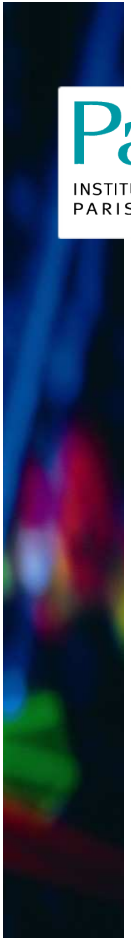
# Applications of machine learning in computational biology

Edouard Pauwels



ParisTech
INSTITUT DES SCIENCES ET TECHNOLOGIES
PARIS INSTITUTE OF TECHNOLOGY

MINES ParisTech

École doctorale n° 421 :

Sciences des métiers de l'ingénieur

**Doctorat ParisTech**

# T H È S E

**pour obtenir le grade de docteur délivré par**

## l'École nationale supérieure des mines de Paris

**Spécialité " Bio-informatique "**

*présentée et soutenue publiquement par*

**Edouard PAUWELS**

le 14 novembre 2013

# Applications de l'apprentissage statistique à la biologie computationnelle

~~~

# Applications of machine learning in computational biology

Directeur de thèse : **Véronique Stoven**

**Jury**

| | | |
|---|---|---|
| **Christophe Ambroise**, | Professeur, Laboratoire Statistique et Génome Université d'Evry | Rapporteur |
| **Didier Rognan**, | Directeur de recherche, Laboratoire D'Innovation Therapeutique, Univeristé de Strasbourg | Rapporteur |
| **Sandrine Dudoit**, | Professeur, Division of Biostatistics, University of California Berkeley | Examinateur |
| **Stéphane Robin**, | Directeur de recherche, Applied Mathematics and Computer Science Unig, AgroParisTech | Examinateur |
| **Yoshihiro Yamanishi**, | Professeur, institute for advanced study, Kyushu University | Examinateur |
| **Véronique Stoven**, | Professeur, Center for Computational Biology, Mines ParisTech | Directeur de thèse |

**Centre de Bio-Informatique**
35 rue Saint-Honoré, 77300 Fontainebleau, France

T H È S E
~~~

# Remerciements

Je remercie Véronique Stoven qui a dirigé cette thèse. Son soutien, ses conseils et l'écoute dont elle a fait preuve à mon égard durant ces trois années ont été déterminants pour mon positionnement scientifique, professionnel et personnel.

Ma reconnaissance va à Christophe Ambroise et Didier Rognan qui ont pris le temps d'être rapporteurs pour cette thèse. Je tiens également à remercier chaleureusement les éxaminateurs, Stéphane Robin, Sandrine Dudoit qui m'a accueilli, conseillé et encadré à l'université de Berkeley, et Yoshihiro Yamanishi qui a su me guider à mes débuts et avec qui j'ai eu le plaisir de collaborer par la suite.

Les collaborations avec des chercheurs et des étudiants bienveillants et attentifs ont grandement contribué à orienter mon parcours et à enrichir mon travail et ma culture. Je tiens donc adresser de sincères remerciements à Jean-Philippe Vert qui dirige l'équipe au sein de laquelle s'est déroulé cette thèse; Christian Lajaunie sans qui je ne l'aurais probablement pas rejoint; Gautier Stoll, Didier Surdez et Anne-Claire Haury avec qui j'ai eu plaisir à collaborer à l'institut Curie; Laurent Jacob, Davide Risso et Miles Lopes que j'ai eu la chance de rencontrer à Berkeley; Emile Richard et Jérôme Bôlte qui m'a régulièrement accueilli à l'université Toulouse Capitole.

Mes amitiés et ma considération vont à mes collègues du CBIO et de l'U900 pour leur présence et leur sympathie. J'espère avoir l'occasion de vous croiser à nouveau sur mon chemin. Je pense particulièrement à Pierre Chiche avec qui j'ai partagé, entre autres choses, le calendrier et les échéances du travail de thèse.

Merci à mes proches: ma famille et mes amis, auxquels je pense ici sans les nommer, pour leur chaleur et leur soutien. Merci enfin à Hélène.

# Contents


**1 Introduction** **1**

1.1 System theoretic approaches in biology . . . . . 1

1.2 Machine learning . . . . . 4

1.3 Contribution and organization of the thesis . . . . . 6

**2 Methods** **9**

2.1 Sparse matrix factorization . . . . . 10

2.1.1 Singular value decomposition . . . . . 10

2.1.2 Sparse matrix decomposition . . . . . 12

2.1.3 Application in data analysis . . . . . 13

2.1.4 Validation in the context of supervised prediction context . . . 15

2.2 Supervised learning . . . . . 16

2.2.1 Empirical risk minimization . . . . . 16

2.2.2 Choice of the loss function . . . . . 17

2.2.3 Kernels and kernel methods . . . . . 17

2.2.4 Nearest neighbours . . . . . 18

2.3 Inference and learning in probabilistic models . . . . . 18

2.3.1 EM algorithm . . . . . 19

2.3.2 Markov Chain Monte Carlo and importance sampling . . . . . 20

2.4 Empirical model evaluation and cross validation . . . . . 21

**3 Protein-ligand interactions** **22**

3.1 Introduction . . . . . 24

3.1.1 Protein-ligand interactions . . . 24
3.1.2 Existing approaches and motivations . . . 24
3.1.3 Content of the chapter . . . 26
3.2 Materials . . . 27
3.3 Methods . . . 29
3.3.1 Sparse canonical correspondence analysis (SCCA) . . . 29
3.3.2 Evaluation of extracted components by reconstruction of drug-target interactions . . . 30
3.3.3 Supervised methods to reconstruct drug-target interactions . . . 31
3.4 Results . . . 33
3.4.1 Performance evaluation for the SCCA method . . . 33
3.4.2 Comparison with other supervised methods . . . 36
3.4.3 Biological interpretation of extracted drug substructures and protein domains using SCCA . . . 38
3.4.4 Comments on $L1$-PLSVM method . . . 44
3.5 Discussion and Conclusion . . . 46
3.5.1 Applications of SCCA in drug development . . . 47

**4 Phenotypic response to molecular perturbations 50**
4.1 Introduction . . . 52
4.1.1 Content of the chapter . . . 53
4.2 Sparse canonical correlation based drug side-effect analysis . . . 55
4.2.1 Background . . . 55
4.2.2 Materials . . . 58
4.2.3 Methods . . . 58
4.2.4 Results . . . 61
4.2.5 Discussion and conclusion . . . 70
4.2.6 Integrating different sources of information . . . 72
4.3 Cell population phenotyping . . . 74
4.3.1 Introduction . . . 74
4.3.2 Materials and Methods . . . 79

4.3.3 Results . . . . . 86
4.3.4 Conclusion and discussion . . . . . 95

**5 Dynamical system parameter identification under budget constraints 99**
5.1 Introduction . . . . . 100
5.1.1 Evaluation of experimental design strategies . . . . . 101
5.1.2 Proposed strategy . . . . . 101
5.1.3 Content of the chapter . . . . . 102
5.2 Methods . . . . . 104
5.2.1 In silico network . . . . . 104
5.2.2 Implementation . . . . . 107
5.3 Results and discussion . . . . . 112
5.3.1 Experimental results . . . . . 112
5.3.2 Discussion . . . . . 116
5.4 Conclusion . . . . . 120

**6 Conclusion 122**

**Bibliography 125**

**A Copulas 141**
A.1 Definition and example . . . . . 141
A.2 Copula models parameters estimation . . . . . 143
A.2.1 Maximum likelihood . . . . . 144
A.2.2 Inference functions for margin . . . . . 144
A.2.3 Semi-parametric estimation . . . . . 145

**B DREAM7 parameter estimation challenge dynamic equations 146**

# List of Tables


3.1 Cross validation results for drug-target interactions . . . . . 34
3.2 Examples of canonical component content . . . . . 39
3.3 Pairwise cross validation results . . . . . 45

4.1 Nitrogen-containing rings of size 5 substructure . . . . . 66
4.2 Association between population descriptors . . . . . 81

5.1 Results of DREAM7 challenge . . . . . 115
5.2 Estimation of the expected risk . . . . . 117

# List of Figures


3.1 An illustration of the proposed method. . . . . 28
3.2 OCCA index plot . . . . . 35
3.3 SCCA index plot . . . . . 36
3.4 Computational cost . . . . . 37
3.5 Drug specificity based on canonical components . . . . . 40
3.6 Chemical structure of Tamoxifen . . . . . 42
3.7 Reconstruction of thiazide based on canonical components . . . . . 43
3.8 Pair-wise and block-wise cross validation . . . . . 45

4.1 ROC curve for side-effect prediction . . . . . 61
4.2 OCCA index plot . . . . . 63
4.3 SCCA index plot . . . . . 64
4.4 Computational cost . . . . . 65
4.5 Risperidone chemical structure . . . . . 67
4.6 Rimonabant substructure selected to be a clue of psychoacticity . . . 70
4.7 HCS data acquisition and processing . . . . . 76
4.8 Within population variability . . . . . 82
4.9 Model fitting . . . . . 90
4.10 Novelty detection and positive controls . . . . . 91
4.11 Model and empirical distributions . . . . . 92
4.12 Negative control outliers . . . . . 93
4.13 Relation between classes and population descriptors . . . . . 95
4.14 Example of a well . . . . . 96

5.1 Gene network for DREAM7 challenge . . . . . 103
5.2 Log likelihood surface . . . . . 110
5.3 Comparison of risk evolution between different strategies . . . . . 113
5.4 Comparison of final risks . . . . . 114
5.5 Comparison between true and predicted time course trajectory . . . . 116
5.6 Trajectories from posterior sample . . . . . 117
5.7 Comparison of parameter and trajectory variability . . . . . 118

# Chapter 1

# Introduction

Biotechnologies came to an era where the amount of information one has access to allows to think about biological objects as complex systems. In this context, the phenomena emerging from these systems are tightly linked to their organizational properties. This raises methodological challenges which are precisely the focus of study of the machine learning community. This thesis is about applications of machine learning methods to study biological phenomena from a complex systems viewpoint. This introduction specifies what is meant by system based and machine learning approaches. The biological applications treated in this manuscript are presented in a second phase.

## 1.1 System theoretic approaches in biology

The general idea that a system being made of different parts is, as a whole, something different from the sum of its parts dates back to antiquity and is rather well accepted today, specially in the scientific community. This idea is to be linked with the notion of scale. At any scale, natural objects are parts of larger systems and are themselves made of different parts. For example, a population of living organisms gathers different individuals which could be pluricellular, each cell consisting of an accumulation of big molecules themselves made of atoms etc ... This naturally leads to a hierarchy of objects, which behaviour is related to the scale at which they are observed.

The first aspect in the understanding of natural objects at a given scale, is the characterization of the object of study as an isolated entity. For example, in cell biology, a living cell is made of different compartments, evolves in time following cell cycle, and could potentially undergo a division. At a different scale, in molecular biology, a protein is a polymer of amino acids that is characterized by a three dimensional structure and has a specific function such as catalysis of a reaction (enzymes), transcription of DNA, or transportation of smaller molecules. The study of these individual characteristics at one scale helps understanding behaviours at higher levels in the hierarchy. In relation with the two previous examples, understanding a phenomenon occurring at the cell level, such as switching from one cell cycle state to another, requires to decipher the chemical reactions that underlie this process.

However, cell biology can not be reduced to the application of molecular principles. The reactions affecting cell cycle do not occur at random but rather in a well organized manner. Taking this global organization into account is crucial in order to understand molecular basis of cell cycle. It requires tools and methods that go beyond the field of molecular biology. More generally, the study of interactions between different levels in the hierarchy of scales requires to consider organization between scales, how small entities organize themselves to constitute bigger entities. This is referred to as a system approach. Observing and explaining several phenomena imposes to consider a set of entities as a whole complex system with specific organizational properties.

What could be the benefits of such an approach? In some cases, one cannot get by without these considerations. For example, if one is interested in characterization of the effect of a given chemical compound at the cell level, it is pointless to draw conclusions based on a single cell experiment. Different types of cells could eventually respond in different ways. Additionally, cells of the same type coming from the same population may exhibit variability in behaviour. The structure of inter cell variability within this population holds information about the underlying biological process. For instance, cells might not react in the same way whether they are surrounded by many other cells or not. Taking into account this variability by considering different cell types and several populations of each cell type is necessary in order to draw reproducible conclusions from such experiments. This is illustrated in one chapter of

this thesis.

In other cases the need to consider a system based approach arises from the highly non linear behaviours of interactions between parts constituting a bigger entity. Many observable phenomena are non linear in nature. For instance, mass action law, modelling chemical kinetics, can lead to highly non linear dynamics. Accumulating these non linearities leads to complex behaviours at the system level which investigation is impossible without such a system approach. Striking examples of this non linearity are bistable systems which can rest in either of two distinct states. Such systems are at the heart of triggering mechanisms that occur at the cellular level, such as cell division, cell differentiation or apoptosis. Chapter 5 of the thesis is dedicated to the optimization of experimental design for the characterization of non linear dynamical system.

System based approaches are not necessary in all circumstances. If not required, they could, however, be beneficial. For example, studying the spatial conformation at the site where a ligand binds to a protein does not require to take into account the whole surrounding molecular environment. However considering this specific interaction occurrence as a member of a set of interactions can shed light on the underlying molecular recognition mechanisms it involves. While all interactions are different, there might exist general trends in what characterizes an interaction. A system based approach can take advantage of this property and use it in a prediction context. These ideas are illustrated in the context of protein ligand interaction prediction and drug side effect prediction in chapters 3 and 4.

While considering the structural organization of groups of objects can be essential or beneficial to comprehend complex systems, it is not possible to apply such ideas in any context. The two main bottlenecks are knowledge and computational methods. It is required to grasp the characteristics of individual small entities in order to understand how they are organized at a larger scale. Advances in biological knowledge and technological abilities to carry out large scale experiments make it possible today to apply system based reasoning to the study of molecular, cellular and macroscopic phenomena. For example, large amount of data is now available about molecular interactions. Moreover, high throughput technologies, such as next

generation sequencing or high content screening, have substantially accelerated the data acquisition process and broadened the spread of experiments that could be carried out in a row. This considerably increases the size and complexity of encountered statistical and computational problems. The so-called machine learning field is exactly at the interface between these two concepts and provides solutions to tackle problems one faces in applying complex systems reasoning in a biological context.

## 1.2 Machine learning

While the field of machine learning is very broad, the problems of interest go from theoretical and empirical performances to statistical and computational trade off, this paragraph focuses on the specific aspects needed to introduce the work presented further in this thesis. Broadly speaking, there exists mainly two different ways of formulating problems addressed in this manuscript. They are referred to as supervised and unsupervised problems, as informally described here.

**Supervised problems** consist in the estimation of a functional relation between two categories of objects based on noisy observations of this relation. Informally, it is assumed that there exists spaces $\mathcal{X}$ and $\mathcal{Y}$ and a function $f\colon \mathcal{X} \to \mathcal{Y}$. One has access to noisy realizations of this functional relationship, meaning that several data points are available which are assumed to be of the form:

$$y_i = f(x_i) + \epsilon_i$$

where $x_i$ is in $\mathcal{X}$, $y_i$ is in $\mathcal{Y}$ and $\epsilon_i$ is random noise. The quantity of interest is the function $f$ which is unknown. The objective is to estimate this function or to accurately reproduce it. Many prediction problems can be formalized this way. The typical example in biology is the response of a system to an external stimulus. In this case $\mathcal{X}$ is a set of possible stimuli, for example chemical compounds, $\mathcal{Y}$ is the set of possible responses of an organism to these stimuli. The nature of the function $f$ and of the estimation procedure will depend on the structure of the input space $\mathcal{X}$ and

the structure of the output space $\mathcal{Y}$. Classification methods are dedicated to discrete and known possible values of $\mathcal{Y}$. For example, in the context of predicting protein targets for ligands, the input space $\mathcal{X}$ is the space of chemical compounds, and the output space $\mathcal{Y}$ reflecting the presence or absence of interaction with a given protein. Regression methods correspond to continuous real output, predicting affinity of a given compound for a certain protein would fall into this class. These methods are used in the context of drug target interaction and side effect prediction in chapters 3 and 4. In chapter 5 we specify the function $f$ as reflecting the dynamics of an underlying dynamical system which we try to characterize based on observation of the system under different experimental conditions.

As discussed in the previous paragraph, taking into account hierarchical structures observed at different scales is potentially beneficial in a biological context. Interestingly, it is possible to relate the structure of the biological problem to the structure of the input space $\mathcal{X}$ and the output space $\mathcal{Y}$ in a supervised setting. To be concrete, define the input space $\mathcal{X}$ as the space of chemical compounds and the output space as a binary digit vectors of size $p$ each entry of the vector reflecting the fact that a given molecule $x$ binds or not to each protein in a set of $p$ proteins. One possibility to solve this problem is to build one classifier for each protein in the considered set. It is also possible to adopt a system approach and share information between different proteins and adapt to the underlying structure of the set of proteins as illustrated in chapter 3.

**Unsupervised problems** consist in searching for hidden structures in a set of unlabelled data. The experimenter is provided with elements $x_i$ taken from an input space $\mathcal{X}$ but no label $y_i$ is given. Broadly speaking, the problem consists in finding meaningful trends that characterize well the typical elements taken from the input space $\mathcal{X}$. These characteristics can be further taken as input to solve supervised problems. A wide variety of methods are available to address unsupervised problems, among which three are informaly presented in this paragraph.

"Principal component-type methods" seek direction of large dispersion in a given dataset. These directions are those that explain observed variability in this dataset.

While typical datasets encountered in biology can be of very dimension, it happens that a few factors can explain most of the variations between data points. This results in a compact representation of the underlying dataset which can be used for prediction purpose or to understand typical trends in the data. This is illustrated in the context of drug target interactions and side effects in chapters 3 and 4.

"Clustering-type" methods seek to partition the input dataset into a few categories. The objective is to output meaningful clusters, data points from the same cluster being very similar and data points from different clusters sharing much less similarity. A specific model based clustering method is presented in the context of cell population phenotyping in chapter 4.

"Density estimation" is the task of defining to which extent typical data points look like, and to differentiate them from data points which are far from them. Methods for solving these problems can be used in the context of outlier detection. In this case one is interested in finding data points that are different from most of the other points, without a predefined criterion to describe these differences. In chapter 4, the model developed for cell population phenotyping is used for such purpose in the context of the study of cell populations.

By nature, unsupervised methods seek to shed light on underlying structure in a dataset. Thus, they are well suited to tackle biological problems at complex systems scale. Furthermore, one can put more emphasis on specific structural assumptions by choosing one type of methods or by adapting a specific method to the context at hand, as specific examples shown in this thesis will illustrate.

## 1.3 Contribution and organization of the thesis

The previous paragraph gave a brief overview of the methods that are used in this work. Chapter 2 provides more technical details and pointers to the literature related to the specific methods used in the projects that are presented. The remaining chapters illustrate these methods on specific biological applications. The biological problems addressed in these chapters constitute relevant questions in drug design (chapter 3 and 4), interpretation of high throughput imaging technologies (chapter 4)

and systems biology (chapter 5). Each chapter introduces the biological questions it addresses as well as related background. The specific focus of interest is put in context and the contributions of the presented work are described. Specific methodological details as well as experimental results are then presented.

**Chapter 3** focuses on the molecular scale and more specifically on the study of protein-ligand interactions. Protein-ligand interactions constitute the key molecular mechanism that drives most important biological processes such as signal transduction or catalysis of metabolic reactions. They are also of interest in the context of drug design, the objective being to disrupt a biological process by modifying the behaviour of one of a protein related to this process, through a molecular interaction with a drug ligand. Sparse matrix factorization techniques are applied to reveal associations between chemical substructures and protein domains underlying these interactions. Examples show that this method extracts relevant information in this context. This information is further used in a prediction context and show performances comparable to state-of-the-art methods. An extension of these ideas using very high dimensional linear classifiers is briefly mentioned.

**Chapter 4** presents results related to phenotyping. In a first part, drug side-effects are investigated using matrix factorization techniques similar to those used in chapter 3. Drug-side effects are the result of the interactions of a small drug molecule with all its potential protein targets. The resulting effect can be viewed as a phenotype for the considered molecules. This chapter focusses on the relation between drug chemical structure and its side-effects. Examples of modulation of side-effects through chemical structure and links with protein interaction profiles are proposed. Performance of the method in the context of supervised side-effect prediction are compared to those of state-of-the-art classification methods. An extension of this work involving the integration of both chemical and protein target interactions information is qualitatively described.

A second part tackles the question of defining cellular phenotypes at the scale of

cell populations based on fluorescent images. Single cell phenotyping based on fluorescent images is a well understood problem. The question of comparing different populations of cells with variable cellular characteristics is less understood. A generative model is proposed to tackle this problem. Numerical experiments are carried out based on high-content screening data recorded to study the Ewing sarcoma disease. Properties of the proposed model are illustrated through various examples.

**Chapter 5** is dedicated to the presentation of a sequential experimental design tool for dynamical systems characterization. The related biological scale lies in between the molecular scale considered in chapter 3 and the phenotypic scale considered in chapter 4. Common models of molecular interactions in systems biology involve non linear dynamical systems with unknown parameters. Estimating these parameters from data is crucial to validate and use these models in practice. The non linearities and the lack of data available lead to parameter non identifiabilities, many different combinations of parameters agree with available data. The objective is to propose a strategy that identifies these non identifiabilities and proposes experiments to mitigate them. The need for design strategies based on numerical criteria is motivated. Bayesian and active learning ideas are used to define such a strategy. Numerical approximations to implement this strategy are provided and simulation results are presented. All the results presented in this chapter are based on numerical simulations of the experimental design process. The motivation and materials for designing and testing this method were provided in the context of DREAM7 Network Parameter Inference Challenge.

# Chapter 2

# Methods

## Résumé

Ce chapitre présente les détails algorithmiques et méthodologiques liés aux travaux présentés dans ce manuscrit. Son contenu est disponible dans la littérature. Ces résultats sont rappelés ici pour mémoire et afin de donner des liens vers la littérature correspondante, sans chercher à être exhaustifs. La première section est dédiée à la présentation de l'algorithme de factorisation parcimonieuse de matrice présenté par [129] et qui est utilisé dans les chapitres 3 et 4. Nous donnons ensuite un aperçu des méthodes d'apprentissage supervisé utilisées dans les mêmes chapitres, ainsi que leurs fondements statistiques. La troisième section présente les algorithmes d'inférence approchée utilisés dans les chapitres 4 et 5 dans le contexte de la modélisation probabiliste et de la modélisation Bayésienne. La dernière section décrit la procédure de validation croisée utilisée pour l'évaluation de modèles dans les chapitres 3 et 4.

## Abstract

This chapter gathers methodological and algorithmic details related to the work presented in this manuscript. All the results presented here are available in the literature. They are recalled here to give pointers to the related articles. The focus is not on being technically exhaustive. The first section is dedicated to the presentation of the

sparse matrix factorization of [129] which is used in chapters 3 et 4. A brief overview of supervised learning methods used in the same chapters is then given, with a rapid overview of the statistical foundations of these methods. The third section presents approximation algorithms used for inference in chapters 4 and 5 in the context of probabilistic and Bayesian modelling. The last section describes the cross validation scheme used for model assessment in chapters 3 and 4.

## 2.1 Sparse matrix factorization

The sparse matrix factorization problem is an area of very intense research. From an algorithmic point of view, the purpose is to approximate an input matrix by the product of two sparse factor matrices or to find a low rank approximation of the input matrix which factors are sparse. From a statistical point of view, this is related to the sparse PCA problem, which is the unsupervised task of separating signal with sparse structure from the environment noise. Various algorithms have been proposed to tackle this problem with different statistical and algorithmic properties. This section describes the algorithm presented in [129] which has low theoretical guaranties but is applicable to large scale problems. This algorithmic choice is briefly compared to other possible choices, and applications in unsupervised data analysis are described. They correspond to the methods used in further chapters. Finally, a validation procedure based on reconstruction ability of the unsupervised factorization is described.

### 2.1.1 Singular value decomposition

The method presented in [129] consists in incorporating sparsity-inducing constraints in a known algorithmic frameworks to solve regular low rank matrix approximation. First, the singular value decomposition is briefly described. Let $Z \in \mathcal{M}_{n\times p}(\mathbb{R})$ be a real matrix with $n$ rows and $p$ columns of rank $K \leq \min(n,p)$. The singular value decomposition of $Z$ is the unique triplet of matrix $(U,D,V)$ satisfying

$$Z = UDV^T, U^TU = I_n, V^TV = I_p, d_1 \geq d_2 \geq \ldots \geq d_K > 0$$

where $U$ is an unitary matrix of size $n \times n$, $V$ an unitary matrix of size $p \times p$ a rectangular matrix of size $n \times p$ which entries are null except for the $K$ diagonal elements given by $d_1, d_2, \ldots, d_K$. Columns of $U$ are eigenvectors of $ZZ^T$, columns of $V$ are eigenvectors of $Z^TZ$. Let $u_k$ be the columns of $U$, $v_k$ the columns $V$ and $d_k$ the $k$-th element of the diagonal of $D$. It is a well known result [30] that

$$\sum_{k=1}^{r} d_k u_k v_k^T = \arg\min_{\widehat{Z} \in M(r)} ||Z - \widehat{Z}||_F$$

where $M(r)$ is the set of matrices of dimension $n \times p$ with rank $r$ and $|| \cdot ||_F$ is the Froebenius norm. The singular value decomposition allows to find the matrix of rank $r$ which is the best approximation to $Z$ in the sense of the Froebenius norm. Interestingly, this is a non convex problem which solution is analytically expressible as an eigenvalue problem. Moreover, it is easy to see that:

$$(u_1, v_1) = \arg\max_{u,v} u^T Z v \qquad ||u||_2 = ||v||_2 = 1. \tag{2.1}$$

A widely used iterative scheme to compute solutions of the previous problem is to initialize $v^{(0)}$, such that $||v^{(0)}||_2 = 1$ and to repeat until convergence :

- $u \leftarrow \arg\max_u u^T Z v \qquad ||u||_2 \leq 1$
- $v \leftarrow \arg\max_v u^T Z v \qquad ||v||_2 \leq 1$

which reduces to the well known power method,

- $v^{(i+1)} = \frac{(Z^TZ)v^{(i)}}{||(Z^TZ)v^{(i)}||_2}$
- $u^{(i+1)} = \frac{(ZZ^T)v^{(i)}}{||(ZZ^T)u^{(i)}||_2}$

which converges to the largest singular value of $Z$ under the condition that it is higher than all the others and that $v^{(0)}$ is not orthogonal to the singular vector associated to this singular value. This scheme allows to compute the leading singular vector, which, together with a deflation procedure, allows to compute the full decomposition.

1. $Z^{(1)} \leftarrow Z$ .

2. for $k \in \{1, \ldots, K\}$

   - find $u_k$, $v_k$ et $d_k$ using the power method with $Z^{(k)}$
   - $Z^{(k+1)} \leftarrow Z^{(k)} - d_k u_k v_k^T$

### 2.1.2 Sparse matrix decomposition

The penalized matrix decomposition algorithm of [129] consists in adding a sparsity constraint to problem (2.1) and uses the same alternate minimization scheme to solve the problem. The formulation is as follows:

$$(u_1, v_1) = \arg\max_{u,v} u^T Z v \qquad ||u||_2 = ||v||_2 = 1, ||u||_1 \leq c_1, ||v||_2 \leq c_2 \tag{2.2}$$

where $c_1$ and $c_2$ are tuning parameters. The iterative scheme becomes:

- $u \leftarrow \arg\max_u u^T Z v \qquad ||u||_2 \leq 1, ||u||_1 \leq c_1$
- $v \leftarrow \arg\max_v u^T Z v \qquad ||v||_2 \leq 1, ||v||_1 \leq c_2.$

The minimization steps in the sub problems is easily solvable. For example, it can be shown that the first step has a solution of the form:

$$u \leftarrow \frac{S(Zv, \delta_1)}{||S(Zv, \delta_1)||_2} \tag{2.3}$$

where $S$ is the soft thresholding operator applied to each entry of the vector. It has the form $S\colon (x, \delta) \rightarrow sign(x)(|x| - \delta)_+$, where $(.)_+$ denotes the positive part. However, the amount of thresholding $\delta$ is unknown. In 2.3, it should be chosen to be null, $\delta_1 = 0$, if the application of 2.3 results in $||u||_1 \leq c_1$, otherwise, $\delta_1$ is a positive constant such that $||u||_1 = c_1$, where the precise value of the constant is found using line search.

As described in [129], the method does not come with any algorithmic or statistical property. It was recently shown in [83] that the proposed alternate minimization scheme converges to a critical point of problem (2.2). Broadly speaking, methods

proposed to solve the sparse matrix factorization problem can be divided into two classes.

1. Convex relaxation methods leading to a semi-definite programming formulation of the problem, whose global minimizer can be computed, for example [26]. State-of-the-art semi-definite programming algorithms are still computationally expensive which limit their application on large datasets.

2. On the other hand, local search iterative methods can only produce local optima for some non convex objective. They are however much less expensive than convex relaxation methods. They can be applied to larger datasets but cannot guaranty that the best solution is found due to multiple critical points.

The algorithm presented here is one of the second kind. It does not guaranty global optimality, but it is much cheaper in term of computation. The algorithm described in [129] was empirically shown to perform well on high dimensional biological applications. This algorithmic choice allowed to perform extensive experiments and parameter tuning on the datasets considered in chapters 3 and 4.

### 2.1.3 Application in data analysis

Many unsupervised data analysis methods amount to solve problem (2.1) for a given input matrix. Canonical correlation analysis [55] and canonical correspondence analysis [45] are described in this paragraph. Sparse version of these methods involving the algorithmic scheme described in the previous paragraph are evaluated on biological examples in chapters 3 and 4.

**Canonical correlation anaylsis.** This method allows to study links between two different representations of the same objects. Let $X \in \mathcal{M}_{n\times p}(\mathbb{R})$ and $Y \in \mathcal{M}_{n\times q}(\mathbb{R})$ be two representations of a set of $n$ objects according to two vector variables of dimension $p$ and $q$ respectively. Columns of $X$ and $Y$ are centered and scaled. The matrix $X^TY$ is an empirical estimate of the correlation structure based on these $n$ individuals. The goal is to study underlying correlations between these two variables.

To do so, we look for linear combinations of variables that are strongly correlated. This amounts to solve

$$\max_{u,v} cor(u,v) = \frac{u^T X^T Y v}{\sqrt{u^T X^T X u} \quad \sqrt{v^T Y^T Y v}}$$

which is equivalent to

$$\max_{u,v} u^T X^T Y v \qquad u^T X^T X u = 1 \qquad v^T Y^T Y v = 1.$$

**Canonical correspondence analysis.** This method allows to study links between two different sets of objects represented by different variables based on a contingency table representing co-occurrences of the considered objects. Let $X \in \mathcal{M}_{n\times p}(\mathbb{R})$ and $Y \in \mathcal{M}_{m\times q}(\mathbb{R})$ represent a first set of $n$ objects in dimension $p$ and a second set of $m$ objects in dimension $q$. A contingency table $A \in \mathcal{M}_{n\times m}(\mathbb{R})$ represents the co-occurrences of the two sets of objects. As for canonical correlation analysis, the goal is to find linear combinations of variables which are strongly correlated by solving

$$\max_{u,v} cor(u,v) = \frac{u^T X^T A Y v}{\sqrt{u^T X^T D_X X u} \quad \sqrt{v^T Y^T D_Y Y v}}$$

where $D_X$ (resp $D_Y$) is a diagonal matrix which entries are the degree of the categorical variables $X$ (resp $Y$). This is equivalent to

$$\max_{u,v} u^T X^T Y v \qquad u^T X^T D_X X u = 1 \qquad v^T Y^T D_Y Y v = 1.$$

**Simplification and addition of sparsity-inducing constraints.** The methods described in this paragraph amount to solve numerical problems which structures are similar to that of (2.1). Empirical evidences suggest that in a high dimensional context, covariance matrices can be approximated by diagonal matrices [120, 29]. This simplification leads to the replacement of constraints of the form $u^T X^T X u = 1$ by a simpler constraint $||u||_2 = 1$. Adding sparsity-inducing constraints leads to sparse variants of canonical correlation analysis and canonical correspondence analysis which have the exact same form as that of (2.2) and can be solved using the numerical scheme

of [129].

The motivation behind the use of sparsity is twofold. Depending on the problem at hand, it could be expected that principal factors should be sparse. For example, in the case of protein ligand interactions, as illustrated in chapter 3, physics underlying the molecular recognition mechanisms suggests that the presence or absence of a limited number of chemical substructures is a key to explain why a given molecule binds to a given protein. One can expect gains in performance by accounting for this hypothesis. Moreover, the factor extracted by the penalised decomposition procedure should be interpretable for practitioners. In high dimensional settings, this analysis is much more challenging when the factors are dense, *i.e.* when they contain many non zero elements. Enforcing sparsity allows eases the interpretation of the extracted factors.

### 2.1.4 Validation in the context of supervised prediction context

Both methods lead to the penalised decomposition of a matrix $Z \in \mathcal{M}_{p\times q}(\mathbb{R})$ of the form $Z \simeq \sum_{i=1}^{K} \rho_i u_i v_i^T$ for a given number of extracted factors $K$. Each factor represents directions of strong correlation in the dataset. Suppose that one is given two new objects $x \in \mathbb{R}^p$ and $y \in \mathbb{R}^q$. Projecting these objects on the subspace spanned by the principal factors leads to the following score

$$s(x,y) = \sum_{i=1}^{K} \rho_i x^T u_i v_i^T y$$

which represents the strength of the association between these new objects based on the extracted principal factors. This score can be used to make predictions. For example if $x$ represents an unseen molecule and $y$ represents a known protein, the score $s$ can be used as a confidence level for the association between $x$ and $y$. It is natural to validate unsupervised methods on supervised reconstruction tasks. This point is illustrated in chapters 3 and 4.

## 2.2 Supervised learning

Many state-of-the-art methods for supervised problems are related to the statistical learning theory. This section proposes an overview of the learning methods used in this manuscript as baselines to evaluate the performance of the methods proposed in this thesis. The principle of empirical risk minimization is described first. Kernel function classes are then briefly described, as well as nearest neighbour methods that are used for comparison purposed in chapters 3 and 4. The purpose is to motivate the choice of these methods as comparison points, and to give a very fast overview of the underlying theory. The book [51] provides detailed materials as well as pointers to the bibliography.

### 2.2.1 Empirical risk minimization

Inductive supervised learning is the task of learning a functional relationship based on examples. Informally, we suppose that there exists an input space $\mathcal{X}$ and an output space $\mathcal{Y}$ (usually $\mathbb{R}$ for regression tasks or $\{-1, 1\}$ for classification tasks) and an hypothetical function $f^*\colon \mathcal{X} \to \mathcal{Y}$ which represents the best that anybody could do when predicting the output from the input. This expression is understood in the following sense. In order to model uncertainty, the relation between $\mathcal{X}$ and $\mathcal{Y}$ is random. $f^*$ is the prediction function that works the best on average.

In practice, one is given a training set of examples of size $n$, $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$, $i = 1 \dots n$, a function class $\mathcal{F}$ of functions $\mathcal{X} \to \mathcal{Y}$ and a loss function $L\colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$. Many learning algorithms consist in solving

$$\hat{f} = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} L(f(x_i), y_i) \tag{2.4}$$

which is minimizing the risk evaluated on a finite sample. The theoretical properties of this procedure is the object of study of statistical learning theory. For example, using tools from empirical process theory, under restrictions on the function class $\mathcal{F}$, the empirical risk minimization procedure is shown to be consistent. This means that when the problem is embedded in a probabilistic structure, as more and more

data comes in, the empirical risk minimization procedure for certain function classes provides an estimate $\hat{f}$ which is close to the best estimate one could expect $f^*$. [123] provides detailed arguments as well as pointers to the literature. Practically, the implication of these results are the following. If there exists a hypothetical true function, the estimated function $\hat{f}$ should be close to it, provided that enough data is available. It can then be used in a prediction context to interpolate the information contained in a training set to unseen examples.

### 2.2.2 Choice of the loss function

Different function classes and different loss functions define different estimators. The loss function depends on the problem at hand. It should reflect how a given function fits to the training set, and allow to solve the problem (2.4) efficiently. For example, in the case of regression, the output space $\mathcal{Y}$ is a continuous subset of $\mathbb{R}$. In this case, the most popular loss function is the square loss $L_s\colon (y_1, y_2) \to (y_1 - y_2)^2$. In binary classification tasks, the output space is binary, identified with $\{-1, 1\}$. Popular loss functions in this context are the logistic loss $L_l(y_1, y_2) = \log(1 + e^{-y_1 y_2})$ which defines logistic regression and the hinge loss of the support vector machine, $L_h(y_1, y_2) = (1 - y_1 y_2)_+$ where $(.)_+$ denotes the positive part. These losses are used in chapters 3 and 4.

### 2.2.3 Kernels and kernel methods

The second aspect of the definition of an estimator in the framework of supervised learning is to choose a function class. One of the simplest example is the class of linear functions which are widely used in regression and classification. Kernels are mathematical objects that have the same properties as an inner product. Positive definite kernels have the appealing property that they allow to define classes of non linear functions that are tractable in the sense that problem of the form of (2.4) can be solved efficiently. They allow to perform learning with objects that cannot be embedded in vector spaces, such as graphs or trees, and to generalize linear methods to non linear settings. Moreover, they exhibit state-of-the-art performances on many

real world problems. Support vector machines involving kernel for molecules and kernel regression are used in chapters 3 and 4. Many more details about positive definite kernels and practical examples are found in [104]. The reproducing Hilbert space theory describes how positive definite kernels define function classes. A detailed exposition is found in [8].

### 2.2.4 Nearest neighbours

Given the training set $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$, when a distance is available on $\mathcal{X}$, the $k$-nearest neighbours methods consists in choosing as an estimate the function

$$\hat{f} \colon x \in \mathcal{X} \to \frac{1}{k} \sum_{x_i \in N_k(x)} y_i$$

where $x_i \in N_k(x)$ when $x_i$ is one of the $k$ nearest neighbours of $x$, and $k$ is a tuning parameter. This constitutes a baseline method that is used in chapters 3 and 4 for performance assessment.

## 2.3 Inference and learning in probabilistic models

This section presents algorithmic details related to the methods that are proposed in chapters 4 and 5. Broadly speaking, inference is the task of computing expectations and learning is the task of computing local optima under some probabilistic distribution. Despite the simplicity of the formulation, this turns out to be often intractable on real world problems. Numerical approximations are required in many practical cases. A first paragraph briefly presents the Expectation Maximization algorithm which is used in chapter 4 and the second paragraph present Monte Carlo Sampling Methods used in chapter 5. The presentations is brief and informal. The book [77] provides details about probabilistic graphical models and inference methods and the article [5] is a good introduction to sampling algorithm for probabilistic modelling.

### 2.3.1 EM algorithm

This is a maximum likelihood procedure that allows to perform learning in generative probabilistic models with partially observed data [27]. A generative model is the definition of a generative stochastic process from which a dataset $\mathcal{D}$ is supposed to be an independent identically distributed sample. In this paragraph, models with hidden random variables are considered, denoted by $Z$. If $\theta \in \Theta$ is a parameter belonging to some space, a generative process defines a density function:

$$P(\mathcal{D}, Z|\theta).$$

This density function assigns a likelihood to the complete data (observed and non observed) given the parameter. This is referred to as complete likelihood. Model fitting consists in finding the parameter value $\theta^*$ that maximizes the marginal likelihood:

$$P(\mathcal{D}|\theta) = \sum_{Z} P(\mathcal{D}, Z|\theta).$$

The fact that some variables are not observed leads to intractable likelihood functions in the sense that their evaluation are computationally very expensive. Indeed, the size of the space in which the variable $Z$ is embedded often grows exponentially with the size of the dataset $\mathcal{D}$. This is the case in clustering applications for example. The EM algorithm is a local search method that is based on the following two steps

- **Expectation step:** Compute a tractable lower bound on the likelihood function. This is done by computing the expectation of the complete log likelihood given the current parameters, integrating out the hidden variable $Z$.
- **Maximization step:** Find the parameters that maximize the expression computed in the previous step.

Alternating these two steps allows to find a local maximum of the log likelihood function. This algorithm is used in chapter 4 for fitting the probabilistic model designed for cell population phenotyping. This algorithm is one of the most simple instances of variational inference methods which are widely used in probabilistic modelling [66].

### 2.3.2 Markov Chain Monte Carlo and importance sampling

The Bayesian approach considers that the parameters of a model are random quantities. If one is given a likelihood function $P(\mathcal{D}|\theta)$ where $\mathcal{D}$ is a dataset and $\theta$ is a vector of parameters. In the Bayesian framework, one defines a prior distribution over the parameter space $\pi_0(\theta)$ and explores the posterior density which takes the form:

$$\pi_1(\theta|\mathcal{D}) \propto P(D|\theta)\pi_0(\theta).$$

Inference under this density is often intractable analytically or even numerically. Sampling methods are designed to sample from a probability distribution which can only be evaluated up to a constant multiplicative factor, which is precisely the case here. The main application is to approximate expectations of given functions based on this sample. Suppose that we are exploring a space $\mathcal{X}$ endowed with a probability measure $\mathbb{P}$. Let $\{x_i\}_{i=1}^n \in \mathcal{X}^n$ be an independent identically distributed sample from $\mathbb{P}$. Given a function $f\colon \mathcal{X} \to \mathbb{R}$, one can approximate the expectation of $f$ under $\mathbb{P}$ by:

$$\int_{\mathcal{X}} f(x)d\mathbb{P}(x) \simeq \frac{1}{n}\sum_{i=1}^{n} f(x_i).$$

The convergence of the approximation to the true numerical value is guaranteed by the strong law of large numbers. Now, suppose that one wishes to compute the expectation of $f$ under another probability distribution $\mathbb{Q}$, one can use the same sample to approximate this value numerically.

$$\int_{\mathcal{X}} f(x)d\mathbb{Q}(x) \simeq \frac{1}{\sum_{i=1}^{n} w_i(x)}\sum_{i=1}^{n} f(x_i)w_i(x).$$

Where $w_i(x) = \frac{d\mathbb{P}(x)}{d\mathbb{Q}(x)}$. This estimate is biased but also converges to the true value as the sample size grows. It is referred to as the importance sampling method. These two methods are used in chapter 5 in the context of Bayesian inference and risk estimation.

In order for those methods to work, one needs to be able to sample from a probability distribution $\mathbb{P}$. In chapter 5, we use a Markov Chain which elements can be proven to be asymptotically distributed according to the distribution of interest. The main sampling algorithm used is Metropolis Hasting which is the most famous Markov Chain Monte Carlo sampling algorithm [11]. It consists in exploring the parameter space based on a random walk guided by the distribution of interest, rejecting or accepting random moves depending on this distribution. [5] is a good introduction to sampling numerical approximation methods.

## 2.4 Empirical model evaluation and cross validation

Many aspects of this thesis relate to the evaluation of machine learning methods in a biological context. By evaluation, we mean how well a method produces correct predictions on an unseen dataset after being trained on a training set of known examples. The first method to empirically assess performance of a statistical procedure is to test it on a synthetic dataset. This approach is used in chapter 5. Empirical evaluation on real data would require to train a method on a training set, to acquire a new dataset of the same kind and evaluate how well the method generalizes on this new dataset. Cross validation mimics this process. $V$-fold cross validation consists in partitioning a dataset in $V$ parts of roughly equal size, train a method on $V-1$ subsets and test it on the held out set. Repeating this procedure multiple times gives an idea of the generalization performance of the method and provides clues about the robustness of these performances. Chapter 7 in [51] provides more details about this procedure. A historical and technical survey is available in [7].

# Chapter 3

# Protein-ligand interactions

## Résumé

L'identification de règles sous-jacentes à la reconnaissance entre les sous structures chimiques d'un ligand et le site fonctionnel d'une protéine constitue un problème important pour la compréhension des mécanismes conduisant à des effets phénotypiques à plus grande échelle. Ce chapitre se concentre sur l'identification de telles règles. Nous décrivons une méthode nouvelle pour extraire des ensembles de sous structures chimiques et de domaines protéiques qui sous tendent les interactions protéine ligand à l'échelle du génome. La méthode est basée sur l'analyse canonique des correspondances parcimonieuse (SCCA) pour l'analyse conjointe de profils de sous structures moléculaires et de domaines protéiques. Une approche par classification $L_1$ pénalisée pour extraire des associations prédictives entre sous-structures et domaines protéiques est également décrite, et comparée à l'approche SCCA.

Les résultats expérimentaux sont basés sur un jeu de données d'interactions protéine-ligand contenant des enzymes, des canaux ioniques, des récepteurs couplés aux protéines G et des récepteurs nucléaires. La méthode SCCA extrait des ensembles de sous structures partagées par des ligands qui peuvent se fixer à un ensemble de domaines protéiques. Ces deux ensembles de sous-structures chimiques et de domaines protéiques forment des composantes qui peuvent être exploitées dans un cadre de découverte de médicaments. Cette approche regroupe des domaines protéiques qui

ne sont pas nécessairement liés d'un point de vue évolutif, mais qui partagent des ligands présentant des structures chimiques similaires. Plusieurs exemples montrent que cette information peut être utile dans un cadre de prédiction d'interaction protéine-ligand ainsi que pour aborder le problème de la spécificité d'un ligand. Nous décrivons une comparaison numérique entre les deux méthodes proposées, SCCA et classification $L_1$ pénalisée, ainsi que des méthodes à l'état de l'art, sur la base du problème de prédiction d'interactions protéine-ligand. Ce chapitre se base principalement sur l'article [137] et une partie des résultats présentés dans [118].

# Abstract

The identification of rules governing molecular recognition between ligand chemical substructures and protein functional sites is a challenging issue for the understanding of molecular mechanisms driving larger scale phenotypic effects. This chapter focuses on the identification of such rules. We describe novel methods to extract sets of chemical substructures and protein domains that govern molecule-target interactions on a genome-wide scale. The method is based on sparse canonical correspondence analysis (SCCA) for analyzing molecular substructure profiles and protein domain profiles simultaneously. An $L_1$ penalized classification approach that extracts predictive associations between substructure and protein domains is also described, and compared to the SCCA approach.

Experimental results are based on a dataset of known protein-ligand interactions including enzymes, ion channels, G protein-coupled receptors and nuclear receptors. SCCA extracts a set of chemical substructures shared by ligands able to bind to a set of protein domains. These two sets of extracted chemical substructures and protein domains form components that can be further exploited in a drug discovery process. This approach successfully clusters protein domains that may be evolutionary unrelated, but that bind a common set of chemical substructures. As shown on several examples, it can also be very helpful for predicting new protein-ligand interactions and addressing the problem of ligand specificity. We describe a numerical comparison between the two proposed methods, SCCA and $L_1$ penalized supervised classification,

and state-of-the-art approaches based on supervised protein ligand interaction prediction task. The chapter is mostly based on [137] and results from [118].

# 3.1 Introduction

## 3.1.1 Protein-ligand interactions

A ligand is a small chemical compounds which interferes with the biological behaviour of its target proteins by direct physical interaction with it. These processes lead to phenotypic effects observed at larger scale (cellular or even macroscopic scale). For example, when a xenobiotic molecule binds to a protein, it perturbs the whole interaction network of this protein, affecting many underlying biological processes. In other words, the effect of a small molecule inside a living organism is the result of the disturbance of all the biological processes which involve a protein to which the molecule binds. It is therefore required to adopt a global point of view and consider all potential targets at the same time when studying these phenomena .

Conventional approaches, such as QSAR and docking can handle only a single protein at a time, and therefore, do not allow to adopt such a global point of view and treat a large set of potential target proteins at a genomic scale. The results discussed in this chapter constitute a methodological contribution toward the adoption of system approaches in the context of protein-ligand interaction predictions. A wide scale of proteins and chemical compounds are considered and jointly analysed. The main application of the approach is the design and optimization of drugs.

## 3.1.2 Existing approaches and motivations

A commonly used computational approach to analyze and predict ligand-protein interactions is docking. Docking approaches consist in finding the preferred orientation of a molecule binding to a protein by modelling the underlying physical energies (see [76] for recent review). Therefore, docking cannot be applied to proteins with unknown 3D structures. Moreover, docking protocols need to be tuned for each protein target, which prevents its use on a large number of proteins at the same time. This

limitation is critical in the case of membrane proteins such as G protein-coupled receptors (GPCRs) which are signal transduction pathway activators, or ion channels which shape electrical signals. Indeed, these membrane proteins are difficult to express, purify and crystallise. Although being both major therapeutic targets, their 3D structure is known to be particularly difficult to determine.

In this context, the importance of chemogenomic approach has grown fast in recent years [67, 117, 28], and a variety of statistical methods based on chemical and genomic information have been proposed to predict drug-target or more generally, ligand-protein interactions. These methods assume that similar proteins bind similar ligands. This assumption is often verified in practical cases and allows to predict protein-ligand interactions for new chemical compounds and new proteins. These methods differ by the underlying description used for proteins and ligands, and by how similarities between these objects are measured. Examples are statistic-based methods that compare target proteins by the similarity of the ligands that bind to them, which can then be used to predict new protein-ligand interactions [46, 70]. Other approaches are the binary classification approaches such as support vector machine with pairwise kernels for compound-protein pairs [86, 33, 59] which we use as a state-of-the-art comparison point in this chapter, and the supervised bipartite graph inference with distance learning based on chemical and genomic similarities [134, 133].

Ligand-protein interactions are often due to the presence of common chemical structures (the pharmacophore) that are usually shared by the ligands of a given protein, whereas this is not expected for random compounds that do not bind to the same protein. Recently, a variety of analyses have been conducted, such as analysis of chemical substructures and biological activity [75], data mining of chemical structural fingerprints and high-throughput screening data in PubChem [50], or extraction of chemical modification patterns in drug development [108]. Ligand-protein interactions are also due to functional sites of proteins (e.g., binding pockets, domains, motifs). Recently, the comparison of binding pockets has been done to investigate the relationship with their ligands [85, 87, 53]. However, these methods require the availability of the 3D structure of proteins. To date, most of the research has been

performed separately from the viewpoints of either ligands or proteins. Yet, the relevant question is how to relate ligand chemical substructures with protein functional sites in terms of ligand-protein interactions. There is therefore a strong incentive to conduct an integrative analysis of both ligand substructures and protein functional sites toward understanding of ligand-protein interactions. In this domain, a challenging issue is to develop methods that identify rules for molecular recognition between ligand chemical substructures and protein functional sites. The proposed methods tackle this problem based on co-occurrences of protein domains and molecular substructures in interacting and non-interacting protein-molecule pairs.

### 3.1.3 Content of the chapter

In this chapter, we describe two novel methods to extract sets of drug chemical substructures and protein domains that govern drug-target interactions. The first one is based on canonical correspondence analysis (CCA) for analyzing drug substructure profiles and protein domain profiles simultaneously. We develop an extension of the CCA algorithm by incorporating sparsity for easier interpretation, which we call sparse canonical correspondence analysis (SCCA). Figure 3.1 shows an illustration of the proposed method. The main interest and originality of the proposed method is that it correlates protein domains to chemical substructures expected to be present in their ligands, based on a learning dataset. From a system point of view, one strength of the method is that it allows to analyse jointly the interactions of many proteins and many molecules. This allows to point out interaction patterns that would not have been foreseen by looking at each interaction separately. Examples illustrate the fact that the method identifies pharmacophores automatically, thus providing structural insights about the mechanisms that govern molecular recognition. These pharcophores are shared by common ligands of a protein. Beyond the protein-ligand interaction problem, the examples we provide illustrate the benefits of system based approach when mining large interaction networks involving many chemical species. This work is based on the article [137].

An extension of this work leads to another method to tackle the same problem. It

uses an explicit representation of the tensor product space of possible pairs of molecular substructure and protein domain together with $L_1$-regularized linear support vector machine. This method is referred to as $L_1$-PLSVM and also extracts pairs of protein domains and chemical substructures that explain protein target interactions. This constitutes a sub-part of the results presented in [118]. Most of the content of the chapter is dedicated to SCCA method, but a comparison between $L_1$-PLSVM and SCCA is made based on their prediction performances.

We first describe the dataset that has been used to conduct those experiments. Then the different methods are presented as well as a description of the numerical experiments that allow to compare different methods from a supervised learning point of view. We compare to the reconstruction performances of a baseline and a state-of-the-art method designed for similar tasks. Biological examples illustrate the ability of the SCCA method to point out meaningful insights when treated as an unsupervised method. Numerical results highlight the different scenarios for which SCCA or $L_1$-PLSVM provide the best of their performances.

## 3.2 Materials

Drug-target interactions were obtained from the DrugBank database which combines detailed data about drugs and drug candidates with comprehensive drug-target information [128]. The version of DrugBank is 2.5. Proteins belong to many different classes, among others, pharmaceutically useful ones such as enzymes, ion channels, G protein-coupled receptors (GPCRs) or nuclear receptors. In this study, we focused on human proteins, which drove us to select all interactions involving human proteins. This led to build a protein-drug dataset containing 4809 interactions involving 1554 proteins and 1862 drugs. The set of interactions is used as gold standard data.

To encode the chemical structures of drugs involved in these interactions, we used a fingerprint corresponding to the 881 chemical substructures defined in the PubChem database [22]. Each drug was represented by an 881 dimensional binary vector whose elements encode for the presence or absence of each PubChem substructure by 1 or 0, respectively. Most of the drugs documented in DrugBank have a link to PubChem,

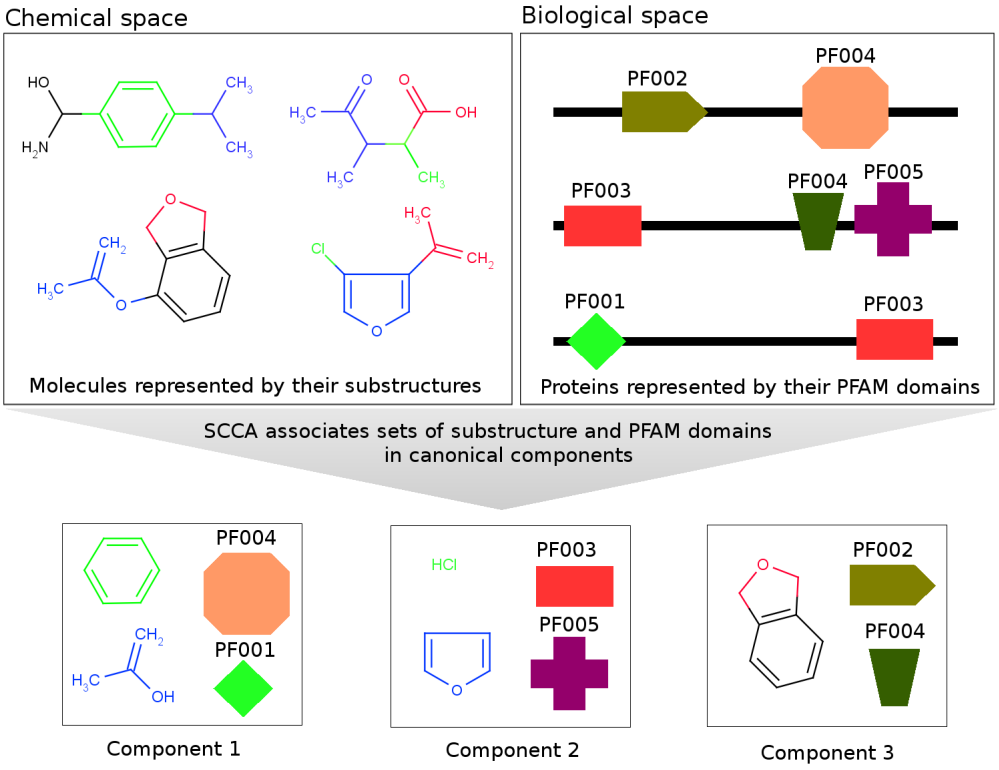



Figure 3.1: An illustration of the proposed method.

but some do not, mainly biotech drugs and mixtures. Interactions involving drugs that have a record in PubChem were kept. Among the 881 substructures used to represent the chemical structures, 663 are actually used, because 218 do not appear in our drug set.

For all proteins, genomic information and annotation were obtained from the UniProt database [6], and associated protein domains were obtained from the PFAM database [35]. PFAM database gathers a large number of protein functional domains. They are regions of the amino acid sequence that are associated with a specific molecular function (*e.g.* recognising site of an enzyme that catalyses a chemical reaction). A protein might have several domains related to several functions which allow to recognize several molecules. These domains implicitly represent proteins by their

functions. In the set of proteins we took into account, 876 PFAM domains are found. Therefore, each protein was represented by a 876 dimensional binary vector whose elements encode for the presence or absence of each of the retained PFAM domain by 1 or 0, respectively.

## 3.3 Methods

We want to extract drug chemical substructures and protein domains which tend to jointly appear in the interacting pairs of drugs and target proteins, and to disappear in the other pairs. This section recalls methods mentioned in chapter 2 and describe how they are adapted in the context of drug target interaction prediction.

### 3.3.1 Sparse canonical correspondence analysis (SCCA)

Suppose that we have a set of $n_x$ drugs with $p$ substructure features, a set of $n_y$ target proteins with $q$ domain features, and information about interactions between the drug set and the target protein set. Note that $n_x \neq n_y$. Each drug is represented by a $p$-dimensional feature vector $\mathbf{x} = (x_1, \cdots, x_p)^T$, and each target protein is represented by a $q$-dimensional feature vector $\mathbf{y} = (y_1, \cdots, y_q)^T$.

Consider two linear combinations for drugs substructure and proteins domains as $u_i = \boldsymbol{\alpha}^T \mathbf{x}_i$ $(i = 1, 2, \cdots, n_x)$ and $v_j = \boldsymbol{\beta}^T \mathbf{y}_j$ $(j = 1, 2, \cdots, n_y)$, respectively, where $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_p)^T$ and $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_q)^T$ are weight vectors. Define the $n_x \times n_y$ adjacency matrix $A$, where element $(A)_{ij}$ is equal to 1 (resp. 0) if drug $\mathbf{x}_i$ and protein $\mathbf{y}_j$ are interact (resp. do not interact). Let $X$ be the $n_x \times p$ matrix defined as $X = [\mathbf{x}_1, \cdots, \mathbf{x}_{n_x}]^T$, and let $Y$ denote the $n_y \times q$ matrix defined as $Y = \left[\mathbf{y}_1, \cdots, \mathbf{y}_{n_y}\right]^T$, where the columns of $X$ and $Y$ are assumed to be centered and scaled. As described in chapter 2, SCCA consists in finding weight vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ which solve the following $L_1$ constrained optimization problem:

$$\begin{aligned} &\max\{\boldsymbol{\alpha}^T X^T A Y \boldsymbol{\beta}\} \quad \text{subject to} \\ &||\boldsymbol{\alpha}||_2^2 \leq 1, \quad ||\boldsymbol{\beta}||_2^2 \leq 1, \quad ||\boldsymbol{\alpha}||_1 \leq c_1\sqrt{p}, \quad ||\boldsymbol{\beta}||_1 \leq c_2\sqrt{q}, \end{aligned} \tag{3.1}$$

where $||\cdot||_1$ is $L_1$ norm (the sum of absolute values of vector entries), $c_1$ and $c_2$ are parameters to control the sparsity and restricted to ranges $0 < c_1 \leq 1$ and $0 < c_2 \leq 1$, where $c_1 = c_2 = 1$ defines the original CCA (OCCA) without sparsity constraint and amounts to compute an SVD (see chapter 2).

Problem (3.1) can be regarded as the problem of penalized matrix decomposition of the matrix $Z = X^T AY$. As mentioned in chapter 2, we can use the penalized matrix decomposition (PMD) proposed by [129]. After $m$ iterations of the algorithm, we obtain $m$ pairs of weight vectors $\boldsymbol{\alpha}_1, \cdots, \boldsymbol{\alpha}_m$ and $\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_m$ which are referred to as components. Components of lower $k$ are called "lower order components", while components of higher $k$ are called "higher order components". High scoring substructures and domains in the weight vectors are considered important in terms of drug-target interactions.

The originality of the SCCA method lies in the development of a sparse version of canonical correspondence analysis to handle the heterogeneous objects and their co-occurence information. It is therefore impossible to directly apply the canonical correlation analysis or its sparse version [93, 124, 129] in the question addressed here.

### 3.3.2 Evaluation of extracted components by reconstruction of drug-target interactions

If the extracted components are biologically meaningful, their use to reconstruct known drug-target interactions should lead to good prediction accuracies. Given a pair of compound $\mathbf{x}$ and protein $\mathbf{y}$, their potential interaction can be estimated based on the chemical substructures present in $\mathbf{x}$, the protein domains present in $\mathbf{y}$, their presence in common extracted components, and their distribution over all the canonical components. We use the following prediction score described in chapter 2. For any given pair of compound $\mathbf{x}$ and protein $\mathbf{y}$:

$$s(\mathbf{x}, \mathbf{y}) = \sum_{k=1}^{m} u_k \rho_k v_k = \sum_{k=1}^{m} \mathbf{x}^T \boldsymbol{\alpha}_k \rho_k \boldsymbol{\beta}_k^T \mathbf{y}, \tag{3.2}$$

where $m$ is the number of canonical components and $\rho_k$ is the $k$-th singular value. If $s(\mathbf{x}, \mathbf{y})$ is higher than a threshold, compound $\mathbf{x}$ and protein $\mathbf{y}$ are predicted to interact with each other.

We perform the following 5-fold cross-validation to evaluate the reconstruction ability. 1) We split drugs and target proteins in the gold standard set into five subsets of roughly equal sizes, and take each subset in turn as a test set. 2) We perform the training of CCA model on the remaining four sets (i.e. we extracted canonical components based on the remaining four sets). 3) We compute the above prediction score for the test set, based on the components extracted from the training set. 4) Finally, we evaluate the prediction accuracy over the five folds.

### 3.3.3 Supervised methods to reconstruct drug-target interactions

Pairewise Support Vector Machine has shown state-of-the-art performances on such supervised tasks. The $L1$-penalized linear SVM is an instance of this specific approach which associates pairs of features that have a highly predictive power. Nearest neighbour is used as a baseline. For all methods, we used the same representations for proteins and ligands, i.e. the feature vectors described in the Materials section 3.2.

#### Pairwise support vector machine (P-SVM and $L1$-PLSVM)

The SVM is a well-known binary classifier, and it is becoming a popular classification method in bioinformatics and chemoinformatics because of its high-performance prediction ability [105]. The use of SVM with pairwise kernels have been proposed to predict new compound-protein interactions [86, 33, 59], which is referred to as pairwise SVM (P-SVM). The pairwise SVM approach reduces the task of predicting interactions to a binary classification task. We consider a training set of drug target pairs $(\mathbf{x}_i, \mathbf{y}_i, \mathbf{i}_i)_{i=1...n}$ where $\mathbf{x}_i$ is a drug, $\mathbf{y}_i$ is a protein represented by their $p$ and $q$ dimensional feature vectors respectively and $\mathbf{i}_i$ is a class variable corresponding to the interaction of drug $\mathbf{x}_i$ and protein $\mathbf{y}_i$. Given $K_d$ and $K_p$, positive definite kernels

for drugs and proteins respectively, the function:

$$K((\mathbf{x}_1, \mathbf{y}_1), (\mathbf{x}_2, \mathbf{y}_2)) = K_d(\mathbf{x}_1, \mathbf{x}_2) K_p(\mathbf{y}_1, \mathbf{y}_2)$$

is a positive definite kernel on pairs of drug protein pairs. Standard supervise learning methods can be applied to discriminate between interacting and non interacting pairs. Applying an SVM to this problem using such a pairwise kernel defines the pairwise SVM. We tested several kernel functions such as linear kernel, Gaussian RBF kernel with various width parameters, polynomial kernel with various degree parameters for drug substructure profiles and protein domain profiles, and the regularization parameter. Those parameters are chosen by cross validation and the corresponding best choices are reported in the result section.

When considering linear kernels, *i.e.* standard scalar product, the proposed kernel between pairs implicitly corresponds to the scalar product between tensor product representations of the protein target pairs. Indeed, in this case:

$$K((\mathbf{x}_1, \mathbf{y}_1), (\mathbf{x}_2, \mathbf{y}_2)) = \mathbf{x}_1^T \mathbf{x}_2 \mathbf{y}_2^T \mathbf{y}_1 = \mathrm{Tr}((\mathbf{x}_1 \mathbf{y}_1^T)^T \mathbf{x}_2 \mathbf{y}_2^T)$$

where Tr denotes the trace operator for matrices. In this case $K$ is the scalar product between the matrices $\mathbf{x}_1 \mathbf{y}_1^T$ and $\mathbf{x}_2 \mathbf{y}_2^T$ which represent both pairs of protein targets. They are binary matrices which represent the presence or absence of all possible (substructure, domain) pairs. This representation is very high dimensional ($663 \times 876$) but also very sparse. Solving a linear SVM with such a representation amounts to find a weight matrix $w$ of the same size that minimizes the empirical loss $L_{SVM}$ (see chapter 2 for details). By adding an $L_1$ sparsity inducing term, we enforce sparsity of the corresponding weights. This is referred to as the $L1$-PLSVM which explicitly compute the tensor product representation behind the pairwise SVM and solves an $L1$ penalized empirical risk minimization problem. As mentioned, the problem is very high dimensional and very sparse. Specific libraries have been designed to take advantage of such a structure [57] which allows to estimate a sparse matrix $w$ using this method in a reasonable time. The resulting estimate represents all possible all possible (substructure, domain) pairs. Because we enforce sparsity, only a few of

these entries are non zero. Thus this extension of the pairwise SVM allows to extract association between chemical substructures and protein domains. However, these pairs are not structured in canonical components as in the SCCA case.

#### Nearest neighbour (NN)

The classical nearest neighbour (NN) method is often used in molecular screening. The proteins potentially interacting with a newly given compound $\mathbf{x}$ are determined as those that interact with the most similar compound in the training set. Given a new compound $\mathbf{x}$, we find $\mathbf{x}'$, its nearest neighbour in the training data according to the fingerprint profile similarity and predict the proteins interacting with $\mathbf{x}'$ to interact with $\mathbf{x}$ with a score reflecting the similarity between $\mathbf{x}$ and $\mathbf{x}'$. Likewise, potential ligands for a newly given protein $\mathbf{y}$ are determined as those that bind to the most similar protein in the training set. Given a new protein $\mathbf{y}$, we find $\mathbf{y}'$, its nearest neighbour in the training data according to the domain profile similarity and predict the molecules interacting with $\mathbf{y}'$ to interact with $\mathbf{y}$ with a score reflecting the similarity between $\mathbf{y}$ and $\mathbf{y}'$. The cosine correlation coefficient is used as a similarity measure for both compounds and proteins.

## 3.4 Results

### 3.4.1 Performance evaluation for the SCCA method

In general, it is difficult to evaluate the performance of an unsupervised feature extraction method in a direct manner. However, if the extracted sets of chemical substructures and proteins domains (the components) are biologically meaningful and capture relevant information with respect to protein-ligand interactions, one would expect that they present good generalization properties. This can be evaluated by testing the ability of the method to reconstruct known drug-target interactions, using the prediction score and the five fold cross-validation scheme described in section 3.3.

We evaluated the performance of the method by the ROC (receiver operating characteristic) curve [48], which is the plot of true positives as a function of false

Table 3.1: Performance evaluation on drug-target interactions reconstruction by 5-fold cross-validation.

| | NN | P-SVM | $L1$-PLSVM | OCCA | SCCA |
|---|---|---|---|---|---|
| AUC | 0.5892 | 0.7504 | 0.7061 | 0.7377 | 0.7497 |
| S.D. | 0.0042 | 0.0064 | 0.0015 | 0.0046 | 0.0057 |

positives based on various thresholds, where true positives are correctly predicted interactions and false positives are predicted interactions that are not present in the gold standard interactions. We summarized the performance by an AUC (area under the ROC curve) score, which is 1 for a perfect inference and 0.5 for a random inference. We repeated the cross-validation experiment five times, and computed the average of the AUC scores over the five cross-validation folds, varying the three parameters of the method. The best results were obtained with $c_1 = 0.1$, $c_2 = 0.2$ for the sparsity parameters, and with $m = 50$ for the number of components in the case of SCCA. The same experiments were repeated for OCCA which has only one parameter, and the best results were obtained for $m = 50$.

The AUC scores for SCCA and OCCA are 0.7497 and 0.7377, respectively. These statistics are summarized in table 3.1. This result shows that both methods perform much better than a random inference, whose AUC score is equal to 0.5. Consequently, this indicates that the proposed prediction score allows to enlighten the good generalization properties of extracted SCCA or OCCA components. Their performance comparison with other methods will be discussed in a later subsection.

Next, we applied SCCA and OCCA on the complete gold standard dataset described in the Materials section, and analyzed the extracted components of drug chemical substructures and protein domains. We used the parameters leading to the best results in the cross-validation experiment.

We examined the resulting weight vectors for drug chemical substructures and protein domains in applying OCCA and SCCA. Figure 3.2 shows the index-plot of weight vectors in applying OCCA, while figure 4.3 shows the index-plot of weight vectors in applying SCCA, where the first six canonical components are shown in

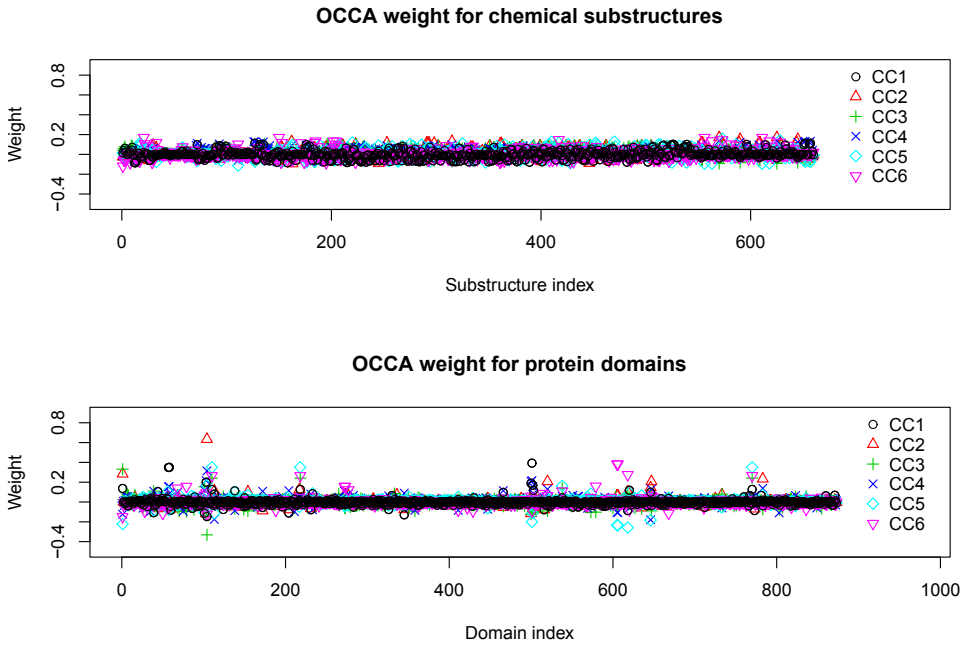



Figure 3.2: Index-plot of weight vectors for drug substructures and protein domains for OCCA. Horizontal axes indicate the index of chemical substructures (upper) or protein domains (bottom), and vertical axes indicate the weight values on the chemical substructures (upper) or protein domains (bottom).

both cases. It seems that almost all elements in the weight vectors in OCCA are non-zero and highly variable, while most of the elements in the weight vectors in SCCA are zero in each component, implying that SCCA can select a very small number of features as informative drug substructures and protein domains.

These results suggest that, although the performance in reconstruction of drug-target interactions of SCCA and OCCA were close, the proposed SCCA provides us with more selective drug substructures and protein domains, without missing important information encoding protein-ligand interaction. In practice, we found that it is very difficult to analyze the extracted components when there are too many high or low scoring weight elements like in OCCA. On the contrary, the advantage of SCCA over OCCA is that it is possible to derive biological interpretations, as shown on a few examples in the next subsection.

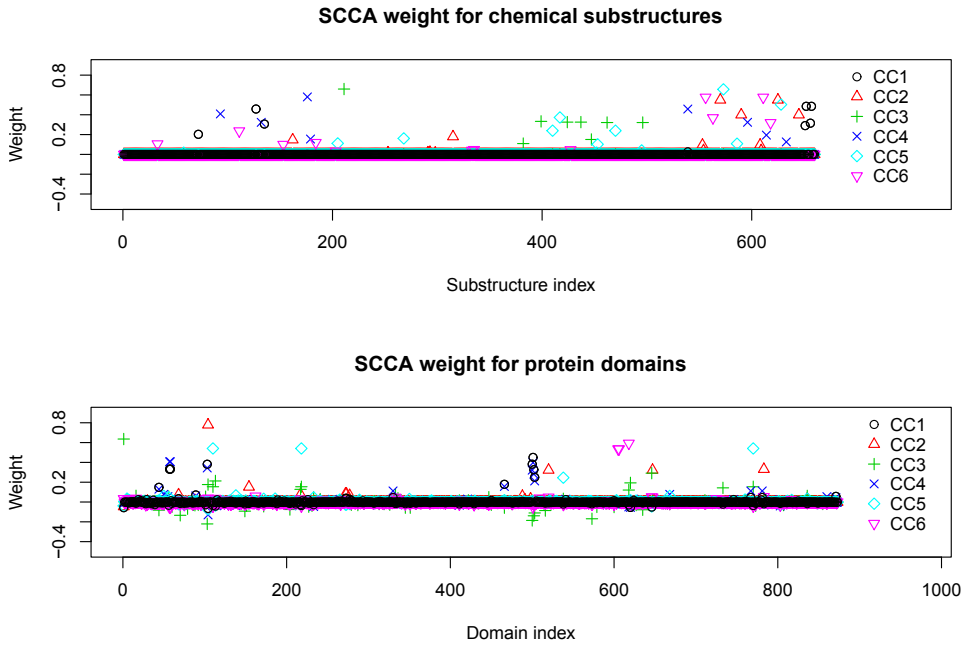



Figure 3.3: Index-plot of weight vectors for drug substructures and protein domains for SCCA. Horizontal axes indicate the index of chemical substructures (upper) or protein domains (bottom), and vertical axes indicate the weight values on the chemical substructures (upper) or protein domains (bottom).

### 3.4.2 Comparison with other supervised methods

If the proposed method captures important features that govern protein-ligand interactions, and if the proposed prediction score is relevant, the performance should be at least as good as those of other methods for predicting protein-ligand interactions, using the same vector descriptions for proteins and ligands.

We performed the same five-fold cross validation experiments for the three other considered prediction methods NN, P-SVM and $L1$-PLSVM on the same protein-ligand dataset, as we did for SCCA and OCCA. The best performance was obtained using the polynomial kernel with degree parameter $d = 3$ and the regularization parameter $C = 1$ in the case of P-SVM .

Table 3.1 shows that SCCA, OCCA and $L1$-PLSVM outperform the baseline, i.e. the NN. Furthermore, the performance of SCCA is similar to that of P-SVM

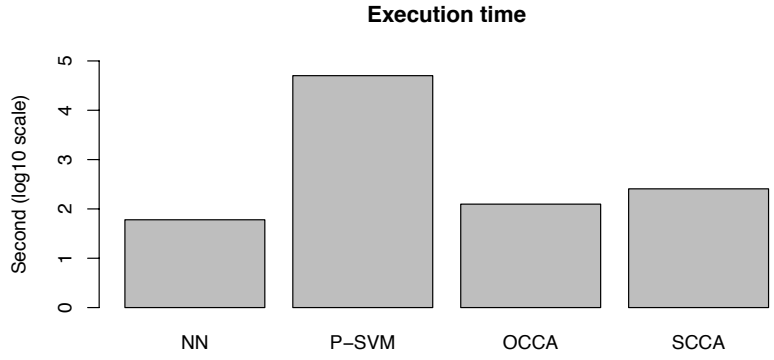



Figure 3.4: Execution time for different methods on a log scale. The pairwise SVM is the most computationally intensive method by at least one order of magnitude.

used as the state-of-the-art prediction method. Performances of $L1$-PLSVM are a bit lower but still much better than NN. These results show that the extracted canonical components contain valuable biological information and underline the interest of the proposed method as a tool for analyzing protein-ligand interactions. In addition, it should be pointed out that P-SVM and NN do not provide any biological interpretation since they only predict interactions, and they do not extract any information about important molecular features for these interactions.

We also investigated the computational cost for each method. Figure 3.4 shows the total execution time of the cross-validation experiment between the four different methods. NN is the fastest, followed by OCCA, SCCA, and P-SVM. As expected, P-SVM is much slower than the other methods, because the complexity of the "learning" phase scales with the *square* of the "number of training compounds *times* the number of training proteins", leading to prohibitive computational difficulties for large-scale problems. These results suggest that SCCA constitutes a good trade-off between prediction accuracy, biological interpretation, and computational efficiency.

### 3.4.3 Biological interpretation of extracted drug substructures and protein domains using SCCA

The SCCA method provides 50 canonical components. The output of the method is a list of canonical components (CCs), each of which contains correlated chemical substructures and protein domains, and a list of proteins and drugs that contributed to extract the chemical substructures and protein domains. All components present a limited number of high scoring chemical substructures and protein domains, which is a consequence of the sparsity of the method. It extracts domains and substructures that summarize the most relevant and consistent information. This allows meaningful analysis of the data for biological interpretation. Table 3.2 shows some examples of extracted chemical substructures (SMILE-like format in PubChem) and protein domains (PFAM IDs) in the first four CCs (CC1, CC2, CC3 and CC4).

We examined the extracted drug substructures and protein domains from biological viewpoints. The results for a few canonical components will be discussed. Analysis of the results shows that the components contain a limited number high scoring protein domains that usually belong to one, or a small number of protein families. For example, most high scoring protein domains of component CC1 belong to nuclear receptors (PF02159: Oestrogen receptor, PF02155: Glucocorticoid receptor, PF00104: Ligand-binding domain of nuclear hormone receptor, PF02161: Progesterone receptor, PF02166: Androgen receptor, PF00105 zinc finger c4 type). Consistent with this observation, the high scoring substructures are typical fragments found in steroids, and the high scoring drugs are steroid-like molecules. The domains from nuclear receptors also appear with high scores in a few other components such as CC4 or CC12. However, these components do not share any of their high scoring chemical substructures, which shows that they are not redundant. We observed that the absence of redundancy between components is a general feature of the method.

Unexpectedly, the annexin domain PF00191 is also present in the top ranked domains of CC1. Annexins are membrane associated proteins that bind phospholipids, inhibit the activity of phospholipase A2, and play a role in the inflammatory response. Annexins and nuclear receptors are evolutionary unrelated proteins with no sequence

Table 3.2: Examples of results for canonical components 1, 2, 3 and 4: high scoring domains, substructures, proteins and drugs

| | |
|---|---|
| Domain | PF02159 (Oestrogen receptor); PF02155 (Glucocorticoid receptor); PF00191 (Annexin); ... |
| Structure | CC1CC(O)CC1; CC1C(O)CCC1; saturated or aromatic carbon-only ring size 9; CC1C(C)CCC1; ... |
| Protein | ESR1_HUMAN (Estrogen receptor); GCR_HUMAN (Glucocorticoid receptor); ... |
| Drug | DB00443 (Betamethasone); DB00823 (Ethynodiol Diacetate); DB00663 (Flumethasone Pivalate)); ... |
| Domain | PF00194 (Carbonic anhydrase); PF08403; PF02254; PF03493 (potassium channel);... |
| Structure | SC1CC(S)CCC1; Sc1cc(S)ccc1; Sc1c(Cl)cccc1; SC1C(Cl)CCCC1; N-S-C:C; N-S; ... |
| Protein | KCMA1_HUMAN (Calcium-activated potassium channel); CAH12_HUMAN (Carbonic anhydrase 12); ... |
| Drug | DB00562 (benzthiazide); DB00232 (Methyclothiazide); DB01324 (Polythiazide); ... |
| Domain | PPF00001 (transmembrane receptor); PF03491 (Serotonin neurotransmitter transporter); ... |
| Structure | C( H)(:C)(:C); C:C-C-C; C-C-C-C:C; C:C-C-C-C; C-C:C-C-C; C-C-C:C-C; ... |
| Protein | TOP2A_HUMAN (DNA topoisomerase); SC6A4_HUMAN (Sodium-dependent serotonin transporter); ... |
| Drug | DB01654 (Thiorphan); DB00743 (gadobenic acid); DB03788 (GC-24); ... |
| Domain | PF00105 (Zinc finger); PF00104; PF02159 (Oestrogen receptor); PF00191 (Annexin); ... |
| Structure | C( C)( C)( C)( C); C-C(C)(C)-C-C; unsaturated non-aromatic carbon-only ring size 6; ... |
| Protein | ESR1_HUMAN (Estrogen receptor); GCR_HUMAN (Glucocorticoid receptor); ... |
| Drug | DB00596 (halobetasol); DB01234 (Dexamethasone); DB00620 (Triamcinolone); ... |

or function similarities. However, annexins and nuclear receptors probably present similar ligand binding pockets in the 3D space, which could not be foreseen from comparison of their primary sequences, and both types of proteins can bind steroid-like ligands. Therefore, the method associated these protein domains in CC1. Some of these steroids ligands are common to both types of proteins. For example DB00443 and DB00663 (respectively PubChem IDs 9782 and 443980) bind to glucocorticoid

receptor and to annexin A1. On the contrary, some steroids only bind to a nuclear receptor and not to annexin. This observation suggests that the method might offer a tool to tackle the important question of specificity

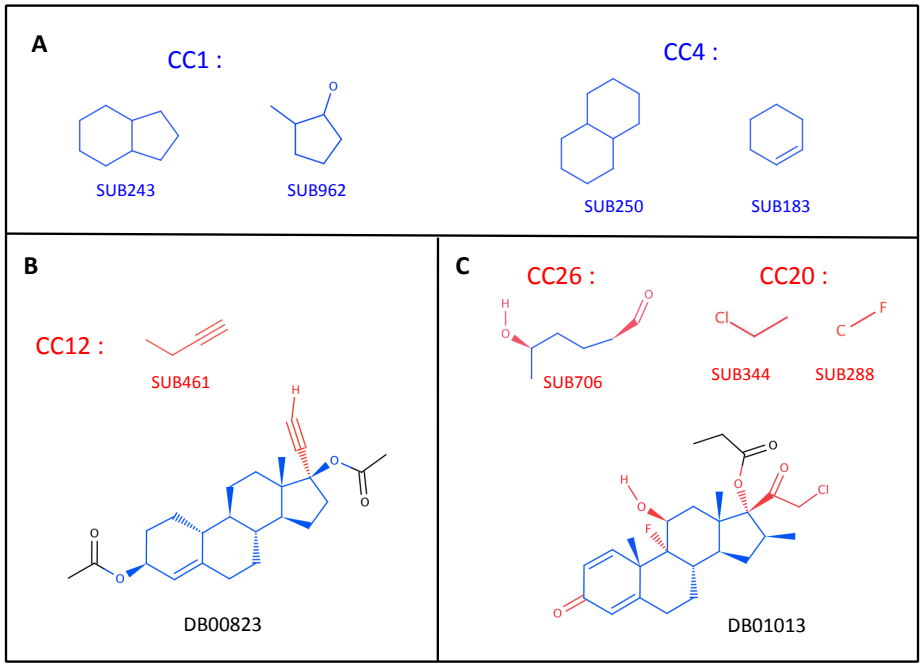



Figure 3.5: (A) Examples of high scoring substructures from components CC1 and CC4 belonging to group A. (B) In blue, part of the molecular structure of DB00823 that can be built using high scoring substructures of components of group A. In red, part of DB00823 that can be built using high scoring substructures of components of group B (specific of estrogen receptor). (C) In blue, part of the molecular structure of DB01013 that can be built using high scoring substructures of components of group A. In red, part of DB01013 that can be built using high scoring substructures of components of group C (specific of annexin). In the case of SUB706, only chemical groups that cannot be built using substructures of group A are colored in red.

To illustrate this point, we will consider the example of the estrogen receptor ESR1_HUMAN (UniProt ID: P03372, Pfam IDs PF00104, PF00105, PF02159) and of annexin A1 ANXA1_HUMAN (UniProt ID: P04083, Pfam ID PF00191). Domains of

these two proteins have high weights in a few common components (CC1, CC4, CC13, CC46, called group A), while only domains of estrogen receptors have high scores in components CC12, CC15, CC29, CC34, CC38 (called group B), and only those of annexin have high scores in components CC19, CC20, CC21, CC26, CC40 (called group C). We will show in the case of drug DB00823 (or PubChem ID 9270) that binds the estrogen receptor but not annexin, and of DB01013 (or PubCHem ID 32798) that binds to annexin but not to the estrogen receptor, how analysis of the substructures belonging to groups A, B and C can be used to explain the specificity of these two drugs. The parts of the chemical structure of DB00823 and DB01013 that can be built using high scoring substructures belonging to group A (components common to estrogen receptor and annexin) is colored in blue in figure 3.5. They correspond to the main steroid scaffolds of these two molecules, as expected for proteins sharing similar types of ligands. However, additional chemical structures of the DB00823 molecule, colored in red in figure 3.5, can only be built by using high scoring substructures found in components of group B, where only estrogen receptor domains have high scores. Similarly, additional chemical structures of DB01013, colored in red in figure 3.5, can only be built using high scoring substructures found in components of group C, where only annexin domains have high scrores. Note that none of the high scoring substructures of components specific of estrogen receptor (group B) are present in DB01013 that only bind annexin, and that reciprocally, none of the high scoring substructures of components specific of annexin (group C) are present in DB00823 that only bind estrogen receptor. In other words, the method allows to highlight the parts of the molecules that encode for their specificity to bind only to estrogen receptor, or only to annexin.

One additional comment should be made: all known annexin ligands are steroids, while estrogen receptor domains also bind other types of molecules such as Tamoxifen (DB00675, PubChem ID 2733526) or other similar molecules such as Raloxifen. As shown in figure 3.6, these molecules are very different from steroids. They lead to the CC34 and CC38 above mentioned components, that have a high score for estrogen receptor domains and not for annexin domains because the latter do not bind these molecules. In figure 3.6, the part of the Tamoxifen molecule that can be built using

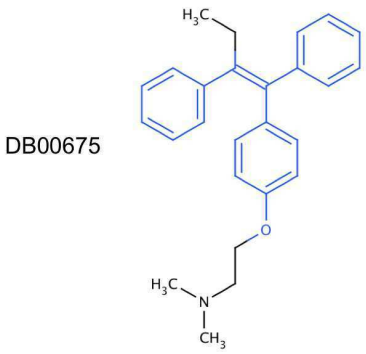



Figure 3.6: Chemical structure of Tamoxifen (DB00675). Parts of the molecule that can be built from chemical substructures of components CC34 and CC38 are colored in blue.

high scoring substructures from CC34 and CC38 is colored in blue.

We will comment more briefly on component CC2, in order to show that the above observations also apply to other components and families of proteins. Component CC2 contains the carbonic anhydrase domain PF00194 , which belongs to zinc metalloenzymes catalyzing reversible hydration of carbon dioxide to bicarbonate. It also contains calcium-dependent potassium channel domains (PF03493, PF02254). Carbonic anhydrase inhibitors are used as anti-glaucoma agents, diuretics and anti-epileptics. Interestingly, the human potassium channel KCMA1_HUMAN (UnitProt ID: Q12791), one of the high scoring proteins in CC2, is also known to be involved in epilepsy. Domains of the carbonic anhydrase and of the calcium-dependent potassium channel also appear together with high scores in a few other components (namely CC7, CC16, CC22, CC27 and CC43), whereas components CC3 and CC25 are specific of carbonic anhydrase, and component CC20 is specific of calcium-dependent potassium channel.

Although different types of drugs are known to bind human calcium-dependent potassium channel and carbonic anhydrase proteins, these two proteins share drugs from the thiazide family. Figure 3.7 shows the general scaffold of thiazide molecules.

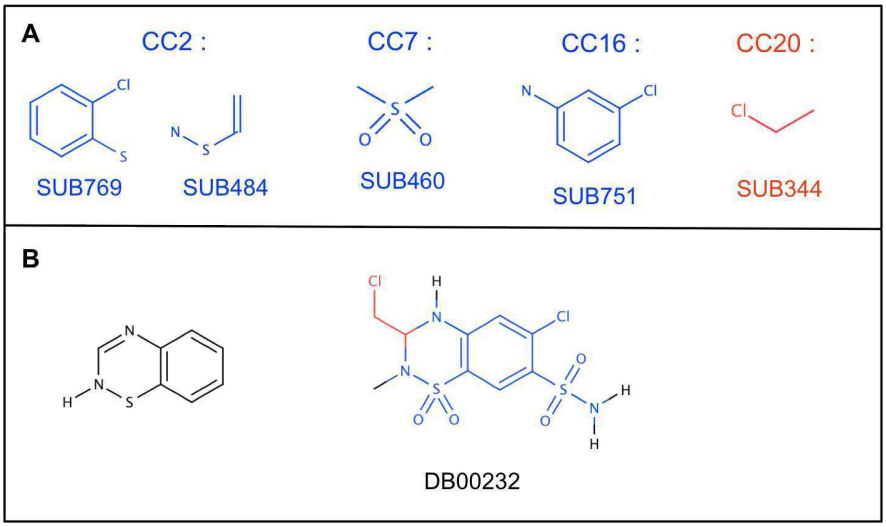



Figure 3.7: (A) In blue, examples of high scoring substructures of components common to calcium-dependent potassium channel and carbonic anhydrase proteins. In red, example of a high scoring substructure from component CC20 specific to calcium-dependent potassium channel. (B) On the left, in black, the basic thiazide scaffold. On the right, in blue, part of the molecular structure of DB00232 that can be built using high scoring substructures from components common to calcium-dependent potassium channel and carbonic anhydrase proteins. In red, part of DB00232 that can be built using high scoring substructures of components CC20, specific of calcium-dependent potassium channel.

All known thiazide ligands of carbonic anhydrase also bind calcium-dependent potassium channel (for example DB00436 or DB00562, among others). However, the thiazide molecule DB00232 (PubChem ID 4121) only binds to KCMA1_HUMAN, the human calcium-dependent potassium channel and not to human carbonic anhydrase. As in the case of annexins and nuclear receptors, although carbonic anhydrase and potassium channel present no sequence similarity, they share similar ligand binding pockets, and are able to bind similar molecules. Therefore, the method associates them in CC2 and in a few other common components, namely CC7, CC16, CC22, CC27, CC43. In figure 3.7, the part of the DB00232 molecule that can be built using substructures of these components is colored in blue. However, the calcium-dependent

potassium channel domains have high scores in component CC20, but this is not the case of carbonic anhydrase. In figure 3.7, the part of DB00232 that can only be built using substructures of CC20 is colored in red. As in the case of estrogen receptor and annexin, the method allows to highlight the parts of the DB00232 that encode for its specificity to bind to calcium-dependent potassium channel and not to carbonic anhydrase.

Finally, we would like mention that component CC20 appears in the two cases discussed above, because the presence of SUB344 in steroid or thiazide molecules happens to modulate their specificity, respectively for annexin or calcium-dependent channel. The fact that a molecule contains substructure SUB344 does not necessarily mean that it will bind to all high scoring proteins of CC20. Indeed, more generally, the protein binding profile of a molecule depends on its complete substructure profile which is not limited to an individual substructure.

### 3.4.4 Comments on $L1$-PLSVM method

#### Performances

The cross validation results presented in the previous paragraphs correspond to a block wise split of the training and test set (see figure 3.8 for a graphical illustration). This choice of validation is closer to reality than choosing the training pairs uniformly. Indeed, in field applications, one is interested by finding candidate drugs for a given protein or by finding targets for a given drug. However, this corresponds to a split that is not uniform over all possible pairs but follows the structure of the matrix and supervised learning method suffer from this bias. Indeed, the underlying assumption behind the supervised learning framework is that the training examples are taken uniformly at random from the space of possible examples (see chapter 2 for a brief overview). Table 3.3 shows the cross validation results in a pairwise setting. In this setting, the performance of the $L1$-PLSVM method are close to those of the P-SVM method and significantly higher than those of SCCA. In such a setting, if the interest is in reconstruction performances, such binary classification approaches should be preferred.

Table 3.3: Performance evaluation on drug-target interaction reconstruction by 5-fold cross-validation in a pairwise setting.

| | $L1$-P-LSVM | P-SVM | SCCA |
|---|---|---|---|
| AUC | 0.8301 | 0.8339 | 0.7975 |
| S.D. | 0.0006 | 0.0005 | 0.0018 |

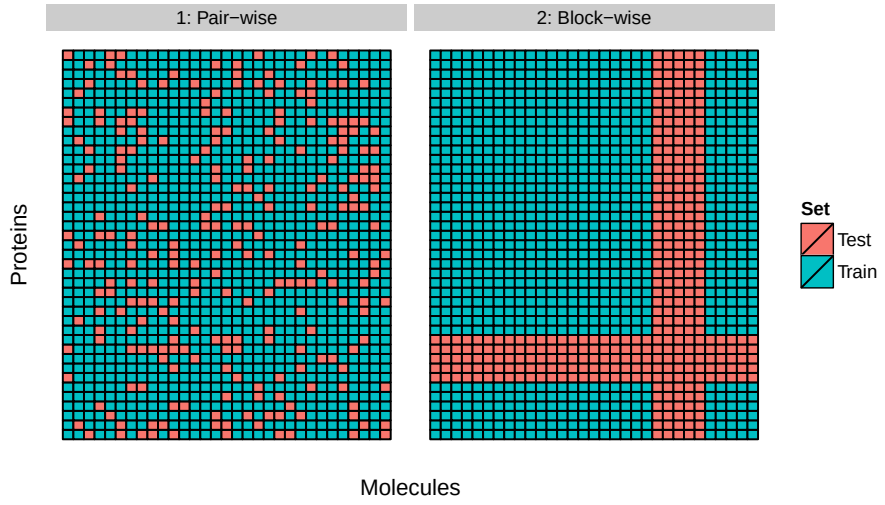



Figure 3.8: Illustration of pair-wise and block-wise cross validation. In the former case, randomly selected pairs are used as a training set while in the latter, the training and test sets reflect the row and column structure of the interaction matrix.

### Extracted substructure domain pairs

Extracted pairs of substructure and domains using the $L1$-PLSVM method are consistent with the underlying biology. This means that they are found in molecules and proteins that indeed interact. However, these associations are not organized in components as for the SCCA method. It is therefore more difficult to discuss deeper examples of associations between chemical substructures and protein domains using this method. It should be noted that the $L1$-PLSVM method extracts a smaller number of associated substructures and domains compared to SCCA (around two order

of magnitude smaller).

It appears that the interest of the two methods depend on the focus point (supervised reconstruction or unsupervised feature extraction) and the learning setting (pair wise or block wise). Unsupervised feature extraction abilities of SCCA are of greater interest. However it does not perform as well on reconstruction task in the pair wise supervised setting. The results presented here shed light on which method should be preferred depending on those parameters.

## 3.5 Discussion and Conclusion

In this chapter we described methods to extract drug chemical substructures and protein domains that govern drug-target interactions. The methods use known protein-ligand interactions as a learning dataset to extract ligand substructures and their associated protein domains, and importantly, they do not require information about proteins 3D structures. From a system point of view, they provide integrative analysis tools to study interactions between chemical and genomic spaces in a unified framework. Since the methods can handle learning datasets containing many proteins and molecules they constitute a contribution to the development of system based approaches for protein-ligand interaction.

Quantitative structure-activity relationship (QSAR) methods are similar in spirit, but they are designed handle a single protein at a time. The approach consists in using molecular similarities to predict activity of new molecule against a given target based on classification or regression methods. Such an approach cannot take advantage of emerging information from large protein target interaction datasets at a genome-wide level. Similar comments can be made regarding docking strategies which rely on known 3D structure of proteins. Moreover, even for proteins of known 3D structures, it is extremely difficult to automate the set-up of these methods for many different binding sites, leading to docking and scoring inaccuracies when they are used on large scale [71].

Prediction of all protein targets for a molecule is the goal of chemogenomics. Various studies report algorithms that implement such methods, and up to now, they

have restrained the search of off-targets within a given family of proteins such as GPCRs [126, 58, 135]. Nonetheless, they demonstrate the advantage of the approach over QSAR like methods. The proposed method is a new contribution to the field of chemogenomics which relies on learning databases of known protein-ligand interactions and which can take into account evolutionary unrelated proteins.

### 3.5.1 Applications of SCCA in drug development

The method could be of interest in various ways in the drug development process. First, given protein target of therapeutic interest, one can identify the components into which this protein domains are found with high scores. Then, one can build a ligand for this target protein using high scoring substructures of these components, potentially with the help of other recent developments in the domain of fragment-based drug discovery [31]. For example, in the Results section, we showed for several drugs (DB00823, DB01013, DB00675, DB00232), that one could build their molecular structure using high scoring substructures of components in which their protein targets have a high score.

Second, for a given drug that binds to a protein target of interest, the method can help to identify off-target proteins: protein domains that are found with high scores in the same components as these of the protein of interest are potential off-targets. Trivial off-targets are proteins that share high sequence similarity with the target protein, and are otherwise easy to identify using classical algorithms such as BLAST [3]. However, two unrelated proteins that underwent convergent evolution may present similar pockets in the 3D space, allowing binding of similar ligands, although they may share no significant sequence similarity. The proposed method can handle such cases by learning ligand similarities from a database. Examples are shown in the Result section for estrogen receptor and annexin, or for calcium-dependent potassium channel and carbonic anhydrase. Various methods have been developed to predict protein-ligand interactions, but predicting off-targets for a drug has been much less studied. The use of drug side-effect similarity has been proposed to identify potential off-targets, but the method can therefore only be used for molecules

of known side-effect profile, i.e. mainly for marketed drugs [20].

Finally, the method can also help to tackle the problem of drug specificity, which is in fact related to the topic of off-target identification. Here, we are interested in a given drug developed against a given target. The drug happens to also bind some off-target proteins. The method could help to optimize the structure of this drug. The principle would be to add chemical substructures that have high scores in components where only the target protein has a high score, and not the off-target proteins. As shown in section 3.4 for estrogen receptor and annexin, drugs that bind only one of these proteins contain substructures present in components where only one of these two proteins has a high score. The same situation was shown for DB00232 that only bind calcium-dependent potassium channel, and not carbonic anhydrase. As shown in section 3.4, the proposed method allows to identify such cases: non trivial off-targets are expected to appear in the same components as the main target. Among several drug candidates, the method could help to eliminate molecules with too many potential off-target interactions, or with potential off-target expected to lead to severe side-effects. On the contrary, the drug candidates whose chemical substructures are not found in canonical components that do not contain its targeted protein domain are expected to be of greater interest.

From technical viewpoints, there are several limitations on the proposed SCCA method. One main difficulty of using SCCA is to choose appropriate sparsity parameters and appropriate number of components. High sparsity promoting parameters would lead to an over-sparse model in all the cases, which might be misleading in the interpretation if the degree of sparsity was not tuned carefully. According to a cross-validation, we used top 50 components, but other components may contain biologically meaningful information. The definition of an appropriate objective function to be maximized or minimized in the cross-validation is an important issue. There remains much room to develop a more appropriate way to choose the parameters. Another pitfall of SCCA is that it might not work well when sparsity is not a relevant characteristic arising from the data. For example, it cannot deal with hierarchically correlated features in the descriptors of drugs or proteins. An extension of the method would be to additionally use other constraints which can deal with such a hierarchical

effect such as the ones presented in [61].

# Chapter 4

# Phenotypic response to molecular perturbations

## Résumé

Caractériser la réponse phénotypique à un stimulus est primordial pour la compréhension du comportement d'organismes pluri-cellulaires. Parmis les applications majeures, on peut citer la mise au point de thérapies moléculaires efficaces. Cependant, les phénotypes sont bien plus complexes que les interactions moléculaires, comme celles considérées dans le chapitre 3. En effet, il existe de multiples possibilités pour définir un phénotype qui implique souvent un grand nombre de cellules. Par ailleurs, la réponse à un stimulus peut varier de manière importante d'un individu à l'autre. Ce chapitre est donc dédié à l'étude de phénotypes à l'échelle d'organismes et à l'échelle de populations de cellules.

Dans une première partie, les effets secondaires de médicaments sont considérés comme des phénotypes macroscopiques causés par les interactions avec un ensemble de protéines. Les liens entre la structure chimique et les effets secondaires sont analysés avec une méthode d'analyse canonique des corrélations parcimonieuse. Cette approche permet d'analyser conjointement les relations entre structure chimique et effets secondaires. Des exemples de modulations d'effets secondaires par la structure chimique et des relations avec les profils de cibles protéiques sont donnés pour

illustrer les résultats produits par cette méthode. La performance prédictive de ces associations entre structure chimique et effets secondaires sont comparées à celles de méthodes à l'état de l'art.

Dans une seconde partie, la question de la comparaison de l'effet de siRNA sur des phénotypes de populations de cellules est explorée. Cette étude se base sur un jeu de données provenant d'une expérience de microscopie fluorescente à haut débit dans le contexte de l'étude de la maladie du sarcome d'Ewing. Des indices provenant de ces données sont proposés pour motiver le besoin de prendre en compte la variabilité intra-population de la réponse cellulaire et la structure de corrélation entre les descripteurs utilisés pour décrire chaque cellule individuellement. Un modèle probabiliste est proposé afin de prendre en compte ces considérations. Des expériences numériques montrent que le modèle proposé possède de meilleures propriétés que celles d'approches plus naïves pour décrire la réponse à un siRNA à l'échelle de populations de cellules.

# Abstract

Characterization of the phenotypic effect in response to a stimulus is a key problem in the understanding of multi-cellular organisms behaviour. This has great implications, for example in the development of efficient molecular therapies. However phenotypes are much more complex to describe and study than the molecular mechanisms considered in chapter 3. Indeed, there are multiple ways to define phenotypes. They usually involve more than a single cell and there might exist a great variability in phenotypic responses to stimuli between different individuals. This chapter is dedicated to the study of phenotypes at the organism level and at the cell population level.

In a first part, drug side effects are considered as macroscopic phenotypes induced by drugs through interaction with a set of proteins. The relationship between chemical structure and side effects is investigated using a sparse canonical correlation analysis method. This approach allows to jointly analyse relations between chemical structure and side effect profiles for drugs at a large scale. Examples of side effect modulation through chemical structure variations and relation with protein target profiles are

provided to illustrate the results provided by the method. Predictive performances of the extracted associations between structures side-effects are compared to these of state-of-the-art methods.

In a second part, the question of phenotypic comparison of siRNA effect on cell populations is investigated. This study is based on data arising from a high throughput fluorescent microscopy experiment in the context of studying Ewing sarcoma disease. Evidences arising from data are presented to support the need for taking into account both variability of the cell response within a population and the correlation structure of individual cell descriptors. A probabilistic model that takes these questions into consideration is proposed. Numerical experiments demonstrate that the model has better properties than those of more naive approaches to study phenotypic response to siRNA at the level of cell populations.

## 4.1 Introduction

This chapter is dedicated to the study of biological response to a stimulus in terms of phenotype. By stimulus, it is meant the exposition to perturbing compounds (marketed drug and siRNA in this chapter). A phenotype results from the effect of these stimuli on a living system (animal or population of cells). Therefore, the scale of interest here is at a higher hierarchical level compared to those considered in chapter 3. By nature, phenotypes are more complex to study, some reasons being that

- Phenotypes are result from a whole system of molecules interacting together.
- There is no universal definition of a phenotype and the boundary between two phenotypes might be fuzzy.
- Phenotypes are variable in the sense that two similar organisms exposed to the same stimulus might respond in different ways.

The consequences of these remarks are, first that it is necessary to consider a large number of interacting physical objects (molecules, proteins, cells ...) in order to understand how phenotypic effects emerge when a living organism is confronted

to a stimulus. Second, a phenotype definition should take into account response variability. The results presented in this chapter illustrate these ideas based on real world data.

As we mentioned, the concept of phenotype is broad. In this chapter, we consider two different scales. At the macroscopic level, a phenotype can be understood as a characteristic of a multi-cellular organism. The example treated in the first section of this chapter is related to marketed drug side-effects. These are compounds which have been available on the drug market (which may not be available any more) and for which side effect data have been collected. A side-effect of a drug can be thought of as a phenotype. Side-effects are due to interactions of a drug with off targets, *i.e.* proteins that do not constitute the main therapeutic target of the drug. They are observable at the macroscopic scale of a human being.

At the level of a cell, cell cycle phases can be understood as phenotypes. Those are particularly interesting in cancer therapies, the objective being to disrupt cell division processes which went out of control. When considering population of cells, even though cells have been exposed to the same stimulus, there is variability in cell responses. This variability is also part of the phenotypic response at the cell population level. These phenomena make it difficult to characterize and compare stimuli effects at the level of cell populations.

### 4.1.1 Content of the chapter

The first section of this chapter is dedicated to the analysis of correlations between drug chemical sub structures and their side-effects. A penalised matrix decomposition is used to investigate these correlations which are further used in side-effect prediction contexts. This is compared to state-of-the-art supervised methods. The results presented in this part are based on the work presented in [94] together with results from [136].

In a second section, we consider cell population phenotypes. Numerical experiments are based on control data from siRNA knock down high content screening experiments in the context of Ewing sarcoma. This study is part of a wider research

project that aims at identifying genes that play a role in Ewing sarcoma. A probabilistic model is proposed to tackle the issue of comparing images of cell populations with heterogeneous individual phenotypes. This section is based on the results presented in [95].

## 4.2 Sparse canonical correlation based drug side-effect analysis

This section is mostly based on the results presented in [94]. The focus is on the relation between adverse drug reactions and drug chemical structure. A matrix decomposition method is applied to extract associations between chemical substructures and drug side-effects which are further used in a supervised prediction context. [136] extended this work by integrating both chemical and protein target profiles in a side-effect prediction context. These results are qualitatively described in the last paragraph.

### 4.2.1 Background

#### Drug side-effects

Drug side-effects, or adverse drug reactions, have become a major public health concern. It is one of the main causes of failure in the process of drug development, and of drug withdrawal once they have reached the market. As an illustration of the extent of this problem, serious drug side-effects are estimated to be the fourth largest cause of death in the United States, resulting in 100,000 deaths per year [42]. In order to reduce these risks, many efforts have been devoted to relate severe side-effects to some specific genetic biomarkers. This so-called pharmacogenomics strategy is a rapidly developing field, especially in oncology [56]. The aim is to prescribe a drug to patients who will benefit from it, while avoiding life threatening side-effects [84].

From a system viewpoint, drugs can be regarded as molecules that induce perturbations to biological systems consisting of various molecular interactions such as protein-protein interactions, metabolic pathways and signal transduction pathways, leading to the observed side-effects [119]. The most common perturbation mechanism is to bind to a protein, thereby modifying its function. These mechanisms have been investigated in chapter 3. Actually, the body's response to a drug reflects not only the expected favorable effects due to the interaction with its target, but also integrates the overall impact of off-target interactions. Indeed, even if a drug has a strong affinity for

its target, it also often binds to other protein pockets with varying affinities, leading to potential side-effects. This concept has been illustrated by comparing pathways affected by toxic compounds and those affected by non-toxic compounds, establishing links between drug side-effects and biological pathways [103].

### *In silico* prediction

Although preclinical *in vitro* safety profiling can be used to predict side-effects by testing compounds with biochemical and cellular assays, experimental detection of drug side-effects remains very challenging in terms of cost and efficiency [127]. Therefore, *in silico* prediction of potential side-effects early in the drug discovery process, before reaching the clinical stages, is of great interest to improve this long and expensive process and to provide new efficient and safe therapies for patients. This task intrisically requires to consider the whole system of proteins in order to determine macroscopic consequences of disturbing biological processes at the molecular scale. Expert systems based on the knowledge of human experts have been developed to predict the toxicity of molecules based on the presence or absence of toxic moieties in their chemical structure. For example, they predict potential toxicity such as mutagenicity, but they do not provide prediction for numerous potential side-effects in human [12]. Recently, several computational methods for predicting side-effects have been proposed, which can be categorized into pathway-based approaches and chemical structure-based approaches.

The principle of pathway-based approaches is to relate drug side-effects to perturbed biological pathways or sub-pathways because these pathways involve proteins targeted by the drug. In a pioneer work to illustrate this concept, it has been shown that drugs with similar side-effects tend to share similar profiles of protein targets [19]. The authors further exploited this characteristic to predict missing drug targets for known drugs using side-effect similarity. [38] proposed a method for relating side-effects to cooperative pathways defined as sub-pathways sharing correlated modifications of gene expression profiles in presence of the drug of interest. However, this method requires gene expression data observed under chemical perturbation of the drug. [131] developed a method to identify off-targets for a drug by docking this

drug into proteins binding pocket similar to that of its primary target. The drug-protein interactions with the best docking scores are incorporated to known biological pathways, which allows to identify potential off-target binding networks for this drug. However, the performance of this method depends heavily on the availability of protein 3D structures and known biological pathways, which limits its large-scale applicability.

The principle of chemical structure-based approaches is to relate drug side-effects to their chemical structures. [102] developed a method that identifies chemical substructures associated to side-effects. However, this method does not provide an integrated framework to predict side-effects for any drug molecule. [135] proposed a method to predict pharmacological and side-effect information using chemical structures, which is then used to infer drug-target interactions. However, the method cannot be applied to predict high-dimensional side-effect profiles.

### Content of this section

This section describes a canonical correlation based approach (CCA) to predict potential side-effect profiles of drug candidate molecules based on their chemical structures, which is applicable on large molecular databanks. In addition, the proposed method is able to extract correlated sets of chemical substructures (or chemical fragments) and side-effects. Sparsity inducing constraints allows to ease the interpretation of the extracted correlated sets. The corresponding method is described in chapter 2 and is referred to as sparse canonical correlation analysis (SCCA).

Numerical experiments show the usefulness of the proposed method on the prediction of 1385 side-effects in the SIDER database from the chemical structures of 888 approved drugs. These predictions are performed with simultaneous extraction of correlated ensembles formed by a set of chemical substructures shared by drugs that are likely to have a set of side-effects. The relevance of the information extracted by the method is demonstrated on specific examples. We also conduct a comprehensive side-effect prediction for many uncharacterized drug molecules stored in DrugBank database, and were able to confirm interesting predictions using independent source of information.

The natural step forward is the integration of both chemical structure and protein interaction profiles to predict uncharacterized drug side-effects. This has been done in a further study [136], which results are qualitatively described at the end of the section.

### 4.2.2 Materials

Side-effect keywords were obtained from the SIDER database which contains information about marketed medicines and their recorded adverse drug reactions [82]. This led to build a dataset containing 888 drugs and 1385 side-effect keywords. Each drug was represented by a 1385 dimensional binary profile $\mathbf{y}$ whose elements encode for the presence or absence of each of the side-effect keywords by 1 or 0, respectively. There are 61,102 associations between drugs and side-effect terms in the dataset, and each drug has 68.8 side-effects on average. This dataset is used to evaluate the performance of the proposed methods in this study.

To encode the drug chemical structure, we used a fingerprint corresponding to the 881 chemical substructures defined in the PubChem database [22]. Each drug was represented by an 881 dimensional binary profile $\mathbf{x}$ whose elements encode for the presence or absence of each PubChem substructure by 1 or 0, respectively. A description of the 881 chemical substructures can be found at the PubChem website [22]. There are 107,292 associations between drugs and chemical substructures in the dataset, and each drug has 120.8 substructures on average.

The other drug information (e.g., ATC code, drug category, protein target) was obtained from DrugBank [128]. This information is used to ease biological interpretation in the side-effect prediction for uncharacterized drugs.

### 4.2.3 Methods

We propose five possible methods to predict drug side-effect profiles from the chemical structures. We have a training set of $n$ drugs with $p$ substructure features and $q$ side-effect features. Each drug is represented by a chemical substructure feature vector $\mathbf{x} = (x_1, \cdots, x_p)^T$, and by a side-effect feature vector $\mathbf{y} = (y_1, \cdots, y_q)^T$. We consider

$X$ the $n \times p$ matrix defined as $X = [\mathbf{x}_1, \cdots, \mathbf{x}_n]^T$, and $Y$ the $n \times q$ matrix defined as $Y = [\mathbf{y}_1, \cdots, \mathbf{y}_n]^T$, where the columns of $X$ and $Y$ are assumed to be centered and scaled.

### Sparse canonical correlation analysis (SCCA) and side-effect prediction

Consider two linear combinations for chemical substructures and side-effects as $u_i = \boldsymbol{\alpha}^T \mathbf{x}_i$ and $v_i = \boldsymbol{\beta}^T \mathbf{y}_i$ $(i = 1, 2, \cdots, n)$, where $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_p)^T$ and $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_q)^T$ are weight vectors. As presented in chapter 2, the SCCA method we consider here consists in finding the weight vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ which solve the following $L_1$ constrained optimization problem:

$$\begin{aligned} &\max\{\boldsymbol{\alpha}^T X^T Y \boldsymbol{\beta}\} \quad \text{subject to} \\ &||\boldsymbol{\alpha}||_2^2 \leq 1, \quad ||\boldsymbol{\beta}||_2^2 \leq 1, \quad ||\boldsymbol{\alpha}||_1 \leq c_1\sqrt{p}, \quad ||\boldsymbol{\beta}||_1 \leq c_2\sqrt{q}, \end{aligned} \tag{4.1}$$

where $||\cdot||_1$ is $L_1$ norm (the sum of absolute values of vector entries), $c_1$ and $c_2$ are parameters to control the sparsity and restricted to range $0 \leq c_1 \leq 1$ and $0 \leq c_2 \leq 1$. For simplicity, we use the same value for $c_1$ and $c_2$ in this study. The sparse version of CCA is referred to as sparse canonical correlation analysis (SCCA). Setting $c_1 = c_2 = 1$ defines the original CCA (OCCA) without sparsity constraint and amounts to compute an SVD (see chapter 2). Problem (4.1) can be regarded as the problem of penalized matrix decomposition of the matrix $Z = X^T Y$. As mentioned in chapter 2, we can use the penalized matrix decomposition (PMD) proposed by [129]. After $m$ iterations of the algorithm, we obtain $m$ pairs of weight vectors $\boldsymbol{\alpha}_1, \cdots, \boldsymbol{\alpha}_m$ and $\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_m$ which are referred to as components with associated weights $\rho_1, \cdots, \rho_m$. Components of lower $k$ are called "lower order components", while components of higher $k$ are called "higher order components".

If the extracted sets of chemical substructures and side-effects are biologically meaningful, potential side-effects for a new drug candidate molecule should be predicted comparing the extracted chemical substructures to its chemical structure. Given the chemical structure profile $\mathbf{x}$ of a new drug candidate molecule, its potential side-effect profile $\mathbf{y}$ can be predicted based on the extracted sets of chemical

substructures and side-effects encoded in $\{\boldsymbol{\alpha}_k\}_{k=1}^m$ and $\{\boldsymbol{\beta}_k\}_{k=1}^m$. We use the prediction score described in chapter 2.

$$\hat{\mathbf{y}} = \sum_{k=1}^{m} \boldsymbol{\beta}_k \rho_k \boldsymbol{\alpha}_k^T \mathbf{x}, \tag{4.2}$$

Note that $\hat{\mathbf{y}}$ is the $q$-dimensional vector whose $j$-th element represents a prediction score for the $j$-th side-effect. If the $j$-th element in $\hat{\mathbf{y}}$ has a high score, the new molecule $\mathbf{x}$ is predicted to lead to the $j$-th side-effect ($j = 1, 2, \cdots, q$).

### Support vector machine (SVM)

The side-effect prediction problem can be viewed as a supervised binary classification problem consisting in predicting whether a given drug $\mathbf{x}$ has a side-effect or not. This should be repeated for all $q$ side-effects. The support vector machine (SVM) is a well-known binary classifier, and it has become a popular classification method in bioinformatics [39] and chemoinformatics [79] because of its high-performance prediction ability [105]. We test several kernel functions such as linear kernel, Gaussian RBF kernel with various width parameters, and polynomial kernel with various degree parameters. Note that this strategy needs to construct $q$ individual SVM classifiers for $q$ side-effects, so it will require considerable computational burden, because $q$ is quite huge in practical applications ($q \sim 1000$ in this study). Choice of kernel and its parameter were made using cross validation.

### Nearest neighbor (NN)

The most straightforward approach is to apply the nearest neighbor (NN), which predicts a given drug $\mathbf{x}$ to have the same side-effects as those of the drug (in a training set) whose chemical substructure profile is the most similar. For each query drug, we look for $k$ nearest neighbors, and if $k'$ of $k$ have a side-effect, we assign the prediction score of $k'/k$ to the query drug. We repeat this procedure for $q$ side-effects. The number of neighbours $k$ was obtained by cross validation.

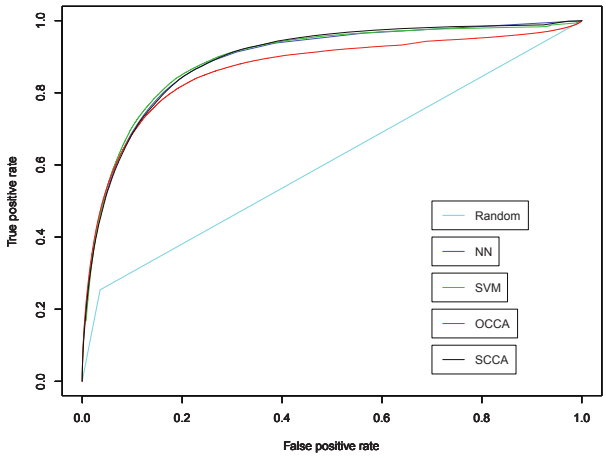



Figure 4.1: ROC curves comparing the performances of nearest neighbor (NN), support vector machine (SVM), ordinary canonical correlation analysis (OCCA) and sparse canonical correlation analysis (SCCA) for side-effect prediction.

#### Random assignment (Random)

To evaluate how difficult the problem considered is, we apply a random assignment procedure, that is, we use the 0/1 ratio to assign a binary label to each test drug randomly. For example, if the ratio in given training data is 90%, we can assign zero for 90% of examples in test; otherwise 1. This method is used as a baseline method in this study.

### 4.2.4 Results

#### Performance evaluation

We applied nearest neighbor (NN), support vector machine (SVM), ordinary canonical correlation analysis (OCCA), and sparse canonical correlation analysis (SCCA) to predict drug side-effect profiles. We also applied random assignment procedure (Random) as a baseline method. First we tested the five methods: Random, NN, SVM, OCCA and SCCA for their abilities to predict known side-effects profiles by the following 5-fold cross-validation. Drugs in the side-effect data were split into 5

subsets of roughly equal size, each subset was then taken in turn as a test set, and we performed the training on the remaining 4 sets. For accurate comparison, we kept the same experimental conditions, where the same training drugs and test drugs are used across the different methods in each cross-validation fold. We evaluated the performance of each method by the ROC (receiver operating characteristic) curve [47], which is a graphical plot of the sensitivity, or true positive rate, against false positive rate (1-specificity or 1-true negative rate). The ROC curve can be represented by plotting the fraction of true positives out of the positives (true positive rate) vs. the fraction of false positives out of the negatives (false positive rate), where true positives are correctly predicted side-effects and false positives are incorrectly predicted side-effects based on the prediction score for various threshold values above which the output is predicted as positive and negative otherwise.

Figure 4.1 shows the ROC curves for the five different methods based on the cross-validation experiment, where the prediction scores for all side-effects were merged and a global ROC curve was drawn for each method. Parameters in each method were chosen by using the AUC (area under the ROC curve) score as an objective function. The best result for NN was obtained by the number of neighbors $k = 50$. The best result for SVM was obtained by Gaussian RBF kernel with width parameter $\sigma = 0.2$ and regularization parameter $C = 1$. The best result for OCCA was obtained by $m = 20$. The best result for SCCA was obtained by the following parameters: $c_1 = c_2 = 0.05$ and $m = 20$. The resulting AUC scores for Random, NN, SVM, OCCA and SCCA are 0.6088, 0.8917, 0.8930, 0.8651 and 0.8932, respectively. From this figure 4.1 we see that the proposed SCCA method outperforms OCCA and its performance is at a competitive level with NN and SVM.

We are also interested in biological interpretability of the outputs of the proposed method to understand the relationship between chemical substructures and side-effects. We focused on OCCA and SCCA, because they are the only methods which can correlate two heterogeneous high-dimentional data sets. We examined the weight vectors for drug chemical substructures and drug side-effects in OCCA and SCCA. Figure 4.2 shows the index-plot of weight vectors in OCCA, and figure 4.3

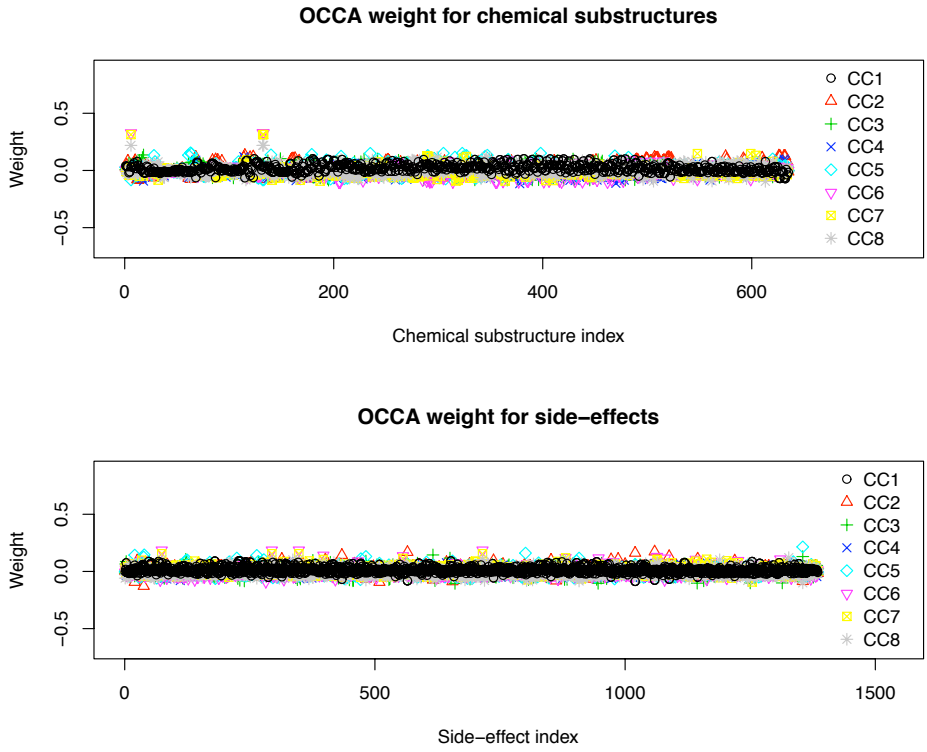



Figure 4.2: Index-plot of weight vectors for drug substructures (upper) and side-effects (lower) extracted by ordinary canonical correlation analysis (OCCA).

shows the index-plot of weight vectors in SCCA, where the first eight canonical components are shown. Almost all elements in the weight vectors extracted using OCCA are non-zero and highly variable, while most of the elements in the weight vectors extracted using SCCA are zero in each component. This underlines the fact that SCCA can select a small number of features as informative drug substructures and side-effects. This result suggests that the proposed SCCA method provides more selective and informative correlation between drug substructures and side-effects without loosing performance. In addition, it should be pointed out that the other methods NN and SVM do not provide any clue for biological interpretation.

Finally, we investigated the computational cost for each method. Figure 4.4 shows

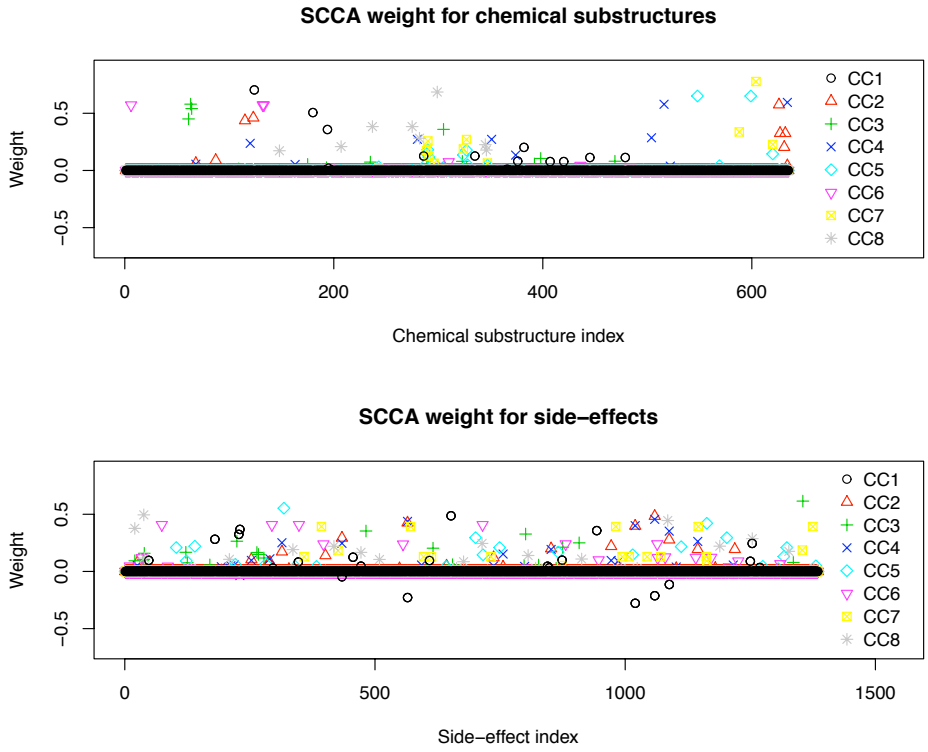



Figure 4.3: Index-plot of weight vectors for drug substructures (upper) and side-effects (lower) extracted by sparse canonical correlation analysis (SCCA).

the total execution times of the cross-validation experiment between the four different methods. NN is the fastest, followed by OCCA, SCCA, and SVM. As expected, SVM is much slower than the other methods, because it requires individual classifiers for all side-effect keywords ($\sim$ 1000 SVM classifiers are required).

### Extracted sets of drug substructures and side-effects

From biological viewpoints, we examined the extracted sets of drug substructures and drug side-effects in each canonical component extracted using SCCA. Note that the other methods (NN, SVM, and OCCA) do not enable us to interpret the biological features. Each component consists of only a small number of substructures and a

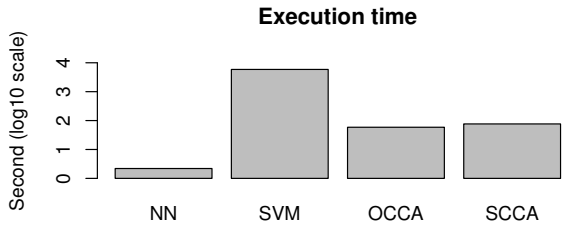


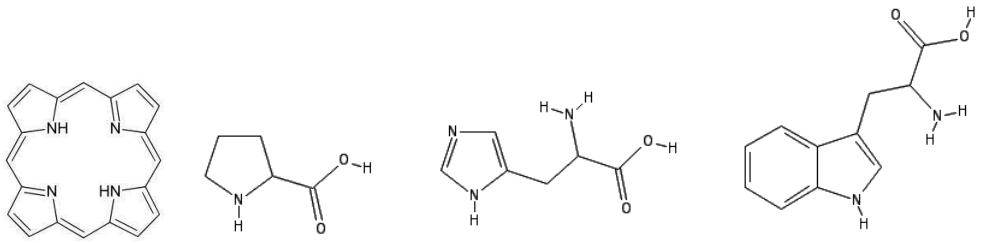

Figure 4.4: Total execution time of the cross-validation experiment for the four methods (log10 scale).

small number of side-effects that are correlated with each other according to SCCA. For each component, two lists of drugs are provided: one containing drugs with a high score for the associated substructures, and one containing drugs with a high score for the associated side-effects. We examined the results when we used the best parameters which provided the highest AUC for all side-effect terms. The content of a few canonical components are discussed to illustrate the ability of the method to extract meaningful biological information.

A canonical correlation coefficient is computed to evaluate the importance of each component. This value corresponds to the value of the objective of (4.1) for each component. The components with high canonical correlation tend to contain rare substructures present only in very few drugs, which are associated to rare side-effects mainly observed for these drugs. These components contain quite specific substructure/side-effect canonical correlations whose interpretation is straightforward. For example, component 6 associates the presence of a boron atom, only found in the bortezomid molecule in the SIDER database, to a short list of neurological side-effects observed only for this drug. Similarly, component 20 essentially clusters a substructure defined by a carbon atom bearing both a bromide atom and a nitrogen atom. This substructure is found only in the bromocriptine molecule of the SIDER database, with two side-effects observed only for this drug (namely, pregnancy induced hypertension and toxemia of pregnancy).

Table 4.1: Nitrogen-containing rings of size 5: (A) Porphyrin group, (B) Proline residue, (C) Histidine residue, (D) Tryptophane residue.

A B C D



In the general case of components containing more frequent substructures, drugs that contain these substructures tend to present side-effects associated to this component, but this correspondence is not strict. Reciprocally, most drugs that have high scores for the side-effects contain the chemical substructures of this component, but not all. Analysis of component 18 can illustrate these points.

This component contains two substructures, the major one being the presence of "four or more saturated or aromatic nitrogen-containing rings of size 5", associated to four side-effects. This substructure is present in five drugs of the SIDER database: verteporfin, porfimer, goserelin, buserelin, and leuprolide. Verteporfin and porfimer contain a porphyrin group displaying four nitrogen-containing rings of size 5, as shown in figure 4.1 (A). Goserelin, buserelin, and leuprolide are synthetic 9-residue peptide analogues of the gonadotropin releasing hormone. Their sequences contain amino-acids whose chemical structures present nitrogen-containing rings of size 5, found in side chains of proline, histidine or tryptophane residues, as shown in figure 4.1 (B), (C) and (D). Overall, four or more nitrogen-containing rings of size 5 are indeed present in their structures.

Note however that these rings are different from those of the porphyrin group. Although goserelin, buserelin and leuprolide on the one hand, and verteporfin and porfimer on the other hand, belong to totally unrelated families of molecules, they share common substructures, at least according to their definition in the present study. All drugs from these two families, but verteporfirin, have high scores for

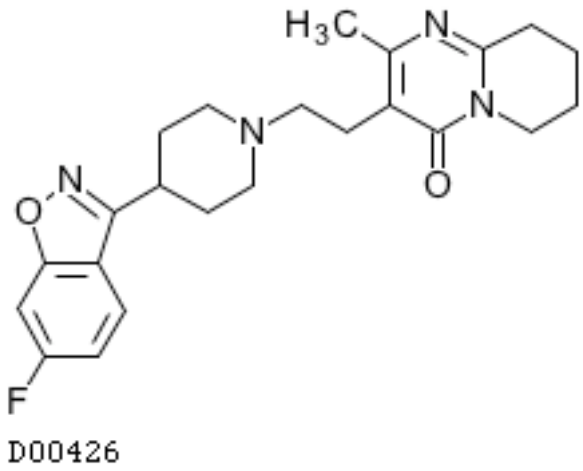



Figure 4.5: Two dimensional graph structure of risperidone.

side-effects of this component. This result illustrates the fact that side-effects of a drug is usually associated to the presence of given substructures, although it may be modulated by the overall molecular structure, as in the case of verteporfirin. This property is also well known in the context of drug structure-activity relationship, which usually depends on given molecular scaffolds, but which is modulated by the presence of additional chemical groups.

Reciprocally, all drugs that have high scores for side-effects of component 18 contain the chemical substructures of this component, but risperidone, as shown in figure 4.5. Its structure is very different from those of porphyrins or gonadotropin analogues. It is an antagonist of the dopamine and of the serotonine receptors. It belongs to the class of antipsychotic agents (see DrugBank), and its high score for side-effects of component 18 cannot be explained in a straightforward manner.

However, in some cases, we were able to relate such unexpected results to the targets of these drugs, as illustrated by component 13. This component contains

substructures that are essentially present in proton pump inhibitors used as anti-ulcer agents like omeprazole. It is also present in a small number of drugs from other families like pramipexole (an antiparkinson agent) or riluzone (a neuroprotective agent). As expected, these anti-ulcer agents are found in the high scoring drugs for side-effects in component 13, together with pramipexole and riluzone, although with lower scores. As for component 18, other drugs that do not contain the high scoring substructures of component 13 are however found among high scoring drugs for side-effects in this component. This is the case of ropinirole. Interestingly, ropinirole is an antiparkinson agent that targets the same protein as pramipexole, namely dopamine receptor.

This result suggests that drugs sharing some protein targets may also share some side-effects. It is also consistent with the idea that the global biological effect of a molecule (both beneficial effects and adverse side-effects) is related to its overall profile of protein targets. Taken together, our results provide examples for which the side-effects of a drug are modulated both by its substructures and by its targets. Note that these two factors are connected since similar molecules tend to share similar protein targets, but this property was not exploited in the present study.

**Comprehensive side-effect prediction for uncharacterized drugs**

We then evaluated the interest of the proposed method for prediction of side-effects for uncharacterized drugs. We predicted potential side-effects for drugs in DrugBank for which side-effect information was not available in the SIDER database. We focused on 2883 drugs which are labeled as "small molecules" in DrugBank. We first make general comments on the results and then present more details for a few well-known specific examples.

Very frequent side-effects, such as "headache" or "nausea" are found in SIDER, and they occur with many drugs. These side-effects are not specific, and they do not appear for a well defined drug category. They are the most frequently predicted side-effects, but they hardly appear with the highest prediction scores for a given drug, which is consistent with the fact that they are common reactions. However, we also find more specific side-effects which are related to special types of drugs.

For example, steroids may lead to "striae", or "linear atrophy", which results in local dermal structure atrophy and skin depigmentation [80]. Indeed, this keyword is mainly found for steroid molecules in SIDER. The top 30 drugs predicted to have this side-effect are also steroids, which is consistent with literature and training data. Moreover, "global amnesia", a very specific keyword in SIDER, is one of the most striking syndromes in clinical neurology whose underlying causes are not well known [101]. 14 drugs catch a high prediction score for this keyword. Among them, one is anticholesteremic, three are antipsychotics, and the others are experimental molecules whose categories are not known. Therefore, three out of four drugs with known indications are related to cognitive functions, which is consistent with the predicted side-effect nature. Although the accuracy of all the predictions was not discussed here, the results are consistent with the available biological and medical information.

We also checked famous examples of withdrawn drugs. Rimonabant (DB06155 in DrugBank) is an anti-obesity agent. It was rejected for approval in the United States, but it was accepted in Europe in 2006. In october 2008, the European Medicines Agency recommended suspension of its marketing authorization because of serious psychic side-effects, mainly severe depression. Indeed, this drug is active in the central nervous system, which may trigger very broad and complex psychic mechanisms. Consistent with this, in our prediction profile, the "borderline personality disorder" and "posttraumatic stress disorder" keywords are found in the ten top ranking keywords for this drug. In other words, our method would have foreseen potential psychoactivity for rimonabant. Furthermore, the method provides a potential rationale for appearance of these psychotic effets. Rimonabant contains the substructure shown in figure 4.6. This substructure is also found in the alprazolam molecule used in the treatment of psychic disorders (a molecule in SIDER). Interestingly, among the 165 molecules of PubChem that also share this substructure and for which pharmacological annotation is available, 40 are classified as "anti-anxiety agents". A reasonable hypothesis to explain rimonabant's severe side-effects may be the presence of this substructure, together with the nature of its protein target (namely, the cannabinoid receptor).

Terfenadine (DB00342 in DrugBank) is an anti-allergic agent which was withdrawn


Figure 4.6: The substructure of Rimonabant selected to be a clue of psychoacticity.

by the U.S. Food and Drug Administration in 1997 because of toxic effects on heart rhythm. The "Aortic stenosis" and "aortic valve incompetence" keywords rank 9-th and 11-th among the predicted side-effects for this drug. These related side-effects are known to often lead to arrhythmias [98], as observed for this drug. In this case again, our method would have foreseen potential severe cardiac side-effects.

### 4.2.5 Discussion and conclusion

In this section we investigate the question of predicting potential side-effect profiles of drug candidate molecules based on their chemical structures using sparse canonical correlation analysis (SCCA). The method is computationally efficient and is applicable on large datasets. From a system perspective, the originality of the proposed method lies in the integration of chemical space and pharmacological space in a unified framework, in the extraction of correlated sets of chemical substructures and side-effects, and in the prediction of a large number of potential side-effects in a row.

Numerical experiments suggest that the method is competitive with state-of-the-art approaches in the task of predicting drug side-effects based on chemical structure. After training the method using publicly available data, we could predict side-effect for out of sample molecules which we could confirm using independent information sources.

One main difficulty of using SCCA is to choose appropriate sparsity parameters and appropriate number of components. High sparsity promoting parameters would

lead to an over-sparse model in all the cases, which might be misleading in the interpretation if the degree of sparsity was not tuned carefully. The optimal parameters value depends highly on the definition of the objective function to be investigated in the cross validation. We evaluated global prediction accuracy, involving all possible drug-side-effect associations. The definition of an appropriate objective function in the cross-validation is an important issue. There remains much room to develop a more appropriate way to choose the parameters, depending on the goal of the analysis.

The proposed method can be applied at various stages of the drug development process. At early stages, among several active drug candidates, the method could help to choose the molecules that should further continue the process and those that should be dropped. It could also help to find new indications for known drugs, a process named drug repurposing. Indeed, side-effects of drugs used in a given pathology can be viewed as a beneficial effect in another pathology. Sildenafil is a famous example of such drug repositioning. The method could help to identify chemical substructures of known drugs that might participate in the appearance of a given side-effect. These substructures could be used as building blocks in fragment-based drug discovery approaches [44] for pathologies in which this side-effect could be positively exploited.

Experimental results suggest that the SCCA approach is competitive with state-of-the-art methods in term of prediction accuracy. It associates chemical substructures and protein domains in components. From a biological point of view, we provide examples that illustrate the modulation of drug side-effects by chemical structure. These examples were provided by the analysis of extracted components. The underlying mechanical mechanism is the protein ligand interaction that we investigated in chapter 3. Drugs perturb biological process by interacting with the corresponding proteins. Therefore a logical continuation of this work is to integrate both chemical and protein target information in order to predict side-effects. The next section describes an extension of this work based on this idea.

### 4.2.6 Integrating different sources of information

An extension of this work was proposed in [136] which we briefly outline in this paragraph. In this section, we considered predicting drug side-effects from chemical structure. As mentioned in the introduction and in chapter 3, drug side-effects are the result of the interactions between the drug and all its targets. This point was further illustrated by giving examples of chemical substructure and side-effect association which were coherent in terms of protein target interaction profiles. These examples were extracted using SCCA method without incorporating this target profile information in the analysis. The next step is to include protein target information for side-effect prediction. Moreover, both chemical information and protein target information can be combined to achieve potentially greater accuracy on this prediction task.

[136] proposes to use kernel regression to integrate chemical and protein target information to predict side-effects. The protein and ligand information used is similar to those used in chapter 3, and the chemical and side-effect information are the same as in the present section (see section 4.2.2). Kernels offer a flexible framework to combine different sources of information to make predictions. Using properties such as additivity, it is possible to build kernel reflecting both chemical similarity and target protein profile similarity for drugs. Kernel regression consists in combining the square loss with a class of functions defined by a positive definite kernel in a supervised learning framework. Solving the kernel regression problem can be done in closed form and requires a matrix inversion.

Based on an extension of the dataset presented in this section (including protein target information from DrugBank database [128] and Matador database [49]), the work of [136] compares the CCA based methods and kernel regression based methods when considering chemical information or protein target information separately or considering the integration of both sources of information. The results presented in [136] suggest that using both sources of information significantly improve the performances in predicting side-effects for drugs when using kernel regression methods. Although the problem tackled in the present section is very similar to that of this extension, it is not possible to compare them quantitatively. Indeed, [136] is based on

a slightly different dataset and filters very frequent side-effects. Moreover, the performance metrics used are different (precision recall curve instead of Receiver Operating Characteristic).

Finally, the methods proposed in [136] only consider performance on side-effect prediction tasks and do not provide any interpretable biological feature. Therefore, they do not provide any clue to associate chemical fragments with specific side-effect as we presented in the previous section. However, the study presented in [136] demonstrates on the drug side-effect prediction task that the integration of different sources of informations improves the prediction accuracy. This is a natural extension of the study presented in this section and a contribution to the development of system based approaches in the context of phenotypic characterization of un-marketed small molecules.

## 4.3 Cell population phenotyping

This section is dedicated to the study of the phenotypic effect of siRNA gene knock down experiments based on high content screening data. A siRNA is a small RNA molecule which is able to bind a specific messenger RNA in order to prevent the translation of the underlying gene. This process results in the silencing of further molecular mechanisms and may finally cause a phenotypic change at a higher level in the hierarchy of biological scales. The biological scale considered here is that of populations of cells. At such a scale, a group of cells can be viewed as a system which organization holds information about the behaviour of the group as a whole. The introduction describes the technology and motivates the necessity to take into account variability within a population of cells and descriptors dependence structure. These specific considerations arise from the fact that we consider groups of cell as complex entities. Indeed, they do not apply when considering cells as separate entities. Preliminary numerical analyses reinforce these remarks based on experimental data. A probabilistic model is described and its properties are investigated in terms of phenotypes at the level of cell populations. The results presented are based on the work published in [95].

### 4.3.1 Introduction

#### High content screening data processing

Fluorescent markers allow to label virtually any cellular structure in living cells [43]. Recent advances in sample preparation and microscopy automation allow cell population imaging on a large scale [96]. Both technologies lead to the development of High Content Screening (HCS) platforms which allow screening living cells under a wide range of experimental conditions. Classically, the aim is to identify a therapeutic target, or a drug candidate. One screen consists in taking several pictures of a large number of cell populations, for example, transfected with RNAi tools or exposed to small molecules. Each experiment is performed in a well in which several pictures are taken, called fields. It gives access to a whole panel of cellular responses to a specific

manipulation. The outcome is a series of cell population images which holds much more information than the single averaged value of the cellular response, classically recorded in HTS screens. The available information accounts for cell variability according to various features, which is precious to characterize a population of cells. However, the heterogeneity of cellular responses makes it difficult to compare and interpret experiments. In the framework we consider in this work, processing the outputs of such experiments requires three steps as illustrated in Figure 4.7.

- **Step 1, segmentation:** This step consists in identifying cells in images and extract features that characterize the shape and texture for each individual cell.
- **Step 2, cellular phenotyping:** This step usually involves machine learning algorithms that classify cells according to different predefined cellular phenotypes based on cellular features and on a training set of annotated cells for which this phenotype is known.
- **Step 3, population phenotyping:** This step aims at defining phenotypes (or classes) at a population level, using population descriptors derived from cellular phenotypes, in order to describe and compare different experiments.

Segmentation and cellular phenotyping steps have been well studied (steps 1 and 2). There has been a huge amount of work to apply image processing tools to cell segmentation and cell features extraction from cell population images. Typical cellular features used in this context are nucleus and cytoplasm size, texture and shape. The cellular phenotyping step aims at converting, for each single cell, the numerical values corresponding to its cellular features into predefined biological phenotypes that are relevant at the cellular scale and characterize the cell status. Typical examples are cell morphology classes, such as shape and appearance or cell cycle state (G1, S, G2 or M phases). Coupling the image segmentation step with supervised machine learning algorithms, many authors proposed methods to classify cells according to various predefined cellular phenotypes using HCS data and a training set of annotated cells [24, 125, 65]. These applications developed in the last decade demonstrate empirically the effectiveness of machine learning algorithms in this setting. Example

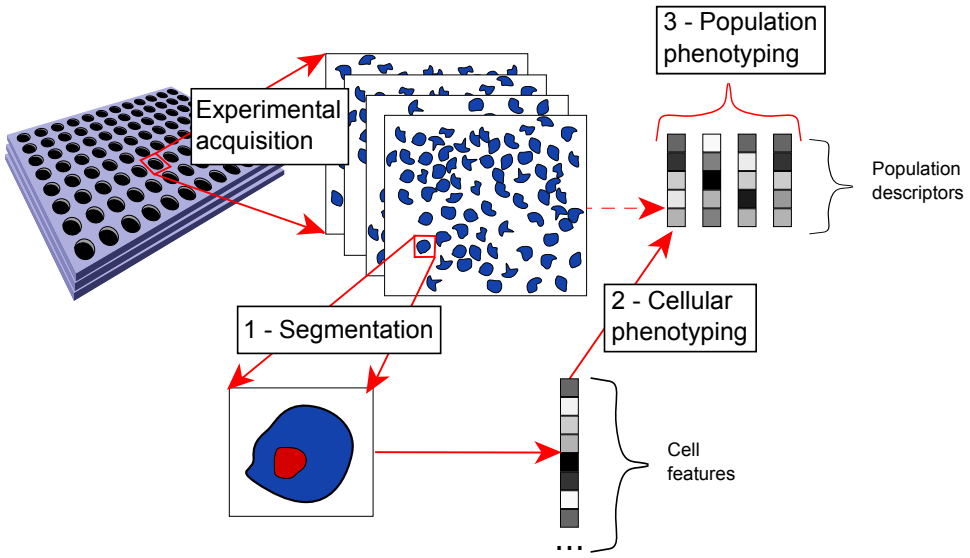



Figure 4.7: HCS data acquisition and processing. After experimental acquisition, we have four images, or fields, per well. Step 1 consists of isolating each cell in each image and computing cell features by means of image processing tools. These features are used to classify cells in each image according to different predefined cell phenotypes in a second step (for example M, G2 phases or apoptosis). This classification provides population descriptors that can be used to define population phenotypic classes.

of algorithms used in this context are state-of-the-art classification algorithms such as support vector machines [112] or boosting [36].

### Analysis methods

A common practice in HCS analysis is to inspect univariate cell features averaged over wells [99, 13, 17]. This is suited for analysis of a single channel (for example corresponding to a single cellular phenotype such as "apoptosis state"). However, analysis of multiple cellular phenotypes may require to take into account their joint distributions. In our setting, cells from images are phenotyped in steps 1 and 2. As there are many potential cell phenotypes of interest, the multivariate setting must be considered, which constitutes a characteristic of the proposed method. Moreover, our model accounts for the field variability in each well, not only averaged values over wells. This constitutes a step toward cell variability characterization within each

well, since within a population of cells in a well, one may observe a range of cellular responses to a given experimental condition. Indeed [113] observe a significant impact of the cell population context on the cellular phenotypes in siRNA screens. They found that local cell density, position of a cell in the local cell population or cell size significantly influence phenotypes such as viral infection or endocytosis.

In an HCS experiment, once, all cells of a population (namely all cells of a well) have been assigned cellular phenotypes, the aim is to characterize this cell population (step 3). In other words, we would like to define a population phenotype based on the cellular phenotypes of all the individual cells it contains, since different cells taken from the same well can display different cellular phenotypes, even when exposed to the same experimental conditions. Therefore, cellular phenotypes cannot be used as population phenotypes in a straightforward manner, and it remains a challenging issue to fill the gap between phenotypical characterization of a population of cells and single cell phenotypes. In particular, definition of population phenotypes is an important issue that one must solve in order to compare cell populations subject to different treatments. For example [37] carry out the segmentation and cellular phenotyping steps and propose a distance learning method to compare different cell populations, and generalize known relations between experiments in a third step. In a different experimental setting, [89] use trajectories defined by time varying cell population responses to compare treatments.

### Content of this section

In this section, we present a method to describe and compare populations of cells in HCS experiments by defining population phenotypes. The input of the proposed method is a table in which each row is a field (an image) and each column is a population descriptor for these fields. For each well several fields are recorded, and well assignment information is available for each field. However, the behaviour of fields within the same well might be different. We refer to this aspect as "within-population variability". Moreover, population phenotyping should not only take into account each single population descriptor individually, but also the joint distribution of these descriptors. We refer to this as "dependence structure of cell population

descriptors". Taking this dependence structure into account improves the description power of the model. Illustration of these aspects of our HCS dataset and further biological motivations will be presented in the method section.

Going back to the HCS data analysis framework presented in Figure 4.7, the proposed method tackles step 3: population phenotyping. A natural approach to characterize a population of cells is to consider the output of the first two steps as descriptors for the population of cells. The total number of cells and the proportion of cells assigned to each predefined cellular phenotype describe the joint behaviour of all cells in a given population. The problem is now to assign a population phenotype based on the descriptors of this cell population. For example, in the present study, we aim at defining population phenotypes based on the following population descriptors : cell count, and cellular phenotypes which are represented by proportions of cells in the different stages of the cell cycle. A population phenotype is meant to characterize the biological state of the cell population in a given experiment. Each population phenotype (or class) should gather cells which behaviours are similar, and population of cells showing dissimilar behaviours should be assigned to different classes. The conceptual difference with the cellular phenotyping step (step 2) is that we do not have predefined population phenotypes, nor do we have annotated cell populations according to population phenotypes. Indeed, the question of how to define such population phenotypes is still open. Therefore, while a supervised framework is suited for solving step 2, because there exists predefined cellular phenotypes, we propose an unsupervised method to tackle the population phenotyping step (stpe 3) where predefined cell population phenotypes are unknown.

We model a cell population using a hierarchical mixture model which is a specific kind of bayesian network, a widely-used class of probabilistic models [78]. "Within-population variability" is modelled using a hierarchical structure and "dependence structure of cell population descriptors" is modelled using multivariate probability distributions. The output of the method characterizes the density of the input fields in the population descriptors space and assigns a phenotypic class to each well. A copula-based parametrization is compared to a gaussian parametrization of the proposed mixture model (details are found in the methods section). To validate our

hypotheses regarding "within-population variability" and "dependence structure of cell population descriptors", we compare performances of the two preceding models to a baseline gaussian mixture model with diagonal covariance matrix, which would correspond to ignoring these two aspects of the data.

In summary, the proposed method is a tool for analysing cell population data. It relies on prior image segmentation and cellular phenotype assignment which corresponds to steps 1 and 2 of this analysis framework. The main purpose of the method is to extract cell population phenotypes and to assess phenotypic variability at the level of cell populations. The model allows to take advantage of HCS specific information: "dependence structure of population descriptors" and "within-population variability", which our experiments suggest to consider in our context. This can be used to tackle the problem of novelty detection (for example, outlier genes in a siRNA experiment), which is one of the main goals of HCS experiments. We validate a HCS data analysis method based on control experiments. It accounts for HCS specificities that were not taken into account by previous methods but have a sound biological meaning. Biological validation of previously unknown outputs of the method constitutes a future line of work.

### 4.3.2 Materials and Methods

#### Experimental acquisition

siRNA screening was performed on shA673-1C Ewing sarcoma derived cell line [121] by the Biophenics platform at Institute Curie. Two experimental conditions were considered: cells were either transfected with a negative siRNA controls (Luciferase GL2 siRNA, Qiagen) or a positive siRNA control (KIF11). Cell numeration and mitotic figures were determined using DAPI staining, cycle phases distinction were determined using EdU (for S Phase) and Cyclin B1 (for G2-M transition) immunofluorescence staining. Apoptosis was detected by cleaved caspase 3 immunofluorescence staining. Images were acquired on IN Cell1000 Analyzer (GE Healthcare Life Sciences) and segmented using IN Cell Investigator software.

### Dataset

Our dataset is comprised of 2688 fields belonging to 672 wells for which we have total cell count and proportions in S, G2, M and Apoptotic phases. Each well is either related to a GL2 or a KIF11 experiment. In addition , we have well assignment information for each field (336 wells x 4 fields per well x 2 manipulations = 2688 fields). Note that cellular phenotypes are not exclusive here. This dataset is one example of output of the two first steps we mentioned in the introduction and the purpose of this study is to validate our method based on it. The proportion of cells in the G0/G1 phases is deduced from the total of those in the S, G2, M.

### Preliminary data analysis

To motivate the need for accounting for "dependence structure of population descriptors" and "within-population variability", we present two simple observations arising from the dataset described in the previous paragraph.

First, we studied the association between cell population descriptors. More precisely, we searched for potential positive or negative correlation between the cell count and the other population descriptors. As shown in Table 4.3.2, there is no association between number of cells and S-phase proportion, as expected: DNA replication is a process of quite constant duration because it mainly depends on the species and the size of the genome. Therefore, the length of the S phase should not depend on the proliferation status or the size of the cell population, as observed. There is a slight positive association between cells number and G1/G0-phase proportion. A plausible biological interpretation is that, at a higher number of cells in a well, the population tends to reach confluence, a situation in which the cell cycle is arrested and cells are known to accumulate in phases G0/G1. In addition, we observed stronger dependences between population descriptors. A positive association is observed between cells number and G2-phase proportion, as well as a negative association between cell number and M-phase proportion. This is a biological observation which has not been generally reported, at least to our knowledge. It may be specific to our experimental design, in which the fields with the highest numbers of cells are reaching the limit

Table 4.2: Association between population descriptors. Association between cell count and proportion of cells in different states based on negative controls. The measure of association is Spearman's rho and the p-value is computed via the asymptotic t approximation [54].

| | S | G2 | M | Apoptosis | G0/G1 |
|---|---|---|---|---|---|
| rho | 0.04 | 0.51 | -0.44 | -0.09 | 0.07 |
| p-value | 0.144 | 2e-16> | 2e-16> | 0.0006 | 0.01 |

of confluence, potentially leading to slow the G2 phase and consequently displaying a reduced number of cells underlying mitosis. Whatever the interpretation of the above observations might be, these results indicate that the cell descriptors used in this study present a dependence structure, and this justifies the choice of a model that can account for this dependency.

Second, we compared the dispersion of fields belonging to the same well to that of fields randomly selected in the dataset. By dispersion, we mean how close a set of fields are one to the other. The distance used is the euclidean distance and the population descriptors used are cell count and proportion of cells in S, G2, M and apoptotic phases. We scaled the data beforehand and used the measure of dispersion of multivariate analysis of variance proposed in [4]. This is the sum of squared pairwise distances. If we consider the set of fields $\{\mathbf{x}_1, \ldots, \mathbf{x}_4\}$, then the dispersion measure is:

$$\sum_{i,j=1}^{4} ||\mathbf{x}_i - \mathbf{x}_j||^2.$$

This is equal to the sum of squared distances of each point from the mean, up to a constant multiplicative factor, and therefore measures how dispersed the fields are. Figure 4.8 indicates that : (i) fields belonging to the same wells do display some variability, which should be taken into account by the model, (ii) this variability is smaller from that of randomly selected fields. Indeed, fields belonging to the same well are part of the same experiment and therefore, are expected to display less phenotypic variability than randomly selected fields. Taken together, these two observations are respectively in good agreement with the ideas of modelling the experimental data

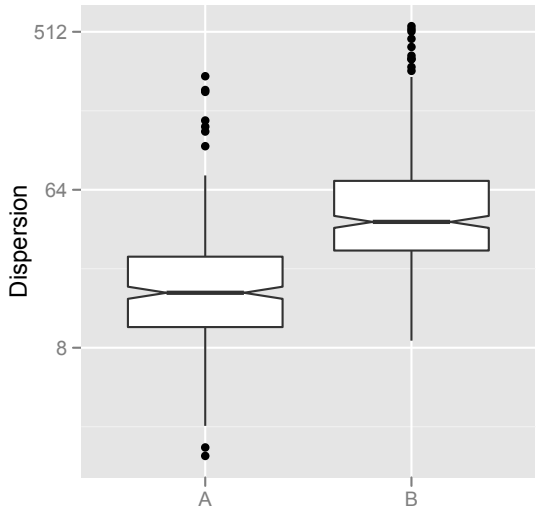



Figure 4.8: Within population variability. Comparison of the dispersion of fields belonging to the same wells (boxplot A) and randomly selected fields (boxplot B). The measure of dispersion is the sum of squared pairwise distances. The population descriptors (cell count and proportions of cells in S, G2, M and apoptotic states) have been scaled beforehand.

taking into account (i) "within-population variability" (ii) within a hierarchical model.

### Model

The proposed model aims at describing HCS data, i.e. a set of wells, each of them containing four fields. The input of the method is a representation based on cell descriptors at the field level (cell count, proportion of cells in S, G2, M and apoptotic phases in our case), coupled to well assignment information. The output is an ensemble of population phenotypes (classes) represented by multivariate distributions. To each image (field), the method assigns a distribution over population phenotypes. We added the quite natural constraint that fields belonging to the same well should correspond to the same class. This hypothesis allows to take into account the "within-population variability" in a given well, which should be part of the population phenotype characterization. This is made possible thanks to the hierarchical

structure of the proposed model.

We tested two different parametrizations for this model : copula-based and gaussian-based. In the former case, the dependence structure of the multi variate class conditional densities are modelled using a copula independently from univariate marginals which shape can be chosen separately. Copulas have been studied since the middle of the 20-th century [40] and have been successfully applied to finance [115], hydrology [41], meteorology [106], neurosciences [91] or gene expression data [73]. We introduce copula-based distributions, to build probabilistic densities that represent cell population phenotypic classes. The use of copula for model-based clustering has been suggested by [60], and proposed by [21] in a semi-parametric framework.

**Formulation of the model:** We observe $\mathbf{X_o}$ which is composed of $\mathbf{N}$ wells $\{\mathbf{X_1}, \ldots, \mathbf{X_N}\}$. We assume that we have $M$ fields in each well $\mathbf{X_n}$. Each field is a vector in $\mathbb{R}^d$. Therefore we represent each well $\mathbf{X_n}$ by a $M$-tuple of vectors $\mathbf{X_n} = \{\mathbf{x_{n}}1, \ldots, \mathbf{x_{n}}_M\}$ where $\mathbf{x_{n}}_i \in \mathbb{R}^d$ for $\mathbf{n} = \mathbf{1} \ldots \mathbf{N}$ and $i = 1, \ldots, M$. In our application, we have $d = 5$ and $M = 4$. The components of this representation are cell counts and proportions of cells in different phases of the cell cycle. In order to model different classes of wells, we introduce the latent variable $\mathbf{Z} \in \{1, \ldots, K\}$ associated to each well where $K$ is fixed in advance. We also assume that given the value of $\mathbf{Z}$, the fields belonging to one well are independent and that wells are independent and identically distributed. These are typical assumptions made in graphical models literature. If $\mathbf{\Theta}$ represents the parameters of this model, the density associated to $\mathbf{X_n} = \{\mathbf{x_{n}}1, \ldots, \mathbf{x_{n}}_M\}$ is then

$$P(\mathbf{X_n}|\mathbf{\Theta}) = \sum_{\mathbf{Z}=1}^{K} P(\mathbf{Z}|\mathbf{\Theta}) \prod_{j=1}^{M} P(\mathbf{x_{n}}_j|\mathbf{Z}, \mathbf{\Theta}) \tag{4.3}$$

With this definition, the likelihood of the total dataset $\mathbf{X_o}$ becomes

$$P(\mathbf{X_o}|\mathbf{\Theta}) = \prod_{\mathbf{n}=\mathbf{1}}^{\mathbf{N}} P(\mathbf{X_n}|\mathbf{\Theta}) = \prod_{\mathbf{n}=\mathbf{1}}^{\mathbf{N}} \sum_{\mathbf{Z}=1}^{K} P(\mathbf{Z}|\mathbf{\Theta}) \prod_{j=1}^{M} P(\mathbf{x_{n}}_j|\mathbf{Z}, \mathbf{\Theta}) \tag{4.4}$$

Given $\mathbf{\Theta}$, this model can be viewed as a generative process which explains how to generate the data from a probabilistic point of view. To generate a cell population (a well $\mathbf{X_n}$), this process takes the following form:

- Choose a population phenotype (a class) from a fixed list. This amounts to sample $\mathbf{Z} \sim P(\mathbf{Z}|\mathbf{\Theta})$.
- Given the population phenotype, generate several sub-populations (fields) according to the multivariate distribution related to this population phenotype. This amounts to sample for $1 \leq j \leq M$, $\mathbf{x_n}_j \sim P(\mathbf{x_n}_j|\mathbf{Z}, \mathbf{\Theta})$

Given $P(\mathbf{Z}|\mathbf{\Theta})$ (a multinomial) and the class conditional density $P(\mathbf{x_n}_j|\mathbf{Z}, \mathbf{\Theta})$, the main issue is to perform inference and learning, which is reversing the generative process defined above to estimate the class distribution related to each well, $P(\mathbf{Z}|\mathbf{X_n}, \mathbf{\Theta})$ , and estimate the parameters of the distributions representing phenotypic classes. We propose gaussian class conditional distributions and copula-based distributions which we now describe.

**Copula-based class conditional distributions:** Copulas became popular in statistical literature at the end of the twentieth century. However, the study of these probabilistic objects goes back to the middle of the century, see [88] for a general review about copulas. The usefulness of copulas comes from Sklar's theorem which states that multivariate distributions can be formalized in term of copula and univariate marginal [110] (see also appendix A).

We use the gaussian copula family which has been introduced in 2000 by [132]. We use the density function formulation of these copulas which let us work with probabilistic densities. A gaussian copula density function is parametrized by a correlation matrix $R$. We refer to the gaussian copula density function as $c_{gR}$. Let $\{F_{\theta_1}, \ldots, F_{\theta_d}\}$ be a set of univariate marginal distributions, $\{f_{\theta_1}, \ldots, f_{\theta_d}\}$ the corresponding univariate densities, such that $\theta_i \in \mathbb{R}^{+*} \times \mathbb{R}^{+*}$ for all $i$. We parametrize $f_{\theta_1}$ as a gamma distribution with parameters $\{\theta_{11}, \theta_{12}\}$. This is a distribution over strictly positive numbers which represents cell counts here. Moreover, we parametrize $f_{\theta_i}, i > 1$ as a beta distribution with parameters $\{\theta_{i1}, \theta_{i2}\}$. This is a distribution

over $(0, 1)$ which represents proportions of cells showing different cellular phenotypes. Plugging these marginals in the gaussian copula $c_{gR}$, whose correlation matrix is $R$, allows to parametrize a distribution whose support is exactly the one our variables are limited to, and to model the dependence structure between univariate marginals. If $\mathbf{x} = (x_1, \ldots, x_n) \in \mathbb{R}^{+*} \times (0, 1)^{d-1}$, it takes the form:

$$P(\mathbf{x}|R, \theta_1, \ldots, \theta_d) = c_{gR}(F_{\theta_1}(x_1), \ldots, F_{\theta_d}(x_d)|R) \prod_{j=1}^{d} f_{\theta_i}(x_i)$$

Moreover, we notice that such a parametrization of the class conditional distribution involves exactly the same number of parameters as a standard gaussian model: one correlation matrix and two parameters per univariate marginal. For copula-based densities, standard parameter estimation by maximum likelihood [25] requires computationally intensive numerical optimization. Approximations of this procedure have been proposed to avoid this. Among them, inference function for margin [109, 63] consists in estimating univariate marginal parameters first and to estimate copula parameters given the marginal parameters in a second step. A semi-parametric procedure [40] has been proposed. This is similar in spirit, except that a non parametric estimate of the univariate marginal distribution function is used (normalized ranks) before estimating the copula parameters based on these non parametric estimates. We used the latter method to estimate copulas parameters together with standard univariate maximum likelihood estimates for marginals. [72] observed empirically that this procedure is more robust to marginals misspecification than the standard maximum likelihood and inference functions for margin. Moreover, this method is less computationally intensive than the two former ones. Details about parameter estimation procedures are provided in appendix A.

**Inference and learning:** Assume that we have a parametrized class probability distribution $P(\mathbf{Z}|\mathbf{\Theta})$ (a multinomial) and a class parametrized conditional distribution $P(\mathbf{x}|\mathbf{Z}, \mathbf{\Theta})$, gaussian or copula-based in our case. Finding the best parameters for our mixture model amounts to maximize (4.4) or the logarithm of (4.4). Optimizing this objective with respect to $\mathbf{\Theta}$ is made difficult by the presence of a sum over latent

classes. Approximate inference has shown to be efficient in this kind of setting. [66] provides a general framework for EM type inference among others (see also chapter 2), which we used to learn the parameters of the model and to infer phenotypic classes of wells in our dataset. Sufficient statistics can be used in the gaussian case. In the copula based model case, we implemented the semi-parametric estimation procedure of [40]. The maximization step is very similar to standard maximization step for Gaussian mixture models. First, univariate marginals parameters are estimated using sufficient statistics and second, correlation matrix of copulas are estimated using scaled ranks instead of absolute values of marginal variables. After optimizing the model parameter $\mathbf{\Theta}$, we obtain $K$ classes represented by class proportions $P(\mathbf{Z}|\mathbf{\Theta})$ and class distribution $P(\mathbf{x}|\mathbf{Z}, \mathbf{\Theta})$. Each well $X$ can be represented as a mixture of cell population phenotypes given by $P(\mathbf{Z}|\mathbf{X}, \mathbf{\Theta})$, which is inferred during the optimization process.

**Baseline comparison:** The proposed model accounts for "within population variability" through its hierarchical structure and "dependence structure of cell population descriptors" through multivariate probability distributions that model dependence between variables. Those two aspects of the model are motivated by observations arising from the data. In order to validate those hypotheses, we compare the performances of those two models to a standard gaussian mixture model with diagonal correlation matrices. This model does not take into account the fact that different fields come from the same well. It also assumes an absence of dependency between population descriptors, because the gaussian class conditional distribution covariances matrices are constrained to be diagonal.

### 4.3.3 Results

Data was generated from a siRNA based HCS on a Ewing sarcoma derived cell line. The considered population descriptors were cell count and proportion of cells showing different cellular phenotypes (S, G2, M phase or apoptotic state). From these data, positive and negative siRNA controls were used in this work to illustrate our approach. GL2 siRNA is a negative control that does not affect proliferation and

cellular phenotypes. KIF11 siRNA is a positive control that induces cell death and therefore leads to massive alteration of cellular phenotypes.

As presented in the "Preliminary data analysis" of Materials and Methods section, we observed that the population descriptors displayed a dependent structure, and that fields belonging to the same well presented less dispersion than fields randomly selected from the dataset (see Figure 2). These preliminary results justify the use of the proposed gaussian or copula based hierarchical models.

We first compare the gaussian and copula based parametrizations of the model in terms of model fitting and generalization properties (See model fitting section). Once parameters of the model are fitted to the data, we build an object representing the density of the data we considered. This is useful in term of novelty discovery. In our case, it would correspond to finding cell populations that are different from the negative control population (GL2 silencing siRNA transfected cells), which behaviour is supposed to be hardly affected by this transfection. Confronting a test dataset to the model, evaluating the likelihood of this new data with respect to this model, allows to measure how different from the training set the test set is. We observe that the proposed method allows to separate positive and negative controls (see section "Novelty detection and positive controls").

Moreover, given the training set, the model classes define the population phenotypes and account for the joint distribution of cell population descriptors. We investigate the properties of these phenotypic classes and underline that the copula based parametrization extracts more meaningful phenotypic classes (see section "Model classes as population phenotypes"). Moreover, we show how those population phenotypes account for different cell behaviours by relating the population phenotypes to cellular phenotypes (see section "Relation between population phenotypes (classes) and cellular phenotypes").

We discuss the advantage of the proposed model compared to previous approaches focusing on one specificity of the approach, "within-population variability" consideration (see section accounting for "within population variability"). We first describe the cross validation experiment that was carried out to evaluate properties of the model.

### Cross validation

We performed 5-fold cross validation experiments on the negative controls dataset composed of 336 wells which represents 1344 fields. This set is split into five subsets of roughly equal sizes. Each subset is taken in turn as a test set, the model is trained on the remaining four sets, and the likelihood of the test set is then evaluated with respect to the model built with the training set. Because the optimization result relies on the initial parameter value, we performed five random restarts for each fold. This allows to evaluate the generalization performances of the model for the whole dataset. We performed this experiment for the gaussian and copula-based models, as well as the baseline model, for a number of population phenotypes ranging from 2 to 20. The number of classes is a parameter of the proposed method. We repeated this experiment ten times over different splits of the dataset. The model giving the best generalization property, i.e. the model with the highest test likelihood, was then trained on the whole negative controls set and the corresponding classes were analysed.

### Model fitting

The cross validation experiment allows to compare different model performances on this dataset. Because all the proposed model are probabilistic in nature, the first criterion we choose to compare different models is the likelihood computed for a test dataset . We proposed two parametrization of the mixture models, a gaussian and a copula-based parametrization which we review in the method section. We compare those two parametrization to the baseline model using this criterion .

Figure 4.9-a shows the training log likelihood of the two models and the baseline model for different numbers of classes. This training likelihood was evaluated using the whole training set. It appears that the copula-based model results in a higher value of the training likelihood. This observation is valid for the whole range of number of classes we considered. It also appears that the baseline fits much less to training data .

Figure 4.9-b represents the test log likelihood, evaluated by cross validation, for

the two models with different numbers of classes. Again, it appears that the copula-based model has better generalization properties independently from the number of classes. Here again the baseline model provides worse fit on test data .

This experiment shows that the proposed model outperforms the baseline model on both training and test datasets for both gaussian and copula based parametrization. This observation validates assumptions encoded in the model which we referred to as "dependence structure of population descriptors" and "within-population variability". We consider now comparing in more details the two parametrizations of the proposed model.

The copula-based model outperforms the gaussian model providing better fit on training data and higher generalization properties on the negative control dataset, while involving exactly the same number of parameters. The copula-based density support matches the domain where our dataset is spread, while the gaussian support is the whole space. Similar results have been reported in other comparative studies of copula models based on different datasets: [32] is an example.

Based on these results, we pick up the model providing the best generalization performances and fit it to the whole negative control set, restarting randomly the algorithm 10 times to avoid local optimum for the parameters values. The results are presented in the two following sections.

### Novelty detection and positive controls

One of the objectives of modelling the negative controls density is to show that we can detect cell populations that are different from these controls, because they could correspond to experiments that are relevant for the studied biological question. To illustrate this point, we used, as controls, cell populations that were transfected with a KIF11 silencing siRNA. We refer to these cell populations as positive controls. It is known that these controls should have a very different behaviour compared to negative controls. Panels **(a)** to **(e)** in Figure 4.10 represent the densities of positive and negative controls univariate cell population descriptors averaged over wells. Panel **(f)** in Figure 4.10 represents the densities of positive and negative control log likelihood. Here the model is trained on negative controls.

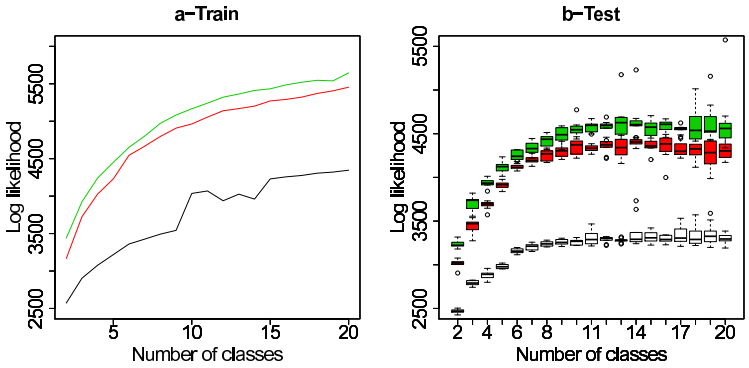



Figure 4.9: Model fitting. Train (a) and test (b) log likelihood of the negative control data for the two proposed models, and the baseline, varying the number of phenotypic classes. Green corresponds to the copula based model, red corresponds to the gaussian model, and black corresponds to the baseline model. For training log likelihood, we picked the best model among 10 random restarts of the algorithm. For the test log likelihood, the boxes account for the variability among ten different splits of the data in a cross validation setting. Given a data split, for each fold and each number of classes, we picked the best model among 5 random restarts of the algorithm.

Positive controls are found to be very different from negative controls. It is easy to distinguish them from negative controls only looking at cell count, for example. The panel **(f)** in Figure 4.10 represents the distribution of log likelihood over wells. The log likelihood given by the model separates the two types of controls. Training our multivariate model on negative controls and testing it on experiments is not less powerful than using univariate methods.

### Model classes as population phenotypes

We propose to use several densities in a mixture model to define population phenotypes by the classes of the model, which corresponds to a mathematical definition. The number of classes was chosen by cross validation.

We inspected the univariate marginal densities. Figure 4.11 compares the empirical density and the density fitted by the model for one population descriptor, proportions of apoptotic cells, for two phenotypic classes. We notice that the model

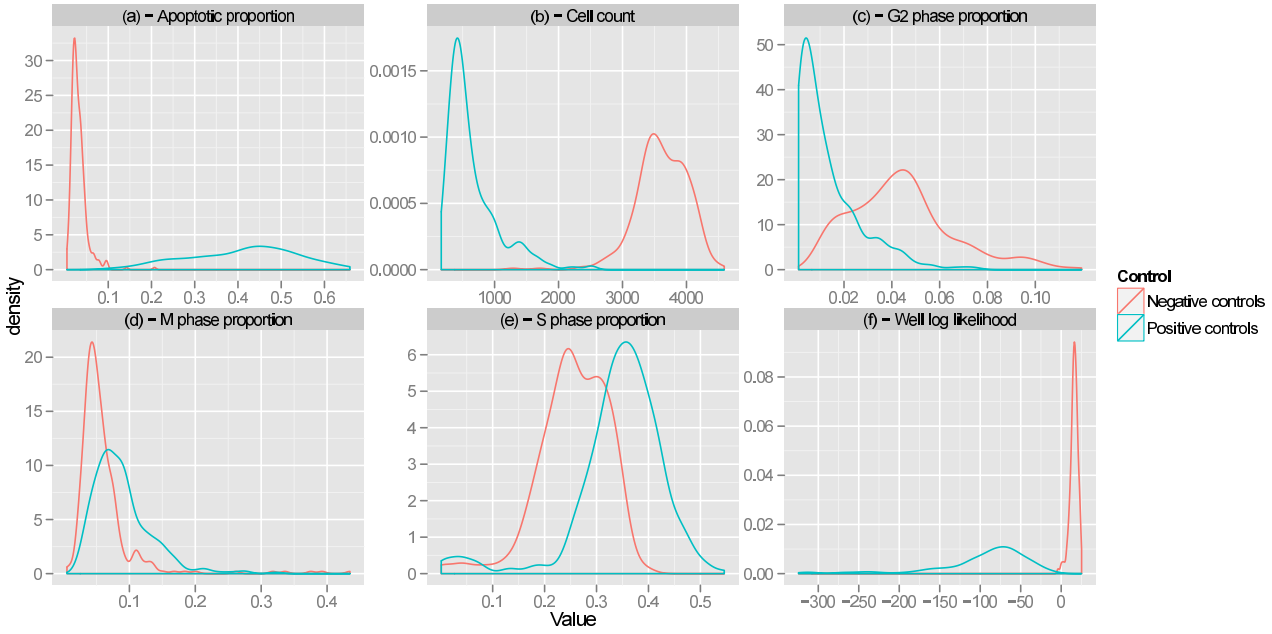



Figure 4.10: Novelty detection and positive controls. Density plot of cell population descriptors averaged over wells (panel **(a)** to **(e)**) and log likelihood (panel **(f)**) given by the model trained on negative controls. Positive controls are very different from negative controls. It is easy to distinguish them from negative controls only looking at cell count. The log likelihood given by the model separates the two type of controls. We observe that the discriminative power of the univariate descriptors is not lost when considering the model likelihood.

densities fitted by the copula model are closer to the empirical density compared to those fitted by the gaussian model. In addition, the parameters of the copula distributions represent physically valid distributions. For example, proportions of cells in apoptosis is higher than 0. As shown in Figure 4.11, the copula-based model accounts for this, while the gaussian model does not.

One example of use of the classes proposed by our model is the detection of atypical behaviours in the training set. Indeed, we inspected visually the cell images of negative control wells that were found in classes containing very few wells (3 classes with 5, 6 and 9 wells respectively over a total of 336). We found that 17 among these 20 wells were not relevant for the negative control modelling because they were

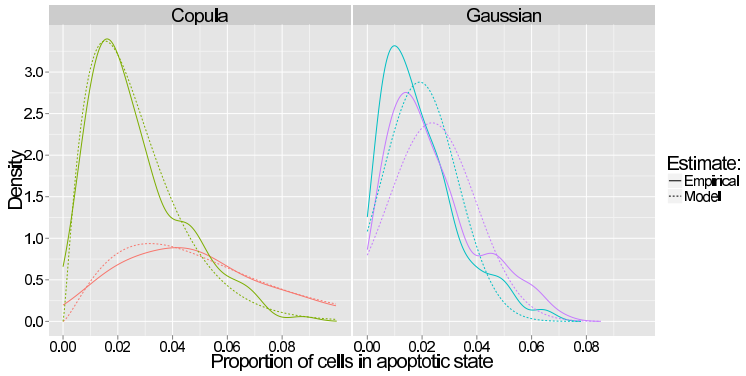



Figure 4.11: Model and empirical distributions. Examples of classes found by the model (Copula model on the left, gaussian model on the right). The proportion of cells in apoptotic state is represented for the cell populations belonging to these classes. We compare for two classes the univariate marginal densities. For each class, the empirical density is represented with a solid line and the density fitted by the model is represented with a broken line.

experimental outliers. These wells presented a recurrent atypical behaviour, and therefore, a few small classes were inferred to account for this during the learning procedure. Figure 4.12 shows bivariate scatter plots of the negative control fields and with these outliers. The proposed method provides clues for detection of such cases.

Moreover, we observed that the other classes containing a higher number of wells could account for experimental variability over cell populations. For example some particular classes contained mainly fields in which cell populations had reached confluence, while others did not, as we could observe in the corresponding images. All the classes do not necessarily account for biologically interpretable differences, because the diversity of cell population showing the same behaviours may require several classes to model it accurately. The number of classes was inferred based on cross validation generalization accuracy which is a much more objective criterion.

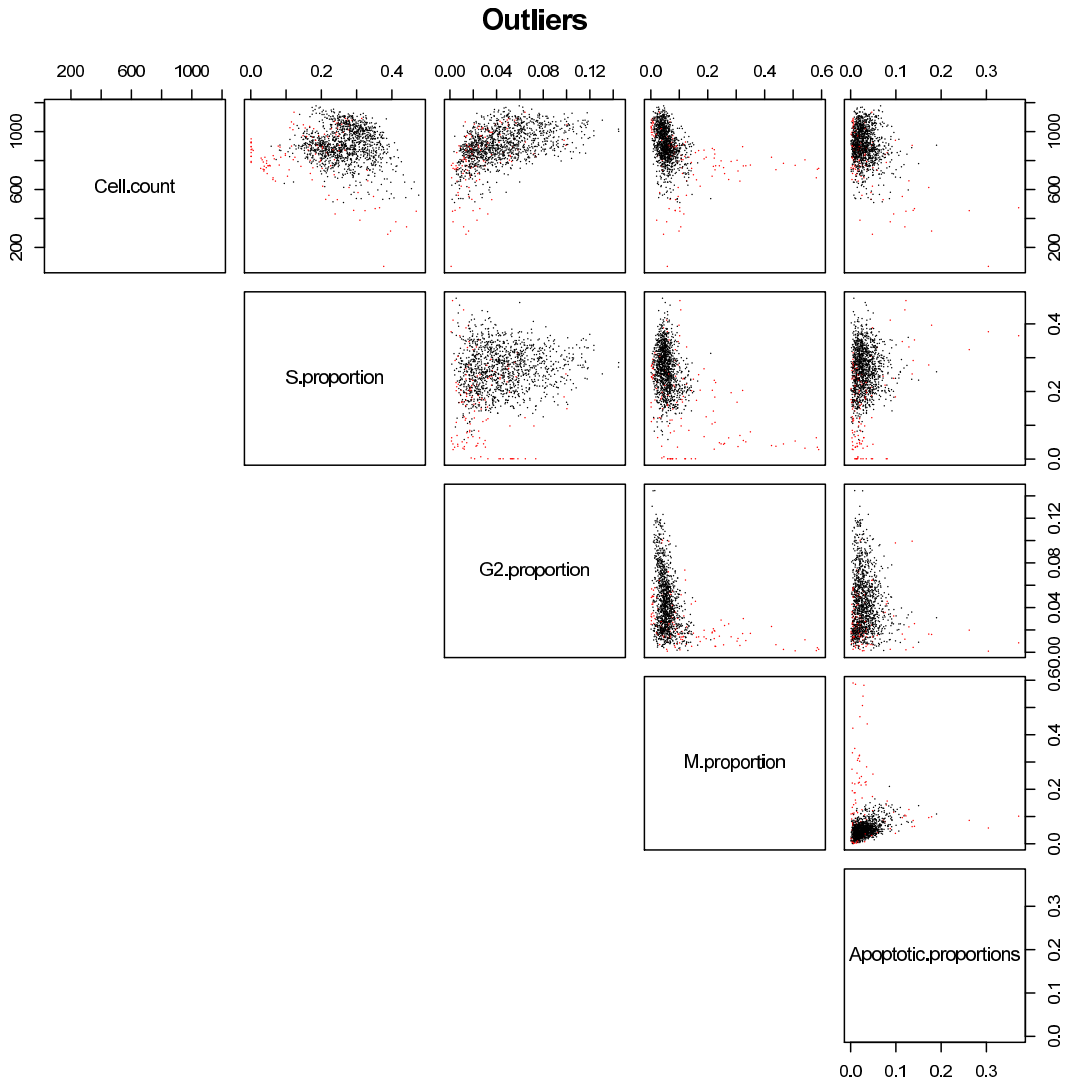



Figure 4.12: Negative control outliers. Bivariate scatter plots of negative controls. The red points correspond to fields belonging to small classes. They were indeed considered as outliers after checking the images (they were found to be irrelevant). Enough of these wells were present in the dataset so that separate classes were inferred by the model to account for this atypical behaviour.

### Relation between population phenotypes (classes) and cellular phenotypes

We considered negative controls and removed the outlier classes since, as mentioned above, they corresponded to irrelevant fields. We inspected differences between remaining classes based on the population descriptors (which were defined from cellular

phenotypes), because this could provide some clues about the biological interpretation of population phenotypes. Figure 4.13 represents the field distribution of population descriptors for each class. It shows that each population descriptor can separates some of the classes, but that none of the descriptors separates all of the classes on its own. This suggests that there is no redundancy between the population descriptors and that the classes reflect possible combinatorial association between population descriptors. The multivariate character of the proposed model allows to account for this fact, while it would not be possible using each population descriptor individually.

**Accounting for "within population variability"**

HCS experiments do not provide an average behaviour characterization, but a whole panel of cell responses within different sub populations (fields) taken from the same well. This information is much richer than a simple average response. The data account for the variability of the responses within a given population. As observed in section 4.3.2, this variability is not the same as the global field variability. The hierarchical structure of the model allows to take this into account which cross validation suggested to be a correct modeling assumption. Indeed, since all fields of a given well correspond to the same experiment, we therefore impose that they belong to the same phenotypic class. The corresponding density must account for the observed variability between those fields.

We illustrate this point in Figure 4.14 which compares one particular negative control well with the whole set of negative controls. Vertical red bars represented in Figure 4.14 show that population descriptors averaged over wells do not account for field variability (see Figure legend). Looking at panel **(a)** to **(e)** and blue vertical lines, the well looks similar to the majority of the negative controls. This would correspond to the single descriptor averaged over wells approach. However the red bars in those panels show that there is a lot of variation between the fields taken from this well, and some fields actually fall in tails of the distribution. This is reflected in the **(f)** panel where the vertical blue line is close to the tail of the distribution. Thus the methods could help to eliminate a potential experimental bias while a simpler approach would not.

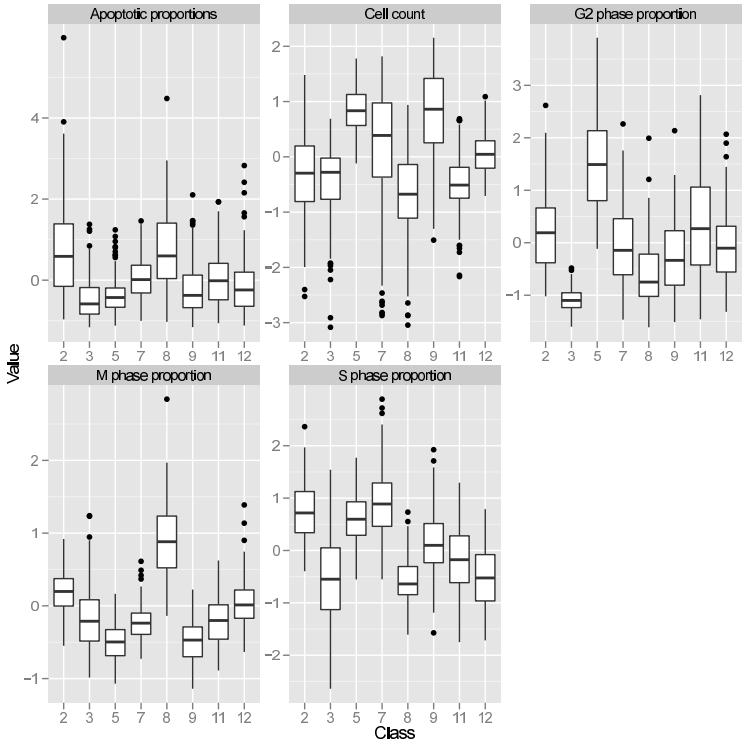



Figure 4.13: Relation between classes and population descriptors. The classes are represented on the $x$ axis. For each class, the boxplot shows the distribution of population descriptors among the fields of this class. Outlier classes were removed. The cell count descriptor has a similar distribution for classes 2 and 3, but other descriptors also allow to differentiate them. Similarly, classes 3 and 5 have a similar proportion of apoptotic cells, but other descriptors also allow to differentiate them. More generally, each descriptor separates different classes. This suggests that there is no redundancy between population descriptors, and that the classes reflect the combinatorial association between population descriptors.

### 4.3.4 Conclusion and discussion

In this section, we tackled the cell population phenotyping step in the HCS data analysis framework. This step is performed after image segmentation and cellular phenotyping (steps 1 and 2). It aims at comparing experiments, and gathering cells with similar behaviours in the same class (i.e. assigning them to the population phenotype). The main difficulties in achieving this task are linked to "dependence

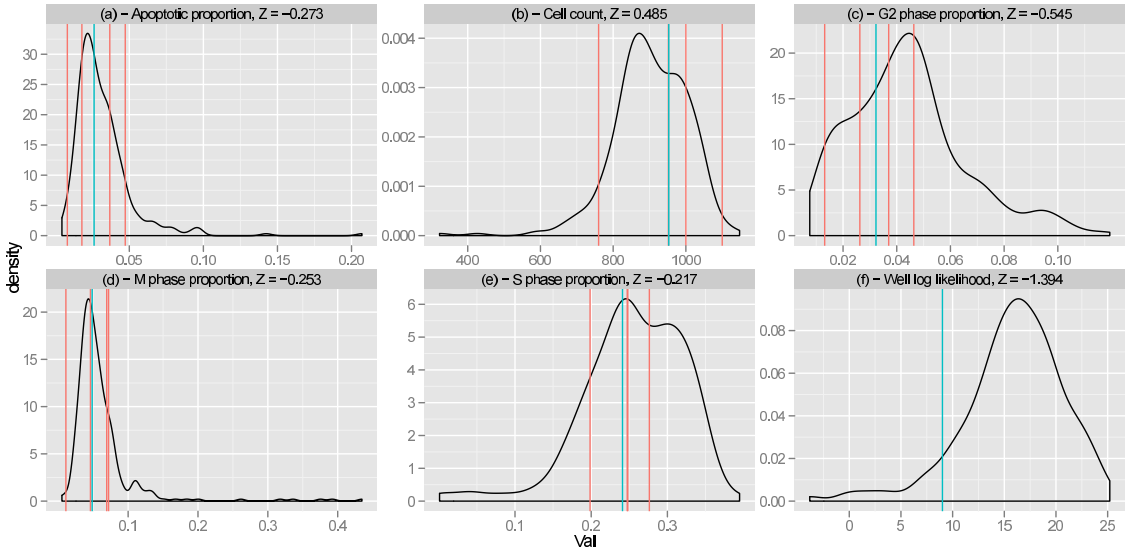



Figure 4.14: Example of a well. **Panels (a) to (e)**, the density plots represent the distribution of cell population descriptors averaged over wells for the negative control dataset. Red lines are the values of the 4 fields of the considered well and the blue lines are the population descriptors averaged over the 4 fields. **Panel (f)** represents the density of the log likelihood for all negative controls. The blue vertical line represents the log-likelihood of the considered well.

structure of population descriptors" and "within-population variability" which should be taken into account. Simple observations showed that these are naturally occurring facts observed in our HCS data. The necessity to take them into account precisely come from the fact that we are considering populations of cells as systems which are groups of smaller species (cells) and which organization is part of the characterization of the behaviour at the population level.

We implemented and compared the performances of two different parametrization of a mixture model, and baseline model that does not account for the specific aspects of the data underlined above. This was performed based on a dataset comprised of two types of cell populations. A comparison of model fitting on test data, using cross validation, suggest that the two specific aspect of the data we focused on when building the model should be considered when studying this kind of data.

Moreover, the copula-based parametrization of the proposed model outperforms the gaussian parametrization. However this copula-based model has some disadvantages from the computational point of view, model fitting being much slower and requiring approximations compared to the gaussian formulation.

The main features of cell populations that the model is able to describe are:

- Univariate variables (cell count or cellular phenotype proportions in our case), described by parametric densities
- Multivariate dependence structure, described by a copula
- Variability within a cell population, described by the hierarchical structure of the mixture model

These features constitute the specificity of HCS data. The proposed model takes them into account to build a phenotypic characterization at the population level. Cross validation experiments suggest that taking into account these aspects of the data provides better models. The literature is very scarce regarding population phenotypes definition. To our knowledge, none of the proposed methods take into account the "within-population variability", which underlines the originality of the proposed model. Pushing this idea further, a future line of work includes modelling at the cell level. [111] propose to infer cell classes from HCS data using single cell measurement. A future work direction is to add a level in the model to infer cell phenotypic classes and population phenotypic classes at the same time in a global model. However the inference computational cost increases a lot and online inference should be used such as in [52].

One application of this model is novelty detection, which is measuring how a cell population related to a given experimental condition is different from a control population. Once a control density is estimated, one can attribute a likelihood to each test experiment which allows to rank them according to how different they are from the controls. For example, the model can detect which siRNA phenotypes are different from a set of controls, and provide orientations toward the most relevant wells in a set of test experiments. The present work constitutes a preliminary validation of this procedure based on two limit cases.

Moreover, the method can help gathering cell populations that show similar behaviours into phenotypic classes. We observed that it can be useful for detection of irrelevant pictures gathered in separate phenotypic classes. The most important future work is to assess to which extend the inferred phenotypic classes are biologically meaningful. For example, wells in which siRNAs target genes with similar biological functions or incubated with drugs with the same target should belong to the same phenotypic class. Future work also include application of the model to target identification. This would require further experimental study for the validation of potential target genes.

# Chapter 5

# Dynamical system parameter identification under budget constraints

## Résumé

Le point de vue adopté dans ce chapitre se situe à l'interface entre l'échelle moléculaire considérée au chapitre 3 et l'échelle phénotypique présentée au chapitre 4. Nous nous intéressons ici aux comportements dynamiques émergeant d'intéractions non linéaires entre des espèces moléculaires. Ces propriétés dynamiques sont cependant trop abstraites pour être traitées comme des phénotypes. Le problème considéré est l'identification de paramètres d'un système dynamique et la construction d'un plan d'expérience afin de faciliter cette identification. Tous les résultats présentés se basent sur des simulations numériques, la stratégie d'identification ainsi que la dynamique moléculaire d'un réseau d'interactions sont simulés afin de comparer différents choix stratégiques. Nous discutons l'importance de définir des critères numériques bien fondés dans le contexte de la recherche du meilleur plan d'expérience. Une stratégie générale pour attaquer ce problème est présentée et une implémentation numérique de cette stratégie est proposée. La comparaison de différentes approximations numériques est donnée ainsi que les résultats du challenge DREAM 7 "Network

Topology and Parameter Inference" pour lequel la méthode a été conçue initialement. Le contenu de ce chapitre n'a pas encore été publié.


# Abstract

The point of view adopted in this chapter lies between the molecular scale treated in chapter 3 and the phenotypic scale presented in chapter 4. We are concerned with dynamical behaviours emerging from non linear interactions between molecular species. These dynamical properties are too abstract to be considered as phenotypes. The problem is to identify unknown dynamical parameters and to design experiments that make this identification more efficient. All the results presented are based on numerical simulations, both the identification strategies and the molecular dynamics of a relatively small interaction network are simulated to compare different strategical choices. We discuss the necessity to provide well defined numerical criteria in order to optimize experimental design. A general strategy to tackle this problem is presented and a numerical implementation of this strategy is proposed. Comparison of different numerical approximation is provided as well as results from the DREAM 7 Network Topology and Parameter Inference Challenge results for which this method was initially designed. The material presented in this chapter has not yet been published.


## 5.1 Introduction

Systems biology emerged a decade ago as the study of biological systems in which interactions between relatively simple biological species (genes, proteins, metabolites ...) lead to overall complex behaviours [74]. Such studies require to specify network structure and dynamical models. Typical descriptions of the dynamics of the system can be stochastic or deterministic [130]. Both descriptions involve unknown parameters. This motivates the design of methods that optimize the choice of experiments to be conducted in order to estimate unknown parameters from data [81]. Sequential methods constitute a promising line of research for these questions [62], a problem which is very close in its formulation to the active learning problem [107].

Various methods have been proposed for sequential experimental design in systems biology [62, 122, 9, 10, 116]. Most of them only take into account local uncertainty about parameters true values, *i.e.* strong multi-modality of the objective function is not considered. As pointed out in the literature, this is a very challenging inverse problem and it might even be pointless from a dynamical point of view [97], or from a biological point of view [34] when applied to real data. Indeed, it is shown in [97] based on numerical simulations on molecular biology dynamical systems that widely different parameter values produce very similar dynamical behaviours. Moreover [34] show based on real time course data that using inverse problem methods to estimate a unique parameter vector produces values which are biologically meaningless.

### 5.1.1 Evaluation of experimental design strategies

The experimental design problem considered in this chapter is broader than the parameter estimation problem. It relies on the ability to estimate parameters accurately. In addition, the design of experiments requires to figure out how further experiments could mitigate uncertainty and under-determinedness of the system. The performances of a strategy regarding this problem are difficult to evaluate. Moreover, different strategies might perform differently on different dynamical systems. It is therefore crucial to be able to evaluate, simulate the experimental design process in order to choose a relevant strategy when faced with a problem of this type.

As a consequence, proposing a strategy that could be reproducibly evaluated based on simulations requires the design choices to rely on well defined numerical criteria. The design strategy proposed in this chapter relies on numerical approximation of a unique numerical criterion. Therefore, it is possible to carry out simulation of the design process and compare different numerical approximation schemes based on this criterion.

### 5.1.2 Proposed strategy

This work is motivated by DREAM7 Network Topology and Parameter Inference Challenge [1, 2] which focuses on ordinary differential equation parameters estimation

from *in silico* experiments with budget constraints (*i.e.* limit of the quantity of data one has access to). This challenge is a simulation of an experimentation process in the context of systems biology. One has to estimate dynamical parameters of a system and can require to perform experiments on this system with different costs. The objective is to estimate the parameters as well as possible given a fixed budget by choosing sequentially experiments to be performed on the system to get new data. We propose a strategy adapted from active learning paradigm [100] which can be cast in the Bayesian framework of [81] for experimental design. This formulation has two appealing conceptual advantages:

- We provide a unique numerical criterion to discriminate between several experiments
- Bayesian paradigm relies on distributions over parameter space instead of a single value

The motivation behind the first idea is to construct a method that can be completely automatized without any human intervention. The focus is not on being optimal for a specific problem instance, rather on the generality of the method to make it easier to transpose to other problems and to make it potentially usable on large scale problems which sizes do not allow to deal with systematic human intervention. The second idea is motivated by the remarks of [97, 34] which were presented in the beginning of this introduction.

### 5.1.3 Content of the chapter

DREAM7 challenge was the occasion to implement this formulation and compare it to other methods. In addition, we implement a fully automatized simulation of the experimental design process. This allows to make extensive comparisons between different space search procedures and to compare with a blind random design strategy. We underline that this is a necessary condition to provide reproducible comparison between different strategies. As our results suggest, the proposed strategy reproducibly outperforms random design. Moreover, exploring several modes of the

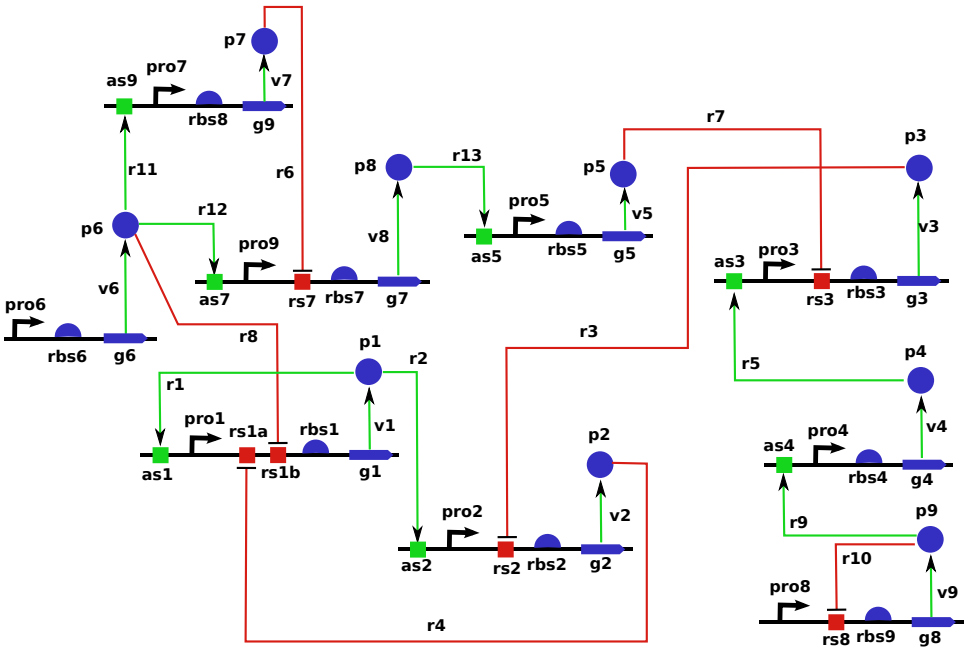



Figure 5.1: Gene network for DREAM7 Network Topology and Parameter Inference Challenge. Promoting reactions are represented by green arrows and inhibitory reactions are depicted by red arrows.

objective function has a large impact on the performance leading to results that are more reproducible.

To summarize, we describe an implementation of a general scheme that could in theory be applied to other dynamical system parameter estimation problems under budget constraints. The procedure can be completely automatized which allows to compare different strategies and verify the reproducibility of the results. Finally, we question the approach of parameter inference and present a slightly different perspective on the problem of dynamics characterization.

## 5.2 Methods

### 5.2.1 In silico network

We took part in the first challenge of DREAM7 Network Topology and Parameter Inference Challenge and used the system provided to carry out our experiments. The network is composed of 9 genes. We were provided with the complete network structure, including expression of the kinetic laws governing the dynamics of this network for which parameters had to be estimated. For each of the 9 genes, both protein and messenger RNA were explicitly modelled and therefore the model contained 18 continuous variables. Promoter strength controls the transcription reaction and ribosomal strength controls the protein synthesis reaction. Decay of messenger RNA and protein concentrations is controlled through degradation rates. A complete description of the underlying differential equations is found in appendix B. The complete network description and implementations of integrators to simulate its dynamics are available from [2]. A picture of the network is provided in figure 5.1. Promoting reactions are represented by green arrows and inhibitory reactions are depicted by red arrows.

Various experiments can be performed on the network producing new time course trajectories in unseen experimental conditions. An experiment consists in choosing an action to perform on the system. The possible actions are

- Nothing (Wild type)
- Delete a gene (remove the corresponding species).
- Knock down a gene (increase the messenger RNA degradation rate by ten folds).
- Decrease gene ribosomal activity (decrease the parameter value by 10 folds).

These actions were coupled with observable quantities

- Messenger RNA concentration for all genes at two possible time resolutions.
- Protein concentration for a pair of proteins at a single resolution.

Purchasing data consists in selecting an action and an observable quantities. In addition, it was possible to estimate the constants (binding affinity and hill coefficient) of one of the 13 reactions in the system. Different experiments and observable quantities have different costs, the objective being to estimate unknown parameters as accurately as possible, given a fixed initial credit budget.

### Notations

Model parameters are denoted by $\theta \in \mathbb{R}^p$. $e$ denotes an experiment to be performed, *i.e.* choice of a specific perturbation. The choice of one parameter value $\theta$ and experiment $e$ leads to time trajectories which we denote by $Y(\theta, e)$. In our case, they are positive quantities since we consider concentrations of chemical species. In practice, we can obtain them using differential equation solvers. The underlying dynamical system does not play a significant role here and we consider that the only access we have to it is through $Y(\theta, e)$. In addition to perturbation, we need to choose an observable $o^{bs}$ from a set of observables $O$. The point here is that we cannot observe the whole system in one experiment, and in particular, we cannot observe the whole set of time trajectories $Y(\theta, e)$. We only have access to a subset of these trajectories discretized with respect to time at a given resolution. The choice of an observable $o^{bs}$ leads to an observations $o$ which is a noisy realization of a sub-part of the true underlying dynamical system $Y(\theta^*, e)$ where $\theta^*$ is an unknown parameter. The problem is to choose a series of experiments and observables in order to infer $\theta^*$ as well as possible, given the cost constraints.

### Model

$P$ denotes a likelihood model related to a single experiment. Roughly speaking, for a given parameter value $\theta$, experiment $e$ and observable $o^{bs}$, $P(o|\theta, e, o^{bs})$ reflects how well data $o$ fits to $Y(\theta, e)$. In our setting, the noise model was specified by challenge organisers and we took the corresponding likelihood to model data (heteroscedastic Gaussian noise which variance is an affine transform of the mean value, see the challenge web page for details). If we have $K$ experiments $e_1, \ldots, e_K$ and observables

$o_1^{bs}, \ldots, o_K^{bs}$ with corresponding observations $o_1, \ldots, o_K$, the joint likelihood function has the form:

$$P(o_1, \ldots, o_K|\theta, e_1, \ldots, e_K, o_1^{bs}, \ldots, o_K^{bs}) = P(o_1|\theta, e_1, o_1^{bs}) \times \ldots \times P(o_K|\theta, e_K, o_K^{bs})$$

(we assume independence of noise realizations). We denote by $\pi_0$ a prior distribution over parameter space $\Theta$. We define sequentially $\pi_K$ to be the posterior distribution over model parameters when we get data from experiments 1 to $K$,

$$\begin{aligned} \pi_K(\theta) &= P(\theta|o_1, \ldots, o_K, e_1, \ldots, e_K, o_1^{bs}, \ldots, o_K^{bs}) \\ &\propto \pi_0(\theta) \times P(o_1, \ldots, o_K|\theta, e_1, \ldots, e_K, o_1^{bs}, \ldots, o_K^{bs}) \\ &\propto \pi_{K-1}(\theta) \times P(o_K|\theta, e_K, o_K^{bs}). \end{aligned} \tag{5.1}$$

**Risk function**

We focus on parameter inference. Let $\theta^* \in \Theta \subseteq \mathbb{R}^p$ denote the true model parameters and $\theta$ another parameter value. We implemented the risk used by challenge organizers to evaluate parameter estimates, namely $r(\theta^*, \theta) = \sum_{i=1}^{p} \log(\frac{\theta_i}{\theta_i^*})^2$. We do not have any access to the true model parameters, but we can estimate posterior distributions related to it. Given such a distribution $\pi$, we define the expected risk at $\theta_0$ by $R(\theta_0, \pi) = \mathbb{E}_{\theta \sim \pi} r(\theta, \theta_0)$.

**Sequential experimental design**

The purpose of the proposed framework is to choose experiment $e$ and observable $o^{bs}$ in an optimal way, and iterate. The proposed methodology is very close in spirit to that proposed in [81]. However, the design is made sequentially in a greedy fashion, and not globally, because of high computational cost. Although the setting is slightly different, the proposed methodology can be viewed as an application of the strategy of [100] using a loss that is adapted to our setting. Formally, at each step $k$, we have the knowledge of $e_1, \ldots, e_k, o_1^{bs}, \ldots, o_k^{bs}, o_1, \ldots, o_k$ which defines a posterior distribution $\pi_k$ on the parameter space $\Theta$ as given in (5.1). Given $\theta_T$ parameter values, an experiment $e$ and an observable $o^{bs}$, we note $o \sim P_{\theta_T, e, o^{bs}}$ to define that $o$ follows the distribution

which density is $P(o|\theta_T, e, o^{bs})$ as given by our likelihood model. We denote by $\pi_o$ the posterior probability over parameter space given the observation $o$. Next perturbation $e_{k+1}$ and observable $o^{bs}_{k+1}$ are chosen such that:

$$(e_{k+1}, o^{bs}_{k+1}) = \arg\min_{e,o^{bs}} \mathbb{E}_{\theta_T \sim \pi_k} \mathbb{E}_{o \sim P_{\theta_T, e, o^{bs}}} R(\pi_o, \theta_T)\,. \tag{5.2}$$

where $\pi_o(\theta) \propto \pi_k(\theta) \times P_{\theta,e,o^{bs}}(o)$. This formulation follows the principle of integrating out what is unknown, namely the true parameter value and the noise. If different experiments and observables have different cost $C_{e,o_{bs}}$, then we can choose the combination with the largest marginal expected risk reduction. In this case, we have:

$$(e_{k+1}, o^{bs}_{k+1}) = \arg\max_{e,o^{bs}} \frac{\mathbb{E}_{\theta_T \sim \pi_k} \mathbb{E}_{\theta \sim \pi_k} r(\theta_T, \theta) - \mathbb{E}_{\theta_T \sim \pi_k} \mathbb{E}_{o \sim P_{\theta_T, e, o^{bs}}} R(\pi_o, \theta_T)}{C_{e,o_{bs}}}\,. \tag{5.3}$$

In both cases, the most important task is to estimate the expected risk for unseen experiments and observable combinations. Given an experiment $e$ and observable $o_{bs}$, in order to estimate the expected risk of this combination, we consider all $\theta_T \in \Theta$ and estimate the risk related to each of them if it was the true parameter value, integrating out noise. We then integrate out $\theta_T$ and choose the best combination. Choosing the next experiment according to this procedure would require to compute a triple expectation which is of course not tractable analytically. This formulation is general and provides an objective criterion that could be applied to many different problems. However, implementing this method requires very strong approximations where algorithmic details play an important role.

### 5.2.2 Implementation

We denote by $E_1, \ldots, E_N$, the set of possible future actions. In general, $E_i$ describes a perturbation $e$ of the system and an observable $o^{bs}$. At each iteration $K$, we have the knowledge of the past, $o_1, \ldots, o_K, e_1, \ldots, e_K, o^{bs}_1, \ldots, o^{bs}_K$. This defines posterior distribution over parameter space $\pi_K$. This distribution cannot be directly manipulated and needs to be approximated in some way. For this, we rely on sampling. The two building blocks of the method are sampling from $\pi_K$ given the past (a brief

introduction to sampling is given in chapter 2), and evaluating the risk of each sample point. Algorithm 1 summarizes the procedure to choose the next action. We note that the structure of the algorithm is very similar to what is proposed in literature (*e.g.* [62, 122, 9, 10, 116]), we first evaluate parameter uncertainty given the data at hand and then optimize the choice of the next experiment given this uncertainty. The specificity of the proposed scheme is that it takes advantage of the theoretical and practical tools available for Bayesian inference. Moreover, a unique numerical criterion is used to discriminate between experiments. This criterion is based on expected risk and, by construction, takes interactions between all components of the problem into account. An illustration of this fact is given in section 5.3.1.

**Algorithm 1:** NextExperiment

**input** : $o_1, \ldots, o_K, e_1, \ldots, e_K, o_1^{bs}, \ldots, o_K^{bs}, \{E_1, \ldots, E_N\}$
**output**: $e_{K+1}$

```
begin
    Θ ← sample(π_K)
    for e ∈ {E_1, ..., E_N} do
        for θ_T ∈ Θ do
            R_{θ_T} ← evalRisk(θ_T)
        R_e ← mean(R_{θ_T} | θ_T ∈ Θ)
    e_{K+1} ← choose({R_e})
```

### Practical details

We implemented the proposed strategy using the same noise model and risk function that were used to evaluate candidates in the DREAM7 Network Topology and Parameter Inference Challenge [1, 2]. This section describes approximations that were made in our implementation of the proposed strategy. As we mentioned in previous sections, the problem cannot be solved exactly and these approximations are needed to apply our strategy. They could be replaced by other approximations. The appealing specificity of what is proposed here is that one uses the same sample to estimate the current posterior and the expected risks of all possible actions at each step of

algorithm 1. We refer the reader to [5] for an introduction to sampling schemes and to [90] for a description of finite difference approximation and BFGS algorithm. A brief introduction to sampling approximations is given in chapter 2.

### Enforcing regularity through the prior distribution

Evaluation of parameter likelihood requires to solve several initial value problems. We used the implementation of the method proposed in [15] provided in the package [114]. The prior distribution $\pi_0$ plays a crucial role at early stages of the design when no data is available about many aspects of the system's behaviour. It penalizes parameters leading to dynamical behaviours that we consider unlikely. In addition to a large variance log normal prior, we considered penalizing parameters leading to non smooth time trajectories. This is done by adding to the prior log density a factor that depends on the total variation of time course trajectories. The advantage of this is twofold. First, it is reasonable to assume that variables we do not observe in a specific design vary smoothly with time. Second, this penalization allows to avoid regions of the parameter space corresponding to very stiff systems, which are poor numerical models of reality, and which simulation are computationally demanding or simply make the solver fail. This is very beneficial to guide the search in the sampling phase.

### Sampling from posterior distribution

The likelihood surface shows multi-modality, plateaus and abrupt jumps as illustrated in figure 5.2. Traditional sampling techniques tend to get stuck in local optima, not accounting for the diversity of high likelihood areas of the parameter space. Finite difference calculation allows to compute an approximation of the gradient of the log-likelihood function which is used together with BFGS algorithm to find local maxima of the posterior distribution. Finite difference approximation is known not to be the most stable numerical method for gradient computation. However, this is the method that provided us with the best trade off between computation, sample size, sample diversity and data fitting. In order to ease sampling, we use a local search method

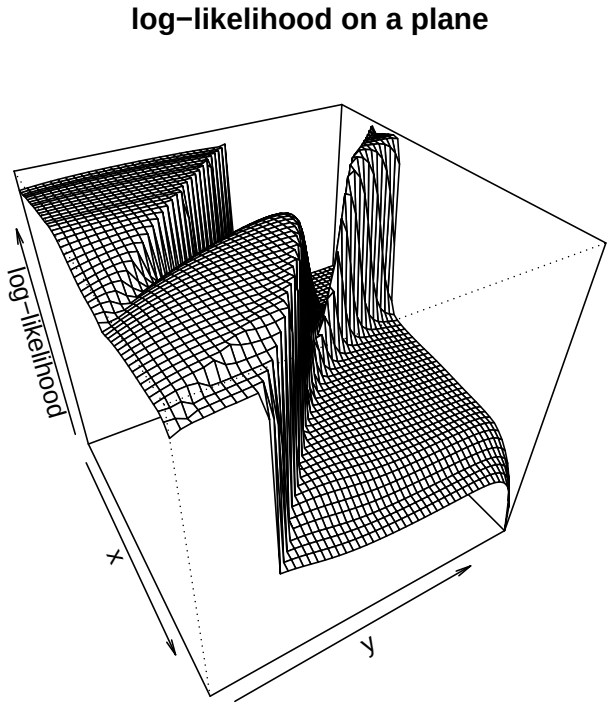



Figure 5.2: Log likelihood surface for parameters living on a restricted area of a two dimensional plane. For clarity, scale is not shown. Areas with low log-likelihood correspond to dynamics that do not fit the data at all, while areas with high log-likelihood fit the data very well. The surface shows multi-modality, plateaus and abrupt jumps which makes it difficult to sample from this density. When parameters do not live on a plane, these curses have even higher effect.

to provide an initial value for a Metropolis Hastings sampler (see also chapter 2). We combine isotropic Gaussian proposal and single parameter modifications. This strategy can be repeated several times to get samples from different modes.

### Estimating the expected risk

**Experiments that provide a time course output:** suppose that $e$ is a perturbation of the system and $o^{bs}$ is an observable. The risk related to this experiment is expressed as:

$$R = \mathbb{E}_{\theta_T \sim \pi_K} \mathbb{E}_{o \sim P_{\theta_T, e, o^{bs}}} \mathbb{E}_{\theta \sim \pi_o} r(\theta, \theta_T)$$

where $\pi_o(\theta) \propto \pi_K(\theta) \times P_{\theta,e,o^{bs}}(o)$. Suppose that we have a sample $\Theta$, of size $N$, drawn from $\pi_K$. Choosing the next experiment requires to compute $\mathbb{E}_{o\sim P_{\theta_T,e,o^{bs}}} \mathbb{E}_{\theta\sim\pi_o} r(\theta,\theta_T)$ for each $\theta_T \in \Theta$. For a fixed $\theta_T$, the output variable $o$ is random. The inner expectation can be approximated by importance weighting:

$$\mathbb{E}_{\theta\sim\pi_o} r(\theta,\theta_T) \simeq \sum_{\theta\in\Theta} w_o(\theta) r(\theta,\theta_T)$$

where $w_o(\theta) \propto P(o|\theta,e,o^{bs})$ (prior terms cancel out, see also chapter 2). It is important to normalize these weights correctly since dimensionality of different experiments will be different. Consider the term $\mathbb{E}_{o\sim P_{\theta_T,e,o^{bs}}}[w_o(\theta)]$. If we could compute this term, the risk would be approximated as:

$$R \simeq \frac{1}{N} \sum_{\theta_T\in\Theta} \sum_{\theta\in\Theta} \mathbb{E}_{o\sim P_{\theta_T,e,o^{bs}}}[w_o(\theta)]\, r(\theta,\theta_T).$$

The inner expectation is difficult to compute due to normalization of the weight. We evaluated it by drawing several outputs $o$ for each $\theta_T$, and tacking the average of the corresponding normalized weights.

**Estimation of model parameters:** we have the possibility to determine some parameters of the dynamical model. Suppose that the experiment consists in evaluating parameter $i$, the "posterior" probability distribution (the distribution which represents our uncertainty after seeing the result) becomes $\pi_{\theta_T}(\theta_{-i}) = \pi_t(\theta_{-i}, \theta_i = \theta_{Ti})$ (where the "minus" subscript indicates that we remove the $i$-th parameter). As in the previous section, suppose that we have a sample $\Theta$, of size $N$, drawn from $\pi_t$. The expected risk of this experiment becomes:

$$R = \mathbb{E}_{\theta_T\sim\pi_t} \mathbb{E}_{\theta_{-i}\sim\pi_{\theta_T}} r(\theta,\theta_T)$$

where $\pi_{\theta_T}(\theta) = \pi_t(\theta_{-i}, \theta_i = \theta_{Ti})$. Using importance sampling we now can approximate this risk as:

$$R \simeq \frac{1}{N} \sum_{\theta_T\in\Theta} \sum_{\theta\in\Theta} w_{\theta_T}(\theta) r(\theta,\theta_T)$$

where $w_{\theta_T}(\theta) \propto \frac{\pi_t(\theta_{-i}, \theta_i = \theta_{Ti})}{\pi_t(\theta)}$ (see also chapter 2). Weights should also be normalized properly in order to compare different experiments.

## 5.3 Results and discussion

### 5.3.1 Experimental results

#### Sub network

We considered testing the proposed method with a small sub-network. We took the same architecture as in figure 5.1 only considering proteins 6, 7 and 8. There are 6 variables which behaviour is governed by 13 parameters in this network. Since the method is based on random-walk type exploration of probability distributions, we simulated the design process 10 times with different pseudo-random number generator seeds and same initial credit budget. We compared sampling from a single mode and sampling from several modes as well as random and active experimental design. Sampling from several modes means that we combine samples from several Markov chains initialized using BFGS local search starting from different initialization points. The results are presented in figure 5.3 and figure 5.4. It appears that given the same sampling method (exploring several modes), our strategy leads to a better use of available credits to estimate parameters. Indeed, the risk of the estimated parameter value is significantly higher in the case of random design (see figure 5.4). Additionally, exploring only a single mode at each step leads to much more variables estimates, making it difficult to reproduce potential good results. The boxplot width is much larger in this case.

#### Dream challenge

DREAM7 challenge was the occasion to compare this formulation to other methods proposed by different teams. Table 5.1 presents the result of all participants. The evaluation was based on both parameter evaluation and prediction of protein time course in an unseen experimental setting. First, our method did not perform as well as other methods. As mentioned in the previouse paragraph, the sampling strategy

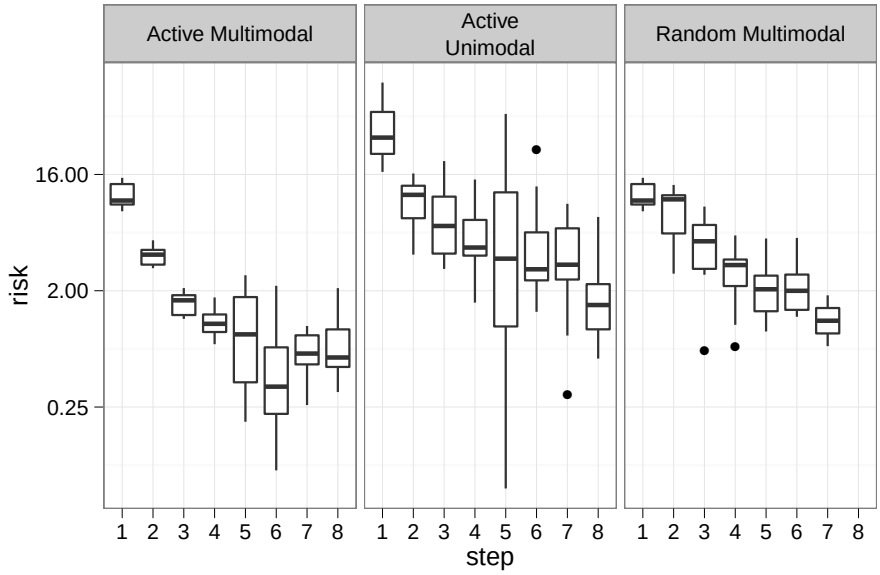



Figure 5.3: Comparison of risk evolution between different strategies on a subnetwork. The figure shows the true risk at each step of the procedure, *i.e.* the approximate posterior distribution is compared to the true underlying parameter which is unknown during the process. The risk is computed at the center of the posterior sample. The boxplots represent 10 repeats of the design procedure given the same initial credit budget. Active design is our strategy while random design consists in choosing experiments randomly. Multimodal means that we explore several modes, *i.e.* combine several Markov chain samples with different starting points. Unimodal means that we only consider one chain. The latest strategy leads to highly variable results. Our active strategy outperforms the random design, we choose experiments that leads to a better use of available credits, making it possible to perform more experiments at the end.

greatly affects the performances of the method. One of the reasons for this result is that we did not carefully tune the sampling and prior parameters when purchasing data during the challenge. Second, the ability to predict well dynamical behaviour does not seem to be strongly linked to the ability to infer parameters accurately. Additional comments on this are given in section 5.3.2. Figure 5.5 compares predicted and true trajectories in the unseen experimental setting used for the evaluation at

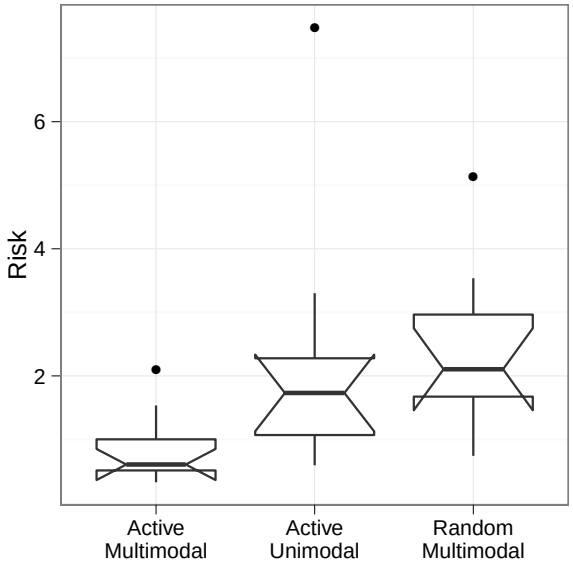



Figure 5.4: Comparison of final risks after all credit has been spent for different strategies on a subnetwork. The figure shows the true risk at the end of the procedure, *i.e.* the approximate posterior distribution is compared to the true underlying parameter which is unknown during the process. The risk is computed at the center of the posterior sample. The boxplots represent 10 repeats of the design procedure given the same initial credit budget. Active design is our strategy while random design consists in choosing experiments randomly. Multimodal means that we explore several modes, *i.e.* combine several Markov chain samples with different starting points. Unimodal means that we only consider one chain. It is clear from this plot that combining multimodal search and active design is beneficial.

the end of the challenge.

### An illustration of method behaviour

The first data we had at hand were low resolution mRNA time courses for the wild type (no perturbation of the system). The first experiments chosen by the method were wild-type protein concentration time courses. This makes sense since we have enormous uncertainty about proteins time courses because we do not know anything

| Team | Dparam | p-value | Dprot | p-value | Score |
|---|---|---|---|---|---|
| orangeballs | 0.0229 | 3.25E-03 | 0.0024 | 1.21E-25 | 27.40 |
| Team 387 | 0.8404 | 1.00E+00 | 0.0160 | 3.39E-18 | 17.47 |
| Team 93 | 0.1592 | 6.00E-01 | 0.0354 | 4.45E-15 | 14.57 |
| Team 509 | 0.0899 | 1.88E-01 | 0.0475 | 6.28E-14 | 13.93 |
| Team 374 | 0.1683 | 6.45E-01 | 0.0979 | 4.01E-11 | 10.59 |
| Team 197 | 0.0453 | 1.37E-02 | 0.1988 | 1.93E-08 | 9.58 |
| Team 202 | 0.1702 | 6.45E-01 | 0.3625 | 2.90E-06 | 5.73 |
| **Team 450** | **0.8128** | **1.00E+00** | **0.3564** | **2.53E-06** | **5.60** |
| Team 111 | 0.3766 | 9.99E-01 | 0.8180 | 1.34E-03 | 2.87 |
| Team 78 | 0.0699 | 9.83E-02 | 19.3233 | 1.00E+00 | 1.01 |
| Team 408 | 0.1883 | 7.29E-01 | 3.2228 | 6.90E-01 | 0.30 |
| Team 626 | 5.0278 | 1.00E+00 | 14.7744 | 1.00E+00 | 0 |

Table 5.1: Results of DREAM7 Network Topology and Parameter Inference Challenge. Two criteria were used to compare teams. The first one (column 2) is related to precision of parameter estimation. The second one (Column 4) is related to prediction of protein time course for an unseen experimental setting. p-values were computed using a bootstrap procedure. The global score aggregates both criteria.

about them. Once we have purchased these datasets we ran our procedure to determine what the next experiment should be. Interestingly, the perturbations with the lowest risk were related to gene 7 which is on the top of the cascade (see figure 5.1) as shown in table 5.2. Moreover it seemed obvious from table 5.2 that we had to observe protein 8 concentration. Indeed, figure 5.6 shows that there is a lot of uncertainty about protein 8 evolution when we remove gene 7. Moreover, our criterion determined that it was better to observe protein 3 than protein 5, which makes sense since the only protein which affects protein 5 evolution is protein 8 (see figure 5.1). Therefore uncertainty about protein 5 time course is tightly linked to protein 8 time course, and observing protein 3 brings more information than observing protein 5. This might not be obvious when looking at the graph in figure 5.6 and could not have been foreseen by a method that considers uncertainty about each protein independently. At this point, we purchased protein 3 and 8 time courses for gene 7 deletion experiment which results are available in figure 5.6.

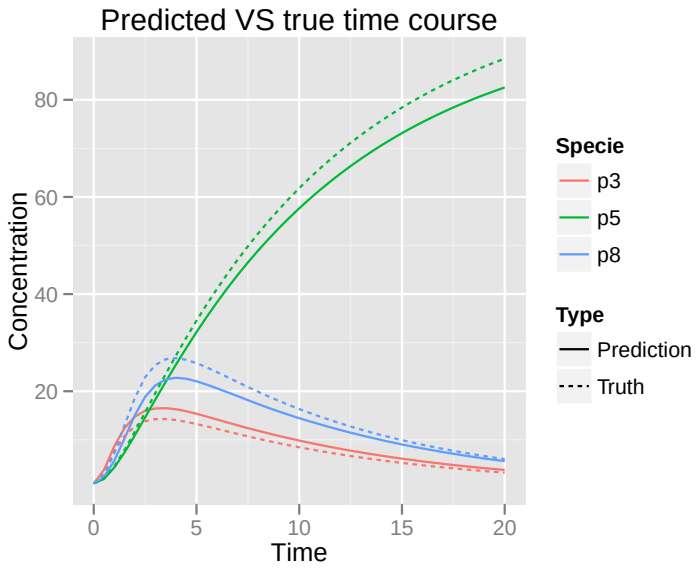



Figure 5.5: Comparison between the true time course trajectory and protein time course prediction. This corresponds to the second criterion for the evaluation of participants to the challenge (table 5.1).

#### Parameter and time trajectory variability

Figure 5.7 represents a sample from the posterior distribution after all credits had been spent and no further experiments could be conducted. Both parameter values and protein time course for the unseen experiment are presented. All parameters have been optimized using local optimization. Some parameter clearly concentrate around a single value while some other have very wide range with multiple accumulation points. Despite this variability in parameter values, the protein time course trajectories are very similar.

### 5.3.2 Discussion

#### Generalizability of the proposed framework

As stated in the introduction, one of the motivation for the approach we consider was to build a method that is not specific to one instance of the problem and could

| Risk | Cost | Experiment | Observe proteins |
|---|---|---|---|
| 771 | 1200 | Delete gene 7 | 3-8 |
| 1196 | 850 | Decrease gene 7 RBS activity | 3-8 |
| 1290 | 750 | Knock down gene 7 | 3-8 |
| 1957 | 850 | Decrease gene 7 RBS activity | 3-7 |
| 2254 | 850 | Decrease gene 7 RBS activity | 7-8 |
| 2554 | 1200 | Delete gene 9 | 3-8 |
| 2867 | 750 | Knock down gene 7 | 8-9 |
| 4647 | 1200 | Delete gene 7 | 8-9 |
| 4798 | 850 | Decrease gene 7 RBS activity | 8-9 |
| 4928 | 850 | Decrease gene 7 RBS activity | 5-8 |

Table 5.2: Estimation of the expected risk at a certain stage of the experimentation, ten lowest values. There is consistency in the type of experiment to be conducted (targeting gene 7 which expression impacts on a big part of the network) and the quantities to measure (protein 8 almost all the time and protein 3 quite often). Figure 5.6 illustrate this point further.

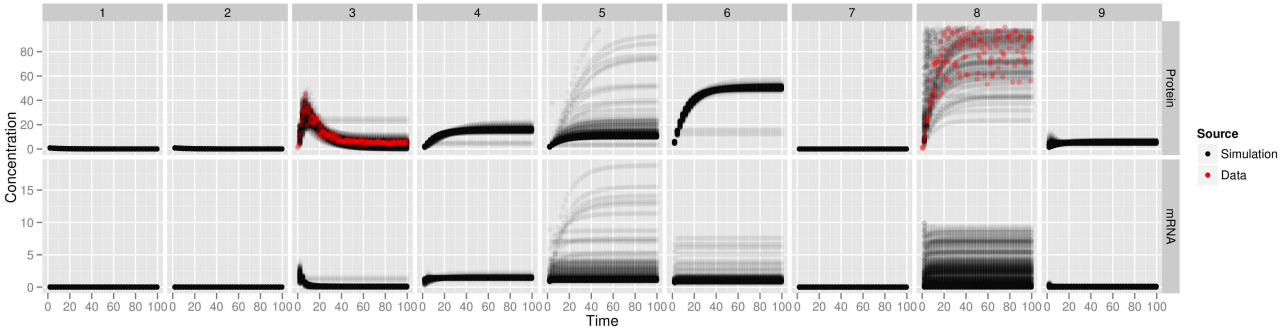



Figure 5.6: Corresponds to table 5.2 figures. We plot trajectories from our posterior sample (protein 8 concentration was divided by 2 and we do not plot concentrations higher than 100). The quantities with the highest variability are protein 8 and 3 concentrations. This is consistent with the estimated risks in 5.2. There is quite a bit of uncertainty in protein 5 concentration, however this is related to protein 8 uncertainty as protein 8 is an inhibitor of protein 5. Moreover, mRNA concentration have much lower values and are not as informative as proteins concentrations. Red dots shows the data we purchased for this experiment after seeing these curve and in accordance with results in table 5.2.

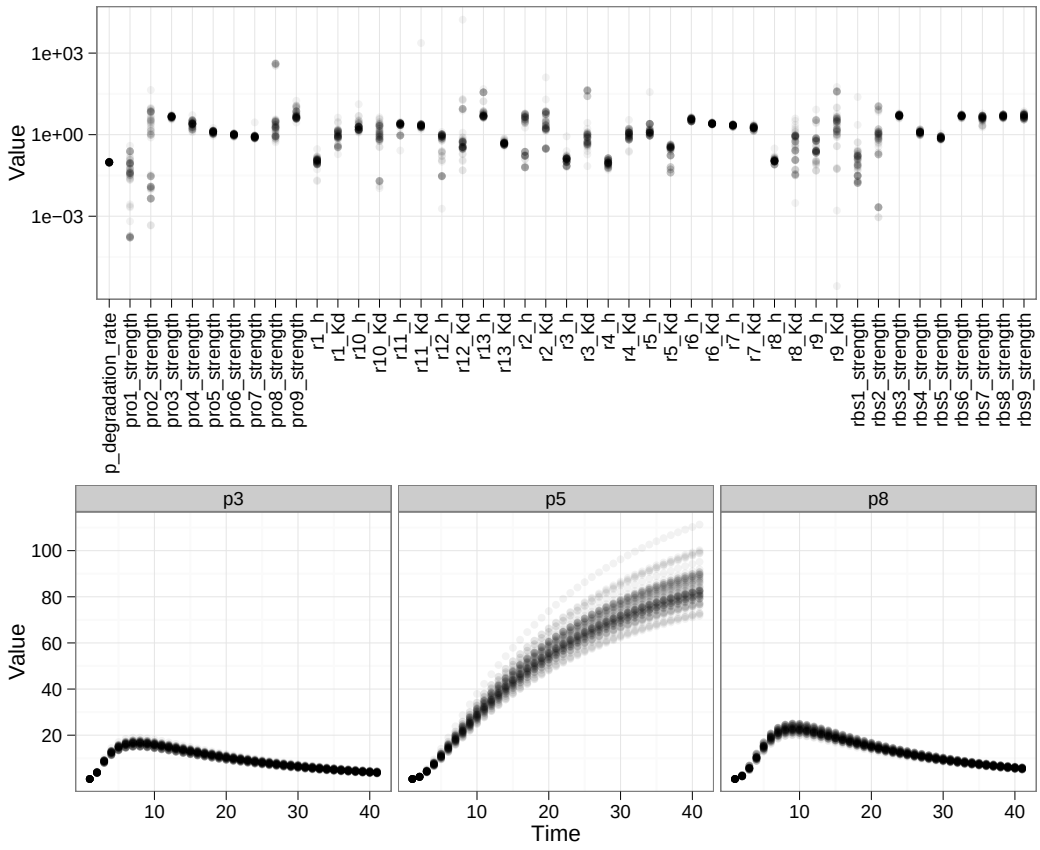



Figure 5.7: Comparison of parameter variability and time course trajectory variability. This is a sample from the posterior distribution after spending all the credits in the challenge. The top of the figure shows parameter values on log scale, while the bottom shows prediction of protein time courses for an unseen experiment. The range of some parameter values is very wide while all these very different values lead to very similar protein time course predictions.

handle different structures using the same tools. A good experimental design strategy should perform well on large variety of dynamical systems and not rely on specific assumptions that are valid only for a few network structures. This is in our opinion a key point in order to be able to deal with large scale networks of size comparable that of a cell for example. Moreover, the experimental design strategy should produce results that are reproducible.

In order to be able to reproduce experimental results related to the experimental design problem (for example to compare between different strategies), one needs to make decision based on a well specified numerical criterion, which could aggregate several criteria, not requiring human subjectivity. For DREAM7 challenge, our strategy did not compare well to other strategies, as results presented in table 5.1 suggest. Indeed, the strategy relies on sampling methods which should be tuned to perform well on specific problems. Comparison of different sampling strategies was done a posteriori. For the challenge, we used a sampling strategy that did not give results with accurate reproducibility.

Our method would theoretically handle more complex systems, the limit being the computational cost of simulating larger networks and exploring higher dimensional parameter spaces. Different strategies for parameter space exploration and uncertainty evaluation could be investigated for larger networks. We chose to use samples from the parameter space because of the high multimodality of the likelihood function. A common approach in active design is to use single parameter value and to measure dispersion of the likelihood around this point of the space. Our experiments suggest that this approach does not lead to consistently reproducible results.

**Parameter estimation versus network behaviour inference**

As observed from the results of different challenge participants in table 5.1, better parameter estimates do not necessarily lead to better reproduction of the original system's behaviour. Moreover, in our setting, as posterior distributions do not concentrate strictly around a single mode, many different parameters from different areas of the parameter space lead to similar fit to the data. This was illustrated in section 5.3.1 (see also figure 5.7) and has already been observed on similar numerical simulations in [97]. These facts question the idea of single parameter estimation and highlight the importance of reproducing the original dynamical system behaviours compared to parameter estimation. Such observations are based on computer experiments where the data is 'clean', meaning that the noise model and dynamics are well specified. The effect highlighted here might be even more important in the study of

real dynamical systems where data is not generated from a model. Indeed, experiments conducted in [34] applying inverse methods to inference of kinetic parameters based on real data suggest that finding a single parameter value is out of reach and leads to biologically unrealistic parameter values. They point out the need to address this problem. The Bayesian framework is an elegant alternative replacing the single parameter strategy by probability distributions over parameter space. Its main drawback lies in higher computational cost. The natural research direction suggested here is the design of efficient Bayesian inference methods for this specific kind of setting.

## 5.4 Conclusion

Computational systems biology is a promising line of research based on the heavy use of computational resources to improve the understanding of the complexity underlying cells biology. The most widespread approach is to specify a dynamical model of the studied biological process based on biochemical knowledge, and consider that the real system follows the same dynamics for some kinetic parameter value. Recent reports suggest that this has benefits in practical applications (*e.g.* [68]). Systematic implementation of the approach requires to deal with the fact that some kinetic parameters might be unknown. This raises the issue of estimating these parameters from experimental data as efficiently as possible. An obvious sanity check is to recover kinetic parameters from synthetic data where dynamic and noise model are well specified. This is already quite a challenge. The method we proposed takes advantage of the Bayesian framework to sequentially choose experiments to be performed, in order to estimate these parameters subject to cost constraints. The method relies on a single numerical criterion and does not depend on specific a specific instance of this problem. Experimental results suggest that the strategy works better than random experimental choice, even though it is not optimal on specific networks where more fine tuned strategies improve accuracy of kinetic parameter inference. We evidenced the mechanisms underlying these observations.

The approach focusing on kinetic parameter estimation is questionable. We give empirical evidences similar to these of [97] that raises the issue of well posedness of

this approach showing that very different parameter values could produce very similar dynamical behaviours, potentially leading to non-identifiabilities. Moreover, focusing on parameter estimation supposes that the dynamical model represents the true underlying chemical process. In some cases, this might be false. For example, hypotheses underlying the law of mass action are not satisfied in the gene transcription process. However, simplified models might still be good proxies to characterise dynamical behaviours we are interested in. The problem of interest here is to reproduce dynamics of a system in terms of observable quantities, and to predict the system behaviour for unseen manipulations. Parameters can be treated as latent variables which impact the dynamics of the system but cannot be observed. In this framework, the Bayesian formalism described here is well suited to tackle the problem of experimental design.

The natural continuity of this work is to adapt the method to treat larger problems. This raises computational issues and requires to develop numerical methods that scale well with the size of the problem. The main bottlenecks are the cost of simulating large dynamical systems, and the need for large sample size in higher dimension for accurate posterior estimation. Promising research directions are parameter estimation methods that do not involve dynamical system simulation such as [18] or differential equation simulation methods that take into account both parameter uncertainty and numerical uncertainty such as the probabilistic integrator of [23].

# Chapter 6

# Conclusion

The common denominator to the studies presented in this thesis is the use of computational statistics for studying biological phenomena. The main motivation behind the use of such methods is the complexity of the underlying tasks. The example of unsupervised correlation based methods and model based clustering methods given in chapters 3 and 4 show that it is possible to use such methods in order to extract relevant information. Here, the complexity of the task is related to the quantity of data that renders non automatized inspection prohibitive. In chapter 5, the complexity of the task stems from highly non linear behaviour of the studied system. In these cases computer aided methods are essential to tackle the underlying question.

Computer aided biology dates back to the middle of the nineties and has now specialized into several specific biology related tasks such as data management and interpretation, technology specific workflows or genomic medicine [92]. Most of the biomedical research currently heavily relies on the integration of technological plat forms, database knowledge, biology and computation. This makes biology one of the largest field of application of computational statistics. The increasing complexity of tasks related to biological research suggests that both fields will remain intimately linked in the future.

Extension of the proposed methods are related to the challenges that the field has to face for the next years. At the molecular scale, more and more information about chemical and biological assays of various nature are publicly available. Molecular

representation is the key issue in prediction. Indeed, poor data representation sets the best performance achievable by any prediction algorithm at a very low level. Stacking information from many sources can help improve the representation of small molecules. Among the relevant related questions are:

- How to integrate heterogeneous molecular information in a scalable way?
- Does this integration improve performances over specific prediction task?
- Does richer representation systematically leads to better prediction performances?

The last question is of interest beyond the field of computational chemistry and chemogenomics. Indeed, motivated by the explosion of the quantity of data available, a general trend in many related fields, is to integrate as many sources of information as possible regardless of their relevance to the task at hand. A natural question is whether this is a good idea or not. Indeed stacking sources of information increases the likelihood of including the most relevant information in the set of available sources but also diminishes the ability of anyone to distinguish between relevant and irrelevant information for a specific task. It might be needed to consider how information sources are related to the physical reality of the studied process. In the field of computational chemistry for example, an interesting research direction is to build and evaluate a similarity measure between molecules that reflect the underlying chemistry. Using expert systems such as in [69] could for example allow to predict the chemical complexity of turning one specie into another.

Building generative models based on cell imaging assays is becoming a popular method to explore and summarize results of fluorescent microscopy cell imaging experiments (for example [138]). Generative models allow to represent the hierarchy of scales explicitly in their structure. Such models are very popular in unsupervised modelling of text corpora [14]. Using such hierarchy in the context of fluorescent cell microscopy could allow to perform phenotypic inference jointly at different levels of the phenotypic scale and to share information between different scales in a well defined manner.

As for dynamical systems as considered in chapter 5, the main bottleneck is the high computational cost of simulating the system accurately. However, given that

the system parameter values are also uncertain, one may not need to perform very accurate simulations in order to perform accurate inference. A very fruitful area of research in machine learning is the investigation of tradeoffs between statistical and computational accuracy. Understanding these tradeoffs leads to massive saving in computational requirement of statistical tasks [16]. Similar tradeoffs might exist when performing inference based on undetermined control problems such as the one considered in chapter 5. At the heart of these tradeoffs lies the question of propagating uncertainty in highly non-linear systems. Efficient methods to achieve this task will probably increase significantly the size of problems of the type of the one considered in 5 that can be tractably handled.

The results presented in this thesis are based on numerical experiments. We demonstrated that the output of these experiments were relevant for the related biological question. These manipulations suggest that the methods described in this thesis could be used for field applications. However, we did not explicitly provide examples of biological discoveries made possible thanks to them. This would be the ultimate goal, a longer term challenge. The only way to actually show that the proposed approaches are useful is to implement them as part of the tools available for projects which ultimate objective is to solve a biological problem. The time scales for this types of projects much wider.

Lastly, a critical point for the usefulness of a method is its computational tractability. This is most striking in the work presented in chapter 5 where the amount of computation required to implement the method on a medium scale problem was considerable. When designing a protocol to solve a biology related problem, one is provided with state-of-the-art algorithmic tools that come from different fields such as computer science, optimization or control. Understanding how efficiently a problem can be solved on a computer and proposing efficient algorithms for abstract problem, in addition to be interesting questions by themselves, directly influence the ability of solving numerical problems. Among others, the field of computer aided biology benefits directly from these advances. Considering the need to adopt systems point of views to improve the understanding of complex biological systems behaviours, algorithmic and complexity questions are keys for future biological advances.

# Bibliography

[1] The dream project website, available at http://www.the-dream-project.org/, sep 2012.

[2] Dream7 network topology and parameter inference challenge website, available at http://www.the-dream-project.org/challenges/network-topology-and-parameter-inference-challenge, sep 2012.

[3] Stephen F Altschul, Warren Gish, Webb Miller, Eugene W Myers, and David J Lipman. Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410, 1990.

[4] Marti J. Anderson. A new method for Non-Parametric multivariate analysis of variance. *Austral Ecology*, 26(1):32–46, February 2001.

[5] Christophe Andrieu, Nando De Freitas, Arnaud Doucet, and Michael I Jordan. An introduction to mcmc for machine learning. *Machine learning*, 50(1-2):5–43, 2003.

[6] Rolf Apweiler, Maria Jesus Martin, Claire O'Donovan, Michele Magrane, Yasmin Alam-Faruque, Ricardo Antunes, Daniel Barrell, Benoit Bely, Mark Bingley, David Binns, et al. The universal protein resource (uniprot) in 2010. *Nucleic Acids Research*, 38(Database issue):D142–8, 2010.

[7] Sylvain Arlot and Alain Celisse. A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4:40–79, 2010.

[8] Nachman Aronszajn. Theory of reproducing kernels. *Transactions of the American mathematical society*, 68(3):337–404, 1950.

[9] Eva Balsa-Canto, Antonio Alonso, and Julio Banga. An iterative identification procedure for dynamic modeling of biochemical networks. *BMC Systems Biology*, 4(1):11, 2010.

[10] Samuel Bandara, Johannes P. Schlöder, Roland Eils, Hans Georg Bock, and Tobias Meyer. Optimal experimental design for parameter estimation of a cell signaling model. *PLoS Comput Biol*, 5(11):e1000558, 11 2009.

[11] Isabel Beichl and Francis Sullivan. The metropolis algorithm. *Computing in Science & Engineering*, 2(1):65–69, 2000.

[12] Emilio Benfenati and Giuseppina Gini. Computational predictive programs (expert systems) in toxicology. *Toxicology*, 119(3):213–225, 1997.

[13] Amanda Birmingham, Laura M. Selfors, Thorsten Forster, David Wrobel, Caleb J. Kennedy, Emma Shanks, Javier Santoyo-Lopez, Dara J. Dunican, Aideen Long, Dermot Kelleher, Queta Smith, Roderick L. Beijersbergen, Peter Ghazal, and Caroline E. Shamu. Statistical methods for analysis of high-throughput RNA interference screens. *Nature Methods*, 6(8):569–575, July 2009.

[14] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[15] P. Bogacki and L. F. Shampine. A 3(2) pair of Runge - Kutta formulas. *Applied Mathematics Letters*, 2:321–325, 1989.

[16] Léon Bottou and Olivier Bousquet. The tradeoffs of large-scale learning. *Optimization for Machine Learning*, page 351, 2011.

[17] Michael Boutros, Lígia P Brás, and Wolfgang Huber. Analysis of cell-based rnai screens. *Genome biology*, 7(7):R66, 2006.

[18] Ben Calderhead, Mark Girolami, and Neil D Lawrence. Accelerating bayesian inference over nonlinear differential equations with gaussian processes. In *Advances in neural information processing systems*, pages 217–224, 2008.

[19] M. Campillos, M. Kuhn, A.C. Gavin, L.J. Jensen, and P. Bork. Drug target identification using side-effect similarity. *Science*, 321(5886):263–6, 2008.

[20] Monica Campillos, Michael Kuhn, Anne-Claude Gavin, Lars Juhl Jensen, and Peer Bork. Drug target identification using side-effect similarity. *Science*, 321(5886):263–266, 2008.

[21] George T. Chang and Guenther Walther. Clustering with mixtures of log-concave distributions. *Computational Statistics & Data Analysis*, 51(12):6242–6251, 2007.

[22] Bin Chen, David Wild, and Rajarshi Guha. Pubchem as a source of polypharmacology. *Journal of chemical information and modeling*, 49(9):2044–2055, 2009.

[23] Oksana Chkrebtii, David A Campbell, Mark A Girolami, and Ben Calderhead. Bayesian uncertainty quantification for differential equations. *arXiv preprint arXiv:1306.2365*, 2013.

[24] Christian Conrad, Holger Erfle, Patrick Warnat, Nathalie Daigle, Thomas Lörch, Jan Ellenberg, Rainer Pepperkok, and Roland Eils. Automatic identification of subcellular phenotypes on human cell arrays. *Genome Research*, 14(6):1130–1136, 2004.

[25] D. R Cox and D. V. Hinkley. *Theoretical statistics*. Springer. 511p, 1974.

[26] Alexandre d'Aspremont, Laurent El Ghaoui, Michael I Jordan, and Gert RG Lanckriet. A direct formulation for sparse pca using semidefinite programming. *SIAM review*, 49(3):434–448, 2007.

[27] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

[28] Christopher M Dobson. Chemical space and biology. *Nature*, 432(7019):824–828, 2004.

[29] Sandrine Dudoit, Jane Fridlyand, and Terence P Speed. Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American statistical association*, 97(457):77–87, 2002.

[30] C. Eckart and G. Young. The approximation of one matrix by another of low rank. *Psychometrika*, 1:211, 1936.

[31] Krystian Eitner and Uwe Koch. From fragment screening to potent binders: strategies for fragment-to-lead evolution. *Mini Reviews in Medicinal Chemistry*, 9(8):956–961, 2009.

[32] G. Elidan. Copula bayesian networks. In J. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 559 – 567. MIT Press, 2010.

[33] Jean-Loup Faulon, Milind Misra, Shawn Martin, Ken Sale, and Rajat Sapra. Genome scale enzyme–metabolite and drug–target interaction predictions using the signature molecular descriptor. *Bioinformatics*, 24(2):225–233, 2008.

[34] Diego Ferández Slezak, Cecilia Suárez, Guillermo A. Cecchi, Guillermo Marshall, and Gustavo Stolovitzky. When the optimal is not the best: Parameter estimation in complex biological models. *PLoS ONE*, 5(10):e13283, 10 2010.

[35] Robert D Finn, John Tate, Jaina Mistry, Penny C Coggill, Stephen John Sammut, Hans-Rudolf Hotz, Goran Ceric, Kristoffer Forslund, Sean R Eddy, Erik LL Sonnhammer, et al. The pfam protein families database. *Nucleic acids research*, 36(suppl 1):D281–D288, 2008.

[36] J.H. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: A statistical view of boosting. *Annals of Statistics*, 28(2):337–407, 2000.

[37] Florian Fuchs, Gregoire Pau, Dominique Kranz, Oleg Sklyar, Christoph Budjan, Sandra Steinbrink, Thomas Horn, Angelika Pedal, Wolfgang Huber, and Michael Boutros. Clustering phenotype populations by genome-wide rnai and multiparametric imaging. *Molecular systems biology*, 6(1), 2010.

[38] M. Fukuzaki, M. Seki, H. Kashima, and J. Sese. Side effect prediction using cooperative pathways. *IEEE International Conference on Bioinformatics and Biomedicine 2009 (IEEE BIBM 2009)*, pages 142–147, 2009.

[39] Terrence S Furey, Nello Cristianini, Nigel Duffy, David W Bednarski, Michel Schummer, and David Haussler. Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, 16(10):906–914, 2000.

[40] C. Genest, K. Ghoudi, and L.-P. Rivest. A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. *Biometrika*, 82(3):543–552, 1995.

[41] Christian Genest and Anne-Catherine Favre. Everything you always wanted to know about copula modeling but were afraid to ask. *Journal of Hydrologic Engineering*, 12:347–368, 2007.

[42] K. M. Giacomini, R. M. Krauss, D. M. Roden, M. Eichelbaum, M. R. Hayden, and Y. Nakamura. When good drugs go bad. *Nature*, 446 (7139):975–977, 2007.

[43] Ben N. G. Giepmans, Stephen R. Adams, Mark H. Ellisman, and Roger Y. Tsien. The fluorescent toolbox for assessing protein location and function. *Science*, 312(5771):217–224, 2006.

[44] R. Gozalbes, R.J. Carbajo, and A. Pineda-Lucena. From fragment screening to potent binders: strategies for fragment-to-lead evolution. *Mini Reviews in Medicinal Chemistry*, 9(8):956–961, 2009.

[45] Michael J Greenacre. *Theory and applications of correspondence analysis*. 1984.

[46] Elisabet Gregori-Puigjané and Jordi Mestres. A ligand-based approach to mining the chemogenomic space of drugs. *Combinatorial chemistry & high throughput screening*, 11(8):669–676, 2008.

[47] M. Gribskov and N.L. Robinson. Use of receiver operating characteristic (ROC) analysis to evaluate sequence matching. *Comput Chem*, 20:25–33, 1996.

[48] Michael Gribskov and Nina L Robinson. Use of receiver operating characteristic (roc) analysis to evaluate sequence matching. *Computers & chemistry*, 20(1):25–33, 1996.

[49] Stefan Günther, Michael Kuhn, Mathias Dunkel, Monica Campillos, Christian Senger, Evangelia Petsalaki, Jessica Ahmed, Eduardo Garcia Urdiales, Andreas Gewiess, Lars Juhl Jensen, et al. Supertarget and matador: resources for exploring drug-target relationships. *Nucleic acids research*, 36(suppl 1):D919–D922, 2008.

[50] Lianyi Han, Yanli Wang, and Stephen H Bryant. Developing and validating predictive decision tree models from mining chemical structural fingerprints and high–throughput screening data in pubchem. *BMC bioinformatics*, 9(1):401, 2008.

[51] Trevor. Hastie, Robert. Tibshirani, and J Jerome H Friedman. *The elements of statistical learning*, volume 1. Springer New York, 2001.

[52] Matthew Hoffman, Francis R Bach, and David M Blei. Online learning for latent dirichlet allocation. In *advances in neural information processing systems*, pages 856–864, 2010.

[53] Brice Hoffmann, Mikhail Zaslavskiy, Jean-Philippe Vert, and Véronique Stoven. A new protein binding pocket similarity measure based on comparison of clouds of atoms in 3d: application to ligand prediction. *BMC bioinformatics*, 11(1):99, 2010.

[54] Myles Hollander and Douglas A. Wolfe. *Nonparametric statistical methods*. Wiley. 787 p, 1973.

[55] Harold Hotelling. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377, 1936.

[56] D. Houtsma, H.J. Guchelaar, and H. Gelderblom. Pharmacogenetics in oncology: a promising field. *Curr Pharm Des*, 16(2):155–163, 2010.

[57] Cho-Jui Hsieh, Kai-Wei Chang, Chih-Jen Lin, S Sathiya Keerthi, and Sellamanickam Sundararajan. A dual coordinate descent method for large-scale linear svm. In *Proceedings of the 25th international conference on Machine learning*, pages 408–415. ACM, 2008.

[58] Laurent Jacob, Brice Hoffmann, Véronique Stoven, and Jean-Philippe Vert. Virtual screening of gpcrs: an in silico chemogenomics approach. *BMC bioinformatics*, 9(1):363, 2008.

[59] Laurent Jacob and Jean-Philippe Vert. Protein-ligand interaction prediction: an improved chemogenomics approach. *Bioinformatics*, 24(19):2149–2156, 2008.

[60] Krzysztof Jajuga and Daniel Papla. Copula functions in model based clustering. In Myra Spiliopoulou, Rudolf Kruse, Christian Borgelt, Andreas N$\tilde{\text{A}}\frac{1}{4}$rnberger, and Wolfgang Gaul, editors, *From Data and Information Analysis to Knowledge Engineering*, Studies in Classification, Data Analysis, and Knowledge Organization, pages 606–613. Springer Berlin Heidelberg, 2006.

[61] Rodolphe Jenatton, Julien Mairal, Guillaume Obozinski, and Francis Bach. Proximal methods for hierarchical sparse coding. *arXiv preprint arXiv:1009.2139*, 2010.

[62] Xiao jiang Feng and Herschel Rabitz. Optimal identification of biochemical reaction networks. *Biophysical Journal*, 86(3):1270–1281, 2004.

[63] H. Joe and J.J. Xu. The estimation method of inference functions for margins for multivariate models. Technical Report 166, Department of Statistics, University of British Columbia, 1996.

[64] Harry Joe. Asymptotic efficiency of the two-stage estimation method for copula-based models. *Journal of Multivariate Analysis*, 94(2):401 – 419, 2005.

[65] Thouis R. Jones, Anne E. Carpenter, Michael R. Lamprecht, Jason Moffat, Serena J. Silver, Jennifer K. Grenier, Adam B. Castoreno, Ulrike S. Eggert, David E. Root, Polina Golland, and David M. Sabatini. Scoring diverse cellular morphologies in image-based screens with iterative feedback and machine learning. *Proceedings of the National Academy of Sciences*, 106(6):1826–1831, 2009.

[66] Michael I. Jordan, Zoubin Ghahramani, Tommi S. Jaakkola, and Lawrence K. Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37:183–233, November 1999.

[67] Minoru Kanehisa, Susumu Goto, Masahiro Hattori, Kiyoko F Aoki-Kinoshita, Masumi Itoh, Shuichi Kawashima, Toshiaki Katayama, Michihiro Araki, and Mika Hirakawa. From genomics to chemical genomics: new developments in kegg. *Nucleic acids research*, 34(suppl 1):D354–D357, 2006.

[68] Jonathan R Karr, Jayodita C Sanghvi, Derek N Macklin, Miriam V Gutschow, Jared M Jacobs, Benjamin Bolival, Nacyra Assad-Garcia, John I Glass, and Markus W Covert. A whole-cell computational model predicts phenotype from genotype. *Cell*, 150(2):389–401, 2012.

[69] Matthew A Kayala, Chloé-Agathe Azencott, Jonathan H Chen, and Pierre Baldi. Learning to predict chemical reactions. *Journal of chemical information and modeling*, 51(9):2209–2222, 2011.

[70] Michael J Keiser, Vincent Setola, John J Irwin, Christian Laggner, Atheir I Abbas, Sandra J Hufeisen, Niels H Jensen, Michael B Kuijer, Roberto C Matos,

Thuy B Tran, et al. Predicting new molecular targets for known drugs. *Nature*, 462(7270):175–181, 2009.

[71] Esther Kellenberger, Nicolas Foata, and Didier Rognan. Ranking targets in structure-based virtual screening of three-dimensional protein libraries: methods and problems. *Journal of chemical information and modeling*, 48(5):1014–1025, 2008.

[72] G. Kim, M. J. Silvapulle, and P. Silvapulle. Comparison of semiparametric and parametric methods for estimating copulas. *Computational Statistics & Data Analysis*, 51(6):2836 – 2850, 2007.

[73] Jong-Min Kim, Yoon-Sung Jung, Engin Sungur, Kap-Hoon Han, Changyi Park, and Insuk Sohn. A copula method for modeling directional dependence of genes. *BMC Bioinformatics*, 9:225. Available: http://www.biomedcentral.com/1471–2105/9/225. Accessed 30 September 2012, 2008.

[74] Hiroaki Kitano. Systems biology: A brief overview. *Science*, 295(5560):1662–1664, 2002.

[75] Justin Klekota and Frederick P Roth. Chemical substructures that enrich for biological activity. *Bioinformatics*, 24(21):2518–2525, 2008.

[76] Peter Kolb, Rafaela S Ferreira, John J Irwin, and Brian K Shoichet. Docking and chemoinformatic screens for new ligands and targets. *Current opinion in biotechnology*, 20(4):429–436, 2009.

[77] Daphne Kollar and Nir Friedman. *Probabilistic graphical models: principles and techniques*. The MIT Press, 2009.

[78] D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. Cambridge: MIT Press. 1231 p, 2009.

[79] S Kramer, E Frank, and C Helma. Fragment generation and support vector machines for inducing sars. *SAR and QSAR in Environmental Research*, 13(5):509–523, 2002.

[80] M.A. Kravette. Perilymphatic atrophy of skin. an adverse side effect of intralesional steroid injections. *Clin Podiatr Med Surg*, 3:457–62, 1986.

[81] Clemens Kreutz and Jens Timmer. Systems biology: experimental design. *FEBS Journal*, 276(4):923–942, 2009.

[82] M. Kuhn, M. Campillos, I. Letunic, L.J. Jensen, and P. Bork. A side effect resource to capture phenotypic effects of drugs. *Mol Syst Biol*, 6:343, 2010.

[83] Ronny Luss and Marc Teboulle. Conditional gradient algorithmsfor rank-one matrix approximations with a sparsity constraint. *SIAM Review*, 55(1):65–98, 2013.

[84] Sarah R McWhinney, Richard M Goldberg, and Howard L McLeod. Platinum neurotoxicity pharmacogenetics. *Molecular cancer therapeutics*, 8(1):10–16, 2009.

[85] Richard J Morris, Rafael J Najmanovich, Abdullah Kahraman, and Janet M Thornton. Real spherical harmonic expansion coefficients as 3d shape descriptors for protein binding pocket and ligand comparisons. *Bioinformatics*, 21(10):2347–2355, 2005.

[86] Nobuyoshi Nagamine and Yasubumi Sakakibara. Statistical prediction of protein–chemical interactions based on chemical structure and mass spectrometry data. *Bioinformatics*, 23(15):2004–2012, 2007.

[87] Rafael Najmanovich, Natalja Kurbatova, and Janet Thornton. Detection of 3d atomic similarities and their use in the discrimination of small molecule protein-binding sites. *Bioinformatics*, 24(16):i105–i111, 2008.

[88] Roger B. Nelsen. *An Introduction to Copulas (Lecture Notes in Statistics)*. Springer. 269 p, 1998.

[89] Beate Neumann, Thomas Walter, Jean-Karim K. Hériché, Jutta Bulkescher, Holger Erfle, Christian Conrad, Phill Rogers, Ina Poser, Michael Held, Urban Liebel, Cihan Cetin, Frank Sieckmann, Gregoire Pau, Rolf Kabbe, Annelie Wünsche, Venkata Satagopam, Michael H. Schmitz, Catherine Chapuis, Daniel W. Gerlich, Reinhard Schneider, Roland Eils, Wolfgang Huber, Jan-Michael M. Peters, Anthony A. Hyman, Richard Durbin, Rainer Pepperkok, and Jan Ellenberg. Phenotypic profiling of the human genome by time-lapse microscopy reveals cell division genes. *Nature*, 464(7289):721–727, April 2010.

[90] Jorge Nocedal and Stephen J Wright. *Numerical optimization*. Springer Science+ Business Media, 2006.

[91] Arno Onken, Steffen Grünewälder, Matthias H. J. Munk, and Klaus Obermayer. Analyzing short-term noise dependencies of spike-counts in macaque prefrontal cortex using copulas and the flashlight transformation. *PLoS Comput Biol*, 5(11):e1000577. Accessed 30 September 2011, 2009.

[92] Christos A Ouzounis. Rise and demise of bioinformatics? promise and progress. *PLoS computational biology*, 8(4):e1002487, 2012.

[93] Elena Parkhomenko, David Tritchler, and Joseph Beyene. Genome-wide sparse canonical correlation of gene expression with genotypes. In *BMC proceedings*, volume 1, page S119. BioMed Central Ltd, 2007.

[94] Edouard Pauwels, Véronique Stoven, and Yoshihiro Yamanishi. Predicting drug side-effect profiles: a chemical fragment-based approach. *BMC bioinformatics*, 12(1):169, 2011.

[95] Edouard Pauwels, Didier Surdez, Gautier Stoll, Aurianne Lescure, Elaine Del Nery, Olivier Delattre, and Véronique Stoven. A probabilistic model for cell population phenotyping using hcs data. *PloS one*, 7(8):e42715, 2012.

[96] Rainer Pepperkok and Jan Ellenberg. High-throughput fluorescence microscopy for systems biology. *Nature Reviews Molecular Cell Biology*, 7(9):690–696, 2006.

[97] Matthew Piazza, Xiao-Jiang Feng, Joshua D. Rabinowitz, and Herschel Rabitz. Diverse metabolic model parameters generate similar methionine cycle dynamics. *Journal of Theoretical Biology*, 251(4):628 – 639, 2008.

[98] C. Richard and M.D. Klein. Ventricular arrhythmias in aortic valve disease: Analysis of 102 patients. *The American Journal of Cardiology*, 53(8):1079 – 1083, 1984.

[99] Nora Rieber, Bettina Knapp, Roland Eils, and Lars Kaderali. Rnaither, an automated pipeline for the statistical analysis of high-throughput rnai screens. *Bioinformatics*, 25(5):678–679, 2009.

[100] Nicholas Roy and Andrew Mccallum. Toward optimal active learning through sampling estimation of error reduction. In *In Proc. 18th International Conf. on Machine Learning*, pages 441–448. Morgan Kaufmann, 2001.

[101] Kerstin Sander and Dirk Sander. New insights into transient global amnesia: recent imaging and clinical findings. *The Lancet Neurology*, 4(7):437–444, 2005.

[102] J. Scheiber, J.L. Jenkins, S.C. Sukuru, A. Bender, D. Mikhailov, M. Milik, K. Azzaoui, S. Whitebread, J. Hamon, L. Urban, M. Glick, and J.W. Davies. Mapping adverse drug reactions in chemical space. *J Med Chem*, 52(9):3103–7, 2009.

[103] Josef Scheiber, Bin Chen, Mariusz Milik, Sai Chetan K Sukuru, Andreas Bender, Dmitri Mikhailov, Steven Whitebread, Jacques Hamon, Kamal Azzaoui, Laszlo Urban, et al. Gaining insight into off-target mediated effects of drug candidates with a comprehensive systems chemical biology analysis. *Journal of chemical information and modeling*, 49(2):308–317, 2009.

[104] Bernhard Schölkopf and Alexander J Smola. *Learning with kernels*. MIT Press, 2002.

[105] Bernhard Schölkopf, Koji Tsuda, and Jean-Philippe Vert. *Kernel methods in computational biology*. The MIT press, 2004.

[106] C. Schölzel and P. Friederichs. Multivariate non-normally distributed random variables in climate research – introduction to the copula approach. *Nonlinear Processes in Geophysics*, 15(5):761–772, 2008.

[107] Burr Settles. Active learning literature survey. Technical report, Computer Sciences Technical Report 1648, University of Wisconsin-Madison, 2010.

[108] Daichi Shigemizu, Michihiro Araki, Shujiro Okuda, Susumu Goto, and Minoru Kanehisa. Extraction and analysis of chemical modification patterns in drug development. *Journal of chemical information and modeling*, 49(4):1122–1129, 2009.

[109] Joanna H. Shih and Thomas A. Louis. Inferences on the association parameter in copula models for bivariate survival data. *Biometrics*, 51:1384–399, 1995.

[110] A. Sklar. Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut de statistique de l'Université de Paris*, 8:229–231, 1959.

[111] Michael D. Slack, Elisabeth D. Martinez, Lani F. Wu, and Steven J. Altschuler. Characterizing heterogeneous cellular responses to perturbations. *Proceedings of the National Academy of Sciences*, 105(49):19306–19311, 2008.

[112] A.J. Smola and B. Schölkopf. On a kernel-based method for pattern recognition, regression, approximation and operator inversion. *Algorithmica*, 22:211–231, 1998.

[113] Berend Snijder, Raphael Sacher, Pauli Rämö, Eva-Maria Damm, Prisca Liberali, and Lucas Pelkmans. Population context determines cell-to-cell variability in endocytosis and virus infection. *Nature*, 461(7263):520–523, 2009.

[114] Karline Soetaert, Thomas Petzoldt, and R Woodrow Setzer. Solving differential equations in r: Package desolve. *Journal of Statistical Software*, 33(9):1–25, 2010.

[115] Wall St, David X. Li, and David X. Li. On default correlation: A copula function approach. *Journal of Fixed income*, 9(4):43–54, 2000.

[116] Bernhard Steiert, Andreas Raue, Jens Timmer, and Clemens Kreutz. Experimental design for parameter estimation of gene regulatory networks. *PLoS ONE*, 7(7):e40052, 07 2012.

[117] Brent R Stockwell. Chemical genetics: ligand-based discovery of gene function. *Nature Reviews Genetics*, 1(2):116–125, 2000.

[118] Yasuo Tabei, Edouard Pauwels, Véronique Stoven, Kazuhiro Takemoto, and Yoshihiro Yamanishi. Identification of chemogenomic features from drug–target interaction networks using interpretable classifiers. *Bioinformatics*, 28(18):i487–i494, 2012.

[119] N.P. Tatonetti, T. Liu, and R.B. Altman. Predicting drug side-effects by chemical systems biology. *Genome Biol*, 10:238, 2009.

[120] Robert Tibshirani, Trevor Hastie, Balasubramanian Narasimhan, and Gilbert Chu. Class prediction by nearest shrunken centroids, with applications to dna microarrays. *Statistical Science*, pages 104–117, 2003.

[121] Franck Tirode, Karine Laud-Duval, Alexandre Prieur, Bruno Delorme, Pierre Charbord, and Olivier Delattre. Mesenchymal stem cell features of ewing tumors. *Cancer Cell*, 11(5):421 – 429, 2007.

[122] Mark Transtrum and Peng Qiu. Optimal experiment selection for parameter estimation in biological differential equation models. *BMC Bioinformatics*, 13(1):181, 2012.

[123] Vladimir N Vapnik. An overview of statistical learning theory. *Neural Networks, IEEE Transactions on*, 10(5):988–999, 1999.

[124] Sandra Waaijenborg, Philip C Verselewel de Witt Hamer, and Aeilko H Zwinderman. Quantifying the association between gene expressions and dna-markers by penalized canonical correlation analysis. *Statistical Applications in Genetics and Molecular Biology*, 7(1), 2008.

[125] Jun Wang, Xiaobo Zhou, Pamela L. Bradley, Shih-Fu Chang, Norbert Perrimon, and Stephen T.C. Wong. Cellular phenotype recognition for high-content rna interference genome-wide screening. *Journal of Biomolecular Screening*, 13(1):29–39, 2008.

[126] Nathanael Weill and Didier Rognan. Development and validation of a novel protein- ligand fingerprint to mine chemogenomic space: Application to g protein-coupled receptors and their ligands. *Journal of chemical information and modeling*, 49(4):1049–1062, 2009.

[127] Steven Whitebread, Jacques Hamon, Dejan Bojanic, and Laszlo Urban. Keynote review:*in vitro* safety pharmacology profiling: an essential tool for successful drug development. *Drug discovery today*, 10(21):1421–1433, 2005.

[128] David S Wishart, Craig Knox, An Chi Guo, Savita Shrivastava, Murtaza Hassanali, Paul Stothard, Zhan Chang, and Jennifer Woolsey. Drugbank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic acids research*, 34(suppl 1):D668–D672, 2006.

[129] Daniela M Witten, Robert Tibshirani, and Trevor Hastie. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, 10(3):515–534, 2009.

[130] Olaf Wolkenhauer, Mukhtar Ullah, Walter Kolch, and Kwang-Hyun Cho. Modelling and simulation of intracellular dynamics: Choosing an appropriate framework. *IEEE Transactions on NanoBioscience*, 3:200–207, 2004.

[131] Li Xie, Jerry Li, Lei Xie, and Philip E Bourne. Drug discovery using chemical systems biology: identification of the protein-ligand binding network to explain the side effects of cetp inhibitors. *PLoS computational biology*, 5(5):e1000387, 2009.

[132] Peter Xue-Kun Song. Multivariate dispersion models generated from gaussian copula. *Scandinavian Journal of Statistics*, 27(2):305–320, 2000.

[133] Yoshihiro Yamanishi. Supervised bipartite graph inference. In *Advances in Neural Information Processing Systems*, pages 1841–1848, 2008.

[134] Yoshihiro Yamanishi, Michihiro Araki, Alex Gutteridge, Wataru Honda, and Minoru Kanehisa. Prediction of drug–target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics*, 24(13):i232–i240, 2008.

[135] Yoshihiro Yamanishi, Masaaki Kotera, Minoru Kanehisa, and Susumu Goto. Drug-target interaction prediction from chemical, genomic and pharmacological data in an integrated framework. *Bioinformatics*, 26(12):i246–i254, 2010.

[136] Yoshihiro Yamanishi, Edouard Pauwels, and Masaaki Kotera. Drug side-effect prediction based on the integration of chemical and biological spaces. *Journal of chemical information and modeling*, 52(12):3284–3292, 2012.

[137] Yoshihiro Yamanishi, Edouard Pauwels, Hiroto Saigo, and Véronique Stoven. Extracting sets of chemical substructures and protein domains governing drug-target interactions. *Journal of chemical information and modeling*, 51(5):1183–1194, 2011.

[138] Ting Zhao and Robert F Murphy. Automated learning of generative models for subcellular location: Building blocks for systems biology. *Cytometry Part A*, 71(12):978–990, 2007.

# Appendix A

# Copulas

Copulas became popular in statistical litterature at the end of the twentieth century. However, the study of these objects goes back to the middle of the century [88]. We present here a brief review of the copula framework.

## A.1 Definition and example

A copula is a joint multivariate (cumulative) distribution which univariate marginals are uniform over $[0, 1]$. More formaly, let $U_1, \ldots, U_d \in [0, 1]^d$ $d$ uniform random variables, $U_i \sim U$. A copula $C : [0, 1]^d \to [0, 1]$ is a joint distribution function

$$C(u_1, \ldots, u_d) = P(U_1 \leq u1, \ldots, U_d \leq u_d)$$

Let $\mathcal{X} = \{X_1, \ldots, X_d\}$ a finite set of real valued random variables and let $F_{\mathcal{X}}(x_1, \ldots, x_d) = P(X_1 \leq x_1, \ldots, X_d \leq x_d)$ a cumulative distribution function over $\mathcal{X}$. The importance of copulas comes from the fact that any distribution can be formalized in term of copula [110].

**Sklar theorem (1959)** Let $F_{\mathcal{X}}$ be any multivariate distribution over real-valued random variables, then there exists a copula function such that

$$F_{\mathcal{X}}(x_1, \ldots, x_d) = C(F_1(x_1), \ldots, F_d(x_d))$$

Where the $F_i$ are the correponding univariate marginals. If each $F_i$ is continuous then $C$ is unique.

Conversely, since for any univariate random variable $X$ with cumulative distribution $F$, the random variable $F(X)$ is uniformly distributed over $[0, 1]$, any copula taking these transformed univariate marginals $\{F_1(X_1), \ldots, F_d(X_d)\}$ as its arguments defines a valid distribution function which univariate marginals are specified. From a modeling point of view, this allows to separate the choice of univariate marginals and the choice of the joint dependence structure when constructing a distribution.

**Copula based densities** Since the joint cumulative distribution function can be expressed in term of copula, the joint density can also be specified in term of copula density. If $\mathbf{x} = (x_1, \ldots, x_d)$, let $F(\mathbf{x}) = C(F_1(x_1), \ldots, F_d(x_d))$ a distribution over $\mathcal{X}$ with $d$-order partial derivatives. The corresponding density can be expressed as

$$\begin{aligned} f(\mathbf{x}) &= \frac{\partial^d F(\mathbf{x})}{\partial x_1, \ldots, \partial x_d} \\ &= \frac{\partial^d C(F_1(x_1), \ldots, F_d(x_d))}{\partial F_1(x_1), \ldots, \partial F_d(x_d)} \prod_i f_i(x_i) \\ &= c(F_1(x_1), \ldots, F_d(x_d)) \prod_i f_i(x_i) \end{aligned} \tag{A.1}$$

Where $c$ is a *copula density function* and $f_i$ is the univariate density corresponding to $F_i$. There are many ways to define and construct copulas, see [88] for a general review. We focus here on one specific copula family which is defined from multivariate gaussian distributions.

**The gaussian copula** This copula family describes the denpendance structure of multivariate gaussians. It has been introduced in 2000 [132]. Let $\phi$ and $\Phi$ the standard (zero mean, unit variance) univariate normal density and cumulative distribution respectively. $\Phi^{-1}$ is the corresponding quantile function. A general univariate normal with mean $\mu$ and standard deviation $\sigma$ has a density function $f_1(x) = \frac{\phi(x_s)}{\sigma}$ and

cumulative distribution function $F_1(x) = \Phi(x_s)$ where $x_s = \frac{x-\mu}{\sigma}$ corresponds to $x$ scaled.

Let $\mathbf{X} = \{X_1, \ldots, X_d\}$ a $d$-dimensional random variable which follows a multivariate normal distribution with correlation matrix $R$. If $\mathbf{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d$, let $\{\mu_i, \sigma_i\}$ the mean and standard deviation of $X_i$ and $x_{si} = \frac{x_i-\mu_i}{\sigma_i}$ for all $i$, $\mathbf{x_s} = (x_{s1}, \ldots, x_{sd}) \in \mathbb{R}^d$. The $d$ dimensional multivariate normal density describing $\mathbf{X}$ can be expressed as

$$\begin{aligned} f_d(\mathbf{x}) &= \frac{1}{(2\pi)^{\frac{d}{2}}|R|^{\frac{1}{2}}\prod_{i=1}^d \sigma_i} \exp(-\frac{1}{2}\mathbf{x}_s^T R \mathbf{x}_s) \\ &= \frac{1}{|R|^{\frac{1}{2}}} \exp(-\frac{1}{2}\mathbf{x}_s^T (R^{-1} - I)\mathbf{x}_s) \prod_{i=1}^d \frac{\phi(x_{si})}{\sigma_i} \end{aligned} \tag{A.2}$$

We have $x_{si} = \Phi^{-1}(\Phi(x_{si})) = \Phi^{-1}(F_i(x_i))$ where $F_i$ is the univariate marginal distribution related to $x_i$. Moreover , we have seen that $\frac{\phi(x_{si})}{\sigma_i}$ is the univariate marginal density related to $x_i$. Identifying the density formulation in (A.2) with the general formulation in (A.1) allows to define the gaussian copula density. Let $R$ a correlation matrix and $\{U_1, \ldots, U_d\}$, $d$ random variables uniform over $[0, 1]$.

$$c_g : (u_1, \ldots, u_d) \to \frac{1}{|R|^{\frac{1}{2}}} \exp(-\frac{1}{2}u_s^T (R^{-1} - I)u_s)$$

where $u_{si} = \Phi^{-1}(u_i)$, is the copula density function describing the joint dependence structure shared by the multivariate normal distributions which correlation matrix is $R$. This definition is independent of the univariate marginal, therefore pluging any univariate marginal in this formulation allows to construct density functions which have a gaussian dependence structure and a different domain for example.

## A.2 Copula models parameters estimation

Let $\{F_{\theta_1}, \ldots, F_{\theta_d}\}$ a set of univariate marginal distributions and $\{f_{\theta_1}, \ldots, f_{\theta_d}\}$ the corresponding univariate densities parametrized by $\theta_F = \{\theta_1, \ldots, \theta_d\}$. Let $c_{\theta_c}$ a copula

density function parameterized by $\theta_c$. We are concerened with finding the parameters $\Theta = \{\theta_F, \theta_c)\}$ which fit the best to a set of $n$ $d$-dimensional data points: $\{\mathbf{x_1}, \ldots, \mathbf{x_n}\}$. We review three common procedures. Let $P_{F,c,\Theta} : \mathbf{x} \to c_{\theta_c}(F_{\theta_1}(x_1), \ldots, F_{\theta_d}(x_d)) \prod_i f_{\theta_i}(x_i)$ the density function associated to the parameter $\Theta$.

### A.2.1 Maximum likelihood

The most logical approach is to find $\hat{\Theta}$ which is defined as

$$\hat{\Theta} = \underset{\Theta}{\operatorname{argmax}} \prod_{k=1}^{n} P_{F,c,\Theta}(\mathbf{x_k})$$

This estimator enjoys consistency and assymptotical normality under regularity conditions [25]. It is well suited to problems which optimal parameters can be estimated in closed form which is not the case here. Numerical optimization is needed, wich becomes very heavy when the dimension increases.

### A.2.2 Inference functions for margin

This two stage estimation has been introduced by [109] for the bivariate case and by [63] in the general case. First the univariate marginal parameters are estimated independently from the copula. In a second stage, the copula parameters are estimated given the univariate marginal parameters. For all $i$,

$$\hat{\theta}_i = \underset{\theta_i}{\operatorname{argmax}} \prod_{k=1}^{n} f_{\theta_i}(\mathbf{x_k}_i)$$

Wich defines $\hat{\theta}_F = \{\hat{\theta}_1, \ldots, \hat{\theta}_d\}$ and fixes the univariate marginals. $\theta_c$ is then estimated as follows.

$$\hat{\theta}_c = \underset{\theta_c}{\operatorname{argmax}} \prod_{k=1}^{n} P_{F,c,\theta_c,\hat{\theta}_F}(\mathbf{x_k})$$

The copula parameter estimate is consistent with assymptotic normality [109]. Moreover, the whole set of estimators $\hat{\Theta} = \{\hat{\theta}_F, \hat{\theta}_c\}$ has the same property [64].

This procedure is way less expensive computationally than the maximum likelihood estimator.

### A.2.3 Semi-parametric estimation

The inference functions for margin method relies on univariate marginals parameter estimation as a first stage. This estimations procedure relies on the parametrization of univariate marginals. [40] proposes a semi-parametric alternative using a non parametric estimate of the univariate marginals. The univariate marginals are replaced by their scaled empirical cumulative distribution function. This amounts to work with standardized ranks instead of original variables. For all $i$, let $F_i^*$ stand for $\frac{n}{n+1}$ times the $i$-th univariate marginal empirical distribution function. The scaling factor ensures that our observations strictly stay in $]0,1[$. The estimator $\hat{\theta}_c$ of the copula parameter is then

$$\hat{\theta}_c = \underset{\theta_c}{\operatorname{argmax}} \prod_{k=1}^{n} c_{\theta_c}(F_1^*(\mathbf{x}_{\mathbf{k}1}), \ldots, F_d^*(\mathbf{x}_{\mathbf{k}d}))$$

This estimator is consistent and assymptotically normal [40]. The criterion to optimize is sometimes called *pseudo-likelihood* due to its similarities with conventional likelihood. In this formulation, the copula parameter estimation does not rely on the univariate marginal parameters, both can be estimated separately. It can be shown that the asymptotic variance of this estimator is higher than that of maximum likelihood estimation. However, [72] showed empirically that this last procedure is more robust to marginals mispecification than the two preceding ones.

# Appendix B

# DREAM7 parameter estimation challenge dynamic equations

$$[\text{as1}] = \frac{\left(\frac{[\text{p1}]}{\text{r1}_{\text{Kd}}}\right)^{\text{r1}_{\text{h}}}}{1 + \left(\frac{[\text{p1}]}{\text{r1}_{\text{Kd}}}\right)^{\text{r1}_{\text{h}}}}$$

$$[\text{as2}] = \frac{\left(\frac{[\text{p1}]}{\text{r2}_{\text{Kd}}}\right)^{\text{r2}_{\text{h}}}}{1 + \left(\frac{[\text{p1}]}{\text{r2}_{\text{Kd}}}\right)^{\text{r2}_{\text{h}}}}$$

$$[\text{as3}] = \frac{\left(\frac{[\text{p4}]}{\text{r5}_{\text{Kd}}}\right)^{\text{r5}_{\text{h}}}}{1 + \left(\frac{[\text{p4}]}{\text{r5}_{\text{Kd}}}\right)^{\text{r5}_{\text{h}}}}$$

$$[\text{as5}] = \frac{\left(\frac{[\text{p8}]}{\text{r13}_{\text{Kd}}}\right)^{\text{r13}_{\text{h}}}}{1 + \left(\frac{[\text{p8}]}{\text{r13}_{\text{Kd}}}\right)^{\text{r13}_{\text{h}}}}$$

$$[\text{as6}] = \frac{\left(\frac{[\text{p9}]}{\text{r9}_{\text{Kd}}}\right)^{\text{r9}_{\text{h}}}}{1 + \left(\frac{[\text{p9}]}{\text{r9}_{\text{Kd}}}\right)^{\text{r9}_{\text{h}}}}$$

$$[\text{as7}] = \frac{\left(\frac{[\text{p6}]}{\text{r12}_{\text{Kd}}}\right)^{\text{r12}_{\text{h}}}}{1 + \left(\frac{[\text{p6}]}{\text{r12}_{\text{Kd}}}\right)^{\text{r12}_{\text{h}}}}$$

$$[\text{as9}] = \frac{\left(\frac{[\text{p6}]}{\text{r11}_{\text{Kd}}}\right)^{\text{r11}_{\text{h}}}}{1 + \left(\frac{[\text{p6}]}{\text{r11}_{\text{Kd}}}\right)^{\text{r11}_{\text{h}}}}$$

$$[\text{rs1a}] = \frac{1}{1 + \left(\frac{[\text{p2}]}{\text{r4}_{\text{Kd}}}\right)^{\text{r4}_{\text{h}}}}$$

$$[\text{rs1b}] = \frac{1}{1 + \left(\frac{[\text{p6}]}{\text{r8}_{\text{Kd}}}\right)^{\text{r8}_{\text{h}}}}$$

$$[\text{rs2}] = \frac{1}{1 + \left(\frac{[\text{p3}]}{\text{r3}_{\text{Kd}}}\right)^{\text{r3}_{\text{h}}}}$$

$$[\text{rs3}] = \frac{1}{1 + \left(\frac{[\text{p5}]}{\text{r7}_{\text{Kd}}}\right)^{\text{r7}_{\text{h}}}}$$

$$[\text{rs7}] = \frac{1}{1 + \left(\frac{[\text{p7}]}{\text{r6}_{\text{Kd}}}\right)^{\text{r6}_{\text{h}}}}$$

$$[\text{rs8}] = \frac{1}{1 + \left(\frac{[\text{p9}]}{\text{r10}_{\text{Kd}}}\right)^{\text{r10}_{\text{h}}}}$$

$$
\begin{aligned}
[\mathrm{g1}] &= [\mathrm{as1}] \cdot [\mathrm{rs1a}] \cdot [\mathrm{rs1b}] \\
[\mathrm{g2}] &= [\mathrm{as2}] \cdot [\mathrm{rs2}] \\
[\mathrm{g3}] &= [\mathrm{as3}] \cdot [\mathrm{rs3}] \\
[\mathrm{g4}] &= [\mathrm{as4}] \cdot [\mathrm{rs4}] \\
[\mathrm{g5}] &= [\mathrm{as5}] \\
[\mathrm{g7}] &= [\mathrm{as7}] \cdot [\mathrm{rs7}] \\
[\mathrm{g8}] &= [\mathrm{rs8}] \\
[\mathrm{g9}] &= [\mathrm{as9}] \\
\frac{\mathrm{d}\left([\mathrm{p1}]\right)}{\mathrm{d}t} &= \mathrm{rbs1}_{\mathrm{strength}} \cdot [\mathrm{v1}_{\mathrm{mrna}}] - \mathrm{p1}_{\mathrm{degradationRate}} \cdot [\mathrm{p1}] \\
\frac{\mathrm{d}\left([\mathrm{p1}]\right)}{\mathrm{d}t} &= \mathrm{rbs1}_{\mathrm{strength}} \cdot [\mathrm{v1}_{\mathrm{mrna}}] - \mathrm{p1}_{\mathrm{degradationRate}} \cdot [\mathrm{p1}] \\
\frac{\mathrm{d}\left([\mathrm{p2}]\right)}{\mathrm{d}t} &= \mathrm{rbs2}_{\mathrm{strength}} \cdot [\mathrm{v2}_{\mathrm{mrna}}] - \mathrm{p2}_{\mathrm{degradationRate}} \cdot [\mathrm{p2}] \\
\frac{\mathrm{d}\left([\mathrm{p3}]\right)}{\mathrm{d}t} &= \mathrm{rbs3}_{\mathrm{strength}} \cdot [\mathrm{v3}_{\mathrm{mrna}}] - \mathrm{p3}_{\mathrm{degradationRate}} \cdot [\mathrm{p3}] \\
\frac{\mathrm{d}\left([\mathrm{p4}]\right)}{\mathrm{d}t} &= \mathrm{rbs4}_{\mathrm{strength}} \cdot [\mathrm{v4}_{\mathrm{mrna}}] - \mathrm{p4}_{\mathrm{degradationRate}} \cdot [\mathrm{p4}] \\
\frac{\mathrm{d}\left([\mathrm{p5}]\right)}{\mathrm{d}t} &= \mathrm{rbs5}_{\mathrm{strength}} \cdot [\mathrm{v5}_{\mathrm{mrna}}] - \mathrm{p5}_{\mathrm{degradationRate}} \cdot [\mathrm{p5}] \\
\frac{\mathrm{d}\left([\mathrm{p6}]\right)}{\mathrm{d}t} &= \mathrm{rbs6}_{\mathrm{strength}} \cdot [\mathrm{v6}_{\mathrm{mrna}}] - \mathrm{p6}_{\mathrm{degradationRate}} \cdot [\mathrm{p6}] \\
\frac{\mathrm{d}\left([\mathrm{p7}]\right)}{\mathrm{d}t} &= \mathrm{rbs8}_{\mathrm{strength}} \cdot [\mathrm{v7}_{\mathrm{mrna}}] - \mathrm{p7}_{\mathrm{degradationRate}} \cdot [\mathrm{p7}] \\
\frac{\mathrm{d}\left([\mathrm{p8}]\right)}{\mathrm{d}t} &= \mathrm{rbs7}_{\mathrm{strength}} \cdot [\mathrm{v8}_{\mathrm{mrna}}] - \mathrm{p8}_{\mathrm{degradationRate}} \cdot [\mathrm{p8}] \\
\frac{\mathrm{d}\left([\mathrm{p9}]\right)}{\mathrm{d}t} &= \mathrm{rbs9}_{\mathrm{strength}} \cdot [\mathrm{v9}_{\mathrm{mrna}}] - \mathrm{p9}_{\mathrm{degradationRate}} \cdot [\mathrm{p9}] \\
\frac{\mathrm{d}\left([\mathrm{v1}_{\mathrm{mrna}}]\right)}{\mathrm{d}t} &= \mathrm{pro1}_{\mathrm{strength}} \cdot [\mathrm{g1}] - \mathrm{v1}_{\mathrm{mrnaDegradationRate}} \cdot [\mathrm{v1}_{\mathrm{mrna}}] \\
\frac{\mathrm{d}\left([\mathrm{v2}_{\mathrm{mrna}}]\right)}{\mathrm{d}t} &= \mathrm{pro2}_{\mathrm{strength}} \cdot [\mathrm{g2}] - \mathrm{v2}_{\mathrm{mrnaDegradationRate}} \cdot [\mathrm{v2}_{\mathrm{mrna}}] \\
\frac{\mathrm{d}\left([\mathrm{v3}_{\mathrm{mrna}}]\right)}{\mathrm{d}t} &= \mathrm{pro3}_{\mathrm{strength}} \cdot [\mathrm{g3}] - \mathrm{v3}_{\mathrm{mrnaDegradationRate}} \cdot [\mathrm{v3}_{\mathrm{mrna}}]
\end{aligned}
$$

$$\frac{\mathrm{d}\left([\mathrm{v4_{mrna}}]\right)}{\mathrm{d}t} = \mathrm{pro4_{strength}} \cdot [\mathrm{g4}] - \mathrm{v4_{mrnaDegradationRate}} \cdot [\mathrm{v4_{mrna}}]$$
$$\frac{\mathrm{d}\left([\mathrm{v5_{mrna}}]\right)}{\mathrm{d}t} = \mathrm{pro5_{strength}} \cdot [\mathrm{g5}] - \mathrm{v5_{mrnaDegradationRate}} \cdot [\mathrm{v5_{mrna}}]$$
$$\frac{\mathrm{d}\left([\mathrm{v6_{mrna}}]\right)}{\mathrm{d}t} = \mathrm{pro6_{strength}} \cdot [\mathrm{g6}] - \mathrm{v6_{mrnaDegradationRate}} \cdot [\mathrm{v6_{mrna}}]$$
$$\frac{\mathrm{d}\left([\mathrm{v7_{mrna}}]\right)}{\mathrm{d}t} = \mathrm{pro7_{strength}} \cdot [\mathrm{g9}] - \mathrm{v7_{mrnaDegradationRate}} \cdot [\mathrm{v7_{mrna}}]$$
$$\frac{\mathrm{d}\left([\mathrm{v8_{mrna}}]\right)}{\mathrm{d}t} = \mathrm{pro9_{strength}} \cdot [\mathrm{g7}] - \mathrm{v8_{mrnaDegradationRate}} \cdot [\mathrm{v8_{mrna}}]$$
$$\frac{\mathrm{d}\left([\mathrm{v9_{mrna}}]\right)}{\mathrm{d}t} = \mathrm{pro8_{strength}} \cdot [\mathrm{g8}] - \mathrm{v9_{mrnaDegradationRate}} \cdot [\mathrm{v9_{mrna}}]$$

## Applications de l'apprentissage statistique à la biologie computationnelle

**RÉSUMÉ :** Les biotechnologies sont arrivées au point ou la quantité d'information disponible permet de penser les objets biologiques comme des systèmes complexes. Dans ce contexte, les phénomènes qui émergent de ces systèmes sont intimement liés aux spécificités de leur organisation. Cela pose des problèmes computationnels et statistiques qui sont précisément l'objet d'étude de la communauté liée à l'apprentissage statistique. Cette thèse traite d'applications de méthodes d'apprentissage pour l'étude de phénomène biologique dans une perspective de système complexe. Ces méthodes sont appliquées dans le cadre de l'analyse d'interactions protéine-ligand et d'effets secondaires, du phenotypage de populations de cellules et du plan d'expérience pour des systèmes dynamiques non linéaires partiellement observés.

D'importantes quantités de données sont désormais disponibles concernant les molécules mises sur le marché, tels que les profils d'interactions protéiques et d'effets secondaires. Cela pose le problème d'intégrer ces données et de trouver une forme de structure sous tendant ces observations à grandes échelles. Nous appliquons des méthodes récentes d'apprentissage non supervisé à l'analyse d'importants jeux de données sur des médicaments. Des exemples illustrent la pertinence de l'information extraite qui est ensuite validée dans un contexte de prédiction.

Les variations de réponses à un traitement entre différents individus posent le problème de définir l'effet d'un stimulus à l'échelle d'une population d'individus. Par exemple, dans le contexte de la microscopie à haut débit, une population de cellules est exposée à différents stimuli. Les variations d'une cellule à l'autre rendent la comparaison de différents traitement non triviale. Un modèle génératif est proposé pour attaquer ce problème et ses propriétés sont étudiées sur la base de données expérimentales.

A l'échelle moléculaire, des comportements complexes émergent de cascades d'interactions non linéaires entre différentes espèces moléculaires. Ces non linéarités engendrent des problèmes d'identifiabilité du système. Elles peuvent cependant être contournées par des plans expérimentaux spécifiques, un des champs de recherche de la biologie des systèmes. Une stratégie Bayésienne itérative de plan expérimental est proposée est des résultats numériques basés sur des simulations in silico d'un réseau biologique sont présentées.

**Mots clés :** Apprentissage statistique, biologie computationnelle, conception de médicaments, microscopie haut débit, biologie des systèmes.

## Applications of machine learning in computational biology

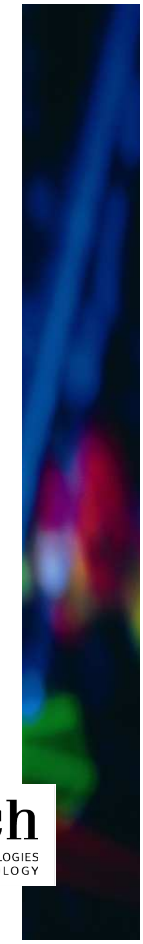
**ABSTRACT :** Biotechnologies came to an era where the amount of information one has access to allows to think about biological objects as complex systems. In this context, the phenomena emerging from those systems are tightly linked to their organizational properties. This raises computational and statistical challenges which are precisely the focus of study of the machine learning community. This thesis is about applications of machine learning methods to study biological phenomena from a complex systems viewpoint. We apply machine learning methods in the context of protein-ligand interaction and side effect analysis, cell population phenotyping and experimental design for partially observed non linear dynamical systems.

Large amount of data is available about marketed molecules, such as protein target interaction profiles and side effect profiles. This raises the issue of making sense of this data and finding structure and patterns that underlie these observations at a large scale. We apply recent unsupervised learning methods to the analysis of large datasets of marketed drugs. Examples show the relevance of extracted information which is further validated in a prediction context.

The variability of the response to a treatment between different individuals poses the challenge of defining the effect of this stimulus at the level of a population of individuals. For example in the context High Content Screening, a population of cells is exposed to different stimuli. Between cell variability within a population renders the comparison of different treatments difficult. A generative model is proposed to overcome this issue and properties of the model are investigated based on experimental data.

At the molecular scale, complex behaviour emerge from cascades of non linear interaction between molecular species. These non linearities leads to system identifiability issues. These can be overcome by specific experimental plan, one of the field of research in systems biology. A Bayesian iterative experimental design strategy is proposed and numerical results based on in silico biological network simulations are presented.

**Keywords** : Machine learning, Computational Biology, Drug Design, High Content Screening, Systems biology



MINES ParisTech

ParisTech
INSTITUT DES SCIENCES ET TECHNOLOGIES
PARIS INSTITUTE OF TECHNOLOGY