Doctoral Dissertation
Doctoral Program in Control and Computer Engineering (33^{th} cycle)

Algorithms for complex systems in the life sciences

AI for gene fusion prioritization and multi-omics data integration

Marta Lovino

* * * * * *

Supervisors
Prof. E.Ficarra, Supervisor
Prof. E.Macii, Supervisor

Doctoral Examination Committee:
Prof. Piero Fariselli, Referee, Università degli Studi di Torino
Prof. Michele Caselle, Referee, Università degli Studi di Torino
Prof. Benno Schwikowski, Pasteur Institute, Paris
Prof. Laura Cantini, IBENS, Paris
Prof. Andrea Giuseppe Bottino, Politecnico di Torino

Politecnico di Torino
June 22^{nd}, 2021

I hereby declare that, the contents and organisation of this dissertation constitute my own original work and does not compromise in any way the rights of third parties, including those relating to the security of personal data.

. . . . . .

Marta Lovino

Turin, June 22nd, 2021

Summary

Due to the continuous increase in the number and complexity of the genomics and biological data, new computer science techniques are needed to analyse these data and provide valuable insights into the main features. The thesis research topic consists in designing and developing bioinformatics methods for complex systems in life sciences to provide informative models about biological processes. The thesis is divided into two main sub-topics. The first sub-topic concerns machine and deep learning techniques applied to the analysis of aberrant genetic sequences like, for instance, gene fusions. The second one is the development of statistics and deep learning techniques for heterogeneous biological and clinical data integration.

Referring to the first sub-topic, a gene fusion is a biological event in which two distinct regions in the DNA create a new fused gene. Gene fusions are a relevant issue in medicine because many gene fusions are involved in cancer, and some of them can even be used as cancer predictors. However, not all of them are necessarily oncogenic. The first part of this thesis is devoted to the automated recognition of oncogenic gene fusions, a very open and challenging problem in cancer development analysis.

- In this context, an automated model for the recognition of oncogenic gene fusions relying exclusively on the amino acid sequence of the resulting proteins has been developed. The main contributions consist of: 1. creation of a proper database used to train and test the model; 2. development of the methodology through the design and the implementation of a predictive model based on a Convolutional Neural Network (CNN) followed by a bidirectional Long Short Term Memory (LSTM) network; 3. extensive comparative analysis with other reference tools in the literature; 4. engineering of the developed method through the implementation and release of an automated tool for gene fusions prioritization downstream of gene fusion detection tools.

- Since the previous approach does not consider post-transcriptional regulation effects, new biological features have been considered (e.g., micro RNA data, gene ontologies, and transcription factors) to improve the overall performance, and a new integrated approach based on MLP has explicitly been designed.

In the end, extensive comparisons with other methods present in the literature have been made. These contributions led to an improved model that outperforms the previous ones, and it competes with state-of-the-art tools.

The rationale behind the second sub-topic of this thesis is the following: due to the widespread of Next Generation Sequencing (NGS) technologies, a large amount of heterogeneous complex data related to several diseases and healthy individuals is now available (e.g., RNA-seq, gene expression data, miRNAs expression data, methylation sequencing data, and many others). Each one of these data is also called omic, and their integrative study is called multi-omics. In this context, the aim is to integrate multi-omics data involving thousands of features (genes, microRNA) and identifying which of them are relevant for a specific biological process. From a computational point of view, finding the best strategies for multi-omics analysis and relevant features identification is a very open challenge.

- The first chapter dedicated to this second sub-topic focuses on the integrative analysis of gene expression and connectivity data of mouse brains exploiting machine learning techniques. The rational behind this study is the exploration of the capability to evaluate the grade of physical connection between brain regions starting from their gene expression data. Many studies have been performed considering the functional connection of two or more brain areas (which areas are activated in response to a specific stimulus). While, analyzing physical connections (i.e., axon bundles) starting from gene expression data is still an open problem. Despite this study is scientifically very relevant to deepen human brain functioning, ethical reasons strongly limit the availability of samples. For this reason, several studies have been carried out on the mouse brain, anatomically similar to the human one. The neuronal connection data (obtained by viral tracers) of mouse brains were processed to identify brain regions physically connected and then evaluated with these areas' gene expression data. A multi-layer perceptron was applied to perform the classification task between connected and unconnected regions providing gene expression data as input. Furthermore, a second model was created to infer the degree of connection between distinct brain regions. The implemented models successfully executed the binary classification task (connected regions against unconnected regions) and distinguished the intensity of the connection in low, medium, and high.

- A second chapter describes a statistical method to reveal pathology-determining microRNA targets in multi-omic datasets. In this work, two multi-omics datasets are used: breast cancer and medulloblastoma datasets. Both the datasets are composed of miRNA, mRNA, and proteomics data related to the same patients. The main computational contribution to the field consists

of designing and implementing an algorithm based on the statistical conditional probability to infer the impact of miRNA post-transcriptional regulation on target genes exploiting the protein expression values. The developed methodology allowed a more in-depth understanding and identification of target genes. Also, it proved to be significantly enriched in three well-known databases (miRDB, TargetScan, and miRTarBase), leading to relevant biological insights.

- Another chapter deals with the classification of multi-omics samples. The literature's main approaches integrate all the features available for each sample upstream of the classifier (early integration approach) or create separate classifiers for each omic and subsequently define a consensus set rules (late integration approach). In this context, the main contribution consists of introducing the probability concept by creating a model based on Bayesian and MLP networks to achieve a consensus guided by the class label and its probability. This approach has shown how a probabilistic late integration classification is more specific than an early integration approach and can identify samples out of the training domain.

- To provide new molecular profiles and patients' categorization, class labels could be helpful. However, they are not always available. Therefore, the need to cluster samples based on their intrinsic characteristics is revealed and dealt with in a specific chapter. Multi-omic clustering in literature is mainly addressed by creating graphs or methods based on multidimensional data reduction. This field's main contribution is creating a model based on deep learning techniques by implementing an MLP with a specifically designed loss function. The loss represents the input samples in a reduced dimensional space by calculating the intra-cluster and inter-cluster distance at each epoch. This approach reported performances comparable to those of most referred methods in the literature, avoiding pre-processing steps for either feature selection or dimensionality reduction. Moreover, it has no limitations on the number of omics to integrate.

Acknowledgements

If I wanted to summarize the Ph.D. path, I would use the word growth. Growth as such a process takes time and dedication. Time to sink your roots in knowledge and dedication to extend your branches where I would not have thought.

In the similarity with a growing tree, the first thank goes to the gardener, Prof. Elisa Ficarra, for putting support posts, watering the soil, and pruning the dead branches. If I think the structure of this tree is composed and ordered, I owe it to her.

With her, I thank Prof. Enrico Macii too and the professors of the EDA research group.

In the last year, a vital graft has been brought by the Signaling and Development in Brain Tumor research group of the Institute Curie in Paris led by Director Olivier Ayrault.

At his laboratory, I was immediately welcomed with great enthusiasm, and I increased many of my skills. A grateful thank goes in particular to Loredana, who introduced me to this reality and above all to Olivier for taking care of the graft, making sure it took root well. Thanks go to the whole research group and those I worked most closely with, Flavia, Jacob, and Gabriele.

I also thank the various professors with whom I have had the opportunity to work over the years, particularly Prof. Giansalvo Cirrincione with his collaborators.

On a personal basis, I thank the EDA research group members and the doctoral students of other groups, particularly Davide. Discussing the technical aspects allowed us to grow together, be supported by older colleagues, and support the youngest.

A special thank goes to the students and undergraduates with whom I have enjoyed interacting over the years. It is a great satisfaction to have contributed to your growth and to see some of you become our collaborators.

Continuing the similarity with a tree's growth, my family, who patiently waited for the sprout to fortify, had a fundamental role. Thanks to them for being close to me on this journey. Your support has been essential.

Thanks to Marco, who moves and makes the branches of the tree shine like a light breeze.

To my parents

Contents

List of Tables xiii

List of Figures xv

1 Introduction 1

2 Background 5
2.1 Computational methods in biology . . . 5
2.1.1 Deep learning on genomic sequences . . . 6
2.1.2 Deep learning for the prioritization of gene fusions . . . 7
2.1.3 Learning connectivity in the mouse brain . . . 9
2.1.4 Algorithms for multi-omics data integration . . . 10
2.2 Main data types . . . 11
2.2.1 Gene fusions data . . . 11
2.2.2 Omics data . . . 12

I Gene Fusions Prioritization 17

3 Gene fusion prioritization based on the genomic sequence 19
3.1 Methodological contribution . . . 19
3.2 Introduction . . . 20
3.3 Data . . . 23
3.3.1 Training set . . . 23
3.3.2 Data-set 1 . . . 23
3.3.3 Data-set 2 . . . 24
3.4 Method . . . 24
3.5 The tool . . . 26
3.5.1 Inference mode (Default one) . . . 26
3.5.2 Retraining mode . . . 29
3.6 Results . . . 30
3.7 Additional experiments . . . 31
3.7.1 Case study . . . 31

3.7.2 *NotOnco* dataset . . . 32
3.8 Conclusions . . . 33

4 Identifying the oncogenic potential of gene fusions exploiting miRNAs 35
4.1 Background . . . 35
4.2 Methods . . . 37
4.2.1 Feature selection . . . 37
4.2.2 Dataset . . . 38
4.3 Results . . . 39
4.3.1 Architecture overview and results on the test set . . . 39
4.3.2 miRNA impact on the classification performance . . . 40
4.3.3 Comparison with state of the art . . . 41
4.3.4 Case study . . . 44
4.4 Discussion . . . 45
4.5 Conclusions . . . 48

II Multi-Omics **51**

5 Automated Prediction of Connectivity between Mouse Brain Regions 53
5.1 Methodological contribution . . . 53
5.2 Introduction . . . 54
5.3 Overview on the proposed method . . . 57
5.4 Material . . . 58
5.4.1 Allen Mouse Brain Atlas . . . 58
5.4.2 BAMS . . . 59
5.5 Method . . . 60
5.5.1 Download of Grid Data . . . 60
5.5.2 Generation Source-Target vectors and corresponding connectivity labels . . . 65
5.5.3 MLP Predictive Model . . . 67
5.6 Results . . . 68
5.6.1 Classification performance . . . 68
5.7 Conclusions . . . 72

6 miRNA-target predictions in a multi-omics dataset 75
6.1 Methodological contribution . . . 75
6.2 Introduction . . . 75
6.3 Methods . . . 77
6.4 Results . . . 79

6.4.1 Results on breast cancer dataset . . . 79
6.4.2 Validation on medulloblastoma dataset . . . 81
6.5 Discussion . . . 82
6.6 Conclusions . . . 84

7 The challenge of multiomics data for classification task 87
7.1 Methodological contribution . . . 87
7.2 Background . . . 88
7.3 Methods . . . 89
7.3.1 Biological data . . . 89
7.3.2 Classification assessment: late integration using different classification models . . . 92
7.3.3 Early Integration . . . 95
7.4 Results . . . 96
7.4.1 Late integration . . . 96
7.4.2 Classification assessment: late integration using different classification models . . . 98
7.4.3 Early Integration . . . 99
7.4.4 Performances of late and early integration method on independent datasets . . . 99
7.5 Discussion . . . 100
7.6 Conclusions . . . 101

8 The challenge of multi-omics data for clustering 103
8.1 Methodological contribution . . . 103
8.2 Introduction . . . 104
8.3 Background . . . 105
8.3.1 Joint Dimensionality Reduction for Data Fusion . . . 107
8.4 The NGL-F neural network . . . 107
8.5 Experiments . . . 109
8.5.1 Adjacency Matrix Based Comparison . . . 111
8.5.2 jDR Based Method Comparison . . . 113
8.5.3 Final Considerations . . . 114
8.6 Conclusions . . . 114

9 Conclusions 117
9.1 Global considerations . . . 119

A List of the published works 121

Bibliography 123

List of Tables

2.1 Example of a typical gene fusion detection tool's output. 12

3.1 Example of the general input file format, in case the user would like to process gene fusions obtained with a gene fusion detection tool different than the supported ones. The first two columns refer to chromosome number and breakpoint coordinate of 5p gene, while third and fourth columns refer to 3p gene. 27

3.2 Relevant fields in the DEEPrior output file for the Inference mode. Fusion Pair indicates the common names of the genes involved in the fusion; Onc Prob is the oncogenic probability value reported by the tool; Main Protein Length is the length of the fused protein; Trunc Protein reports if the fused protein is truncated (an early stop codon occurs in the protein) or not; 5p gene comp indicates if 5p gene is complete in the fusion (stop codon of the upstream gene is present in the protein); 3p gene compl indicates if 3p gene is complete in the fusion (start codon of the downstream gene is present in the protein); Main Protein is the protein reconstructed by DEEPrior. 5p and 3p gene info fields stand for a list of many other useful information about the genes involved in the fusion. 28

3.3 Example of the input file in the retraining mode, in case the user would like to include in the prediction model new validated gene fusions (e.g. a new cancer or new gene fusion variants) The first two columns refer to chromosome number and breakpoint coordinate of 5p gene, while third and fourth columns refer to 3p gene. *label* column must be 0 if the gene fusion is related to the not oncogenic class, 1 otherwise. 30

3.4 Sample tissue type (breast or prostate), sample SRA accession, highly probable oncogenic gene fusion identified by DEEPrior in that sample and validated label. More in detail, I checked if the reported gene fusion has been validated in studies [61] and [216]. *Unknown* label in the *Validated* column means that the gene fusion was not considered for validation in studies [61] and [216]. 32

5.1 Training parameters for multi-class classification with the Nadam optimizer. . . . 70
5.2 Confusion matrix for multi-class classification . . . 70
5.3 Quality metrics for multi-class classification . . . 71
5.4 Training parameters for binary classification with the Nadam optimizer 71
5.5 Quality metrics for binary classification . . . 72
6.1 **Breast cancer.** Overlap and statistic measures to test the enrichment in the three databases: miRDB, TargetScan, mirTarBase. . . 80
6.2 **Medulloblastoma cancer.** Overlap and statistic measures to test the enrichment in the three databases: miRDB, TargetScan, mirTarBase. . . . 81
7.1 Structure of each MLP node used to build the tree-MLP architecture. X size is the total numer of features for mRNA, meth, and miRNA data. The y size depends on how many classes must be predicted (2 for root MLP and healthy leaf MLP, and 3 for tumor leaf MLP). . . 91
7.2 Late Integration using MLP model with PCA dimensionality reduction technique. . . . 97
7.3 Late Integration using MLP model with ICA dimensionality reduction technique. . . . 97
7.4 Comparison between PCA and ICA preprocessed methods on the kidney test set using MLP model. All the reported metrics are computed on *not Unknown* samples. The support metric (i.e. the number of not Unknown samples) is the value on which the other metrics are computed. . . . 97
7.5 Comparison between all the methods on the kidney test set. All the reported metrics are computed on *not Unknown* samples. The support metric (the number of not Unknown samples) is the value on which the other metrics are computed. . . . 98
7.6 Early Integration using MLP model . . . 99
7.7 Percentage of samples predicted as *Unknown* in the stomach and lung datasets. . . . 100

List of Figures

2.1 Typical tabular file. Features (in this case, genes) are reported on the rows, samples (in this case, patients) on the columns. x_{ij} represents the expression value of feature i in patient j. 13

3.1 Architecture of the deep learning model in DEEPrior. 25

3.2 **Workflow of DEEPrior tool.** For each gene fusion (see different colors in the figure), DEEPprior generates all possible proteins, considering all transcripts of the fused genes. In the end, the amino-acid sequences are fed into the deep learning model, obtaining a 0-1 value for each protein. The oncogenic probability of each gene fusion is obtained as the maximum of all these values. 26

4.1 Confusion matrices reporting the MLP results including miRNAs (on the left) and excluding miRNA features (on the right). 41

4.2 The green bars correspond to the results reported by Shugay M. et al. in their paper. In blue the results obtained by ChimerDriver are displayed. 42

4.3 The 24 oncogenic gene fusions validated in prostate and breast tumor samples are reported. STAR-fusion did not detect the three gene fusions marked in gray hence were not available to ChimerDriver for further processing. ChimerDriver correctly classified as oncogenic 18 out of the 21 available gene fusions. 45

4.4 Here I report the distribution of the false positives (FP), false negatives (FN), true positives (TP) and true negatives (TN) regarding Cancermine information for both 5p' and 3p' genes(respectively Figure 4.4a) and 4.4b)). Noticeably, FPs are never tumor suppressors, drivers or oncogenes. 48

5.1 Scheme of the analysis pipeline. For each Source-Target pair (N, in total), I randomly select M voxel combinations. Per each combination, I generate two 3,318 gene expression vectors (Source and Target, respectively) with information taken from AMBA. The concatenation of the two vectors represents the Source-Target vector given as input to our MLP model. A categorical label describing the Source-Target connection degree is obtained by setting empirical thresholds on the connectivity values provided by MBCA. 58
5.2 Schematic diagram of Source-Targets projections. A Source is a brain region where the viral tracer is injected (*inj*, in the figure). As a result of the injection, multiple axonal projections are produced in so-called Target regions. 59
5.3 Reference space 62
5.4 Elaboration of gene expression data: main phases. (1) Retrieve a SectionDataset for each of the 3318 genes; (2) download grid expression data in the form of an energy.raw file; (3) reconstruct a 3,318×159,326 matrix of gene expression levels, with rows corresponding to genes and columns to 3D voxels; (4) store data into a .csv file. 63
5.5 Elaboration of connectivity data: main phases. (1) Retrieve a SectionDataset for each of the 2333 primary injections; (2) projection grid-data in the form of an .Nrrd file; (3) reconstruct a projection volume, unpacked into a vector of 1,203,840 elements; (4) store data into a .csv file. 64
5.6 Interactive matrix from BAMS. Each element of the matrix represents the connection between two regions, reported in rows and columns. Different colours encode different connection intensities, with white corresponding to unknown connections. 65
5.7 SQLite database tables generated to store all the gene expression and connectivity data. 67
5.8 MLP architecture for classification tasks. For each layer, I report number of nodes, activation function and dropout value. When values are different for binary and multi-class tasks, we report them both, separated by a slash symbol. 68
5.9 Training performance curves (loss on the left, accuracy on the right) of the binary classifier. 72
5.10 Receiver operating characteristic (ROC) curve on the test set for binary classification. 73
6.1 Partial correlation estimate value e_{ij} computed for gene i and miRNA j. These values are easily stored in a table. 78
6.2 Partial correlation p-value p_{ij} computed for each gene i and miRNA j. These values are easily stored in a table. 78

6.3 Breast cancer. Histogram of partial correlation estimates. 80
6.4 Breast cancer. Histogram of partial correlation p-values in logarithmic scale (-log10). 80
6.5 Medulloblastoma cancer. Histogram of partial correlation estimates. 81
6.6 Medulloblastoma cancer. Histogram of partial correlation p-values in logarithmic scale (-log10). 81
6.7 Visual representation of the sets. The blue set contains all possible gene-miRNA couples available in the dataset, and the red area contains the gene-miRNA couples selected by the method as significant. The green set represents all the gene-miRNA couples available in a database (e.g., miRDB). Consequently, gene-miRNA couples selected as relevant by the method can be present or not in the database. The hypergeometric test computes the significance of this overlap. 83
6.8 **Breast cancer heatmap** of selected miRNA targets. Gray areas represent selected miRNA-gene target. Red areas represent a miRNA-gene target validated in one of miRDB, TagetScan, miRTarBase. Genes and miRNAs are sorted according to their genomic coordinates. 85
6.9 **Medulloblastoma cancer heatmap** of selected miRNA targets. Gray areas represent selected miRNA-gene target. Red areas represent a miRNA-gene target validated in one of miRDB, TagetScan, miRTarBase. Genes and miRNAs are sorted according to their genomic coordinates. 86
7.1 Proposed tree MLP model: i) each node is trained on three different subsets of the original dataset. (X, y) aims to distinguish between healthy and tumor samples, (X', y') between subtypes of healthy samples and (X", y") between subtypes of tumor samples; ii) the output of each node consists of the predicted label y_{pred} and the class-membership probability P. 95
8.1 NGL-F architecture: N datasets are fed in input to NGL-F. For each dataset, a multi-layer perceptron employed and customized according to dataset complexity. Clustering outputs are at the end combined in order to create a sample graph built from the adjacency matrix S. 108

8.2 NGL-F network architecture as used in the experiments. Between brackets, the dimensionality of input/output data of each layer is reported. Regarding the matrices, the dimensions are defined as features x samples since the matrix is transposed. Instead, each dense and output layers is reported the dimensionality of the associated weight matrix. It should also be noticed the different dimensionality of the two input sources, miRNA (top) and mRNA (bottom) maintained through the layers. 110
8.3 Adjacency matrix of the sample using (left) SNF and (right) NGL-F algorithms 112
8.4 Kamada-Kawai path-length graph of the sample adjacency matrix computed by NGL-F (left) and SNF (right) algorithms. 112
8.5 Harmonic mean of cluster efficiency and purity computed on the spectral clusters, computed on the adjacency matrix produced by the different algorithms. 113

Chapter 1

Introduction

In life sciences, understanding and analyzing the main biological phenomena is of crucial importance. A myriad of data ranging from genomics, image analysis, medical record traces, and various state-of-the-art databases can describe biological phenomena [229]. Therefore, tackling a biophysical problem involves information of different types. In this sense, it is necessary to consider what data is available to analyze the actual problem and which computational issues it involves[101].

Besides the diversity of information, biological systems are inherently involved as they are made up of multiple components. Some of these are regulatory aspects and are well known, while the role of other molecules and their interactions is still a large area of discovery [138].

Therefore, it is necessary to consider the information available and the meaning of interpreting and analyzing a complex system. Complex systems typically require complex modeling to obtain biological information relevant and specific to the problem[125, 60].

Among the many types of data concerning biological processes, spatial data, information from various databases, and omics data have been considered in this thesis.

The term omics refers to a set of complex data that includes, for example, genomics (i.e., information on DNA), transcriptomics (mainly RNA), proteomics, and phospho-proteomics (proteins)[111, 29]. Therefore, some complex computational approaches will be presented to process biological data and significantly contribute to significant biological problems. These methods involve both analytical techniques and machine and deep learning techniques for interpreting various complex systems.

The first part of this thesis is devoted to studying and analyzing gene fusions,

a biological phenomenon in which, following a genetic alteration, two genes join and can generate an oncogenic protein molecule. However, not all gene fusions are oncogenic. Therefore it is a significant challenge to create automatic tools for the prioritization of gene fusions, i.e., to show which of these have a higher probability of being involved in oncogenic processes.

The main methodological contribution consists of creating ad-doc models for the analysis of gene fusion proteins. These methods involve convolutional networks and long short-term memory networks applied directly to the fusion sequence instead of the gene fusion product's protein domains. These approaches outperform the main state-of-the-art tools and contribute significantly to this sector.

Also, post-transcriptional regulation features that characterize gene fusions have been considered. By exploiting ad-hoc methods for integrating this information, it has been possible to identify the relevance of microRNAs (miRNAs) to prioritize gene fusions.

A second part of the thesis is dedicated to integrating various types of data. The first chapter concerns integrating gene expression data with spatial information to predict connected and unconnected areas within the mouse brain. In this context, the main methodological contribution consisted of identifying a strategy to integrate axon connectivity data with the point values of gene expression within the mouse brain. Therefore, the developed method identifies if two brain regions are connected by axons analyzing the gene expression data alone.

Another computational problem regards integrating multi-omics data, involving gene expression (mRNA) data, microRNA expression (miRNA) data, protein expression, and DNA methylation data.

In this context of integration, two main ways are possible: the first concerns the analysis of a multi-omic dataset for the discovery of regulatory phenomena involving genes, miRNAs, and proteins, while the second consists of integrating the various features of mRNA, miRNA, and methylation to improve the classification and clustering between the samples.

To discover regulatory phenomena, a specific method that exploits partial correlation statistical test was implemented to bring out the post-transcriptional function of miRNAs concerning their target genes. The method proved to be enriched in three databases: miRDB, TargetScan, miRTarBase.

Another computational problem regards integrating multi-omics data, involving gene expression (mRNA) data, microRNA expression (miRNA) data, protein expression, and DNA methylation data.

In this context of integration, two main ways are possible: the first concerns the analysis of a multi-omic dataset for the discovery of regulatory phenomena involving genes, miRNAs, and proteins, while the second consists of integrating the

various features of mRNA, miRNA, and methylation to improve the classification and clustering between the samples.

Instead, integrating different data to get single clustering information is still a challenging problem regarding clustering in a multi-omic context. In this thesis, the methodological approach is based on designing a specific loss function for a multilayer perceptron method to cluster samples summarizing the information coming from all the omics.

In the end, Chapter 2 will present the background of the main computational approaches mentioned in this thesis and, section by section, the specific data's computational problems and the various methodological contributions.

Next, Chapters 3 and 4 are dedicated to gene fusions, with the description of the various implemented approaches and the achieved results. In Chapter 6, a multi-omic integration method based on statistical techniques is presented to reveal the mRNA expression adjustments determined by the post-transcriptional action of miRNAs.

Chapters 7 and 8 deal respectively with the machine and deep learning techniques for classifying and clustering patients based on multiple omics data (in this case, genomics, transcriptomics, and proteomics).

Finally, Chapter 9 reports this thesis's conclusions, underlining each section's main innovative contributions.

Chapter 2

Background

In the last century, an exciting relationship binds biology and computational algorithms. Many of the optimization and machine and deep learning algorithms draw inspiration from biological processes. An example of this relationship consists of the genetic algorithms inspired by a biological population's evolution mechanisms to optimize the search for a satisfactory solution among the many possible ones. Another example is convolutional neural networks (CNNs) that draw inspiration from the visual cortex to learn specific patterns encoded in the images [175]. Simultaneously, computational algorithms have become necessary to process, analyze, and understand the main biological phenomena related to cell life and oncogenic processes. In the last decades, multiple algorithms have been used to unravel biological insights in complex systems, ranging from statistical to neural network approaches. In the first part of this section, the main algorithmic methods will be described, focusing on the still open challenges. This thesis's relevant data will be presented in the second part, providing a brief biological overview and their computational issues.

2.1 Computational methods in biology

Recently, machine and deep learning techniques spread in almost all fields of science and bioinformatics, exploiting the growth of data-driven methodological approaches.

Although the back-propagation algorithm was first presented in 1962 by Dreyfus, its first uses for artificial neural networks date back to 1981 by Werbos [58, 214]. The first implementation of a multilayer convolutional neural network takes place in 1989 by LeCun, in which the network was applied to actual problems of classifying handwritten images [130]. One of the most famous implementations is by LeCun in 1998, who implemented the popular LeNet-5. It is a particular convolutional network architecture dedicated to the MNIST dataset classification, which

subsequently became the most famous dataset for testing algorithms' performance for pattern recognition.

One of the critical points of convolutional neural networks is creating an accurate classification without building ad hoc features for each type of problem. These networks are based on automated learning, allowing the user to overcome one of the challenging aspects of training automatic classifiers, namely the feature extraction problem [129]. In the biomedical context, the extraction of relevant features for the classification of biological images is a particularly challenging problem. This is particularly true, for instance, for histological images. Indeed, it is difficult for the expert to identify exact measures of the tissue's morphological and structural characteristics under analysis. Besides, all known and valid rules in a healthy context are no longer proper in the pathological context, making the design of a reliable model more difficult. Another aspect to consider is the intra-class variability within the same data set, and among different data sets too.

Although the aforementioned issues, convolutional neural networks have been successfully applied to some biological context, such as the segmentation of glioma tumors in the brain by Hussain et al. [100], the prediction of retinopathy in diabetic patients [172], and the segmentation of brain tumors in MRI images [90]. Other examples include the diagnosis of breast cancer, the identification of interstitial lung diseases, and the identification and classification of nuclei in routine histological images of colon cancer [197, 14, 191].

2.1.1 Deep learning on genomic sequences

Convolutional neural networks have experienced significant growth in applications and continuous refinements. However, they are not the only models that have received substantial attention in recent years. Recurrent neural networks have been specifically designed to process sequential data and have found significant applications in natural language processing, language translation, and speech recognition [82, 187, 218, 123].

Among the many applications found in the literature, the property of considering temporal and spatial sequences has made convolutional and recurrent networks particularly suitable for analyzing biological sequences [139, 198, 91]. Several works exploit LSTM networks to identify gene motifs within the genomic sequences and analyze the structure of the regulatory sequences that characterize the DNA [118]. Other works, for example, concern the visualization of these genomic sequences to make the human genome more interpretable [126].

A field in which convolutional and recurrent methods have reached a mature

stage predicts the secondary structure of proteins starting from its genomic sequence [228]. The prediction of the secondary structure of proteins is essential. It is possible to verify protein complexes' interactions and identify potential drugs binding to specific proteins.

Also, various papers focus on predicting the specificities of DNA, and RNA binding proteins [220, 6], enabling personalized medicine and the discovery of new potential therapeutic drugs.

2.1.2 Deep learning for the prioritization of gene fusions

One area in which computer algorithms have contributed significantly to biology is that of the study of cancer, which afflicts millions of people every year and is the second leading cause of death immediately after cardiovascular diseases[105, 176]. Cancer can be characterized by a progressive accumulation of mutations at the genomic level. This accumulation of mutations can take various forms, including deletions or insertions or point mutations in the coding sequence, amplifying portions of the genetic code, and chromosomal rearrangements. In particular, chromosomal rearrangements can cause the juxtaposition of enhancer elements or the creation of a fusion gene [151, 148].

Over time, fusion genes have been associated with oncogenic processes. Some of these are exploited as biomarkers for particular types of tumors. Particularly famous gene fusions are those involving the BRCA1 gene for breast cancer prognosis, the TMPRSS2-ERG genes that identify 50% of prostate cancers [54], EWS-FLI1 characteristic of Ewing's sarcoma [205]. Gene fusions are made up of the juxtaposition of two different genes. Depending on the genes involved, their functionality, and the mutations of the genes mentioned above, the gene fusions can have an oncogenic effect strongly characterizing a particular pathology. At the same time, not all gene fusions are always associated with oncogenic phenomena. Hence the fundamental need to identify correctly those fusion genes associated with a higher probability of oncogenic processes. It is essential to recap briefly how such biological alterations are identified to understand building algorithms' computational challenges.

Commonly, in a fusion genes detection routine, the RNA is extracted from the biological sample and inserted into a genomic sequencing machine. This machine's output consists of a .fasta or .fastq file containing all the biological sample reads. Then, gene fusions detection tools are used to have a list of candidate gene fusions for that sample. Among the most famous tools, ChimeraScan, Defuse, STARfusion, FusionHunter, Tophat-fusion and SOAP-fusion deserve to be mentioned [102, 153, 56, 85, 137, 120, 217]. These tools aim to identify most of the gene fusions

present in the sample supplied to the algorithm. Some of these tools have a high sensibility with low specificity (i.e., ChimeraScan), others have a low sensibility with high specificity (i.e., FusionHunter). Gene fusion detection tools can report false positives, which are gene fusions that are not validated once analyzed by PCR in a biological laboratory. Also, the fact of finding a gene fusion within a sample and validating it with PCR does not necessarily imply that this fusion gene is the driver of a particular oncogenic pathology.

Therefore, this assumption implies that not all the gene fusions detected by a gene fusion detection tool are oncogenic and drivers in pathological processes given a biological tumor sample. Although the number of gene fusions drivers of oncogenic processes is continuously growing and are reported in many databases, their number is still limited [70, 131]. Also, many PCR-validated gene fusions are present in tumor samples, but in this case, the driver oncogenic function is supposed. However, in all practical purposes, gene fusions found in healthy samples are considered as non-oncogenic while the PCR-validated fusion genes in tumor samples are labeled as oncogenic.

Various approaches have been proposed over time to recognize among the multiple gene fusions those responsible for oncogenic processes, starting from the manual analysis of the genes involved in the fusion to develop automatic methods based on machine learning techniques [26, 155]. In both approaches, the simplification of labeling as healthy and oncogenic the gene fusions coming respectively from healthy and tumor samples is used only in the computational models' validation phase. This assumption would be insufficient to train efficient machine learning models. Therefore, only gene fusions with an irrefutable oncogenic driver capability are used to train the automatic models.

The oncogenic driver capability's certainty appears to be one of the most problematic aspects in the realization and training of the automatic models from the computational point of view. Machine learning models allow training and testing of the models provided that a fair number of high-quality samples are available.

In the literature, machine learning models to identify gene fusions are based on the analysis of protein domains held within the fusion to train naïve Bayesian and random forest classifiers to determine a fusion's oncogenicity. In this thesis [3, 189], deep learning methods applied directly to the genomic sequence of the fusion gene will be addressed to prioritize potentially oncogenic gene fusions.

From a computational point of view, besides the challenge given by the certainty in the oncogenic driver capability, a crucial aspect is the information in the fusion gene sequence. Although various deep learning algorithms have proved helpful in

predicting genomic patterns, so far, it has not been shown how much the genomic sequence of a gene fusion alone could be sufficiently predictive of such a complex phenomenon as the oncogenicity of a gene fusion.

Also, the specific case of gene fusion analysis is challenging as the genes involved in the training database are entirely disjoint from the genes used in the test database. This aspect is standard in the definition of automatic algorithms; however, it is required that the test set's data belong to the same domain and have the same characteristics as the data of the training set. It is usually required that the test set data are independent of those of the training set but share the same properties. As described above, it is not known how much the oncogenicity property is represented by the genomic sequence alone. Indeed, by completely changing the genes used for the test set, there is no guarantee that the working domain between train and test set is homogeneous.

In this thesis, the previously described challenges have been collected by showing deep learning algorithms based on convolutional and recurrent networks to perform the prioritization of gene fusion and thus identify the most likely oncogenic ones.

2.1.3 Learning connectivity in the mouse brain

A diversity of heterogeneous and different data has become available regarding the same context or pathology in the biological field. Integrating the data mentioned above is always an exciting and challenging activity, as integration methods depend on the data itself and the result obtained. It is common to have both genomic information and images available. In this context, machine and deep learning models find application in the translation of information between different domains, thus allowing for transferring knowledge from one domain to another [15, 88]. In this thesis, one aspect addressed was integrating genomic and spatial data regarding the mouse brain. In particular, the work aimed to identify a physical connection between two or more regions of the mouse brain, starting from the genomic data alone.

From a computational point of view, the main challenges of this section consist of mapping genomic information in the three-dimensional space together with spatial images representing the intensity of a viral tracer to infer which axons connect specific regions of the brain to the central nervous system.

In this context, the significant difficulty is finding the information to create a predictive model and establish the physical intensity of the neural connection starting from the gene expression data only in the three-dimensional form.

As discussed in the previous section regarding the genomic sequence, also, in this case, the direct connection between gene expression and the physical connectivity of two or more regions in the brain is not known. Indeed, this connection still needs to be further investigated.

2.1.4 Algorithms for multi-omics data integration

In biology, it is expected within a specific pathology finding more molecular subtypes. Patients with the same clinical manifestation can have a different genetic heritage and response to drugs. Therefore, the treatment of the disease follows specific protocols and may affect the duration of survival. Although the identification of molecular profiles is possible sometimes by yielding a single omic data (for example, transcription), the joint integration of several omic data allows most times to carry out integrated and more detailed profiling.

Although integrating different data is a strength in the biological field, the computational effort necessary to obtain helpful information starting from genomic data in an integrated form is considerable [223, 196]. From a computational point of view, each data has specific dynamics of the values (range values). It is fundamental to understand how to scale the input value to not condition the result based on a single omic.

The most frequently used omics comprise the transcriptome, micro RNA (miRNA), proteome, methylation data, single point mutation, and many others in the genomic field. Their high dimensionality characterizes all these omics. Omics data have tens to hundreds of thousands of features for each sample [9].

More detailed information on the type and dimensionality of the data processed is shown in Section 2.2.

Moreover, the complexity of integrating multi-omics data derives not only from the intrinsic dimensionality of each omic but from the modeling of a biological phenomenon that is complex by its nature, whose potential is not yet fully known. Therefore, it is necessary to clearly understand the aim of the multi-omics integration to filter out the noise and work with the signal of interest [30].

From a computational point of view, the information coded at the multi-omic level is sometimes redundant. The selection from those redundant parts is not always automatic, on the contrary.

When working with multi-omics datasets, two main strands can be identified. The first strand consists of the samples' integrated analysis to infer the subtype

or sample classification. The second strand instead identifies genes, miRNAs, proteins, or methylation probes closely related to the pathology in question. Both these strands are affected by dimensionality and the complexity of the system under examination[168].

In this thesis, both aspects have been considered by presenting works of classification and multi-omic clustering of patients and the extraction of particular features, particularly miRNA, to investigate specific biological problems.

2.2 Main data types

In this section, the primary data used in this thesis will be summarized. The aim is not to provide a detailed description of each data but to summarize each data's origin and meaning. In particular, I will not focus on the biological and biotechnological details but on the computational problems that may arise from the data. A more precise overview will be reserved for the number of features and samples available for each of this information. Section 2.2.1 will concern the data used in this thesis regarding the problem of gene fusions. Subsequently, the main multi-omics data will be detailed in Section 2.2.2.

2.2.1 Gene fusions data

Gene fusions can be studied considering various biological information, such as analyzing the translocations in wet laboratories or using automated tools that allow the identification of candidate gene fusions in fasta/fastq samples.
Because of the cost reduction of NGS sequencing experiments, the availability of RNA sequencing data has become relevant. Subsequently, several tools were explicitly designed to infer the candidate fusions for each sample [102, 153, 85]. Although the strategies used by the gene fusion tools are partially different, all these align the reads of the fasta/fastq files on the reference genome and subsequently identify those reads that are a symptom of a fusion phenomenon. Typically, the result of such software comprises a tabular file showing the genes involved in the fusion (gene at 5p 'and gene at 3p') together with the coordinate of the breakpoint on both genes plus other biological information related to alignment. An example of this type of data is shown in Table 2.1.

Within this thesis, all the information related to gene fusions is obtained from the chromosomes and each fusion's breakpoint coordinates. However, this approach does not allow for the exploitation of additional information, such as knowing which

5p' gene	3p' gene	5p' chr	5p' coordinate	3p' chr	3p' coordinate	Additional output
TMPRSS2	ERG	21	41,480,305	21	38,567,261	...
ACACA	STAC2	17	37,184,992	17	39,215,536	...
RPS6KB1	SNF8	17	59,933,046	17	48,929,555	...

Table 2.1: Example of a typical gene fusion detection tool's output.

specific transcript is involved in gene fusion. Therefore, in the absence of data on the particular transcript, all known transcripts for that pair of genes are considered by evaluating all possible transcript combinations.

From a computational perspective, the proteins resulting from the gene fusions process vary from a few amino acids to a size comparable to healthy proteins (several hundred to a few thousand amino acids). As anticipated in Section 2.1.2, the total number of oncogenic validated gene fusions is limited to a few thousand sequences. In contrast, the total number of gene fusions identified as candidate gene fusions is approximately 100 times greater.

All the details concerning the reconstruction of the fusion sequence are reported in Chapter 3.

2.2.2 Omics data

In biology, a lot of omics data and information may be available for the same patient. However, this wealth of information can be difficult to manage as the number of features available for each omic varies from hundreds to hundreds of thousands. This section illustrates the characteristics of the major data structures to deal with when analyzing omics data. The aim is to provide a general overview and illustrate the specific computational problems related to RNA, miRNA, methylation (meth), and proteomics (prot).

mRNA

The most abundant and most studied type of RNA is messenger RNA. Its measure is closely related to the number of proteins inside the cell [122]. It is possible to use mainly two techniques to quantify genes' expression level, quantification by microarray and NGS sequencing techniques. Although the technological functioning of these two platforms is very different, starting from both, it is possible to obtain a tabular file as in Figure 2.1 in which the value x_{ij} represents the amount

of expression of the gene i in patient j.

Figure 2.1: Typical tabular file. Features (in this case, genes) are reported on the rows, samples (in this case, patients) on the columns. x_{ij} represents the expression value of feature i in patient j.

miRNAs

Micro RNAs (miRNAs) are short RNA molecules mainly involved in the post-transcriptional processes of the cell [89]. They have a central region of about seven

nucleotide long (seed region) responsible for the bond that miRNA can form with some genes (typically in the region close to the promoter), thus inhibiting mRNA translation into protein. Therefore, the study and quantification of these molecules are fundamental in deciphering and revealing biological phenomena' complexity. The expression level of miRNAs is calculated through NGS techniques, and the form of such data is a tabular file (Figure 2.1) where the value at position x_{ij} represents the amount of expression of miRNA i in patient j.

Methylation data

Another biological phenomenon that plays a fundamental role during post-transcriptional processes is methylation [180]. This phenomenon allows altering the functioning of the resulting protein by inserting a methyl group near some specific regions (typically regions of DNA enriched with cytosine and guanine) without modifying the sequence of nucleotide bases. Some DNA areas have a high level of methylation on a physiological level, while other regions are typically poorly methylated. However, changes in the level of methylation can lead to altered proteins' expression, thus making it necessary to study the changes in methylation regarding basal conditions. The quantification of the methylation level occurs through tabular files (Figure 2.1) in which the x_{ij} value represents the methylation value of the DNA region (called a probe) i in patient j.

Proteomics

It is possible to state that most of the cell's life depends on its proteins, simplifying the biological processes to the extreme. It has become increasingly necessary to quantify proteins within tissue in recent years as they are directly connected to the main cellular processes [127]. Unlike DNA or RNA molecules that the PCR process can amplify, proteins cannot be amplified despite having very low concentrations within the sample. Therefore, the techniques used so far for the quantification of these molecules are based on the microarray (in the case in which the proteins to be investigated are completely known in their sequence and structure) or on the mass spectrometer, which can quantify the peptides within a sample. The reconstruction of the quantification of proteins from the quantification of the mass spectrometer's peptides requires adequate processing. At the end of this process, a tabular file (Figure 2.1) is obtained in which the value x_{ij} represents the quantity of protein i in the sample j.

General computational considerations

After presenting the omics' critical characteristics in this thesis, it is appropriate to provide some considerations regarding the main computational challenges deriving from them. The major challenge comes from the abundance (and redundancy) of the features available for each sample. Typically the number of features is about 10 or 100 times the number of samples. Therefore, proper feature selection or dimensionality reduction techniques must be considered for classification and clustering samples. Deciphering the regulatory cascades between the various genes, proteins, and miRNAs becomes complex because of the high dimensionality.

Part I

Gene Fusions Prioritization

Chapter 3

Gene fusion prioritization based on the genomic sequence

3.1 Methodological contribution

This chapter is devoted to studying and analyzing gene fusions, a biological phenomenon in which, following a genetic alteration, two genes join and can generate an oncogenic protein molecule. However, not all gene fusions are oncogenic. Therefore it is a significant challenge to create automatic tools for the prioritization of gene fusions, i.e., to show which of these have a higher probability of being involved in oncogenic processes.

The oncogenic driver capability's certainty appears to be one of the most problematic aspects in the realization and training of the automatic models from the computational point of view. Machine learning models allow training and testing of the models provided that a fair number of high-quality samples are available.

From a computational point of view, besides the challenge given by the certainty in the oncogenic driver capability, a crucial aspect is the information in the fusion gene sequence. Although various deep learning algorithms have proved helpful in predicting genomic patterns, so far, it has not been shown how much the genomic sequence of a gene fusion alone could be sufficiently predictive of such a complex phenomenon as the oncogenicity of a gene fusion.

Also, the specific case of gene fusion analysis is challenging as the genes involved in the training database are entirely disjoint from the genes used in the test database. This aspect is standard in the definition of automatic algorithms; however, it is required that the test set's data belong to the same domain and have the same characteristics as the data of the training set. It is usually required that the

test set data are independent of those of the training set but share the same properties. As described above, it is not known how much the oncogenicity property is represented by the genomic sequence alone. Indeed, by completely changing the genes used for the test set, there is no guarantee that the working domain between train and test set is homogeneous.

In this context, an automated model for the recognition of oncogenic gene fusions using exclusively on the amino acid sequence of the resulting proteins has been developed. The main contributions consist of: 1. creation of a proper database used to train and test the model; 2. development of the methodology through the design and the implementation of a prediction model based on a Convolutional Neural Network (CNN) followed by a bidirectional Long Short Term Memory (LSTM) network; 3. extensive comparative analysis with other reference tools in the literature; 4. engineering of the developed method through the implementation and release of an automated tool for gene fusions prioritization downstream of gene fusion detection tools.

3.2 Introduction

Nowadays, the increased availability of Next Generation Sequencing (NGS) data enables new unforeseen insights into the relationship between some genetic rearrangements and cancer development. In this regard, a challenging area is represented by the study of gene fusions. In this genetic aberration, two separate DNA regions (usually two distinct genes) join together into a hybrid gene. The genes retained at 5p' and 3p' of the fused sequence are conventionally called 5p' gene and 3p' gene, respectively. If the promoter region of at least one of the two genes is retained in the fusion, the erroneous sequence is transcribed at the RNA level, and the aberrated transcript can result in an abnormal protein[156].
Since the discovery of the first genetic rearrangement by Nowell and Hungerford in 1960, a large number of gene fusions have been associated with cancer development and used as cancer predictors[156]. However, gene fusions do not automatically relate to oncogenic processes, as they can be found in large numbers even in non-tumoral samples [19]. Therefore, predicting whether an altered transcript will result in an oncogenic protein is a very critical and challenging task in the study of cancer development.

Traditionally, many methodologies have been used for the identification of fusion genes (e.g. fluorescence in situ hybridization (FISH) [10] or comparative genomic hybridization (CGH) [112]). In recent years, the spreading of NGS technologies has enabled gene fusion detection tools, whose aim is to identify chimeric transcripts exploiting information coming from RNA paired-end sequencing data [185].

Typically, the analysis of such data consists of three main phases:

1. primary identification of candidate gene fusions,
2. filtering of the fusion candidates, based on the number of reads mapping to a specific region and the functional annotation of the involved genes. The outcome of this phase is a sub-set of candidates with the best read quality mappings and the highest probability of resulting in a functional oncogenic product,
3. in-situ validation of the fusions resulting from phase 2.

The first phase of the analysis is performed using fusion detection tools (among the others, Chimerascan [102], Defuse [153], Prada [206] and many more [163, 108, 2]). Nowadays, the major issue with these tools' outcome is related to the interpretation of the found chimeric transcripts. Given each gene fusion's high validation costs, extensive post-processing efforts are devoted to distinguishing driver fusions from passenger mutations to reduce the number of false positives in the last part of the pipeline. This approach makes the second phase of the analysis particularly critical and challenging.

Indeed, the majority of the tools in the literature apply filtering criteria based on the read mapping quality (among the others, Tophat-fusion [120] and Star-fusion[85]). A complementary approach for the interpretation of gene fusion candidates consists of a functional study of the chimeric transcript, looking at possible similarities with cancer genes: the higher the similarity, the higher the probability of developing into cancer. This similarity analysis involves specific functional annotations, protein interactions as well as protein domain analysis [128].

With the aim to achieve a fully functional study of a chimeric transcript, all the available literature approaches reconstruct the candidate fusions and then apply different types of machine learning methods to perform protein domain analysis [189, 3]. Given the uncertainty on the training set, these tools mainly use predictive models to derive conserved and lost protein domains in fusions and then exploit the outcome of such predictions to train a machine learning method. The most popular tool in this category is Oncofuse [189] that assigns a functional prediction score (oncogenic potential, i.e., the probability of being driver events) to the fusion sequences exploiting a naive bayesian classifier.

While the information on conserved or lost protein domains is generally successful in prioritizing the candidate fusions, this approach's well-known drawback is its lack of flexibility. Indeed, any change in the classification problem (either a

different type of cancer or newly acquired information) requires significant efforts devoted to re-parametrizing the model and laborious re-derivation of the protein domains. This process is a very inconvenient trait, especially if I consider that the study of cancer development is built on top of continuously evolving information.

In this work, I focused on the functional annotation of the chimeric transcript (phase 2 of the analysis pipeline) and a more flexible approach. I exploit human reference sequences, relying only on the raw fusion sequence information, with no additional input about conserved or lost protein domains. The aim is to avoid any possible bias that the prediction models leveraged by protein domain analysis may introduce into the classification task and improve the generalization capabilities and ease-of-retraining of the classifier.

The proposed solution is based on a Convolutional Neural Network (CNN) and a bidirectional Long Short Term Memory (LSTM) network to handle the prioritization problem. CNN is a class of deep, feed-forward neural networks with the inbuilt ability to automatically learning the most significant classification features directly from the raw input data [221, 222]. Hence, they avoid the necessity of designing handcrafted descriptors, which may be difficult to generalize to different classification problems. Thanks to these peculiar characteristics, they can quickly adapt to newly acquired information by merely re-running the automated back-propagation algorithm on the new training data. Initially designed for image classification tasks, CNNs and LSTM are now successfully applied to most pattern recognition and classification problems, from computer vision [55] and natural language processing [49] to bioinformatics (for example, to the prediction of single-cell DNA methylation states and microRNA targets, as well as to the recognition of splice junction sites and promoter sequence regions [158]).

In this work, I feed the model directly with the real amino-acid composition of the fused proteins, with no additional data interpretation to design a model that is entirely independent of protein domain information. The network's output consists of a $0 - 1$ score, which can be interpreted as the probability of the chimeric input transcript to be oncogenic. This score can also be translated into a categorical class label, partitioning the input gene fusions into two different groups (oncogenic or not oncogenic, respectively), with a corresponding confidence level.

The algorithm is the core part of DEEPrior, a simple and easy-to-use tool for prioritizing gene fusions. In addition to the tool, another relevant contribution consists of releasing to the scientific community a database with 4779 amino-acid protein sequences that I collected and reconstructed for this work by combining the information reported by multiple sources [70, 19, 131].

3.3 Data

Although many databases related to gene fusions have been released recently, the availability of databases reporting the proteins resulting from annotated and validated gene fusions is still a critical issue. Here:
https://github.com/bioinformatics-polito/DEEPrior/tree/master/DEEPrior/data, I release the protein fusions datasets specifically reconstructed from multiple sources and used to assess DEEPrior performances to the community. Overall, I used three datasets (one for training and two different ones for performance assessment), described with more details in the following. A label is associated to each gene fusion of the data-sets, respectively *Onco* for the oncogenic and *NotOnco* for the not oncogenic. In this thesis, a fusion pair is defined as the union of the 5p gene name with the 3p one.

3.3.1 Training set

This set consists of 786 fusion pairs and 2118 sequences, respectively 1059 *Onco* and 1059 *NotOnco*, obtained from two different sources.
The *Onco* sequences were obtained from COSMIC, Catalog of Somatic Mutations in Cancer [70]. Among all the mutations involved in oncogenic processes, COSMIC also provides a list of validated gene fusions in the Complete Fusion Export Table. Among all the instances reported, I selected only the ones for which complete information was provided about the transcripts and the exact breakpoint positions in order to be able to reconstruct the resulting amino-acid sequence.

The *NotOnco* sequences, on the other hand, were obtained from work by Babicenau et al.[19], where more than 10000 gene fusions were obtained by applying SOAPfuse gene fusions detection tool to 171 non-neoplastic tissues. Among all the gene fusions reported in the paper, I first discarded the ones not belonging to the human species or coming from cell lines (ESC, MSC, MFC10). As the *NotOnco* gene fusions were over-represented concerning the *Onco* ones by one order of magnitude, I selected the *NotOnco* gene fusions that were present in at least four different tissues or different patients.

To complete the dataset, I added gene fusions present in at least three different tissues or different patients. The selection proceeded until I obtained a total number of *NotOnco* sequences equal to the number of the *Onco* sequences.

3.3.2 Data-set 1

This set was used to test DEEPrior performances, and it is composed of 142 fusion pairs and 156 gene fusions,122 *Onco* and 34 *NotOnco*. As there are no fusion

pairs in common with the training set, this set is completely statistically independent.

Overall, the data were extracted from three different sources.

The sequences associated with *Onco* gene fusions were extracted from the ChimerDB2.0 database [121]. The genomic positions were obtained by taking the gene fusions from ChimerDB3.0-ChimerSeq [131] that originate from ChimerDB2.0.

33 of the *NotOnco* gene fusions were the false positives reported by TopHat-Fusion [120]. They were obtained from two healthy samples (testis and thyroid), with corresponding data published by Illumina within the BodyMap 2.0 project [63]. The other *NotOnco* gene fusions were obtained by applying STAR-Fusion on the Illumina BodyMap 2.0 samples for which information about the originating tissue was provided.

3.3.3 Data-set 2

This set was used to test DEEPrior performances, and it is composed of 2595 fusion pairs and 2623 gene fusions, all belonging to the *Onco* category. This dataset was built starting from Gao et al.[79], who published a fusion call set of more than 25000 gene fusions, obtained by applying three fusion detection tools on the entire TCGA database and appropriately filtering the fusions that are found in healthy samples.

Also, for the samples for which WGS data were available, the presence of gene fusions was validated at the DNA level. The validated gene fusions dataset was kindly provided by the Authors on request. The 1,78% of fusion pairs are in common with the training set.

3.4 Method

The model consists of a CNN followed by a bidirectional LSTM, trained on the entire training set.
Data representation leverages on top of a token embedding learned during the training, where the tokens (i.e., the individual amino acids) are mapped onto a geometric space so that similar tokens are geometrically close.

The model processes sequences between 6 and 4000 amino acids in length. Shorter sequences are not considered as they can hardly be functional, while sequences longer than 4000 amino acids are truncated before being processed by the

model since only the 0,22% of Uniprot sequences are longer than 4000 amino-acids. Furthermore, as the model has been defined, all sequences undergo a padding process.

Different configurations of the number of layers, nodes per layer, and dropout were evaluated to optimize the model, running 10-fold cross-validation for each configuration and repeating each fold ten times to establish the dependence on the initialization. In the end, the optimal model was the following. Embedding layer initialized randomly with size 16; One-dimensional convolution layer with 128 filters with size five kernel and Relu activation function. Max pooling with three window sizes and 0.3 dropouts. Bidirectional LSTM with 32 nodes with tanh activation function and 0.3 dropouts. Final dense layer with a sigmoid activation function.

The number of epochs was set to 100, batch size to 64. In the training phase, I used Keras callback EarlyStopping with patience (number of epochs with no improvement after which training will be stopped) equal to 30 and minimum change in the monitored quantity to qualify as an improvement equal to 0.

The network was implemented in Python 3.7 with Keras library [44] and its architecture is summarized in Figure 3.1.

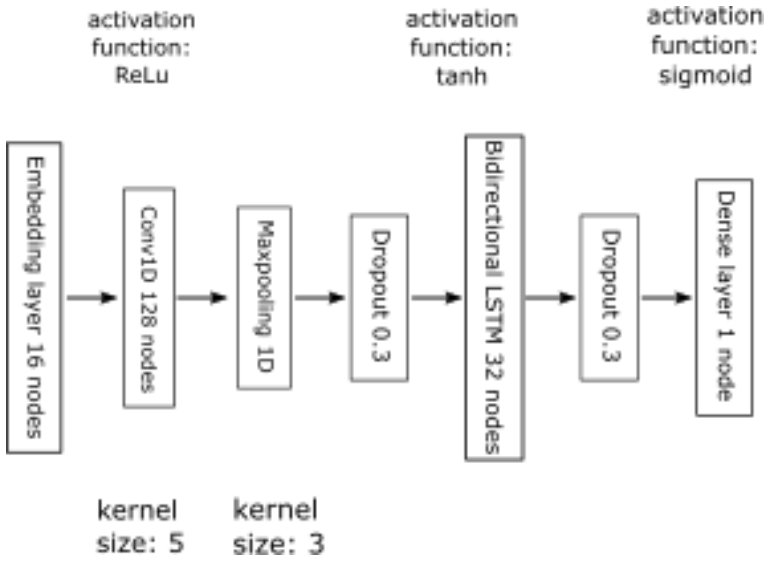


Figure 3.1: Architecture of the deep learning model in DEEPrior.

3.5 The tool

DEEPrior is a user-friendly tool for gene fusions prioritization downstream of gene fusion detection tools. It is implemented in Python 3.7 with minimal additional libraries, and it is available both for CPU and GPU. DEEPrior workflow is summarized in the Figure 3.2.
After executing a fusion detection tool, for each gene fusion, DEEPrior constructs all possible proteins (all coding transcripts of each gene are considered). All resulting amino-acid sequences are then fed into the prediction model, which provides a score for each sequence. The final oncogenic probability value of the gene fusion is obtained as the maximum among these scores.

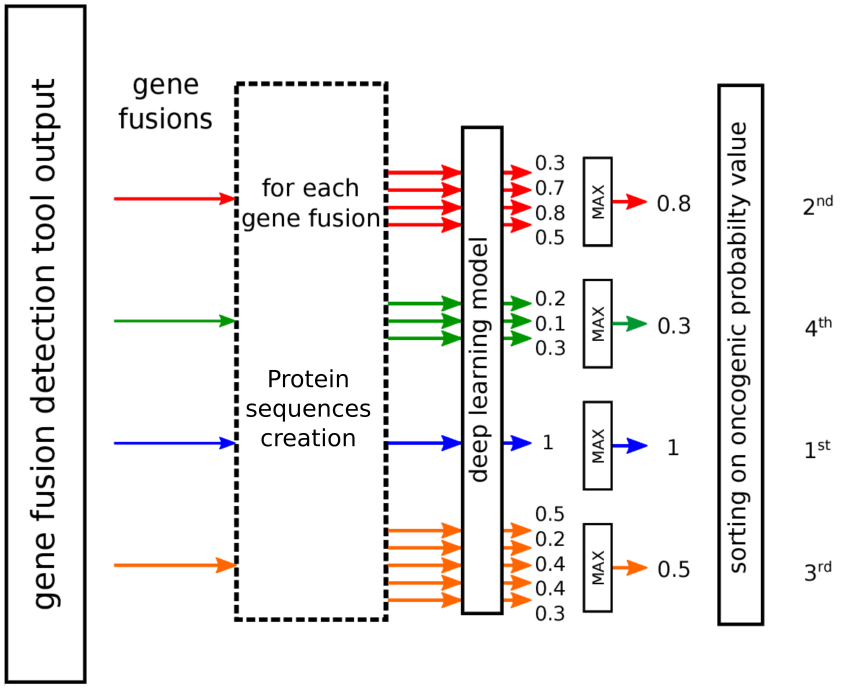


Figure 3.2: **Workflow of DEEPrior tool.** For each gene fusion (see different colors in the figure), DEEPprior generates all possible proteins, considering all transcripts of the fused genes. In the end, the amino-acid sequences are fed into the deep learning model, obtaining a 0-1 value for each protein. The oncogenic probability of each gene fusion is obtained as the maximum of all these values.

The tool supports two different modes: inference and retraining.

3.5.1 Inference mode (Default one)

This mode performs a prioritization of an input set of gene fusions, exploiting a given prediction model.

Input

The input consists of an $N \times 4$ tabular file with rows corresponding to gene fusions, each with four respective attributes: chromosome and coordinate of 5p' end, chromosome and coordinate of 3p' end. Alternatively, the user can provide as input directly the outcome of many gene fusion detection tools. DEEPrior is designed to support as input to the *Inference mode* the most prevalent gene fusion detection tools.

Supported tools are:

- ChimPIPE [183]
- DeFuse [153]
- EricScript [27]
- FusionCatcher [163]
- InFusion [167]
- JAFFA [53]
- SOAPfuse [108]
- STAR-Fusion [85]
- TopHat [120]

Therefore, the user can choose among the most prevalent gene fusion detection tools with no effort. However, I underline that any gene fusion of which the genomic breakpoints are known can be processed, providing a tab-separated file of the breakpoints' genomic coordinates. The first two columns refer to chromosome number and breakpoint coordinate of 5p gene, while the third and fourth columns refer to 3p gene. Coordinates can be entered in genome version grch37 or grch38. An example of the general input file format is reported in Table 3.1.

chr5p	coord5p	chr3p	coord3p
chr7	1000000	chr4	1000000
chr9	2555965	chr6	56444888

Table 3.1: Example of the general input file format, in case the user would like to process gene fusions obtained with a gene fusion detection tool different than the supported ones. The first two columns refer to chromosome number and breakpoint coordinate of 5p gene, while third and fourth columns refer to 3p gene.

Output

A tabular file with N rows corresponding to gene fusions, where for each gene fusion, an oncogenic probability value (a value in [0-1] range) is provided, together with additional information, such as name, description, and ENSEMBL identifier of 5p' and 3p' genes, and the following specific information about the fusion: length of the fused protein, whether the protein is predicted to be truncated, whether 3p or 5p genes are complete, and the corresponding fusion protein sequence. The relevant fields of DEEPrior output can be found in Table 3.2.

Fusion Pair	Onc Prob	5p gene info	3p gene info	Main Protein Length	Trunc Protein	5p gene compl	3p gene compl	Main Protein
TMPRSS2_ERG	0.80	...	...	123	Yes	No	Yes	PYSHEK...
RPS6KB1_SNF8	0.24	...	...	20	Yes	No	No	PGARVRL...
ACACA_STAC2	0.88	...	...	70	Yes	No	No	WGIPLPW...

Table 3.2: Relevant fields in the DEEPrior output file for the Inference mode. Fusion Pair indicates the common names of the genes involved in the fusion; Onc Prob is the oncogenic probability value reported by the tool; Main Protein Length is the length of the fused protein; Trunc Protein reports if the fused protein is truncated (an early stop codon occurs in the protein) or not; 5p gene comp indicates if 5p gene is complete in the fusion (stop codon of the upstream gene is present in the protein); 3p gene compl indicates if 3p gene is complete in the fusion (start codon of the downstream gene is present in the protein); Main Protein is the protein reconstructed by DEEPrior. 5p and 3p gene info fields stand for a list of many other useful information about the genes involved in the fusion.

The complete list of the output information is the following:

- **fusion_pair**: name of the gene fusion with common gene names
- **oncogenic probability value**: oncogenic probability value reported by the tool. It is a number between 0 and 1. Closer is the number to 1, higher is the probability to be oncogenic
- **version**: grch37 or grch38 depending on the genome version parameter defined during the running of DEEPrior. Remember that hg19 is equivalent to grch37 and hg38 is equivalent to grch38
- **chr5p**: chromosome number of 5p gene
- **coord5p**: breakpoint coordinate of 5p gene on chromosome 5p (1-based coordinate system)

- **5p strand**: strand of 5p gene
- **5p common name**: common name of 5p gene
- **5p ensg**: ENSEMBL gene identifier of 5p gene
- **5p gene functionality**: functionality of 5p gene (e.g. proteing coding or not)
- **5p gene description**: additional information about 5p gene provided by ENSEMBL, usually a description of the biological process in which the gene is involved
- **chr3p**: chromosome number of 3p gene
- coord3p: breakpoint coordinate of 3p gene on chromosome 3p (1-based coordinate system)
- **3p strand**: strand of 3p gene
- **3p common name**: common name of 3p gene
- **3p ensg**: ENSEMBL gene identifier of 3p gene
- **3p gene functionality**: functionality of 3p gene (e.g. protein coding or not)
- **3p gene description**: additional information about 3p gene provided by ENSEMBL, usually a description of the biological process in which the gene is involved
- **MainProteinLength**: length of the fused protein
- **TruncatedProtein**: Yes if the fused protein is truncated (an early stop codon occurs in the protein). No otherwise.
- **5p_gene_complete**: Yes if 5p gene is complete in the fusion (stop codon in upstream gene is present in the protein). No otherwise.
- **3p_gene_complete**: Yes if 3p gene is complete in the fusion (start codon in downstream gene is present in the protein). No otherwise.
- **main protein**: the protein with no skipped exons

3.5.2 Retraining mode

In case new validated gene fusions are available (e.g., new cancer or new gene fusion variants), this mode can be optionally employed to generate a custom prediction model easily.

Input

The input consists of an $N \times 5$ tabular file, with rows corresponding to new gene fusions to be added to the prediction model. The required attributes are chromosome and coordinate of both 5p' and 3p' end, and a label of that gene fusion (0 for not oncogenic and 1 for oncogenic).

In this case, the input file is tab-separated and contains validated gene fusions to be included in the retraining of the model for which the label (oncogenic or not oncogenic) is known. The file is similar to the one reported in Table 3.1, and also it contains the *Label* column, which indicates the class to which that gene fusion belongs. 0 means not oncogenic, and one oncogenic. An example of this file is provided in Table 3.3.

chr5p	coord5p	chr3p	coord3p	label
chr7	1000000	chr4	1000000	0
chr9	2555965	chr6	56444888	1

Table 3.3: Example of the input file in the retraining mode, in case the user would like to include in the prediction model new validated gene fusions (e.g. a new cancer or new gene fusion variants) The first two columns refer to chromosome number and breakpoint coordinate of 5p gene, while third and fourth columns refer to 3p gene. *label* column must be 0 if the gene fusion is related to the not oncogenic class, 1 otherwise.

Output

A *.hdf5* file corresponding to the newly generated model. The new model can further be selected as the model in the Inference mode.

3.6 Results

The following refers to the GPU version. However, similar results can be obtained with the CPU version, as they share the same architecture. The experiments were performed on two different data sets, namely *Data set 1* and 2.

To assess the performance of DEEPrior, I first exploited Data set 1, which is completely independent of the training set. It consists of 156 fusions, 122 oncogenic and 34 not oncogenic (see Section 3.3.2 for details). To decide whether a gene fusion is relevant, I set a threshold $thr = 0.8$ on the oncogenic probability value returned by the tool.

By doing so, I obtained that 39.74% of the predictions were selected as relevant. Among them, 9.67% were false positives.

To assess this result's goodness, I run on the same data set two state-of-the-art algorithms (Oncofuse and Pegasus) which provide a score of relevance in the range $[0, 1]$. To be consistent with the test, I set $thr = 0.8$. Oncofuse returned 10.71% of the fusions with 6.67% of false positives. Pegasus returned 8.97% of gene fusions, with 0 false positives.

Besides, I evaluated DEEPrior on Data set 2, consisting of 2623 oncogenic gene fusions from the TCGA validated via WGS (see details in Section 3.3.3). DEEPrior provided 32.48% of the fusions above the threshold, against the 23.55% of Oncofuse and the 15.36% of Pegasus.

3.7 Additional experiments

In order to assess the relevance of my results, I first applied DEEPrior to 6 RNA-seq samples of breast cancer published by Edgren *et al.*[61].

I processed the samples using STAR-fusion [85] and then DEEPrior with $thr = 0.8$. DEEPrior identified 9 gene fusions as highly probable oncogenic. 6 of them were reported in the original study [61] as validated.

Concerning the remaining 3 gene fusions, I remark that no experiment for their validation was provided in the original study.

Besides, I evaluated DEEPrior performance onto 4 RNA-seq samples of prostate cancer studied by Wu *et al.* [216]. In this case, DEEPrior identified TMPRSS2_ERG gene fusion as highly probable oncogenic. This fusion was validated by Wu *et al.* [216] Furthermore, its functional impact on prostate cancer is well known.

3.7.1 Case study

I selected two well-known studies to assess DEEPrior performances: 6 breast cancer samples [61] and 4 prostate cancer samples [216]. The samples are all RNA-seq data and are processed with STAR-fusion and then with DEEPrior. The SRA accession number of each sample and highly probable oncogenic gene fusions identified by DEEPrior ($thr = 0.8$) are reported in Table 3.4. Note that *Unknown* label in the *Validated* column means that the gene fusion was not considered for validation in studies [61] and [216].

tissue	SRA	Gene Fusion	Validated
breast	SRR064286	BCAS4_BCAS3	Yes
breast	SRR064287	BSG_NFIX	Yes
breast	SRR064287	PPP1R12A_SEPTIN10	Yes
breast	SRR064438 SRR064439	ACACA_STAC2	Yes
breast	SRR064438	LAMP1_MCF2L	Yes
breast	SRR064438 SRR064439	PIP4K2B_RAD51C	Unknown
breast	SRR064440 SRR064441	TATDN1_GSDMB	Unknown
breast	SRR064440 SRR064441	CYTH1_EIF3H	Yes
breast	SRR064440	ATAD5_TLK2	Unknown
prostate	SRR496597		
prostate	SRR496595		
prostate	SRR496581 SRR496580	TMPRSS2_ERG	Yes

Table 3.4: Sample tissue type (breast or prostate), sample SRA accession, highly probable oncogenic gene fusion identified by DEEPrior in that sample and validated label. More in detail, I checked if the reported gene fusion has been validated in studies [61] and [216]. *Unknown* label in the *Validated* column means that the gene fusion was not considered for validation in studies [61] and [216].

For breast cancer tissue, 9 gene fusions were identified as highly probable oncogenic, and 6 of them are reported in the original study [61] as validated. I have to remark that, concerning the remaining 3 gene fusions, the validation information was not available in [61]. On the other hand, on prostate cancer samples, DEEPrior identified TMPRSS2_ERG gene fusion as highly probable oncogenic. This fusion was validated by [216] Moreover, its functional impact on prostate cancer is well known.

3.7.2 *NotOnco* dataset

Since in the real world the number of not oncogenic gene fusions is at least one order of magnitude greater than the number of oncogenic gene fusions, I additionally tested the performance of DEEPrior on a set of not oncogenic gene fusions published by Babicenau et al. [19].

I selected a total of 5436 not oncogenic gene fusions. These fusions were not

included in the training set and occurred only once among all samples and all tissues.
DEEPrior identified as not oncogenic the 75,02% of the gene fusions. Almost 80% of these fusions were predicted to be strongly not oncogenic (oncogenic probability value ≤ 0.2). These results suggested that DEEPrior can filter out the most considerable portion of the not oncogenic fusions.

3.8 Conclusions

In this thesis, I suggested that the only amino-acid sequence is enough to predict the oncogenic potential of a protein sequence resulting from a gene fusion. Based on this hypothesis, I proposed a CNN plus LSTM model that takes the amino-acid sequence as input, without any additional information about protein domains. This approach is much more flexible than the available annotation tools, as the algorithm can be easily re-adapted to different cancers or to newly acquired information by simply re-running the automated backpropagation algorithm on a new training set.

Even though the scarcity of the training data intrinsically limits the model, it achieved a good classification accuracy different the test sets and case studies, overcoming the predictions obtained by Oncofuse both in terms of classification accuracy and reliability of the prediction.

Chapter 4

Identifying the oncogenic potential of gene fusions exploiting miRNAs

4.1 Background

Gene fusions are one of the most common somatic mutations and are considered to be responsible for 20% of global human cancer morbidity [160, 68]. A gene fusion is a biological event where two independent genes fuse together to form a hybrid gene. In the most common case, one gene retains the promoter region and the other one provides the end of the hybrid gene. The former is referred to as 5p' gene, while the latter is called 3p' gene. The position where the break occurs is called breakpoint.

The advent of next-generation sequencing (NGS) and the development of fusion detection algorithms [153, 102, 85, 64] led to the discovery of hundreds of novel fusion sequences.

However, not all gene fusions are oncogenic. Indeed, some are genuinely expressed in normal human cells [73] or constitute passenger events [195]. At the same time, other gene fusions are considered to be responsible for a significant percentage of specific kind of tumors [141, 154, 204, 115].

A precise diagnosis of oncogenic gene fusions can inform therapeutics treatments [59, 188] and be used to predict prognosis, patient survival, and treatment response [68]. Additionally, focusing the research on a smaller number of putative oncogenic fusions a diagnosis could take less time; thus, the risks related to misdiagnosis and waiting may be significantly reduced for the patients.

However, discriminating between cancer-driver fusions and non-driver events is not a trivial task.

The first necessary step to solve this problem is performed by the fusion detection tools [153, 102, 85], that identify the candidate gene fusions relying on the

sample's reads, trying to reduce as much as possible the number of false positives (i.e., detected gene fusions that are not found in the sample in later lab validation). Additional studies proposed more sophisticated approaches based on machine learning (ML) techniques applied on top of fusion detection tools' output. Specifically, Oncofuse [189] and Pegasus [3] are noteworthy and use protein domains of the fusion proteins to train the models and predict the oncogenic potential of a fusion. Undoubtedly protein domains are highly informative for the characterization of gene fusions. However, the use of such information as a feature for the ML model requires careful processing from scratch whenever the training database is updated with novel validated fusions.

Recently, previous works explored deep-learning (DL) techniques [145] and presented DEEPPrior [146], a DL model to perform gene fusion prioritization using amino acid sequences of the fusion proteins, based on a Convolutional Neural Network (CNN) and a bidirectional Long Short Term Memory (LSTM) network. Compared to the state-of-the-art tools, this approach is highly effective in accomplishing the classification task with the advantage of avoiding labor-intensive processing of the protein domains.

However, it is known that the oncogenic potential of a molecule depends not only on the sequence itself but also on the effect of post-transcriptional regulatory processes[69].

Transcription Factors (TFs) and micro-RNAs (miRNAs) play a decisive role in the transcriptional and post-transcriptional regulatory processes [152] and can contribute to determining the gene fusion outcome.

To date, most of the available tools exploit transcriptional information and common gene properties to accomplish this task, without considering the post-transcriptional regulators affecting the oncogenic processes.

Here, I present ChimerDriver, a new DL architecture based on a Multi-Layer Perceptron (MLP) which integrates gene-related information with miRNAs and TFs including then in the model transcriptional and post-transcriptional regulative information. Indeed, ChimerDriver exploits the knowledge about TFs and miRNAs targetting each of the genes involved in the fusion to perform gene fusion classification.

ChimerDriver was tested on multiple publicly available datasets and exhibited better classification performance with respect to the state-of-the-art tools. In the end, post-transcriptional regulators confirm the central role in the discovery of oncogenic processes and miRNAs, in particular, are a precious source of information to improve the prediction of the oncogenic gene fusions.

In the following, the proposed method is illustrated alongside with the results in Results section. The discussion and conclusion are reported in Discussion and Conclusions sections, respectively, while a detailed description of model, its architecture and the input datasets is provided into the Methods section.

4.2 Methods

The oncogenic potential of gene fusions was assessed through an MLP to approach this challenging task. The MLP was believed to be a suitable method since a simple yet effective configuration can characterize it. The process of adjusting the hyper-parameters of the MLP allowed for fruitful research for the best model to achieve the highest possible performances.

The chosen dataset used for training and testing the model is broad and extensively described in Dataset subsection.

The features used by the MLP were carefully constructed to obtain the highest possible degree of information concerning the gene fusion samples. Subsequently, the high number of features was reduced through the Random Forest feature selection technique.

4.2.1 Feature selection

All the available features come from multiple sources, and they are related to different characteristics of the gene fusions.

The first five features are obtained from gene fusion structure, and Cancermine [104], a literature-mined database of drivers, oncogenes, and tumor suppressors in cancer. Two features correspond to the retained percentage of 5p' and 3p' genes in the gene fusion, given the breakpoint coordinates. One additional feature controls for the strands of 5p' and 3p' genes, and it is equal to 1 if the two strands are concordant (the two genes transcribe in the same direction), 0 otherwise. The remaining two features correspond to the nature of each gene according to Cancermine [104]: 'Oncogenic', 'Driver', 'Tumor suppressor' or 'Other' when no other option was available.

Other studies [189] have already covered the impact of TF [50] and GOs [219] in the gene fusion classification, which proved to be extremely useful for this purpose. Therefore, TF and GOs have been included in the ChimerDriver model. Specifically, a set of 181 TFs was extracted from the ENCODE database [50] and only those related to the gene in the 5p' position were considered.

Additionally, since each gene can be involved in many different GOs, all of them have been selected. This approach resulted in an extensive amount of GOs to consider, that is, 5125 features.

Besides, my main contribution consists of including miRNAs post-transcriptional regulation in the model. Specifically, all miRNAs predicted to target all 5p' and 3p' genes have been considered. This information was extracted from TargetScan, a popular state-of-the-art database that predicts biological targets of miRNAs by searching for the presence of sites that match the seed region of each miRNA[5], reporting for each miRNA all possible target genes. A set of 333 miRNAs was obtained by investigating the probability of both genes belonging to the gene fusion.

In case of ambiguity, only the highest probability was retained.

The final feature set was considerably lengthy. Thus, I performed feature selection to reduce the 5644 total features to a more reasonable number. The chosen feature selection method was the random forest in which the number of features was lowered according to a threshold. The higher the threshold, the lower the number of retained features. For this study, the already stated threshold was kept in the range 0.0001-0.0005.

4.2.2 Dataset

I retrieved 1765 samples for the training set, 1059 labeled as oncogenic gene fusions and the remaining 706 as not oncogenic. On the other hand, the testing set consisted of 2622 positive samples and 2624 negative samples. A set of 156 gene fusions (122 positives and 34 negative samples) was used in combination with 200 randomly selected samples of the training set as a validation set during the neural network training phase. The processing performed by ChimerDriver to build the features for these samples caused a modest amount of gene fusions to be discarded. It was due to unrealistic values obtained from the calculation of the percentage of the retained gene due to occasional errors in retrieving the correct breakpoint value or strand of a limited amount of genes. However, the majority of the samples adopted in this study are in common with DEEPrior work[146].

Training set

The oncogenic samples of the training set were extracted from COSMIC (Catalog of Somatic Mutations in Cancer). This popular database includes information on gene fusions involved in solid tumors and leukaemias. [70] The chosen 1059 oncogenic gene fusions were already experimentally validated. Moreover, the exact breakpoint positions were provided for each of them. On the other hand, the 706 not oncogenic gene fusions were reported by Babicenau et al. [19] and detected by a gene fusion detection tool in non-neoplastic tissues.

Test set

To build the test set, I used the database provided by Gao et al. [79] which is the result of three fusion detection tools applied on the TCGA database. Among these samples, the authors kindly provided validated gene fusions upon request. These samples were the ones for which WGS data were available. From this collection, I extracted 2622 oncogenic gene fusions for which only the 1,78% of fusion pairs were in common with the training set. Besides, I incorporated a comparable number of negative samples to better attest my tool's performances. The set of 2624 not oncogenic gene fusions was reported by Babicenau et al. [19]. These gene fusions

were found in healthy tissues and stored in a different portion of the database with respect to the one used for the training set.

Model architecture

As previously stated, an MLP was explicitly designed to evaluate the oncogenic potential of the gene fusions. Four layers characterized the final model. For each layer, the number of nodes was respectively: 512, 256, 128, 64.

The activation functions were varied to check which configuration would return the highest performances. Different combinations of the following functions were used: Sigmoid, Tanh and Relu.

Other parameters that have been varied were the learning rate, the dropout, and the number of epochs.

- Learning rate: 0.0001, 0.001, 0.01
- Dropout: 0, 0.1, 0.2, 0.3, 0.4
- Number of epochs: 500 - 1000

The final tool can either take advantage of an early stopping module that stops the training when the accuracy does not improve for 50 consequent epochs or train for a fixed number of epochs.

As already stated, the tool's validation set is a combination of 156 gene fusions and 200 randomly selected samples from to the training set. The goal was to validate the model on a set that included information coming from a different source concerning the training set. In this case, the 200 samples used for validation were not considered in the training phase.

4.3 Results

This section presents an overview of ChimerDriver and of the datasets used in the training, testing, and validation phases. Additionally, the results obtained with ChimerDriver and the comparison with the state-of-the-art tools are discussed. In the end, a case study in which ChimerDriver was applied on a pair of well-known datasets is presented.

4.3.1 Architecture overview and results on the test set

ChimerDriver is based on a feed-forward multi-layer perceptron (MLP). The input feature set combines structural properties of 5p' and 3p' genes with a comprehensive profiling of transcriptional and post-transcriptional regulators (TFs and miRNAs).

In details, I considered as input features the retained percentage of genes, the strand and the relevance in cancer of both 5p' and 3p' genes. Additionally, the oncogenic role of each gene was taken into account. This information was extracted by Cancermine [104], a database which classifies genes as drivers, oncogenes or tumor suppressors. This feature contributes to the assessment of the functional profiling of the gene fusion. TFs and Gene Ontologies (GOs) were also included in the feature set due to their importance in assessing the oncogenic potential of gene fusions[189]. Finally, I added miRNAs to consider post-transcriptional regulation, which influences the translation of gene fusions and therefore their actual oncogenic activity. Specifically, I considered all miRNAs which target 5p' and 3p' genes according to Targetscan database[5].

Due to the high complexity of the model and extensive amount of features, a feature selection process has been performed on the input feature set. The model was trained on 1765 gene fusions, obtained from COSMIC, Catalog of Somatic Mutations in Cancer [70] and from Babicenau et al. work [19]. Given each gene fusion's breakpoint, the aforementioned features are extracted and then fed to the MLP.

According to the cross-validation results, the best network configuration was characterized by four layers with respectively 512, 256, 128, 64 nodes. For each node, the associated activation function was the tanh. The best learning rate was found to be 0.01, while the best dropout value applied to each layer was 0.2. Therefore the model was tested on 5246 gene fusions. 2622 oncogenic gene fusions were retrieved from the work of Gao et al.[79] and the remaining 2624 were gene fusions found in healthy tissues and reported by Babicenau et al. [19]. I ensured that these 2624 gene fusions are entirely independent of involved genes from the training set's ones. The model returned a 0.81 f1-score and 83% precision when tested on this set of gene fusions.

4.3.2 miRNA impact on the classification performance

The miRNA features were extracted from TargetScan [5], a popular database that maps gene-miRNA pairs providing various kinds of information. I mainly focused on the probability that the miRNA would target the specific gene during post-transcriptional regulation. This value was extracted for both 5p' and 3p' genes and it is intended to represent the involvement of the miRNAs in the gene fusion processes. In figure 4.1 I highlight the impact of the miRNA features in the classification by displaying the confusion matrices including and excluding miRNAs from the evaluation. The impact of miRNAs is particularly evident when looking at the number of false-positive gene fusions, which is about three times higher when miRNAs are not taken into consideration. Including miRNAs in the classification task, the AUC value increases from 0.75 to 0.81, and the precision as well from 68% to 83%.

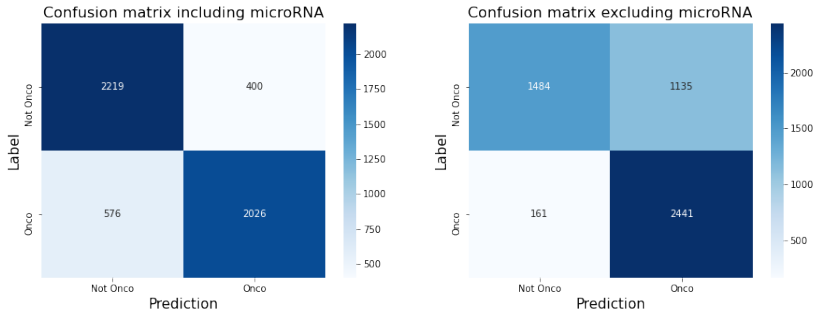


Figure 4.1: Confusion matrices reporting the MLP results including miRNAs (on the left) and excluding miRNA features (on the right).

4.3.3 Comparison with state of the art

ChimerDriver performances were compared to the ones reported by three related works: Oncofuse[189], DEEPrior[146], and Pegasus[3]. To compare the results in the most unbiased way, the experimental conditions of the three tools were reproduced and ChimerDriver was applied.

Oncofuse

To test the robustness of the proposed method, I extrapolated the training set and testing set used by Oncofuse [189]. Those samples were used to train and test my model and then to compare Oncofuse and ChimerDriver performances.

Oncofuse training samples were extracted from TICDB [164], a curated database that contains gene fusions found in tumor samples, and from a collection of fusion genes [74], and read-through transcripts [159] found in normal cells named NORM-RTH. Oncofuse's authors then built the oncogenic testing set by merging oncogenic gene fusions from CHIMERDB [121] and NGS, respectively oncogenic fusions predicted by gene fusion detection tools and fusions discovered and published in NGS studies about cancer [62, 186, 17, 28]. On the other hand, not oncogenic testing samples were taken from Refseq [119] and CGC [192], two databases that report unbroken gene fusions. In particular, the samples that belong to CGC involve unbroken oncogenic genes.

All the previously listed features were processed and gathered, except for the two features related to the retained percentage of genes. These features could not be considered in the evaluation since the provided dataset omitted the breakpoint information.

ChimerDriver model was tailored to this comparison. I obtained 281 input features: the strands and the involvement in oncogenic processes of both 5p' and 3p' genes, 93 TFs, 155 miRNAs, and 30 GOs. The maximum number of epochs was set to 50, and the number of nodes per layer was 256, 128, 64, and 32 (the associated

activation functions were the relu, sigmoid, relu, and sigmoid respectively). The learning rate was fixed to 0.03, while the dropout value applied to each layer was 0.4.

Figure 4.2 shows the comparison of the classification results obtained by ChimerDriver and Oncofuse. Precisely, the green bars correspond to the results reported by Shugay M. et al. [189] for Oncofuse performances. In blue, the results obtained by ChimerDriver are displayed.

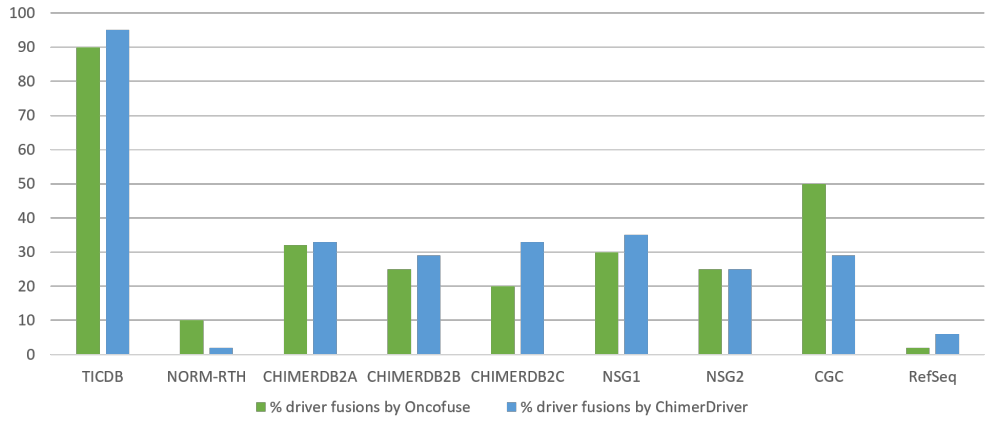


Figure 4.2: The green bars correspond to the results reported by Shugay M. et al. in their paper. In blue the results obtained by ChimerDriver are displayed.

The results of ChimerDriver, when trained and tested on the samples provided by Oncofuse, were able to outperform the ones illustrated in the original paper. Similarly to the research conducted by Shugay et al., the results for each database are displayed separately. The bar diagram shows the percentage of driver gene fusions detected by the model.

Indeed, ChimerDriver results reported in 4.2 show high performances for the training set, outperforming the ones obtained by Oncofuse. 95% of TICDB samples were correctly classified as driver gene fusions by ChimerDriver as opposed to the assumed 90% reported by Oncofuse, furthermore 2% of the NORM-RTH samples were incorrectly classified as driver gene fusions by ChimerDriver as opposed to the assumed 10% reported by Oncofuse.

ChimerDriver successfully outperformed Oncofuse in the oncogenic gene fusion databases used as a test set, namely ChimerDB2A, ChimerDB2B, ChimerDB2C, NGS1 and NGS2. ChimerDriver identified more or a comparable amount of oncogenic gene fusions in each database with respect to Oncofuse, correctly classifying about 1/3 of the samples.

ChimerDriver minimized the number of detected driver fusions of unbroken oncogenic genes, identifying a lower number of driver gene fusions in CGC database, as additional test set.

When tested on the not-oncogenic samples in RefSeq database, Chimerdriver returned a slightly higher number of driver gene fusions.

In general, I may conclude that even without the information on the retained percentage of genes, ChimerDriver outperformed Oncofuse in the great majority of cases.

DEEPrior

DEEPrior is a DL-based classifier which performs gene prioritization using protein sequences obtained from the gene fusion samples. Its architecture is based on a CNN and an LSTM network. It was trained on a dataset extracted from COSMIC [70], and Babicenau et al.'s study [19] and tested on part of the oncogenic gene fusion collection validated by Gao et al. [79]. DEEPrior reconstructs the protein sequences from gene fusion breakpoint information and assigns to each gene fusions an oncogenic score defining its oncogenic probability. Gene fusions are ordered according to the oncogenic score and highly scored fusions are prioritized as drivers. In this sense, DEEPrior main aim consists in providing a reliable classification prediction (oncogenic or not) according to the oncogenic score.

I trained and tested ChimerDriver on DEEPrior training set and test set (*Dataset 2* in DEEPrior paper). As a result, ChimerDriver correctly classified 78% of oncogenic gene fusions from the test set. On the contrary, DEEPrior prioritized as driver the 32.48% of gene fusions found in the test set and among them, 9.67% were false positives. Since DEEPrior aims at classifying only highly probable oncogenic fusions, the percentage of prioritized gene fusions is not directly comparable with the classification performances obtained with ChimerDriver. ChimerDriver provides a classification result for each gene fusion, while DEEPrior classifies a very small percentage of gene fusions in the dataset.

I can conclude that ChimerDriver approach exploits different sources of information (TFs, GOs, miRNAs) while DEEPrior focuses on identifying the oncogenic potential of a gene fusion through its protein sequence without considering the effect of post-transcriptional regulators.

At the same time, ChimerDriver ensures a less computationally intensive approach in the training phase compared to DEEPrior.

Pegasus

To further assess ChimerDriver classification performances, I took into account Pegasus [3], a state-of-the-art tool for gene fusion detection and classification purposes. Pegasus exploits a traditional machine learning model for the prediction of driver gene fusion, namely a gradient tree boosting algorithm.

Also in this case, ChimerDriver was trained and tested on the gene fusion samples used to develop and validate Pegasus.

I observed that the training dataset was strongly unbalanced towards the negative samples, comprising of over 9923 negative samples out of 10162 gene fusions. In order not to penalize the MLP architecture which is particularly sensible to class unbalance, I lowered the number of negative gene pairs to 239, namely the number of positive samples.

ChimerDriver was cross-validated on 10 folds using the aforementioned training samples. It should be noted that, as a result of balancing the classes, the model was given a fairly small number of training examples. The maximum f1-score mean obtained with different values of learning rate and dropout was equal to 0.89 and corresponded to learning rate and dropout, respectively equal to 0.001 and 0.

Pegasus's test set accounted for 78 gene fusions, 39 oncogenic and 39 not oncogenic respectively. According to Pegasus authors, the curated subset of 39 oncogenic gene fusions were almost entirely correctly classified by Pegasus that reported 0.97 of AUC and 0.95 of AUC for the not oncogenic samples.

ChimerDriver correctly classified 27 out of 39 not oncogenic gene fusions enforcing the notion that the model is able to generalize well even with not oncogenic gene fusions. On the other hand, the oncogenic test samples represented a more difficult classification task for ChimerDriver, which detected 17 oncogenic gene fusions. It should be noted that ChimerDriver model was originally trained and tested on a wide variety of gene fusions proving its ability to learn and generalize well when given a fair amount of examples. On the contrary, Pegasus was developed and refined to address gene fusions related to particular tissues like reactive lymph node tissue (used in the training set), glioblastoma multiforme, and anaplastic large cell lymphoma which globally involve a reduced number of samples. In my opinion, the small number of samples in Pegasus training set negatively impacted ChimerDriver performances, hindering the likelihood of reaching the outcome reported by the Pegasus authors.

4.3.4 Case study

Finally, to assess ChimerDriver's performances in a clinical context, I selected two well-known studies: 6 breast cancer samples [61] and 4 prostate cancer samples [216] in which 24 gene fusions are reported to be experimentally validated. The samples are all RNA-seq data. I processed them with STAR-fusion [85] to identify which gene fusions were found in these samples by a standard and accurate fusion detection tool. 21 out of the 24 validated gene fusions were actually detected with STAR-fusion and subsequently processed with ChimerDriver to confirm the ability in correctly detecting oncogenic gene fusions in a real-world case. Figure 4.3 shows the results of this assessment. Specifically, the gene fusions marked in gray were not detected by STAR-fusion hence were not available to ChimerDriver for further processing.

On the 21 samples, ChimerDriver wrongly classified as not oncogenic the three

Validated gene fusions	Detected by FeatureFusion
ANKHD1_PCDH1	Yes
CCDC85C_SETD3	Yes
WDR67_ZNF704	Not detected by Starfusion
CYTH1_EIF3H	Yes
DHX35_ITCH	Yes
BSG_NFIX	Yes
PPP1R12A_SEPT10	Yes
NOTCH1_NUP214	Yes
BCAS4_BCAS3	Yes
ARFGEF2_SULF2	Yes
RPS6KB1_TMEM49	Not detected by Starfusion
TMPRSS2_ERG	Yes

Validated gene fusions	Detected by FeatureFusion
ACACA_STAC2	No
RPS6KB1_SNF8	Yes
VAPB_IKZF3	Yes
ZMYND8_CEP25	Not detected by Starfusion
RAB22A_MYO9B	Yes
SKA2_MYO19	Yes
STARD3_DOK5	Yes
LAMP1_MCF2L	Yes
GLB1_CMTM7	No
CPNE1_PI3	No
TATDN1_GSDMB	Yes
RARA_PKIA	Yes

Figure 4.3: The 24 oncogenic gene fusions validated in prostate and breast tumor samples are reported. STAR-fusion did not detect the three gene fusions marked in gray hence were not available to ChimerDriver for further processing. ChimerDriver correctly classified as oncogenic 18 out of the 21 available gene fusions.

oncogenic gene fusions marked in orange. By inspecting the oncogenic role of 5p' and 3p' genes, but also the retained percentage in the gene fusion, a possible explanation for the wrong classification could be hypothesized. Concerning the ACACA-STAC2 gene fusion, no information on the involvement of any of the two genes was provided to the algorithm. So, although most of the portion of both genes was retained after the gene fusion event, ChimerDriver was probably unsure about their role in oncogenic processes. As for the GLB1-CMTM7 fusion, the algorithm was aware that the latter gene is involved in tumor suppression, on the other hand the retained percentage of CMTM7 was less than 45%. This probably led to the conclusion that there was not enough gene left in the gene fusion to cause issues. Similarly, in the CPNE1-PI3 fusion the percentage of retained genes (respectively 25% and 40%) was probably too low to label the gene fusion as oncogenic even if the genes were associated to the roles oncogenic and driver respectively. Finally, ChimerDriver correctly classified the 18 remaining gene fusions as oncogenic. Hence, ChimerDriver correctly classified 18 out of 21 oncogenic gene fusions, demonstrating that the specifically designed neural network is proficient in learning and generalizing from a consistent number of gene fusion samples. Moreover, the information gathered from the different sources and provided to the tool as features proved to be particularly effective in discerning between oncogenic and not-oncogenic fusions even in a realistic circumnstance.

4.4 Discussion

Identifying oncogenic gene fusions is of crucial importance in cancer detection and prognosis. To date, state-of-the-art tools exploit transcriptional and

GOs information, without considering the post-transcriptional regulators in predicting the oncogenic potential of a gene fusion. Here, I presented ChimerDriver, a novel tool to accomplish the aforementioned task exploiting transcriptional and post-transcriptional regulators. In details, ChimerDriver focuses on miRNAs post-transcriptional effect as a key feature to performe the prediction.

ChimerDriver is based on an ad-hoc designed neural network embedding miRNAs, transcription factors, gene ontologies, and gene-specific information to predict gene fusions' oncogenic potential. The model is stable and exhibits excellent classification performance (f1-score = 0.98).

I tested my classifier against three state-of-the-art tools: Oncofuse, DEEPrior, and Pegasus.

With respect to Oncofuse, I introduced post-transcriprional regulation to perform the classification and, as a result, ChimerDriver outperformed Oncofuse in the great majority of tested cases.

In particular, ChimerDriver performed better than Oncofuse on the test set, correctly classifying as oncogenic about 1/3 of the oncogenic gene fusions. ChimerDriver identified a comparable or higher amount of oncogenic gene fusions outperforming Oncofuse results in each of the positive test cases. ChimerDriver minimized the number of detected driver fusions in 'unbroken oncogenic genes' (negative testing samples) extracted from CGC compared to Oncofuse. This result confirmed the ability of ChimerDriver in generalizing and taking advantage of the given set of features to make a correct prediction.ChimerDriver returned a slightly higher number of oncogenic gene fusions than Oncofuse when tested on RefSeq database of 'unbroken not-oncogenic genes' . I recall that the breakpoint information was not available in Oncofuse datasets. Therefore, to perform an unbiased comparison with Oncofuse, the breakpoint information was neglected by ChimerDriver model. Consequently, the percentage of driver gene fusions detected by ChimerDriver on RefSeq was slightly higher than expected probably because the tool could no profit from the breakpoint information.

ChimerDriver also outperformed DEEPrior in terms of the number of classified gene fusion. In particular, ChimerDriver correctly identified 78% of oncogenic gene fusions in the dataset used to test DEEPrior, which prioritized as oncogenic only 32.48% of the samples. It should be noted that the goals of DEEPrior and ChimerDriver are slightly different. The first performs a prioritization of gene fusions, returning those with an oncogenic probability greater than a threshold (typically 80%). ChimerDriver instead performs an immediate classification of each gene fusion by integrating transcriptional and post-transcriptional features in the assessement. The final outcome of ChimerDriver is remarkable in terms of number of oncogenic samples that were correctly classified while also enlightening because it stresses the extent in which miRNAs are involved in the oncogenic processes of

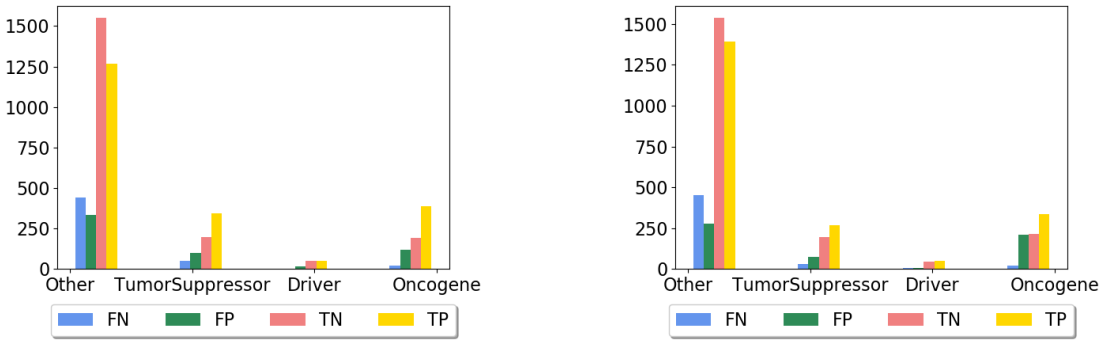
gene fusions.

Moreover, the performances of ChimerDriver were compared to the ones reported by Pegasus authors. According to their research, the latter was able to correctly classify almost all of the test samples. After training and testing ChimerDriver on the gene fusions provided by the authors, it was observed that the number of detected oncogenic samples was lower than the results reported by Pegasus. As already stated, the number of training samples was lowered in order to balance the oncogenic and not oncogenic classes. However, the limited number of samples processed by ChimerDriver in the training phase has probably inhibited the neural network from learning efficiently. In addition, Pegasus's authors specify that the negative validation samples included at least one oncogene or tumor suppressor domain. I remind that, in order to make a prediction, ChimerDriver relies also the role of each gene in oncogenic processes (e.g. driver, oncogene or tumor suppressor), making the classification task particularly arduous to tackle. Nevertheless, ChimerDriver correctly classified a most of the not oncogenic gene fusions enforcing the notion that the model is able to generalize well in this situation.

In this work, I focused on the integration of information coming from different databases to improve the current state-of-the-art research on classifying oncogenic gene fusions. Additionally, a neural network was specifically designed for this task. However, the main contribution of the present work is the introduction of miRNAs in the classification model. In fact, despite miRNAs role in determining the oncogenic potential of gene fusions has been demonstrated, they had never been considerated in such a task. In the present work, I showed that they could significantly improve the model performance. In particular, they reduced by two-thirds the number of false positives and improved the AUC and the precision of the model. I can conclude that miRNAs, being involved in the regulation of gene fusion-related protein, are a promising indicator of the oncogenic potential of gene fusions.

The main limitation of the proposed method is that some gene fusions are misclassified. To better investigate ChimerDriver classification with respect to the Cancermine [104] role, I reported in Figure 4.4 the distribution of the Cancermine roles (e.g. tumor suppressor, driver, oncogene, other) for 5p' gene (Figure 4.4a) and 3p' gene (Figure 4.4b). In addition, test set samples are divided in each role according to the classification results (false positives (FP), false negatives (FN), true positives (TP) and true negatives (TN)). TP samples are characterized both for 5p' ad 3p' genes by a prevalence of suppressors and oncogenes. On the contrary, TN mostly refer to the 'other' CancerMine role. As a consequence, FP samples could consist of oncogenes (in particular for 3p' gene) and FN samples are hardly ever related to tumor suppressors, drivers, or oncogenes. In this sense, FP and FN samples reflects ChimerDriver behaviour on TP and TN respectively. In a clinical context, FN misclassified samples are unlikely to be tested for in lab validation, since

most of them involve genes with not a specific oncogene/tumor suppressor role. FP samples instead would have been considered for an experimental validation, that in the the end would exclude them from oncogenic fusions. However, laboratories would still benefit from a selection of putative oncogenic gene fusions.



(a) (b)

Figure 4.4: Here I report the distribution of the false positives (FP), false negatives (FN), true positives (TP) and true negatives (TN) regarding Cancermine information for both 5p' and 3p' genes(respectively Figure 4.4a) and 4.4b)). Noticeably, FPs are never tumor suppressors, drivers or oncogenes.

4.5 Conclusions

Gene fusions are a common mutation that is nowadays known to be responsible for about 1/5 of human cancers. It is of the uttermost importance to correctly identify gene fusions to improve cancer detection and prognosis. Considering that the state-of-the-art tools exploit transcriptional and gene information neglecting post-transcriptional regulations, I combined this knowledge and established the value of miRNAs in achieving superior classification performances.

To conclude, I presented ChimerDriver, a novel and stable DL architecture based on a Multi-Layer Perceptron (MLP) which, for the first time, combines gene-level features with TFs and miRNAs targetting the gene fusion to perform its classification and prioritization.

ChimerDriver was trained and tested on a consistent number of gene fusions. The final results highlight the impact of miRNAs in the evaluation of the oncogenic potential of gene fusions. I can infer that the inclusion of miRNAs represents a valuable advantage in the identification of oncogenic gene fusions.

ChimerDriver can become a valuable tool for research laboratories to predict the oncogenic potential of gene fusions. Indeed, the expensive validations could be targeted cost-effectively with this easy-to-use tool; additionally, it may speed

up identifying novel and potentially oncogenic gene fusions, allowing for better diagnosis, classification, and treatment of cancer patients.

Part II

Multi-Omics

Chapter 5

Automated Prediction of Connectivity between Mouse Brain Regions

5.1 Methodological contribution

This chapter focuses on the integrative analysis of mouse brains' gene expression and connectivity data exploiting machine learning techniques. Many studies have been performed considering the functional connection of two or more brain areas (which areas are activated in response to a specific stimulus) while analyzing physical connections (i.e., axon bundles) starting from gene expression data is still an open problem.

In this context, the main methodological contribution consisted of identifying a strategy to integrate axon connectivity data with the point values of gene expression within the mouse brain. Therefore, the developed method identifies if two brain regions are connected by axons analyzing the gene expression data alone.

From a computational point of view, this section's main challenges consist of mapping genomic information in the three-dimensional space together with spatial images representing the intensity of a viral tracer to infer which axons connect specific regions of the brain to the central nervous system. As discussed in the previous chapters regarding the genomic sequence, also, in this case, the direct connection between gene expression and the physical connectivity of two or more regions in the brain is not known. Indeed, this connection still needs to be further investigated.

In this work, the neuronal connection data (obtained by viral tracers) of mouse

brains were processed to identify brain regions physically connected and then evaluated with these areas' gene expression data. A multi-layer perceptron was applied to perform the classification task between connected and unconnected regions providing gene expression data as input. Furthermore, a second model was created to infer the degree of connection between distinct brain regions. The implemented models successfully executed the binary classification task (connected regions against unconnected regions) and distinguished the intensity of the connection in low, medium, and high.

5.2 Introduction

The brain is a complex organ comprised of more than 100 billion neurons grouped into many functional regions that communicate through electrochemical signals.
When referring to the brain, physical connectivity refers to the pattern of anatomical links constituted by the neurons' axons and connected to the dendrites of postsynaptic neurons [114]. The physical connections that link numerous groups of neurons constitute a network that, on a larger scale, constitutes the so-called anatomical brain connectivity.

It is shown in the literature that functional properties of neurons and neuronal systems depend on neural connectivity patterns [23, 25]. This idea has long attracted the attention of neuro-anatomists, who dedicated their studies to the new field of science dealing with the assembly, mapping, and analysis of the connectome [193].

The anatomical connectivity in the brain is constituted by fibers that propagate from the neuronal bodies. These, in turn, contain the nucleus and all the nuclear components that contribute to cellular differentiation and morphogenesis. Accordingly, the main factors influencing connectivity patterns have to be searched at the cellular scale, meaning that cellular activity influences physical brain connectivity patterns at the anatomical level. Hence, the analysis of the cellular activity in the form of neuronal gene expression profiles may represent an effective way of understanding the physical connectome more in depth [116, 190].

Gene expression is how information from a gene is used to synthesize a functional gene product such as a protein. Gene regulation gives the cell control over structure and function and is the basis for cellular differentiation, morphogenesis, and adaptability of any organism.
Gene expression profiling is the measurement of the activity (i.e., expression)

of genes. Sequence-based techniques such as RNA-Seq provide information on the sequences of genes, from which their expression level can be derived. Nonetheless, they extract information through a disruptive process of the tissue under investigation, providing gene expression levels averaged over the whole cellular population without any spatial information. On the other hand, Single-cell RNA-seq (scRNA-seq) [135], relying on the separation of single cells from the tissue by enzymatic or mechanical dissociation, provide cell-specific information but even in this case with lack of information on spatial location and the micro-environment [117].
Instead, using *in-situ* techniques, it is possible to detect the spatial distribution of gene expression levels in the tissue. Fluorescence *in situ* hybridization (FISH or ISH) uses RNA or DNA complementary hybridization probes labeled to fluorescent molecules. Once the probes have hybridized the fixed tissue target, the transcript can be localized and quantified through fluorescence microscopy images. Thanks to this process, FISH allows us to maintain both spatial and morphological information. On top of that, it generally generates better-quality images than other *in situ* techniques [135], which makes it the ideal source of information for connectome studies.

Due to the crucial role of anatomical brain connectivity, scientists generated and made available some brain atlases, modeling the axonal connections between different brain regions [124, 34]. Upon these connectivity models, the scientific community conducted several studies to detect the existence of anatomical neural connections or spatial correlations between the brain tissue's intrinsic properties. Studies on the mouse brain based on visualization and clustering showed that gene expression and connectivity information have significant spatial auto-correlation levels, which needs to be accounted for through integrative analysis [67]. Based on these studies' results, brain regions with similar expression profiles tend to have similar connectivity profiles. Likewise, brain regions that are anatomically connected have gene expression patterns that are remarkably similar [67]. Some studies have also identified genes responsible for the relationship between cellular activity and connectivity, as they are directly involved in neuronal development and axon guidance [72]. With more in-depth investigations of the specific relationship between gene expression and connectivity in the mouse brain, gene expression is predictive of the connectivity pattern when the connectivity signals are in a discrete form. Also, most of the predictive power resides in the expression data from a relatively small number of genes, suggesting that very few genes are responsible for generating brain connectivity in each specific brain structure [107].

All these findings stem from the analysis of data from the mouse brain. Nonetheless, many mouse brain genes find a direct correspondence in the human brain, and regionally enriched genes were demonstrated to be conserved when shifting from one species to another [194]. Based on this evidence, the combination of human

and mouse single-cell transcriptomic profiles, through the application of feature selection and linear modeling, was used to provide better insights into human brain connectivity. The combined data were then used to demonstrate that gene expression is a better indicator of cellular localization than the location of cell nuclei, especially for cells with large and irregularly shaped cell bodies such as the neurons [110].

Upon these considerations, gene expression data can be effectively used to predict brain connectivity automatically. While most of the works in the literature focus on either analyzing the most relevant genetic signatures of neuronal connectivity [77, 184], or on investigating the direct relationships between gene expression and brain functionality [178, 13, 78], lesser attention is devoted to predicting anatomical connectivity at a cellular resolution, directly using the transcriptomic profile as the input baseline. The most representative works in this regard use model-based techniques (e.g., sparse linear models [107, 66]), obtaining a good prediction accuracy level (between 80% and 90%) at the well-known cost of difficult parametrization and non-obvious selection of the variables.

In this thesis, I push forward the path of predicting the degree of anatomical connection of brain areas by performing integrative analysis of gene expression profiles and connectivity data. To do so, I interpret the prediction as a classification problem, where the input feature vectors describing gene expression profiles of brain region pairs are automatically grouped into multiple classes based on their level of physical connectivity. To solve this classification problem, I exploit neural networks, which, compared to traditional model-based techniques, have the advantage of being non-parametric and do not require a priori definition of the mathematical relationships among variables. As such, our tool is developed on a case-study of mouse brain data, but it can be ideally applied to any other application of interest.

The developed method implements all the stages essential to solving the connectivity classification problem in a fully automated way, including data collection, storage, pre-processing, and the in-depth analysis of the prediction outcome, aiming at the investigation of anatomical connections between the brain macro-regions.

Based on the nature and complexity of the analysis to perform, I chose to implement a Multilayer Perceptron (MLP), a class of feed-forward artificial neural networks that is often used both for data classification and regression [109].
This network is fed with feature vectors representing the gene expression profiles of two different brain regions. Each feature vector element corresponds to a gene, specifically to that gene's expression level in a low-dimensional spot (i.e., a voxel) of a region. The available spatial connectivity data are aggregated to obtain classification labels for the feature vectors, obtaining a unique value representing the

connection between two regions.

Then, I investigated the network's outcome on two different tasks: a multi-class classification task (with three classes corresponding to unconnected, weakly connected, and strongly connected areas, respectively) and a more straightforward binary classification task (unconnected and connected). The analysis was performed on an extensive dataset from the cortex and the cerebellum (58 regions in total). These specific regions were selected because the corresponding annotated datasets ensure a comprehensive representation of connectivity degrees.

5.3 Overview on the proposed method

The proposed methodology consists of an MLP classifier, where the input is a vector (so-called *Source-Target* vector), representing the gene expression levels of two regions of the mouse brain, respectively called *Source* and *Target*. The classifier's output is a unique categorical label, representing the overall connectivity degree of all the voxels corresponding to the input Source-Target pair.

To generate the *Source-Target* vectors and corresponding connectivity labels, in this study, I used gene expression and connectivity values available from the *Allen Mouse Brain Atlas (AMBA)* [133] and the *Mouse Brain Connectivity Atlas (MBCA)* [124] resources, combined with the connections' intensities reported by the *Brain Architecture Management System (BAMS)* database [34].

An extensive dataset was generated, choosing as representative brain areas the cortex and the cerebellum. These areas include significant and independent functional regions. Hence, from the analysis of such areas, I expect to find i) dense connectivity between internal sub-regions of the same area and ii) low connectivity degree between the two areas as a whole. The overall procedure is represented in Figure 5.1. I considered 58 different regions (8 from the cerebellum and 50 from the cortex) and randomly selected 21 voxels for each possible Source-Target combination. By doing so, I obtained a total number of 54,495 Source-Target vectors (M value in Figure 5.1), with corresponding connectivity labels. Each of two parts of a Source-Target vector is the expression level of a set of genes within a particular voxel of the mouse brain, where the first and second half of the vector contain values belonging to the Source or the Target regions, respectively. More details about the databases, as well as on the specific methodology applied to generate Source-Target vectors and labels, will be provided in Sections 5.4 and 5.5, respectively.

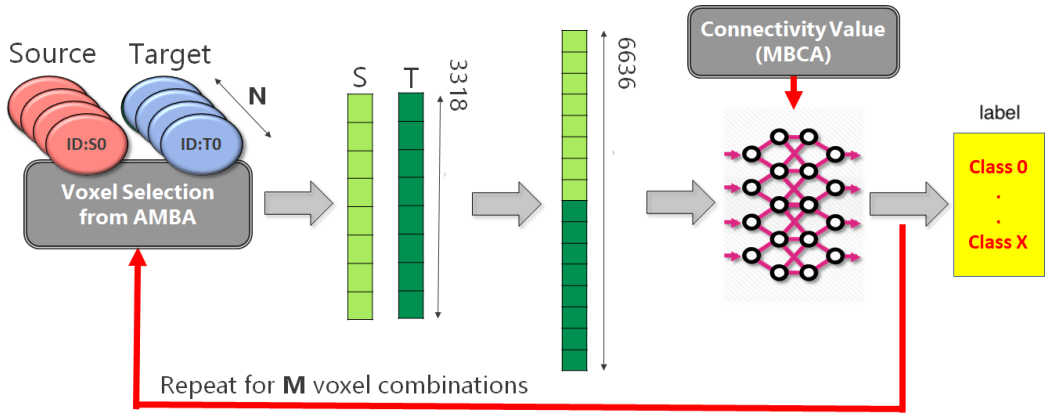


Figure 5.1: Scheme of the analysis pipeline. For each Source-Target pair (N, in total), I randomly select M voxel combinations. Per each combination, I generate two 3,318 gene expression vectors (Source and Target, respectively) with information taken from AMBA. The concatenation of the two vectors represents the Source-Target vector given as input to our MLP model. A categorical label describing the Source-Target connection degree is obtained by setting empirical thresholds on the connectivity values provided by MBCA.

5.4 Material

In the following, I describe in more detail the datasets from which the Source-Target vectors given as input to the MLP model and corresponding connectivity labels shown in Figure 5.1 were obtained.

5.4.1 Allen Mouse Brain Atlas

The *Allen Mouse Brain Atlas* (AMBA) represents an integration between transcriptomic and neuroanatomic mouse brain data. It is a complete high-resolution atlas of gene expression throughout the adult mouse brain composed of different sections and tools that enable easy data navigation and analysis. Gene expression patterns are available as images obtained by *in-situ* hybridization (ISH) technique [200] applied on full brains of 56-day old C57BL/6J male mice. For each gene, expression levels are provided as grid data, in the form of a 3D matrix representing the mouse brain's three-dimensional structure. Each element of the matrix is a voxel at 200 μm resolution, storing a gene expression level. In the study, this information is used as input to the classification model.

The *Allen Mouse Brain Connectivity Atlas* (MBCA) consists of connectivity values in the form of axonal projections labeled by rAAV, a viral tracer injected in a specific site and then detected through two-photon tomography. When the viral tracer is injected in a brain region, referred to as *Source region*, it produces axonal

projections in several *Target regions* (see Figure 5.2 for a schematic representation). These projection data are provided for more than 200 mouse brain regions in the coronal section. In the Allen MBCA database, more than one injection site can be found for a single brain region. Section 5.5.1 will describe how multiple injection sites were handled in the proposed methodology.

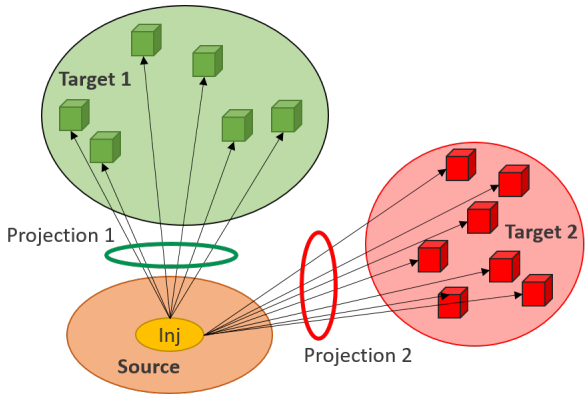


Figure 5.2: Schematic diagram of Source-Targets projections. A Source is a brain region where the viral tracer is injected (*inj*, in the figure). As a result of the injection, multiple axonal projections are produced in so-called Target regions.

Like gene expression, connectivity information is available for each injection site in the form of grid data. Each element of the corresponding 3D matrix is a voxel (in this case, provided at 100 μm resolution) whose value represents the connectivity degree in that specific 3D position. In our study, the connectivity data is used to obtain a classification label for each couple of gene expression profiles from two different brain regions.

Corresponding gene expression and connectivity data (respectively from AMBA and MBCA) of each brain region can be coupled at different spatial resolutions, using structural annotation files. The cerebral regions, grouped into hierarchical layers, consist of several voxels of gene expression and connectivity values, both referred to as a reference space created by the Allen team for mouse brain modeling.

5.4.2 BAMS

As an additional source of information for our study, I used neural circuitry data provided by the *Brain Architecture Management System* database (*BAMS*) to select the most significant brain areas for the investigation [34].
This database contains about 45,000 connection reports between different gray matter regions of the rat, in the form of interactive matrices showing each brain region

pair's strength of the connection.
Several studies demonstrate that mouse and rat brains share the same anatomical features, only at a different scale [67, 166]. Hence, the connection reports can select the most promising brain areas even in the mouse.

5.5 Method

Besides the MLP prediction model, I implemented a complete automated pipeline to handle dataset collection and the organization and processing of the gene expression and connectivity data into Source-Target vectors with corresponding class labels, to be given as input to the MLP. The main steps of this pipeline, implemented exploiting the Knime framework [31] and the SeqAn library [177], are the following:

1. download of grid data from the available data sources;
2. processing of the raw grid data to integrate the gene expression and the connectivity information;
3. generation of a full and coherent dataset of Source-Target gene expression vectors and corresponding connectivity labels, ready to be cropped into training, validation, and test sets for the MLP.

5.5.1 Download of Grid Data

The Allen Brain Atlas provides grid-data at different resolutions, consisting of 3D summaries of both the gene expression and connectivity data, re-sampled to a Common Coordinate Space of the 3D reference brain model [7]. The database provides a structural grid data annotation system at each resolution scale to enable spatially coherent processing of these two sets of data. This annotation allows linking mouse brain voxels to anatomical structures in the Common Coordinate Space.

Grid data is downloadable through an API service by queries. The queries were implemented through a web application (the RMA BUILDER) that is freely accessible on the Allen Brain Atlas's API section.

Gene expression

As mentioned in Section 5.4, the Allen Institute Mouse Gene expression data consist of whole-brain *in-situ* hybridization data obtained from brains of 56-day old C57BL/6J male mice [7]. The grid-data of the detected expression levels are provided for coronal and sagittal sections. Even though the sagittal section counts

more than 20,000 genes, connectivity data are available only for the coronal section. Thus, in this study, I focused on the 3,318 gene expression grid-data corresponding to this specific section.

The main phases of the elaboration of gene expression data are represented in Figure 5.4. Each gene's expression profile throughout the mouse's brain is associated with a SectionDataSet, a specific data object of the Allen Brain Atlas framework where all the experiment's information is stored. I first build a query to retrieve the SectionDataSet unique identifiers (IDs) for the gene expression experiments in the form of an XML document. Then, to retrieve the corresponding gene-expression grid-data, I build a query with the RMA BUILDER and obtain in return an *energy.raw* file for each of the 3,318 gene expression experiments. This file contains a vector of 159,326 elements corresponding to the 3D voxels of the mouse brain model (67×41×58 voxels at 200 μm resolution) that can be reconstructed leveraging the reference information provided by the database, as shown in Figure 5.3 [4].

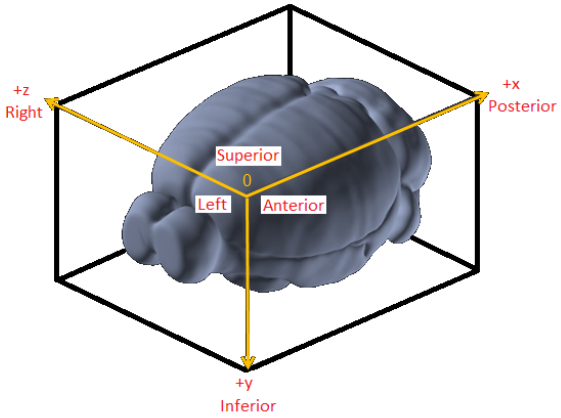
At the end of the download procedure from the Allen Brain website, [7], I obtain a 3,318×159,326 matrix of gene expression levels, with rows corresponding to genes and columns to 3D voxels. This matrix is stored into a single .csv file, as represented in Figure 5.4.

Allen Connectivity

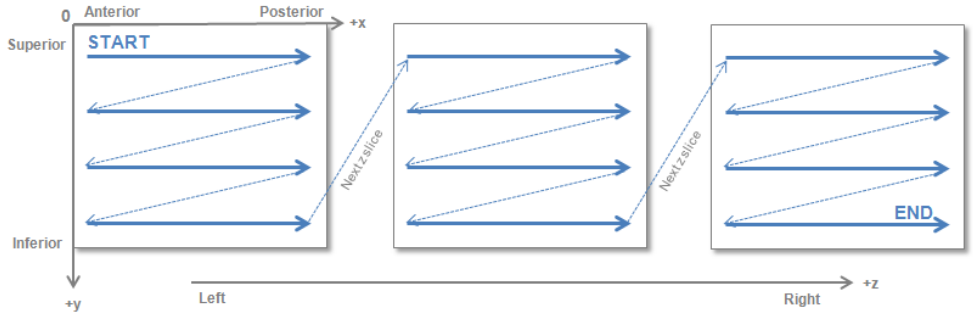
As outlined in Section 5.4, The Mouse Brain Connectivity Atlas provided connectivity information in the form of axonal projections labeled by rAAV viral tracer and detected through two-photon tomography for more than 200 mouse brain regions in the coronal section. Injection sites refer to the spots where the viral tracer is injected. The region where a specific injection site is placed and the region where the injection produced axonal projections are referred to as *Source* and *Target* regions, respectively. Injections involving a single region are called primary. Nonetheless, because of the small size of the mouse brain, a single injection can involve more than one region. These are called secondary injections.

In this work, we focused only on the primary injection sites and considered connectivity data at 100 μm resolution, which is the closest to the 200 μm gene expression resolution among all the available ones (10, 25, 50, 100 μm, respectively).

The main phases of the elaboration of connectivity data are represented in Figure 5.5. Again, each primary injection site corresponds to a SectionDataSet. Hence, I designed a query to retrieve the SectionDataSet IDs of injection experiments through the API service, which returns an XML document with 2333 primary



(a) 3D Volume



(b) Packing criteria of the volumetric data into a 1-dimensional array.

Figure 5.3: The common reference space is in PIR orientation where x axis = Anterior-to-Posterior, y axis = Superior-to-Inferior and z axis = Left-to-Right

injection IDS. Such IDs are exploited to build a query with the RMA BUILDER and retrieve the connectivity grid data in return.

By doing so, I obtain 2,333.Nrrd files, each representing the axonal projections of a specific primary injection site. This approach provides a correspondence between the 2,333 primary injections and their corresponding target regions.

For connectivity data, the 3-D volumetric grid-level information at 100 μm are provided in the form of a $13 \times 80 \times 114$ numerical array, as represented in Figure 5.3.a.

Maintaining the spatial reference provided by the Allen Brain Atlas, each 3D matrix was unpacked into a vector of 1,203,840 elements. This way, I obtained 2,333 vectors in total that were stored into a single *.csv* file along with the source region indication (see the last phase of Figure 5.5).

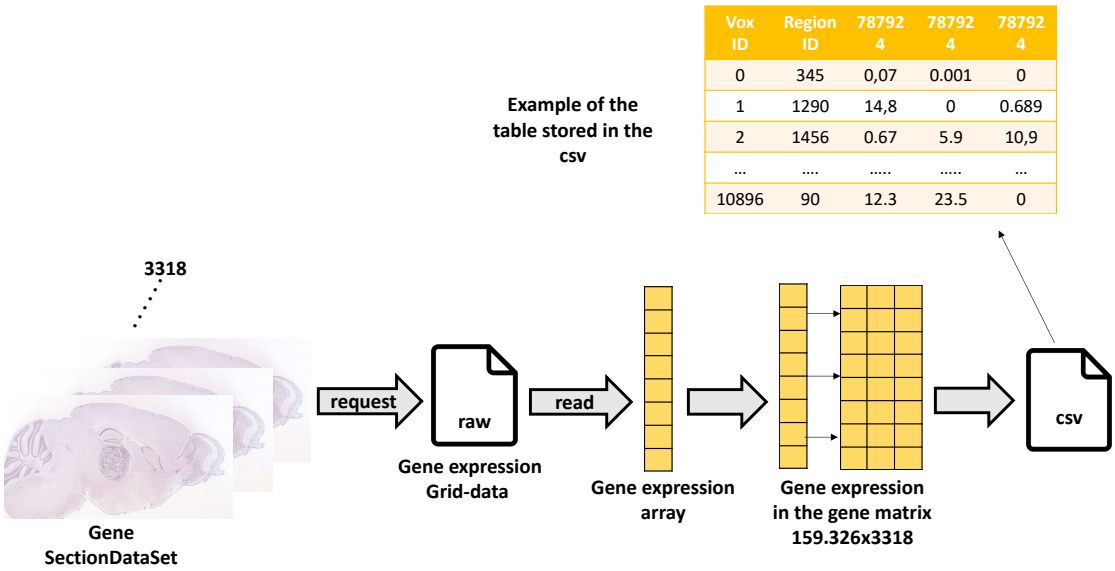


Figure 5.4: Elaboration of gene expression data: main phases. (1) Retrieve a SectionDataset for each of the 3318 genes; (2) download grid expression data in the form of an energy.raw file; (3) reconstruct a 3,318×159,326 matrix of gene expression levels, with rows corresponding to genes and columns to 3D voxels; (4) store data into a .csv file.

Structural annotation file

An annotation volume is a 3D raster image that partitions the reference space into structures, whose number of voxels depends on the size of the structure and the model's specific resolution. Each voxel is assigned to a specific brain structure employing a region ID [7].

Brain structures in the Allen reference spaces are arranged in trees, with leaf nodes representing very fine anatomical partitioning and nodes closer to the root corresponding to gross partitioning. The annotation file reports region IDs together with the details of the finest anatomical partitioning.
Hence, gene expression and connectivity data can be mapped to several common reference spaces. To link each data voxel to the corresponding membership brain region, the Allen Brain Atlas provides a structural annotation file at different resolutions, where the $i-th$ annotation element allows to map the $i-th$ voxel in the data array to its brain structure.

Like the gene expression data, the annotation is provided at 200 μm resolution,

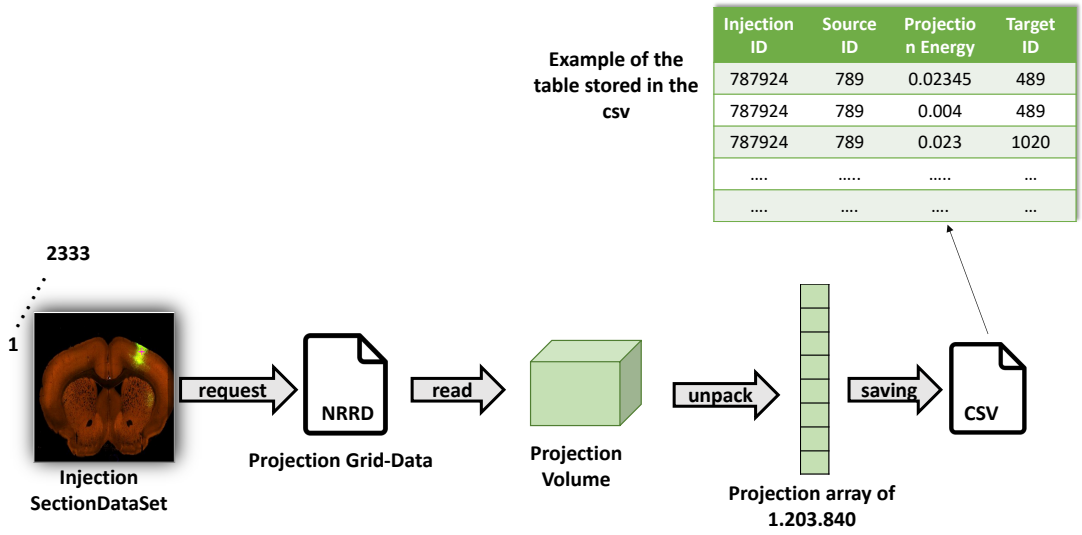


Figure 5.5: Elaboration of connectivity data: main phases. (1) Retrieve a SectionDataset for each of the 2333 primary injections; (2) projection grid-data in the form of an .Nrrd file; (3) reconstruct a projection volume, unpacked into a vector of 1,203,840 elements; (4) store data into a .csv file.

in the form of a vector of 159,326 elements.
Likewise, the connectivity annotation (CA) is provided at 100 μm resolution, reshaped in the form of a vector of 1,203,840 elements. Different from the other data, the primary injection structure annotation is not provided at the finest annotation level. I implemented a procedure to trace both the annotations back to the same resolution level in the annotation tree. To do so, I exploited a list of dictionaries provided by the Allen Brain Atlas, documenting brain structures and their hierarchical relationships in the form of a structure graph.

Brain Architecture Management System (BAMS)

As mentioned in Section 5.4, additional neural circuitry data collected from BAMS were used as a reference to decide which brain structures are most significant for our analysis. To date, BAMS includes about 45,000 connection reports between different gray matter regions, leveraging information on connections that were demonstrated by previous studies.

The reports can be freely downloaded from the site of BAMS [208] in the form of an interactive matrix (see Figure 5.6), where each element (i, j) defines the existence and the intensity (encoded by a value in a $< 1-9 >$ range) of the connection between two specific brain regions i and j, identified by the same universal acronyms used by the Allen Brain Atlas. Unknown connections are assigned a 0 value.

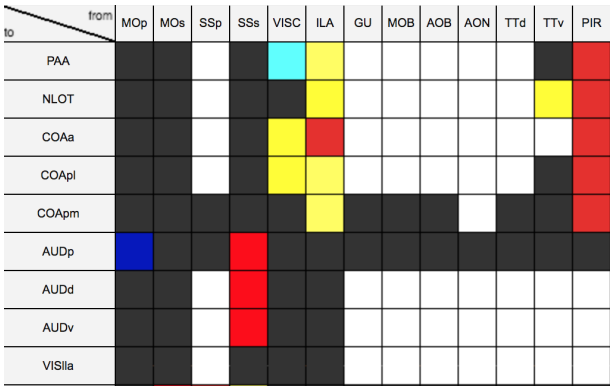


Figure 5.6: Interactive matrix from BAMS. Each element of the matrix represents the connection between two regions, reported in rows and columns. Different colours encode different connection intensities, with white corresponding to unknown connections.

5.5.2 Generation Source-Target vectors and corresponding connectivity labels

The last step of the dataset generation consists of assigning a unique connectivity label to the Source-Target gene expression vectors.

All the connectivity values reported for a specific injection ID (i.e., experiment) and a specific Source-Target combination are first aggregated based on their median value. This solution is preferred to others (e.g., mean value) because the median value is inherently robust to the presence of outliers and noise. Nonetheless, as like this technology, a specific region may be a site of injection of multiple experiments. Hence, for each experiment, the connectivity of the axonal projections produced in the corresponding target regions will be stored in a specific SectionDataSet. Then, if a specific source has targeted the same region in different experiments, that specific combination of Source-Target regions will correspond to more than one median value. To tackle this issue, I implemented the second level of aggregation and obtained the final connectivity value as the maximum of all the multiple median values. This choice stems from the empirical observation that the connectivity network detected in each experiment (and hence, the corresponding connectivity value) is highly dependent on the specific position of the injection. Hence, using the maximum as the most representative value has a two-fold advantage: i) it filters out small connectivity values possibly due to peripheral injection sites, and ii) allows to select the experiments with the best spatial conditions as the most representative

of a specific source-target combination.

Based on the empirical connectivity thresholds defined in Section 2, this connectivity value is transformed into a categorical label representing the strength of the connection: either (0,1,2) for multi-class classification, or (0,1) for binary classification.

To allow further processing and easy access of the data, in this solution, the full and coherent dataset of Source-Target gene expression vectors and the corresponding connectivity labels were stored into four tables of an SQLite database shown in Figure 5.7:

1. Table *voxID2Annotation* carries the spatial information and contains the voxel ID and corresponding brain structure annotation.

2. Table *voxID2GenExpr* was obtained by filtering out the voxels with gene expression level value equal to 0. It is made of columns reporting gene expression value, voxel ID, and gene ID, respectively.

3. Table *injection2regionID* was obtained by grouping all the voxels by Source and injection ID. Hence, it reports the Source region ID for each injection.

4. Table *injection2target* was obtained by grouping each voxels' connection values by the Target ID. More specifically, voxels belonging to the same Target region were aggregated by the median of the values associated with each of these voxels. The final table is then composed of three columns: injection ID, the median of the values obtained for a specific Target ID, and its annotation ID, respectively.

This database solution allows the quick generation of custom datasets to be given as input to prediction models, avoiding re-processing the raw-data.

A custom dataset leveraging such a database can be built as follows. First, N Source-Target regions are selected, based on the specific analysis to perform. Gene expression and connectivity data of the selected pairs undergo the following pipeline, as represented in Figure 5.1:

1. for each Source-Target pair, M voxels belonging to the source region and M voxels to the target regions are selected on the expression gene annotation.

2. for each selected voxel, a vector composed of 3,318 elements is generated, where each element corresponds to the expression level of a specific gene. Hence, M vectors representing the gene expression profile of the Source and M vectors representing the gene expression profile of the Target are obtained.

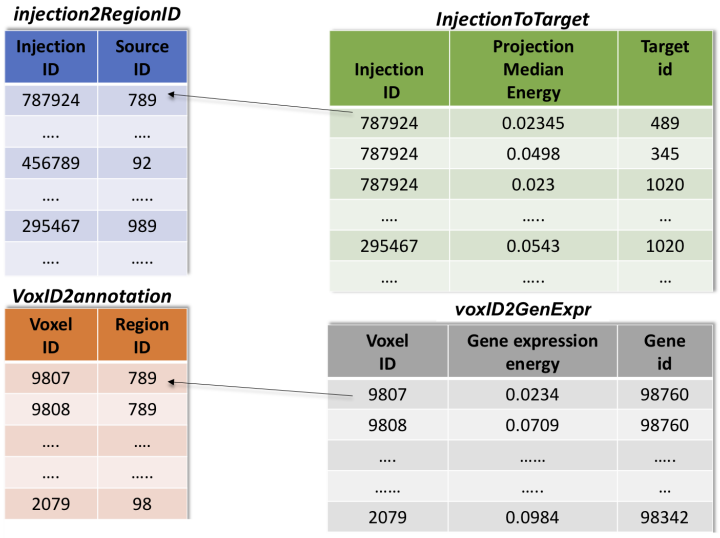


Figure 5.7: SQLite database tables generated to store all the gene expression and connectivity data.

3. A dataset is created by selecting P combinations among all possible Source-Target voxel combinations. More specifically, the gene expression vector corresponding to the Source voxel is concatenated with the gene expression vector corresponding to the Target voxel. Hence, the obtained dataset will be made of P vectors.

4. In the end, a unique categorical label representing the Source-Target connectivity is assigned to each combination.

These steps are repeated for all N number of Source-Target regions.

The obtained dataset undergoes a normalization process by scaling input vectors in a (0,1) range. They can then be divided into training, validation, and test sets fed into the predictive model.

5.5.3 MLP Predictive Model

As a predictive model, I designed a Multilayer Perceptron. In the following, I describe in detail the MLP architectures and corresponding design parameters that provided the best performance values for the multi-class and binary classification tasks discussed in Section 5.6.

This MLP architecture, represented in Figure 5.8, is composed of a hidden layer with 64 nodes and two hidden layers with 32 nodes each. The first hidden layer applies a 'sigmoid' activation function on the entries. In the following hidden layers,

nodes apply the 'ReLu' (rectified linear unit) activation function on their inputs. Three Dropout layers are placed after the hidden layers to avoid the overfitting phenomenon, occurring when the MLP specializes too much on the training set, losing its ability to generalize on the validation set. When the error on the validation set starts to increase, indicating possible overfitting, the dropout layers "drop out" random neurons, temporally removing their contribution to downstream the activation of neurons. It has been widely demonstrated to improve the generalization capabilities of the network [92]. Notably, two options are given for the activation function of the output layer: softmax and sigmoid, respectively for multi-class and binary classifications.

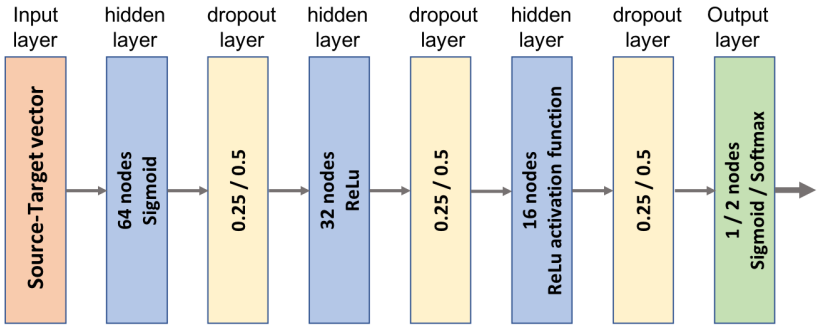


Figure 5.8: MLP architecture for classification tasks. For each layer, I report number of nodes, activation function and dropout value. When values are different for binary and multi-class tasks, we report them both, separated by a slash symbol.

5.6 Results

In this section I will focus on the results of the methodology. More specifically, I will assess the proposed solution in verifying whether gene expression profiles contain enough information to predict the intensity of anatomical connections between brain regions. On top of that, I will provide a quantitative evaluation of a classification system's performance leveraging gene expression profiles as input and connectivity as the classification label.

5.6.1 Classification performance

The classification consists of a multi-class and a binary classification task, respectively, where the Source-Target vectors are grouped into a corresponding number of categories representing their connectivity degree, leveraging the MLP architectures described in *Section 5.5.3*.

In this section, I will focus on the results of the methodology. More specifically, I will assess the proposed solution in verifying whether gene expression profiles contain enough information to predict the intensity of anatomical connections between brain regions. On top of that, I will provide a quantitative evaluation of a classification system's performance leveraging gene expression profiles as input and connectivity as the classification label.

Multi-class classification task

To generate a dataset for the multi-class task, all the available Source-Target vectors were divided into three categories based on empirical thresholds on the connectivity values provided by the MBCA database:

1. **Class label "0" (unconnected)**: 5,000 Source-Target vectors with connectivity equal to 0
2. **Class label "1" (weekly connected)**: 5,000 Source-Target vectors with connectivity values in the range [0.006, 0.1)
3. **Class label "2" (strongly connected)**: 4,583 Source-Target vectors with connectivity > 0.1

Therefore, the whole dataset was composed of 14,583×6,636 vectors with their corresponding class labels.
This dataset was randomly split into three disjoint subsets used for training, validation, and testing purposes. The three sets contained 10,499 vectors, 1,167 vectors, and 2,917 vectors, respectively.

The MLP architecture implemented to solve the multi-class classification problem will be described in detail in *Section 5.5.3*. The training phase consisted of 200 epochs in total, during which the dataset was propagated in batches of size 6. At the end of each propagation, the error between predicted values and desired outputs was quantified in terms of the *categorical cross-entropy* loss function. Working towards the minimization of the error, Nesterov-accelerated Adaptive Moment Estimation (*Nadam* [57]) optimizer updated parameters with a *learning rate* of 0.002 for each training example. The training procedure's full parameter set is summarized in Table 5.1 to ensure the experiment's full reproducibility.

MLP performance was computed in terms of classification errors (i.e., the fraction of input instances that were correctly assigned to their specific class category). After 200 learning epochs, MLP training accuracy reached an accuracy value on the training set of 0.914, ensuring the model's convergence. Nonetheless, the classification accuracy of the trained MLP decreased to 0.764 when computed on the

Table 5.1: Training parameters for multi-class classification with the Nadam optimizer.

epochs	learning rate (Lr)	**decay**	**beta1**	**beta9**	**Loss function**	**batch size**
200	0.002	0.004	0.9	0.999	categorical cross entropy	6

test dataset containing completely unseen data, suggesting an over-fitting problem.

To have a more in-depth view of the classification outcome, in Table 5.2 I show a confusion matrix, with rows and columns representing respectively items in the real and the predicted class. Hence, the main diagonal of the matrix reports the percentage of instances correctly classified, separately for the three different class categories, while the other elements of the matrix show the misclassified items and their respective distributions.

Table 5.2: Confusion matrix for multi-class classification

		Predicted class		
		unconnected	***weakly connected***	***strongly connected***
Real class	***unconnected***	75%	23%	2%
	weakly connected	13%	81%	6%
	strongly connected	2%	31%	67%

As it can be gathered from the confusion matrix, the classifier had heterogeneous classification outcomes, with the best classification accuracy (81%) for the instances with weak connection levels and the lowest accuracy (67%) for the ones with a strong connection. Unconnected instances were detected with an adequate level of accuracy (75%).

In general, very few misclassifications happened involving two class categories at the extremes: only 2% of the unconnected instances and the strong samples were wrongly assigned to the strongly connected class and the unconnected class, respectively. The most frequent misclassifications (31%) consisted of strongly connected samples classified in the weakly connected class. This phenomenon is probably due to the slight overfitting of the MLP towards this class, suggesting that the training data were not representative enough for a three-class categorization.

In Table 5.3, I report the whole set of quality metrics (i.e., recall, precision, F1 score, and accuracy [171]) obtained for each class, which confirm the analysis provided above.

The following consideration can be drawn observing the overall outcome of the classification. While the MLP classifier provides only partial discrimination of the connectivity degree, it has an acceptable accuracy in differentiating between

Table 5.3: Quality metrics for multi-class classification

		Quality metrics			
		Recall	***Precision***	***F1_score***	***Accuracy***
Class	***unconnected***	75%	86%	81%	76%
	weakly connected	81%	66%	73%	
	strongly connected	67%	83%	74%	

zones with connection (i.e., weakly or strongly connected class) and zones without connection (i.e., unconnected class).

Binary classification task

In light of the results obtained in the multi-class predictions, to boost the classifier capabilities in discriminating between connected and unconnected areas, I designed a binary MLP. To perform the binary classification task, this time, I divided the available dataset into two sub-sets, as follows:

1. **Class label "0" (unconnected)**: 20,000 Source-Target vectors with connectivity values equal to 0. This sub-set is composed of gene expression vectors obtained, selecting only unconnected Source-Target region pairs.

2. **Class label "1" (connected)**: 17,136 Source-Target vectors with connectivity values > 0.006.

Therefore, the overall dataset contained 37,136×6,636 vectors, with their corresponding binary labels. The whole dataset was divided into training, validation, and test set, containing 26,737 vectors, 2,516 vectors, and 7,428 vectors.

The training phase consisted of 100 epochs, during which the dataset was propagated in batches of size 32. Again, at the end of each propagation, the error between predicted values and desired outputs was calculated by the *binary cross entropy* loss function. The learning procedure leveraged the *Nadam* optimizer, updating parameters with a learning rate of 0.002 for each training example. The comprehensive set of the training parameters are shown in Table 5.4.

Table 5.4: Training parameters for binary classification with the Nadam optimizer

epochs	**learning rate (Lr)**	**decay**	**beta1**	**beta9**	**Loss function**	**batch size**
100	0.002	0.004	0.9	0.999	binary cross entropy	32

As shown in the training curves of Figure 5.9, after the 100 epochs of training, the MLP reached 0.89 training accuracy with 0.247 loss. On the other hand, the validation accuracy turned out to be not much lower than the training accuracy (around 0.85), suggesting a correct convergence without overfitting.

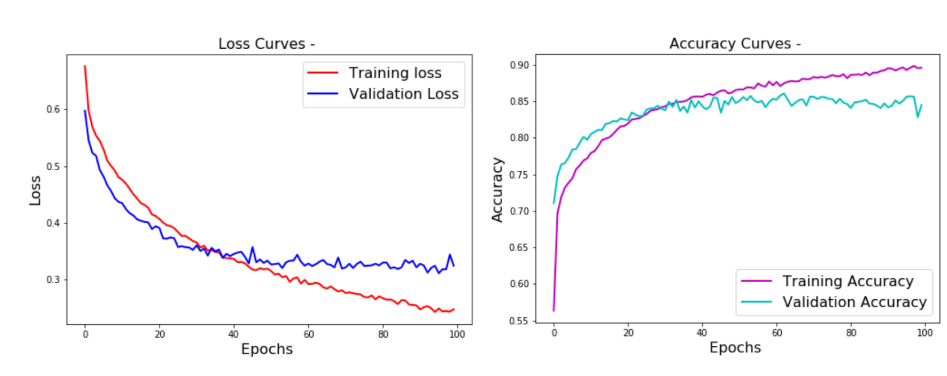


Figure 5.9: Training performance curves (loss on the left, accuracy on the right) of the binary classifier.

To assess the performance of the trained MLP on both the classes, besides accuracy, I quantified precision, recall, and F1 score on the test set (see values reported in Table 5.5).

Table 5.5: Quality metrics for binary classification

	Quality metrics			
	Precision	***Recall***	***F1_score***	***Accuracy***
unconnected	94%	75%	84%	85%
connected	77%	95%	85%	

As can be gathered from the table, the overall classification outcome was positive (85% accuracy), with a reasonable balance between precision and recall in both the class categories. The unconnected class has a better precision value (94%), and vice-versa the connected class has higher recall (95%), but both classes have similarly high values of F1 scores (84% and 85%, respectively).

The classification system's good performance is also confirmed by the ROC curve's shape (in Figure 5.10), with the area under the curve equal to 0.943.

This last experiment demonstrates that it is possible to distinguish between connected and unconnected regions reliably. The fine discrimination between different intensities of physical connections is also possible, but with more uncertainty, most probably due to the training data's technological noise.

5.7 Conclusions

As demonstrated by our results, our gene expression data-driven approach allows distinguishing between connected and unconnected brain areas at a cellular

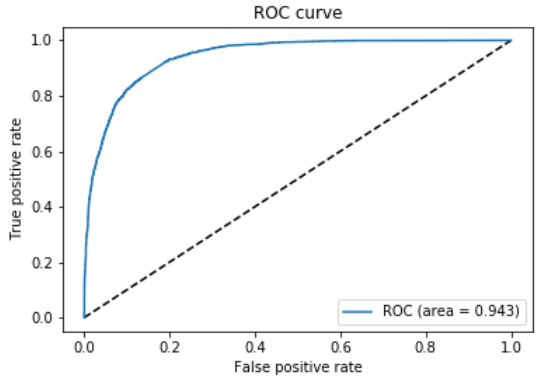


Figure 5.10: Receiver operating characteristic (ROC) curve on the test set for binary classification.

resolution scale, with no need for extensive parametrization or a priori knowledge of the process.

This approach opens the way to more in-depth investigations on brain connectome and brain functionality's genetic footprint. The possible directions for this study are mainly two. The first is aimed at extending the available knowledge on brain connectivity. Indeed, structural information on neural circuitry (e.g., BAMS) is to this date characterized by a large number of unknown connections and missing data. In the long term, this work's second direction will investigate the connectivity prediction model's transferability from mice to other mammalians (especially humans).

In this regard, the main research question is how and to which extent the prediction model trained on the Allen Mouse Brain Atlas can be applied to infer the anatomical connectivity of more complex brains, possibly exploiting not only gene expression levels from in-situ hybridization but also RNA-Seq data. The model could be used either as-is or after partial fine-tuning of the network on new training data.

Chapter 6

miRNA-target predictions in a multi-omics dataset

6.1 Methodological contribution

This chapter describes a statistical method to reveal determining microRNA targets in multi-omic datasets.

The main computational problem regards the selection of relevant patterns in a multi-omics dataset. Indeed, many feature selection algorithms have been proposed. However, given the high dimensionality of the data, it is not an easy task.

Here, two multi-omics datasets are used: breast cancer and medulloblastoma datasets. Both the datasets are composed of miRNA, mRNA, and proteomics data related to the same patients. The main computational contribution to the field consists of designing and implementing an algorithm based on the statistical conditional probability to infer the impact of miRNA post-transcriptional regulation on target genes exploiting the protein expression values. The developed methodology allowed a more in-depth understanding and identification of target genes. Also, it proved to be significantly enriched in three well-known databases (miRDB, TargetScan, and miRTarBase), leading to relevant biological insights.

6.2 Introduction

The cost reduction of next-generation sequencing techniques has recently facilitated generating a significant amount of genomic data. Mass spectrometry has also advanced considerably, allowing in recent years an increase in the proteomics data[162].

Therefore, it is a well-established practice to analyze each omic individually. Besides, several multi-omics data integration techniques have been proposed. In this context, the most common objectives consist of identifying patient clusters or clusters of relevant characteristics such as genes, miRNA, CpG probes, and many others [211, 16, 181, 199, 38, 20].

However, technologies for quantifying proteins in a given sample mainly consist of mass spectrometry or microarrays. Although proteomics provides complementary information to the transcriptome, the quantification of proteins is affected by a method's inexistence to amplify the biological material (such as PCR for the transcriptome). These technological limitations have meant that proteomics has developed more recently than transcriptomics. Consequently, the availability of multi-omics data in which proteomics is present in a massive way is still limited[210, 87, 207].

In the basic formulation of molecular biology's central dogma, DNA is first transcribed into mRNA and then subsequently translated into proteins[51]. As stated by Vogel et al., transcriptional regulation is only half the story[209, 170]. In fact, given a late amount of mRNA, the corresponding protein amount is not necessarily proportional. Indeed, various post-transcriptional processes intervene in the regulatory chain, modifying the expression of a protein[18].

Among the many post-transcriptional processes, miRNAs are distinguished, proven to be involved in various regulated processes, and directly responsible for some pathologies[202, 165]. The mechanism of regulation of miRNAs occurs through the silencing of genes. This process occurs either through the degradation of the mRNA or by preventing the mRNA from being translated. In the first case, the miRNA is complementary to a part of the target gene and directly degrades the mRNA molecule. In the second case, however, the mRNA is not degraded, but the translation is inhibited, limiting the amount of protein produced.

In the literature, there are several tools to identify miRNA targets. Such tools are often based on quantifying the correlation between mRNA and miRNA expression[39, 150]. Moreover, they often focus attention on the effect of a single miRNA rather than on their combinatorial function in which a single miRNA can have multiple targets. Indeed, many miRNAs are organized in genomic clusters, and co-regulated miRNAs can jointly target different molecular pathways.

Among the many post-transcriptional processes, here is a new method for identifying miRNA targets by exploiting the proteome's information. This method is based on the computation of partial correlation as a measure of relevance of the

miRNA targets considering the transcriptome and proteomics' expression values simultaneously. In this way, it is possible to overcome the previous models' limitations, which traditionally did not consider the proteome.

Below, the Methods section illustrates the proposed approach's details, while the Results and Discussion sections summarize the main findings obtained.

6.3 Methods

In this section, the central part of the method will be discussed. In particular, the method devised for identifying target genes will be presented, as well as the three databases used to verify their enrichment and biological relevance.

This method exploits the statistical measure of partial correlation, which measures the correlation between two variables (x and y) by checking the effects of a third variable, z.

The partial correlation between x and y by controlling for z is computed according to the following formula:

$$r_{yx.z} = \frac{r_{yx} - (r_{yz})(r_{xz})}{\sqrt{1 - r_{yz}^2}\sqrt{1 - r_{xz}^2}}$$

where r_{ab} represents the bi-variate correlation between a and b variables.

In this way, it is possible to measure the correlation between mRNA and the related protein by attributing the discrepancy between mRNA and protein to the miRNA's inhibitory action on that particular gene. In detail, for each coding gene whose protein expression value was known, the partial correlation between the gene (variable x) and the relative protein (variable y) was calculated for each miRNA (variable z).

Concerning the partial correlation calculation, both the intensity of the correlation (estimate) and the associated p-value is obtained for each possible gene-miRNA pair (defined as target-miRNA). These measures can be inserted for convenience in a table like in Figures 6.1, 6.2.

It is necessary to identify among the many possible pairs those significantly relevant in our dataset. For this purpose, three conditions are imposed:

1. significant p-value of partial correlation (e.g., 10-7)

Figure 6.1: Partial correlation estimate value e_{ij} computed for gene i and miRNA j. These values are easily stored in a table.

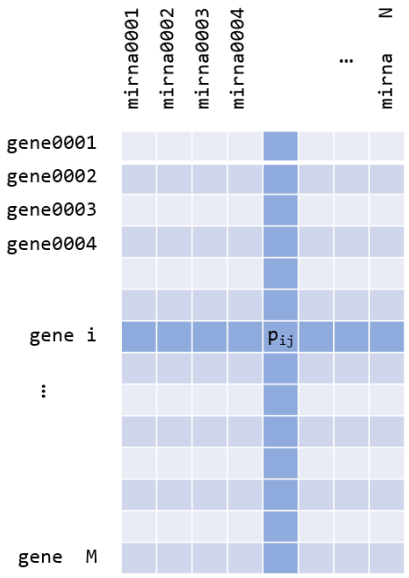


Figure 6.2: Partial correlation p-value p_{ij} computed for each gene i and miRNA j. These values are easily stored in a table.

2. high estimate (e.g., 0.6). This condition is not independent of the previous one. Indeed, selecting pairs with a significant p-value leads to the identification of high estimate values. The "high estimate" condition serves to restrict the potential targets further.

3. Estimate of partial correlation greater than or equal to the bivariate correlation between mRNA and protein:

$$r_{yx.z} > r_{yx}$$

This constraint ensures that, by unbundling miRNA's inhibitory effect, the correlation between mRNA and protein is improved.

Therefore, the proposed model considers potential targets the genes and miRNAs that satisfy all three conditions described above.

Once the potential targets have been obtained, it is essential to verify that they are enriched in one or more of the following databases:

- miRDB: an online database for predicting functional microRNA targets based on recently updated transcriptome-wide target prediction data[43, 213, 215].

- TargetScan: it predicts biological targets of miRNAs by searching for the presence of conserved 8mer, 7mer, and 6mer sites that match the seed region of each miRNA[75, 136, 83].
- miRTarBase: it is a curated database of MicroRNA-Target Interactions[93, 45, 46, 94].

In addition, to evaluate the significance of such enrichment, the Fisher test, and the hypergeometric test are performed.

Finally, to make this information accessible to the user, a heatmap is created, showing the miRNAs on the abscissa and the potential target genes on the ordinate. Both the miRNAs and the genes on both axes are sorted according to their genomic position. In this way, the user can visually check for the presence of multiple miRNAs targeting multiple genes by considering potential horizontal or vertical bands in the heatmap.

6.4 Results

In this section, the results obtained on a dataset of 77 breast cancer patients are presented. Subsequently, by way of validation, the results are reported on a cohort of 26 medulloblastoma samples.

6.4.1 Results on breast cancer dataset

The proposed method was first applied to a dataset of 77 breast cancer patients. For each of these patients, the expression value of mRNA, miRNA, and proteomics are available [157]. The partial correlation value and the relative p-value were then calculated for each gene corresponding to its protein, considering all possible miRNAs as z control variables.

Figure 6.3 and 6.4 show the histograms containing the distribution of the estimates and p-values for all the possible miRNA target pairs.

Subsequently, those satisfying the three criteria previously described in the method were selected among all the miRNA target pairs. The thresholds' chosen value corresponds to 10-7, 0.6, and 0 respectively for the first, second, and third conditions. In this way, 155738 miRNA target pairs were selected by the method over more than 3 millions pairs.

Furthermore, to verify which miRNA target pairs were already known or predicted in the literature, the enrichment of these pairs was verified in miRDB, TargetScan, and miRTarBase. Besides, to evaluate the significance of enrichment, the

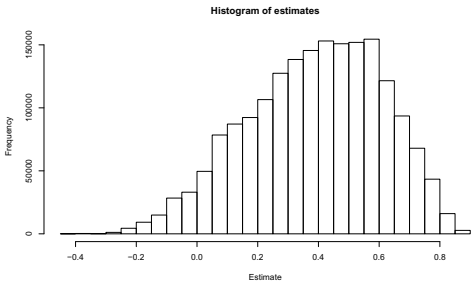


Figure 6.3: Breast cancer. Histogram of partial correlation estimates.

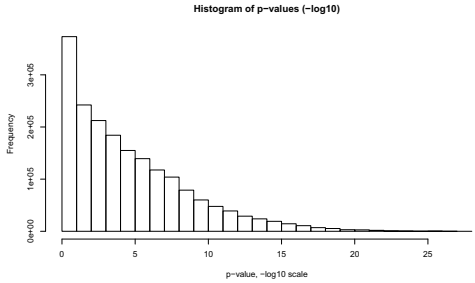


Figure 6.4: Breast cancer. Histogram of partial correlation p-values in logarithmic scale (-log10).

	Selected area (three condistions)	**Negative control (estimate around 0)**
miRDB	**6342/155738**	**1067/31728**
hypergeometric p-value	2.6245e-05	0.9999996
TargetScan	**1708/155738**	**270/31728**
hypergeometric p-value	0.529972	0.9999948
miRTarBase	**2835/155738**	**435/31728**
hypergeometric p-value	0.1342582	1

Table 6.1: **Breast cancer.** Overlap and statistic measures to test the enrichment in the three databases: miRDB, TargetScan, mirTarBase.

Fisher test, and the hypergeometric test were conducted. Table 6.1 shows the values of each intersection and the respective p-values.

Besides, to assess the method's relevance, a control zone was selected, corresponding to the region in which the correlation estimate value of the partial correlation is not significant and is around zero ([-0.2, 0.2]). Also, in this area, enrichment in miRDB, TargetScan, and miRTarBase was evaluated. The hypergeometric test was calculated. The results are reported in the last column in Table 6.1.

Finally, the Figure 6.8 represents the enrichment heatmap. For each target-miRNA pair, the color is gray if that pair is considered significant by the proposed method, while it is in red if it is enriched in at least one of the three databases.

6.4.2 Validation on medulloblastoma dataset

Another dataset containing 26 medulloblastoma samples was used to validate the proposed methodology [71]. On it, the method described above was applied. Figures 6.5 and 6.6 shows the distribution of estimates and p-values. Based on the histogram distribution, selecting the relevant miRNA target pairs is made using 10-7, 0.8, and 0.01 for the first, second, and third conditions, respectively. As regards the control zone, the region [-0.2, 0.2] was chosen.

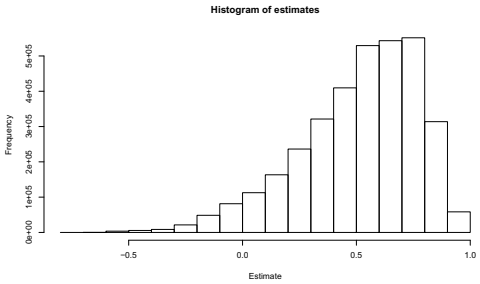


Figure 6.5: Medulloblastoma cancer. Histogram of partial correlation estimates.

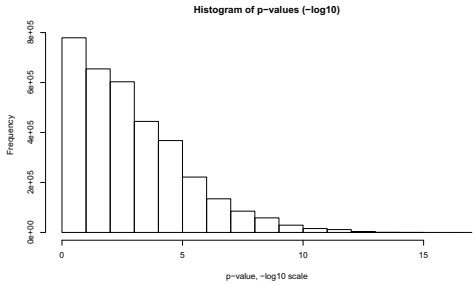


Figure 6.6: Medulloblastoma cancer. Histogram of partial correlation p-values in logarithmic scale (-log10).

All possible miRNA target pairs were verified for enrichment in miRDB, TargetScan, and miRTarBase, and the hypergeometric test was calculated for both the region of interest and the control region. All values are shown in Table 6.2.

	Selected area (three condistions)	**Negative control (estimate around 0)**
miRDB	**1553/36952**	**11530/405757**
hypergeometric p-value	2.333484e-12	1
TargetScan	**242/36952**	**1412/405757**
hypergeometric p-value	1.037228e-09	1
miRTarBase	**539/36952**	**5095/405757**
hypergeometric p-value	0.002015418	0.9776983

Table 6.2: **Medulloblastoma cancer.** Overlap and statistic measures to test the enrichment in the three databases: miRDB, TargetScan, mirTarBase.

Finally, the heatmap representing the selected and validated miRNA target

pairs is shown in Figure 6.9.

6.5 Discussion

In this section, the main results obtained will be discussed.

The first consideration concerns the distribution of partial correlation estimates for both the breast cancer dataset and medulloblastoma. In both cases, this measure is predominantly positive. This fact reflects the existing evidence indicating a positive correlation between target and miRNA.

The two histograms have a different kurtosis, hence the need to insert specific significance thresholds for the method depending on the dataset. However, the minimal variation of the thresholds does not produce different effects in terms of results.

As anticipated in the results section, it is advisable to verify the significance of enriching the targets found in the various databases (miRDB, TargetScan, miRTarBase). Given a situation like the one in Figure 6.7, for each database, it has been obtained:

- the number of target-miRNA pairs selected in the dataset,
- the number of target-miRNA pairs present in that database out of the total number of target-miRNA pairs selected by the method and
- the total number of target-miRNA pairs present in that database.

This information allows calculating the significance of the overlap through the Fisher and the hypergeometric test. As can be seen from Tables 6.1 and 6.2 in the results, most of the overlaps are significant. In all circumstances, the number of target-miRNA pairs found in miRDB is always higher than those found in TargetScan or miRTarBase. This circumstance can be attributed to the fact that miRDB contains a certain number of targets predicted through an algorithm based on artificial intelligence, and it contains a large number of predictions.

The hypergeometric test was calculated to investigate this enrichment better, i.e., the probability that an overlap equal to or greater than that tested is significant.

It is possible to note how the hypergeometric test's significance is better in the selected intersection than in control one. Indeed, although the intersection with the database (e.g., miRDB) in the control region at the turn of 0 is significant, the event of finding a larger overlap between the control region and the database is not

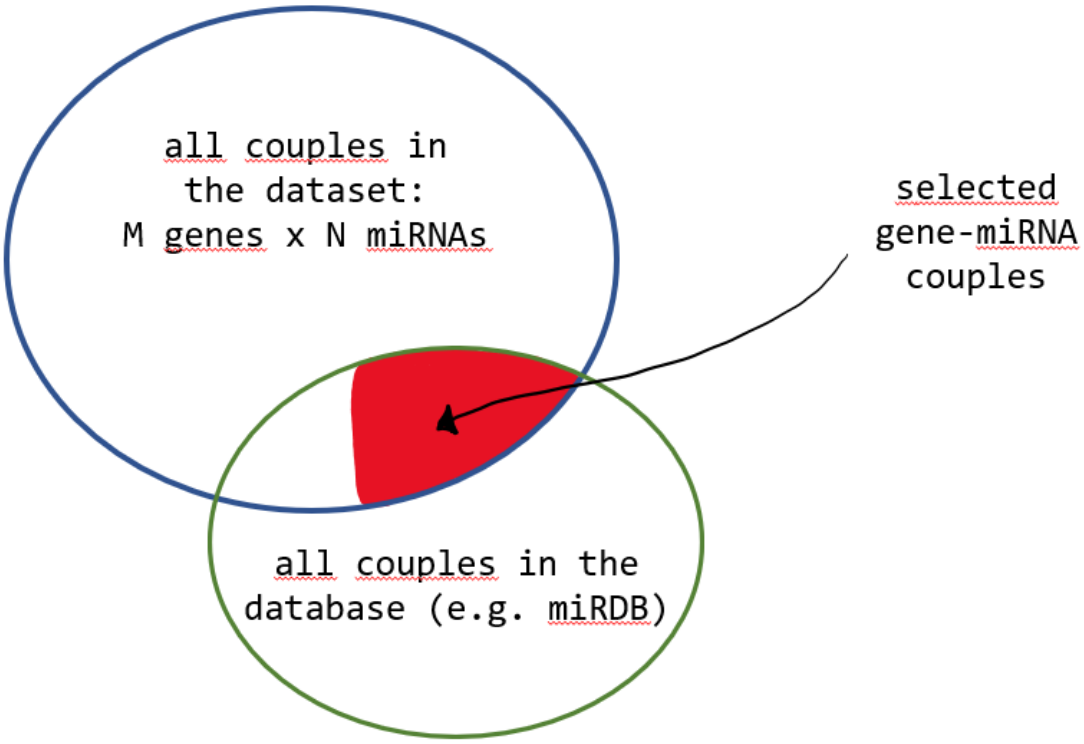


Figure 6.7: Visual representation of the sets. The blue set contains all possible gene-miRNA couples available in the dataset, and the red area contains the gene-miRNA couples selected by the method as significant. The green set represents all the gene-miRNA couples available in a database (e.g., miRDB). Consequently, gene-miRNA couples selected as relevant by the method can be present or not in the database. The hypergeometric test computes the significance of this overlap.

statistically significant.

Finally, this method is beneficial for identifying potential new targets in a pathology. Being a purely statistical method and not based on previous information allows for exploratory analyses to identify new targets. Besides, the user can quickly inspect the result of the investigation through the final plot. Target genes and miRNAs are sorted based on their genomic coordinates, highlighting miRNA families that co-regulate the same targets.

6.6 Conclusions

Identifying miRNA targets is a challenging problem as it is useful to identify new targets and evaluate them experimentally for research and disease treatment purposes. In the literature, various approaches have been proposed by exploiting the information contained in the transcriptome.

In this work, a new approach is proposed and is based on the statistical metric of partial correlation. This model, therefore, described the phenomenon of translation prevention by miRNAs towards target genes.

The model was presented on a breast cancer dataset and further validated on an additional cohort of 26 medulloblastoma patients. The proposed model effectively identifies miRNA targets, which are significantly enriched in three literature databases: miRDB, TargetScan, miRTarBase. Besides, the final heatmap allows you to graphically view the new targets and simultaneously view which of these are validated.

Therefore, this approach uses the proteome data to successfully identify new potential targets, bringing new and significant knowledge in this area, which can also be used for clinical purposes.

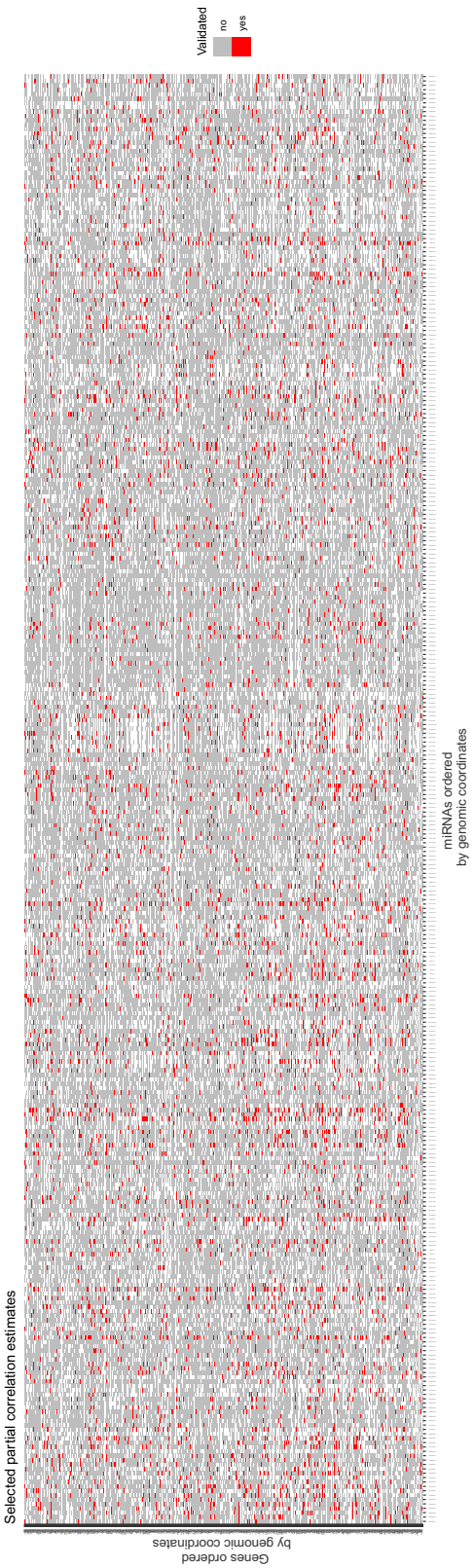


Figure 6.8: **Breast cancer heatmap** of selected miRNA targets. Gray areas represent selected miRNA-gene target. Red areas represent a miRNA-gene target validated in one of miRDB, TagetScan, miRTarBase. Genes and miRNAs are sorted according to their genomic coordinates.

Figure 6.9: **Medulloblastoma cancer heatmap** of selected miRNA targets. Gray areas represent selected miRNA-gene target. Red areas represent a miRNA-gene target validated in one of miRDB, TagetScan, miRTarBase. Genes and miRNAs are sorted according to their genomic coordinates.

Chapter 7

The challenge of multiomics data for classification task

7.1 Methodological contribution

Although integrating different data is a strength in the biological field, the computational effort necessary to obtain helpful information starting from genomic data in an integrated form is considerable. From a computational point of view, each data has specific dynamics of the values (range values). It is fundamental to understand how to scale the input value to not condition the result based on a single omic.

Moreover, the complexity of integrating multi-omics data derives not only from the intrinsic dimensionality of each omic but from the modeling of a biological phenomenon that is complex by its nature, whose potential is not yet fully known. Therefore, it is necessary to clearly understand the aim of the multi-omics integration to filter out the noise and work with the signal of interest.

From a computational point of view, the information coded at the multi-omic level is sometimes redundant. The selection from those redundant parts is not always automatic, on the contrary.

This chapter deals with the classification of multi-omics samples. The literature's main approaches integrate all the features available for each sample upstream of the classifier (early integration approach) or create separate classifiers for each omic and subsequently define a consensus set rules (late integration approach). In this context, the main contribution consists of introducing the probability concept by creating a model based on Bayesian and MLP networks to achieve a consensus guided by the class label and its probability. This approach has shown how a probabilistic late integration classification is more specific than an early integration approach. Also, the proposed model can better identify anomalous samples

concerning the training domain. This tool is potent, as, in addition to recognizing outliers belonging to the same tissue used in training, it excludes from the classification samples that belong to a different tissue or tumor subtype than that with which the model was trained. This aspect represents a significant advantage in the clinics.

7.2 Background

In recent years, the reduction of costs for the sequencing of biological molecules, including DNA, RNA, and proteins, has allowed the widespread of vast amounts of data both in the form of large structured databases and in the form of repositories specially created for the study of particular pathologies [22, 147, 36, 84, 134].

In this context, various omic data can be taken into account for the study and analysis of samples, either tumor or healthy data: gene expression (mRNA), microRNA expression (miRNA), methylation (meth), copy number alterations (CNA), single nucleotide polymorphism (SNV), proteomics and phosphoproteomics [111, 96].

Two strands are typically available in the multi-omics analysis: first, the subdivision of samples into classes [211, 16, 181, 199] and second, the identification of specific pathways and gene patterns in the dataset [38, 20]. This work focuses exclusively on the first strand; in particular, a method for the classification of cancer samples by simultaneously exploiting the information from different omics is presented. Although the work relies on mRNA, miRNA, and meth, it must be noticed that the same algorithm can be extended to other omics.

In the multi-omics classification approach, a crucial step is represented by the algorithm used to integrate the different omics.

To this aim, the intuitive approach consists of training the model on a dataset obtained by concatenating all the features available for a unique sample. This procedure is also called early integration, as all the information is merged before training the model [173].

By contrast, another approach is to create an individual classifier for each omic and subsequently integrate the classification result. This approach is defined as late integration. One of the late integration techniques consists in making a consensus among the various omics, in such a way that the multi-omic class is the most voted class among the outputs on the individual omics [76, 149]. The main drawback of the majority voting techniques is the difficulty of assigning the multi-omic class in case the output on each omic is different, or if more than one class is equally voted among all the omics.

To date, most of the integration techniques in biology are based either on the

early or late integration approach, without considering the certainty in the classification process. This aspect is particularly relevant since different biological features may characterize each omic. Then, depending on the specific context, the information carried by some omics could be more relevant than others for classification. Therefore, considering the classification's certainty allows one to discern each omic's actual contribution and provide a more conservative classification.

This work proposes the use of a late integration learning method that for each omic returns not only the class of a sample but also the class membership probability, allowing a deeper understanding of the classification certainty.

The proposed probability aware late integration method has been compared to an early integration method built with the same late integration method architecture. The results suggested that the late integration approach can provide a more conservative classification discarding uncertain predictions and identifying outlier samples.

In particular, the use of the class-membership probability allows to filter samples according to the class probability and consequently postpone for further analyses those samples on which there is not enough certainty in the classification across all the omics. This approach is advantageous in creating automatic tools that, integrating different omic information, may favor the clinical practice by proposing a classification label when all the omics are confident enough in their classification and an *Unknown* label when discrepancies are found across the omics. In this way, physicians can quickly look at well-defined samples and focus more on the most exciting and challenging cases where human control is crucial.

7.3 Methods

7.3.1 Biological data

Although the proposed method can be applied to any tissue and pathology, this work deals with the study of kidney tumor samples freely available in the Genomic Data Commons (GDC) database [84]. The samples used to train and test the method belong to three main kidney tumor subtypes: kidney renal papillary cell carcinoma (KIRP), kidney renal clear cell carcinoma (KIRCH), and kidney chromophobe (KICH). Besides, a reduced number of healthy samples is available for both KIRP and KICH subtypes (usually, these tissues are healthy areas surrounding a KIRP or KICH tumor). For KIRP, KIRCH, and KICH subtypes, only samples available for mRNA, miRNA, and meth data are selected, obtaining a final dataset of 909 kidney samples.

The mRNA, miRNA, and meth data are tabular data commonly represented

as matrices, where the value in position *(i,j)* represents the amount of a specific biological product or the intensity of a phenomenon (mRNA, miRNA, and meth, respectively) in a specific sample. The mRNA, miRNA, and meth matrices carry different biological information. The mRNA expression value is strictly related to the amount of its protein (higher is the number, higher the protein) which regulates a specific pathway in the cell life cycle.

The miRNA expression value indicates the amount of a specific miRNA, a small non-coding RNA molecule that intervenes in the post-transcriptional process, regulating the amount of final protein produced.

Methylation value refers to the methylation beta value, an estimate of the methylation level computed as the ratio of intensities between methylated and unmethylated alleles. The biological effect of methylation consists of the change of a DNA segment's activity without changing its sequence (when methylation occurs, it reduces the DNA transcription, thus reducing the amount of protein).

It must be noticed that many biological molecules act together in order to regulate the cell activity and that changes in the values of one or more omics can be correlated to a specific pathology or a tumor subtype.

Data preprocessing

After downloading and selecting samples for which both mRNA, miRNA, and meth data are available, the following preprocessing is performed:

- **mRNA:** Raw count data have originally about 60000 mRNA genes and have been normalized using the Variance Stabilizing Transformation (VST) [98]. Not protein-coding genes have been discarded, reaching about 20000 mRNA genes, and z-score transformation has been performed.
- **miRNA:** The miRNA data have about 2000 miRNAs and have been normalized using deseq [144]. Then pseudo-counts have been computed as $log_2(count\_value+1)$. In the end, z-score transformation has been performed.
- **meth:** Among the 27000 features in methylation array data obtained with Illumina Human Methylation 27 platform (450000 features for the Illumina Human Methylation 450), the batch effect due to the use of different platforms has been corrected using the limma package [179]. Since original data are intrinsically normalized, no further normalization is required.

The 909 kidney samples belong to 5 classes: **tumor KIRCH:** 509, **tumor KIRP:** 288, **tumor KICH:** 65, **healthy KIRCH:** 24, **healthy KIRP:** 23. They have been further divided into a training set (75% of the samples) and test set (25% of the samples) such that the latter includes the same proportion of samples belonging to the different classes. Furthermore, to overcome the problem of a

too unbalanced dataset, the SMOTE over-sampling technique was used [42]. In this way, it was possible to perform an augmentation operation on the training set. In order to test the model on samples that do not belong to the kidney classes, 37 stomach samples and 817 lung samples have been obtained from GDC [84], by applying the same preprocessing steps described at the beginning of this paragraph. These datasets are used only as test sets without re-training the kidney model. The aim is to evaluate the ability of the probabilistic approaches in recognizing unseen classes.

MLP Model

Since an MLP equipped with a cross-entropy loss function, with associated either logistic sigmoid (two-class problem) or softmax (multiclass problem), outputs the class-membership posterior probabilities of the inputs [33], the MLP classifier is, therefore, able to return the class label and the associated probability of the sample belonging to a class.

As shown in Table 7.1, minimal architectures in the implemented neural networks were used (e.g., MLP with a single hidden layer with 20 neurons and a single activation layer). Many hyper-parameters configurations have been considered. In the end, gradient descent, with backpropagation and the cross-entropy as loss function, was used. The optimizer was Adam.

Layer	Activation function	Input Size	Output Size
fully connected	Relu	X	20
fully connected	SoftMax	20	y

Table 7.1: Structure of each MLP node used to build the tree-MLP architecture. X size is the total numer of features for mRNA, meth, and miRNA data. The y size depends on how many classes must be predicted (2 for root MLP and healthy leaf MLP, and 3 for tumor leaf MLP).

Late Integration

In this paragraph, the late integration approach is described. Before performing the integration, the dimensionality of each omic was reduced through the PCA. Subsequently, a tree-MLP model was applied to each omic.

Once the classification on each omic is performed, the final consensus is built considering the final probabilities on each omic. Given:

- n: the number of the omics,

- m: the number of the classes,
- th: threshold on the omics, in order to filter predictions with low probabilities across all the omics,
- tr: threshold on the classes, in order to select only samples with a not uniform distribution of the class-membership probabilities across the m classes,
- P_{ij}: the class membership probability for class i and omic j,
- $S_i = \sum_{j=1}^{n} P_{ij}$: the sum of the probabilities on all the omics for a single class,
- $S_a = \sum_{i=1}^{m} S_i$: the sum of the probabilities on all the omics and all the samples,
- $S_m = S_i/n$: the mean of the probabilities on all the omics for a single class.

The consensus for a sample is built according to the following formula:

$$y_{consensus} = \begin{cases} Unknown, & \text{if } \max_i(S_m) < th \text{ or } \max_i(S_i)/S_a < tr \\ \arg\max_i(S_i), & \text{otherwise} \end{cases}$$

In that way, a sample with a low mean probability across all the omics is labeled as *Unknown*. Besides, when a sample receives similar S_i values for more than one class, the model is uncertain in its prediction. Therefore, a tr threshold is set to select only samples with a non-uniform distribution of the class-membership probabilities across the m classes.

This final consensus can be applied using any number of omics as long as each omic represents different views of the same sample. Obviously, the larger the number of the omics, the more reliable the consensus prediction can be.

7.3.2 Classification assessment: late integration using different classification models

All models have been tested on both the test set, consisting of kidney samples belonging to the five classes of the training set, and on indipendent datasets like stomach and lung cancer samples. All models were implemented in Pytorch framework [169]. The Pyro library [32] was used for the BNN to transform the parameters into random variables and run stochastic variational inference. The methods compared for late integration classification assessment were the deterministic SVM and RF approach and the probabilistic MLP, tree-MLP, BNN, and Rotational Forest approach (see later for details). The final consensus for the probabilistic methods was evaluated according to the metric described in the previous paragraph. Then, it was the same as for the MLP model.

Support Vector Machine, Random Forest

In order to have a baseline for the results, a support vector machine (SVM) and a random forest (RF) classifier have been applied to the training set (with hyper-parameters optimization) [35, 52]. Unless these models do not output a class-membership probability, they can provide valuable insights into the data. Since they cannot estimate the certainty of their prediction, the implementation of the consensus has been slightly modified. The final consensus for SVM and RF classifiers is given by the majority voting between the different omics.

tree-MLP model

On the other hand, to compare the MLP architecture with other methods that return a class-membership probability, a tree-MLP classifier has been built. An extension of the multi-layer perceptron (MLP) combining several MLPs in a tree architecture (tree-MLP) is proposed here as a baseline. The aim is to evaluate a more flexible architecture's performances than the MLP model, which can be more easily updated in new subtyping.

As it can be seen in Fig. 7.1, a tree-like architecture was created with MLP models as nodes and trained separately on subsets of the training set. Each node is made up of an MLP whose architecture is the same as the one presented above. There is a root node (trained to recognize healthy from tumor samples) and two leaf nodes for this specific problem. The former is trained on healthy samples and classifies them into KIRP and KIRCH healthy tissues. The latter is trained on tumor samples and classifies them into KIRP, KIRCH, and KICH tumors. Therefore, given a new sample S, it will be classified by the root MLP as healthy or tumor (y_{root}) with a class-membership probability P_r. After selecting the leaf node corresponding to y_{root}, it returns the subclass label y_{leaf} (tumor_KIRP, tumor_KIRCH, and tumor_KICH for tumor leaf MLP; healthy_KIRP and healthy_KIRCH for normal leaf MLP) with its class-membership probability P_{leaf}. The final class y_{pred} is equal to y_{leaf}. The algorithm is formalized in Algorithm 1.

Bayesian Neural Network

Also, a Bayesian neural network (BNN) has been tested [24]. The BNN model has the same structure as the MLP; however, it works differently. Indeed, as the loss is modified with a Bayesian regularization term, its weights are no longer deterministic like an MLP but probabilistic, and each neuron learns to follow probabilistic distributions. Therefore, it is possible to infer the level of uncertainty of the class-membership probability estimation of the input, representing how much a sample belongs to a given class. The model is applied to the sample n times. The median

Algorithm 1: Algorithm of tree-MLP.

input: X, y ; `// whole dataset, where X is {`$x_1..x_n$`} and y is {tumor or healthy}`
input: X', y' ; `// healthy samples subset, where X' is {`$x'_1..x'_n$`} and y' is {healthy_KIRP, healthy_KIRCH}`
input: X", y"; `// tumor samples subset, where X" is {`$x''_1..x''_n$`} and y' is {tumor_KIRP, tumor_KIRCH, tumor_KICH}`
input: S ; `// sample to classify`
require: Θ_{root} ; `// MLP root model trained on X, y`
require: Θ_{leaf1} ; `// MLP leaf model trained on X', y'`
require: Θ_{leaf2} ; `// MLP leaf model trained on X", y"`
$P_r = \Theta_{root}(S)$;
$y_{root} = argmax(P_r)$;
if y_{root}==*0* **then**
  $\Theta_{chosen\_leaf} = \Theta_{leaf1}$;
else
  $\Theta_{chosen\_leaf} = \Theta_{leaf2}$;
end
$P_{leaf} = \Theta_{chosen\_leaf}(S)$;
$y_{pred} = argmax(P_{leaf})$;

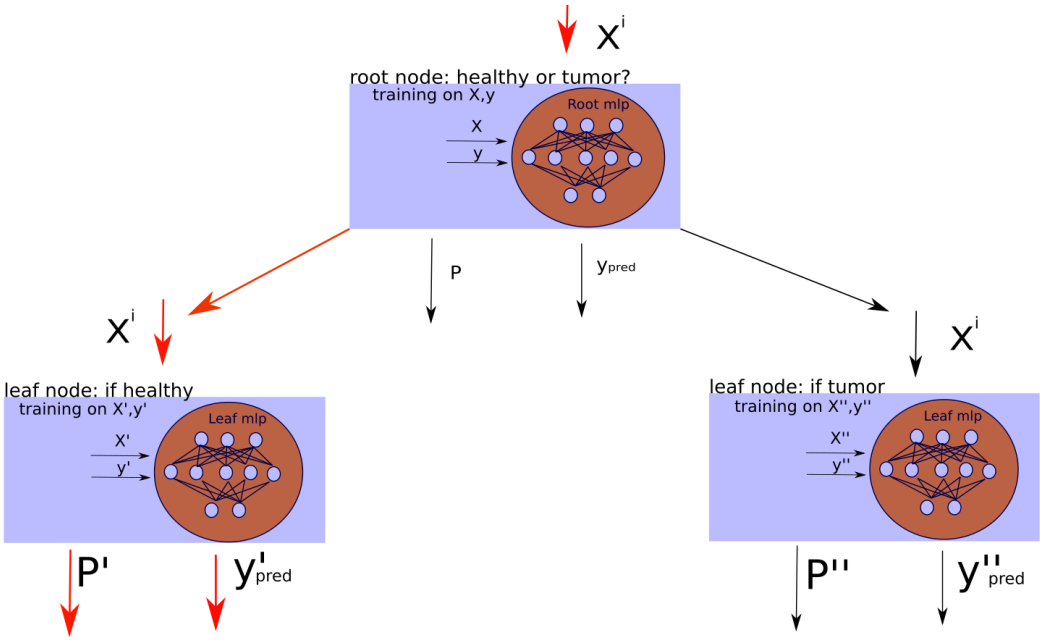


Figure 7.1: Proposed tree MLP model: i) each node is trained on three different subsets of the original dataset. (X, y) aims to distinguish between healthy and tumor samples, (X', y') between subtypes of healthy samples and (X", y") between subtypes of tumor samples; ii) the output of each node consists of the predicted label y_{pred} and the class-membership probability P.

value among all the output probabilities is selected as the final probability. For instance, if the median value is 0.95, the output is highly stable, and its classification uncertainty is very low.

Rotation Forest

The Rotation Forests method uses several estimators (in this application n= 10) to calculate a probability that a sample belongs to a class. Each estimator (Decision Tree) is trained over a rotation matrix obtained applying a feature extractor (PCA) over a subset of the original features. In the end, all the estimators participate in the final probability value computing the average result[182].

7.3.3 Early Integration

In this paragraph, the early integration approach is presented. The first step consists of normalizing the data separately on each omic through the z-score so that omics on different scales receive the same importance. Then all the omics are

grouped into a single dataset by concatenating all the features for a single sample. A data reduction process on the entire dataset takes place considering the first 21 main components obtained through the PCA method. In the end, for each type of architecture (MLP, tree-MLP, BNN, Rotation Forest), one model corresponding to the entire dataset has been trained. For each sample, the output consists of the class membership probability and the class label considering the entire dataset. To perform a fair comparison with the late integration approach, all the results reported in the Discussion Section refer to the MLP architecture.

7.4 Results

This section presents the results related to the proposed late integration method on kidney, stomach, and lung datasets. Kidney cancer is divided into 5 classes: tumor kidney chromophobe (KICH), tumor kidney renal clear cell carcinoma (KIRC), tumor kidney renal papillary cell carcinoma (KIRP), healthy KIRC, and healthy KIRP. For detailed information about the datasets refer to Section 7.3.

7.4.1 Late integration

The proposed approach consists of a probability-aware late integration method, based on a multi-layer perceptron (MLP) model with Principal Component Analysis (PCA) dimensionality reduction technique. The integration method exploits the class membership probability of each sample to provide an integrative class label. In particular, once the classification on each omic is performed, the final consensus is built combining the final probabilities on each omic, as detailed later. Two thresholds (th and tr) defining the sample classification uncertainty were set, and an uncertain sample was then classified as *Unknown*. A detailed description of the late integration approach and the formulation of the final consensus probability is reported in the Section Methods.
The model has been trained on a kidney dataset made up of 909 samples for which the information about mRNA, miRNA, and meth was available in the GDC repository[84]. Among them, 681 samples were used for training the late integration model and 228 for testing it.

The results related to the late integration MLP model have been obtained setting $th = 0.9$, and $tr = 0.25$. Principal Component Analysis (PCA) and Independent Component Analysis (ICA) have been tested on the kidney dataset to evaluate the effect of the dimensionality reduction techniques on the classifier [227, 226, 225, 95, 224]

The confusion matrices in Tables 7.2 and 7.3 report the details of the classification results respectively for PCA and ICA dimensionality reduction methods. As it

		Predicted					
		Healthy		Tumor			
		KIRC	KIRP	KICH	KIRC	KIRP	Unknown
Actual	healthy KIRC	3	0	0	0	0	2
	healthy KIRP	0	5	0	0	0	1
	tumor KICH	0	0	15	0	0	2
	tumor KIRC	0	0	2	114	1	11
	tumor KIRP	0	0	0	0	69	3

Table 7.2: Late Integration using MLP model with PCA dimensionality reduction technique.

		Predicted					
		Healthy		Tumor			
		KIRC	KIRP	KICH	KIRC	KIRP	Unknown
Actual	healthy KIRC	4	0	0	0	0	1
	healthy KIRP	0	4	0	0	0	2
	tumor KICH	0	0	15	0	0	2
	tumor KIRC	0	0	2	110	1	15
	tumor KIRP	0	0	0	0	65	7

Table 7.3: Late Integration using MLP model with ICA dimensionality reduction technique.

	Precision	Recall	F1-score	Accuracy	Support	*Unknown*
PCA	99%	99%	99%	99%	209	**8.33**%
ICA	99%	99%	99%	99%	201	11.84%

Table 7.4: Comparison between PCA and ICA preprocessed methods on the kidney test set using MLP model. All the reported metrics are computed on *not Unknown* samples. The support metric (i.e. the number of not Unknown samples) is the value on which the other metrics are computed.

can be seen in Table 7.4, independently from the type of dimensionality reduction technique, the MLP method reached 99% of accuracy and 99% of weighted average f1-score (see also results in Table 7.5, 4th row). The metrics were computed disregarding *Unknown* samples as they had not been assigned to any class. The MLP classifier with PCA selected as *Unknown* the 8.33% of kidney test set samples compared to the 11.84% of the MLP model equipped with ICA dimensionality reduction algorithm.

In the end, although the performance metrics precision, recall, F1-score, accuracy are comparable both for PCA and ICA MLP models, the final MLP selected

model is based on the PCA dimensionality reduction technique since the support metric is the highest. Consequently, the percentage of *Unknown* samples is lower.

7.4.2 Classification assessment: late integration using different classification models

Probability-aware and unaware methods have been used to evaluate the proposed model's performances versus other classification models. We considered Support Vector Machines (SVM) and Random Forest (RF) classifiers [35, 52] that do not consider the class-membership probability in the former class. For the latter, architectures based on the class-membership probability like Bayesian Neural Networks (BNN) [24], Rotation Forest [182], and a particular multi-layer perceptron (named tree-MLP) ad-hoc designed have been explored [212, 140]. Details about all the methods are reported in Section Methods. The performances on the kidney test set are reported in Table 7.5.

	Precision	Recall	F1-score	Accuracy	Support	*Unknown*
RF	97%	98%	97%	96%	228	-
SVM	97%	98%	98%	97%	228	-
tree-MLP	98%	98%	98%	98%	197	13.59%
MLP	99%	99%	99%	**99**%	209	**8.33**%
BNN	99%	99%	99%	99%	207	9.21%
Rotation Forest	99%	99%	99%	99%	179	21.49%

Table 7.5: Comparison between all the methods on the kidney test set. All the reported metrics are computed on *not Unknown* samples. The support metric (the number of not Unknown samples) is the value on which the other metrics are computed.

Regarding SVM and RF, the accuracy and weighted average f1-score have been obtained with (97%, 97%) and (96%, 96%) respectively. It should be noticed that the consensus creation for SVM and RF is different from that used in the MLP model since SVM and RF do not output the class-membership probabilities. Therefore, for SVM and RF classifiers, the consensus is based on the majority voting on the three omics without considering class probabilities.

Concerning the results of the four methods based on class-membership probability, all the metrics in Table 7.5 were computed disregarding *Unknown* samples. Indeed, the MLP model reached 99% of accuracy, precision, recall, and weighted average f1-score. It then classified as *Unknown* the 8,33% of kidney samples. The metric for evaluating the final consensus for tree-MLP, BNN, and Rotation Forest was the same as the MLP model. So, no biases were introduced. The tree-MLP

model reached 98% of accuracy and weighted average f1-score. The tree-MLP model classified as *Unknown* the 13.59% of kidney samples.
The BNN model classified as *Unknown* the 9.21% of kidney samples, and it achieved the 99% of accuracy and the 99% of weighted average f1-score. The Rotation Forest model classified as *Unknown* the 21.49% of kidney samples. Accuracy and weighted average f1-score were 99% and 99%, respectively. In conclusion, even if the four probabilistic models' performances were comparable, the MLP model labeled as *Unknown* the lowest percentage of samples.

7.4.3 Early Integration

In this section, the effect of combining the features from different omics before training the model is explored. More details about the early integration technique are reported in Section Methods. The confusion matrix on the kidney test set is presented in Table 7.6. The early integration method based on the MLP model predicts a class label for almost all the samples, reaching 95% of accuracy and 96% of f1-score.

		Predicted					
		Healthy		Tumor			
		KIRC	KIRP	KICH	KIRC	KIRP	Unknown
Actual	healthy KIRC	5	0	0	0	0	0
	healthy KIRP	0	6	0	0	0	0
	tumor KICH	0	0	17	0	0	0
	tumor KIRC	1	0	5	119	2	1
	tumor KIRP	0	0	1	3	68	0

Table 7.6: Early Integration using MLP model

7.4.4 Performances of late and early integration method on independent datasets

Besides, to test the late and early MLP models on samples from a different domain, the MLP model has been evaluated on 37 stomach samples and 817 lung cancer samples. This section explores the model performances on completely independent datasets compared to the one on which the models were trained. As it can be seen in Table 7.7, the late integration approach recognizes a higher percentage of samples as *Unknown* (43% and 29.7% for the stomach and lung datasets, respectively). On the contrary, the early integration method labels as *Unknown* the 2.7% and 5.3% of stomach and lung samples, respectively.

Dataset	Integration		
	Early	Late	Total samples
stomach	2.7%	43%	37
lung	5.3%	29.7%	817

Table 7.7: Percentage of samples predicted as *Unknown* in the stomach and lung datasets.

7.5 Discussion

As reported above, all the classifiers perform generally well, independently from the kind of integration technique (early or late). In fact, the accuracy and weighted average f1-score is quite always higher or equal to 96% (see Tables 7.2,7.5,7.6,7.7).

As shown in Table 7.5, the performance on the kidney test set of MLP and BNN methods is very similar. Therefore, MLP was selected as the proposed model for late integration since its training time is lower than BNN one, and the percentage of samples classified as *Unknown* is slightly lower. SVM and RF classifiers report marginally lower performances than MLP and BNN algorithms. They do not benefit from class-membership probability evaluation, and so they can not avoid misclassifying uncertain samples. The tree-MLP algorithm obtains a comparable accuracy, but it is less precise. The rationale behind the implementation of the tree-MLP model for comparison is the possibility to retrain one of its nodes separately. This aspect can be crucial in the biological domain since new molecular subtypes of the same tumor are continually redefined. In this case, the tree-MLP model can be updated on the new classes retraining only the involved nodes and not the entire classifier, avoiding spare time.

After selecting MLP architecture, the multi-omics late integration approach versus the early integration one has been explored. Regarding the kidney test set, the early integration approach reached 94% of accuracy and 78% of f1-score compared to 99% and 98%, respectively, for the late integration.

Although the late integration method is more accurate than the early integration one, it discards a higher percentage of samples from the classification by labeling them as *Unknown* (8.3% of all the samples versus less than 1%). About one-third of the kidney test set healthy samples are selected as *Unknown* from the late integration method, whereas they are all correctly classified with the early integration approach.

The late integration approach considers the classification of a few KICH samples not sufficiently confident, thus labeling them as *Unknown*, while the early integration labels as *Unknown* only one sample. As for the KIRC and KIRP tumor classes,

the main difference between early and late integration is the percentage of *Unknown* samples, which is higher for the late integration approach. However, the *Unknown* samples for KIRC and KIRP tumor classes identified by the late integration model are mainly misclassified in the early integration approach. Indeed, for these two classes, the late integration model's overall precision is higher compared to the early integration one.

Also, the late integration method exploits the class membership probabilities to classify the 43% of stomach and 29.7% of lung samples as *Unknown*. In contrast, early integration fails since it classifies as *Unknown* only the 2.7% of stomach and 5.3% of lung samples. This aspect is particularly crucial in avoiding the classification of samples that belong to different domains than the one used to train the model. It should be noticed that in the late integration model, a few samples from different tissues (i.e., stomach or lung) are still classified in one of the kidney classes. This event could be attributed to the fact that some basal biological processes (e.g., the ones related to the cell life cycle) are common between different tissues and cancers. In the end, since most of the stomach and lung samples are classified as uncertain samples, this method is suited especially for screening purposes.

However, it must be noticed that the type of integration (early or late) depends on the purpose of the classification. When the samples' provenance is known and belongs to the same domain of the training dataset, an early integration approach is to be preferred as the number of *Unknown* samples is generally low. On the contrary, the late integration approach is more suitable for performing a conservative classification since it can consider the uncertainty coming from even a single omic. However, as it classifies a percentage of samples as *Unknown*, the proposed late integration method can be used, for instance, for screening purposes and for detecting the most relevant correlated features.

7.6 Conclusions

In the multi-omics classification context, various implementations of early or late integration of the features are proposed. To date, it is not known how the class membership probabilities affect the classification determining a conservative approach or not. In this work, a late integration method based on an MLP architecture is presented. This approach allows for a more precise and conservative classification compared to the early integration technique. Also, it can consider the uncertainty signal coming from even a single omic and thus performing a more certain classification.

Besides, the late integration approach significantly outperforms the early integration approach in classifying samples coming from a tissue on which the model has not been trained. This aspect is particularly relevant in clinical practice since usually, it is preferable to receive an *Unknown* label instead of a wrong prediction. Moreover, the MLP architecture is particularly effective in applications with ever-evolving knowledge, such as genetic complex disease studies, preventing the classifier from being trained from scratch.

Chapter 8

The challenge of multi-omics data for clustering

8.1 Methodological contribution

In biology, it is expected within a specific pathology finding more molecular subtypes. Patients with the same clinical manifestation can have a different genetic heritage and response to drugs. Therefore, the treatment of the disease follows specific protocols and may affect the duration of survival. Although the identification of molecular profiles is possible sometimes by yielding a single omic data (for example, transcription), the joint integration of several omic data allows most times to carry out integrated and more detailed profiling.

From a computational point of view, integrating different data to get single clustering information is still a challenging problem regarding clustering in a multi-omic context.

To provide new molecular profiles and patients' categorization, class labels could be helpful. In this chapter, the main contribution consists of creating a model based on deep learning techniques by implementing an MLP with a specifically designed function. The loss represents the input samples in a reduced dimensional space by calculating the intra-cluster and inter-cluster distance at each epoch. This approach reported performances comparable to those of most referred methods in the literature, avoiding pre-processing steps for either feature selection or dimensionality reduction. Moreover, it has no limits on the number of omics to integrate.

8.2 Introduction

In recent years high throughput techniques have both driven down the costs and increased the speed in biological data acquisition[162]. Several types of "omic" data (such as genomics, epigenomic, and proteomic data) can be acquired for a single sample. Currently, standardized databases are being built as a cooperative effort to make data from different omics available for research, considerably speeding up progress in biology and medicine.
The availability and standardization of data are opening avenues to data-driven research, from statistical analysis to supervised and unsupervised machine learning.

Supervised learning is limited to the fields where it is possible to obtain accurate labels, like in survival studies or the prediction of other hard outcomes[40]. Conversely, unsupervised learning, especially clustering analysis, can lead to discovering new classes that may have biological relevance. For instance, clustering of RNA expression data can lead to the discovery of cancer subtypes. [80].
While studying single omics can provide valuable insight, for example, for the discovery or classification on cancer [81], single omics each carry partial information, and thus using multi-omic data integration is of fundamental importance in order to get more accurate analyses and predictions. However, the integration is not trivial and represents an open computational problem.

Attempts to solve it are merging all the features from different omics in a single feature space or performing a consensus clustering among the different input datasets through network-based techniques, joint dimensionality reduction techniques, or other types like Bayesian Consensus clustering. However, the current state of the art does not consistently perform better than single omic analysis on the best performing omic [174]. The development of new data fusion techniques is an open research problem. Here the proposed method to address it is an in-depth learning approach called Neural Graph Learning Fusion (NGL-F) that attempts to perform the fusion to reflect the topology of the input spaces.

One of this work's main contributions is to propose an original neural approach for modeling multi-omic datasets. Compared to the state-of-the-art algorithms, this approach exploits the manifold topology of the input space. This approach's main advantage is extending the algorithm to the case of omics having a different number of samples; this is not possible using the existing techniques, which are not tailored to the problem at the end.

8.3 Background

Given the greater availability of omic data, thanks to high throughput techniques, data-driven biology has dramatically expanded with the help of creating open databases and the development and improvement of algorithms.

A cooperative effort has led to large scale projects aiming to provide a unified basis for omic data collection and study. Examples are the Ensembl Genome project and the Human Proteome Project, providing a growing data set for the main eucaryotic genes and an attempt to create a map of the cell's protein-based molecular architecture. [97, 132]. Similarly, several public databases combine multiple information like omic data, clinical data, and histological images in the medical field, providing the foundation for data-driven medical research. Among such projects, the National Cancer Institute Genomic Data Commons (GDC) is a unified data-sharing platform for multiple cancer genomic projects. It provides standards for data collection to minimize inconsistencies due to the procedures used. More than 80'000 samples constitute a valuable resource for data-driven medical research [106]. Projects like those mentioned above have opened several avenues for computational studies, from statistical analysis to machine learning. The typical problems to be solved are classification and clustering. Clustering problems are of great interest because they allow new classes from data beyond human capability. For example, discovering new cancer subtypes plays an essential role in designing effective therapies that account for resistances. Clustering is an unsupervised learning approach to partitioning sample sets to maximize some similarity score among samples in the same subset and minimize it between different subset [103]. While different computational approaches have produced significant results even with single omics, [81], any omic taken by itself provides an incomplete picture. For example, greater gene expression values for protein-coding genes correlate with higher protein counts for the protein they code. However, there are regulatory mechanisms that inhibit the translation of mRNA into proteins. One such regulatory element is a small non-coding RNA molecule (miRNA). Thus combining mRNA and miRNA data should provide a better insight into the cell activity. Combining the information from multiple omics is crucial to discover patterns and generate insights at a system level. However, there are significant difficulties to be overcome.

Different approaches are available, focusing on multi-omic clustering. One distinction is between early integration and late integration algorithms: the former unites the features from different omics in a single matrix then performs the clustering; the latter performs clustering separately on the omics then merges the information. Early integration might reveal problems when the number of samples is much less than the number of features because it increases the dimensionality of the feature space significantly. Late integration is a complex theoretical and computational problem requiring discovering new and better algorithms to perform the fusion of the clusterings obtained from every omic individually. The difficulties in the use

of multi-omic data emerge when widely used techniques are benchmarked on real clinical cases and are shown not to perform consistently better than single omic data, especially if the comparison is with the best performing omic [174].
One relevant class of techniques is that of network-based techniques. An essential technique in this class is Similarity Network Fusion (SNF) [211], which starts from the similarity matrices of the original data and creates a consensus through an iterative algorithm. The matrices from individual omics are updated at each step, accounting for relevant contributions from the others. This approach has outperformed single-omic studies in some problems, such as identifying cancer subtypes and predicting survival rates when combining mRNA expression, DNA methylation, and miRNA expression. The method is fast and straightforward; however, it has limitations like requiring to have the same samples across all omics. Although the proposed NGL-F method has been trained on datasets containing the same samples, in principle, this is not a strict requirement. Another approach consists of applying dimensionality reduction techniques on the input space, accounting for the features of the different omics. This step is achieved through several algorithms to extend to multiple-input, datasets the techniques applied to a single matrix. This approach is called "Joint Dimensionality Reduction" (jDR). Finally, another attempt at multi-omic clustering comes from Bayesian methods, like Bayesian Consensus Clustering (BCC), which utilize a priori assumptions on the underlying distribution of the data to create a statistical model [37]. However, the latter has shown a lower accuracy when compared to dimensionality reduction techniques [41] and so is not considered in our analysis. Both the SNF method and a selection of dimensionality reduction techniques, combined with well-estabilished clustering algorithms, are used as a benchmark to test the algorithm's effectiveness described in this paper.

Neural networks offer an ample development space for the near future: the ability to fit complex data distributions make gradient-based methods very suitable for capturing the underlying topology of input data. If combined with competitive learning, by properly defining the output layer and the loss function, one obtains networks, which determine the position of cluster centroids through backpropagation and create a weighted graph establishing the connection strength among centroids. The strength of this type of method applied to multi-omic data is twofold: the presence of the weights allows to use of different cutoffs for building adjacency matrices for the data fusion and the design of a proper global loss function for networks processing the data from different omics can allow, through backpropagation, to take into account data from multiple omics in determining the weights for each clustering.

8.3.1 Joint Dimensionality Reduction for Data Fusion

In this section a description of some of the state-of-the-art dimensionality reduction tools, is provided. The goal of those approaches is to reduce high dimensonal omics into a low dimensional space. This is done by decomposing the matrices representing each of the L different omic matrices M_i with $i = 1, ..., L$, each of dimensions $n_i \times m$ (where m is the number of samples and n_i the number of features) into the product of a $k_i \times m$ factor matrix (F)and $n_i \times k_i$ omics-specific weight/projection matrices (A_i).
There are many methods based on different mathematical formulations. Here are the ones implemented for the comparison:

- Joint and Individual Variation Explained (JIVE) is an extension of PCA to multi-omic data. PCA seeks to describe the data with a reduced number of meta-features obtained by linear combination under the condition that the new meta-features are orthogonal and variance is maximized. JIVE decomposes each omic matrix into a joint factor matrix U, an omics-specific factor matrix A and residual noise E: $X_i = U_i S + A_i + E_i$ for $i = 1...L$. S is a score matrix explaining variability across multiple types of data. The objective function $||E||^2$ is minimized with E = $[E_1...E_L]^T$ [142].

- Regularized Generalized Canonical Correlation Analysis (GCCA) is a generalization of CCA, a method looking for linear combination of two matrices with the greatest correlation. GCCA determines a factorization of the same form as JIVE but maximizes the correlation between omic specific factors by finding projection vectors u^i such that the correlation between projected data is maximized: $\underset{i,j}{\text{argmax}}(Corr(X_i u_i, X_j, u_j))$ for all $i, j = 1...L$ [201].

8.4 The NGL-F neural network

The Neural Graph Learning for data Fusion (NGL-F) is a gradient-based competitive neural network [21], which uncovers topological sample-to-sample relationships using multiple data sources. Given two or more data types for the same set of samples (e.g., patients), NGL-F learns the mutual relationships among samples taking into account such heterogeneous information simultaneously. The output of NGL-F is a set of graphs. For each data set, NGL-F aims to find a graph where nodes represent cluster centroids while edges represent cluster topological properties. The learned topology described by such graphs is used to create the sample adjacency matrix (S). The information contained in the matrix represents all datasets, and it can be used to uncover latent patterns among samples. In this sense, the sample adjacency matrix is used to build a unique graph (sample graph) in which nodes represents samples, and the edges are derived from S.

NGL-F is composed of a set of dual multi-layer perceptrons (MLPs) [21], one for each dataset, equipped with a final competitive layer. Weights are estimated by backpropagation. The activation functions are ReLU for the hidden layers and linear for the competitive output units. The input of each network is a data set represented as a matrix $X_z \in \mathbb{R}^{d,n}$, where n is the number of samples and d the number of features. Each MLP provides as output a set of vectors $w_i \in \mathbb{R}^d$ representing cluster centroids for the input data. For each data source taken into consideration, a multi-layer neural network is instantiated. The architecture of each network can be customized according to the complexity of its data set (see Fig. 8.1).

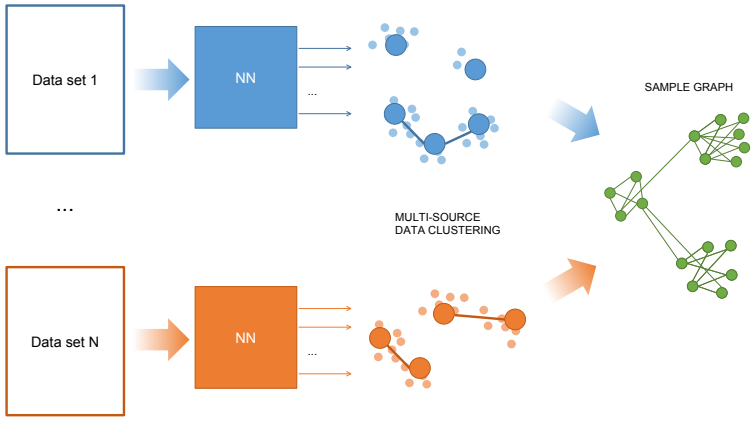


Figure 8.1: NGL-F architecture: N datasets are fed in input to NGL-F. For each dataset, a multi-layer perceptron employed and customized according to dataset complexity. Clustering outputs are at the end combined in order to create a sample graph built from the adjacency matrix S.

The loss function of NGL-F takes into account, at the same time, the quality of clusters found by each MLP and their underlying topology. The relationships among clusters are modeled using an adjacency matrix E, where $E(i, j)$ represents the number of samples for which w_i and w_j are the two closest centroids. The higher $E(i, j)$, the more their respective clusters are related. Metric E represents a graph on the neural network, where the nodes are the neurons, and the edges are inter-neuron connections. These links represent the topology of the input data.

The loss function of each MLP is composed of four terms taking into account inter- and intra-cluster distances, quantization error, and parsimony in representing the underlying topology:

$$\mathcal{L}_z = \frac{\max_k d_{intra}(C_k)}{\max_{i,j} d_{inter}(C_i, C_j)} + Q + ||E|| \tag{8.1}$$

where $d_{intra}(C_k)$ is the intra-cluster distance, $d_{inter}(C_i, C_j)$ the inter-cluster distance, and Q the quantization error.

The complete diameter distance is used as an intra-cluster quality index, representing the distance between two most remote samples belonging to the same cluster:

$$d_{intra}(C_i) = \max_{x,y \in C_i} d(x, y) \tag{8.2}$$

The single linkage distance, representing the closest distance between two samples belonging to two different clusters, is used to model inter-cluster distance:

$$d_{inter}(C_i, C_j) = \min_{x \in C_i, y \in C_j} d(x, y) \tag{8.3}$$

The quantization error is computed as the norm of the distances between cluster centroids (w_i) and cluster points (C_i):

$$Q = ||d(w_i, x)||_2 \qquad \forall x \in C_i \tag{8.4}$$

The NGL-F loss function is the linear combination of MLPs' losses:

$$\mathcal{L} = \sum_z \mathcal{L}_z \tag{8.5}$$

Once all networks terminate the training procedure, the resulting clusters are analyzed. For each data set, two samples are considered near each other if they belong to the same cluster, far from each other, in case they belong to different clusters. A sample adjacency matrix S is then computed as follow:

$$S(i, j) = \sum_{d=1}^{n} near_d(i, j), \tag{8.6}$$

where $near_d(i, j)$ is a boolean function calculating the samples' proximity as previously explained, and n is the number of data set taken into consideration. This matrix is the result of the fusion process. Its quality can be analyzed and compared to other methods in different ways, as shown in the next section.

8.5 Experiments

Data have been downloaded from the NIH Genomic Data Commons [161] and have been collected in tabular form, resulting in an mRNA and a miRNA transcriptome profiling matrix.
The mRNA matrix consists of raw counts gene expression values [12]. For protein-coding genes, a higher value represents a more significant amount of protein produced. This statement is true unless regulatory mechanisms inhibit the translation

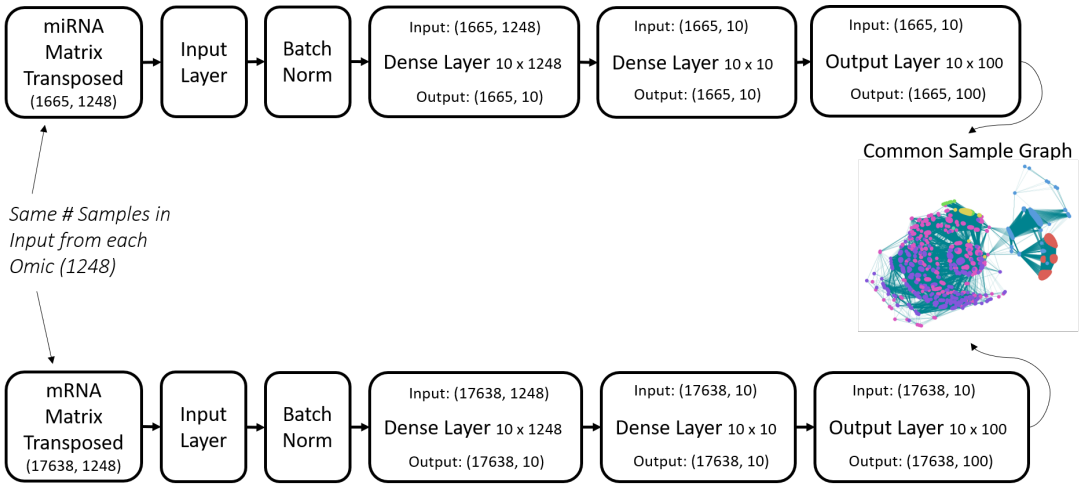


Figure 8.2: NGL-F network architecture as used in the experiments. Between brackets, the dimensionality of input/output data of each layer is reported. Regarding the matrices, the dimensions are defined as features x samples since the matrix is transposed. Instead, each dense and output layers is reported the dimensionality of the associated weight matrix. It should also be noticed the different dimensionality of the two input sources, miRNA (top) and mRNA (bottom) maintained through the layers.

of the mRNA.
The miRNA matrix consists of raw counts miRNA values [47]. In this case, a higher expression value corresponds to a lower presence of the proteins related to that sequence because miRNA inhibits mRNA translation.

The data were preprocessed as follows:

- For the mRNA matrix, the genes with a zero expression value across all the samples were deleted; then, the normalization was performed through a variance stabilizing transformation [99], and only protein-coding genes were selected. This approach resulted in 17682 genes for which the expression value is reported.
- For the miRNA matrix, the sequences with zero expression value across the samples were deleted, and the matrix was normalized through DESeq2 [143]. The final values were obtained as $log_2(exprValue + 1)$ [11].

The patients for which either the mRNA or the miRNA data were missing were deleted from the matrices. This approach resulted in 1248 miRNA and mRNA sequences for which the expression value is reported. This deletion is not a strict requirement for NGL-F, but it is necessary to compare it with SNF on the same dataset.

Data samples come from either healthy or cancerous lung tissue belonging to two types: Lung Adenocarcinoma (LUAD) or Lung Squamous cells Carcinoma (LUSC). The healthy tissue has been taken from non-tumoral tissue samples, usually close to the tumor's position.
Data were acquired from three projects: TCGA-LUAD [203] and CPTAC-3, with samples from adenocarcinoma patients, and TCGA-LUSC, with samples from squamous cells carcinoma patients. Overall this resulted in six different annotations, all reported as the project's name followed by either the "tumoral" or "healthy" annotation.

All the code for the experiments has been implemented in Python 3, relying upon open-source libraries [1, 86]. All the experiments have been run on the same machine: Intel® Core™ i7-8750H 6-Core Processor at 2.20 GHz equipped with 8 GB RAM.

The two datasets previously described have been fed as input to the NGL-F algorithm. The structure of the networks employed in this paper is reported in Fig. 8.2. NGL-F is a single neural network that employs a set of dual multi-layer perceptrons, one for each analyzed omic. The use of dual networks is justified, given the high-dimensionality of the data sources [8, 21]. The number of features may vary between different omic, and it is maintained through the layers, as dual networks are trained on the transposed matrix [21]. In this way, output nodes preserve input dimensionality and can be used as cluster centroids for each input matrix. In this implementation, the only requirement is on the number of samples (1248) that need to be identical among the omics. As mentioned in Sec. 8.4, the fusion process consists of creating a unique sample adjacency matrix that considers the information extracted from every omic data. To compare the results of the proposed method, the experiment was repeated using the SNF algorithm. As explained in Sec. 8.3.1, NGL-F is also compared with JIVE, tICA, and GCCA joint dimensionality reduction techniques; however, since they do not yield an adjacency matrix, only the clustering quality is considered as relevant for measuring their performance.

8.5.1 Adjacency Matrix Based Comparison

The adjacency matrices built by NLG-F and SNF methods are depicted in Fig. 8.3. Observing the two plots shows that the results are similar to both methods capable of identifying similarities among data. This approach is a significant result as it shows the quality of the fusion process carried out by the proposed method compared to a state-of-the-art algorithm.

It was decided to plot the sample adjacency matrix through the Kamada-Kawai path-length algorithm [113] to analyze this result better. This algorithm is a force-directed graph drawing method that can visualize undirected graphs in a two-dimensional space. The main characteristics of this class of algorithms are that

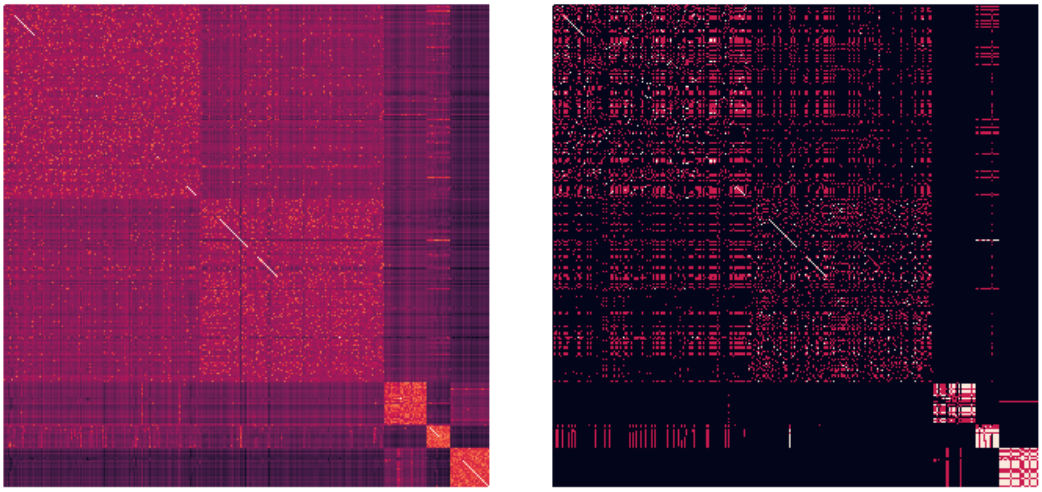


Figure 8.3: Adjacency matrix of the sample using (left) SNF and (right) NGL-F algorithms

edges are displayed so that the number of crossings is the lowest possible. In the two plots of Fig. 8.4, it is clear that the number of connections found by SNF is redundant: even isolated samples as the LUAD tumoral ones on the top and left edges are connected with many other samples. Conversely, NGL-F better identifies outliers as seen with the tumoral CPTAC3 on the top right corner. However, the sample adjacency matrix plot produced by SNF, better separates LUAD from LUSC tumoral data, while in the plot concerning NGL-F the samples belonging to the two classes are quite confused.

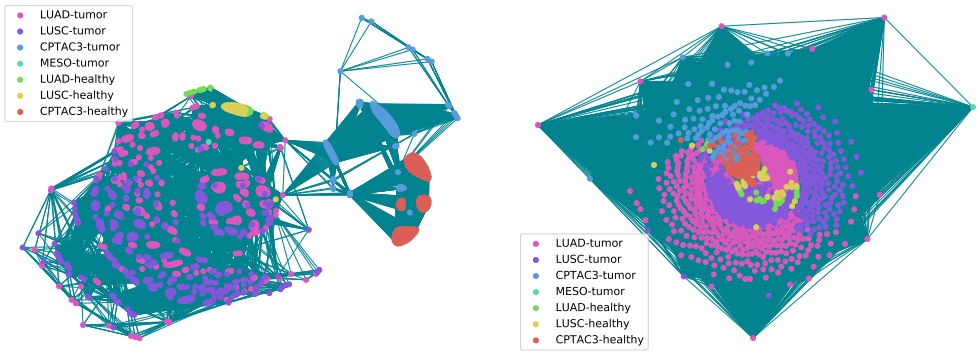
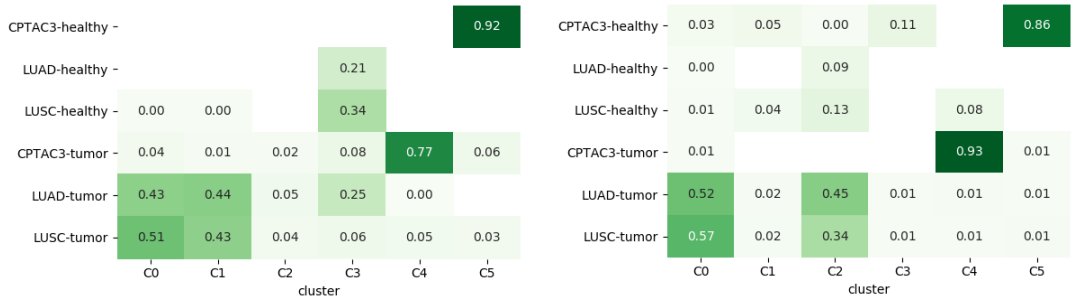


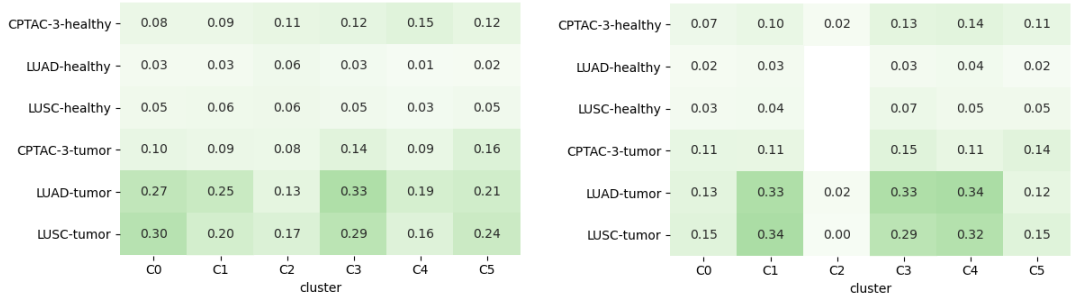
Figure 8.4: Kamada-Kawai path-length graph of the sample adjacency matrix computed by NGL-F (left) and SNF (right) algorithms.

Finally, the proposed algorithm's quality is validated through a comparison of the spectral clustering executed on the two adjacency matrices. Figs. 8.5a and 8.5b show the clustering quality for NGL-F and SNF, respectively; the harmonic mean of purity and efficiency of the clusters is computed according to the class of the samples belonging to each cluster. Both clustering techniques can precisely identify CPTAC3 healthy samples grouped in the C5 cluster. Also, CPTAC3 tumor samples are mostly collected in a single cluster, C4; however, SNF's adjacency matrix seems to separate these samples better as the corresponding cluster quality is higher than the previous. Instead, samples belonging to LUAD and LUSC (both tumor and healthy) seem to be more challenging to identify. For both tissues, tumor samples are collected together in the C0 and C2 clusters for SNF and C0 and C1 clusters for NGL-F. At last, the few LUAD and LUSC healthy samples are mostly placed in the C3 cluster for NGL-F, while they are split among all the clusters in the case of SNF.



(a) NGL-F

(b) SNF



(c) JIVE

(d) GCCA

Figure 8.5: Harmonic mean of cluster efficiency and purity computed on the spectral clusters, computed on the adjacency matrix produced by the different algorithms.

8.5.2 jDR Based Method Comparison

The jDR algorithms accomplish the data fusion through projecting datasets into a unique lower-dimensional space (see Sec. 8.3.1); in this sense, they yield a

single reduced-matrix that takes into consideration the information coming from all omics. It can be argued that the strategy employed by such methods for discovering the information underlying data is quite different w.r.t. both SNF and NGL-F; indeed, they do not output an adjacency matrix natively and, to assess their spectral cluster performances, it should be derived indirectly from the network graph.
JIVE (Fig. 8.5c) and GCCA (Fig. 8.5d) behave in a very similar way w.r.t. all the input categories. Their performances are better on LUAD and LUSC tumors but not sufficient to correctly discriminate the input data; they are both unable to discover any underlying pattern on the remaining labels. Fig. 8.5a shows NGL-F outperforms the jDR methods for all the input clusters. The CPTAC3 category, both in the healthy (0.92) and tumor (0.77) cases, is very well recognized; furthermore, for LUAD and LUSC tumors, NGL-F reaches a good values algorithm of understanding some pattern about these pathologies. Instead, the LUSC-healthy condition is much better recognized than in any of the other methods but not sufficiently enough to assess the network has learned it. Finally, it is interesting to notice that none of the proposed techniques can deal with the LUAD-healthy category, and it will be investigated in future work.

8.5.3 Final Considerations

Summing up, the results produced by SNF and NGL-F algorithms are very similar. It is noticeable to point out the importance of this result, as NGL-F is an entirely new algorithm based on a recent neural theory [21]. Compared to state-of-the-art methods such as SNF, the neural network structure of NGL-F shows higher flexibility and can be easily extended to omics with a different number of samples. Furthermore, both techniques outperformed all jDR algorithms in the clustering performance comparison. As shown in Fig. 8.5, no factorization matrices produced by these algorithms allow the spectral clustering to extract meaningful biological patterns.

8.6 Conclusions

Since data interpretation from multiple data sources is still an open and challenging problem, some multi-omic approaches have been recently proposed. However, these methods do not take into account the intrinsic topology of each omic. Therefore, NGL-F has been designed to tackle this issue. It is an unsupervised deep learning neural network endowed with an original final layer that is competitive because of the loss function's choice. Indeed, it considers both the quantization and the clustering, and the onset of the edges. The training procedure is repeated for all input datasets generating for each one a network of centroids to which the samples are assigned competitively, with criteria for creating and decaying connections

between the centroids (prototypes) themselves. The outcome is a connected graph for each input, which is merged to obtain the final graph from which the clusters are derived. Experimental results show its competitiveness with state-of-the-art algorithms. However, NGL-F is more flexible as it allows working with omics having a different number of samples.
Since the proposed method works in the original feature space, omics' relevance can be retrieved from the model and directly investigated for further biological considerations. Hence, the proposed algorithm is suitable for a broader range of applications.

Further developments of the proposed approach will deal with implementing convolutional layers into the neural architecture and with a more in-depth analysis of the loss function, taking into account cluster densities [65]. Besides, the development of an incremental, hierarchical [48], and biclustering versions of NGL-F will also be studied.

Chapter 9

Conclusions

Due to the continuous increase in the number and complexity of the genetic and biological data, new computer science techniques are needed to analyze these data and provide valuable insights into the main features. The thesis research topic consists of studying complex systems in life sciences to offer informative models about biological processes. The thesis focuses on two main sub-topics.

The first sub-topic concerns machine and deep learning techniques applied to the analysis of aberrant genetic sequences like, for instance, gene fusions. The second one is the development of statistic and deep learning techniques for heterogeneous biological and clinical data integration.

In aberrant genetic sequence analysis and gene fusions prioritization, machine and deep learning models have been explored, leading to two main contributions. The first contribution consists of designing a deep learning model to recognize oncogenic gene fusions using the resulting proteins' amino acid sequence exclusively. This model is based on a CNN followed by a bidirectional LSTM that outperformed the art tools' state.

The second contribution does not exploit the gene fusion sequence, but it included post-transcriptional regulators like transcription factors and mainly miRNA. This approach is based on an MLP architecture and led to an improved model that outperforms the previous ones, and it competes with state-of-the-art tools.

The rationale behind the thesis's second sub-topic is the following: due to the widespread of Next Generation Sequencing (NGS) technologies, a large amount of heterogeneous complex data related to several diseases and healthy individuals is now available. In this context, the aim is to integrate multi-omics data involving thousands of features (e.g. genes, micro-RNA) and identifying which of them are

relevant for a specific biological process. From a computational point of view, finding the best strategies for multi-omics analysis and relevant features identification is a very open challenge. Four main aspects have been considered: analysis of the connectivity in the mouse brain, miRNA target identification, multi-omics sample classification, and multi-omics sample clustering.

The first aspect is the integrative analysis of gene expression and connectivity data of mouse brains exploiting machine learning techniques. The neuronal connection data (obtained by viral tracers) of mouse brains were processed to identify brain regions physically connected and then evaluated with these areas' gene expression data. A multi-layer perceptron was applied to perform the classification task between connected and unconnected regions providing gene expression data as input. Furthermore, a second model was created to infer the degree of connection between distinct brain regions. The implemented models successfully executed the binary classification task (connected regions against unconnected areas) and distinguished the relationship's intensity in low, medium, and high.

The second aspect focuses on a statistical method to reveal pathology-determining microRNA targets in multi-omic dataset. In this work, two multi-omics datasets are used: breast cancer and medulloblastoma datasets. Both the datasets are composed of miRNA, mRNA, and proteomics data related to the same patients. The main computational contribution to the field consists of designing and implementing an algorithm based on the statistical conditional probability to infer the impact of miRNA post-transcriptional regulation on target genes exploiting the protein expression values. The developed methodology allowed a more in-depth understanding and identification of target genes. Also, it proved to be significantly enriched in three well-known databases (miRDB, TargetScan, and miRTarBase), leading to relevant biological insights.

The third aspect deals with the classification of multi-omics samples. The literature's main approaches integrate all the features available for each sample upstream of the classifier (early integration approach) or create separate classifiers for each omic and subsequently define a consensus set rules (late integration approach). In this context, the main contribution consists of introducing the probability concept by creating a model based on Bayesian and MLP networks to achieve a consensus guided by the class label and its probability. This approach has shown how a probabilistic late integration classification is more specific than an early integration approach. Also, the proposed model can better identify anomalous samples concerning the training domain. This tool is potent, as, in addition to recognizing outliers belonging to the same tissue used in training, it excludes from the classification samples that belong to a different tissue or tumor subtype than that with which the model was trained. This aspect represents a significant advantage in the clinics.

To provide new molecular profiles and patients' categorization, class labels could be helpful. However, they are not always available. Therefore, the need to cluster samples based on their intrinsic characteristics is revealed and dealt with in a specific chapter. Multi-omic clustering in literature is mainly addressed by creating graphs or methods based on multidimensional data reduction. This field's main contribution is creating a model based on deep learning techniques by implementing an MLP with a specifically designed loss function. The loss represents the input samples in a reduced dimensional space by calculating the intra-cluster and inter-cluster distance at each epoch. This approach reported performances comparable to those of most referred methods in the literature, avoiding pre-processing steps for either feature selection or dimensionality reduction. Moreover, it has no limits on the number of omics to integrate.

9.1 Global considerations

This thesis presents statistical and deep learning models to address various biological problems, from aberrant sequence analysis to sample classification and clustering.

All the sections present a significant contribution to the bioinformatics community. Indeed, the proposed methods aim to allow physicians, biologists, and researchers to identify new unforeseen aspects. These methods allow the selection of the driver elements of a disease in a more specific way. Therefore, they can be extended in the clinical context in the future.

This thesis's main novelties consist of models closer to the complexity of the investigated data. Indeed, although the biological questions addressed in this thesis are not entirely new, the use of new tools and more complex models (able to handle heterogeneous and multi-omic data) has allowed us to identify the relevant signal with greater precision.

In the end, the study of the biological phenomena addressed in this thesis is not yet exhausting. They can still be investigated, taking advantage of new computational models representing their complexity more realistically and precisely.

Appendix A

List of the published works

This Ph.D. thesis aims to present the main research activities and personal contributions to the scientific community in recent years.

These works have been disseminated through the following publications:

- Lovino, M., Ciaburri, M. S., Urgese, G., Di Cataldo, S., & Ficarra, E. (2020). DEEPrior: a deep learning tool for the prioritization of gene fusions. *Bioinformatics*, 36(10), 3248-3250.
- Lovino, M., Urgese, G., Macii, E., Di Cataldo, S., & Ficarra, E. (2019). A deep learning approach to the screening of oncogenic gene fusions in humans. *International journal of molecular sciences*, 20(7), 1645.
- Roberti, I., Lovino, M., Di Cataldo, S., Ficarra, E., & Urgese, G. (2019). Exploiting Gene Expression Profiles for the Automated Prediction of Connectivity between Brain Regions. *International journal of molecular sciences*, 20(8), 2035.
- Lovino, M., Bontempo, G., Cirrincione, G., & Ficarra, E. (2020, October). Multi-omics Classification on Kidney Samples Exploiting Uncertainty-Aware Models. In *International Conference on Intelligent Computing* (pp. 32-42). Springer, Cham.
- Barbiero, P., Lovino, M., Siviero, M., Ciravegna, G., Randazzo, V., Ficarra, E., & Cirrincione, G. (2020, October). Unsupervised Multi-omic Data Fusion: The Neural Graph Learning Network. In *International Conference on Intelligent Computing* (pp. 172-182). Springer, Cham.
- Lovino, M., Urgese, G., Macii, E., Di Cataldo, S., & Ficarra, E. (2018, September). Predicting the Oncogenic Potential of Gene Fusions Using Convolutional Neural Networks. In *International Meeting on Computational Intelligence Methods for Bioinformatics and Biostatistics* (pp. 277-284). Springer,

Cham.

These works have recently been submitted and are still under review:

- Lovino, M., Randazzo, V., Ciravegna, G., Barbiero, P., Ficarra, E., & Cirrincione, G. (2021) A survey on data integration for multi-omicssample clustering. Submitted at *Neurocomputing*.
- Lovino, M., Bontempo, G., Cirrincione, G., & Ficarra, E. (2021). An uncertainty-aware late integration method formulti-omics sample classification. Submitted at *BMC Bioinformatics*.
- Barrese V. S., Montemurro M., Lovino, M. & Ficarra, E. (2021). Identifying the oncogenic potential of genefusions exploiting miRNAs. Submitted at *BMC Bioinformatics*.

Bibliography

[1] Martin Abadi et al. "Tensorflow: A system for large-scale machine learning". In: *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*. 2016, pp. 265–283.

[2] Francesco Abate et al. "Bellerophontes: an RNA-Seq data analysis framework for chimeric transcripts discovery based on accurate fusion model". In: *Bioinformatics* 28.16 (2012), pp. 2114–2121.

[3] Francesco Abate et al. "Pegasus: a comprehensive annotation and prediction tool for detection of driver gene fusions in cancer". In: *BMC systems biology* 8.1 (2014), pp. 1–14.

[4] Nitin Agarwal, Xiangmin Xu, and Meenakshisundaram Gopi. "Geometry processing of conventionally produced mouse brain slice images". In: *Journal of neuroscience methods* 306 (2018), pp. 45–56.

[5] Vikram Agarwal et al. "Predicting effective microRNA target sites in mammalian mRNAs". In: *eLife* 4 (Aug. 2015). Ed. by Elisa Izaurralde, e05005. ISSN: 2050-084X. DOI: 10.7554/eLife.05005. URL: https://doi.org/10.7554/eLife.05005.

[6] Babak Alipanahi et al. "Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning". In: *Nature biotechnology* 33.8 (2015), pp. 831–838.

[7] Atlas Allen-Brain-Institute. *Allen Mouse Brain Atlas*. http://mouse.brain-map.org/static/brainexplorer/. 2018. (Visited on 09/30/2018).

[8] Naomi Altman and Martin Krzywinski. "The curse (s) of dimensionality". In: *Nat Methods* 15.6 (2018), pp. 399–400.

[9] Naomi Altman and Martin Krzywinski. "The curse(s) of dimensionality". en. In: *Nature Methods* 15.6 (June 2018), pp. 397–397. ISSN: 1548-7091, 1548-7105. DOI: 10.1038/s41592-018-0013-3. URL: http://www.nature.com/articles/s41592-018-0013-3 (visited on 12/04/2020).

[10] Rudolf Amann and Bernhard M Fuchs. "Single-cell identification in microbial communities by improved fluorescence in situ hybridization techniques". In: *Nature Reviews Microbiology* 6.5 (2008), p. 339.

[11] Simon Anders and Wolfgang Huber. "Differential expression of RNA-Seq data at the gene level–the DESeq package". In: *Heidelberg, Germany: European Molecular Biology Laboratory (EMBL)* 10 (2012), f1000research.

[12] Simon Anders, Paul Theodor Pyl, and Wolfgang Huber. "HTSeq—a Python framework to work with high-throughput sequencing data". In: *Bioinformatics* 31.2 (2015), pp. 166–169.

[13] Kevin M Anderson et al. "Gene expression links functional networks across cortex and striatum". In: *Nature communications* 9.1 (2018), p. 1428.

[14] Marios Anthimopoulos et al. "Lung pattern classification for interstitial lung diseases using a deep convolutional neural network". In: *IEEE transactions on medical imaging* 35.5 (2016), pp. 1207–1216.

[15] Laura Antonelli et al. "Integrating imaging and omics data: A review". In: *Biomedical Signal Processing and Control* 52 (2019), pp. 264–280.

[16] Ricard Argelaguet et al. "Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets". In: *Molecular Systems Biology* 14.6 (June 2018). Publisher: John Wiley & Sons, Ltd, e8124.

[17] Yan W Asmann et al. "Detection of redundant fusion transcripts as biomarkers or disease-specific therapeutic targets in breast cancer". In: *Cancer research* 72.8 (2012), pp. 1921–1928.

[18] Yann Audic and Rebecca S Hartley. "Post-transcriptional regulation in cancer". In: *Biology of the Cell* 96.7 (2004), pp. 479–498.

[19] Mihaela Babiceanu et al. "Recurrent chimeric fusion RNAs in non-cancer tissues and cells". In: *Nucleic acids research* 44.6 (2016), pp. 2859–2872.

[20] Albert-László Barabási, Natali Gulbahce, and Joseph Loscalzo. "Network medicine: a network-based approach to human disease". In: *Nature Reviews Genetics* 12.1 (Jan. 2011). Number: 1 Publisher: Nature Publishing Group, pp. 56–68.

[21] Pietro Barbiero et al. *Topological Gradient-based Competitive Learning*. 2020.

[22] Pietro Barbiero et al. "Unsupervised Multi-omic Data Fusion: The Neural Graph Learning Network". In: *International Conference on Intelligent Computing*. Springer. 2020, pp. 172–182.

[23] Cornelia I Bargmann and Eve Marder. "From the connectome to brain function". In: *Nature methods* 10.6 (2013), p. 483.

[24] Andrew Bate et al. "A Bayesian neural network method for adverse drug reaction signal generation". In: *European journal of clinical pharmacology* 54.4 (1998), pp. 315–321.

[25] Karla Batista-Garcia-Ramo and Caridad Fernandez-Verdecia. "What We Know About the Brain Structure–Function Relationship". In: *Behavioral Sciences* 8.4 (2018), p. 39.

[26] Marco Beccuti et al. "The structure of state-of-art gene fusion-finder algorithms". In: *Genome Bioinformatics* 1.2 (2013).

[27] Matteo Benelli et al. "Discovering chimeric transcripts in paired-end RNA-seq data by using EricScript". In: *Bioinformatics* 28.24 (2012), pp. 3232–3239.

[28] Matteo Benelli et al. "Discovering chimeric transcripts in paired-end RNA-seq data by using EricScript". In: *Bioinformatics* 28.24 (2012), pp. 3232–3239.

[29] Bonnie Berger, Jian Peng, and Mona Singh. "Computational solutions for omics data". In: *Nature reviews genetics* 14.5 (2013), pp. 333–346.

[30] Matteo Bersanelli et al. "Methods for the integration of multi-omics data: mathematical aspects". en. In: *BMC Bioinformatics* 17.2 (Jan. 2016), S15. ISSN: 1471-2105. DOI: 10.1186/s12859-015-0857-9. URL: https://doi.org/10.1186/s12859-015-0857-9 (visited on 11/16/2020).

[31] Michael R Berthold et al. "KNIME-the Konstanz information miner: version 2.0 and beyond". In: *AcM SIGKDD explorations Newsletter* 11.1 (2009), pp. 26–31.

[32] Eli Bingham et al. "Pyro: Deep Universal Probabilistic Programming". In: *arXiv:1810.09538 [cs, stat]* (Oct. 2018). arXiv: 1810.09538.

[33] Christopher M. Bishop. *Pattern recognition and machine learning*. 2006.

[34] Mihail Bota and Larry W Swanson. "BAMS neuroanatomical ontology: design and implementation". In: *Frontiers in neuroinformatics* 2 (2008), p. 2.

[35] Leo Breiman. "Random forests". In: *Machine learning* 45.1 (2001), pp. 5–32.

[36] Cancer Genome Atlas Research Network, John N. Weinstein, and Gordon B. Mills. "The Cancer Genome Atlas Pan-Cancer analysis project". In: *Nature Genetics* 45.10 (Oct. 2013), pp. 1113–1120.

[37] Laura Cantini et al. "Benchmarking joint multi-omics dimensionality reduction approaches for cancer study". In: *bioRxiv* (2020).

[38] Laura Cantini et al. "Detection of gene communities in multi-networks reveals cancer drivers". In: *Scientific Reports* 5.1 (Dec. 2015). Number: 1 Publisher: Nature Publishing Group.

[39] Laura Cantini et al. "MMRA MicroRNA master regulator analysis". In: (2015).

[40] Kumardeep Chaudhary et al. “Deep learning–based multi-omics integration robustly predicts survival in liver cancer”. In: *Clinical Cancer Research* 24.6 (2018), pp. 1248–1259.

[41] Cecile Chauvel et al. “Evaluation of integrative clustering methods for the analysis of multi-omics data”. In: *Briefings in Bioinformatics* 21.2 (2020), pp. 541–552.

[42] Nitesh V Chawla et al. “SMOTE: synthetic minority over-sampling technique”. In: *Journal of artificial intelligence research* 16 (2002), pp. 321–357.

[43] Yuhao Chen and Xiaowei Wang. “miRDB: an online database for prediction of functional microRNA targets”. In: *Nucleic acids research* 48.D1 (2020), pp. D127–D131.

[44] François Chollet et al. *Keras*. 2015.

[45] Chih-Hung Chou et al. “miRTarBase 2016: updates to the experimentally validated miRNA-target interactions database”. In: *Nucleic acids research* 44.D1 (2016), pp. D239–D247.

[46] Chih-Hung Chou et al. “miRTarBase update 2018: a resource for experimentally validated microRNA-target interactions”. In: *Nucleic acids research* 46.D1 (2018), pp. D296–D302.

[47] Andy Chu et al. “Large-scale profiling of microRNAs for the cancer genome atlas”. In: *Nucleic acids research* 44.1 (2016), e3–e3.

[48] Giansalvo Cirrincione et al. “The GH-EXIN neural network for hierarchical clustering”. In: *Neural Networks* 121 (2020), pp. 57–73.

[49] Ronan Collobert and Jason Weston. “A unified architecture for natural language processing: Deep neural networks with multitask learning”. In: *Proceedings of the 25th international conference on Machine learning*. ACM. 2008, pp. 160–167.

[50] ENCODE Project Consortium. “The ENCODE (ENCyclopedia Of DNA Elements) Project”. In: *Science* 306.5696 (Oct. 2004), pp. 636–40. DOI: 10.1126/science.1105136.

[51] Francis Crick. “Central dogma of molecular biology”. In: *Nature* 227.5258 (1970), pp. 561–563.

[52] Nello Cristianini, John Shawe-Taylor, et al. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press, 2000.

[53] Nadia M Davidson, Ian J Majewski, and Alicia Oshlack. “JAFFA: High sensitivity transcriptome-focused fusion gene detection”. In: *Genome medicine* 7.1 (2015), p. 43.

[54] Jose D Debes and Donald J Tindall. "Mechanisms of androgen-refractory prostate cancer". In: *New England Journal of Medicine* 351.15 (2004), pp. 1488–1490.

[55] Santa Di Cataldo and Elisa Ficarra. "Mining textural knowledge in biological images: Applications, methods and trends". In: *Computational and structural biotechnology journal* 15 (2017), pp. 56–67.

[56] Alexander Dobin et al. "STAR: ultrafast universal RNA-seq aligner". In: *Bioinformatics* 29.1 (2013), pp. 15–21.

[57] Timothy Dozat. "Incorporating nesterov momentum into adam". In: *Proceedings of the 4th International Conference on Learning Representations* (2016).

[58] Stuart Dreyfus. "The numerical solution of variational problems". In: *Journal of Mathematical Analysis and Applications* 5.1 (1962), pp. 30–45.

[59] Brian J Druker. "Imatinib as a paradigm of targeted therapies". In: *Advances in cancer research* 91.1 (2004), pp. 1–30.

[60] Ali Ebrahim et al. "Multi-omic data integration enables discovery of hidden biological regularities". In: *Nature communications* 7.1 (2016), pp. 1–9.

[61] Henrik Edgren et al. "Identification of fusion genes in breast cancer by paired-end RNA-sequencing". In: *Genome biology* 12.1 (2011), R6.

[62] Henrik Edgren et al. "Inga Rye, Sandra Nyberg, Maija Wolf, Anne Lise Borresen Dale et Olli Kallioniemi: Identification of fusion genes in breast cancer by paired-end RNA-sequencing". In: *Genome Biology* 12.1 (2011), R6.

[63] Ensembl. *The Illumina Body Map 2.0 Project*. https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-513/. Accessed: 2019-06-20. 2010.

[64] Heyer Erin E. et al. "Diagnosis of fusion genes using targeted RNA sequencing". In: *Nature Communications* (2019). DOI: https://doi.org/10.1038/s41467-019-09374-9.

[65] Martin Ester et al. "A density-based algorithm for discovering clusters in large spatial databases with noise." In: *Kdd*. Vol. 96. 34. 1996, pp. 226–231.

[66] Ahmed Fakhry and Shuiwang Ji. "High-resolution prediction of mouse brain connectivity using gene expression patterns". In: *Methods* 73 (2015), pp. 71–78.

[67] Ahmed Fakhry et al. "Global analysis of gene expression and projection target correlations in the mouse brain". In: *Brain informatics* 2.2 (2015), pp. 107–117.

[68] Mitelman Felix, Johansson Bertil, and Mertens Fredrik. "The impact of translocations and gene fusions on cancer causation". In: *Nature Reviews Cancer* 7 (2007), pp. 233–245.

[69] Witold Filipowicz, Suvendra N Bhattacharyya, and Nahum Sonenberg. "Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight?" In: *Nature reviews genetics* 9.2 (2008), pp. 102–114.

[70] Simon A Forbes et al. "COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer". In: *Nucleic acids research* 39.suppl_1 (2010), pp. D945–D950.

[71] Antoine Forget et al. "Aberrant ERBB4-SRC signaling as a hallmark of group 4 medulloblastoma revealed by integrative phosphoproteomic profiling". In: *Cancer cell* 34.3 (2018), pp. 379–395.

[72] Leon French and Paul Pavlidis. "Relationships between gene expression and brain wiring in the adult rodent brain". In: *PLoS computational biology* 7.1 (2011), e1001049.

[73] Milana Frenkel-Morgenstern et al. "Chimeras taking shape: potential functions of proteins encoded by chimeric RNA transcripts". In: *Genome research* 22.7 (2012), pp. 1231–1242.

[74] Milana Frenkel-Morgenstern et al. "Chimeras taking shape: potential functions of proteins encoded by chimeric RNA transcripts". In: *Genome research* 22.7 (2012), pp. 1231–1242.

[75] Robin C Friedman et al. "Most mammalian mRNAs are conserved targets of microRNAs". In: *Genome research* 19.1 (2009), pp. 92–105.

[76] Mathias Fuchs et al. "Connecting high-dimensional mRNA and miRNA expression data for binary medical classification problems". In: *Computer Methods and Programs in Biomedicine* 111.3 (Sept. 2013), pp. 592–601.

[77] Ben D Fulcher and Alex Fornito. "A transcriptional signature of hub connectivity in the mouse connectome". In: *Proceedings of the National Academy of Sciences* 113.5 (2016), pp. 1435–1440.

[78] Florian Ganglberger et al. "Predicting functional neuroanatomical maps from fusing brain networks with genetic information". In: *NeuroImage* 170 (2018), pp. 113–120.

[79] Qingsong Gao et al. "Driver fusions and their implications in the development and treatment of human cancers". In: *Cell reports* 23.1 (2018), pp. 227–238.

[80] Shaowei Gao et al. "Unsupervised clustering reveals new prostate cancer subtypes". In: *Translational Cancer Research* 6.3 (2017), pp. 561–572.

[81] Todd R Golub et al. "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring". In: *science* 286.5439 (1999), pp. 531–537.

[82] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. "Speech recognition with deep recurrent neural networks". In: *2013 IEEE international conference on acoustics, speech and signal processing*. IEEE. 2013, pp. 6645–6649.

[83] Andrew Grimson et al. "MicroRNA targeting specificity in mammals: determinants beyond seed pairing". In: *Molecular cell* 27.1 (2007), pp. 91–105.

[84] Robert L. Grossman et al. "Toward a Shared Vision for Cancer Genomic Data". In: *New England Journal of Medicine* 375.12 (Sept. 2016). Publisher: Massachusetts Medical Society, pp. 1109–1112.

[85] Brian J Haas et al. "STAR-Fusion: fast and accurate fusion transcript detection from RNA-Seq". In: *BioRxiv* (2017), p. 120295.

[86] Aric Hagberg, Pieter Swart, and Daniel S Chult. *Exploring network structure, dynamics, and function using NetworkX*. Tech. rep. Los Alamos National Lab.(LANL), Los Alamos, NM (United States), 2008.

[87] Sam Hanash. "Disease proteomics". In: *Nature* 422.6928 (2003), pp. 226–232.

[88] Yehudit Hasin, Marcus Seldin, and Aldons Lusis. "Multi-omics approaches to disease". In: *Genome biology* 18.1 (2017), pp. 1–15.

[89] Jean Hausser and Mihaela Zavolan. "Identification and consequences of miRNA–target interactions—beyond repression of gene expression". In: *Nature Reviews Genetics* 15.9 (2014), pp. 599–612.

[90] Mohammad Havaei et al. "Brain tumor segmentation with deep neural networks". In: *Medical image analysis* 35 (2017), pp. 18–31.

[91] John Hawkins and Mikael Bodén. "The applicability of recurrent neural networks for biological sequence analysis". In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 2.3 (2005), pp. 243–253.

[92] Geoffrey E Hinton et al. "Improving neural networks by preventing co-adaptation of feature detectors". In: *arXiv preprint arXiv:1207.0580* (2012).

[93] Sheng-Da Hsu et al. "miRTarBase: a database curates experimentally validated microRNA–target interactions". In: *Nucleic acids research* 39.suppl_1 (2011), pp. D163–D169.

[94] Hsi-Yuan Huang et al. "miRTarBase 2020: updates to the experimentally validated microRNA–target interaction database". In: *Nucleic acids research* 48.D1 (2020), pp. D148–D154.

[95] De-Shuang Huang and Chun-Hou Zheng. "Independent component analysis-based penalized discriminant method for tumor classification using gene expression data". In: *Bioinformatics* 22.15 (2006), pp. 1855–1862.

[96] De-Shuang Huang et al. "Classifying protein sequences using hydropathy blocks". In: *Pattern recognition* 39.12 (2006), pp. 2293–2300.

[97] Tim Hubbard et al. "The Ensembl genome database project". In: *Nucleic acids research* 30.1 (2002), pp. 38–41.

[98] Wolfgang Huber et al. "Variance stabilization applied to microarray data calibration and to the quantification of differential expression". In: *Bioinformatics* 18.suppl_1 (July 2002). Publisher: Oxford Academic, S96–S104.

[99] Wolfgang Huber et al. "Variance stabilization applied to microarray data calibration and to the quantification of differential expression". In: *Bioinformatics* 18.suppl_1 (2002), S96–S104.

[100] Saddam Hussain, Syed Muhammad Anwar, and Muhammad Majid. "Segmentation of glioma tumors in brain using deep convolutional neural network". In: *Neurocomputing* 282 (2018), pp. 248–261.

[101] Daehee Hwang et al. "A data integration methodology for systems biology". In: *Proceedings of the National Academy of Sciences* 102.48 (2005), pp. 17296–17301.

[102] Matthew K Iyer, Arul M Chinnaiyan, and Christopher A Maher. "ChimeraScan: a tool for identifying chimeric transcription in sequencing data". In: *Bioinformatics* 27.20 (2011), pp. 2903–2904.

[103] Anil K Jain, M Narasimha Murty, and Patrick J Flynn. "Data clustering: a review". In: *ACM computing surveys (CSUR)* 31.3 (1999), pp. 264–323.

[104] Lever Jake et al. "CancerMine: a literature-mined resource for drivers, oncogenes and tumor suppressors in cancer". In: *Nature Methods* 16 (2019), pp. 505–507. DOI: https://doi.org/10.1038/s41592-019-0422-y.

[105] Ahmedin Jemal et al. "Global cancer statistics". In: *CA: a cancer journal for clinicians* 61.2 (2011), pp. 69–90.

[106] Mark A Jensen et al. "The NCI Genomic Data Commons as an engine for precision medicine". In: *Blood, The Journal of the American Society of Hematology* 130.4 (2017), pp. 453–459.

[107] Shuiwang Ji, Ahmed Fakhry, and Houtao Deng. "Integrative analysis of the connectivity and gene expression atlases in the mouse brain". In: *NeuroImage* 84 (2014), pp. 245–253.

[108] Wenlong Jia et al. "SOAPfuse: an algorithm for identifying fusion transcripts from paired-end RNA-Seq data". In: *Genome biology* 14.2 (2013), R12.

[109] ZQ John Lu. "The elements of statistical learning: data mining, inference, and prediction". In: *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 173.3 (2010), pp. 693–694.

[110] Travis S Johnson et al. "Integration of Mouse and Human Single-cell RNA Sequencing Infers Spatial Cell-type Composition in Human Brains". In: *bioRxiv* (2019). DOI: 10.1101/527499. eprint: https://www.biorxiv.org/content/early/2019/01/22/527499.full.pdf. URL: https://www.biorxiv.org/content/early/2019/01/22/527499.

[111] Andrew R Joyce and Bernhard Ø Palsson. "The model organism as a system: integrating omics data sets". In: *Nature reviews Molecular cell biology* 7.3 (2006), pp. 198–210.

[112] Anne Kallioniemi et al. "Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors". In: *Science* 258.5083 (1992), pp. 818–821.

[113] Tomihisa Kamada and Satoru Kawai. "An algorithm for drawing general undirected graphs". In: *Information Processing Letters* 31.1 (1989), pp. 7–15. ISSN: 0020-0190. DOI: https://doi.org/10.1016/0020-0190(89)90102-6. URL: http://www.sciencedirect.com/science/article/pii/0020019089901026.

[114] Eric R Kandel et al. *Principles of neural science*. Vol. 4. McGraw-hill New York, 2000.

[115] Kalpana Kannan et al. "Recurrent BCAM-AKT2 fusion gene leads to a constitutively activated AKT2 fusion kinase in high-grade serous ovarian carcinoma". In: *Proceedings of the National Academy of Sciences* 112.11 (2015), E1272–E1277. ISSN: 0027-8424. DOI: 10.1073/pnas.1501735112. eprint: https://www.pnas.org/content/112/11/E1272.full.pdf. URL: https://www.pnas.org/content/112/11/E1272.

[116] Alon Kaufman et al. "Gene expression of Caenorhabditis elegans neurons carries information on their synaptic connectivity". In: *PLoS computational biology* 2.12 (2006), e167.

[117] Rongqin Ke et al. "Fourth Generation of Next-Generation Sequencing Technologies: Promise and Consequences". In: *Human mutation* 37.12 (2016), pp. 1363–1367.

[118] David R Kelley, Jasper Snoek, and John L Rinn. "Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks". In: *Genome research* 26.7 (2016), pp. 990–999.

[119] W James Kent et al. "The human genome browser at UCSC". In: *Genome research* 12.6 (2002), pp. 996–1006.

[120] Daehwan Kim and Steven L Salzberg. "TopHat-Fusion: an algorithm for discovery of novel fusion transcripts". In: *Genome biology* 12.8 (2011), R72.

[121] Pora Kim et al. "ChimerDB 2.0 a knowledgebase for fusion genes updated". In: *Nucleic acids research* 38.suppl_1 (2009), pp. D81–D85.

[122] Alwin Köhler and Ed Hurt. "Exporting RNA from the nucleus to the cytoplasm". In: *Nature reviews Molecular cell biology* 8.10 (2007), pp. 761–773.

[123] Weicong Kong et al. "Short-term residential load forecasting based on LSTM recurrent neural network". In: *IEEE Transactions on Smart Grid* 10.1 (2017), pp. 841–851.

[124] Leonard Kuan et al. "Neuroinformatics of the allen mouse brain connectivity atlas". In: *Methods* 73 (2015), pp. 4–17.

[125] Zoe Lacroix. "Biological data integration: wrapping data and tools". In: *IEEE Transactions on information technology in biomedicine* 6.2 (2002), pp. 123–128.

[126] Jack Lanchantin et al. "Deep motif: Visualizing genomic sequence classifications". In: *arXiv preprint arXiv:1605.01133* (2016).

[127] Mark Larance and Angus I Lamond. "Multidimensional proteomics for cell biology". In: *Nature reviews Molecular cell biology* 16.5 (2015), pp. 269–280.

[128] Natasha S Latysheva and M Madan Babu. "Discovering and understanding oncogenic gene fusions through data intensive computational approaches". In: *Nucleic acids research* 44.10 (2016), pp. 4487–4503.

[129] Yann LeCun et al. "Gradient-based learning applied to document recognition". In: *Proceedings of the IEEE* 86.11 (1998), pp. 2278–2324.

[130] Yann LeCun et al. "Handwritten digit recognition with a back-propagation network". In: *Advances in neural information processing systems* 2 (1989), pp. 396–404.

[131] Myunggyo Lee et al. "ChimerDB 3.0: an enhanced database for fusion genes from cancer transcriptome and literature data mining". In: *Nucleic acids research* 45.D1 (2016), pp. D784–D789.

[132] Pierre Legrain et al. "The human proteome project: current state and future direction". In: *Molecular & cellular proteomics* 10.7 (2011).

[133] Ed S Lein et al. "Genome-wide atlas of gene expression in the adult mouse brain". In: *Nature* 445.7124 (2007), p. 168.

[134] Rasko Leinonen et al. "The sequence read archive". In: *Nucleic Acids Research* 39.Database issue (Jan. 2011), pp. D19–21.

[135] Jeffrey M Levsky and Robert H Singer. "Gene expression and the myth of the average cell". In: *Trends in cell biology* 13.1 (2003), pp. 4–6.

[136] Benjamin P Lewis, Christopher B Burge, and David P Bartel. "Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets". In: *cell* 120.1 (2005), pp. 15–20.

[137] Yang Li et al. "FusionHunter: identifying fusion transcripts in cancer using paired-end RNA-seq". In: *Bioinformatics* 27.12 (2011), pp. 1708–1710.

[138] Yifeng Li, Fang-Xiang Wu, and Alioune Ngom. "A review on machine learning principles for multi-view biological data integration". In: *Briefings in bioinformatics* 19.2 (2018), pp. 325–340.

[139] Jianxiao Liu et al. "Application of deep learning in genomics". In: *Science China Life Sciences* (2020), pp. 1–19.

[140] Kun-Hong Liu and De-Shuang Huang. "Cancer classification using rotation forest". In: *Computers in biology and medicine* 38.5 (2008), pp. 601–610.

[141] M Natividad Lobato et al. "Modeling chromosomal translocations using conditional alleles to recapitulate initiating events in human leukemias". In: *Journal of the National Cancer Institute Monographs* 2008.39 (2008), pp. 58–63.

[142] Eric F Lock et al. "Joint and individual variation explained (JIVE) for integrated analysis of multiple data types". In: *The annals of applied statistics* 7.1 (2013), p. 523.

[143] Michael I Love, Wolfgang Huber, and Simon Anders. "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2". In: *Genome biology* 15.12 (2014), p. 550.

[144] Michael I. Love, Wolfgang Huber, and Simon Anders. "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2". In: *Genome Biology* 15.12 (Dec. 2014), p. 550.

[145] Marta Lovino et al. "A deep learning approach to the screening of oncogenic gene fusions in humans". In: *International journal of molecular sciences* 20.7 (2019), p. 1645.

[146] Marta Lovino et al. "DEEPrior: a deep learning tool for the prioritization of gene fusions". In: *Bioinformatics* 36.10 (Feb. 2020), pp. 3248–3250. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btaa069. eprint: https://academic.oup.com/bioinformatics/article-pdf/36/10/3248/33204199/btaa069.pdf. URL: https://doi.org/10.1093/bioinformatics/btaa069.

[147] Marta Lovino et al. "Multi-omics Classification on Kidney Samples Exploiting Uncertainty-Aware Models". In: *International Conference on Intelligent Computing*. Springer. 2020, pp. 32–42.

[148] Christopher A Maher et al. "Transcriptome sequencing to detect gene fusions in cancer". In: *Nature* 458.7234 (2009), pp. 97–101.

[149] S. Mallik et al. "Integrated analysis of gene expression and genome-wide DNA methylation for tumor prediction: An association rule mining-based approach". In: Apr. 2013, pp. 120–127.

[150] Loredana Martignetti et al. "ROMA: representation and quantification of module activity from target expression data". In: *Frontiers in genetics* 7 (2016), p. 18.

[151] Iñigo Martincorena and Peter J Campbell. "Somatic mutation in cancer and normal cells". In: *Science* 349.6255 (2015), pp. 1483–1489.

[152] Natalia J Martinez and Albertha JM Walhout. "The interplay between transcription factors and microRNAs in genome-scale regulatory networks". In: *Bioessays* 31.4 (2009), pp. 435–445.

[153] Andrew McPherson et al. "deFuse: an algorithm for gene fusion discovery in tumor RNA-Seq data". In: *PLoS Comput Biol* 7.5 (2011), e1001138.

[154] Fredrik Mertens, Cristina R Antonescu, and Felix Mitelman. "Gene fusions in soft tissue tumors: recurrent and overlapping pathogenetic themes". In: *Genes, Chromosomes and Cancer* 55.4 (2016), pp. 291–310.

[155] Fredrik Mertens and Johnbosco Tayebwa. "Evolving techniques for gene fusion detection in soft tissue tumours". In: *Histopathology* 64.1 (2014), pp. 151–162.

[156] Fredrik Mertens et al. "The emerging complexity of gene fusions in cancer". In: *Nature Reviews Cancer* 15.6 (2015), p. 371.

[157] Philipp Mertins et al. "Proteogenomics connects somatic mutations to signalling in breast cancer". In: *Nature* 534.7605 (2016), pp. 55–62.

[158] Seonwoo Min, Byunghan Lee, and Sungroh Yoon. "Deep learning in bioinformatics". In: *Briefings in bioinformatics* 18.5 (2017), pp. 851–869.

[159] Serban Nacu et al. "Deep RNA sequencing analysis of readthrough gene fusions in human prostate adenocarcinoma and reference samples". In: *BMC medical genomics* 4.1 (2011), pp. 1–22.

[160] Mridula Nambiar, Vijayalakshmi Kari, and Sathees C Raghavan. "Chromosomal translocations in cancer". In: *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer* 1786.2 (2008), pp. 139–152.

[161] National Cancer Institute. *GDC Data Portal*. https://portal.gdc.cancer.gov/, last accessed on 2020-06-14.

[162] National Human Genome Research Institute. *The Cost of Sequencing a Human Genome*. https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost, last accessed on 2020-06-14.

[163] Daniel Nicorici et al. "FusionCatcher-a tool for finding somatic fusion genes in paired-end RNA-sequencing data". In: *BioRxiv* (2014), p. 011650.

[164] Francisco J Novo, Iñigo Ortiz de Mendıbil, and José L Vizmanos. "TICdb: a collection of gene-mapped translocation breakpoints in cancer". In: *BMC genomics* 8.1 (2007), pp. 1–5.

[165] Gregor Obernosterer et al. "Post-transcriptional regulation of microRNA expression". In: *Rna* 12.7 (2006), pp. 1161–1167.

[166] Seung Wook Oh et al. "A mesoscale connectome of the mouse brain". In: *Nature* 508.7495 (2014), p. 207.

[167] Konstantin Okonechnikov et al. "InFusion: advancing discovery of fusion genes and chimeric transcripts from deep RNA-sequencing data". In: *PloS one* 11.12 (2016), e0167417.

[168] Bernhard Palsson and Karsten Zengler. "The challenges of integrating multi-omic data sets". en. In: *Nature Chemical Biology* 6.11 (Nov. 2010), pp. 787–789. ISSN: 1552-4450, 1552-4469. DOI: 10.1038/nchembio.462. URL: http://www.nature.com/articles/nchembio.462 (visited on 11/16/2020).

[169] Adam Paszke et al. "Automatic differentiation in PyTorch". In: (Oct. 2017).

[170] Joshua B Plotkin. "Transcriptional regulation is only half the story". In: *Molecular systems biology* 6.1 (2010), p. 406.

[171] David Martin Powers. "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation". In: (2011).

[172] Harry Pratt et al. "Convolutional neural networks for diabetic retinopathy". In: *Procedia Computer Science* 90 (2016), pp. 200–205.

[173] Nimrod Rappoport and Ron Shamir. "Multi-omic and multi-view clustering algorithms: review and cancer benchmark". en. In: *Nucleic Acids Research* 46.20 (Nov. 2018), pp. 10546–10562. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gky889. URL: https://academic.oup.com/nar/article/46/20/10546/5123392 (visited on 11/07/2020).

[174] Nimrod Rappoport and Ron Shamir. "Multi-omic and multi-view clustering algorithms: review and cancer benchmark". In: *Nucleic acids research* 46.20 (2018), pp. 10546–10562.

[175] Waseem Rawat and Zenghui Wang. "Deep convolutional neural networks for image classification: A comprehensive review". In: *Neural computation* 29.9 (2017), pp. 2352–2449.

[176] Hilda Razzaghi et al. “Leading causes of cancer mortality—Caribbean region, 2003–2013”. In: *Morbidity and Mortality Weekly Report* 65.49 (2016), pp. 1395–1400.

[177] Knut Reinert et al. “The SeqAn C++ template library for efficient sequence analysis: a resource for programmers”. In: *Journal of biotechnology* 261 (2017), pp. 157–168. DOI: https://doi.org/10.1016/j.jbiotec.2017.07.017.

[178] Jonas Richiardi et al. “Correlated gene expression supports synchronous activity in brain networks”. In: *Science* 348.6240 (2015), pp. 1241–1244.

[179] Matthew E. Ritchie et al. “limma powers differential expression analyses for RNA-sequencing and microarray studies”. en. In: *Nucleic Acids Research* 43.7 (Apr. 2015). Publisher: Oxford Academic, e47–e47. ISSN: 0305-1048. DOI: 10.1093/nar/gkv007. URL: https://academic.oup.com/nar/article/43/7/e47/2414268 (visited on 11/09/2020).

[180] Keith D Robertson and Alan P Wolffe. “DNA methylation in health and disease”. In: *Nature Reviews Genetics* 1.1 (2000), pp. 11–19.

[181] Ana I. Robles et al. “An Integrated Prognostic Classifier for Stage I Lung Adenocarcinoma based on mRNA, microRNA and DNA Methylation Biomarkers”. In: 10 (July 2015), pp. 1037–1048.

[182] Juan José Rodriguez, Ludmila I Kuncheva, and Carlos J Alonso. “Rotation forest: A new classifier ensemble method”. In: *IEEE transactions on pattern analysis and machine intelligence* 28.10 (2006), pp. 1619–1630.

[183] Bernardo Rodriguez-Martin et al. “ChimPipe: Accurate detection of fusion genes and transcription-induced chimeras from RNA-seq data”. In: *BMC genomics* 18.1 (2017), p. 7.

[184] Marcia Roy et al. “Regional diversity in the postsynaptic proteome of the mouse brain”. In: *Proteomes* 6.3 (2018), p. 31.

[185] Nand K Roy et al. “Techniques to Identify Novel Fusion Genes and to Detect Known Fusion Genes”. In: *Fusion Genes And Cancer*. World Scientific, 2017, pp. 59–79.

[186] Onur Sakarya et al. “RNA-Seq mapping and detection of gene fusions with a suffix array algorithm”. In: *PLoS Comput Biol* 8.4 (2012), e1002464.

[187] Mike Schuster and Kuldip K Paliwal. “Bidirectional recurrent neural networks”. In: *IEEE transactions on Signal Processing* 45.11 (1997), pp. 2673–2681.

[188] Alice T Shaw et al. “Effect of crizotinib on overall survival in patients with advanced non-small-cell lung cancer harbouring ALK gene rearrangement: a retrospective analysis”. In: *The lancet oncology* 12.11 (2011), pp. 1004–1012.

[189] Mikhail Shugay et al. “Oncofuse: a computational framework for the prediction of the oncogenic potential of gene fusions”. In: *Bioinformatics* 29.20 (2013), pp. 2539–2546.

[190] Derek Sieburth et al. “Systematic analysis of genes required for synapse structure and function”. In: *Nature* 436.7050 (2005), p. 510.

[191] Korsuk Sirinukunwattana et al. “Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images”. In: *IEEE transactions on medical imaging* 35.5 (2016), pp. 1196–1206.

[192] Zbyslaw Sondka et al. “The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers”. In: *Nature Reviews Cancer* 18.11 (2018), pp. 696–705.

[193] Olaf Sporns, Giulio Tononi, and Rolf Kötter. “The human connectome: a structural description of the human brain”. In: *PLoS computational biology* 1.4 (2005), e42.

[194] Andrew D Strand et al. “Conservation of regional gene expression in mouse and human brain”. In: *PLoS genetics* 3.4 (2007), e59.

[195] Nicolas Stransky et al. “The landscape of kinase fusions in cancer”. In: *Nature communications* 5.1 (2014), pp. 1–10.

[196] Indhupriya Subramanian et al. “Multi-omics Data Integration, Interpretation, and Its Application”. In: *Bioinformatics and Biology Insights* 14 (Jan. 2020). ISSN: 1177-9322. DOI: 10.1177/1177932219899051. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7003173/ (visited on 11/16/2020).

[197] Wenqing Sun et al. “Enhancing deep convolutional neural network scheme for breast cancer diagnosis with unlabeled data”. In: *Computerized Medical Imaging and Graphics* 57 (2017), pp. 4–9.

[198] Binhua Tang et al. “Recent advances of deep learning in bioinformatics and computational biology”. In: *Frontiers in genetics* 10 (2019), p. 214.

[199] Wei Tang et al. “Tumor origin detection with tissue-specific miRNA and DNA methylation markers”. In: *Bioinformatics* 34.3 (Feb. 2018). Publisher: Oxford Academic, pp. 398–406.

[200] Diethard Tautz and Christine Pfeifle. “A non-radioactive in situ hybridization method for the localization of specific RNAs in Drosophila embryos reveals translational control of the segmentation gene hunchback”. In: *Chromosoma* 98.2 (1989), pp. 81–85.

[201] Arthur Tenenhaus and Michel Tenenhaus. “Regularized generalized canonical correlation analysis”. In: *Psychometrika* 76.2 (2011), p. 257.

[202] J Michael Thomson et al. "Extensive post-transcriptional regulation of microRNAs and its implications for cancer". In: *Genes & development* 20.16 (2006), pp. 2202–2207.

[203] Katarzyna Tomczak, Patrycja Czerwinska, and Maciej Wiznerowicz. "The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge". In: *Contemporary oncology* 19.1A (2015), A68.

[204] Scott A. Tomlins et al. "Role of the TMPRSS2-ERG Gene Fusion in Prostate Cancer". In: *Neoplasia* 10.2 (2008), 177–IN9. ISSN: 1476-5586. DOI: https://doi.org/10.1593/neo.07822. URL: http://www.sciencedirect.com/science/article/pii/S1476558608800644.

[205] Jeffrey A Toretsky et al. "Oncoprotein EWS-FLI1 activity is enhanced by RNA helicase A". In: *Cancer research* 66.11 (2006), pp. 5574–5581.

[206] Wandaliz Torres-Garcia et al. "PRADA: pipeline for RNA sequencing data analysis". In: *Bioinformatics* 30.15 (2014), pp. 2224–2226.

[207] Mike Tyers and Matthias Mann. "From genomics to proteomics". In: *Nature* 422.6928 (2003), pp. 193–197.

[208] The University-of-sout-California. *The BAMS Rat Connectome Project*. https://bams2.bams1.org/connections/grid/80/. 2013.

[209] Christine Vogel and Edward M Marcotte. "Insights into the regulation of protein abundance from proteomic and transcriptomic analyses". In: *Nature reviews genetics* 13.4 (2012), pp. 227–232.

[210] John M Walker. *The proteomics protocols handbook*. Springer, 2005.

[211] Bo Wang et al. "Similarity network fusion for aggregating data types on a genomic scale". In: *Nature Methods* 11.3 (Mar. 2014). Number: 3 Publisher: Nature Publishing Group, pp. 333–337.

[212] Shu-Lin Wang et al. "Tumor classification by combining PNN classifier ensemble with neighborhood rough set based gene reduction". In: *Computers in Biology and Medicine* 40.2 (2010), pp. 179–189.

[213] Xiaowei Wang. "miRDB: a microRNA target prediction and functional annotation database with a wiki interface". In: *Rna* 14.6 (2008), pp. 1012–1017.

[214] Paul J Werbos. "Applications of advances in nonlinear sensitivity analysis". In: *System modeling and optimization*. Springer, 1982, pp. 762–770.

[215] Nathan Wong and Xiaowei Wang. "miRDB: an online resource for microRNA target prediction and functional annotations". In: *Nucleic acids research* 43.D1 (2015), pp. D146–D152.

[216] Chunxiao Wu et al. "Poly-gene fusion transcripts and chromothripsis in prostate cancer". In: *Genes, Chromosomes and Cancer* 51.12 (2012), pp. 1144–1153.

[217] Jikun Wu et al. "SOAPfusion: a robust and effective computational fusion discovery tool for RNA-seq reads". In: *Bioinformatics* 29.23 (2013), pp. 2971–2978.

[218] Yonghui Wu et al. "Google's neural machine translation system: Bridging the gap between human and machine translation". In: *arXiv preprint arXiv:1609.08144* (2016).

[219] Andrew D Yates, Premanand Achuthan, Akanni, et al. "Ensembl 2020". In: *Nucleic Acids Research* 48.D1 (Nov. 2019), pp. D682–D688. ISSN: 0305-1048. DOI: 10.1093/nar/gkz966. eprint: https://academic.oup.com/nar/article-pdf/48/D1/D682/31697830/gkz966.pdf. URL: https://doi.org/10.1093/nar/gkz966.

[220] Haoyang Zeng et al. "Convolutional neural network architectures for predicting DNA–protein binding". In: *Bioinformatics* 32.12 (2016), pp. i121–i127.

[221] Wei Zhang. "Shift-invariant pattern recognition neural network and its optical architecture". In: *Proceedings of annual conference of the Japan Society of Applied Physics*. 1988.

[222] Wei Zhang et al. "Parallel distributed processing model with local space-invariant interconnections and its optical architecture". In: *Applied optics* 29.32 (1990), pp. 4790–4797.

[223] Jing Zhao et al. "Multi-view learning overview: Recent progress and new challenges". en. In: *Information Fusion* 38 (Nov. 2017), pp. 43–54. ISSN: 15662535. DOI: 10.1016/j.inffus.2017.02.007. URL: https://linkinghub.elsevier.com/retrieve/pii/S1566253516302032 (visited on 11/16/2020).

[224] Xing-Ming Zhao, Yiu-Ming Cheung, and De-Shuang Huang. "A novel approach to extracting features from motif content and protein composition for protein sequence classification". In: *Neural Networks* 18.8 (2005), pp. 1019–1028.

[225] Chun-Hou Zheng, De-Shuang Huang, and Li Shang. "Feature selection in independent component subspace for microarray data classification". In: *Neurocomputing* 69.16-18 (2006), pp. 2407–2410.

[226] Chun-Hou Zheng et al. "Gene expression data classification using consensus independent component analysis". In: *Genomics, proteomics & bioinformatics* 6.2 (2008), pp. 74–82.

[227] Chun-Hou Zheng et al. "Metasample-based sparse representation for tumor classification". In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 8.5 (2011), pp. 1273–1282.

[228] Jian Zhou and Olga Troyanskaya. "Deep supervised and convolutional generative stochastic network for protein secondary structure prediction". In: *International conference on machine learning*. PMLR. 2014, pp. 745–753.

[229] Zhi-Hua Zhou et al. "Big data opportunities and challenges: Discussions from data analytics perspectives [discussion forum]". In: *IEEE Computational intelligence magazine* 9.4 (2014), pp. 62–74.

This Ph.D. thesis has been typeset by means of the TEX-system facilities. The typesetting engine was pdfLATEX. The document class was `toptesi`, by Claudio Beccari, with option `tipotesi=scudo`. This class is available in every up-to-date and complete TEX-system installation.