear close to each other).

The approach proposed here only depends on known relations between the modeled biological entities, instead of molecular properties/features (e.g. sequence or structural information) as used in conventional machine learning based gene/protein annotation methods. As a result, we believe that our prediction results will be complementary to these conventional methods. Therefore, ensemble-based predictors that will both include NMF models and conventional machine learning models is expected to reach increased prediction performances. Another possibility would be the direct use of the latent vector representations of biological entities in machine learning based predictors, as input feature vectors. We hope that our study will contribute to the scientific literature with the results obtained and their discussion, in terms of modeling relational biological data.

REFERENCES

Abdi, H., & Williams, L. J. (2010). Principal component analysis. In *Wiley Interdisciplinary Reviews: Computational Statistics*. https://doi.org/10.1002/wics.101

Adhikari, V. K., Guo, Y., Hao, F., Varvello, M., Hilt, V., Steiner, M., & Zhang, Z. L. (2012). Unreeling netflix: Understanding and improving multi-CDN movie delivery. *Proceedings - IEEE INFOCOM*. https://doi.org/10.1109/INFCOM.2012.6195531

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., & Walter, P. (2008). Molecular Biology of the Cell, 5th edition by B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter. In *Biochemistry and Molecular Biology Education*. https://doi.org/10.1002/bmb.20192

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., & Sherlock, G. (2000). Gene ontology: Tool for the unification of biology. In *Nature Genetics*. https://doi.org/10.1038/75556

Berman, H. M., Battistuz, T., Bhat, T. N., Bluhm, W. F., Bourne, P. E., Burkhardt, K., Feng, Z., Gilliland, G. L., Iype, L., Jain, S., Fagan, P., Marvin, J., Padilla, D., Ravichandran, V., Schneider, B., Thanki, N., Weissig, H., Westbrook, J. D., & Zardecki, C. (2002). The protein data bank. *Acta Crystallographica Section D: Biological Crystallography*. https://doi.org/10.1107/S0907444902003451

Binns, D., Dimmer, E., Huntley, R., Barrell, D., O'Donovan, C., & Apweiler, R. (2009). QuickGO: A web-based tool for Gene Ontology searching. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btp536

Bottou, L. (2012). *Stochastic Gradient Descent Tricks*. https://doi.org/10.1007/978-3-642-35289-8_25

Bouchard, G., Guo, S., & Yin, D. (2013). Convex collective matrix factorization. *Journal of Machine Learning Research*, *31*, 144–152.

Cai, D., He, X., Wu, X., & Han, J. (2008). Non-negative matrix factorization on manifold. *Proceedings - IEEE International Conference on Data Mining, ICDM*. https://doi.org/10.1109/ICDM.2008.57

Ceddia, G., Pinoli, P., Ceri, S., & Masseroli, M. (2019). Non-negative Matrix Tri-Factorization for Data Integration and Network-based Drug Repositioning. *2019 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology, CIBCB 2019*. https://doi.org/10.1109/CIBCB.2019.8791474

Cherney, D., Denton, T., Thomas, R., & Waldron, A. (2013). Linear algebra. In *World*. https://doi.org/10.1007/978-88-470-1839-6

Cobanoglu, M. C., Liu, C., Hu, F., Oltvai, Z. N., & Bahar, I. (2013). Predicting drug-target interactions using probabilistic matrix factorization. *Journal of Chemical Information and Modeling*, *53*(12), 3399–3409. https://doi.org/10.1021/ci400219z

Dessimoz, C., Škunca, N., & Thomas, P. D. (2013). CAFA and the Open World of protein function predictions. In *Trends in Genetics*. https://doi.org/10.1016/j.tig.2013.09.005

Devarajan, K. (2008). Nonnegative matrix factorization: An analytical and interpretive tool in computational biology. *PLoS Computational Biology*, *4*(7). https://doi.org/10.1371/journal.pcbi.1000029

Dissez, G., Ceddia, G., Pinoli, P., Ceri, S., & Masseroli, M. (2019). Drug repositioning predictions by non-negative matrix tri-factorization of integrated association data. *ACM-BCB 2019 - Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, 25–33. https://doi.org/10.1145/3307339.3342154

Doolittle, R. F. (1985). Proteins. *Scientific American*. https://doi.org/10.1038/scientificamerican1085-88

Gaulton, A., Bellis, L. J., Bento, A. P., Chambers, J., Davies, M., Hersey, A., Light, Y., McGlinchey, S., Michalovich, D., Al-Lazikani, B., & Overington, J. P. (2012). ChEMBL: A large-scale bioactivity database for drug discovery. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkr777

Genetic Disorders | NHGRI. (n.d.). Retrieved January 23, 2020, from

https://www.genome.gov/For-Patients-and-Families/Genetic-Disorders

Golub, G. H., & Loan, C. F. Van. (1996). Matrix Computations (Johns Hopkins Studies in Mathematical Sciences)(3rd Edition). In *Linear Algebra and its Applications*. https://doi.org/10.1016/0024-3795(94)90446-4

Gönen, M. (2012). Predicting drug-target interactions from chemical and genomic kernels using Bayesian matrix factorization. *Bioinformatics*, *28*(18), 2304–2310. https://doi.org/10.1093/bioinformatics/bts360

Gordon, S. E., & Perugini, M. A. (2016). Protein-ligand interactions. In *Analytical Ultracentrifugation: Instrumentation, Software, and Applications*. https://doi.org/10.1007/978-4-431-55985-6_16

Hajian-Tilaki, K. (2013). Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation. In *Caspian Journal of Internal Medicine*.

Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A., & McKusick, V. A. (2005). Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gki033

Hennig, S., Groth, D., & Lehrach, H. (2003). Automated gene ontology annotation for anonymous sequence data. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkg582

Huntley, R. P., Sawford, T., Mutowo-Meullenet, P., Shypitsyna, A., Bonilla, C., Martin, M. J., & O'Donovan, C. (2015). The GOA database: Gene Ontology annotation updates for 2015. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gku1113

Jiang, Y., Oron, T. R., Clark, W. T., Bankapur, A. R., D'Andrea, D., Lepore, R., Funk, C. S., Kahanda, I., Verspoor, K. M., Ben-Hur, A., Koo, D. C. E., Penfold-Brown, D., Shasha, D., Youngs, N., Bonneau, R., Lin, A., Sahraeian, S. M. E., Martelli, P. L., Profiti, G., … Radivojac, P. (2016). An expanded evaluation of protein function prediction methods shows an improvement in accuracy. *Genome Biology*. https://doi.org/10.1186/s13059-016-1037-6

Koren, Y., Bell, R., & Volinsky, C. (2009). Matrix factorization techniques for recommender systems. *Computer*. https://doi.org/10.1109/MC.2009.263

Lang, S. (2013). Linear Algebra - Matrix calculus. *Methods in Molecular Biology (Clifton, N.J.)*. https://doi.org/10.1007/978-1-62703-059-5_19

Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, *401*(6755), 788–791. https://doi.org/10.1038/44565

Lee, D. D., & Seung, H. S. (2001). Algorithms for non-negative matrix factorization. *Advances in Neural Information Processing Systems*.

Lewis. (2005). Genes XI. In *Transfusion Medicine*. https://doi.org/10.1111/j.1365-3148.2005.00580.x

Lippert, C., Weber, S. H., Huang, Y., Tresp, V., Schubert, M., & Kriegel, H. (2008). *Relation Prediction in Multi-Relational Domains using Matrix Factorization*. *Siso*.

Liu, J. S., & Jiang, B. (2013). Statistical methods in bioinformatics. In *Basics of Bioinformatics: Lecture Notes of the Graduate Summer School on Bioinformatics of China* (Vol. 9783642389). https://doi.org/10.1007/978-3-642-38951-1_4

Luo, Y., Mao, C., Yang, Y., Wang, F., Ahmad, F. S., Arnett, D., Irvin, M. R., & Shah, S. J. (2019). Integrating hypertension phenotype and genotype with hybrid non-negative matrix factorization. *Bioinformatics*, *35*(8), 1395–1403. https://doi.org/10.1093/bioinformatics/bty804

Montani, D., Günther, S., Dorfmüller, P., Perros, F., Girerd, B., Garcia, G., Jaïs, X., Savale, L., Artaud-Macari, E., & Price, L. (2013). Orphanet Journal of Rare Diseases. *Orphanet Journal of Rare Diseases*.

Oxford. (n.d.). *Overfitting | Meaning of Overfitting by Lexico*. Retrieved February 23, 2020, from https://www.lexico.com/definition/overfitting

Pehkonen, P., Wong, G., & Törönen, P. (2005). Theme discovery from gene lists for identification and viewing of multiple functional groups. *BMC Bioinformatics*, *6*, 1–18. https://doi.org/10.1186/1471-2105-6-162

Piñero, J., Bravo, Á., Queralt-Rosinach, N., Gutiérrez-Sacristán, A., Deu-Pons, J., Centeno, E., García-García, J., Sanz, F., & Furlong, L. I. (2017). DisGeNET: A comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkw943

Piñero, J., Queralt-Rosinach, N., Bravo, À., Deu-Pons, J., Bauer-Mehren, A., Baron, M., Sanz, F., & Furlong, L. I. (2015). DisGeNET: A discovery platform for the dynamical exploration of human diseases and their genes. *Database*. https://doi.org/10.1093/database/bav028

Piñero, J., Ramírez-Anguita, J. M., Saüch-Pitarch, J., Ronzano, F., Centeno, E., Sanz, F., & Furlong, L. I. (2019). The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkz1021

Radivojac, P., Clark, W. T., Oron, T. R., Schnoes, A. M., Wittkop, T., Sokolov, A., Graim, K., Funk, C., Verspoor, K., Ben-Hur, A., Pandey, G., Yunes, J. M., Talwalkar, A. S., Repo, S., Souza, M. L., Piovesan, D., Casadio, R., Wang, Z., Cheng, J., … Friedberg, I. (2013). A large-scale evaluation of computational protein function prediction. *Nature Methods*. https://doi.org/10.1038/nmeth.2340

Ripley, L. S. (2013). Mutation. In *Brenner's Encyclopedia of Genetics: Second Edition*. https://doi.org/10.1016/B978-0-12-374984-0.01007-X

Sharan, R., Ulitsky, I., & Shamir, R. (2007). Network-based prediction of protein function. In *Molecular Systems Biology*. https://doi.org/10.1038/msb4100129

Singh, A. P., & Gordon, G. J. (2008). Relational Learning via Collective Matrix Factorization Categories and Subject Descriptors. *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 650–658. https://doi.org/10.1145/1401890.1401969

Skrbo, A., Begović, B., & Skrbo, S. (2004). Classification of drugs using the ATC system (Anatomic, Therapeutic, Chemical Classification) and the latest changes. *Medicinski Arhiv*.

Stanford. (2015). 4.2 Matrix Factorization: Objective and ALS Algorithm on a Single Machine. *Stanford Lecture*, *323*, 1–4.

Stuart, J. M., Segal, E., Koller, D., & Kim, S. K. (2003). A gene-coexpression network for global discovery of conserved genetic modules. *Science*. https://doi.org/10.1126/science.1087447

UniProt Consortium, T. (2018). UniProt: the universal protein knowledgebase. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gky092

William Burrows, & Dante G. Scarpelli. (2019). *disease | Definition, Types, & Control | Britannica*. Encyclopædia Britannica, Inc. https://www.britannica.com/science/disease/Metabolic-defects

William S. Klug, Michael R. Cummings, C. A. S. (2006). Conceptos de Genetica. In *Pearson*.

Wishart, D. S. (2012). DrugBank. In *Principles of Pharmacogenetics and Pharmacogenomics*. https://doi.org/10.1017/CBO9781139051194.008

Zheng, X., Ding, H., Mamitsuka, H., & Zhu, S. (2013). Collaborative matrix factorization with multiple similarities for predicting drug-Target interactions. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *Part F1288*, 1025–1033. https://doi.org/10.1145/2487575.2487670

Zhou, N., Jiang, Y., Bergquist, T. R., Lee, A. J., Kacsoh, B. Z., Crocker, A. W., Lewis, K. A., Georghiou, G., Nguyen, H. N., Hamid, M. N., Davis, L., Dogan, T., Atalay, V., Rifaioglu, A. S., Dalkıran, A., Cetin Atalay, R., Zhang, C., Hurto, R. L., Freddolino, P. L., … Friedberg, I. (2019). The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens. *Genome Biology*. https://doi.org/10.1186/s13059-019-1835-8

Žitnik, M., & Zupan, B. (2015). Data fusion by matrix factorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *37*(1), 41–53. https://doi.org/10.1109/TPAMI.2014.2343973

A.1. Confusion Matrices for Baseline NMF

A.1.1. Protein x Cellular Component

a. 10-fold cross-validation (k=50)

Thr.	TP	TN	FP	FN	Recall	Precision	FPR	Accuracy	F-score	MCC
0	52200	0	52200	0	1	0.5	1	0.5	0.67	
1E-06	52026	7805	44395	174	1.00	0.54	0.85	0.57	0.70	0.28
1E-05	51983	11469	40731	217	1.00	0.56	0.78	0.61	0.72	0.34
0.0001	51810	17569	34631	390	0.99	0.60	0.66	0.66	0.75	0.44
0.001	51134	27638	24562	1066	0.98	0.68	0.47	0.75	0.80	0.57
0.01	48077	42216	9984	4123	0.92	0.83	0.19	0.86	0.87	0.73
0.02	45403	46065	6135	6797	0.87	0.88	0.12	0.88	**0.88**	**0.75**
0.03	43064	47854	4346	9136	0.82	0.91	0.08	0.87	0.86	0.74
0.04	40854	48928	3272	11346	0.78	0.93	0.06	0.86	0.85	0.73
0.05	38847	49555	2645	13353	0.74	0.94	0.05	0.85	0.83	0.71
0.06	37168	50031	2169	15032	0.71	0.94	0.04	0.84	0.81	0.69
0.07	35752	50372	1828	16448	0.68	0.95	0.04	0.82	0.80	0.68
0.08	34449	50630	1570	17751	0.66	0.96	0.03	0.81	0.78	0.66
0.09	33244	50840	1360	18956	0.64	0.96	0.03	0.81	0.77	0.65
0.1	32024	51029	1171	20176	0.61	0.96	0.02	0.80	0.75	0.63
0.2	23012	51782	418	29188	0.44	0.98	0.01	0.72	0.61	0.52
0.3	17686	52007	193	34514	0.34	0.99	0.00	0.67	0.50	0.44
0.4	13080	52128	72	39120	0.25	0.99	0.00	0.62	0.40	0.38
0.5	6558	52170	30	45642	0.13	1.00	0.00	0.56	0.22	0.26
0.6	3550	52191	9	48650	0.07	1.00	0.00	0.53	0.13	0.19
0.7	1329	52198	2	50871	0.03	1.00	0.00	0.51	0.05	0.11
0.8	377	52200	0	51823	0.01	1	0	0.50	0.01	0.06
0.9	88	52200	0	52112	0.00	1	0	0.50	0.00	0.03
1	38	52200	0	52162	0.00	1	0	0.50	0.00	0.02

Thr: Threshold

b.3-fold cross-validation (k=150)

Thr.	TP	TN	FP	FN	Recall	Precision	FPR	Accuracy	F-score	MCC
0	34140	0	37299	0	1	0.48	1	0.48	0.65	
1E-06	34094	6025	31274	46	1.00	0.52	0.84	0.56	0.69	0.29
1E-05	33988	9856	27443	152	1.00	0.55	0.74	0.61	0.71	0.37
0.0001	33669	15896	21403	471	0.99	0.61	0.57	0.69	0.75	0.49
0.001	31884	24537	12762	2256	0.93	0.71	0.34	0.79	0.81	0.61
0.01	24626	33443	3856	9514	0.72	0.86	0.10	0.81	0.79	0.63
0.02	18612	35277	2022	15528	0.55	0.90	0.05	0.75	0.68	0.54
0.03	13876	36045	1254	20264	0.41	0.92	0.03	0.70	0.56	0.46
0.04	11108	36408	891	23032	0.33	0.93	0.02	0.67	0.48	0.40
0.05	9410	36639	660	24730	0.28	0.93	0.02	0.64	0.43	0.37
0.06	8227	36774	525	25913	0.24	0.94	0.01	0.63	0.38	0.35
0.07	7531	36878	421	26609	0.22	0.95	0.01	0.62	0.36	0.33
0.08	6982	36938	361	27158	0.20	0.95	0.01	0.61	0.34	0.32
0.09	6632	36979	320	27508	0.19	0.95	0.01	0.61	0.32	0.31
0.1	6348	37023	276	27792	0.19	0.96	0.01	0.61	0.31	0.31
0.2	4799	37198	101	29341	0.14	0.98	0.00	0.59	0.25	0.27
0.3	3791	37250	49	30349	0.11	0.99	0.00	0.57	0.20	0.24
0.4	2843	37283	16	31297	0.08	0.99	0.00	0.56	0.15	0.21
0.5	1120	37297	2	33020	0.03	1.00	0.00	0.54	0.06	0.13
0.6	478	37297	2	33662	0.01	1.00	0.00	0.53	0.03	0.09
0.7	91	37298	1	34049	0.00	0.99	0.00	0.52	0.01	0.04
0.8	18	37299	0	34122	0.00	1	0	0.52	0.00	0.02
0.9	4	37299	0	34136	0.00	1	0	0.52	0.00	0.01
1	0	37299	0	34140	0		0	0.52		

Thr: Threshold

A.1.2. Protein x Molecular Function

a. 10-fold cross-validation (k=50)

Thr.	TP	TN	FP	FN	Recall	Precision	FPR	Accuracy	F-score	MCC
0	34270	0	34270	0	1	0.5	1	0.5	0.67	
1E-06	33671	5024	29246	599	0.98	0.54	0.85	0.56	0.69	0.24
1E-05	33608	6275	27995	662	0.98	0.55	0.82	0.58	0.70	0.27
0.0001	33476	8972	25298	794	0.98	0.57	0.74	0.62	0.72	0.34
0.001	32829	15826	18444	1441	0.96	0.64	0.54	0.71	0.77	0.48
0.01	30392	28185	6085	3878	0.89	0.83	0.18	0.85	0.86	0.71
0.02	29022	30890	3380	5248	0.85	0.90	0.10	0.87	**0.87**	0.75
0.03	27738	32042	2228	6532	0.81	0.93	0.07	0.87	0.86	**0.75**
0.04	26439	32607	1663	7831	0.77	0.94	0.05	0.86	0.85	0.73
0.05	25127	32951	1319	9143	0.73	0.95	0.04	0.85	0.83	0.71
0.06	23864	33160	1110	10406	0.70	0.96	0.03	0.83	0.81	0.69
0.07	22709	33322	948	11561	0.66	0.96	0.03	0.82	0.78	0.67
0.08	21668	33430	840	12602	0.63	0.96	0.02	0.80	0.76	0.65
0.09	20831	33534	736	13439	0.61	0.97	0.02	0.79	0.75	0.63
0.1	20105	33629	641	14165	0.59	0.97	0.02	0.78	0.73	0.62
0.2	14822	34028	242	19448	0.43	0.98	0.01	0.71	0.60	0.51
0.3	11759	34160	110	22511	0.34	0.99	0.00	0.67	0.51	0.45
0.4	9081	34217	53	25189	0.26	0.99	0.00	0.63	0.42	0.39
0.5	5305	34250	20	28965	0.15	1.00	0.00	0.58	0.27	0.29
0.6	3057	34261	9	31213	0.09	1.00	0.00	0.54	0.16	0.22
0.7	1543	34269	1	32727	0.05	1.00	0.00	0.52	0.09	0.15
0.8	562	34269	1	33708	0.02	1.00	0.00	0.51	0.03	0.09
0.9	185	34270	0	34085	0.01	1	0	0.50	0.01	0.05
1	56	34270	0	34214	0.00	1	0	0.50	0.00	0.03

Thr: Threshold

b.3-fold cross-validation (k=150)

Thr.	TP	TN	FP	FN	Recall	Precision	FPR	Accuracy	F-score	MCC
0	22609	0	25443	0	1	0.47	1	0.47	0.64	
1E-06	22488	4351	21092	121	0.99	0.52	0.83	0.56	0.68	0.28
1E-05	22405	5777	19666	204	0.99	0.53	0.77	0.59	0.69	0.33
0.0001	22033	9280	16163	576	0.97	0.58	0.64	0.65	0.72	0.42
0.001	20556	15926	9517	2053	0.91	0.68	0.37	0.76	0.78	0.55
0.01	16560	22774	2669	6049	0.73	0.86	0.10	0.82	0.79	0.64
0.02	14030	24075	1368	8579	0.62	0.91	0.05	0.79	0.74	0.61
0.03	11601	24601	842	11008	0.51	0.93	0.03	0.75	0.66	0.55
0.04	10139	24855	588	12470	0.45	0.95	0.02	0.73	0.61	0.51
0.05	9302	25005	438	13307	0.41	0.96	0.02	0.71	0.58	0.49
0.06	8748	25088	355	13861	0.39	0.96	0.01	0.70	0.55	0.48
0.07	8352	25157	286	14257	0.37	0.97	0.01	0.70	0.53	0.47
0.08	8023	25184	259	14586	0.35	0.97	0.01	0.69	0.52	0.46
0.09	7688	25223	220	14921	0.34	0.97	0.01	0.68	0.50	0.45
0.1	7419	25250	193	15190	0.33	0.97	0.01	0.68	0.49	0.44
0.2	5780	25369	74	16829	0.26	0.99	0.00	0.65	0.41	0.39
0.3	4603	25405	38	18006	0.20	0.99	0.00	0.62	0.34	0.34
0.4	3347	25424	19	19262	0.15	0.99	0.00	0.60	0.26	0.29
0.5	1575	25439	4	21034	0.07	1.00	0.00	0.56	0.13	0.19
0.6	730	25443	0	21879	0.03	1	0	0.54	0.06	0.13
0.7	180	25443	0	22429	0.01	1	0	0.53	0.02	0.07
0.8	44	25443	0	22565	0.00	1	0	0.53	0.00	0.03
0.9	16	25443	0	22593	0.00	1	0	0.53	0.00	0.02
1	3	25443	0	22606	0.00	1	0	0.53	0.00	0.01

Thr: Threshold

A.1.3. Protein x Biological Process (k=100)

a. 10-fold cross-validation

Thr.	TP	TN	FP	FN	Recall	Precision	FPR	Accuracy	F-score	MCC
0	75880	0	75880	0	1	0.5	1	0.5	0.67	
1E-06	75446	10818	65062	434	0.99	0.54	0.86	0.57	0.70	0.26
1E-05	75314	16025	59855	566	0.99	0.56	0.79	0.60	0.71	0.33
0.0001	74904	25194	50686	976	0.99	0.60	0.67	0.66	0.74	0.42
0.001	73652	40312	35568	2228	0.97	0.67	0.47	0.75	0.80	0.56
0.01	68298	61029	14851	7582	0.90	0.82	0.20	0.85	0.86	0.71
0.02	64165	66833	9047	11715	0.85	0.88	0.12	0.86	0.86	0.73
0.03	60724	69540	6340	15156	0.80	0.91	0.08	0.86	0.85	0.72
0.04	57709	71144	4736	18171	0.76	0.92	0.06	0.85	0.83	0.71
0.05	55050	72143	3737	20830	0.73	0.94	0.05	0.84	0.82	0.69
0.06	52615	72828	3052	23265	0.69	0.95	0.04	0.83	0.80	0.68
0.07	50552	73346	2534	25328	0.67	0.95	0.03	0.82	0.78	0.66
0.08	48551	73729	2151	27329	0.64	0.96	0.03	0.81	0.77	0.65
0.09	46769	74014	1866	29111	0.62	0.96	0.02	0.80	0.75	0.63
0.1	45116	74282	1598	30764	0.59	0.97	0.02	0.79	0.74	0.62
0.2	33243	75363	517	42637	0.44	0.98	0.01	0.72	0.61	0.52
0.3	26064	75683	197	49816	0.34	0.99	0.00	0.67	0.51	0.45
0.4	19958	75801	79	55922	0.26	1.00	0.00	0.63	0.42	0.39
0.5	14065	75853	27	61815	0.19	1.00	0.00	0.59	0.31	0.32
0.6	8915	75872	8	66965	0.12	1.00	0.00	0.56	0.21	0.25
0.7	3444	75878	2	72436	0.05	1.00	0.00	0.52	0.09	0.15
0.8	1490	75880	0	74390	0.02	1	0	0.51	0.04	0.10
0.9	666	75880	0	75214	0.01	1	0	0.50	0.02	0.07
1	251	75880	0	75629	0.00	1	0	0.50	0.01	0.04

Thr: Threshold

b.3-fold cross-validation (k=200)

Thr.	TP	TN	FP	FN	Recall	Precision	FPR	Accuracy	F-score	MCC
0	55959	0	59220	0	1	0.49	1	0.49	0.65	
1E-06	55775	9274	49946	184	1.00	0.53	0.84	0.56	0.69	0.28
1E-05	55613	13569	45651	346	0.99	0.55	0.77	0.60	0.71	0.34
0.0001	55137	20970	38250	822	0.99	0.59	0.65	0.66	0.74	0.43
0.001	53562	32586	26634	2397	0.96	0.67	0.45	0.75	0.79	0.55
0.01	47192	48316	10904	8767	0.84	0.81	0.18	0.83	0.83	0.66
0.02	42754	53021	6199	13205	0.76	0.87	0.10	0.83	0.82	0.67
0.03	39253	55190	4030	16706	0.70	0.91	0.07	0.82	0.79	0.65
0.04	36258	56330	2890	19701	0.65	0.93	0.05	0.80	0.76	0.63
0.05	33806	57046	2174	22153	0.60	0.94	0.04	0.79	0.74	0.61
0.06	31593	57506	1714	24366	0.56	0.95	0.03	0.77	0.71	0.59
0.07	29748	57841	1379	26211	0.53	0.96	0.02	0.76	0.68	0.57
0.08	28008	58089	1131	27951	0.50	0.96	0.02	0.75	0.66	0.55
0.09	26479	58266	954	29480	0.47	0.97	0.02	0.74	0.64	0.54
0.1	25190	58404	816	30769	0.45	0.97	0.01	0.73	0.61	0.52
0.2	17120	58985	235	38839	0.31	0.99	0.00	0.66	0.47	0.42
0.3	12444	59126	94	43515	0.22	0.99	0.00	0.62	0.36	0.35
0.4	8560	59179	41	47399	0.15	1.00	0.00	0.59	0.27	0.29
0.5	4786	59216	4	51173	0.09	1.00	7E-05	0.56	0.16	0.21
0.6	2529	59219	1	53430	0.05	1.00	2E-05	0.54	0.09	0.15
0.7	759	59220	0	55200	0.01	1	0	0.52	0.03	0.08
0.8	269	59220	0	55690	0.00	1	0	0.52	0.01	0.05
0.9	83	59220	0	55876	0.00	1	0	0.51	0.00	0.03
1	22	59220	0	55937	0.00	1	0	0.51	0.00	0.01

Thr: Threshold

A.1.4. Protein x Disease (k=100)

a. 10-fold cross-validation

Thr.	TP	TN	FP	FN	Recall	Precision	FPR	Accuracy	F-score	MCC
0	71370	0	71370	0	1	0.5	1	0.5	0.67	
1E-06	69464	15448	55922	1906	0.97	0.55	0.78	0.59	0.71	0.29
1E-05	69272	21083	50287	2098	0.97	0.58	0.70	0.63	0.73	0.36
0.0001	68789	29937	41433	2581	0.96	0.62	0.58	0.69	0.76	0.46
0.001	67399	42725	28645	3971	0.94	0.70	0.40	0.77	0.81	0.58
0.01	62879	58994	12376	8491	0.88	0.84	0.17	0.85	0.86	0.71
0.02	59794	63476	7894	11576	0.84	0.88	0.11	0.86	0.86	0.73
0.03	57197	65816	5554	14173	0.80	0.91	0.08	0.86	0.85	0.73
0.04	55014	67091	4279	16356	0.77	0.93	0.06	0.86	0.84	0.72
0.05	53157	68043	3327	18213	0.74	0.94	0.05	0.85	0.83	0.71
0.06	51415	68672	2698	19955	0.72	0.95	0.04	0.84	0.82	0.70
0.07	49823	69146	2224	21547	0.70	0.96	0.03	0.83	0.81	0.69
0.08	48285	69537	1833	23085	0.68	0.96	0.03	0.83	0.79	0.68
0.09	46897	69823	1547	24473	0.66	0.97	0.02	0.82	0.78	0.67
0.1	45625	70065	1305	25745	0.64	0.97	0.02	0.81	0.77	0.66
0.2	36202	71030	340	35168	0.51	0.99	0.00	0.75	0.67	0.58
0.3	30158	71265	105	41212	0.42	1.00	0.00	0.71	0.59	0.52
0.4	25266	71325	45	46104	0.35	1.00	0.00	0.68	0.52	0.46
0.5	19389	71356	14	51981	0.27	1.00	0.00	0.64	0.43	0.40
0.6	14424	71366	4	56946	0.20	1.00	0.00	0.60	0.34	0.34
0.7	9344	71368	2	62026	0.13	1.00	0.00	0.57	0.23	0.26
0.8	4003	71369	1	67367	0.06	1.00	0.00	0.53	0.11	0.17
0.9	1279	71370	0	70091	0.02	1	0	0.51	0.04	0.10
1	435	71370	0	70935	0.01	1	0	0.50	0.01	0.06

Thr: Threshold

b.3-fold cross-validation (k=200)

Thr	TP	TN	FP	FN	Recall	Precision	FPR	Accuracy	F-score	MCC
0	63121	0	65784	0	1	0.49	1	0.49	0.66	
1E-06	62767	11196	54588	354	0.99	0.53	0.83	0.57	0.70	0.29
1E-05	62422	16149	49635	699	0.99	0.56	0.75	0.61	0.71	0.35
0.0001	61665	23999	41785	1456	0.98	0.60	0.64	0.66	0.74	0.43
0.001	59505	36307	29477	3616	0.94	0.67	0.45	0.74	0.78	0.53
0.01	52857	52637	13147	10264	0.84	0.80	0.20	0.82	0.82	0.64
0.02	48616	57686	8098	14505	0.77	0.86	0.12	0.82	0.81	0.65
0.03	45371	60210	5574	17750	0.72	0.89	0.08	0.82	0.80	0.65
0.04	42819	61788	3996	20302	0.68	0.91	0.06	0.81	0.78	0.64
0.05	40587	62816	2968	22534	0.64	0.93	0.05	0.80	0.76	0.63
0.06	38545	63470	2314	24576	0.61	0.94	0.04	0.79	0.74	0.62
0.07	36857	63968	1816	26264	0.58	0.95	0.03	0.78	0.72	0.61
0.08	35221	64305	1479	27900	0.56	0.96	0.02	0.77	0.71	0.59
0.09	33806	64550	1234	29315	0.54	0.96	0.02	0.76	0.69	0.58
0.1	32577	64768	1016	30544	0.52	0.97	0.02	0.76	0.67	0.57
0.2	23688	65541	243	39433	0.38	0.99	0.00	0.69	0.54	0.48
0.3	18175	65710	74	44946	0.29	1.00	0.00	0.65	0.45	0.41
0.4	13716	65759	25	49405	0.22	1.00	0.00	0.62	0.36	0.35
0.5	8209	65778	6	54912	0.13	1.00	9E-05	0.57	0.23	0.27
0.6	4685	65784	0	58436	0.07	1	0	0.55	0.14	0.20
0.7	2071	65784	0	61050	0.03	1	0	0.53	0.06	0.13
0.8	618	65784	0	62503	0.01	1	0	0.52	0.02	0.07
0.9	161	65784	0	62960	0.00	1	0	0.51	0.01	0.04
1	59	65784	0	63062	0.00	1	0	0.51	0.00	0.02

Thr: Threshold

A.2. Confusion Matrices for HNMF

A.2.1. Protein x Cellular Component and Protein x Disease (k=150)

a. 10-fold cross-validation

1. Protein x Cellular Component

Thr.	TP	TN	FP	FN	Recall	Precision	FPR	Accuracy	F-score	MCC
0	52200	0	52200	0	1	0.5	1	0.5	0.67	
1E-06	40707	49590	2610	11493	0.78	0.94	0.05	0.86	0.85	0.74
1E-05	40707	49590	2610	11493	0.78	0.94	0.05	0.86	0.85	0.74
0.0001	40706	49594	2606	11494	0.78	0.94	0.05	0.86	0.85	0.74
0.001	40699	49609	2591	11501	0.78	0.94	0.05	0.87	0.85	0.74
0.01	40674	49753	2447	11526	0.78	0.94	0.05	0.87	0.85	0.74
0.02	40620	49881	2319	11580	0.78	0.95	0.04	0.87	0.85	0.75
0.03	40495	50009	2191	11705	0.78	0.95	0.04	0.87	0.85	0.75
0.04	40260	50145	2055	11940	0.77	0.95	0.04	0.87	0.85	0.75
0.05	39993	50291	1909	12207	0.77	0.95	0.04	0.86	0.85	0.74
0.06	39700	50435	1765	12500	0.76	0.96	0.03	0.86	0.85	0.74
0.07	39351	50566	1634	12849	0.75	0.96	0.03	0.86	0.84	0.74
0.08	39031	50674	1526	13169	0.75	0.96	0.03	0.86	0.84	0.74
0.09	38696	50803	1397	13504	0.74	0.97	0.03	0.86	0.84	0.73
0.1	38369	50898	1302	13831	0.74	0.97	0.02	0.86	0.84	0.73
0.2	34858	51595	605	17342	0.67	0.98	0.01	0.83	0.80	0.69
0.3	30685	51901	299	21515	0.59	0.99	0.01	0.79	0.74	0.64
0.4	26570	52027	173	25630	0.51	0.99	0.00	0.75	0.67	0.58
0.5	17811	52112	88	34389	0.34	1.00	0.00	0.67	0.51	0.45
0.6	9579	52149	51	42621	0.18	0.99	0.00	0.59	0.31	0.32
0.7	4828	52167	33	47372	0.09	0.99	0.00	0.55	0.17	0.22
0.8	2721	52178	22	49479	0.05	0.99	0.00	0.53	0.10	0.16
0.9	1544	52184	16	50656	0.03	0.99	0.00	0.51	0.06	0.12
1	979	52186	14	51221	0.02	0.99	0.00	0.51	0.04	0.10

Thr: Threshold

2. Protein x Disease

Thr.	TP	TN	FP	FN	Recall	Precision	FPR	Accuracy	F-score	MCC
0	71370	0	71370	0	1	0.5	1	0.5	0.67	
1E-06	54331	66896	4474	17039	0.76	0.92	0.06	0.85	0.83	0.71
1E-05	54331	66897	4473	17039	0.76	0.92	0.06	0.85	**0.83**	0.71
0.0001	54323	66900	4470	17047	0.76	0.92	0.06	0.85	0.83	0.71
0.001	54270	66944	4426	17100	0.76	0.92	0.06	0.85	0.83	0.71
0.01	53766	67572	3798	17604	0.75	0.93	0.05	0.85	0.83	0.71
0.02	53149	68116	3254	18221	0.74	0.94	0.05	0.85	0.83	**0.72**
0.03	52496	68492	2878	18874	0.74	0.95	0.04	0.85	0.83	0.71
0.04	51748	68821	2549	19622	0.73	0.95	0.04	0.84	0.82	0.71
0.05	50924	69096	2274	20446	0.71	0.96	0.03	0.84	0.82	0.70
0.06	50048	69357	2013	21322	0.70	0.96	0.03	0.84	0.81	0.70
0.07	49135	69573	1797	22235	0.69	0.96	0.03	0.83	0.80	0.69
0.08	48163	69759	1611	23207	0.67	0.97	0.02	0.83	0.80	0.68
0.09	47189	69913	1457	24181	0.66	0.97	0.02	0.82	0.79	0.68
0.1	46193	70040	1330	25177	0.65	0.97	0.02	0.81	0.78	0.67
0.2	37584	70813	557	33786	0.53	0.99	0.01	0.76	0.69	0.59
0.3	30358	71125	245	41012	0.43	0.99	0.00	0.71	0.60	0.51
0.4	24570	71232	138	46800	0.34	0.99	0.00	0.67	0.51	0.45
0.5	19665	71299	71	51705	0.28	1.00	0.00	0.64	0.43	0.40
0.6	15880	71332	38	55490	0.22	1.00	0.00	0.61	0.36	0.35
0.7	11001	71345	25	60369	0.15	1.00	0.00	0.58	0.27	0.29
0.8	7385	71351	19	63985	0.10	1.00	0.00	0.55	0.19	0.23
0.9	3737	71360	10	67633	0.05	1.00	0.00	0.53	0.10	0.16
1	2284	71363	7	69086	0.03	1.00	0.00	0.52	0.06	0.13

Thr: Threshold

b. 3-fold cross-validation

1. Protein x Cellular Component

Thr.	TP	TN	FP	FN	Recall	Precision	FPR	Accuracy	F-score	MCC
0	34140	0	37299	0	1	0.48	1	0.48	0.65	
1.00E-06	23800	34696	2603	10340	0.70	0.90	0.07	0.82	0.79	0.65
1.00E-05	23800	34697	2602	10340	0.70	0.90	0.07	0.82	0.79	0.65
0.0001	23797	34705	2594	10343	0.70	0.90	0.07	0.82	0.79	0.65
0.001	23787	34725	2574	10353	0.70	0.90	0.07	0.82	0.79	0.65
0.01	23722	35050	2249	10418	0.69	0.91	0.06	0.82	0.79	0.66
0.02	23610	35270	2029	10530	0.69	0.92	0.05	0.82	0.79	0.66
0.03	23480	35463	1836	10660	0.69	0.93	0.05	0.83	0.79	0.67
0.04	23260	35660	1639	10880	0.68	0.93	0.04	0.82	0.79	0.67
0.05	23061	35824	1475	11079	0.68	0.94	0.04	0.82	0.79	0.67
0.06	22849	35943	1356	11291	0.67	0.94	0.04	0.82	0.78	0.67
0.07	22609	36079	1220	11531	0.66	0.95	0.03	0.82	0.78	0.67
0.08	22424	36183	1116	11716	0.66	0.95	0.03	0.82	0.78	0.67
0.09	22194	36259	1040	11946	0.65	0.96	0.03	0.82	0.77	0.66
0.1	21907	36359	940	12233	0.64	0.96	0.03	0.82	0.77	0.66
0.2	18689	36878	421	15451	0.55	0.98	0.01	0.78	0.70	0.60
0.3	15932	37081	218	18208	0.47	0.99	0.01	0.74	0.63	0.55
0.4	11777	37190	109	22363	0.34	0.99	0.00	0.69	0.51	0.46
0.5	8539	37235	64	25601	0.25	0.99	0.00	0.64	0.40	0.38
0.6	5531	37272	27	28609	0.16	1.00	0.00	0.60	0.28	0.30
0.7	2763	37285	14	31377	0.08	0.99	0.00	0.56	0.15	0.21
0.8	1375	37289	10	32765	0.04	0.99	0.00	0.54	0.08	0.14
0.9	689	37295	4	33451	0.02	0.99	0.00	0.53	0.04	0.10
1	381	37297	2	33759	0.01	0.99	0.00	0.53	0.02	0.08

Thr: Threshold

2. Protein x Disease

Thr	TP	TN	FP	FN	Recall	Precision	FPR	Accuracy	F-score	MCC
0	63121	0	65784	0	1	0.489671	1	0.49	0.66	
1E-06	44429	60702	5082	18692	0.70	0.90	0.08	0.82	0.79	0.64
1E-05	44428	60702	5082	18693	0.70	0.90	0.08	0.82	0.79	0.64
0.0001	44415	60710	5074	18706	0.70	0.90	0.08	0.82	0.79	0.64
0.001	44357	60767	5017	18764	0.70	0.90	0.08	0.82	0.79	0.64
0.01	43562	61723	4061	19559	0.69	0.91	0.06	0.82	0.79	0.65
0.02	42679	62449	3335	20442	0.68	0.93	0.05	0.82	0.78	0.65
0.03	41920	62851	2933	21201	0.66	0.93	0.04	0.81	0.78	0.65
0.04	41190	63169	2615	21931	0.65	0.94	0.04	0.81	0.77	0.65
0.05	40259	63494	2290	22862	0.64	0.95	0.03	0.80	0.76	0.64
0.06	39408	63731	2053	23713	0.62	0.95	0.03	0.80	0.75	0.63
0.07	38556	63983	1801	24565	0.61	0.96	0.03	0.80	0.75	0.63
0.08	37706	64161	1623	25415	0.60	0.96	0.02	0.79	0.74	0.62
0.09	36860	64345	1439	26261	0.58	0.96	0.02	0.79	0.73	0.61
0.1	36000	64491	1293	27121	0.57	0.97	0.02	0.78	0.72	0.61
0.2	28455	65279	505	34666	0.45	0.98	0.01	0.73	0.62	0.53
0.3	22151	65555	229	40970	0.35	0.99	0.00	0.68	0.52	0.46
0.4	16269	65675	109	46852	0.26	0.99	0.00	0.64	0.41	0.38
0.5	11155	65722	62	51966	0.18	0.99	0.00	0.60	0.30	0.31
0.6	7704	65746	38	55417	0.12	1.00	0.00	0.57	0.22	0.26
0.7	4133	65757	27	58988	0.07	0.99	0.00	0.54	0.12	0.18
0.8	2511	65764	20	60610	0.04	0.99	0.00	0.53	0.08	0.14
0.9	1433	65775	9	61688	0.02	0.99	0.00	0.52	0.04	0.11
1	946	65777	7	62175	0.01	0.99	0.00	0.52	0.03	0.09

Thr: Threshold

A.2.2. Protein x Molecular Function and Protein x Disease (k=150)

a. 10-fold cross-validation

1. Protein x Molecular Function

Thr.	TP	TN	FP	FN	Recall	Precision	FPR	Accuracy	F-score	MCC
0	34270	0	34270	0	1	0.5	1	0.5	0.67	
1E-06	25590	32530	1740	8680	0.75	0.94	0.05	0.85	0.83	0.71
1E-05	25590	32530	1740	8680	0.75	0.94	0.05	0.85	0.83	0.71
0.0001	25589	32530	1740	8681	0.75	0.94	0.05	0.85	0.83	0.71
0.001	25584	32543	1727	8686	0.75	0.94	0.05	0.85	0.83	0.71
0.01	25550	32620	1650	8720	0.75	0.94	0.05	0.85	**0.83**	0.71
0.02	25503	32661	1609	8767	0.74	0.94	0.05	0.85	0.83	**0.71**
0.03	25434	32705	1565	8836	0.74	0.94	0.05	0.85	0.83	0.71
0.04	25273	32790	1480	8997	0.74	0.94	0.04	0.85	0.83	0.71
0.05	25070	32890	1380	9200	0.73	0.95	0.04	0.85	0.83	0.71
0.06	24779	33004	1266	9491	0.72	0.95	0.04	0.84	0.82	0.71
0.07	24462	33109	1161	9808	0.71	0.95	0.03	0.84	0.82	0.70
0.08	24099	33213	1057	10171	0.70	0.96	0.03	0.84	0.81	0.70
0.09	23723	33339	931	10547	0.69	0.96	0.03	0.83	0.81	0.69
0.1	23385	33412	858	10885	0.68	0.96	0.03	0.83	0.80	0.69
0.2	20195	33851	419	14075	0.59	0.98	0.01	0.79	0.74	0.63
0.3	16822	34068	202	17448	0.49	0.99	0.01	0.74	0.66	0.56
0.4	14249	34155	115	20021	0.42	0.99	0.00	0.71	0.59	0.51
0.5	8556	34223	47	25714	0.25	0.99	0.00	0.62	0.40	0.37
0.6	3587	34242	28	30683	0.10	0.99	0.00	0.55	0.19	0.23
0.7	2460	34251	19	31810	0.07	0.99	0.00	0.54	0.13	0.19
0.8	1745	34259	11	32525	0.05	0.99	0.00	0.53	0.10	0.16
0.9	1017	34263	7	33253	0.03	0.99	0.00	0.51	0.06	0.12
1	716	34264	6	33554	0.02	0.99	0.00	0.51	0.04	0.10

Thr: Threshold

2. Protein x Disease

Thr.	TP	TN	FP	FN	Recall	Precision	FPR	Accuracy	F-score	MCC
0	71370	0	71370	0	1	0.5	1	0.5	0.67	
1E-06	53568	67087	4283	17802	0.75	0.93	0.06	0.85	0.83	0.70
1E-05	53568	67090	4280	17802	0.75	0.93	0.06	0.85	**0.83**	0.70
0.0001	53564	67094	4276	17806	0.75	0.93	0.06	0.85	0.83	0.70
0.001	53521	67124	4246	17849	0.75	0.93	0.06	0.85	0.83	0.70
0.01	53050	67700	3670	18320	0.74	0.94	0.05	0.85	0.83	0.71
0.02	52468	68233	3137	18902	0.74	0.94	0.04	0.85	0.83	**0.71**
0.03	51798	68575	2795	19572	0.73	0.95	0.04	0.84	0.82	0.71
0.04	51085	68892	2478	20285	0.72	0.95	0.03	0.84	0.82	0.70
0.05	50261	69147	2223	21109	0.70	0.96	0.03	0.84	0.81	0.70
0.06	49365	69372	1998	22005	0.69	0.96	0.03	0.83	0.80	0.69
0.07	48518	69590	1780	22852	0.68	0.96	0.02	0.83	0.80	0.69
0.08	47677	69760	1610	23693	0.67	0.97	0.02	0.82	0.79	0.68
0.09	46765	69915	1455	24605	0.66	0.97	0.02	0.82	0.78	0.67
0.1	45913	70067	1303	25457	0.64	0.97	0.02	0.81	0.77	0.66
0.2	37313	70846	524	34057	0.52	0.99	0.01	0.76	0.68	0.58
0.3	30802	71117	253	40568	0.43	0.99	0.00	0.71	0.60	0.52
0.4	25304	71230	140	46066	0.35	0.99	0.00	0.68	0.52	0.46
0.5	20324	71280	90	51046	0.28	1.00	0.00	0.64	0.44	0.40
0.6	16008	71317	53	55362	0.22	1.00	0.00	0.61	0.37	0.35
0.7	11195	71338	32	60175	0.16	1.00	0.00	0.58	0.27	0.29
0.8	7208	71352	18	64162	0.10	1.00	0.00	0.55	0.18	0.23
0.9	3649	71358	12	67721	0.05	1.00	0.00	0.53	0.10	0.16
1	2193	71361	9	69177	0.03	1.00	0.00	0.52	0.06	0.12

Thr: Threshold

b. 3-fold cross-validation

1. Protein x Molecular Function

Thr.	TP	TN	FP	FN	Recall	Precision	FPR	Accuracy	F-score	MCC
0	22609	0	25443	0	1	0.47	1	0.47	0.64	
1E-06	15496	23937	1506	7113	0.69	0.91	0.06	0.82	0.78	0.65
1E-05	15496	23937	1506	7113	0.69	0.91	0.06	0.82	0.78	0.65
0.0001	15490	23940	1503	7119	0.69	0.91	0.06	0.82	0.78	0.65
0.001	15483	23948	1495	7126	0.68	0.91	0.06	0.82	0.78	0.65
0.01	15421	24051	1392	7188	0.68	0.92	0.05	0.82	0.78	0.66
0.02	15348	24127	1316	7261	0.68	0.92	0.05	0.82	0.78	0.66
0.03	15233	24207	1236	7376	0.67	0.92	0.05	0.82	0.78	0.66
0.04	15117	24276	1167	7492	0.67	0.93	0.05	0.82	0.78	0.66
0.05	14940	24382	1061	7669	0.66	0.93	0.04	0.82	0.77	0.66
0.06	14773	24510	933	7836	0.65	0.94	0.04	0.82	0.77	0.66
0.07	14511	24616	827	8098	0.64	0.95	0.03	0.81	0.76	0.65
0.08	14271	24695	748	8338	0.63	0.95	0.03	0.81	0.76	0.65
0.09	14046	24756	687	8563	0.62	0.95	0.03	0.81	0.75	0.64
0.1	13752	24850	593	8857	0.61	0.96	0.02	0.80	0.74	0.64
0.2	11258	25200	243	11351	0.50	0.98	0.01	0.76	0.66	0.57
0.3	9209	25330	113	13400	0.41	0.99	0.00	0.72	0.58	0.51
0.4	7521	25381	62	15088	0.33	0.99	0.00	0.68	0.50	0.45
0.5	4404	25414	29	18205	0.19	0.99	0.00	0.62	0.33	0.33
0.6	1588	25422	21	21021	0.07	0.99	0.00	0.56	0.13	0.19
0.7	1017	25429	14	21592	0.04	0.99	0.00	0.55	0.09	0.15
0.8	663	25436	7	21946	0.03	0.99	0.00	0.54	0.06	0.12
0.9	440	25438	5	22169	0.02	0.99	0.00	0.54	0.04	0.10
1	205	25439	4	22404	0.01	0.98	0.00	0.53	0.02	0.07

Thr: Threshold

2. Protein x Disease

Thr.	TP	TN	FP	FN	Recall	Precision	FPR	Accuracy	F-score	MCC
0	63121	0	65784	0	1	0.49	1	0.49	0.66	
1E-06	43580	60882	4902	19541	0.69	0.90	0.07	0.81	0.78	0.64
1E-05	43578	60882	4902	19543	0.69	0.90	0.07	0.81	0.78	0.64
0.0001	43571	60891	4893	19550	0.69	0.90	0.07	0.81	0.78	0.64
0.001	43506	60935	4849	19615	0.69	0.90	0.07	0.81	0.78	0.64
0.01	42856	61722	4062	20265	0.68	0.91	0.06	0.81	0.78	0.64
0.02	42038	62410	3374	21083	0.67	0.93	0.05	0.81	0.77	0.64
0.03	41258	62871	2913	21863	0.65	0.93	0.04	0.81	0.77	0.64
0.04	40556	63204	2580	22565	0.64	0.94	0.04	0.80	0.76	0.64
0.05	39753	63517	2267	23368	0.63	0.95	0.03	0.80	0.76	0.63
0.06	38924	63764	2020	24197	0.62	0.95	0.03	0.80	0.75	0.63
0.07	38101	63977	1807	25020	0.60	0.95	0.03	0.79	0.74	0.62
0.08	37283	64160	1624	25838	0.59	0.96	0.02	0.79	0.73	0.62
0.09	36499	64317	1467	26622	0.58	0.96	0.02	0.78	0.72	0.61
0.1	35684	64477	1307	27437	0.57	0.96	0.02	0.78	0.71	0.60
0.2	28220	65310	474	34901	0.45	0.98	0.01	0.73	0.61	0.53
0.3	21479	65598	186	41642	0.34	0.99	0.00	0.68	0.51	0.45
0.4	16458	65688	96	46663	0.26	0.99	0.00	0.64	0.41	0.39
0.5	11468	65734	50	51653	0.18	1.00	0.00	0.60	0.31	0.32
0.6	7654	65761	23	55467	0.12	1.00	0.00	0.57	0.22	0.26
0.7	4182	65769	15	58939	0.07	1.00	0.00	0.54	0.12	0.19
0.8	2505	65778	6	60616	0.04	1.00	0.00	0.53	0.08	0.14
0.9	1429	65778	6	61692	0.02	1.00	0.00	0.52	0.04	0.11
1	945	65779	5	62176	0.01	0.99	0.00	0.52	0.03	0.09

Thr: Threshold

A.2.3. Protein x Biological Process and Protein x Disease (k=200)

a. 10-fold cross-validation

1. Protein x Biological Process

Thr.	TP	TN	FP	FN	Recall	Precision	FPR	Accuracy	F-score	MCC
0	75880	0	75880	0	1	0.5	1	0.5	0.67	
1E-06	58833	71482	4398	17047	0.78	0.93	0.06	0.86	0.85	0.73
1E-05	58833	71482	4398	17047	0.78	0.93	0.06	0.86	0.85	0.73
0.0001	58827	71486	4394	17053	0.78	0.93	0.06	0.86	0.85	0.73
0.001	58812	71522	4358	17068	0.78	0.93	0.06	0.86	0.85	0.73
0.01	58558	71904	3976	17322	0.77	0.94	0.05	0.86	**0.85**	0.73
0.02	58258	72158	3722	17622	0.77	0.94	0.05	0.86	0.85	**0.73**
0.03	57839	72379	3501	18041	0.76	0.94	0.05	0.86	0.84	0.73
0.04	57131	72615	3265	18749	0.75	0.95	0.04	0.85	0.84	0.73
0.05	56217	72867	3013	19663	0.74	0.95	0.04	0.85	0.83	0.72
0.06	55093	73158	2722	20787	0.73	0.95	0.04	0.85	0.82	0.71
0.07	53926	73437	2443	21954	0.71	0.96	0.03	0.84	0.82	0.70
0.08	52631	73678	2202	23249	0.69	0.96	0.03	0.83	0.81	0.69
0.09	51342	73870	2010	24538	0.68	0.96	0.03	0.83	0.79	0.68
0.1	50037	74044	1836	25843	0.66	0.96	0.02	0.82	0.78	0.67
0.2	37176	75139	741	38704	0.49	0.98	0.01	0.74	0.65	0.55
0.3	26938	75550	330	48942	0.36	0.99	0.00	0.68	0.52	0.46
0.4	19225	75721	159	56655	0.25	0.99	0.00	0.63	0.40	0.38
0.5	13934	75808	72	61946	0.18	0.99	0.00	0.59	0.31	0.32
0.6	9828	75841	39	66052	0.13	1.00	0.00	0.56	0.23	0.26
0.7	6001	75856	24	69879	0.08	1.00	0.00	0.54	0.15	0.20
0.8	4078	75865	15	71802	0.05	1.00	0.00	0.53	0.10	0.17
0.9	2810	75870	10	73070	0.04	1.00	0.00	0.52	0.07	0.14
1	1875	75874	6	74005	0.02	1.00	0.00	0.51	0.05	0.11

Thr: Threshold

2. Protein x Disease

Thr.	TP	TN	FP	FN	Recall	Precision	FPR	Accuracy	F-score	MCC
0	71370	0	71370	0	1	0.5	1	0.5	0.67	
1E-06	55186	66008	5362	16184	0.77	0.91	0.08	0.85	0.84	0.71
1E-05	55186	66010	5360	16184	0.77	0.91	0.08	0.85	0.84	0.71
0.0001	55183	66018	5352	16187	0.77	0.91	0.07	0.85	0.84	0.71
0.001	55137	66085	5285	16233	0.77	0.91	0.07	0.85	0.84	0.71
0.01	54408	67101	4269	16962	0.76	0.93	0.06	0.85	**0.84**	0.71
0.02	53643	67838	3532	17727	0.75	0.94	0.05	0.85	0.83	**0.72**
0.03	52828	68340	3030	18542	0.74	0.95	0.04	0.85	0.83	0.71
0.04	52002	68704	2666	19368	0.73	0.95	0.04	0.85	0.83	0.71
0.05	51052	69023	2347	20318	0.72	0.96	0.03	0.84	0.82	0.71
0.06	50083	69297	2073	21287	0.70	0.96	0.03	0.84	0.81	0.70
0.07	49049	69544	1826	22321	0.69	0.96	0.03	0.83	0.80	0.69
0.08	48051	69748	1622	23319	0.67	0.97	0.02	0.83	0.79	0.68
0.09	47033	69916	1454	24337	0.66	0.97	0.02	0.82	0.78	0.67
0.1	46000	70052	1318	25370	0.64	0.97	0.02	0.81	0.78	0.66
0.2	36894	70835	535	34476	0.52	0.99	0.01	0.75	0.68	0.58
0.3	29704	71103	267	41666	0.42	0.99	0.00	0.71	0.59	0.51
0.4	23889	71215	155	47481	0.33	0.99	0.00	0.67	0.50	0.44
0.5	18975	71282	88	52395	0.27	1.00	0.00	0.63	0.42	0.39
0.6	14916	71316	54	56454	0.21	1.00	0.00	0.60	0.35	0.34
0.7	10739	71336	34	60631	0.15	1.00	0.00	0.57	0.26	0.28
0.8	7336	71346	24	64034	0.10	1.00	0.00	0.55	0.19	0.23
0.9	3844	71351	19	67526	0.05	1.00	0.00	0.53	0.10	0.17
1	2337	71356	14	69033	0.03	0.99	0.00	0.52	0.06	0.13

Thr: Threshold

b. 3-fold cross-validation

1. Protein x Biological Process

Thr.	TP	TN	FP	FN	Recall	Precision	FPR	Accuracy	F-score	MCC
0	55959	0	59220	0	1	0.49	1	0.49	0.65	
1E-06	38854	54839	4381	17105	0.69	0.90	0.07	0.81	0.78	0.64
1E-05	38854	54839	4381	17105	0.69	0.90	0.07	0.81	0.78	0.64
0.0001	38851	54845	4375	17108	0.69	0.90	0.07	0.81	0.78	0.64
0.001	38823	54881	4339	17136	0.69	0.90	0.07	0.81	0.78	0.64
0.01	38243	55703	3517	17716	0.68	0.92	0.06	0.82	0.78	0.65
0.02	37770	56124	3096	18189	0.67	0.92	0.05	0.82	0.78	0.65
0.03	37229	56451	2769	18730	0.67	0.93	0.05	0.81	0.78	0.65
0.04	36545	56718	2502	19414	0.65	0.94	0.04	0.81	0.77	0.64
0.05	35777	56942	2278	20182	0.64	0.94	0.04	0.80	0.76	0.64
0.06	34945	57183	2037	21014	0.62	0.94	0.03	0.80	0.75	0.63
0.07	33893	57423	1797	22066	0.61	0.95	0.03	0.79	0.74	0.62
0.08	32839	57601	1619	23120	0.59	0.95	0.03	0.79	0.73	0.61
0.09	31758	57794	1426	24201	0.57	0.96	0.02	0.78	0.71	0.60
0.1	30763	57943	1277	25196	0.55	0.96	0.02	0.77	0.70	0.59
0.2	21444	58759	461	34515	0.38	0.98	0.01	0.70	0.55	0.48
0.3	14662	59020	200	41297	0.26	0.99	0.00	0.64	0.41	0.39
0.4	9752	59129	91	46207	0.17	0.99	0.00	0.60	0.30	0.31
0.5	6733	59179	41	49226	0.12	0.99	0.00	0.57	0.21	0.25
0.6	4316	59194	26	51643	0.08	0.99	0.00	0.55	0.14	0.20
0.7	2684	59208	12	53275	0.05	1.00	0.00	0.54	0.09	0.16
0.8	1674	59210	10	54285	0.03	0.99	0.00	0.53	0.06	0.12
0.9	1101	59214	6	54858	0.02	0.99	0.00	0.52	0.04	0.10
1	692	59216	4	55267	0.01	0.99	0.00	0.52	0.02	0.08

Thr: Threshold

2. Protein x Disease

Thr.	TP	TN	FP	FN	Recall	Precision	FPR	Accuracy	F-score	MCC
0	63121	0	65784	0	1	0.49	1	0.49	0.66	
1E-06	45253	59540	6244	17868	0.72	0.88	0.09	0.81	0.79	0.63
1E-05	45253	59540	6244	17868	0.72	0.88	0.09	0.81	0.79	0.63
0.0001	45244	59546	6238	17877	0.72	0.88	0.09	0.81	0.79	0.63
0.001	45189	59635	6149	17932	0.72	0.88	0.09	0.81	0.79	0.64
0.01	44166	61194	4590	18955	0.70	0.91	0.07	0.82	0.79	0.65
0.02	43069	62107	3677	20052	0.68	0.92	0.06	0.82	0.78	0.65
0.03	42111	62657	3127	21010	0.67	0.93	0.05	0.81	0.78	0.65
0.04	41183	63070	2714	21938	0.65	0.94	0.04	0.81	0.77	0.64
0.05	40238	63481	2303	22883	0.64	0.95	0.04	0.80	0.76	0.64
0.06	39355	63759	2025	23766	0.62	0.95	0.03	0.80	0.75	0.63
0.07	38490	63974	1810	24631	0.61	0.96	0.03	0.79	0.74	0.63
0.08	37512	64165	1619	25609	0.59	0.96	0.02	0.79	0.73	0.62
0.09	36534	64353	1431	26587	0.58	0.96	0.02	0.78	0.72	0.61
0.1	35619	64506	1278	27502	0.56	0.97	0.02	0.78	0.71	0.60
0.2	27988	65286	498	35133	0.44	0.98	0.01	0.72	0.61	0.53
0.3	21460	65560	224	41661	0.34	0.99	0.00	0.68	0.51	0.45
0.4	15798	65665	119	47323	0.25	0.99	0.00	0.63	0.40	0.38
0.5	10977	65717	67	52144	0.17	0.99	0.00	0.59	0.30	0.31
0.6	7528	65742	42	55593	0.12	0.99	0.00	0.57	0.21	0.25
0.7	4374	65764	20	58747	0.07	1.00	0.00	0.54	0.13	0.19
0.8	2644	65770	14	60477	0.04	0.99	0.00	0.53	0.08	0.15
0.9	1601	65774	10	61520	0.03	0.99	0.00	0.52	0.05	0.11
1	1120	65777	7	62001	0.02	0.99	0.00	0.52	0.03	0.09

Thr: Threshold

A. 3. Confusion Matrices for NMTF without PPI matrix

A.3.1. The sets of k values tested for NMTF algorithm without and with PPI matrix.

	k1 (CC)	k2 (Prot)	k3 (MF)	k4 (BP)	k5 (Dis)
k1	50	100	50	150	150
k2	50	200	50	150	150
k3	50	500	50	150	150
k4	50	1500	50	150	150
k5	50	2500	50	150	150
k6	150	100	150	150	150
k7	150	200	150	150	150
k8	150	500	150	150	150
k9	150	1500	150	150	150
k10	150	2500	150	150	150
k11	50	100	50	200	200
k12	50	200	50	200	200
k13	50	500	50	200	200
k14	50	1500	50	200	200
k15	50	2500	50	200	200
k16	150	100	150	200	200
k17	150	200	150	200	200
k18	150	500	150	200	200
k19	150	1500	150	200	200
k20	150	2500	150	200	200

Prot: Protein, Dis: Disease

A.3.2. Error rates for each set of k values according to relation matrices along with their averages, without PPI matrix.

	R12	R13	R14	R15	Avg
k1	0.036273	0.033162	0.026058	0.025634	0.030282
k2	0.035658	0.032333	0.025733	0.025912	0.029909
k3	0.034681	0.031946	0.02559	0.025929	0.029536
k4	0.035074	0.032367	0.025534	0.026204	0.029795
k5	0.034949	0.032196	0.025849	0.02589	0.029721
k6	0.029251	0.030031	0.024164	0.024479	0.026981
k7	0.028154	0.030106	0.02424	0.024898	0.02685
k8	0.028555	0.03033	0.02422	0.024565	0.026917
k9	0.027943	0.030293	0.024259	0.025031	0.026881
k10	0.027782	0.03032	0.024335	0.024669	**0.026777**
k11	0.036028	0.032139	0.025794	0.026353	0.030079
k12	0.036066	0.032338	0.025898	0.025838	0.030035
k13	0.034603	0.032199	0.025964	0.026088	0.029713
k14	0.034544	0.032281	0.02586	0.026227	0.029728
k15	0.034659	0.032429	0.025852	0.025972	0.029728
k16	0.029992	0.030338	0.024085	0.024538	0.027238
k17	0.02848	0.03012	0.024089	0.024498	0.026797
k18	0.028073	0.030099	0.024143	0.024923	0.026809
k19	0.027998	0.030138	0.024077	0.024911	0.026781
k20	0.027894	0.030525	0.024303	0.024787	0.026877

Avg: Average

A.3.3 Confusion matrices of NMTF algorithm without PPI matrix.

1. Protein x Cellular Component

a. 10-fold cross-validation

Thr.	TP	TN	FP	FN	Recall	Precision	FPR	Accuracy	F-score	MCC
0	52200	0	52200	0	1	0.5	1	0.5	0.67	
1E-06	51986	10199	42001	214	1.00	0.55	0.80	0.60	0.71	0.32
1E-05	51910	14212	37988	290	0.99	0.58	0.73	0.63	0.73	0.39
0.0001	51723	20564	31636	477	0.99	0.62	0.61	0.69	0.76	0.48
0.001	51106	30219	21981	1094	0.98	0.70	0.42	0.78	0.82	0.61
0.01	48860	43153	9047	3340	0.94	0.84	0.17	0.88	0.89	0.77
0.02	47368	46223	5977	4832	0.91	0.89	0.11	0.90	0.90	0.79
0.03	46171	47772	4428	6029	0.88	0.91	0.08	0.90	**0.90**	**0.80**
0.04	45206	48608	3592	6994	0.87	0.93	0.07	0.90	0.90	0.80
0.05	44322	49175	3025	7878	0.85	0.94	0.06	0.90	0.89	0.79
0.06	43593	49630	2570	8607	0.84	0.94	0.05	0.89	0.89	0.79
0.07	42802	49970	2230	9398	0.82	0.95	0.04	0.89	0.88	0.78
0.08	42125	50230	1970	10075	0.81	0.96	0.04	0.88	0.87	0.78
0.09	41492	50432	1768	10708	0.79	0.96	0.03	0.88	0.87	0.77
0.1	40861	50597	1603	11339	0.78	0.96	0.03	0.88	0.86	0.77
0.2	35086	51475	725	17114	0.67	0.98	0.01	0.83	0.80	0.69
0.3	29696	51852	348	22504	0.57	0.99	0.01	0.78	0.72	0.62
0.4	24427	52026	174	27773	0.47	0.99	0.00	0.73	0.64	0.55
0.5	19383	52117	83	32817	0.37	1.00	0.00	0.68	0.54	0.47
0.6	14561	52174	26	37639	0.28	1.00	0.00	0.64	0.44	0.40
0.7	9910	52192	8	42290	0.19	1.00	0.00	0.59	0.32	0.32
0.8	6096	52197	3	46104	0.12	1.00	0.00	0.56	0.21	0.25
0.9	3411	52199	1	48789	0.07	1.00	0.00	0.53	0.12	0.18
1	1751	52200	0	50449	0.03	1	0.00	0.52	0.06	0.13

Thr: Threshold

b. 3-fold cross-validation

Thr.	TP	TN	FP	FN	Recall	Precision	FPR	Accuracy	F-score	MCC
0	34140	0	37299	0	1	0.48	1	0.48	0.65	
1E-06	34003	7203	30096	137	1.00	0.53	0.81	0.58	0.69	0.31
1E-05	33905	9985	27314	235	0.99	0.55	0.73	0.61	0.71	0.37
0.0001	33670	14204	23095	470	0.99	0.59	0.62	0.67	0.74	0.45
0.001	32983	20841	16458	1157	0.97	0.67	0.44	0.75	0.79	0.57
0.01	30793	31096	6203	3347	0.90	0.83	0.17	0.87	0.87	0.74
0.02	29376	33423	3876	4764	0.86	0.88	0.10	0.88	0.87	0.76
0.03	28348	34510	2789	5792	0.83	0.91	0.07	0.88	0.87	0.76
0.04	27494	35139	2160	6646	0.81	0.93	0.06	0.88	0.86	0.76
0.05	26725	35518	1781	7415	0.78	0.94	0.05	0.87	0.85	0.75
0.06	26060	35812	1487	8080	0.76	0.95	0.04	0.87	0.84	0.74
0.07	25442	36010	1289	8698	0.75	0.95	0.03	0.86	0.84	0.73
0.08	24897	36180	1119	9243	0.73	0.96	0.03	0.85	0.83	0.73
0.09	24378	36324	975	9762	0.71	0.96	0.03	0.85	0.82	0.72
0.1	23862	36426	873	10278	0.70	0.96	0.02	0.84	0.81	0.71
0.2	19252	36994	305	14888	0.56	0.98	0.01	0.79	0.72	0.62
0.3	15189	37170	129	18951	0.44	0.99	0.00	0.73	0.61	0.54
0.4	11408	37241	58	22732	0.33	0.99	0.00	0.68	0.50	0.45
0.5	7956	37279	20	26184	0.23	1.00	0.00	0.63	0.38	0.37
0.6	5019	37290	9	29121	0.15	1.00	0.00	0.59	0.26	0.29
0.7	2867	37297	2	31273	0.08	1.00	0.00	0.56	0.15	0.21
0.8	1533	37298	1	32607	0.04	1.00	0.00	0.54	0.09	0.15
0.9	785	37298	1	33355	0.02	1.00	0.00	0.53	0.04	0.11
1	410	37299	0	33730	0.01	1	0	0.53	0.02	0.08

Thr: Threshold

2. Protein x Molecular Function

a. 10-fold cross-validation

Thr.	TP	TN	FP	FN	Recall	Precision	FPR	Accuracy	F-score	MCC
0	34270	0	34270	0	1	0.5	1	0.5	0.67	
1E-06	34076	5384	28886	194	0.99	0.54	0.84	0.58	0.70	0.28
1E-05	34010	6584	27686	260	0.99	0.55	0.81	0.59	0.71	0.31
0.0001	33868	8495	25775	402	0.99	0.57	0.75	0.62	0.72	0.35
0.001	33449	13242	21028	821	0.98	0.61	0.61	0.68	0.75	0.45
0.01	30827	24908	9362	3443	0.90	0.77	0.27	0.81	0.83	0.64
0.02	28729	28585	5685	5541	0.84	0.83	0.17	0.84	**0.84**	0.67
0.03	27061	30346	3924	7209	0.79	0.87	0.11	0.84	0.83	**0.68**
0.04	25722	31312	2958	8548	0.75	0.90	0.09	0.83	0.82	0.67
0.05	24554	31925	2345	9716	0.72	0.91	0.07	0.82	0.80	0.66
0.06	23593	32365	1905	10677	0.69	0.93	0.06	0.82	0.79	0.65
0.07	22775	32673	1597	11495	0.66	0.93	0.05	0.81	0.78	0.65
0.08	22022	32925	1345	12248	0.64	0.94	0.04	0.80	0.76	0.64
0.09	21294	33126	1144	12976	0.62	0.95	0.03	0.79	0.75	0.63
0.1	20660	33271	999	13610	0.60	0.95	0.03	0.79	0.74	0.62
0.2	16401	33947	323	17869	0.48	0.98	0.01	0.73	0.64	0.55
0.3	13686	34122	148	20584	0.40	0.99	0.00	0.70	0.57	0.49
0.4	11468	34193	77	22802	0.33	0.99	0.00	0.67	0.50	0.44
0.5	9323	34230	40	24947	0.27	1.00	0.00	0.64	0.43	0.39
0.6	6987	34249	21	27283	0.20	1.00	0.00	0.60	0.34	0.34
0.7	4566	34264	6	29704	0.13	1.00	0.00	0.57	0.24	0.27
0.8	2589	34268	2	31681	0.08	1.00	0.00	0.54	0.14	0.20
0.9	1273	34270	0	32997	0.04	1	0.00	0.52	0.07	0.14
1	534	34270	0	33736	0.02	1	0.00	0.51	0.03	0.09

Thr: Threshold

b. 3-fold cross-validation

Thr.	TP	TN	FP	FN	Recall	Precision	FPR	Accuracy	F-score	MCC
0	22609	0	25443	0	1	0.47	1	0.47	0.64	
1E-06	22398	3808	21635	211	0.99	0.51	0.85	0.55	0.67	0.25
1E-05	22333	4599	20844	276	0.99	0.52	0.82	0.56	0.68	0.28
0.0001	22205	6015	19428	404	0.98	0.53	0.76	0.59	0.69	0.32
0.001	21786	9759	15684	823	0.96	0.58	0.62	0.66	0.73	0.42
0.01	19656	18868	6575	2953	0.87	0.75	0.26	0.80	0.80	0.61
0.02	18048	21620	3823	4561	0.80	0.83	0.15	0.83	0.81	0.65
0.03	16795	22845	2598	5814	0.74	0.87	0.10	0.82	0.80	0.65
0.04	15832	23535	1908	6777	0.70	0.89	0.07	0.82	0.78	0.65
0.05	14977	23964	1479	7632	0.66	0.91	0.06	0.81	0.77	0.64
0.06	14194	24261	1182	8415	0.63	0.92	0.05	0.80	0.75	0.62
0.07	13538	24461	982	9071	0.60	0.93	0.04	0.79	0.73	0.61
0.08	13007	24602	841	9602	0.58	0.94	0.03	0.78	0.71	0.60
0.09	12535	24727	716	10074	0.55	0.95	0.03	0.78	0.70	0.59
0.1	12145	24830	613	10464	0.54	0.95	0.02	0.77	0.69	0.58
0.2	9287	25262	181	13322	0.41	0.98	0.01	0.72	0.58	0.51
0.3	7469	25368	75	15140	0.33	0.99	0.00	0.68	0.50	0.45
0.4	5829	25412	31	16780	0.26	0.99	0.00	0.65	0.41	0.39
0.5	4084	25432	11	18525	0.18	1.00	0.00	0.61	0.31	0.32
0.6	2444	25440	3	20165	0.11	1.00	0.00	0.58	0.20	0.25
0.7	1143	25442	1	21466	0.05	1.00	4E-05	0.55	0.10	0.17
0.8	451	25443	0	22158	0.02	1	0	0.54	0.04	0.10
0.9	145	25443	0	22464	0.01	1	0	0.53	0.01	0.06
1	42	25443	0	22567	0.00	1	0	0.53	0.00	0.03

Thr: Threshold

3. Protein x Biological Process

a. 10-fold cross-validation

Thr.	TP	TN	FP	FN	Recall	Precision	FPR	Accuracy	F-score	MCC
0	75880	0	75880	0	1	0.5	1	0.5	0.67	
1E-06	75621	4887	70993	259	1.00	0.52	0.94	0.53	0.68	0.17
1E-05	75514	5991	69889	366	1.00	0.52	0.92	0.54	0.68	0.19
0.0001	75240	8945	66935	640	0.99	0.53	0.88	0.55	0.69	0.22
0.001	73785	20237	55643	2095	0.97	0.57	0.73	0.62	0.72	0.34
0.01	61968	54691	21189	13912	0.82	0.75	0.28	0.77	0.78	0.54
0.02	52717	64945	10935	23163	0.69	0.83	0.14	0.78	0.76	0.56
0.03	46277	69066	6814	29603	0.61	0.87	0.09	0.76	0.72	0.55
0.04	41426	71170	4710	34454	0.55	0.90	0.06	0.74	0.68	0.53
0.05	37819	72483	3397	38061	0.50	0.92	0.04	0.73	0.65	0.51
0.06	34988	73300	2580	40892	0.46	0.93	0.03	0.71	0.62	0.49
0.07	32634	73847	2033	43246	0.43	0.94	0.03	0.70	0.59	0.48
0.08	30602	74203	1677	45278	0.40	0.95	0.02	0.69	0.57	0.47
0.09	28789	74491	1389	47091	0.38	0.95	0.02	0.68	0.54	0.45
0.1	27276	74720	1160	48604	0.36	0.96	0.02	0.67	0.52	0.44
0.2	18067	75575	305	57813	0.24	0.98	0.00	0.62	0.38	0.36
0.3	13340	75767	113	62540	0.18	0.99	0.00	0.59	0.30	0.31
0.4	10109	75834	46	65771	0.13	1.00	0.00	0.57	0.23	0.27
0.5	7660	75862	18	68220	0.10	1.00	0.00	0.55	0.18	0.23
0.6	5619	75873	7	70261	0.07	1.00	0.00	0.54	0.14	0.20
0.7	3910	75877	3	71970	0.05	1.00	0.00	0.53	0.10	0.16
0.8	2303	75879	1	73577	0.03	1.00	0.00	0.52	0.06	0.12
0.9	1106	75880	0	74774	0.01	1.00	0.00	0.51	0.03	0.09
1	435	75880	0	75445	0.01	1	0.00	0.50	0.01	0.05

Thr: Threshold

b. 3-fold cross-validation

Thr.	TP	TN	FP	FN	Recall	Precision	FPR	Accuracy	F-score	MCC
0	55959	0	59220	0	1	0.49	1	0.49	0.65	
1E-06	55662	3635	55585	297	0.99	0.50	0.94	0.51	0.67	0.15
1E-05	55543	4454	54766	416	0.99	0.50	0.92	0.52	0.67	0.17
0.0001	55198	6546	52674	761	0.99	0.51	0.89	0.54	0.67	0.20
0.001	53853	15204	44016	2106	0.96	0.55	0.74	0.60	0.70	0.31
0.01	43117	43847	15373	12842	0.77	0.74	0.26	0.76	0.75	0.51
0.02	35191	51781	7439	20768	0.63	0.83	0.13	0.76	0.71	0.52
0.03	29972	54766	4454	25987	0.54	0.87	0.08	0.74	0.66	0.50
0.04	26371	56212	3008	29588	0.47	0.90	0.05	0.72	0.62	0.48
0.05	23809	57020	2200	32150	0.43	0.92	0.04	0.70	0.58	0.46
0.06	21772	57557	1663	34187	0.39	0.93	0.03	0.69	0.55	0.45
0.07	20134	57915	1305	35825	0.36	0.94	0.02	0.68	0.52	0.43
0.08	18767	58151	1069	37192	0.34	0.95	0.02	0.67	0.50	0.42
0.09	17566	58348	872	38393	0.31	0.95	0.01	0.66	0.47	0.41
0.1	16473	58503	717	39486	0.29	0.96	0.01	0.65	0.45	0.40
0.2	10144	59055	165	45815	0.18	0.98	0.00	0.60	0.31	0.31
0.3	6970	59158	62	48989	0.12	0.99	0.00	0.57	0.22	0.26
0.4	4674	59200	20	51285	0.08	1.00	0.00	0.55	0.15	0.21
0.5	3034	59215	5	52925	0.05	1.00	8E-05	0.54	0.10	0.17
0.6	1788	59220	0	54171	0.03	1	0	0.53	0.06	0.13
0.7	914	59220	0	55045	0.02	1	0	0.52	0.03	0.09
0.8	349	59220	0	55610	0.01	1	0	0.52	0.01	0.06
0.9	154	59220	0	55805	0.00	1	0	0.52	0.01	0.04
1	50	59220	0	55909	0.00	1	0	0.51	0.00	0.02

Thr: Threshold

4. Protein x Disease

a. 10-fold cross-validation

Thr.	TP	TN	FP	FN	Recall	Precision	FPR	Accuracy	F-score	MCC
0	71370	0	71370	0	1	0.5	1	0.5	0.67	
1E-06	70804	4796	66574	566	0.99	0.52	0.93	0.53	0.68	0.16
1E-05	70587	5820	65550	783	0.99	0.52	0.92	0.54	0.68	0.17
0.0001	70106	8142	63228	1264	0.98	0.53	0.89	0.55	0.68	0.19
0.001	68115	18240	53130	3255	0.95	0.56	0.74	0.60	0.71	0.29
0.01	55481	51784	19586	15889	0.78	0.74	0.27	0.75	**0.76**	0.50
0.02	46929	60903	10467	24441	0.66	0.82	0.15	0.76	0.73	**0.52**
0.03	41068	64765	6605	30302	0.58	0.86	0.09	0.74	0.69	0.51
0.04	36700	66732	4638	34670	0.51	0.89	0.06	0.72	0.65	0.50
0.05	33203	67945	3425	38167	0.47	0.91	0.05	0.71	0.61	0.48
0.06	30323	68747	2623	41047	0.42	0.92	0.04	0.69	0.58	0.46
0.07	27960	69324	2046	43410	0.39	0.93	0.03	0.68	0.55	0.45
0.08	25936	69697	1673	45434	0.36	0.94	0.02	0.67	0.52	0.43
0.09	24125	70051	1319	47245	0.34	0.95	0.02	0.66	0.50	0.42
0.1	22533	70277	1093	48837	0.32	0.95	0.02	0.65	0.47	0.40
0.2	13496	71124	246	57874	0.19	0.98	0.00	0.59	0.32	0.31
0.3	9553	71284	86	61817	0.13	0.99	0.00	0.57	0.24	0.26
0.4	7257	71336	34	64113	0.10	1.00	0.00	0.55	0.18	0.23
0.5	5627	71358	12	65743	0.08	1.00	0.00	0.54	0.15	0.20
0.6	4337	71364	6	67033	0.06	1.00	0.00	0.53	0.11	0.18
0.7	3128	71367	3	68242	0.04	1.00	0.00	0.52	0.08	0.15
0.8	1902	71367	3	69468	0.03	1.00	0.00	0.51	0.05	0.12
0.9	859	71369	1	70511	0.01	1.00	0.00	0.51	0.02	0.08
1	300	71370	0	71070	0.00	1	0.00	0.50	0.01	0.05

Thr: Threshold

b. 3-fold cross-validation

Thr.	TP	TN	FP	FN	Recall	Precision	FPR	Accuracy	F-score	MCC
0	63121	0	65784	0	1	0.49	1	0.49	0.66	
1E-06	62698	2336	63448	423	0.99	0.50	0.96	0.50	0.66	0.10
1E-05	62496	3042	62742	625	0.99	0.50	0.95	0.51	0.66	0.11
0.0001	62053	4823	60961	1068	0.98	0.50	0.93	0.52	0.67	0.13
0.001	60297	13046	52738	2824	0.96	0.53	0.80	0.57	0.68	0.23
0.01	47666	47164	18620	15455	0.76	0.72	0.28	0.74	0.74	0.47
0.02	39638	56024	9760	23483	0.63	0.80	0.15	0.74	0.70	0.49
0.03	34451	59559	6225	28670	0.55	0.85	0.09	0.73	0.66	0.49
0.04	30616	61409	4375	32505	0.49	0.87	0.07	0.71	0.62	0.47
0.05	27509	62535	3249	35612	0.44	0.89	0.05	0.70	0.59	0.45
0.06	25080	63280	2504	38041	0.40	0.91	0.04	0.69	0.55	0.44
0.07	23061	63838	1946	40060	0.37	0.92	0.03	0.67	0.52	0.42
0.08	21285	64222	1562	41836	0.34	0.93	0.02	0.66	0.50	0.41
0.09	19710	64505	1279	43411	0.31	0.94	0.02	0.65	0.47	0.40
0.1	18390	64752	1032	44731	0.29	0.95	0.02	0.64	0.45	0.39
0.2	10688	65566	218	52433	0.17	0.98	0.00	0.59	0.29	0.30
0.3	7014	65723	61	56107	0.11	0.99	0.00	0.56	0.20	0.24
0.4	4757	65761	23	58364	0.08	1.00	0.00	0.55	0.14	0.20
0.5	3082	65778	6	60039	0.05	1.00	9E-05	0.53	0.09	0.16
0.6	1832	65782	2	61289	0.03	1.00	3E-05	0.52	0.06	0.12
0.7	885	65783	1	62236	0.01	1.00	2E-05	0.52	0.03	0.08
0.8	319	65784	0	62802	0.01	1	0	0.51	0.01	0.05
0.9	113	65784	0	63008	0.00	1	0	0.51	0.00	0.03
1	37	65784	0	63084	0.00	1	0	0.51	0.00	0.02

Thr: Threshold

A. 4. Confusion Matrices for NMTF with PPI matrix

A.4.1. The error rates for each matrix for NMTF algorithm with PPI matrix.

	R12	R13	R14	R15	Avg
k1	0.055356	0.037618	0.028647	0.0287	0.03758
k2	0.05509	0.037472	0.028745	0.028391	0.037424
k3	0.055046	0.037071	0.028604	0.028466	**0.037297**
k4	0.055228	0.037821	0.028819	0.028531	0.0376
k5	0.055127	0.037422	0.028771	0.028933	0.037563
k6	0.055414	0.03772	0.0287	0.028916	0.037688
k7	0.055506	0.037917	0.028654	0.028971	0.037762
k8	0.055522	0.037836	0.028721	0.028812	0.037723
k9	0.055591	0.037842	0.028762	0.028758	0.037738
k10	0.055563	0.037867	0.028821	0.029028	0.03782
k11	0.055316	0.037652	0.028915	0.02874	0.037656
k12	0.055076	0.037512	0.028562	0.028238	0.037347
k13	0.055204	0.037659	0.028761	0.028494	0.03753
k14	0.054863	0.037508	0.028762	0.028404	0.037384
k15	0.055227	0.037563	0.028773	0.028449	0.037503
k16	0.055355	0.03774	0.028782	0.028982	0.037715
k17	0.055344	0.037753	0.028714	0.028887	0.037675
k18	0.055405	0.037845	0.028812	0.029025	0.037772
k19	0.055565	0.037968	0.028847	0.028986	0.037841
k20	0.055512	0.037862	0.02887	0.028972	0.037804

Avg: Average

A.4.2. Confusion matrices of NMTF algorithm with PPI matrix.

1. Protein x Cellular Component

a. 10-fold cross-validation

Thr.	TP	TN	FP	FN	Recall	Precision	FPR	Accuracy	F-score	MCC
0	52200	0	52200	0	1	0.5	1	0.5	0.67	
1E-06	52018	7374	44826	182	1.00	0.54	0.86	0.57	0.70	0.27
1E-05	51974	9760	42440	226	1.00	0.55	0.81	0.59	0.71	0.31
0.0001	51871	14361	37839	329	0.99	0.58	0.72	0.63	0.73	0.39
0.001	51434	24488	27712	766	0.99	0.65	0.53	0.73	0.78	0.53
0.01	49394	41211	10989	2806	0.95	0.82	0.21	0.87	0.88	0.74
0.02	47856	45078	7122	4344	0.92	0.87	0.14	0.89	0.89	0.78
0.03	46497	46821	5379	5703	0.89	0.90	0.10	0.89	**0.89**	**0.79**
0.04	45373	47821	4379	6827	0.87	0.91	0.08	0.89	0.89	0.79
0.05	44385	48507	3693	7815	0.85	0.92	0.07	0.89	0.89	0.78
0.06	43442	49013	3187	8758	0.83	0.93	0.06	0.89	0.88	0.78
0.07	42555	49380	2820	9645	0.82	0.94	0.05	0.88	0.87	0.77
0.08	41724	49715	2485	10476	0.80	0.94	0.05	0.88	0.87	0.76
0.09	40967	49974	2226	11233	0.78	0.95	0.04	0.87	0.86	0.75
0.1	40206	50183	2017	11994	0.77	0.95	0.04	0.87	0.85	0.75
0.2	33692	51247	953	18508	0.65	0.97	0.02	0.81	0.78	0.67
0.3	28010	51732	468	24190	0.54	0.98	0.01	0.76	0.69	0.59
0.4	22559	51962	238	29641	0.43	0.99	0.00	0.71	0.60	0.52
0.5	17287	52078	122	34913	0.33	0.99	0.00	0.66	0.50	0.44
0.6	12579	52149	51	39621	0.24	1.00	0.00	0.62	0.39	0.37
0.7	8696	52181	19	43504	0.17	1.00	0.00	0.58	0.29	0.30
0.8	5522	52193	7	46678	0.11	1.00	0.00	0.55	0.19	0.24
0.9	3320	52194	6	48880	0.06	1.00	0.00	0.53	0.12	0.18
1	1873	52199	1	50327	0.04	1.00	0.00	0.52	0.07	0.14

Thr: Threshold

b. 3-fold cross-validation

Thr.	TP	TN	FP	FN	Recall	Precision	FPR	Accuracy	F-score	MCC
0	34140	0	37299	0	1	0.47789	1	0.48	0.65	
1E-06	34028	6673	30626	112	1.00	0.53	0.82	0.57	0.69	0.30
1E-05	33954	9354	27945	186	0.99	0.55	0.75	0.61	0.71	0.36
0.0001	33760	13815	23484	380	0.99	0.59	0.63	0.67	0.74	0.45
0.001	33119	20565	16734	1021	0.97	0.66	0.45	0.75	0.79	0.57
0.01	30963	30896	6403	3177	0.91	0.83	0.17	0.87	0.87	0.74
0.02	29655	33306	3993	4485	0.87	0.88	0.11	0.88	0.87	0.76
0.03	28646	34424	2875	5494	0.84	0.91	0.08	0.88	0.87	0.77
0.04	27854	35039	2260	6286	0.82	0.92	0.06	0.88	0.87	0.76
0.05	27112	35455	1844	7028	0.79	0.94	0.05	0.88	0.86	0.76
0.06	26483	35756	1543	7657	0.78	0.94	0.04	0.87	0.85	0.75
0.07	25840	35978	1321	8300	0.76	0.95	0.04	0.87	0.84	0.74
0.08	25263	36152	1147	8877	0.74	0.96	0.03	0.86	0.83	0.73
0.09	24730	36279	1020	9410	0.72	0.96	0.03	0.85	0.83	0.73
0.1	24165	36393	906	9975	0.71	0.96	0.02	0.85	0.82	0.72
0.2	19519	36997	302	14621	0.57	0.98	0.01	0.79	0.72	0.63
0.3	15512	37185	114	18628	0.45	0.99	0.00	0.74	0.62	0.55
0.4	11578	37242	57	22562	0.34	1.00	0.00	0.68	0.51	0.46
0.5	8090	37279	20	26050	0.24	1.00	0.00	0.64	0.38	0.37
0.6	5055	37291	8	29085	0.15	1.00	0.00	0.59	0.26	0.29
0.7	2867	37295	4	31273	0.08	1.00	0.00	0.56	0.15	0.21
0.8	1539	37296	3	32601	0.05	1.00	0.00	0.54	0.09	0.15
0.9	805	37298	1	33335	0.02	1.00	0.00	0.53	0.05	0.11
1	425	37298	1	33715	0.01	1.00	0.00	0.53	0.02	0.08

Thr: Threshold

2. Protein x Molecular Function

a. 10-fold cross-validation

Thr.	TP	TN	FP	FN	Recall	Precision	FPR	Accuracy	F-score	MCC
0	34270	0	34270	0	1	0.5	1	0.5	0.67	
1E-06	34158	3236	31034	112	1.00	0.52	0.91	0.55	0.69	0.21
1E-05	34123	3986	30284	147	1.00	0.53	0.88	0.56	0.69	0.24
0.0001	34064	5351	28919	206	0.99	0.54	0.84	0.58	0.70	0.28
0.001	33684	10117	24153	586	0.98	0.58	0.70	0.64	0.73	0.38
0.01	30393	23987	10283	3877	0.89	0.75	0.30	0.79	0.81	0.60
0.02	27937	28300	5970	6333	0.82	0.82	0.17	0.82	**0.82**	0.64
0.03	26092	30181	4089	8178	0.76	0.86	0.12	0.82	0.81	**0.65**
0.04	24540	31224	3046	9730	0.72	0.89	0.09	0.81	0.79	0.64
0.05	23354	31878	2392	10916	0.68	0.91	0.07	0.81	0.78	0.63
0.06	22286	32345	1925	11984	0.65	0.92	0.06	0.80	0.76	0.62
0.07	21281	32674	1596	12989	0.62	0.93	0.05	0.79	0.74	0.61
0.08	20425	32917	1353	13845	0.60	0.94	0.04	0.78	0.73	0.60
0.09	19672	33157	1113	14598	0.57	0.95	0.03	0.77	0.71	0.59
0.1	19025	33335	935	15245	0.56	0.95	0.03	0.76	0.70	0.58
0.2	14596	33997	273	19674	0.43	0.98	0.01	0.71	0.59	0.51
0.3	12016	34148	122	22254	0.35	0.99	0.00	0.67	0.52	0.45
0.4	9878	34208	62	24392	0.29	0.99	0.00	0.64	0.45	0.41
0.5	7745	34237	33	26525	0.23	1.00	0.00	0.61	0.37	0.35
0.6	5693	34254	16	28577	0.17	1.00	0.00	0.58	0.28	0.30
0.7	3844	34266	4	30426	0.11	1.00	0.00	0.56	0.20	0.24
0.8	2322	34269	1	31948	0.07	1.00	0.00	0.53	0.13	0.19
0.9	1267	34270	0	33003	0.04	1	0	0.52	0.07	0.14
1	635	34270	0	33635	0.02	1	0	0.51	0.04	0.10

Thr: Threshold

b. 3-fold cross-validation

Thr.	TP	TN	FP	FN	Recall	Precision	FPR	Accuracy	F-score	MCC
0	22609	0	25443	0	1	0.47	1	0.47	0.64	
1E-06	22514	3495	21948	95	1.00	0.51	0.86	0.54	0.67	0.25
1E-05	22478	4317	21126	131	0.99	0.52	0.83	0.56	0.68	0.28
0.0001	22391	5703	19740	218	0.99	0.53	0.78	0.58	0.69	0.33
0.001	22130	9304	16139	479	0.98	0.58	0.63	0.65	0.73	0.43
0.01	20347	18659	6784	2262	0.90	0.75	0.27	0.81	0.82	0.64
0.02	18820	21518	3925	3789	0.83	0.83	0.15	0.84	0.83	0.68
0.03	17609	22785	2658	5000	0.78	0.87	0.10	0.84	0.82	0.68
0.04	16583	23469	1974	6026	0.73	0.89	0.08	0.83	0.81	0.67
0.05	15760	23894	1549	6849	0.70	0.91	0.06	0.83	0.79	0.66
0.06	15043	24223	1220	7566	0.67	0.92	0.05	0.82	0.77	0.65
0.07	14394	24427	1016	8215	0.64	0.93	0.04	0.81	0.76	0.64
0.08	13834	24586	857	8775	0.61	0.94	0.03	0.80	0.74	0.63
0.09	13371	24701	742	9238	0.59	0.95	0.03	0.79	0.73	0.62
0.1	12904	24798	645	9705	0.57	0.95	0.03	0.78	0.71	0.60
0.2	9902	25250	193	12707	0.44	0.98	0.01	0.73	0.61	0.53
0.3	8066	25361	82	14543	0.36	0.99	0.00	0.70	0.52	0.47
0.4	6392	25409	34	16217	0.28	0.99	0.00	0.66	0.44	0.41
0.5	4681	25429	14	17928	0.21	1.00	0.00	0.63	0.34	0.35
0.6	3028	25440	3	19581	0.13	1.00	0.00	0.59	0.24	0.27
0.7	1700	25442	1	20909	0.08	1.00	0.00	0.56	0.14	0.20
0.8	820	25443	0	21789	0.04	1	0	0.55	0.07	0.14
0.9	393	25443	0	22216	0.02	1	0	0.54	0.03	0.10
1	157	25443	0	22452	0.01	1	0	0.53	0.01	0.06

Thr: Threshold

3. Protein x Biological Process

a. 10-fold cross-validation

Thr.	TP	TN	FP	FN	Recall	Precision	FPR	Accuracy	F-score	MCC
0	75880	0	75880	0	1	0.5	1	0.5	0.67	
1E-06	75687	3612	72268	193	1.00	0.51	0.95	0.52	0.68	0.14
1E-05	75624	4532	71348	256	1.00	0.51	0.94	0.53	0.68	0.16
0.0001	75425	6833	69047	455	0.99	0.52	0.91	0.54	0.68	0.20
0.001	74297	16047	59833	1583	0.98	0.55	0.79	0.60	0.71	0.30
0.01	61836	52892	22988	14044	0.81	0.73	0.30	0.76	**0.77**	0.52
0.02	51318	64365	11515	24562	0.68	0.82	0.15	0.76	0.74	**0.53**
0.03	44113	68793	7087	31767	0.58	0.86	0.09	0.74	0.69	0.52
0.04	38961	71064	4816	36919	0.51	0.89	0.06	0.72	0.65	0.50
0.05	35018	72368	3512	40862	0.46	0.91	0.05	0.71	0.61	0.48
0.06	31944	73225	2655	43936	0.42	0.92	0.03	0.69	0.58	0.46
0.07	29397	73785	2095	46483	0.39	0.93	0.03	0.68	0.55	0.44
0.08	27313	74201	1679	48567	0.36	0.94	0.02	0.67	0.52	0.43
0.09	25431	74507	1373	50449	0.34	0.95	0.02	0.66	0.50	0.42
0.1	23713	74726	1154	52167	0.31	0.95	0.02	0.65	0.47	0.40
0.2	13932	75603	277	61948	0.18	0.98	0.00	0.59	0.31	0.31
0.3	9323	75788	92	66557	0.12	0.99	0.00	0.56	0.22	0.25
0.4	6599	75848	32	69281	0.09	1.00	0.00	0.54	0.16	0.21
0.5	4717	75867	13	71163	0.06	1.00	0.00	0.53	0.12	0.18
0.6	3386	75875	5	72494	0.04	1.00	0.00	0.52	0.09	0.15
0.7	2260	75878	2	73620	0.03	1.00	0.00	0.51	0.06	0.12
0.8	1332	75878	2	74548	0.02	1.00	0.00	0.51	0.03	0.09
0.9	666	75880	0	75214	0.01	1	0	0.50	0.02	0.07
1	284	75880	0	75596	0.00	1	0	0.50	0.01	0.04

Thr: Threshold

b. 3-fold cross-validation

Thr.	TP	TN	FP	FN	Recall	Precision	FPR	Accuracy	F-score	MCC
0	55959	0	59220	0	1	0.49	1	0.49	0.65	
1E-06	55708	3420	55800	251	1.00	0.50	0.94	0.51	0.67	0.15
1E-05	55598	4237	54983	361	0.99	0.50	0.93	0.52	0.67	0.17
0.0001	55331	6350	52870	628	0.99	0.51	0.89	0.54	0.67	0.20
0.001	54083	14676	44544	1876	0.97	0.55	0.75	0.60	0.70	0.31
0.01	43706	43561	15659	12253	0.78	0.74	0.26	0.76	0.76	0.52
0.02	35771	51678	7542	20188	0.64	0.83	0.13	0.76	0.72	0.53
0.03	30496	54669	4551	25463	0.54	0.87	0.08	0.74	0.67	0.51
0.04	26788	56082	3138	29171	0.48	0.90	0.05	0.72	0.62	0.49
0.05	24091	56959	2261	31868	0.43	0.91	0.04	0.70	0.59	0.47
0.06	21988	57483	1737	33971	0.39	0.93	0.03	0.69	0.55	0.45
0.07	20258	57850	1370	35701	0.36	0.94	0.02	0.68	0.52	0.43
0.08	18797	58121	1099	37162	0.34	0.94	0.02	0.67	0.50	0.42
0.09	17551	58312	908	38408	0.31	0.95	0.02	0.66	0.47	0.41
0.1	16488	58475	745	39471	0.29	0.96	0.01	0.65	0.45	0.40
0.2	10155	59043	177	45804	0.18	0.98	0.00	0.60	0.31	0.31
0.3	7016	59163	57	48943	0.13	0.99	0.00	0.57	0.22	0.26
0.4	4787	59199	21	51172	0.09	1.00	0.00	0.56	0.16	0.21
0.5	3127	59217	3	52832	0.06	1.00	5E-05	0.54	0.11	0.17
0.6	1814	59220	0	54145	0.03	1	0	0.53	0.06	0.13
0.7	859	59220	0	55100	0.02	1	0	0.52	0.03	0.09
0.8	368	59220	0	55591	0.01	1	0	0.52	0.01	0.06
0.9	163	59220	0	55796	0.00	1	0	0.52	0.01	0.04
1	58	59220	0	55901	0.00	1	0	0.51	0.00	0.02

Thr: Threshold

4. Protein x Disease

a. 10-fold cross-validation

Thr.	TP	TN	FP	FN	Recall	Precision	FPR	Accuracy	F-score	MCC
0	71370	0	71370	0	1	0.5	1	0.5	0.67	
1E-06	71093	2006	69364	277	1.00	0.51	0.97	0.51	0.67	0.10
1E-05	71004	2513	68857	366	0.99	0.51	0.96	0.52	0.67	0.11
0.0001	70739	3748	67622	631	0.99	0.51	0.95	0.52	0.67	0.13
0.001	69343	10545	60825	2027	0.97	0.53	0.85	0.56	0.69	0.21
0.01	54099	49564	21806	17271	0.76	0.71	0.31	0.73	**0.73**	0.45
0.02	43079	61017	10353	28291	0.60	0.81	0.15	0.73	0.69	**0.47**
0.03	36419	65069	6301	34951	0.51	0.85	0.09	0.71	0.64	0.46
0.04	31550	67104	4266	39820	0.44	0.88	0.06	0.69	0.59	0.44
0.05	27825	68294	3076	43545	0.39	0.90	0.04	0.67	0.54	0.42
0.06	24889	69100	2270	46481	0.35	0.92	0.03	0.66	0.51	0.40
0.07	22562	69617	1753	48808	0.32	0.93	0.02	0.65	0.47	0.39
0.08	20594	70012	1358	50776	0.29	0.94	0.02	0.63	0.44	0.37
0.09	18925	70291	1079	52445	0.27	0.95	0.02	0.63	0.41	0.36
0.1	17516	70493	877	53854	0.25	0.95	0.01	0.62	0.39	0.35
0.2	9934	71183	187	61436	0.14	0.98	0.00	0.57	0.24	0.27
0.3	6898	71298	72	64472	0.10	0.99	0.00	0.55	0.18	0.22
0.4	5133	71345	25	66237	0.07	1.00	0.00	0.54	0.13	0.19
0.5	3900	71358	12	67470	0.05	1.00	0.00	0.53	0.10	0.17
0.6	2950	71365	5	68420	0.04	1.00	0.00	0.52	0.08	0.14
0.7	2097	71366	4	69273	0.03	1.00	0.00	0.51	0.06	0.12
0.8	1280	71369	1	70090	0.02	1.00	0.00	0.51	0.04	0.10
0.9	534	71370	0	70836	0.01	1	0	0.50	0.01	0.06
1	174	71370	0	71196	0.00	1	0	0.50	0.00	0.03

Thr: Threshold

b. 3-fold cross-validation

Thr.	TP	TN	FP	FN	Recall	Precision	FPR	Accuracy	F-score	MCC
0	63121	0	65784	0	1	0.49	1	0.49	0.66	
1E-06	62680	2549	63235	441	0.99	0.50	0.96	0.51	0.66	0.11
1E-05	62468	3231	62553	653	0.99	0.50	0.95	0.51	0.66	0.11
0.0001	62053	4778	61006	1068	0.98	0.50	0.93	0.52	0.67	0.13
0.001	60447	11844	53940	2674	0.96	0.53	0.82	0.56	0.68	0.22
0.01	47647	46016	19768	15474	0.75	0.71	0.30	0.73	0.73	0.45
0.02	39036	55891	9893	24085	0.62	0.80	0.15	0.74	0.70	0.48
0.03	33661	59529	6255	29460	0.53	0.84	0.10	0.72	0.65	0.47
0.04	29808	61472	4312	33313	0.47	0.87	0.07	0.71	0.61	0.46
0.05	26684	62627	3157	36437	0.42	0.89	0.05	0.69	0.57	0.44
0.06	24194	63396	2388	38927	0.38	0.91	0.04	0.68	0.54	0.43
0.07	22146	63921	1863	40975	0.35	0.92	0.03	0.67	0.51	0.41
0.08	20363	64324	1460	42758	0.32	0.93	0.02	0.66	0.48	0.40
0.09	18893	64588	1196	44228	0.30	0.94	0.02	0.65	0.45	0.39
0.1	17564	64818	966	45557	0.28	0.95	0.01	0.64	0.43	0.38
0.2	10229	65567	217	52892	0.16	0.98	0.00	0.59	0.28	0.29
0.3	6729	65717	67	56392	0.11	0.99	0.00	0.56	0.19	0.24
0.4	4520	65767	17	58601	0.07	1.00	0.00	0.55	0.13	0.19
0.5	3057	65781	3	60064	0.05	1.00	5E-05	0.53	0.09	0.16
0.6	1814	65784	0	61307	0.03	1	0	0.52	0.06	0.12
0.7	873	65784	0	62248	0.01	1	0	0.52	0.03	0.08
0.8	324	65784	0	62797	0.01	1	0	0.51	0.01	0.05
0.9	107	65784	0	63014	0.00	1	0	0.51	0.00	0.03
1	36	65784	0	63085	0.00	1	0	0.51	0.00	0.02

Thr: Threshold