Aalto University
School of Science
Degree Programme in Computer, Communication and Information Sciences

Heba Sourkatti

Principal Metabolic Flux Mode Analysis

Master's Thesis
Espoo, October 2, 2018

Supervisor: Professor Juho Rousu, Aalto University
Advisor: Dr. Sahely Bhadra, Indian Institute of Technology (IIT), Palakkad

Aalto University
School of Science
Degree Programme in Computer, Communication and Information Sciences

ABSTRACT OF MASTER'S THESIS

Author:	Heba Sourkatti		
Title:	Principal Metabolic Flux Mode Analysis		
Date:	October 2, 2018	**Pages:**	56
Major:	Machine Learning and Data Mining	**Code:**	SCI3044
Supervisor:	Professor Juho Rousu, Aalto University		
Advisor:	Dr. Sahely Bhadra, Indian Institute of Technology (IIT), Palakkad		

In recent years, much progress has been achieved in the computational analysis of the metabolic networks, as a consequence of the rapid growth of the omics database. However, current literature analysis algorithms still lack good biological interpretability of the analysis results. Moreover, they can not be applied on a whole-genome level.

This thesis assesses the potential of the Principal Metabolic Flux Mode Analysis (PMFA). The PMFA is a novel algorithm that was recently developed, which aims to improve the interpretability of Principal Component Analysis (PCA), through including a stoichiometric regularization to the PCA objective function. The PMFA can determine the flux modes that explain the highest variability in the network and it can also scale-up to a whole-genome level using the sparse version of PMFA. Furthermore, this thesis compares the PMFA to the recent approach Principal Elementary Mode Analysis (PEMA), which also tries to enhance the PCA interpretability. However, this approach is computationally heavy and thus fails to handle the large-scale networks (e.g., whole-genome). In order to further determine the feasibility of the PMFA approach for the analysis of metabolism, a Graph-regularized Matrix Factorization (GMF) was developed analogous to PMFA framework, similarly by adding the network stoichiometric matrix to a graph-structured matrix factorization framework.

The results illustrate the potential of PMFA as a metabolic network analysis for identifying fluxes that explain maximum variation in the network and it can be used to analyze whole-genome level. In addition, the results showed that GMF method performed well in predicting active Elementary Modes (EMs) on simulated data but failed to work on large networks, while PEMA had the lowest performance among all methods. Based on the results, future work can be conducted to improve the GMF approach in terms of genome-scale analysis through including sparsity.

Keywords:	metabolic network analysis, PCA, elementary modes, stoichiometric modelling, matrix factorization.
Language:	English

Acknowledgments

All praise is due to Allah, the most gracious and most merciful.

Firstly, I would like to express my sincere gratitude is to my supervisor Prof. Juho Rousu, for his continuous support during this thesis work. He has always brought both new and interesting ideas to the table, delivered with patience, motivation, kindness and fruitful discussions.

Besides my supervisor, I would like to genuinely thank Dr. Sahely Bhadra, for her invaluable and friendly advice during this thesis work. Although she was living across the globe, she was always open and collaborative whenever I ran into difficulties.

My sincere thanks also go to Jonathan Strahl, for his kindness, easy-going, approachable character, and for his diving through almost every theory, which helped me to establish fundamental understanding.

I dedicate this work to my family, for their continuous and unparalleled love and support. I am forever indebted to my parents for giving me the opportunities and experiences that made me who I am, and to my brothers Mohammed, Gamal Eldin and Aladdin and my little sister Fatima. My earnest thanks to Amal, my dear friend and my partner in journey.
To you all, I dedicate this.

Espoo, October 2, 2018

Heba Sourkatti

Abbreviations and Acronyms

PMFA	Principal Metabolic Flux Mode Analysis
GENRE	Genome-scale Network Reconstruction
SNA	Stoichiometric Network Analysis
PCA	Principal Component Analysis
SVD	Singular Value Decomposition
MF	Matrix Factorization
PEMA	Principal Elementary Mode Analysis
EMs	Elementary Modes
GMF	Graph-regularized Matrix Factorization
PCs	Principal Components
MCR-ALS	Multivariate Curve Resolution Alternating Least Squares
SPMFA	Sparse Principal Metabolic Flux Mode Analysis
LOO	Leave-One out cross validation
ROC	Receiver Operating Characteristic
AUC	Area Under the ROC Curve
AUPR	Area Under Precision-Recall curve
AP	Average precision
tp	true positive
fp	false positive
tn	true negative
fn	false negative
tpr	true positive rate
fpr	false positive rate
FoV	Fraction of Variance
MSE	Mean-Square Error
CCP	Concave Convex Procedure

Contents

Abbreviations and Acronyms **4**

1 Introduction **7**
1.1 Problem statement 8
1.2 Structure of the Thesis 8

2 Literature Review **9**
2.1 Metabolic Networks 9
2.1.1 Stoichiometric Matrix 11
2.1.2 The (Right) null space 12
2.1.3 Network Boundaries 13
2.1.4 The Flux Cone 14
2.1.5 Elementary Modes (EM) 16
2.2 Methodological Background 18
2.2.1 Principal Component Analysis (PCA) 18
2.2.2 Matrix Factorization (MF) 21
2.2.3 Multivariate Curve Resolution Alternating Least Squares (MCR-ALS) 22
2.2.4 Principal Elementary Mode Analysis (PEMA) 23

3 Methods and Data **26**
3.1 Datasets used in experiments 26
3.2 Principal Metabolic Flux Mode Analysis (PMFA) 27
3.3 Graph-regularized Matrix Factorization (GMF) 30
3.3.1 Model selection 34

4 Results and Discussion **35**
4.1 Active elementary flux modes (EMs) prediction 35
4.2 Explained variance 40
4.3 Sparse flux modes recovery 42

5 Conclusion **44**

A Appendix **50**

Chapter 1

Introduction

Molecular biology has been the focus of intense study due to its wide range of potential applications, such as disease diagnosis as well as the production of new drugs, fine chemicals, industrial enzymes and biofuel [8, 22, 44]. However, the birth of Systems Biology which requires a full understanding of the whole biological system has led to the development of numerous Omics techniques with a massive data sets, used to identify all elements of the studied level (metabolites, proteins, transcripts and genes) [32]. Since then, the concern of some scientists was to study and use those massive sets in addition to biologically interpret them.

Assembling biochemical reaction networks is one of the approaches to study Omics data-sets by capturing all interactions between all components [32]. The scope of this thesis is focused on metabolic networks analysis. Many approaches were developed for this purpose, one of the first approaches were Stoichiometric Network Analysis (SNA) which is used to identify unique pathways in the network. Different algorithms were developed by modifying SNA such as Elementary Mode Analysis and Extreme Pathway Analysis [43]. However, those approaches fail to work on a large set of samples and can perform only on a single sample at a time.

Consequently, metabolic networks analysis has introduced dimensionality reduction techniques for the analysis of large sample sets. Principal Component Analysis (PCA) [36], Singular Value Decomposition (SVD) [4] and Matrix Factorization (MF) [11] are some of the used dimensionality reduction approaches. Although, PCA has been widely used in molecular biology and can extract the systematic variation in the dataset. PCA results are dense and difficult to be interpreted in terms of biology. Moreover, PCA deals with the reactions independently and ignores the underlying struc-

ture of the network (i.e. the connections between reactions). Accordingly, some approaches were proposed to enhance the PCA interpretability such as Principal Elementary Mode Analysis (PEMA) [16], where it identifies a prior set of Elementary Modes (EMs) as PCs candidates. Hence, it requires prior assumptions and heavy computations in the case of large networks. Therefore, PEMA fails to work on a genome-scale network. Recently, Principal Metabolic Flux Mode Analysis (PMFA) [9] was developed to improve the PCA interpretability while maintaining its applicability on genome-scale networks. Therefore, this thesis determines the feasibility of PMFA methodology on different datasets and compares it to other approaches. Additionally, for the PMFA evaluation, a Graph-regularized matrix factorization (GMF) was developed to analyze metabolic networks. GMF approximates the target matrix by two matrices and the objective is to minimize the approximation error. Furthermore, GMF in this thesis involves the stoichiometric matrix as a side information, which was added as a graph.

1.1 Problem statement

The aim of this thesis is to evaluate the Principal Metabolic Flux Mode Analysis algorithm and assess its performance compared to different techniques. In order to analyze metabolic networks and identify the flux modes in the network, the algorithm combines PCA with stoichiometric network analysis. Additionally, this thesis develops a form of Matrix Factorization for the analysis of metabolic networks, that tries to replicate Principal Metabolic Flux Mode Analysis framework by adding additional information about the network structure and considering the reactions directions. Furthermore, the thesis evaluates the GMF algorithm on different datasets.

1.2 Structure of the Thesis

The rest of this thesis is structured as follows. Chapter 2 summarizes some of the vital biological backgrounds in order to understand the biological concepts relevant to this thesis, as well as, reviews some of the literature methods that are used in metabolic network analysis. Chapter 3 reports the datasets used in this thesis. In addition, Chapter 3 presents the two main algorithms that are evaluated in this thesis and defines their regularized optimization framework. Chapter 4 analyzes the results of the experiments and assesses the findings. Finally, Chapter 5 concludes the thesis with a discussion.

Chapter 2

Literature Review

This chapter is divided into two main sections, *Section 2.1: Metabolic networks* and *Section 2.2: Methodological background*. Firstly, in order to cover the fundamental biological aspects within the scope of this thesis, section 2.1 briefly represents the metabolic network and its mathematical representation in subsection 2.1.1. Subsection 2.1.2 describes the right null space of the stoichiometric matrix S and the network boundaries are defined in subsection 2.1.3. Additionally, subsection 2.1.4 previews stoichiometric models and the effect of the different constraints added to the model where the feasible flux distribution is represented as a cone. The section ends with subsection 2.1.5. This subsection defines the networks pathways and the different types of network-based pathways. Also, it introduces the Elementary Modes (EMs) and discusses how they differ from the extreme pathways.

Secondly, section 2.2 reviews the basic methods as well as the most recent developed algorithms used in the analysis of omics data metabolism. Subsection 2.2.1 summarizes Principal Component Analysis (PCA) approach, PCA formulation techniques and the application of PCA in bioinformatics data. Subsection 2.2.2 describes Matrix Factorization (MF) method and the connection between PCA and MF. Finally, Subsections 2.2.3 and 2.2.4 review the recently developed algorithms for the analysis of metabolism.

2.1 Metabolic Networks

One of the pivotal steps to diagnose new diseases and develop drugs is to understand the physiological and behavioral activity of the human body. Such activities can be measured from the various biological networks, where the

most characterized network among them is the metabolic network. During the last decades, modeling biological networks and especially the reconstruction of metabolic networks has expanded rapidly from small scales to a whole genome scale [1]. Figure 2.1 shows the steps for the genome-scale metabolic reconstruction and some possible uses. The process starts with the modeling step, where the model is built using the information gathered from literature and gene-annotation data. The second step is to convert this information into a mathematical model. Finally, the mathematical model is computationally analyzed.

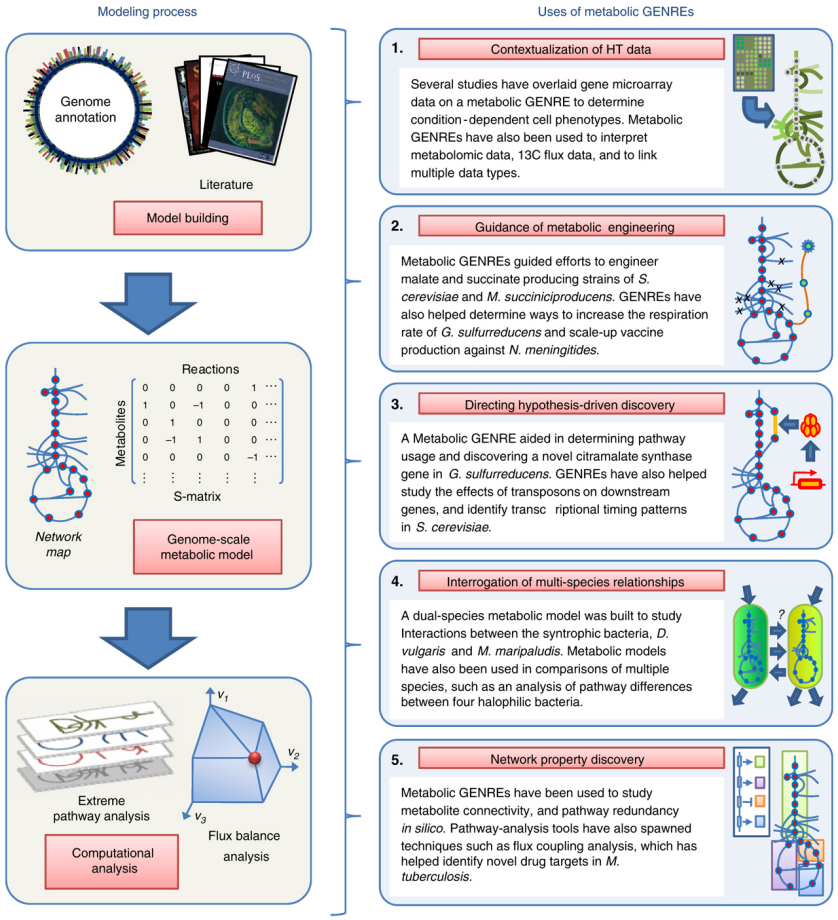


Figure 2.1: The metabolic Genome-scale Reconstructions (GENREs [7]) steps and some applications. The figure is taken from [1]

The metabolic network is composed of chemical reactions that can be represented as equations. These chemical equations have all the stoichiometry

information about the network and it can be represented in a matrix form known as stoichiometric matrix S. Accordingly, the network properties are resolved by mathematically analyzing the stoichiometric matrix.

2.1.1 Stoichiometric Matrix

The stoichiometric matrix contains the stoichiometry of the reactions, where columns represent reactions (constrained by chemical rules e.g. elemental balancing) and rows correspond to metabolites. As a result, each row comprises all reactions that involved the equivalent metabolite and hence illustrates the connectivity between reactions. The matrix entities will reflect the production or the consumption of the compound in the corresponding reaction.

Furthermore, the stoichiometric matrix in mathematical framework represents a linear transformation of the flux vector to a time derivatives vector of the compound concentrations. Where v is the flux vector (consists of the n reactions rates) and x is the concentration vector (consists of m metabolites):

$$v = (v_1, v_2, \cdots, v_n) \tag{2.1}$$
$$x = (x_1, x_2, \cdots, x_m) \tag{2.2}$$

In metabolic networks, the metabolites concentrations change over time, this is shown in the S linear mapping of v to $\frac{dx}{dt}$:

$$\frac{dx}{dt} = Sv \tag{2.3}$$

where the network functional states are identified by the *dynamic mass balances* that are represented by equation 2.3. Additionally, biochemical moieties and electric charge are conserved in the stoichiometric matrix as well as the chemical elements must be balanced. Systems Equation 2.3 can be rewritten for each system:

$$\frac{dx_i}{dt} = \sum_k s_{ik} v_k \tag{2.4}$$

Equation 2.4 shows that metabolite x_i is composed or degraded from the summation of all fluxes v_k.

2.1.2 The (Right) null space

The linear transformation of S forms fundamental four subspaces, two flux spaces (row space and null space) and two concentration spaces (column space and left-null space). Figure 2.2 shows the linear transformation and the four subspaces.

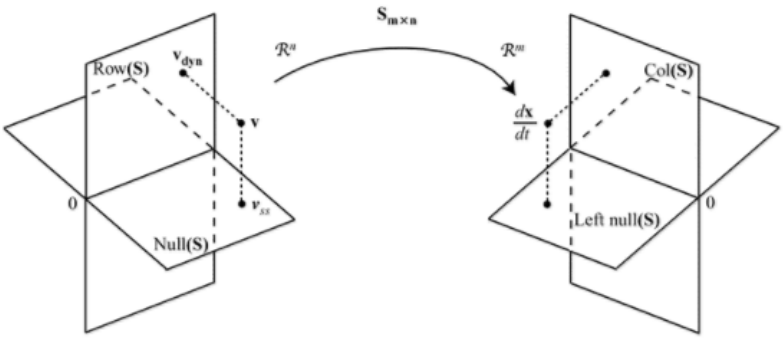


Figure 2.2: The S four subspaces and the linear transformation. The figure is taken from [32]

Understanding the four subspaces reveals essential system properties and provide elaborated interpretation of the networks in terms of dynamics, time-invariant and steady-state features.

The state where the metabolites concentrations are at equilibrium is known as the steady state. Typically, the production fluxes and consumption fluxes are identical and hence the metabolites concentration is constant. It can be represented as [37]:

$$Sv = 0 \tag{2.5}$$

The null space of S consists of all vectors that satisfy equation 2.5, forming a linear homogeneous set of equations. There is a matrix R that spans the

null space, where the columns of R denote a set of basis vectors r_i that satisfy:

$$SR = 0 \tag{2.6}$$

where S is composed of row vectors and R is composed of column vectors. Each set of basis vectors has a unique weight w_i for a certain v. However, there is not a unique set of basis vectors. The importance of null space resides in the fact that all the steady state pathways in a metabolic network are defined by the R basis set. These pathways delineate the connection between the inputs and outputs in the network while the sum of the compound concentrations is constant over time.

2.1.3 Network Boundaries

The stoichiometric matrix has different forms rely on the network scope. The internal stoichiometric matrix S_{int} contains m internal metabolites x_i and n internal reactions v_i that describes all the reactions that happened in the cell, whereas the exchange stoichiometric matrix S_{exch} adds the exchange reactions b_i without considering the external metabolites c_i, which in turn allows the metabolites to transfer in and out of the cell boundary. In addition, the total stoichiometric matrix S_{tot} includes the external metabolites c_i.

$$S_{tot} = \begin{array}{c} \\ x_i \\ \\ c_i \end{array} \begin{array}{c} \begin{array}{cc} v_i & \quad b_i \end{array} \\ \left(\begin{array}{cc|cc} & & & \\ & & & \\ \multicolumn{2}{c|}{0} & & \end{array} \right) \end{array}$$

The partitioning of the flux vector to internal and external fluxes defines the boundaries of the network. The network *systems boundary* can be open as in the exchange stoichiometric matrix or closed system, for instance, the internal and total stoichiometric. Illustration of the open and closed networks is shown in figure 2.3.

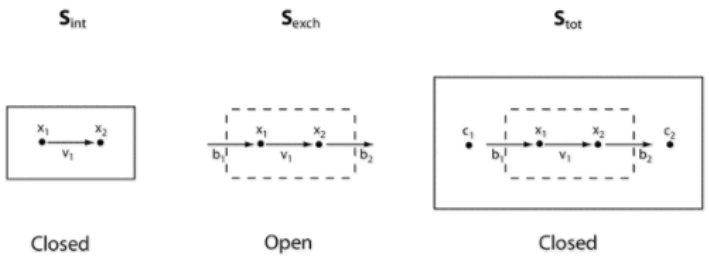


Figure 2.3: Open and closed metabolic networks. The figure is taken from [32]

2.1.4 The Flux Cone

The cellular metabolism and biochemistry can be represented by the stoichiometric models that are based on the *dynamic mass balances* equations 2.3 [26]. The model can include additional constraints for different analysis purposes. For example, in order to determine the network steady-state pathways, the time derivative in the mass balances system is relaxed to zero constraining the determined flux distribution (equation 2.5). The added constraints will specify the feasible flux distributions space, as shown in Figure 2.4. [27] Fig-

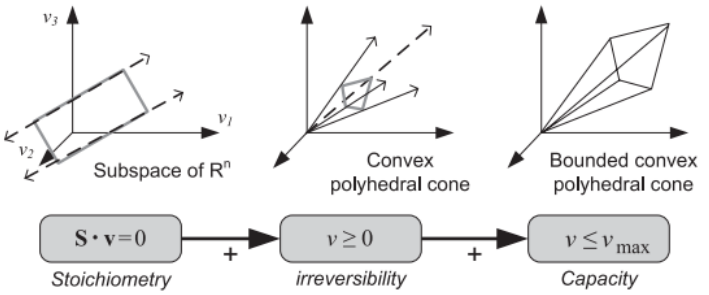


Figure 2.4: Space of feasible steady-state flux distributions. The figure is taken from [27]

ure 2.4 presents the null space at the left, where space is a hyperplane and the middle plot shows the space with an added irreversibility constraint for irreversible fluxes that flow in one direction and hence they need to be non-negative. Inequality constraint converts the problem from a simple linear algebra to a convex analysis and the resulting space of the flux distribution is a convex polyhedral cone. The last space demonstrates the model when

a third constraint is incorporated for the maximum flux values obtained from the enzyme or transporters capacity, in this case, the fluxes values are bounded between 0 and the maximum value constraint forming a bounded convex polyhedral cone.

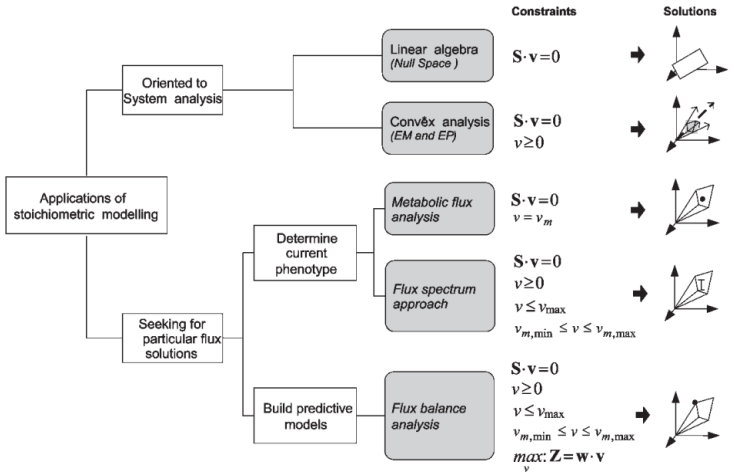


Figure 2.5: Applications of stoichiometric modelling. The figure is taken from [27]

As mentioned earlier each set of constraints serves various purposes (illustrated in Figure 2.5). For instance, the convex analysis allows to include the irreversible fluxes. Consequently, Elementary Modes Analysis and Extreme Pathways Analysis both use convex analysis to identify all the steady-state flux distributions of a metabolic network by producing a unique convex set of vectors, as well as, characterizing the minimal set of systematic pathways. These two methodologies are used for the analysis of pathways.

2.1.5 Elementary Modes (EM)

The smallest elements of a metabolic network are the reactions, where a metabolite is changed into another compound catalyzed by an enzyme. The products of the reactions can be a reactant of another reaction composing a series of consecutive chemical reactions, called *pathways* [30]. Figure 2.6 shows the evolution of pathways from simple reactions and then from pathways to networks.

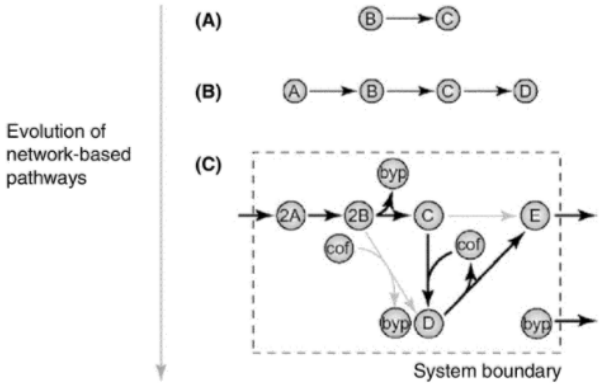


Figure 2.6: The development from reactions (A) to pathways (B) to network (C). The figure is taken from [32]

As mentioned earlier, the constraint-based null space defines the fluxes distribution space with a convex polyhedral cone, where the edges of the cone represent the *extreme pathways*. This can be described mathematically as:

$$C = \{v : v = \sum_{i=1}^{p} \alpha_i \mathbf{p}_i, \alpha_i \geq 0, \text{ for all i}\} \tag{2.7}$$

where C is the flux cone, $\mathbf{p}_i$ are the extreme pathways and α_i are the weights. Another *network-based pathway* is the **elementary modes** defined by S.Schuster [38]. They are similar to extreme pathways and usually both terms used to refer to the same pathways. However, they differ in the reactions (exchange fluxes) representation. The different representations (shown in Figure 2.7) leads to variant polytope forms.

Figure 2.7 illustrates the difference between the extreme pathways and the elementary modes. There are only three extreme pathways whereas, for the same network there are four elementary modes. Having the first three

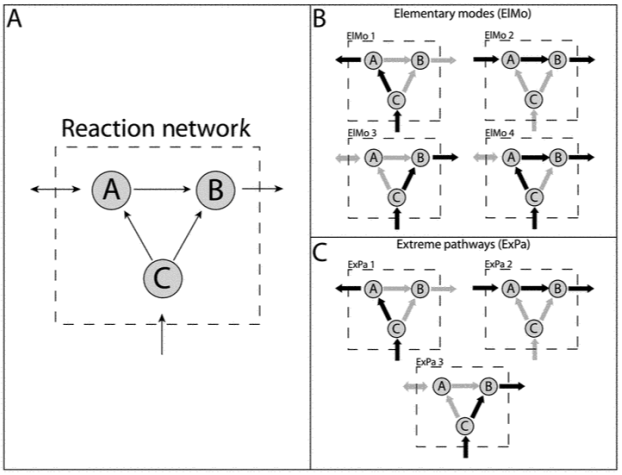


Figure 2.7: The difference between extreme pathways and elementary modes for a simple network, Four elementary modes shown in B) that are similar to extreme pathways in shown C) and the fourth EIMo is a combination of the first two pathways. The figure is taken from [32]

EMs equivalent to the extreme pathways, while the fourth EM is achieved by the nonnegative combination of $EIMo_1$ and $EIMo_2$. By considering the reversible exchange flux of compound A, the number of elementary modes becomes higher than the extreme pathways. It is noticeable that the extreme pathways are just a subset of the elementary modes.

2.2 Methodological Background

This section summarizes some of the widely used approaches for the metabolism analysis and highlights some of the recently developed algorithms.

2.2.1 Principal Component Analysis (PCA)

Principal Component Analysis is a simple unsupervised non-parametric method that retrieves important information from complex data sets. Due to the effectiveness and simplicity of PCA, it has always been one of the most used technique in data analysis. It can be utilized for different purposes, such as dimensionality reduction, feature extraction or data visualization. Basically, PCA algorithm is derived from two different viewpoints, the first viewpoint is based on the variance maximization and the second one is based on the minimization of the approximation cost [3, 10, 40].

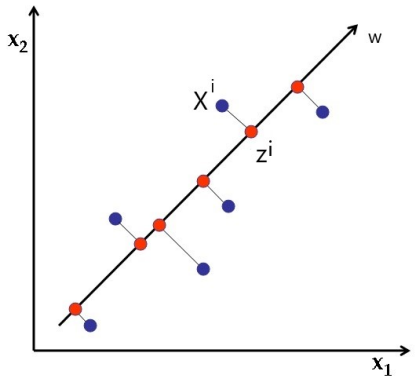


Figure 2.8: The projection of the original data x to the to z

Variance Maximization

PCA projects the original high dimensional data ($\{x^t \in \mathbb{R}^d\}_{t=1}^n$, $x \sim \mathcal{N}(0, \Sigma)$) to a new lower dimensional space ($k \ll d$):

$$z^t = w^T x^t, t = 1, ..., n$$

In matrix form, stack the data into a matrix $X \in \mathbb{R}^{d \times n}$:

$$Z = W^T X, \text{ with } W = (w_1, w_2, ..., w_k)$$

Where, $z^t \in \mathbb{R}^k$ known as principal components (PC) and the projection matrix $W \in \mathbb{R}^{d \times k}$ with vectors w_i called PC directions.

The linear projection is on the direction of w in such away that it will maximize the information retained by $z^t \in \mathbb{R}^k$ about data $x^t \in \mathbb{R}^d$. Therefore [10]:

$$\begin{aligned} w_1 &= \arg\max w^T \Sigma w \\ &\text{s.t. } \|w\| = 1 \end{aligned} \tag{2.8}$$

where, $\Sigma = \frac{1}{n} \sum_{t=1}^{n} x^t (x^t)^T$ and $\|w\| = w^T w = 1$.

Approximation-Error Minimization

In this approach, the data point x_i is approximated by a low-dimensional approximation:

$$x_i \approx \sum_{j=1}^{k} w_j z_{ij} \tag{2.9}$$

where

$$z_{ij} = w_j^T x_i$$

$(w_1 \dots w_d)$ is an orthonormal basis of k-dimensional subspace. w is chosen to minimize the approximation error [10]:

$$\frac{1}{n}\sum \|x-\hat{x}\|_2^2 = \frac{1}{n}\sum_{i=1}^{n} \| \sum_{j=k+1}^{d} w_j z_{ij}\|_2^2 = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=k+1}^{d} (z_{ij})^2$$
$$= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=k+1}^{d} w_j^T x_i {x_i}^T w_j = \sum_{j=k+1}^{d} w_j^T \Sigma w_j \quad (2.10)$$

PCA in bioinformatics

The emergence of genomics data in bioinformatics introduced high dimensional measurements including gene-expression data. PCA is commonly used to reduce the gene-expression dimensions and efficaciously describe the variation of gene-expressions. PCA lower dimensional gene expressions can be utilized for many applications that were inapplicable on the high dimensional gene expression data. For instance, Gene-expression visualization, clustering genes or samples, regression analysis and many other applications. Unfortunately, Principal Components (PCs) of gene expressions data are not biologically interpretable owing to the fact that the PCs are a linear combination of a very high number of genes [28, 29]. In addition, some researchers stated that PCA is not suitable for gene expression analysis since it clusters genes into non-overlapping groups. In fact, genes might be present in different reactions. Moreover, researchers believe that negative values of PCA oppose physics and more complex to be interpreted [18].

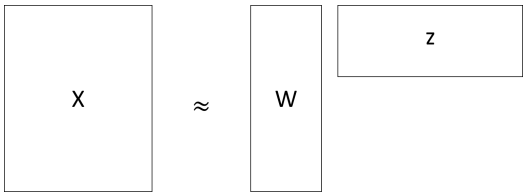
Consequently, many recent studies were conducted to improve the interpretability of PCA for different applications in the bioinformatics field. PEMA is an example of modified PCA for better interpretability on fluxomics data in terms of biological, this method is discussed later in section 2.2.4. Another example of modified PCA in bioinformatics is the sparse non-negative generalized PCA that is applied to metabolomics. This algorithm defines a generalized form of PCA to be a least-square matrix decomposition. Further, a kernel smoother is included to incorporates the structural dependencies by smoothing the distances between the variables. Additionally, sparsity is employed through adding l_1 norm regularization and controlled by regularization parameter. This method has reported good results for clustering high-dimensional data, as well as, better performance than the PCA in dimensionality reduction and the explained variance [2].

Other researchers prefer to employ alternative approaches to overcome the PCA interpretability challenge, including *matrix factorization* and *multivariate curve resolution-alternating least square* that are discussed in the following sections 2.2.2 and 2.2.4 respectively.

2.2.2 Matrix Factorization (MF)

Matrix Factorization is a dimensionality reduction technique that is mostly used in Latent factor models for solving *collaborative filtering* problems [25]. The mathematical formulation of Matrix Factorization is quite simple. It approximates the original data by multiplying two low-rank matrices. Basically, the low-rank factorization of a given data set X is formed as follows:

$$X = WZ^T + E \tag{2.11}$$



where, $W \in \mathbb{R}^{d\times k}, Z \in \mathbb{R}^{n\times k}$,E is the noise and $k \ll d$. It is noticeable that equation 2.11 above is similar to equation 2.9. Thus, PCA can be denoted as a matrix factorization and particularly as an orthogonal matrix factorization.

Then the objective function used to find the optimal solutions for the W and Z matrices can be obtained from equation 2.11 as follows [39]:

$$Obj = \min_{W,Z} \|X - WZ^T\|_F^2 \tag{2.12}$$

However, since this model is a predictive model that learns from a training set and be evaluated on a different unseen test data, it is then prone to overfitting. In order to avoid the possible overfitting, a regularization was added to the model:

$$\min_{W,Z} \|X - WZ^T\|_F^2 + \lambda(\|W\|_F^2 + \|Z\|_F^2) \tag{2.13}$$

This minimization problem can be solved by two approaches, Stochastic Gradient Descent, and Alternating Least Squares [25].

MF in bioinformatics

Matrix Factorization has been widely used in bioinformatics particularly as a clustering tool for genes. Conversely, other clustering techniques have many drawbacks in terms of gene clustering, such as clustering genes based on their global similarities and fail to include local behavior, as well as, grouping genes into a single cluster whereas genes may participate in many clusters [18, 24]. Although Matrix Factorization had overcome these limitations, MF does not have a unique solution. For this reason, MF was improved by enforcing sparseness to both basis factors and encoding vectors [18]. Other researches also promote to include different constraints to enhance the MF performance [19].

2.2.3 Multivariate Curve Resolution Alternating Least Squares (MCR-ALS)

Due to the many disadvantages of the PCA interpretability on biological data and specifically on analyzing fluxomics, a more reasonable approach was proposed to identify flux distribution and to determine pathways such as *Multivariate Curve Resolution-Alternating Least Squares*. MCR-ALS is originally used in chemical analysis and well known as a chemometric method. It performs a linear decomposition of the original matrix into two matrices [5]:

$$X = CP^T + E \tag{2.14}$$

where X is the original data, C and P are the decomposed matrices and E is the experimental error matrix.

The algorithm implements an ALS iterative process by estimating one matrix and then optimizing the other side. In order to improve the optimization results, the algorithm introduces additional information about the system by adding constraints(e.g. Non-negativity, selectivity) to the mathematical model [14]. Equation 2.14 presents a similar formula as shown in

matrix factorization equation 2.11, it can be seen that the MCR is a matrix factorization method.

For the flux analysis, the MCR method can be used by defining the flux distribution of a metabolic network as the linear combination of the pathways present in the network. Additionally, model constraints are allowed to be added. Since pathways do not have to be orthogonal, thus grant more sensible results in the biological view point [17]. Equation 2.14 in [17] is defined as follows: P columns represents the modeled pathways and C is the contribution of modeled pathways in different scenarios. Unlike PCA, MCR can not preselect the number of components, this number can be obtained by initially applying PCA [41] or SVD [21] on the data set.

Further, MCR algorithm can optionally include one of the following constraints: 1) non-negativity: the pathways and their contributions are set to be non-negative, 2) closure on contributions: the balance of mass is achieved for a closed system by setting the sum of the active pathways in each scenario (C matrix rows) to be 1 and 3) selectivity: is a selective constraint for the contribution of pathways in some scenarios, where the selected pathway contributions will be multiplied by 1 and the non-selected pathways relative contributions will be multiplied by 0. This constraint is usually added to reduce the noise. The inclusion of different constraints is an advantage to MCR-ALS over PCA. However, MCR-ALS method fails to determine the total flux flowing through a pathway and hence scenarios comparison cannot be achieved [17].

2.2.4 Principal Elementary Mode Analysis (PEMA)

Principal Elementary Mode Analysis (PEMA) is a recently developed method used on fluxomics data to determine the active elementary flux modes (EMs) in a metabolic network. The algorithm is based on the PCA method and improves the PCA biological interpretability. PEMA pre-selects the EMs as the PCs candidates, this is achieved by the following model:

$$X = \Lambda P_{EM}^{T} + F \tag{2.15}$$

where $X \in \mathbb{R}^{d \times n}$ is the flux dataset with n fluxes, $P_{EM} \in \mathbb{R}^{n \times k}$ is the Principal elementary mode matrix with k pre-selected EMs, $\Lambda \in \mathbb{R}^{d \times k}$ is the

weighting matrix and $F \in \mathbb{R}^{d \times n}$ is the residual matrix [16].

The P_{EM}s loadings are computed from equation 2.15:

$$\Lambda = XP_{EM}(P_{EM}^T P_{EM})^{-1} \tag{2.16}$$

where the weightings are calculated for each EM and then the explained variance (EV) by each EM is computed as follows:

$$EV = 100\%(\|X\|^2 - \|F\|^2)/(\|X\|^2) \tag{2.17}$$

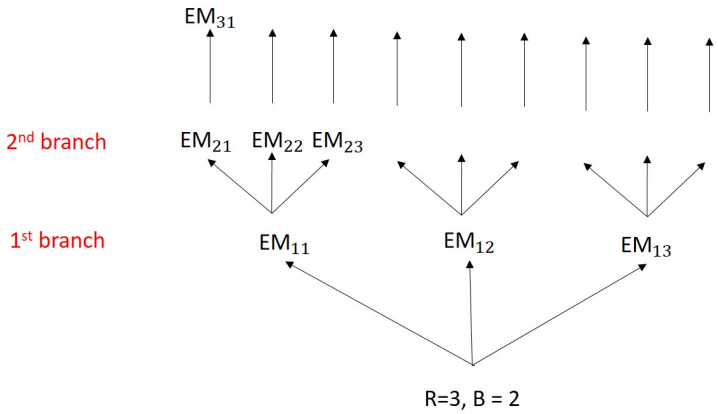


Figure 2.9: Example for the PEMA selection procedure using relaxation parameter $R = 3$ and branch parameter $B = 2$

After sorting EMs by their EV, the first P_{EM} is selected to be the EM with the highest EV value. Then the explained variance by the first P_{EM} and each EM is computed and the second P_{EM} will be the pairs with the most explained variance after sorting the EMs again by the EV, this is repeated for all EMs. Clearly, the selection of the first P_{EM} dominates the upcoming P_{EM}'s and hence their explained variance. Since the EMs are not orthogonal. Therefore, the selected EMs might not be the best.

In order to enhance the selection procedure, PEMA algorithm employed two tuning parameters "Relaxation R" and "Branch number B". The relaxation R specify the number of EM selected at each stage and the branch

number B determine till which (P_{EM}) level the relaxation is applied. For instance, if $R = 3$ and $B = 2$ as shown in figure 2.9, PEMA in the first stage (1st Branch for identifying the 1st P_{EM}) will select 3 EMs (Relaxation = 3) that capture the highest EV. The second stage (2nd Branch) PEMA will select 3 EMs for each EM that was previously selected. For the following selections, only one EM will be chosen for each EM. As a result, for each P_{EM} identification there will be various combinations of EMs with different EV that might explain more variance.

The PEMA algorithm successfully identified the active EMs for experimental and simulated data sets, demonstrating biologically significant results. Unfortunately, the algorithm fails to work on genome-scale networks due to the explicit computation that is needed to be run for all EMs [16].

Chapter 3

Methods and Data

This chapter summarizes the data in section 3.1 that were used for the experiments. Then it describes the main methods of this thesis Principal Metabolic Flux Mode Analysis (PMFA) and Graph-regularized Matrix Factorization (GMF) in sections 3.2 and 3.3, respectively. Both methods are modifications of the previously discussed techniques where PMFA is a PCA modification and GMF is a modification of matrix factorization.

3.1 Datasets used in experiments

The experiments were conducted on different datasets. A simulated data was used to compare the obtained results with the ground truth known by the simulated model, whilst an experimental data set was used to evaluate real-world data performance. Additionally, a whole genome data set was used to assess the efficiency of the algorithms on a large scale.

Simulated Data

The simulated data was taken from [45]. This set simulates the metabolic network of Pichia Pastoris growth on glucose, glycerol and methanol (proposed in [42]). It describes the fundamental catabolic pathways of the Pichia Pastoris with 45 metabolites and 44 reactions. Reactions {2-8, 15, 22-27, 29, 34 and 41} are reversible reactions while the rest are irreversible reactions. The internal stoichiometric matrix consists of 36 internal metabolites and the 44 reactions. The flux data was generated in [45] based on the 98 EMs provided by [42], through simulating 12 experimental scenarios resulting in 16 active EMs {1, 3, 7, 12, 13, 14, 16, 19, 20, 22, 23, 24, 28, 32, 33 and 37}.

Experimental Data

Saccharomyces Cerevisiae metabolic network was chosen as experimental data taken from [45]. The network was reconstructed in [20], it represents the metabolism of the Saccharomyces Cerevisiae and contains glycolysis, the pentose phosphate pathway, anaplerotic carboxylation, fermentative pathways, the TCA cycle, malic enzyme as well as the anabolic pathways. The set comprises 42 metabolites and 47 reactions with a total number of 1182 EMs. The internal stoichiometric matrix contains 30 internal metabolites and the 47 reactions.

Whole-genome Data

Saccharomyces Cerevisiae whole-genome data was provided by [9]. The network comprises 2220 metabolites (2055 internal metabolites) and 3494 reactions catalyzed by 909 genes. The transcriptomic data consist of a steady-state set that was originated in [35] and a time-series set that was originated in [34].

3.2 Principal Metabolic Flux Mode Analysis (PMFA)

Metabolic flux analysis is accomplished using different approaches. PCA is one of the most frequently used ones, due to its simplicity and low computational cost. PCA defines the relevant data from the noise and retains the data that describes the most variation in the set. However, PCA in fluxomic data has some disadvantages. For instance, it can not include the network structure and hence processes the reactions independently. In addition, PCA cannot convey outputs as pathways or EMs due to the dense nature of PCA results. In contrast, Stoichiometric flux analysis algorithms can easily identify the metabolic flux modes. Nevertheless, these methods fail to process in large datasets(i.e. whole-genome scale) and are not suitable for exploratory analysis.

In order to overcome those drawbacks, [9] introduced a novel combination of PCA and Stoichiometric Flux Analysis for the purpose of analyzing metabolic fluxes. The new algorithm is Principal Metabolic Flux Mode Analysis (PMFA). PMFA adds the stoichiometric structure as a regularization to the PCA optimization problem.

As mentioned earlier in Section 2.2.1. PCA finds the first PC by maximizing the variance as shown in equation 2.8. To introduce the stoichiometric structure, PMFA admitted additional constraint ($Sw = 0$) that represents the stoichiometry equation 2.5 in a steady-state. Since this is a hard constraint and restricted to steady-state conditions only. PMFA replaced it with a soft constraint which was relaxed by a regularization parameter λ to allow the algorithm to handle states other than steady-state (i.e. transients). Additionally, irreversible reactions weights w_{ir} were set to be non-negative and expressed by a directionality constraint ($w_{ir} \geq 0$). Considering all these constraints the PMFA optimization problem is formed as follows:

$$\begin{aligned} \max_{w} \quad & w^T \Sigma w - \lambda \|Sw\|_2^2 \\ \text{s.t.} \quad & w_{ir} \geq 0 \\ & \|w\|_2 = 1 \end{aligned} \tag{3.1}$$

Equation 3.1 is denoted as PMFA^{l_2} and PMF refers to the Principal Components produced by PMFA. Assuming that the fluxomic data matrix is $X \in \mathbb{R}^{d \times n}$ where d is the number of reactions and n is the number of samples, Σ is the data covariance matrix, $S \in \mathbb{R}^{m \times d}$ is the stoichiometric matrix where m is the number of metabolites, Sw determines the change of metabolic concentrations for all metabolites and $\|w\|_2$ is the l_2 norm of the flux vector $w \in \mathbb{R}^d$. Additionally, rev-PMFA^{l_2} denotes the reversible version of PMFA^{l_2} where it assumes that all reactions in the metabolic networks are reversible. rev-PMFA^{l_2} is formed by disregarding reaction directionality constraint from equation 3.1 as follows:

$$\begin{aligned} \max_{w} \quad & w^T \Sigma w - \lambda \|Sw\|_2^2 \\ \text{s.t.} \quad & \|w\|_2 = 1 \end{aligned} \tag{3.2}$$

The Sw regularization can be of l_2 norm or l_1 norm (as shown in equation 3.3) depending on the data. While l_2 norm favors numerous small deviations from steady-state and penalize large deviations, l_1 norm allows only a few outliers and hence used in a set with a small number of large steady-state deviations.

$$\begin{aligned} \max_{w} \quad & w^T \Sigma w - \lambda \|Sw\|_1 \\ \text{s.t.} \quad & w_{ir} \geq 0 \\ & \|w\|_2 = 1 \end{aligned} \tag{3.3}$$

Further, PMFA results are dense since the PMF components are a linear combination of all reactions activities. Therefore, [9] proposed another version of PMFA which is the *Sparse Principal Metabolic Flux Mode Analysis*. This version improves the results interpretability (i.e. favors modes with few reactions) and grants the analysis of large scales. The SPMFA optimization uses l_1 norm on w and formed as follows:

$$\begin{aligned} \max_{w} \quad & w^T \Sigma w - \lambda \|Sw\|_2^2 \\ \text{s.t.} \quad & w_{ir} \geq 0 \\ & \|w\|_1 = C \end{aligned} \tag{3.4}$$

Equation 3.4 is denoted as SPMFA^{l_2} and the form without the directionality constraint is rev-SPMFA^{l_2}, Where the degree of sparsity in PMF loadings is controlled by the hyperparameter C and SPMFA^{l_1} is formed as follows [9]:

$$\begin{aligned} \max_{w} \quad & w^T \Sigma w - \lambda \|Sw\|_1 \\ \text{s.t.} \quad & w_{ir} \geq 0 \\ & \|w\|_1 = C \end{aligned} \tag{3.5}$$

3.3 Graph-regularized Matrix Factorization (GMF)

For the analysis of gene expression data, many researchers preferred Matrix Factorization over PCA and SVD, since both PCA and SVD results are not biologically interpretable. In fact, negative values contradict physical realities [23]. However, matrix factorization was mainly used to extract distinct patterns and few studies were done on matrix factorization to capture metabolic flux modes. In this section, a modified form of MF is presented that tries to replicate the PMFA framework by involving the network stoichiometric structure in the optimization scheme.

[33] formed a graph-structured matrix factorization. Where, $X \in \mathbb{R}^{d\times n}$ is the target matrix approximated by two matrices $W \in \mathbb{R}^{d\times k}$ and $Z \in \mathbb{R}^{n\times k}$ as follows:

$$\hat{W}, \hat{Z} = \arg\min_{W,Z} \frac{1}{2}\|X - WZ^T\|_F^2 \tag{3.6}$$

The graph regularized problem assumes that the relationship between the rows of W is encoded in the adjacency matrix of a graph (V^w, E^w), where V^w and E^w represents the graph vertices and edges respectively. Additionally, two rows or columns are near to each other in the Euclidean distance, if they are connected in the graph by an edge:

$$\frac{1}{2}\sum_{i,j} E^w_{ij}(w_i - w_j)^2 = tr(W^T\mathbf{Lap}(E^w)W) \tag{3.7}$$

where the graph Laplacian $\mathbf{Lap}(E^w)$ is computed as follow:

$$\mathbf{Lap}(E^w) = D^w - E^w \tag{3.8}$$

where D^w is the diagonal matrix:

$$D^w_{ii} = \sum_{j\sim i} E^w_{ij}$$

This similarly applies for adding additional information (graph) for Z. By adding those graphs to the matrix factorization problem, the optimization variables W and Z are forced to follow the graph structure [33].

$$\begin{aligned}\min_{W,Z} \frac{1}{2}\|X - WZ^T\|_F^2 + \frac{\lambda_L}{2}\{tr(W^T\mathbf{Lap}(E^w)W) + tr(Z^T\mathbf{Lap}(E^z)Z)\}+ \\ \frac{\lambda_w}{2}\|W\|_F^2 + \frac{\lambda_z}{2}\|Z\|_F^2 \\ = \min_{W,Z} \frac{1}{2}\|X - WZ^T\|_F^2 + \frac{1}{2}\{tr(W^T L_w W) + tr(Z^T L_z Z)\} \quad (3.9)\end{aligned}$$

where:

$$\begin{aligned}L_w =& \lambda_L\mathbf{Lap}(E^w) + \lambda_w I_d \\ L_z =& \lambda_L\mathbf{Lap}(E^z) + \lambda_z I_n\end{aligned}$$

Note:

$$\begin{aligned}\|W\|_F^2 =& tr(W^T I_d W) \\ \|Z\|_F^2 =& tr(Z^T I_n Z)\end{aligned}$$

The problem was then solved in an Alternating Least Squares scheme by optimizing one side and fix the other. The following subproblem was obtained to optimize Z with W fixed:

$$\begin{aligned}\min_Z f(Z) &= \frac{1}{2}\|X - WZ^T\|_F^2 + \frac{1}{2}tr(Z^T L_z Z) \\ &= \frac{1}{2}tr((X - WZ^T)^T(X - WZ^T)) + \frac{1}{2}tr(Z^T L_z Z) \\ &= \frac{1}{2}tr(X^T X - X^T WZ^T - (WZ^T)^T Y + (WZ^T)^T WZ^T) + \frac{1}{2}tr(Z^T L_z Z) \\ &= \frac{1}{2}tr(X^T X - X^T WZ^T - ZW^T X + ZW^T WZ^T) + \frac{1}{2}tr(Z^T L_z Z) \\ &= \frac{1}{2}tr(X^T X - 2X^T WZ^T + ZW^T WZ^T) + \frac{1}{2}tr(Z^T L_z Z)\end{aligned} \quad (3.10)$$

By setting: $\nabla f(Z) = 0$, a Sylvester equation 3.11 was formed for Z:

$$ZW^TW + L_zZ = X^TW \tag{3.11}$$

Sylvester equation can be solved in a closed form using the standard Bartels-Stewart algorithm [6]. Similarly, the following subproblem was obtained to optimize W with Z fixed:

$$\min_W f(W) = \frac{1}{2}\|X - WZ^T\|_F^2 + \frac{1}{2}tr(W^TL_wW) \tag{3.12}$$

$$= \frac{1}{2}tr(X^TX - 2X^TWZ^T + ZW^TWZ^T) + \frac{1}{2}tr(W^TL_wW) \tag{3.13}$$

Setting: $\nabla f(W) = 0$

$$WZ^TZ + L_w^TW = XZ \tag{3.14}$$

In this thesis, graph-regularized matrix factorization method was used to add the side information from the stoichiometric matrix to the optimization problem resulting in an additional stoichiometric regularization.

The target matrix to be optimized is the gene-expression matrix, with reactions as rows and samples as columns. For the reaction side, a graph is constructed from the reaction adjacency matrix (presented in 3.15). The topological properties are driven from the nonzero components in the stoichiometric matrix. For this reason, a binary form of stoichiometric matrix $\hat{S}$ is adopted where the matrix values are 0 or 1 depending on the presence or absence of the compound in the reaction. Furthermore, the diagonal matrix will represent the number of elements in each reaction.

$$E^w = \hat{S}^T\hat{S} \tag{3.15}$$

Neglecting the samples side graph, the graph-regularized matrix factorization problem is formed as follow:

$$\min_{W,Z} \frac{1}{2}\|X - WZ^T\|_F^2 + \frac{\lambda_L}{2}\{tr(W^T\mathbf{Lap}(E^w)W) + tr(Z^T I_n Z)\} + \frac{\lambda_w}{2}\|W\|_F^2 + \frac{\lambda_z}{2}\|Z\|_F^2$$
$$= \min_{W,Z} \frac{1}{2}\|X - WZ^T\|_F^2 + \frac{1}{2}\{tr(W^T L_w W) + tr(Z^T L_z Z)\} \quad (3.16)$$

where:

$$L_w = \lambda_L \mathbf{Lap}(E^w) + \lambda_w I_d$$
$$L_z = \lambda_L I_n + \lambda_z I_n$$

Bartels-Stewart algorithm [6] can be used to solve the closed form Sylvester equations (3.11 for Z and 3.14 for W). FreeLYAP [31] is a freely available MATLAB implementation for the Bartels-Stewart algorithm.

Furthermore, all unnecessary parameters can be dropped to simplify the problem. Since Z does not have additional information. The objective function will be as follows:

$$\min_{W,Z} \|X - WZ^T\|_F^2 + \lambda_L tr(W^T\mathbf{Lap}(E^w)W) \quad (3.17)$$

Where the updating rules for W and Z will be as follows [12]:

$$w_{ij} \leftarrow w_{ij}\frac{(XZ + \lambda_L E^w W)}{(WZ^T Z + \lambda_L D^w W)} \quad (3.18)$$

$$z_{ij} \leftarrow z_{ij}\frac{(X^T W)}{(ZW^T W)} \quad (3.19)$$

Similar to PMFA, additional constraints were added to attain interpretable results. Firstly, the weights are constraints to limit them from scaling up. Secondly, a directionality constraint has been included, where the irreversible

reactions must be positive or zero. The following is the GMF algorithm for metabolic network analysis:

$$\begin{aligned} \min_{W,Z} \quad & \|X - WZ^T\|_F^2 + \lambda_L tr(W^T \mathbf{Lap}(E^w) W) \\ \text{s.t.} \quad & w_{ir} \geq 0 \\ & \|w\|_2 = 1 \end{aligned} \tag{3.20}$$

Furthermore, another version of GMF (reversible GMF) is formed by dropping the directionality constraint, thus allows all reactions to be reversible.

3.3.1 Model selection

In order to select the GMF model optimum regularization parameter λ_L , a Leave-one-out (LOO) cross-validation was used on Pichia Pastoris simulated data. The data was randomly partitioned into training sets and test sets. Furthermore, a predictive model was trained with all samples except one sample that was used for testing, this training was repeated with each sample excluded at a time. The optimum regularization parameter is then selected to minimize the error on test samples.

Chapter 4

Results and Discussion

This chapter evaluates the efficiency of the different methods in the analysis of metabolic networks. The first section, section 4.1 assesses the performance of PMFA, PEMA and GMF classifiers to predict active EMs. Section 4.2 measures the variance explained by the compared approaches (PCA, PMFA, PEMA and GMF). The last section 4.3 validate the PMFA and SPMFA algorithms performance on the genome-scale network.

4.1 Active elementary flux modes (EMs) prediction

In this experiment, the ability of PEMA, PMFA and GMF (with optimum regularization parameters) to correctly predict active elementary flux modes was evaluated through the use of the precision-recall and ROC (Receiver Operating Characteristic) metrics. For this purpose, the experiment was carried out on Pichia Pastoris simulated data set, where the active EMs are known. The correlation between each component of (PMF, PEM or GMF "shown in Appendix A ") and all 98 elementary flux modes were computed to identify active EMs predicted by the different methods. Then the EMs were sorted according to their maximum correlation in a descending order and the top EMs would be the predicted active elementary modes. In other words, each method will binary classify the EMs into active and inactive EMs.

The precision-recall curve is used to evaluate the classifiers outputs. The *precision* measures the positive predicted active EMs (result relevancy), mathematically defined in equation 4.1 as the number of true positive(tp) (i.e., positives correctly classified) divided by the sum of true positive counts and

false positive(fp) counts. Whereas, the *recall* (also known as the true positive rate) measures the truly retrieved active EMs by the classifier and mathematically defined in equation 4.2 as the number of true positives(tp) divided by the total number of positives (i.e., the sum of true positive counts and false negative(fn) counts). The precision and recall values for different thresholds are outlined in the precision-recall curve [13, 15].

$$\text{precision} = \frac{\text{tp}}{\text{tp} + \text{fp}} \tag{4.1}$$

$$\text{recall} = \text{tpr} = \frac{\text{tp}}{\text{tp} + \text{fn}} \tag{4.2}$$

where, tp is the number of active EMs classified as active, fn is the number of active EMs classified as inactive, tn is the number of inactive EMs classified as inactive and fp is the number of inactive EMs classified as active.

Further, Area Under Precision-Recall curve (AUPR) is a measure used to summarize the precision-recall curve, where high AUPR represents high precision and high recall. Another tool to summarize the precision-recall curve is the Average precision (AP). AP is a weighted mean of precisions, for n thresholds the weights are defined as the increase in recall from the previous threshold.

$$\text{AP} = \sum_{\text{n}} (\text{recall}_{\text{n}} - \text{recall}_{\text{n}-1}) \text{precision}_{\text{n}} \tag{4.3}$$

Additionally, the ROC graph was also used to visualize the performance of the different classifiers. The ROC curve plots the true positive rate (tpr) versus the false positive rate (fpr). True positive rate measures the correctly classified active EMs and is defined in equation 4.2 and the false positive rate is the false alarm rate that measures the misclassified inactive EMs, mathematically defined as the false positive counts divided by the total negatives (i.e., the sum of false positive counts and true negative counts) [15].

$$\text{fpr} = \frac{\text{fp}}{\text{fp} + \text{tn}} \tag{4.4}$$

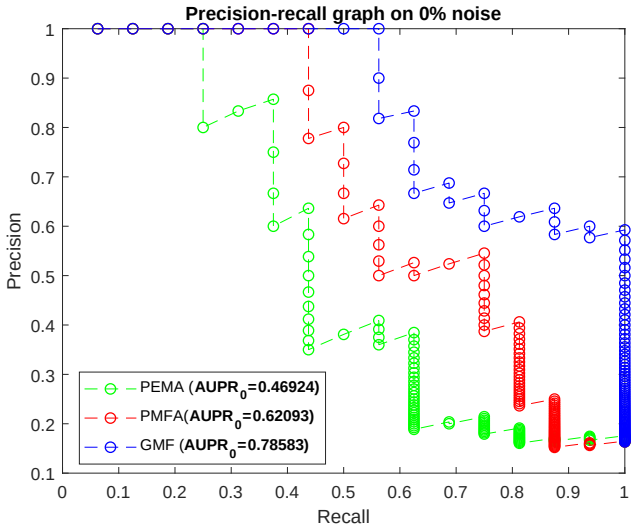
Figure 4.1 shows the precision-recall curves for PEMA, PMFA and GMF on Pichia Pastoris simulated model with 0% noise in (a) and 20% added noise in (b). As observed in Figure 4.1(a) PEMA with 3-factors has very low AUPR (**0.47**) for 0% noise and has dramatically declined to **0.099** with the increase of added noise to 20% in figure 4.1(b). Additionally, it can be seen in figure 4.1(a), 3-components PMFA performance decreased with the increase of added noise in figure 4.1(b), that the area under precision-recall (AUPR) curve was **0.62** for 0% noise and decreased to **0.19** with the increase of added noise to 20%. Figure 4.1 presents the precision-recall curves of the 3-rank GMF performance, where the GMF performance slightly decreased with the addition of noise from **0.79** AUPR to **0.65**.

Figure 4.2 illustrates the ROC curves for PEMA, PMFA and GMF on Pichia Pastoris simulated model with 0% noise in (a) and 20% added noise in (b). The ROC curves shown in Figure 4.2 confirm that the addition of noise drastically degrades the PEMA performance, that the AUC for 0% noise data was **0.69**, it has dropped to **0.22** for 20% noise data in figure 4.2(b). Furthermore, figure 4.2 shows that the ROC curves for PMFA performed better than a random guessing classifier (i.e., above 0.5), with a reported **0.805** AUC for 0% noise and **0.515** for 20% noise.

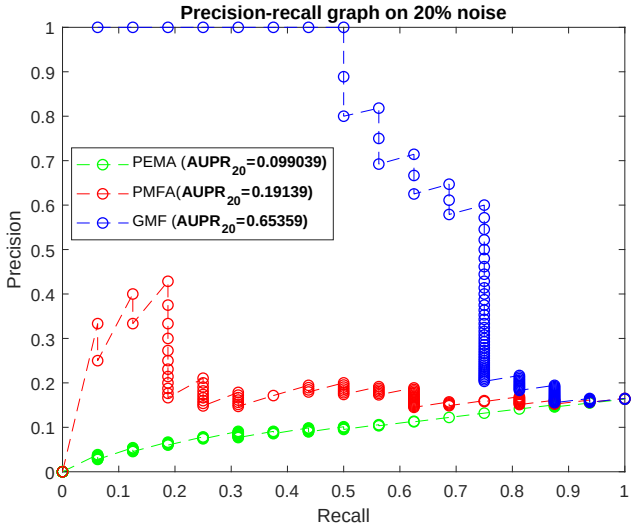
Similarly, Figure 4.2 presents the ROC curves evaluation of the GMF performance. It can be observed that the GMF is robust to noise. Additionally, GMF reported a very high AUC (**0.96** for 0% noise and **0.79** for 20% noise) in comparison with the PMFA and PEMA classifiers. Table 4.1 summarizes the evaluations results of the different methods.

Table 4.1: Classification performance of PMFA, PEMA and GMF.

	AP		**AUPR**		**AUC**	
	0% noise	20% noise	0% noise	20% noise	0% noise	20% noise
PMFA	0.689	0.217	0.621	0.191	0.805	0.515
PEMA	0.540	0.106	0.469	0.099	0.692	0.216
GMF	0.852	0.720	0.786	0.654	0.962	0.785

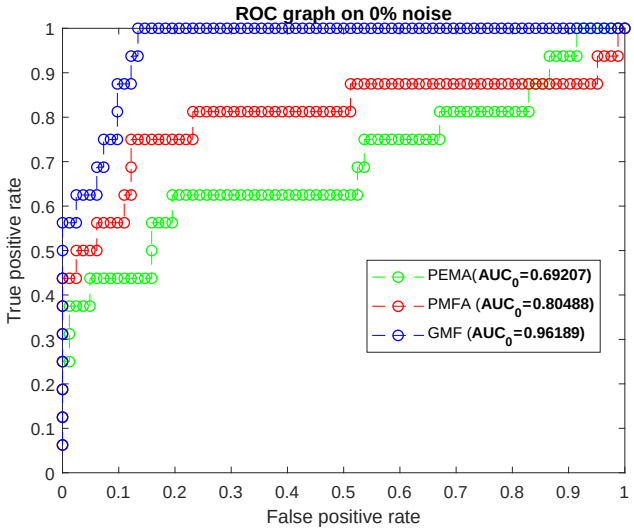


(a) Precision-Recall curves on 0% noise

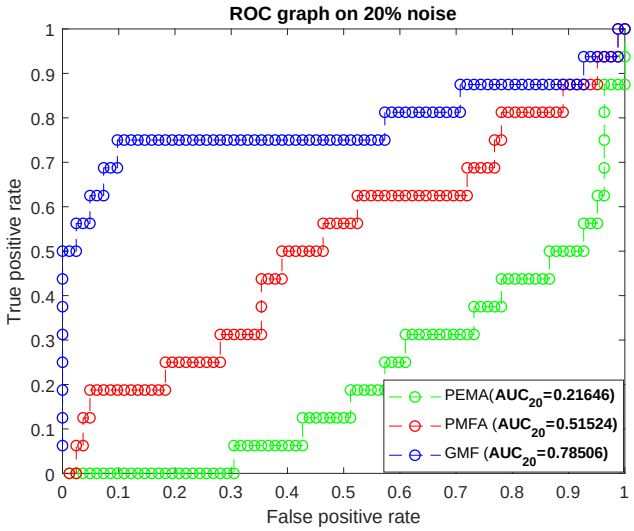


(b) Precision-Recall curves on 20% noise

Figure 4.1: The Precision-Recall curves on 0% noise(a) Pichia Pastoris data set and Precision-Recall curves on 20% noise(b).



(a) ROC curves on 0% noise



(b) ROC curves on 20% noise

Figure 4.2: ROC curves of PEMA, PMFA and GMF on Pichia Pastoris data set with 0% noise(a) and ROC curves on Pichia Pastoris data set with 20% noise(b).

4.2 Explained variance

The main objective of the PMFA algorithm is to extract pathways with the maximum sample variance while penalizing the projections that highly deviate from the steady state. This experiment measures the explained variance captured by the different approaches (PCA, PEMA and GMF) and compares the results. In addition, the penalization of the deviation from steady-state was evaluated. For this purpose, a fraction of variance (FoV) measure was used to explain the captured variance of a sample.

$$FoV = \frac{w^T \Sigma w}{Trace(\Sigma)} \tag{4.5}$$

Figure 4.3 illustrates the relationship between the change in internal metabolites and the fraction of variance (FoV) captured by the different methods as well as the first PMF of the PMFA approach. It can be seen that the FoV captured by PMFA and rev-PMFA are slightly less than the PCA and PCA_{rev}, respectively. Nevertheless, both PMFA and rev-PMFA FoVs are converging towards PCA and PCA_{rev} FoVs values when the regularization parameter value is 0. This observation matches the theoretical PMFA algorithm, where the PMFA turns to PCA when the regularization is disregarded. Additionally, It can be seen from Figure 4.3 that the deviation from the steady state decreased with the increase in the stoichiometric regularization parameter λ. Furthermore, Figure 4.3 shows that rev-PMFA has higher FoV than PMFA. This can be explained that the directionality constraint in PMFA is reducing the captured variance. Additionally, PMFA and rev-PMFA captured higher sample variance than PEMA approach. Similarly, GMF (with 3 components and $\lambda = 5$) captured higher sample variance than PEMA with (1,2 and 10 factors).

To evaluate the performance of PMFA on noisy data, the FoV captured by PMFA was measured on Pichia Pastrios simulated data with 20% added noise shown in Figure 4.4. The noise has affected the amount of the captured sample variance, the FoVs captured on noisy data by all methods are less than the FoVs captured on the dataset with 0% noise. The PMFA and rev-PMFA results are similar to the previous results. However, the PEMA results have decreased dramatically in comparison to the results on the noiseless data set. In addition, Figure 4.4 shows that GMF (with 3 components and $\lambda = 5$) captured higher sample variance than the PEMA approach and slightly less than the PMFA.

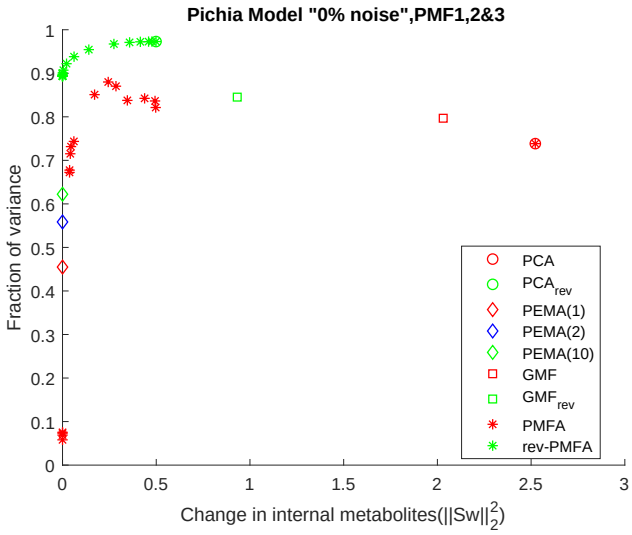


Figure 4.3: Fraction of variance captured by reversible PCA, PCA, PEMA (1,2 and 10 factors), GMF, reversible GMF, PMFA and reversible PMFA with different l_2 regularizer values.

Furthermore, both reversible methods rev-PMFA and GMF$_{rev}$ captures higher FoVs than their corresponding alternative approach PMFA and GMF, respectively. This indicates that the additional directionality constraint reduces the ability of the approaches to capture higher FoVs.

GMF explained variance

The objective of the GMF is to minimize the approximation error. Figure 4.5 illustrate the connection between maximizing the explained variance (PMFA approach) and minimizing the mean-square error (GMF approach), while the GMF Mean-square error (MSE) decreases the Fraction of Variance (FoV) increases and vice versa.

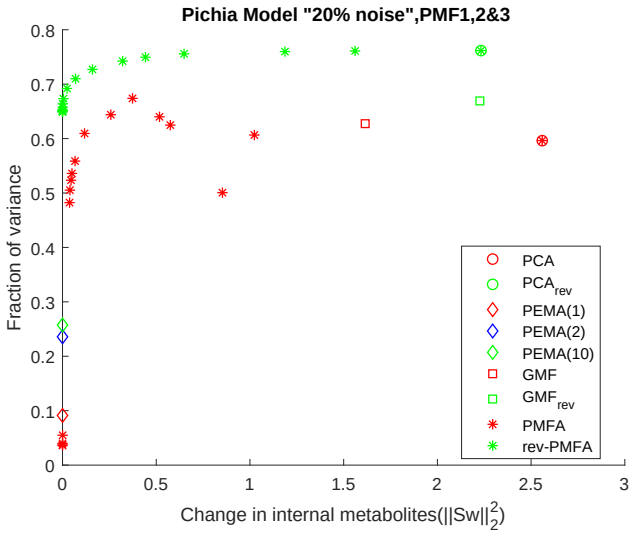
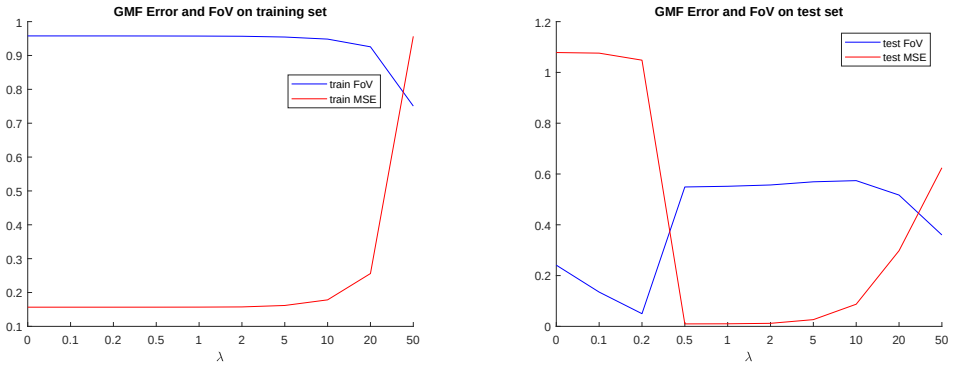


Figure 4.4: Fraction of variance captured by reversible PCA, PCA, PEMA (1,2 and 10 factors), GMF, GMF$_{rev}$, PMFA and reversible PMFA with different l_2 regularization values.

4.3 Sparse flux modes recovery

This experiment was carried out to validate the performance of the different approaches on the full genome metabolic network, where the data is highly sparse. Unfortunately, GMF and PEMA failed to handle the full-genome sparse data. Conversely, PMFA was able to efficiently captures the variance of the samples due to its scalability up to whole-genome sets, on account of the *Concave Convex Procedure* (CCP) optimization.

Figure 4.6 presents the Fraction of Variance (FoV) of the PCA, PMFA (in red) and SPCA, SPMFA(in yellow). both SPMFA and SPCA captured higher FoV values than PMFA and PCA, indicating that sparse versions perform better than the regular approaches on sparse datasets. However, if the directionality constraint is discarded rev-PMFA and PCA$_{rev}$ (in green) captures higher FoV than rev-SPMFA and SPCA$_{rev}$.



(a) MSE on training data (b) MSE on test data

Figure 4.5: Mean-square errors and Fraction of Variances captured by GMF on S. cerevisiea (a) train set, (b) test set.

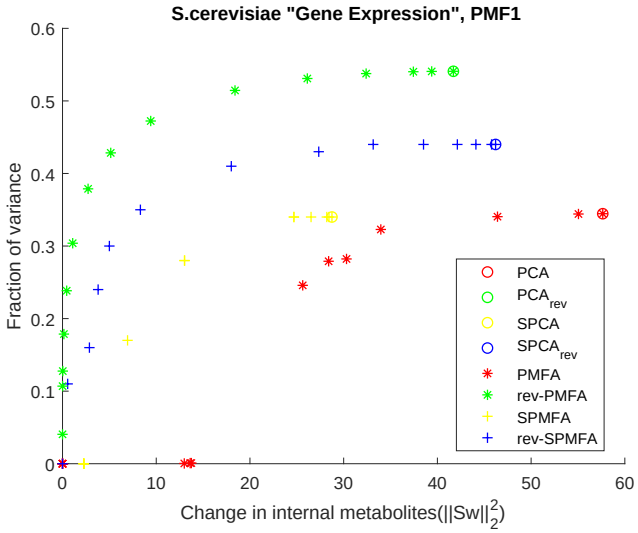


Figure 4.6: PMFA and SPMFA on the whole-genome data set with 1 component.

Chapter 5

Conclusion

This thesis assesses the Principal Metabolic Flux Mode Analysis (PMFA) approach for metabolic network analysis on various datasets types(e.g. Simulated data, Whole-genome data). The algorithm merges the well-known PCA with the stoichiometric network analysis, resulting in a framework that not only can easily determine the network fluxes with the highest variation but also can acclimate to deviations from steady-state due to the regularization parameter. As shown in Chapter 4 PMFA competed with the literature methods(e.g. PEMA) by capturing the highest explained variance with a minimum number of factors. Additionally, PMFA was the only approach capable of analyzing whole-genome scale owing to the Concave-Convex Procedure(CCP).

Further, this thesis develops a Graph regularized Matrix Factorization for the metabolic network analysis, that resembles the PMFA framework through including the stoichiometric network graph as a side information. The GMF achieved good results on experimental data and was able to perfectly predict the active EMs on simulated data. However, it fails to work on the whole-genome dataset. Moreover, since the GMF regularized by the Laplacian of the stoichiometric network graph($tr(W^T LW)$). In other words, it does not directly include the stoichiometric steady-state($Sw = 0$) constraint. The results could not present the fluxes at the steady-state. Hence, gives a reason for further research on including the steady-state for better results. Future research could also develop a sparse version of GMF, as it will make it possible for the GMF to scale-up and be able to work on genome-scale networks.

Bibliography

[1] A Oberhardt, M., Ø Palsson, B., and Papin, J. *Applications of Genome-Scale Metabolic Reconstructions.*, vol. 5. Cambridge University Press, 11 2009.

[2] Allen, G. I., and Maletic-Savatic, M. Sparse non-negative generalized pca with applications to metabolomics. *Bioinformatics 27*, 21 (2011), 3029–3035.

[3] Alpaydin, E. *Introduction to Machine Learning*, 2nd ed. The MIT Press, 2010.

[4] Alter, O., Brown, P. O., and Botstein, D. Singular value decomposition for genome-wide expression data processing and modeling. *Proceedings of the National Academy of Sciences 97*, 18 (2000), 10101–10106.

[5] Azzouz, T., and Tauler, R. Application of multivariate curve resolution alternating least squares (mcr-als) to the quantitative analysis of pharmaceutical and agricultural samples. *Talanta 74*, 5 (2008), 1201 – 1210.

[6] Bartels, R. H., and Stewart, G. W. Solution of the matrix equation ax + xb = c [f4]. *Commun. ACM 15*, 9 (Sept. 1972), 820–826.

[7] Becker, S. A., and Palsson, B. Ø. Genome-scale reconstruction of the metabolic network in staphylococcus aureus n315: an initial draft to the two-dimensional annotation. *BMC Microbiology 5*, 1 (Mar 2005), 8.

[8] Bello, MD, E. A., and Schwinn, MD, D. A. Molecular biology and medicine a primer for the clinician. *Anesthesiology 85*, 6 (1996), 1462–1478.

[9] Bhadra, S., Blomberg, P., Castillo, S., and Rousu, J. Principal metabolic flux mode analysis. *Bioinformatics* (2018), bty049.

[10] Bishop, C. M. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[11] Brunet, J.-P., Tamayo, P., Golub, T. R., and Mesirov, J. P. Metagenes and molecular pattern discovery using matrix factorization. *Proceedings of the National Academy of Sciences 101*, 12 (2004), 4164–4169.

[12] Cai, D., He, X., Han, J., and Huang, T. S. Graph regularized nonnegative matrix factorization for data representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence 33*, 8 (Aug 2011), 1548–1560.

[13] Davis, J., and Goadrich, M. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd International Conference on Machine Learning* (New York, NY, USA, 2006), ICML '06, ACM, pp. 233–240.

[14] de Juan, A., and Tauler, R. Chemometrics applied to unravel multicomponent processes and mixtures: Revisiting latest trends in multivariate resolution. *Analytica Chimica Acta 500*, 1 (2003), 195 – 210. ANALYTICAL HORIZONS - An International Symposium celebrating the publication of Volume 500 of Analytica Chimica Acta.

[15] Fawcett, T. An introduction to roc analysis. *Pattern Recognition Letters 27*, 8 (2006), 861 – 874. ROC Analysis in Pattern Recognition.

[16] Folch-Fortuny, A., Marques, R., Isidro, I. A., Oliveira, R., and Ferrer, A. Principal elementary mode analysis (pema). *Mol. BioSyst. 12* (2016), 737–746.

[17] Folch-Fortuny, A., Tortajada, M., Prats-Montalbán, J., Llaneras, F., Picó, J., and Ferrer, A. Mcr-als on metabolic networks: Obtaining more meaningful pathways. *Chemometrics and Intelligent Laboratory Systems 142* (2015), 293 – 303.

[18] Frigyesi, A., and Höglund, M. Non-negative matrix factorization for the analysis of complex gene expression data: Identification of clinically relevant tumor subtypes. *Cancer Informatics 6* (2008), CIN.S606.

[19] Gaujoux, R., and Seoighe, C. Semi-supervised nonnegative matrix factorization for gene expression deconvolution: A case study. *Infection, Genetics and Evolution 12*, 5 (2012), 913 – 921.

[20] Hayakawa, K., Kajihata, S., Matsuda, F., and Shimizu, H. 13c-metabolic flux analysis in s-adenosyl-l-methionine production by saccharomyces cerevisiae. *Journal of Bioscience and Bioengineering 120*, 5 (2015), 532 – 538.

[21] Jaumot, J., Tauler, R., and Gargallo, R. Exploratory data analysis of dna microarrays by multivariate curve resolution. *Analytical Biochemistry 358*, 1 (2006), 76 – 89.

[22] Junker, B. H., and Schreiber, F. *Analysis of Biological Networks (Wiley Series in Bioinformatics)*. Wiley-Interscience, 2008.

[23] Kim, M. H., Seo, H. J., Joung, J.-G., and Kim, J. H. Comprehensive evaluation of matrix factorization methods for the analysis of dna microarray gene expression data. *BMC Bioinformatics 12*, 13 (Nov 2011), S8.

[24] Kong, W., Mou, X., and Hu, X. Exploring matrix factorization techniques for significant genes identification of alzheimer's disease microarray gene expression data. *BMC Bioinformatics 12*, 5 (Jul 2011), S7.

[25] Koren, Y., Bell, R., and Volinsky, C. Matrix factorization techniques for recommender systems. *Computer 42*, 8 (Aug 2009), 30–37.

[26] Kuepfer, L. *Stoichiometric Modelling of Microbial Metabolism*. Springer New York, New York, NY, 2014, pp. 3–18.

[27] Llaneras, F., and Picó, J. Stoichiometric modelling of cell metabolism. *Journal of Bioscience and Bioengineering 105*, 1 (2008), 1 – 11.

[28] Ma, S., and Dai, Y. Principal component analysis based methods in bioinformatics studies. *Briefings in Bioinformatics 12*, 6 (2011), 714–722.

[29] Ma, S., and Kosorok, M. R. Identification of differential gene pathways with principal component analysis. *Bioinformatics 25*, 7 (2009), 882–889.

[30] Nelson, D., Lehninger, A., and Cox, M. *Lehninger Principles of Biochemistry*. Lehninger Principles of Biochemistry. W. H. Freeman, 2008.

[31] Nick Hale, A. T., and Wilber, H. freelyap. MIT, 2014.

[32] Palsson, B. O. *Systems Biology: Properties of Reconstructed Networks*. Cambridge University Press, New York, NY, USA, 2006.

[33] Rao, N., Yu, H.-F., Ravikumar, P. K., and Dhillon, I. S. Collaborative filtering with graph information: Consistency and scalable methods. In *Advances in Neural Information Processing Systems 28*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds. Curran Associates, Inc., 2015, pp. 2107–2115.

[34] Rintala, E., Jouhten, P., Toivari, M., Wiebe, M. G., Maaheimo, H., Penttilä, M., and Ruohonen, L. Transcriptional responses of saccharomyces cerevisiae to shift from respiratory and respirofermentative to fully fermentative metabolism. *OMICS: A Journal of Integrative Biology 15*, 7-8 (2011), 461–476. PMID: 21348598.

[35] Rintala, E., Toivari, M., Pitkänen, J.-P., Wiebe, M. G., Ruohonen, L., and Penttilä, M. Low oxygen levels as a trigger for enhancement of respiratory metabolism in saccharomyces cerevisiae. *BMC Genomics 10*, 1 (Oct 2009), 461.

[36] Sar?yar, B., Perk, S., Akman, U., and Hortaçsu, A. Monte carlo sampling and principal component analysis of flux distributions yield topological and modular information on metabolic networks. *Journal of Theoretical Biology 242*, 2 (2006), 389 – 400.

[37] Schilling, C. H., Schuster, S., Palsson, B. O., and Heinrich, R. Metabolic pathway analysis: Basic concepts and scientific applications in the post-genomic era. *Biotechnology Progress 15*, 3 (1 1999), 296–303.

[38] SCHUSTER, S., and HILGETAG, C. On elementary flux modes in biochemical reaction systems at steady state. *Journal of Biological Systems 02*, 02 (1994), 165–182.

[39] Shi, J., Zheng, X., and Yang, W. Survey on probabilistic models of low-rank matrix factorizations. *Entropy 19*, 8 (2017), 424.

[40] Shlens, J. A tutorial on principal component analysis. In *Systems Neurobiology Laboratory, Salk Institute for Biological Studies* (2005).

[41] Tauler, R. Multivariate curve resolution applied to second order data. *Chemometrics and Intelligent Laboratory Systems 30*, 1 (1995), 133 – 146. InCINC '94 Selected papers from the First International Chemometrics Internet Conference.

[42] Tortajada, M., Llaneras, F., and Picó, J. Validation of a constraint-based model of pichia pastoris metabolism under data scarcity. *BMC Systems Biology 4*, 1 (Aug 2010), 115.

[43] Trinh, C. T., Wlaschin, A., and Srienc, F. Elementary mode analysis: a useful metabolic pathway analysis tool for characterizing cellular metabolism. *Applied Microbiology and Biotechnology 81*, 5 (Nov 2008), 813.

[44] Van Dien, S. *Metabolic Engineering for Bioprocess Commercialization.* Springer International Publishing, 2016.

[45] von Stosch, M., Rodrigues de Azevedo, C., Luis, M., Feyo de Azevedo, S., and Oliveira, R. A principal components method constrained by elementary flux modes: analysis of flux data sets. *BMC Bioinformatics 17*, 1 (May 2016), 200.

Appendix A

Appendix

```
function [tpr,fpr,prec,AP] = tfpr(I,Pichia)
```

This function measures the tpr, fpr, precision and Average Precision.

```
%Inputs:
% I: Predicted EMs indexes.
% Pichia: Pichia data, Pichia.ActiveEMs: ground-truth active EMs.

%Outputs:
% tpr : true positive rate, recall.
% fpr: false positive rate.
% prec: precision.
% AP: Average Precision.
```

Initialization:

```
tpr=zeros(1,98);
fpr=zeros(1,98);
prec=zeros(1,98);
tp = 0;
fp = 0;
tprl=0;
AP=0;

for l = 1:1:98
    % increment tp count if the ground-truth active EMs is in the top
 i's
    if (ismember(I(l),Pichia.ActiveEMs'))
        tp = tp+1;
    % increment fp count if the ground-truth active EMs isnt in the
 top i's
    else
        fp = fp+1;
    end
    tpr(l) = tp/length(Pichia.ActiveEMs);
    fpr(l) = fp/(length(I)-length(Pichia.ActiveEMs));
    prec(l) =tp/(tp+fp);
    AP = AP+((tpr(l)-tprl)*prec(l));
    % Save the previous recall value
    tprl=tpr(1,l);
end
```

Published with MATLAB® R2017b

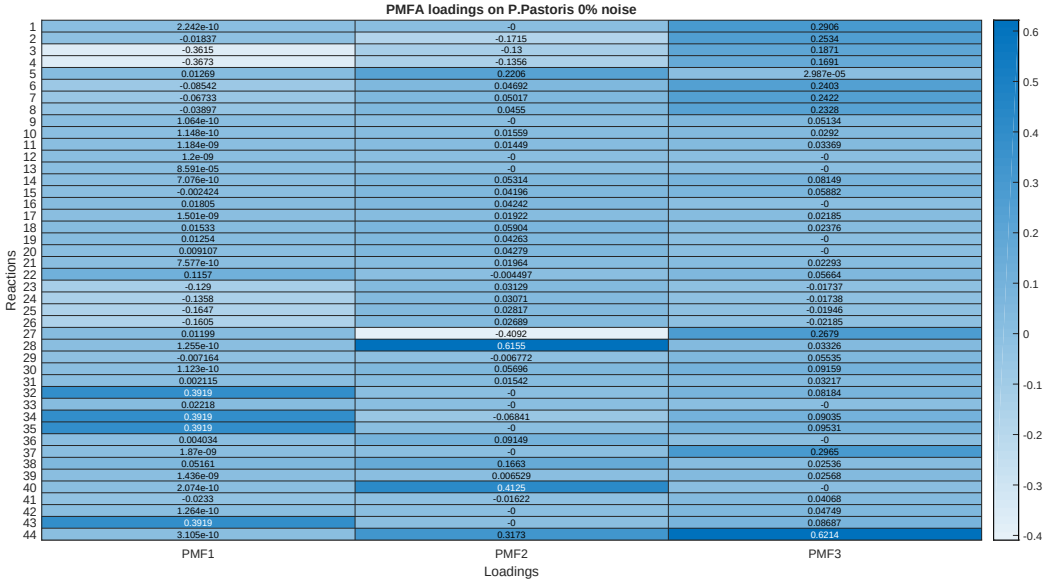


Figure A.1: PMFA$_{l2}$ loadings with $\lambda = 5$ on P.Pastoris with 0% noise

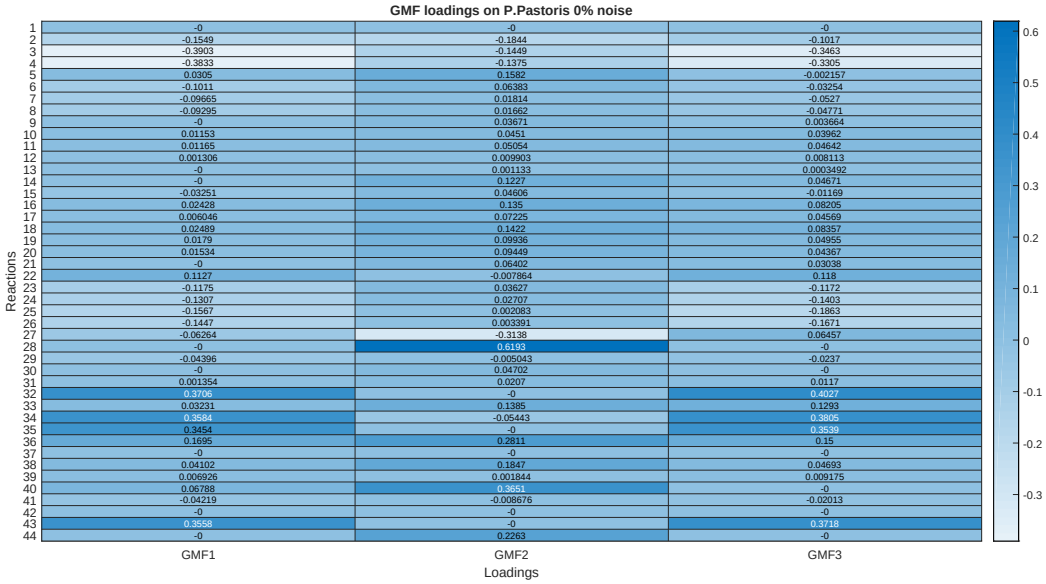


Figure A.2: GMF loadings with $\lambda = 5$ on P.Pastoris with 0% noise

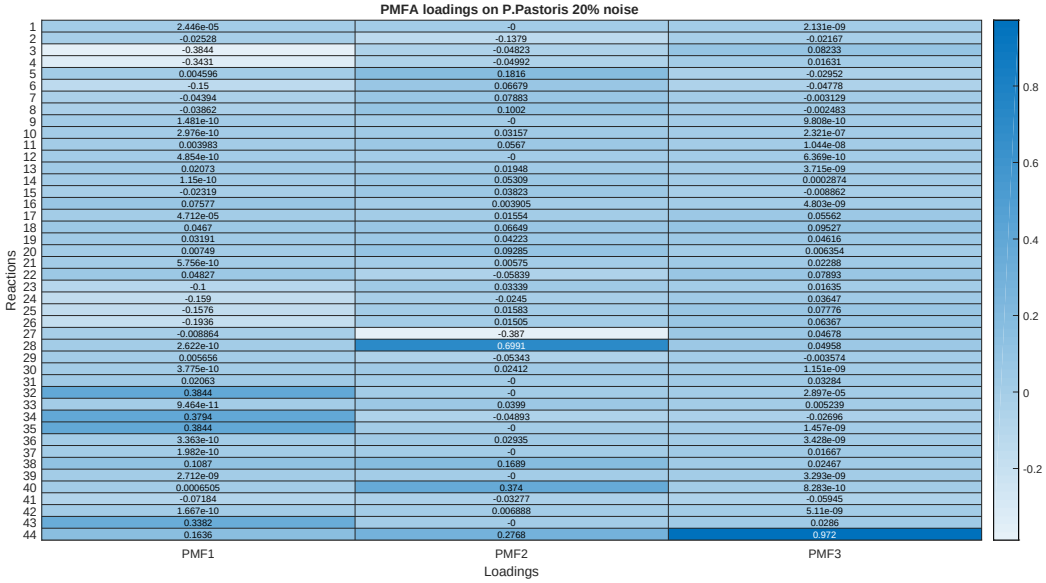


Figure A.3: PMFA$_{l2}$ loadings with $\lambda = 5$ on P.Pastoris with 20% noise

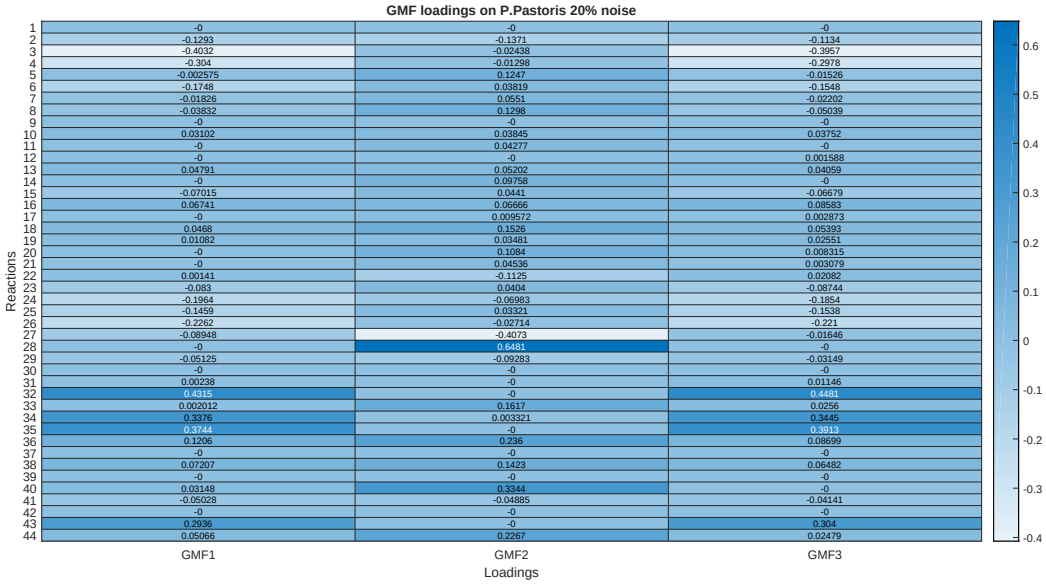


Figure A.4: GMF loadings with $\lambda = 5$ on P.Pastoris with 20% noise

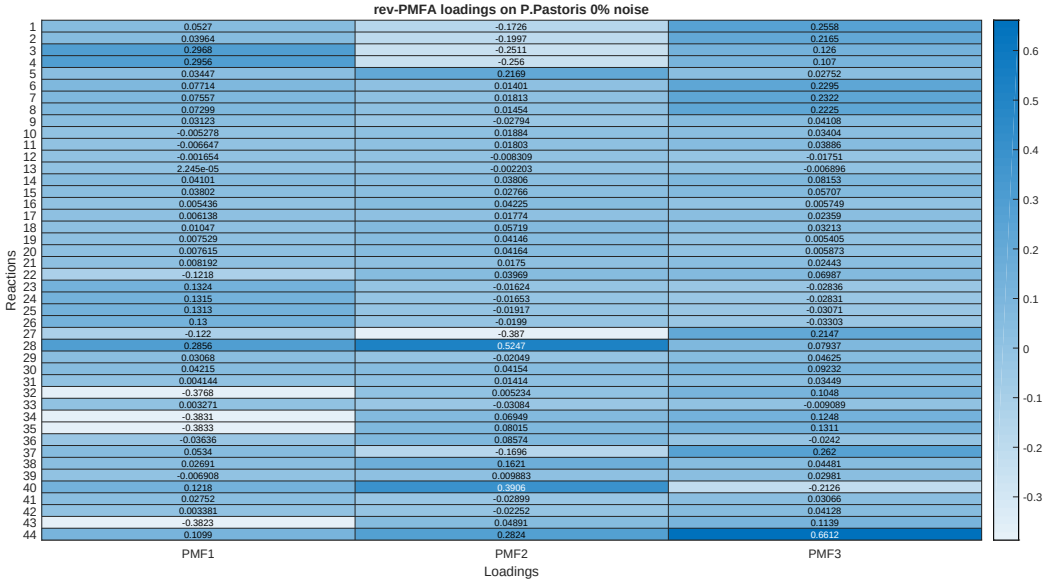


Figure A.5: rev-PMFA$_{l2}$ loadings with $\lambda = 5$ on P.Pastoris with 0% noise

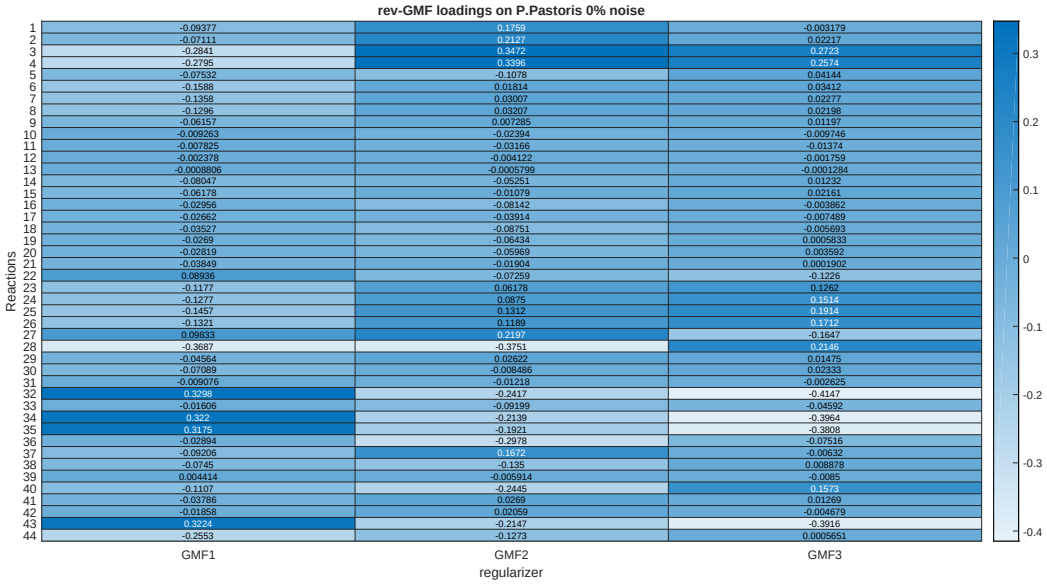


Figure A.6: GMF$_{rev}$ loadings with $\lambda = 5$ on P.Pastoris with 0% noise

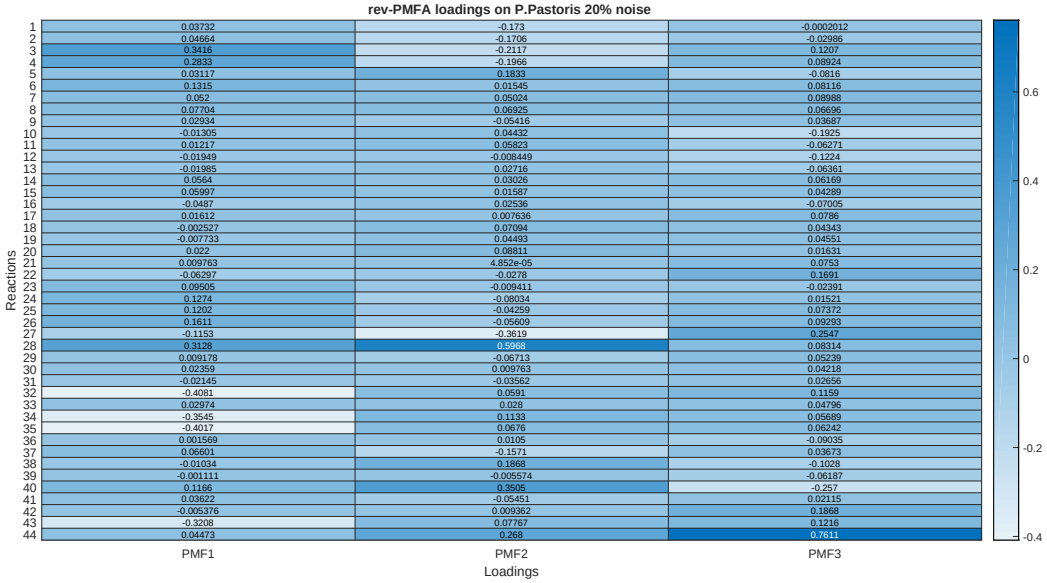


Figure A.7: rev-PMFA$_{l2}$ loadings with $\lambda = 5$ on P.Pastoris with 20% noise

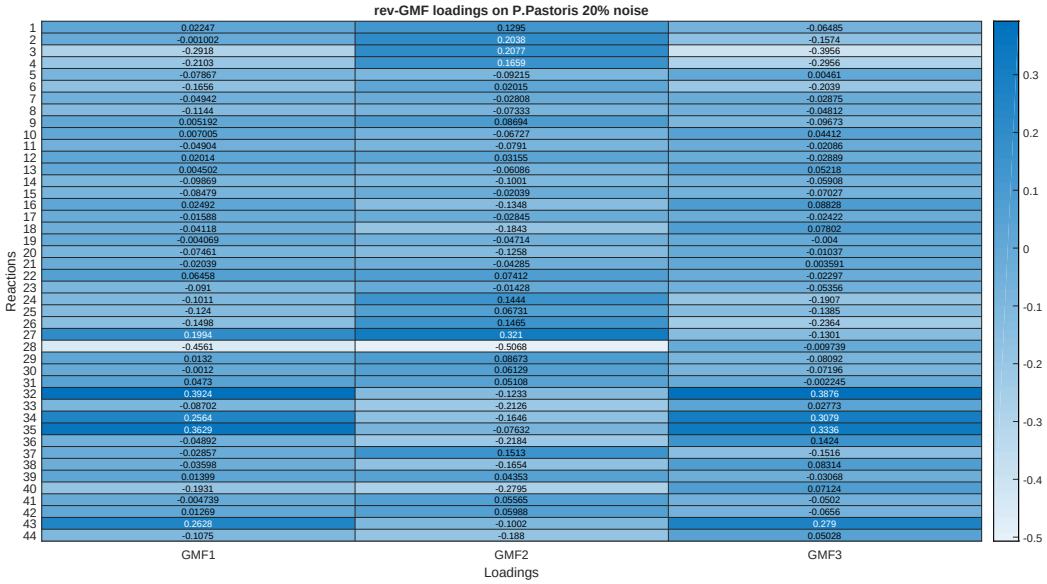


Figure A.8: GMF$_{rev}$ loadings with $\lambda = 5$ on P.Pastoris with 20% noise


Aalto University
School of Science