UNIVERSITA' DEGLI STUDI DI TORINO:

DIPARTIMENTO DI: Fisica

DOTTORATO DI RICERCA IN :
Sistemi complessi per le scienze della vita

CICLO: XXXV

TITOLO DELLA TESI: Topic modeling methods for the analysis of gene expression data

TESI PRESENTATA DA: Filippo Valle

TUTOR(S): Prof. Michele Caselle

COORDINATORE DEL DOTTORATO: Prof. Enzo Medico

ANNI ACCADEMICI: 2019/2020-2020/2021-2021/2022

SETTORE SCIENTIFICO-DISCIPLINARE DI AFFERENZA*: FIS/02

GLOSSARY

ITALIAN	**ENGLISH**
Dipartimento di Fisica	Department of Physics
Dottorato di ricerca in Sistemi Complessi per le scienze della vita	PhD Programme in Complex Systems for Life Sciences
Ciclo XXXV	Cycle XXXV
titolo della tesi Topic modeling methods for the analysis of gene expression data	Topic modeling methods for the analysis of gene expression data
Tesi presentata da Filippo Valle	Thesis' author Filippo Valle
Tutor(s) Michele Caselle	Supervisor(s) Michele Caselle
Coordinatore del Dottorato prof. Enzo Medico	PhD Programme Co-ordinator Enzo Medico
Anni Accademici 2019/2020-2020/20221-2021/2022	Academic years of enrolment 2019/2020-2020/20221-2021/2022
Settore Scientifico Disciplinare di Afferenza FIS/02	Code of scientific discipline FIS/02

University of Turin

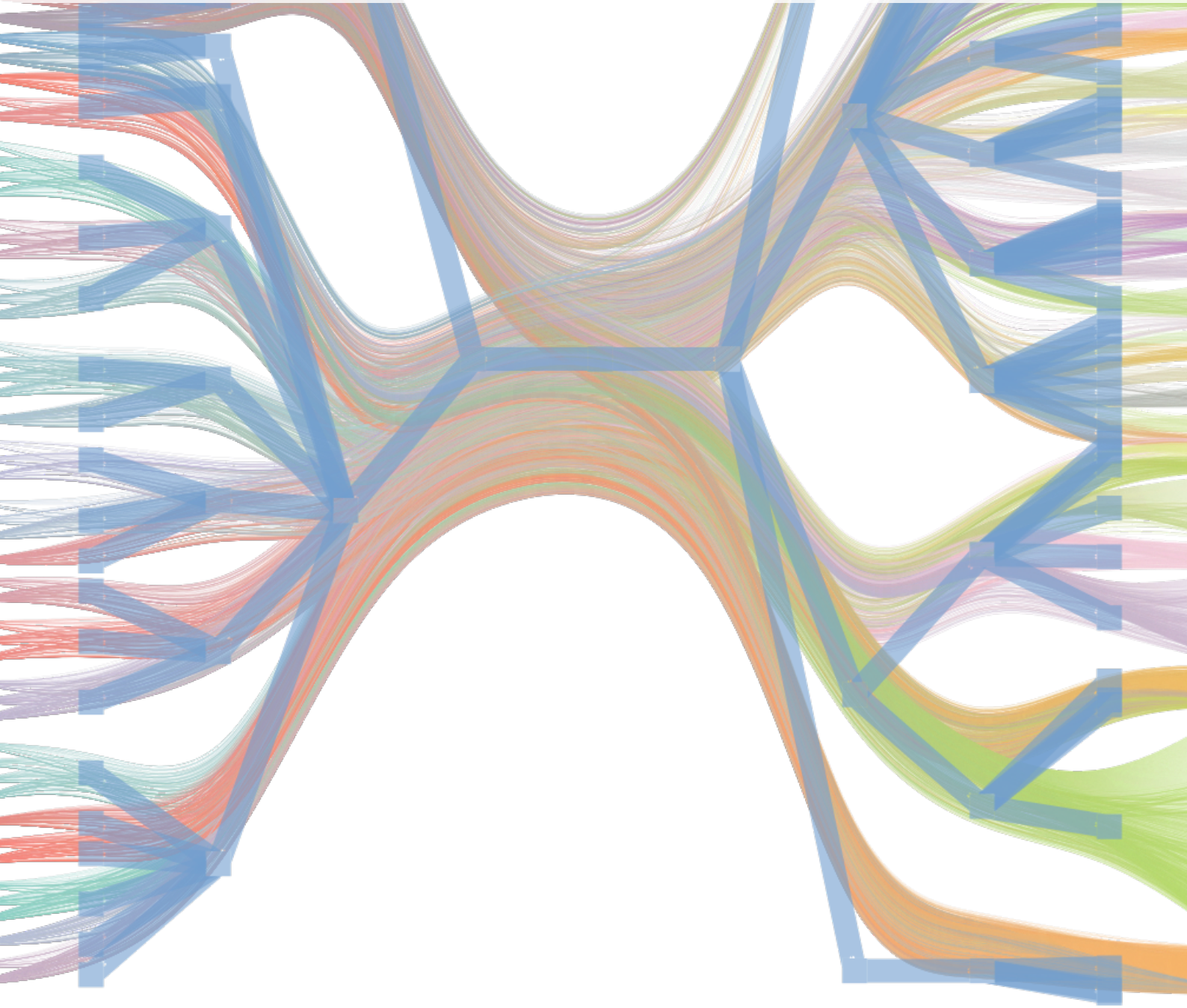
Complex Systems for Life Sciences

Topic Modeling Methods for the Analysis of Gene Expression Data

Filippo Valle

Tutor: Michele Caselle

Ph.D. Thesis, XXXV cycle, Spring 2023



This thesis is submitted under the Ph.D. program *Complex Systems for Life Sciences*, coordinated by Prof. Enzo Medico. I worked at the Physics' Department at the University of Turin during the academic years: 2019-2020 2020-2021 2021-2022; I was supervised by Prof. Michele Caselle. The activity sector of this thesis is *FIS/02*.

The front page depicts a bipartite network.

. . . the imagination of nature is far, far greater than the imagination of man.[1]

[1] *What do you care what other people think?* Richard Feynman

Abstract

Topic modeling is a widely used approach to extract relevant information from large datasets. Recently the problem of finding a latent structure in a dataset was mapped to the community detection problem in network theory and a new class of topic modeling strategies has been introduced to overcome some of the limitations of classical methods. We tested this approach on lung and breast cancer samples from the TCGA and METABRIC databases, using data of messenger RNA, microRNAs and copy number variations. The established cancer subtype organization is well reconstructed in the inferred latent topic structure. Moreover, the "topic" that the algorithm extracts correspond to genes involved in cancer development and they are enriched in genes known to play a role in the corresponding disease; they are strongly related to the survival probability of patients too. In biology, integrating transcriptional data with other layers of information, such as the post-transcriptional regulation mediated by microRNAs, can be crucial in identifying the driver genes and the subtypes of complex and heterogeneous diseases such as cancer. More specifically, we show how an algorithm based on a hierarchical version of stochastic block modeling can be adapted to integrate any combination of data. We will also show that the inclusion of the microRNAs layer significantly improves the accuracy of subtype classification. As a final result, we show how operating in the low dimensional topic space, one can predict the cancer subtype of a new unseen expression sample.

Contents

CHAPTER 1

Introduction

In the recent years the number of gene expression measurements is growing and the challenges of comprehending and interpreting the complexity of these biological networks are growing as well. Clustering, both genes and samples (or single-cells), has been playing a major role in revealing hidden structures in transcriptomic datasets.

The availability of gene expression datasets enables advancement in the understanding of how changes in gene expression affects phenotypes. Recognising more precisely the population a sample belongs to is part of an effort to develop personalised drugs: the so-called precision medicine [6]. Reduce the dimensionality and find a biologically relevant embedding space can also help to describe and to visualise the network of the data [6, 71, 73]. Many different algorithms have been applied to this kind of data so far; these include t-SNE [130], k-means, Nonnegative Matrix Factorisation [12], LDA [10], Topic Mapping [75], WGCNA [76] and others [28, 30, 66, 69, 135].

We propose the use of topic modeling [57] to cluster and to mine information from transcriptomic data and nevertheless to reduce the dimensionality of the datasets. Topic models, in particular in their popular and widely used LDA [10] flavour, are being used in biophysics [137] and in particular by [29] on the GTEx data as a grade of a membership model. They are particularly interesting because they allow a complete description of the data space in terms of probability distributions and they are way more informative than a simple clustering approach.

A few years ago, it has been proved that topic modeling can be related to a community detection [34, 35, 88] problem on bipartite networks [40], at the same time a new algorithm, which solved some of the problems of an LDA-only approach, emerged: it is known as hierarchical Stochastic Block Model (hSBM). This well-developed techniques based on stochastic block modeling [58] can be applied without the need of a aprticular prior or using any hyper-parameter.

During my Ph.D., we investigated the properties of hSBM in classifying RNA-sequencing data mainly from GTEx [45, 86],TCGA [7, 19] and METABRIC [27], in particular studying cancer (sub)types [127, 128] and we learnt some useful properties of this approach. The use of topic modeling in this framework was driven by the fact that we observed emergent statistical laws in single-cell transcriptomic data [3, 77], the most important being the so-called Zipf's law [138] common in many complex systems [4, 25, 42, 93, 95]. This will be described in Chapter 2.

We demonstrated that hierarchical Stochastic Block Model can cluster together samples known to be similar and that this happens at different layers of resolution.

In Chapter 3 we will investigate the behaviour and the properties of topic modeling if one considers healthy tissues. We compare different algorithms to find the one which is optimal for this task [115], then we looked into the latent structure of the data that emerged after each run. When using hSBM, a global pattern emerges over the single-tissue structure, this encouraged us to prefer this approach among LDA or others. Recently [121] suggested that potentially important genes are ignored, the approach hereby described is unbiased in the choice of genes and could overcome this fact.

We show that applying a simple normalisation it is possible to retrieve well-known biological functions and processes. In other words the topics we found are not trivial and they strongly correlate with the biological properties of the samples analysed.

Once the method were stabilised and sufficiently understood, we propose to focus on the analysis of cancer data. Cancer is the cause of 1 over 6 deaths [26] worldwide and if the first part of this work is a

proof of concept on a relatively easy task, carrying this approache to cancer's research really worth an effort [6, 14, 15, 16, 17]. This will be described in Chapter 4.

We also developed a novel approach [126] to integrate multiple kind of data using Stochastic Block Modeling and we called it nSBM (widely described in Chapter 5). It can be useful to incorporate some information (i.e. metadata) in the model in an unsupervised fashion, for instance [65] incorporated lexical priors to add more knowledge into the model. A supervised [84] or a keyword assisted [32] version of LDA emerged. Other approaches to incorporate information made use of Dirichlet Forest [5]. The idea of incorporating annotations in community detection problems has been discussed by [90]; moreover, [62] added metadata nodes before running a Stochastic Block Modeling to reconstruct links. In a different context, it has been proposed to use Stochastic Block Modeling on multilayered networks by [129] and recently [33] investigated the transition that happens when metadata really affects SBM. Integrating information in networks can be useful, for instance, in recommendation systems [41, 47, 63].

Data in this kind of problem are often represented in the form of the so-called Bag of Words [2, 52], in the language of probability theory, this is an assumption of exchangeability for the words in a document and in the context of the component systems [81, 82, 83], it counts the number of components (words) in each realization (book). Although words are not the only source of data available or useful for topic models, recently n-grams were used in Bag of Tricks [68] and Bag of Links has been proposed [67, 107] to improve cross-lingual text classification.

Taking advantage of the network approach to topic models and using the idea of incorporating additional layers of information in networks, we propose a way of incorporating metadata in a bipartite network and we will discuss how it is useful to extract relevant information from texts in the framework of topic modeling.

I packed a python tool nSBM, derived by hSBM [40], ready to install and easily executable on n-partite networks trough graph-tool [99]. Since Stochastic Block Modeling is able to run on the whole network, each new layer contributes to its own topics and distributions.

This multi-feature approach revealed itself very useful to integrate different 'omics. We will show that including microRNA information in a topic modeling analysis improves the performances of this method [126, 127, 128]. Some results of the multiomics topic modeling will be presented in section 6.

A topic modeling approach revealed itself powerful for the analysis of spatial transcriptomics data in [116] and to integrate (using nSBM) lnc-RNA in the analysis of breast cancer single-cells [37].

Finally, let me highlight the fact that during my Ph.D. I contributed a lot to the open source community (see Section 7 for the details, links and code) improving existing projects, developing original code easily executable (packages or recipes) by anyone and developing minor classes to run side projects.

In Appendix A the reader will find all the technical details about the implementation of the models described in this work.

PART I

Component Systems

CHAPTER 2

Emergent laws in single-cell

This whole work is based on the study, the analysis and the inference on RNA-Sequencing data (either single cell or bulk), in this section we will describe how these data can be analysed and studied in the framework of the so-called component-systems.

2.1 The data structure for component systems

A transcriptomic dataset, and more generally a component system, can be described by a matrix $\{n_i^c\}$ where each entry represents the counts relative to transcript (i.e, the component) $i \in \{1, \dots, N\}$ in cell (i.e., the realization) $c \in \{1, \dots, R\}$. N is the total number of different transcripts that could be present (the number of genes as a first approximation), which is essentially the vocabulary of our system. R is the number of cells analyzed. Each column of the data matrix is a vector $\{n_\cdot^c\} = \{n_1^c, \dots, n_N^c\}$ that fully describes the expression profile of a single cell c. The size of the transcriptome of a cell captured in the experiment is defined as $M^c = \sum_{i=1}^N n_i^c$. While in other component systems, such as texts of natural language, this parameter is simply the size of the realization (e.g., the book size), in our context M^c represents the measured transcriptome size. Therefore, it does not necessarily correspond to the total number of transcripts in the cell, that is due to the sampling process involved in RNA capture. For each gene or transcript i, we can define its frequency in cell c as $f_i^c = \frac{n_i^c}{M_c}$ and the abundance of a gene as $E_i = \sum_j n_{ij}$.

2.2 Robust emergence of Zipf's law for the gene expression levels at different scales

One of the hallmarks of complex systems, from real-world networks to natural language, is a high level of heterogeneity, which is often characterized by the emergence of power-law distributions [91]. For component systems in particular, the frequency of components is often well described by a power law known as Zipf's law [4, 82, 91, 138]. In natural language, this law describes the distribution of word frequencies in a corpus of texts, typically reported as a rank plot. In the context of transcriptomics, this would translate in a law for the distribution of gene expression levels in a large-scale dataset. Fig. 2.1A reports the rank-plot of the relative expression levels f_i calculated by averaging across cells belonging to the same organ (different curves correspond to different organs) in the Mouse Cell Atlas (MCA) [50]. The distribution is largely compatible with a power-law decay with an exponent close to -1, as in the classic Zipf's law, followed by an exponential tail. The shape of the distribution does not depend on the specific dataset or on the experimental technique used. An essentially identical plot (Fig. 2.1B) is obtained by looking at the same organs in an alternative mouse expression atlas, i.e., Tabula Muris [123], in which different sequencing methods were adopted. This statistical property seems indeed very general and not limited to scRNAseq data or to the specific species in analysis. For example, the same law emerges considering bulk RNA sequencing measurements across healthy tissues in human from the GTEx database [45] (Fig. 2.1C). This result corroborates previous observations based on microarray and SAGE (serial analysis of gene expression) datasets that reported a power-law distribution of gene expression levels across different species and experimental conditions [36, 55, 125].

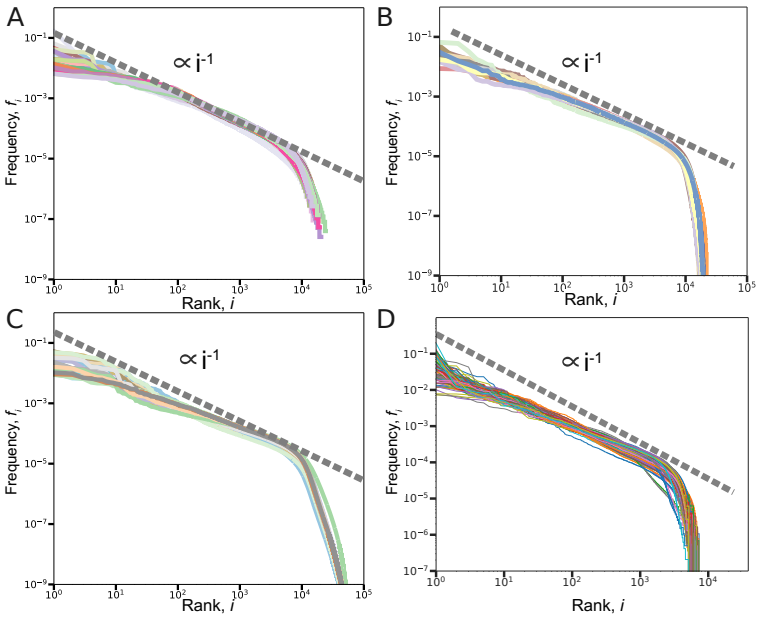


Figure 2.1: **A robust Zipf-like law for gene expression levels**. The average relative expression levels f_i reported as a function of their rank. The distributions reported correspond to averages over single cells belonging to different mouse organs from the Mouse Cell Atlas **(A)**, from Tabula Muris database **(B)**, and to bulk RNA sequencing data from samples of healthy human organs in the GTEx database **(C)**. Each curve corresponds to a single organ or tissue. **(D)** The relative gene expression levels evaluated in single cells (without averaging) follow an analogous Zipf-like trend. We report the distribution relative to 100 cells from the heart sample in Tabula Muris. Similar results can be obtained from other organs or from the Mouse Cell Atlas. The dashed lines are just a reference power-law scaling with exponent -1.

Fig. 2.1D shows an illustrative example of the gene expression distributions in single cells. Besides some variability, the distributions recapitulate the population ones reported in the other panels. Therefore, the Zipf-like behaviour is an inherent property of single-cell expression profiles.

In conclusion, Zipf's law appears to be a robustly emerging statistical property of gene expression data from bulk to single-cell experiments. This empirical law essentially sets the only free parameters f_i of our null model. Since this model only describes a sampling process given the empirical average frequencies of components, we can test what properties of the system can be explained merely by sampling effects and what features are instead potentially due to biological variability.

2.3 A Zipf's law with multiple regimes

At a coarse grained view, the rank plot of gene expression levels can be described as a power law followed by an exponential tail. The presence of a double scaling in the component frequency distribution again is a general feature of several component systems. A similar behaviour can be observed by looking at protein domain frequencies in genomes of different species [82]. A double scaling was also observed in natural language [39], where it was tentatively explained by a model with two different classes of words: common words (high rank) composing a core vocabulary and the rest of more specific words in a vast vocabulary. Analogously, two different groups of genes can be distinguished in bulk transcriptomic data: a core of highly expressed genes with active promoters and a second group of lowly expressed and putatively non-functional transcripts [55].

Considering all the cells in the MCA, highly expressed genes (around 100 genes) follow a power law with exponent close to -0.5, while the central part of the distribution is well described by an exponent close to -1 as in the classic Zipf's law. Interestingly, a very similar law with three regimes

was observed in a quantitative transcriptomic study of fission yeast [80].

The first regime appears to be composed by highly expressed genes related to basic functions that are common across different organs. This second regime is composed of actively expressed genes that are more tissue specific and whose expression approximately follows the classic Zipf's law with exponent -1.

2.4 The average number of detected transcripts follows Heaps' law as predicted by a sampling process

A complex biological system such as an organ is composed by multiple cell types with transcription programs differentiated according to their functional role. Even the repertoire of genes that have to be transcribed is expected to vary from cell to cell as a function, for example, of the level of specialization of the cellular phenotype. Therefore, a basic observable difference between single-cell expression profiles could be the total number of genes that are actually transcribed. Resuming the analogy with texts of natural language, different texts typically use a different vocabulary (i.e., total number of different words), and the size of the vocabulary can depend on several factors such as the author style or the topic complexity. However, the average vocabulary of texts empirically displays a specific and well conserved sublinear scaling with the text size, known as Heaps' law [4, 54, 82]. Again, an analogous law relates the number of different genes or protein domains to the genome size in prokaryotes [25]. Transcriptomic data present the additional complication: the number of detected transcripts also depends on the sampling process due to the RNA capture process. This naturally introduces a dependence on the sampling efficiency which is proportional to M^c, i.e., the total number of captured transcripts from a cell c. Fig. 2.2A shows the number of different mRNAs as a function of the total number of UMI (as an estimate of the total number of detected mRNAs).

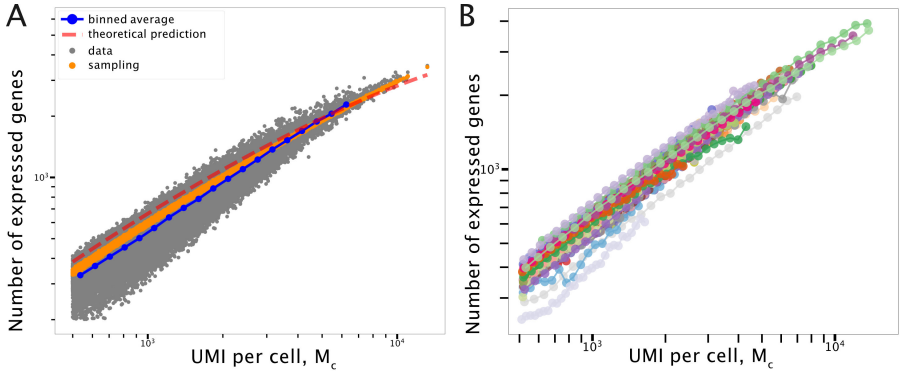


Figure 2.2: **The number of detected different transcripts follows Heaps' law. (A)** The number of mRNAs with at least one detected transcript $h(M)$ is reported as a function of the transcriptome size as measured by the number of UMI, i.e., $M^c = \sum_i n_i^c$. Each point in the scatter plot thus corresponds to a single cell, for the illustrative example of cells in the Bone Marrow from the MCA. The average empirical sublinear scaling (blue dots) is compared to the results of a stochastic sampling process using detailed simulations (orange dots) and analytical predictions (red dashed line). **(B)** The same sublinear average scaling is approximately conserved in all organs reported in the MCA. See Figure B.1 in Section B for the color legend of this Figure.

The sublinear power-law scaling is very similar to the one found in other component systems [4, 25]. This empirical trend can be compared with predictions from the model. The model assumption is that the probability of observing a specific mRNA i in the sampling process is only determined by its empirical average frequency f_i. It is easy to show [83] that according to this model the probability of not observing a mRNA given the total number of transcripts sampled M is well approximated by

$$P_i(0|M) \simeq e^{-f_i M}. \tag{2.1}$$

From this expression, we can calculate the expected number of detected different transcripts h as

$$\langle h(M)\rangle = N - \sum_{i=1}^{N} P_i(0|M) \simeq N - \sum_{i=1}^{N} e^{-f_i M}, \tag{2.2}$$

where N is the total number of possible mRNAs, given by the number of genes considered in the experiment, which is in the order of 10^4. The formula above well reproduces the results of direct simulations of the sampling process reported as orange dots, and also captures quite accurately the empirical average scaling. Therefore, the observed repertoire of expressed genes in these scRNAseq experiments is on average mostly determined by the sampling process. This trend has to be carefully taken into account in order to reliably estimate the biological variability in transcript repertoires.

A quantitative difference between the empirical average number of expressed genes (blue line in Fig. 2.2A) and the expectation from sampling (orange line) can be observed. In fact, the sampling model slightly overestimates the empirical trend. In other words, cells typically express a lower number of genes to a higher expression level than expected.

The model can be simplified by exploiting this observation. Instead of considering all the f_i values as free parameters that have to be inferred from data, we can assume the double power-law scaling, with exponents γ_1 and γ_2 estimated by fitting, and the exponential tail for low frequency components. In this case, it can be shown [83] that the expression for $h(M)$ simplifies to

$$\langle h(M)\rangle = N - \sum_{i=1}^{i^*} \left(1 - Ai^{-\gamma_1}\right)^M - \sum_{i=i^*+1}^{i^{**}} \left(1 - Bi^{-\gamma_2}\right)^M - \sum_{i=i^{**}+1}^{N} \left(1 - Ce^{-k*i}\right)^M. \tag{2.3}$$

The factors A, B, C are defined by imposing normalization and continuity conditions between the three regimes:

$$\begin{cases} A(i^*)^{-\gamma_1} = B(i^*)^{-\gamma_2} \\ B(i^{**})^{-\gamma_2} = Ce^{-k(i^{**})} \\ A\sum_{i=1}^{i^*} i^{-\gamma_1} + B\sum_{i=i^*}^{i^{**}} i^{-\gamma_2} + \sum_{i=i^{**}}^{N} Ce^{-ki} = 1. \end{cases} \tag{2.4}$$

i^* is the rank at which the change of power law exponent is estimated, while i^{**} is the rank at which the exponential regime starts. This is the theoretical prediction reported as a dashed-red line in Fig. 2.2A. If the sampling process is the dominant factor setting the repertoire of observed transcripts, the trend should not depend crucially on the biology of the system in analysis. Indeed, the sublinear scaling is well conserved across different organs as reported in Fig. 2.2B.

2.5 Variability in the repertoire of expressed genes follows Taylor's law and reveals deviations from a sampling process

As discussed in the previous section, the scaling of the average number of detected genes can be well explained as a result of the sampling process. However, there is substantial variability in the empirical data, i.e., cells with the same total number of UMI can have expression repertoires of largely different sizes. The question is if this variability can be again explained as sampling fluctuations. The model provides a precise prediction for the variance σ_h^2 as a function of the average value $\langle h\rangle$. Fig. 2.3A compares the model prediction of a Poisson scaling (blue dashed line) with the empirical scaling (grey dots) evaluated over all the cells in the MCA dataset in order to have large statistics. The empirical variance displays a power-law scaling with the average vocabulary size that is not compatible with a Poisson scaling. Fitting the empirical scaling with the function $C\langle h\rangle^k$ leads to an exponent $k = 1.64 \pm 0.18$. This value is significantly different from the Poisson scaling expected from sampling ($Z = 3.5$) and more similar to the quadratic scaling ($R^2 = 0.94$ and $Z < 2$) that has been observed for several other complex systems [31, 38, 42].

2.6 Validation on Smart-seq3

Single cell transcriptomic data can be generated with different RNA-seq library preparation protocols (10X, Smart-seq2, Smart-seq3), in particular Smart-seq2 detects more genes in a cell, especially

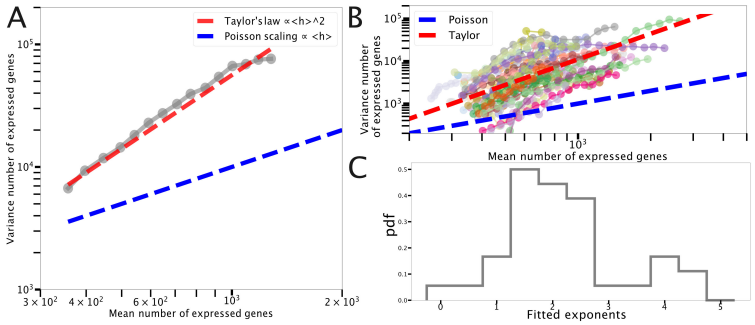


Figure 2.3: **Fluctuation scaling in the number of detected transcripts follows Taylor's law.** **(A)** The variance in the number of measured expressed genes is reported as a function of its average value for all cells in the MCA. Data are compared to a quadratic scaling (red dashed line) and to the Poisson scaling predicted by a sampling process (blue dashed line). **(B)** The fluctuation scaling is conserved by considering separately different organs and tissues. **(C)** Probability density function of the exponents k obtained by fitting the curves in panel **(B)** with $C\langle h\rangle^k$.

low abundance transcripts [133]. We searched the same Zipf and Hepas law on data coming from experiments on a human cell line (HEK) [48]. As shown in Figure 2.4 the observations made on Smart-seq3 seems consistent the ones described before and referred to other techniques.

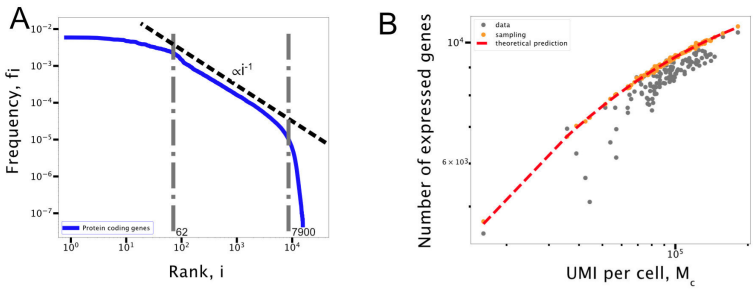


Figure 2.4: **Test consistency of results on data coming from Smart-seq3 sequencing technique.** (**A**)Zipf's law and (**B**) Heaps law referred to the HEK human cell line. They are consistent (same exponent) with the observation made in mouse with another (Smart-seq2) and described in Figure 2.1 and Figure 2.3

PART II

Topic Modeling

CHAPTER 3

Topic modeling: from Natural Language to genomic data

In section 2 we demonstrated that RNA-Sequencing data share some statistical properties with texts datasets. This suggests the possibility of using models developed in text analysis to study RNA-Sequencing data, in particular in this section we will describe how we took advantage of a topic modeling approach [57] to study RNA-sequencing healthy tissues.

Mining GTEx using hierarchical Stochastic Block Model

The scope of this part of the work is to investigate the use of topic modeling on another kind of complex systems: transcriptional datasets. Similarly to a corpus of texts, a set of samples' transcripts can be seen as documents in which every word may appear multiple times, in the context of "component systems" [81] genes are the *components* necessary to build samples that play the role of the *realizations*. Our setting is so far composed by R samples, N genes and E edges that define a many-to-many relationship between the two.

This kind of data can be naturally described as a bipartite network with genes on one side, samples on the other and links between the two sides weighted by the gene expression values. In Figure 3.1 and in the front page illustrative bipartite networks are pictured. It was demonstrated [40] that performing community detection [35] using Stochastic Block Models [58] on this kind of bipartite network is equivalent to performing topic modeling. The model considered and proposed by [40] called hierarchical Stochastic Block Modeling (from now hSBM) does exactly this and it will be deeply discussed in the next chapters of this thesis.

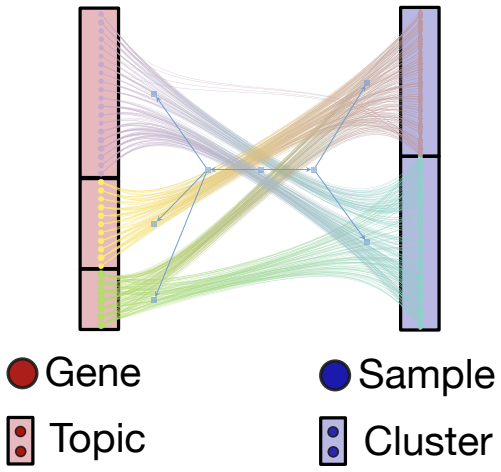


Figure 3.1: **The hSBM partition of samples in clusters and genes in topics.** The lines connecting genes and samples encode the weights of the bipartite network (i.e. the gene expression values in the different samples).

The output of the hSBM model consists of blocks of samples (clusters) and blocks of genes (topics)[1]. An illustration is in Figure 3.1. Topic models are different from other clustering algorithms because in addition to clusters they provide "fuzzy" memberships. A sample can be described by a mixture of topics via the probability distribution $P(\text{topic}|\text{sample})$ defined in Eq. A.6 and widely discussed in the next sections. The membership of genes into topics is even probabilistic, we have the $P(\text{gene}|\text{topic})$ (see Eq. A.7 for a definition) distribution that can be used to define a rank of genes based on their contribution to a particular topic. Moreover, these blocks have a hierarchical structure of L layers. Blocks of layer l, are nodes in layer $l+1$ (see [100] for details and illustrations). There are two extreme layers, by construction, one is composed by two blocks (one with all samples and one with all the genes) and one in which each block is a node.

We picked up a subset of GTEx: 1000 samples from the most represented tissues and 3000 highly variable genes were kept (see appendix A for details). We choose this as the experimental setup for this introductory analyses. We run hSBM and looked at the composition of the clusters, since hSBM is a generative model our expectation was that the algorithm reproduced the structure of the dataset recognising the different tissues of our set. In this setting its output had four layers with 1, 8, 29, 978 clusters and 22, 273, 2725 and 2880 topics respectively. Notice that the first partition involved only the gene side of the network: all samples were in the same cluster and genes were split into 22 topics. This could be driven by the asymmetry of the network, genes were three times the samples.

We choose the Normalised Mutual Information (NMI) as a measure of the clustering quality and to evaluate the performance of the model at each level of the hierarchy. See chapter A.4 for further details about NMI.

The score is estimated across layers and reaches two maxima: one if we consider tissues as ground truth and one if we consider sub-tissues. In other words, the tissue separation is better in the upper levels when the number of clusters is similar to the number of tissues, but the sub-tissues are better classified at deeper levels of the hierarchy. For instance in one layer a cluster is composed by *Brain*'s samples, in the next layer this is splitted and a cluster corresponding to the subtissue *Brain - Cerebellum* emerged; this particular separation of Brain samples was observed by [86] when the dataset was published. For other examples of tissues and subtissues separation we recall the Figure 3.2.

The clusters' tree and the scores confirm the first finding of this chapter: the algorithm is retrieving the hierarchical structure of the data. Never forget another advantage of this approach: hSBM is completely non parametric, the number of clusters and layers is detected autonomously by the model [100].

[1]These blocks can be set to be overlapping. We choose not to add this layer of complexity in this work since the results we obtained so far are still full of information.

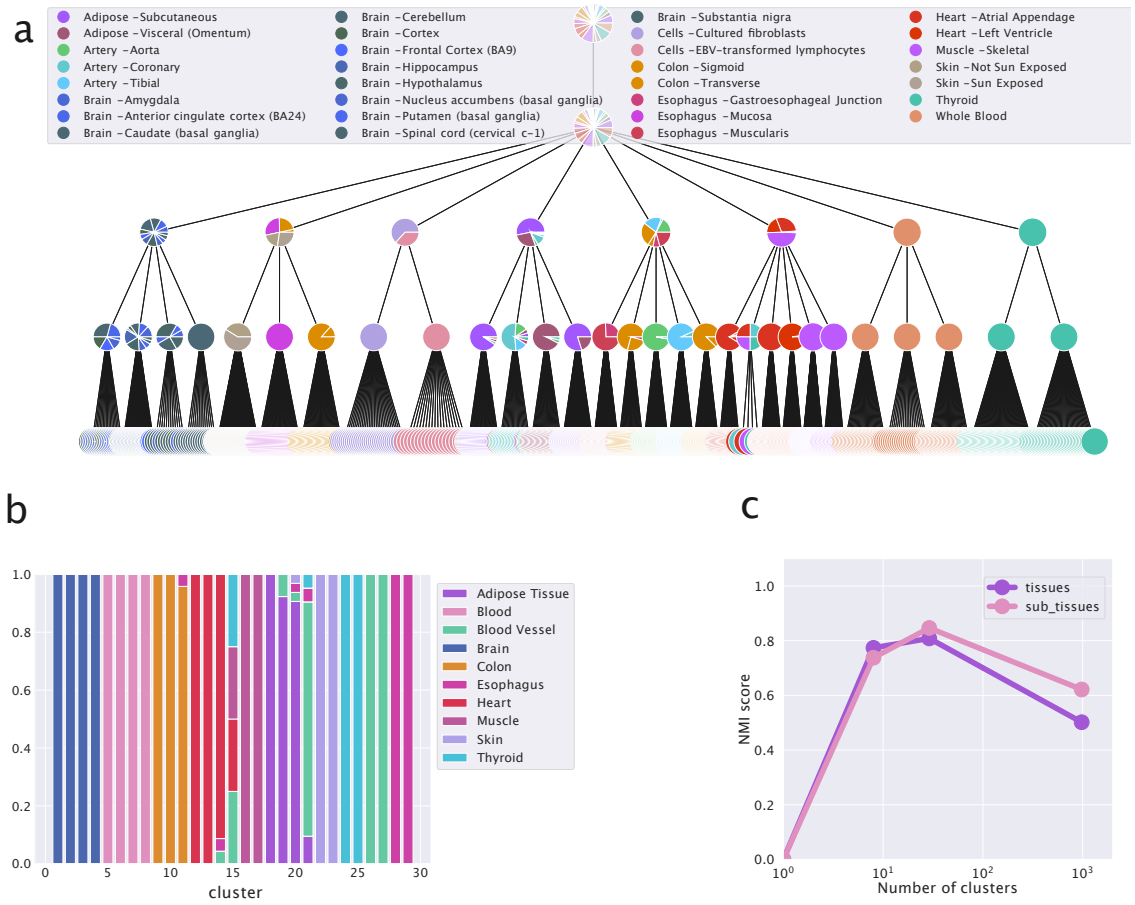


Figure 3.2: **hSBM retrieves tissue separation in GTEx data. (a)** Here we show the hierarchy structure in the form of a tree. The first layer from top is a all-documents node by construction. Then the algorithm found another node with all documents, at this point only the gene-side of the network is partitioned. Then the algorithm found 8 clusters with a rough separation of tissues. From left to right the reader can identify a Brain cluster, a cluster with Colon, Esophagus and Skin, a clusters with Cells, one mostly with Adipose tissue, one with Artery, one with Muscle and Hearth, one of Blood and in the end one with Thyroid. In the next layers the classification become finer, from Brain it emerged a Brain - Cerebellum cluster, Skin and Colon are separated, Cells are split into Fibroblasts and Lymphocytes, Aorta, Coronary, Tibial emerged, Hearth and Muscle are separated. **(b)** The composition of clusters at a given layer. Each column represent a cluster and columns are coloured due to the tissue of the samples. We reported the layer of the hierarchy with the best performances. **(c)** We estimated the *NMI* score considering both tissue and sub-tissue labels. This measures the accordance between the cluster annotations and the ground truth given by the samples' tissue. In the upper levels (when there are few clusters) the tissue separation is better than the sub-tissues' one; on the contrary the sub-tissue separation's performance is better at higher levels.

3.1 Data processing effects on different models

Gene selection

In order to investigate if different gene selections may affect the outcome of different models, because each model makes certain errors and, conversely, is able to correctly capture certain unique features [46]. We run some clustering and topic modeling algorithm with a random selected set of features (genes). In Figure 3.3 the Normalised Mutual Information obtained with different random gene selections. hSBM and LDA topic models seems to be the more stable (obtain an high score with low variance) models [105].

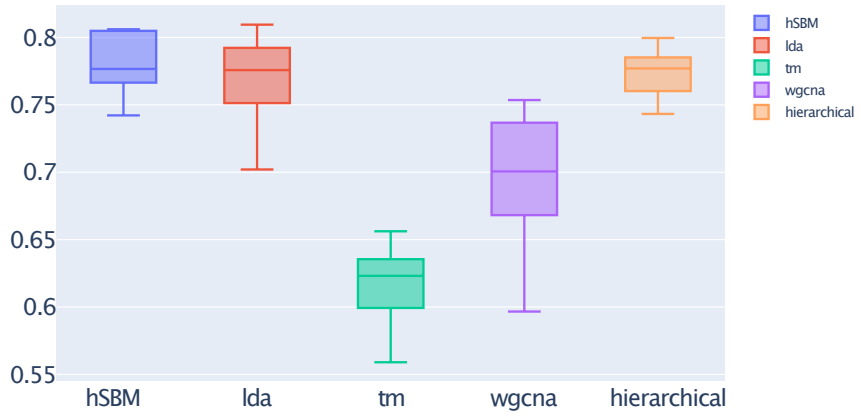


Figure 3.3: **Score obtained by different algorithms with random selected genes.** Box plot of max(NMI) per different algorithms when selecting genes at random.

Log transformation

It is a common approach in processing this kind of data to apply a log-transformation of the data. In Figure 3.4 the score obtained by the same model with and without applying a log transformation to the data before running the models themselves. It is pretty evident that topic models tend to be less susceptible to the pre-processing of the data.

3.2 The effects of priors in the model output: should we use markers?

When running a topic modeling algorithm each sample is described as a mixture of topics. We looked at the topic distribution of samples to investigate the topology of the new *topics' space*. To do this we started considering $P(\text{topic}|\text{sample})$, then we marginalised samples by tissue and obtained the $P(\text{topic}|\text{tissue})$. These two quantities carry the information about the relationship between the topics and the tissues. The goal is to identify how a topic mixture in a sample is related to its tissue.

In Figure 3.5a,c we show that hSBM tends to find a power-law like distribution similar for all the tissues. This means that samples of all tissues are described by a mixture of few important (high $P(\text{topic}|\text{sample})$) topics and a lot of more specific (low $P(\text{topic}|\text{sample})$) ones. Interestingly this distribution is similar for all the tissues.

LDA (3.5b,d) built the space differently and the distribution presents more peaks. In other words, topics are more tissue-specific and a sample's mixture contains almost a single topic. This is easier to be interpreted since the relationship sample-topic is trivial: a sample is described by almost a single

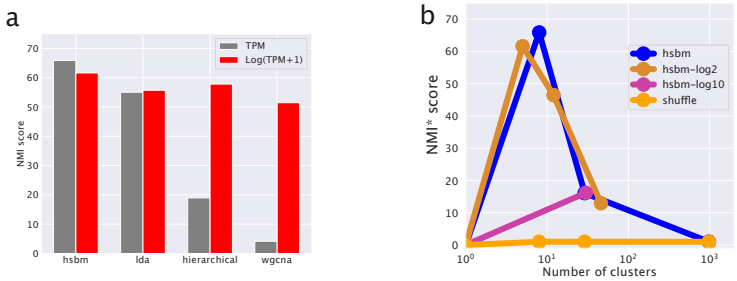


Figure 3.4: **Log normalization effects. (a)** Performances of different algorithms when running with TPM or $\log(TPM+1)$. **(b)** We ran the algorithm on the TPM dataset, on the log transformed $\log_2(TPM+1)$ and $\log_1 0(TPM+1)$ tables. With the $\log_1 0$ transformation we found just a single hierarchy layer. The scores are comparable. We report the score of a simple null model which preserves the number of clusters and their sizes, but reshuffles the samples' labels.

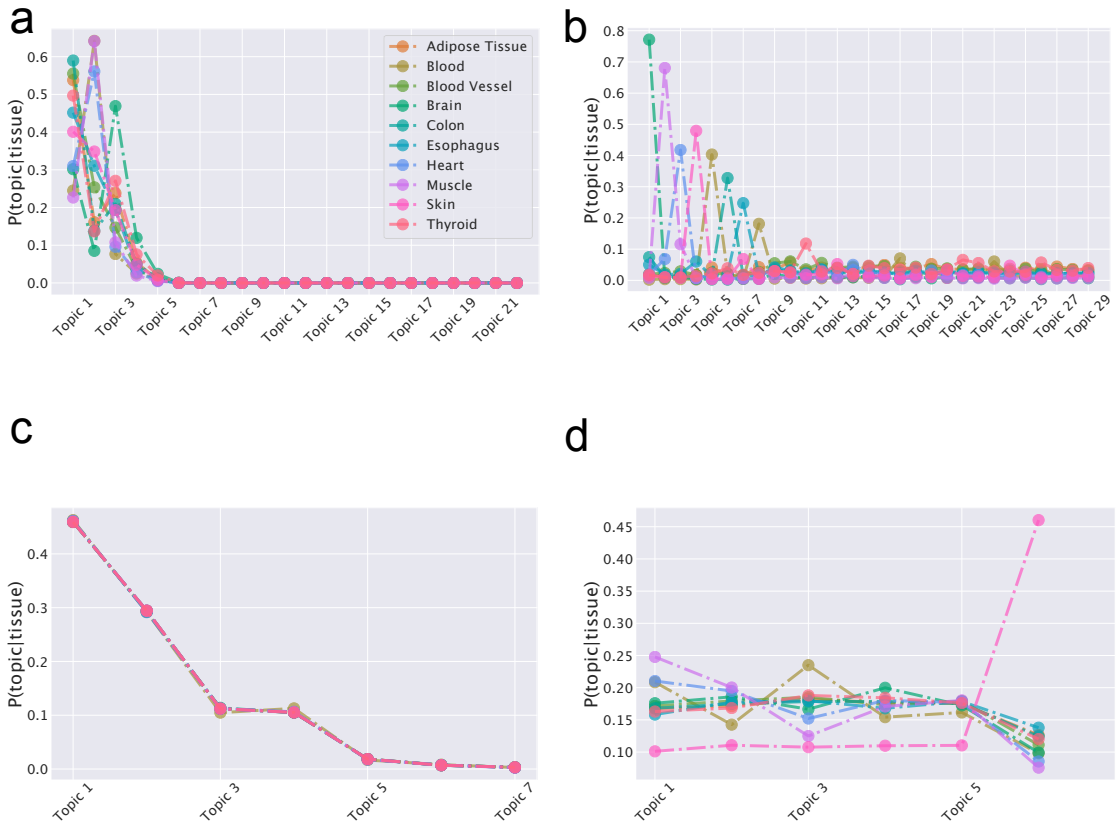


Figure 3.5: P(**topic**|**tissue**) **for different tissues reflects the role of the prior in the inference process.** In **(a)** the output of hierarchical Stochastic Block Model and in **(b)** the output of LDA In the setting with samples from 10 tissues of GTEx. Analysis with housekeeping genes with hSBM **(c)** and LDA **(d)** reveals the differences of the topic space output by the two algorithms.

topic. Nevertheless this behaviour has been forced by the algorithm and it has not been inferred from data.

Gene expression matrices present a power-law distribution in the abundances of genes [36]. This power-law is almost the same for all the tissues (see Figure 3.6). hSBM tends to reconstruct the

behaviour of the data in the inferred topic structure: this is one of the points that guided us to use hSBM in this biological context.

The same fact is observed if one measures the correlation between the average frequency in topics and topics' frequencies. The correlation is clear in hSBM (Figure 3.7a), but it is not, for instance, in LDA or other (see Figure 3.7b for example).

This is interesting biologically: hSBM avoids to find few, specific markers, but it searches for a combination of topics, similar in all samples and tissues. The fact that hSBM has an uniform prior and that the topic structure extracted by it is similar for different tissues suggests that this is a good approach to avoid over-interpretation of the output, in particular regarding the gene side of the network. In fact, if the algorithm injects its prior in the topics' mixtures it is more difficult to separate the relevant biological information from the algorithm biases.

Our point becomes even more clear when one considers housekeeping genes (Figure 3.5c,d). They are, by definition, highly expressed in every sample. So we do not expect that some genes are more differently expressed than others. In this situation the distribution of hSBM's topics is power-law like, the LDA's one have peaks induced by the Dirichlet prior or by other elements of the algorithm itself. Since there do not exist free lunches [115] a model which is good in a task is not the optimal choice for another task, one has to take into account as many algorithm priors as possible when looking at the results obtained with different models.

Finally, [121] suggested that many important genes are potentially ignored in research and the authors supported the hypotheses that an insufficient understanding of the biology of many disease genes has prevented the successful development of therapies. The fact that topics follow a global pattern rather than create a one-to-one relationship between topics and biological features, and the fact that topic modeling works on both random and housekeeping genes without requiring specific filtering could bring to light genes potentially ignored otherwise.

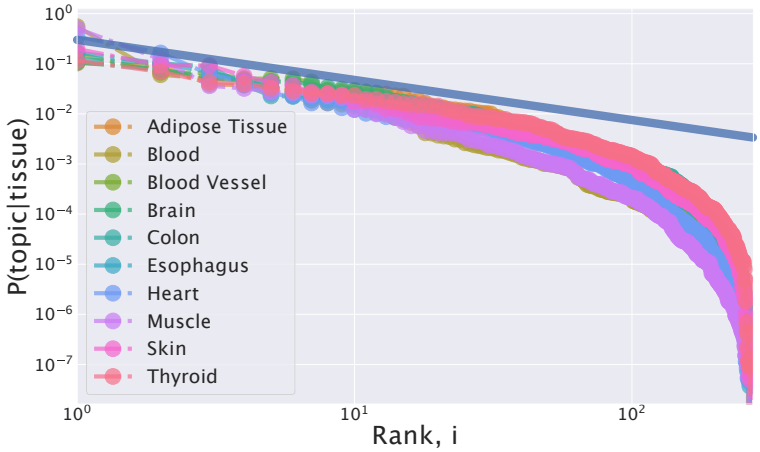


Figure 3.6: P(**topic**|**tissue**) **ranked for different tissues.** Topics are ranked for each tissue. The structure of the topic reconstructed by hSBM reflects the trend in the data (vlue line). This plot would have no meaning for LDA as it would be very similar to Figure 3.5b, being its peaked structure.

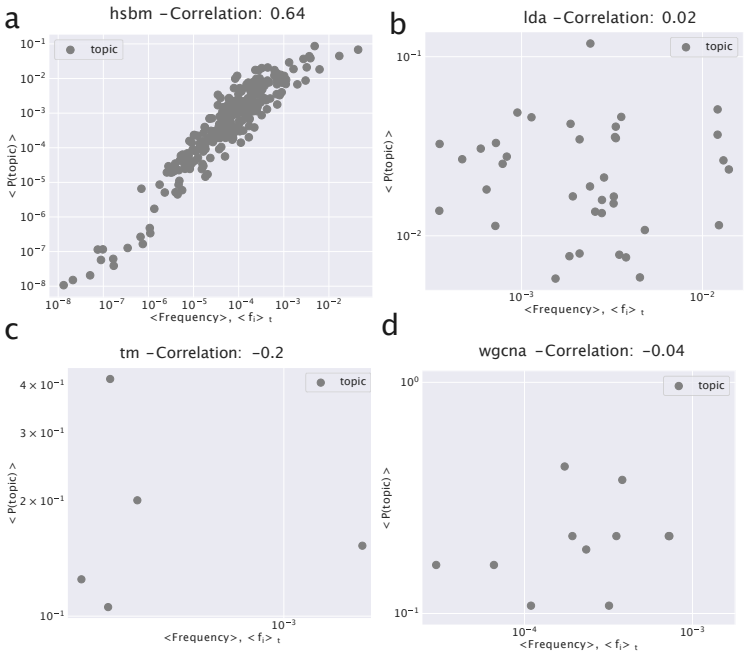


Figure 3.7: **Topic ditributions correlations.** In **(a)**, **(b)**, **(c)** and **(d)** we estimated the average frequency of genes in each topic $\langle f_i \rangle_t = \frac{1}{|t|} \Sigma_{i \in t} \Sigma_{s=1}^{R} n_{ij}/R$ where R is the number of samples, t is a topic with $|t|$ genes. We correlated this with the frequency of the topic $f_t = \frac{1}{R} \Sigma_{s=1}^{R} P(\mathrm{t}|\mathrm{s})$, it is the same measure in the case one considers topics as words. The different approach of the two algorithms is evident in this plots. The topic distribution of hsbm is correlated to the distribution of words in the original data. This recalls the observation, done by in the original paper [40], that topics have different *dissemination*.

3.3 Distances between samples are preserved when using hSBM

As reported originally by [86] there is a hierarchy of distances between tissues. To measure the distance between two tissues in the "topic space" we defined an archetype per each tissue averaging the $P(\text{sample}|\text{topic})$ over all samples of a given tissue. After this, we measured the distances of a tissue versus all the others. We did the same in the original gene expression space, so we ended up with two vectors of distances per each tissue one with the distances from all the other tissues in the original data space, one with the distances from all the other tissues in the topic space. At this point it is possible to measure the Spearman correlation between this two vectors per each tissue, this correlation would be 1 if the ranking of the distances of one tissue versus all the others is consistent after having projected samples in the topic space. As shown in Figure 3.8 using hSBM to project the data in a lower dimension space gives the most consistent results.

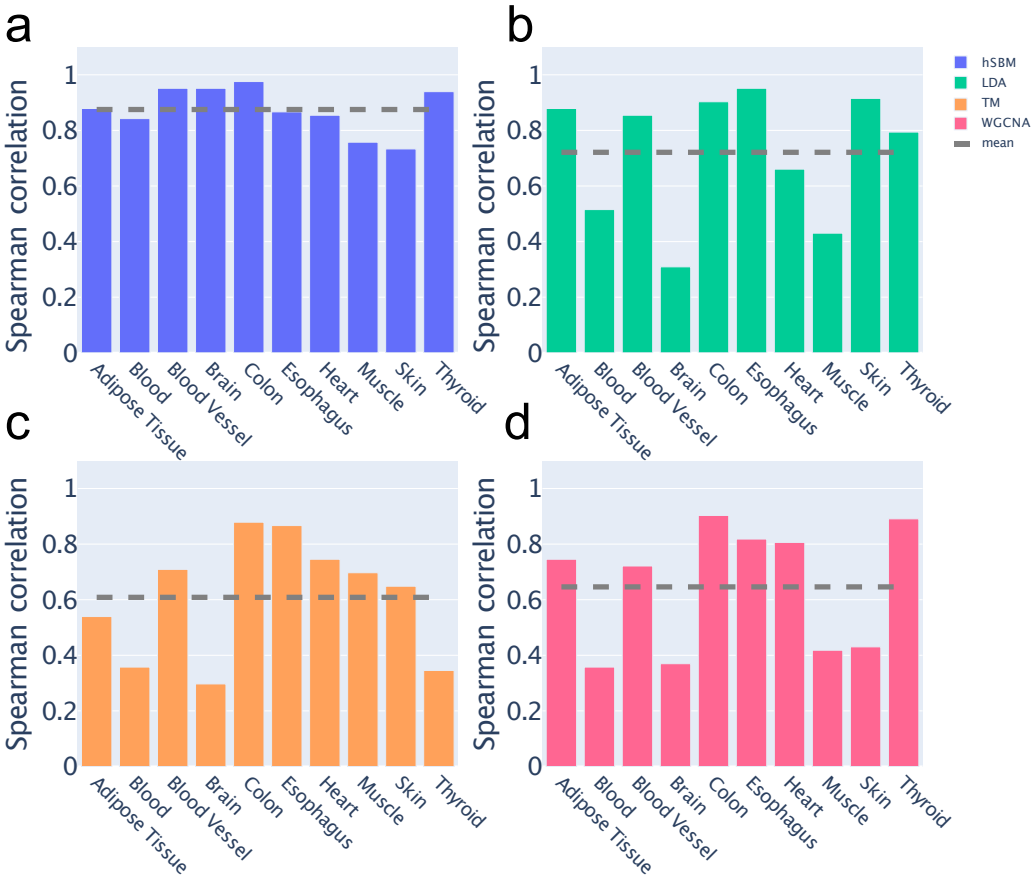


Figure 3.8: **Spearman correlation between vector of distance of one tissue versus all the others.** (**A**) hSBM, (**B**) LDA, (**C**) TM and (**D**) WGCNA.

3.4 Code availability

Code and notebooks to reproduce our analyses are available from https://github.com/fvalle1/topics[2].

[2]Topic and Clusters names (i.e. Cluster 1, Topic 1 may vary between the draft and the repository. We changed them for writing porpuses.)

CHAPTER 4

Topic modeling on breast and lung

In Section 3 we discussed how to apply hSBM-based topic modeling to healthy RNA-Sequencing samples, in this section we will describe the same approach on cancer data from The Cancer Genome Atlas (TCGA). Applying this kind of models on diseased data will, hopefully, shed light on cancer and will help the clinicians with new information about the patients.

4.1 Analysis of breast cancer samples

Breast cancer is the most common malignancy in women and one of the three most common cancers worldwide [11, 26, 51, 74]. It is also one of the few examples of a tumor for which there is a widely accepted subtype classification [106, 110, 120] based on gene expression that has a relevant therapeutic role and is instrumental for better clinical outcomes (in particular for HER2 subtypes). Breast cancer samples are usually divided in 5 different subtypes: Luminal A, Luminal B, Triple-Negative/Basal, HER2 and Normal-like. On the clinical side, this classification is based on the levels of a handful of proteins whose presence in the biopsy is usually detected using immunohistochemistry (IHC) assays. In particular, two hormone-receptors (estrogen-receptor (ER) and progesterone-receptor (PR)), HER2 (the Human Epidermal growth factor Receptor 2 (HER2) is a growth-promoting protein and plays an important role in several signaling pathways) and Ki-67 (Ki-67 is a nuclear antigen expressed by all proliferating cells during late G1 through the M phases of the cell cycle, peaking in the G2-M and with a rapid decline after mitosis and is thus an indicator of cancer cells growth).

Here is some information on these subtypes and their clinical outcomes.

- Luminal A breast cancer is hormone-receptor positive, HER2 negative and has low levels of the protein Ki-67. Luminal A cancers are low grade, tend to grow slowly and have the best prognosis.
- Luminal B is very similar to Luminal A from the gene expression point of view. The main difference is that it can be either HER2 positive or HER2 negative, and is typically characterized by high levels of Ki-67. As a consequence, Luminal B cancers generally grow slightly faster than Luminal A cancers and their prognosis is slightly worse.
- Triple-negative/Basal (which we shall simply denote as Basal in the following) are both hormone-receptor negative and HER2 negative.
- HER2 is hormone-receptor negative and HER2 positive. This class of breast cancers tend to grow faster than the Luminal ones and can have a worse prognosis, but they are often successfully treated with targeted therapies aimed at the HER2 protein.
- Normal-like breast cancer is similar to Luminal A: hormone-receptor positive, HER2 negative and has low levels of the protein Ki-67. However, its prognosis is slightly worse than Luminal A prognosis.

The same classification can be obtained (to a large extent [113]) looking at the expression levels of the well known "Prediction Analysis of Microarray (PAM)50" signature [97, 109]. Given the expression levels of these signature genes, samples are then classified using standard machine learning methods (Classification and Regression Trees (CART), Weighted Voting (WV), Support Vector Machine (SVM),

Nonnegative Matrix Factorization (NMF) or k-Nearest Neighbors (k-NN)) or using methods based on the euclidean distance in the signature space like Nearest Template Prediction (NTP) [61] or with more sophisticated network-based methods like Hope4genes [14]. The agreement among different classifiers and with the IHC-based subtyping is in general reasonably good but far from perfect. The classification task is made particularly difficult by the heterogeneity of cancer tissues (biopsies may contain relevant portions of healthy tissue) and by the intrinsic variability of gene expression patterns in cancer cell lines. For instance, the TCGA samples that we shall use for our analysis have been recently reanalyzed in TCGABiolinks [89] leading to a significant relabeling of samples.

To address this particular issue, we downloaded both the *PAM50* labels from [72], which is the most widely used set of annotations, and the more recent and highly curated *SubtypeSelected* annotation provided by the new functionalities of TCGABiolinks [89]. In the following, we shall compare the performance of our algorithms against both these annotations.

Our main goal in this framework was not to propose a new signature or a new classifier on top of the existing ones, but to show that it is possible to obtain relevant information on the cancer samples, like subtype annotation, the survival probability or lists of potential driver genes and altered pathways, without resorting to the marker genes mentioned above but looking instead at the overall gene expression pattern. We think this is an important achievement since it allows us to address breast cancer (and in principle any other complex pathology or cancer) without being influenced by the expression levels of few, often wildly fluctuating, marker genes, and opens the possibility to find new driver genes and possibly new subtype structures that may have therapeutic relevance.

Clustering of breast cancer samples

We performed the hSBM analysis on a bipartite network starting from all the 1222 samples of the TCGA-BRCA project on one side and a suitable selection of genes on the other side; the links were weighted by the expression values. We used also a set of other state-of-the-art clustering tools: Latent Dirichlet Allocation (LDA) [10], Weighted Gene Correlation Network Analysis (WGCNA) [76] and hierarchical clustering (hierarchical) [69]. We also compared the quality of clustering with the two annotations *PAM50* [72] and *SubtypeSelected* [89]. Table 4.1 reports the number of samples annotated for each subtype in the two annotation systems.

	PAM50	Subtype Selected
Basal	212	188
HER2	91	82
Luminal A	633	576
Luminal B	231	217
Normal-like	42	142

Figure 4.1: **Number of samples per each annotation in Breast cancer.** TCGABiolinks assigns more Normal-like subtypes.

On the gene side, instead of looking to cancer specific markers we selected, as mentioned above, only breast related genes, i.e., genes whose behavior was different in breast tissues with respect to other tissues (see Methods section for a precise definition). Results with the complementary choice of highly variable genes can be found in [127]. After this selection, we ended up with 978 genes. Among these genes, only HER2 among the classic markers discussed above was present.

hSBM finds a first layer of clustering in which the samples are divided into 8 clusters and the genes are organized in 6 topics. It identifies a further more refined level of organization composed by 29 clusters and 41 topics (the algorithm identifies then two further levels of partition with 149 and 1204 clusters and 147 and 399 topics respectively). These partitions on the cluster side convey little information and their score is very low. They correspond to the rightmost points in the Figure 4.3b,c.

Results are reported in Figure 4.3. Figure 4.3a,b report the subtype organization in clusters for the first two layers (8 clusters for Figure 4.3a and 29 clusters for Figure 4.3b). We see looking at these figures that hSBM is able to identify rather well Basal, HER2 and Normal-like samples, while it mixes Luminal A and B samples. For this reason, we shall treat them as a single subtype “Luminal” in the following.

We used the Normalized Mutual Information (NMI) measured compared to a null model as a score to evaluate the performance of the various algorithms in identifying cancer subtypes. The NMI scores are reported in Figure 4.3d for the comparison between different clustering algorithms and in Figure 4.3c for the comparison of hSBM results using the two different sample annotations.

Looking at the figures, we see that the highest NMI is reached for the first layer and that hSBM outperforms Weighted Gene Correlation Network Analysis (WGCNA), Latent Dirichlet Allocation (LDA) and hierarchical clustering. In order to set a comparison between different algorithms, the values of their several free parameters have to be selected. We chose the configuration of WGCNA, LDA and hierarchical clustering that could match more closely the number of topics and clusters obtained with hSBM. The rationale is to compare the different methods at the same resolution level (i.e., number of clusters and/or topics), thus at similar levels of dimensionality reduction. Therefore, it is possible that the algorithms could achieve better performances at different resolutions or using different performance metrics. This is for example the case of WGCNA. Setting WGCNA with different correlation threshold can improve its score but at the cost of producing in output a much larger number of topics and clusters with respect to hSBM.

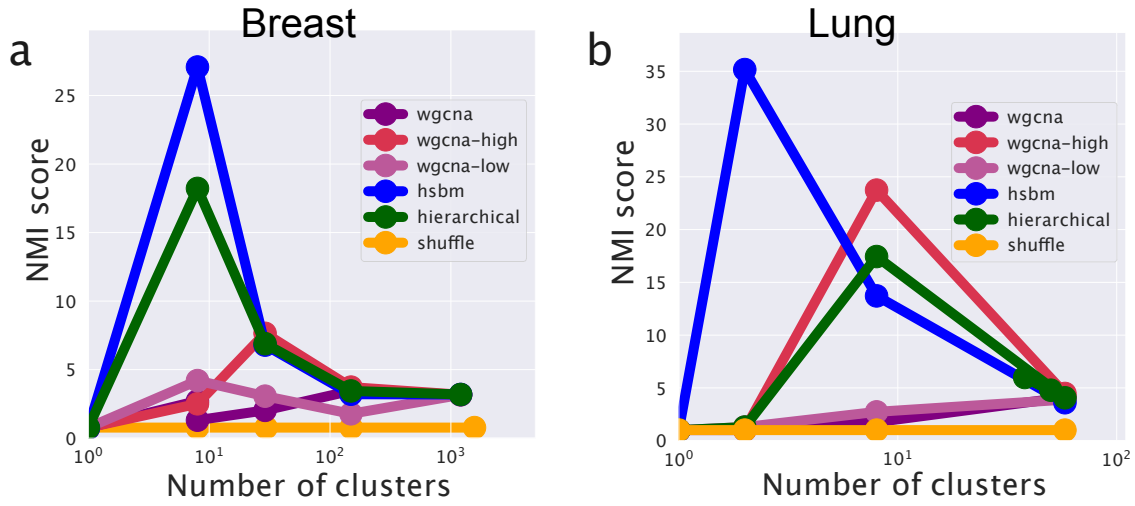


Figure 4.2: **Performance of WGCNA with different settings.** We ran WGCNA with different thresholds. `wgcna-high` represents a configuration in which WGCNA is set to find an high number of modules (topics): this case it similar to hierarchical clustering. `wgcna-low` represents a setting in which the algorithm is set to found few modules. The label `wgcna` represents the setting reported in the paper in which we set it to emulate the number of topics of hSBM (which searches the optimal number of topics itself). When WGCNA searches many modules, its outcome is similar to the hSBM (and of course to hierarchical clustering) one. Therefore, if WGCNA is set to replicate the resolution (i.e., the number of topics and clusters) of hSBM its classification performances are low.

Interestingly, we find a higher value of NMI at both resolution levels for the recent and more refined annotation of samples *SubtypeSelected* [89] with respect to the old one from [72]. In [89], the authors recognized several additional Normal-like samples thanks to an extensive effort to systematically quantify tumor purity with a variety of methods integrated into a consensus approach across TCGA cancer types. Indeed, the tumor microenvironment includes non-cancerous cells of which a large proportion are immune cells or cells that support blood vessels and other Normal-like cell.

Functional enrichment of the topics

Topics are nothing but lists of genes. A common way to investigate their properties is to perform enrichment tests using tools like GSEA [122]. The enrichment analysis on genes associated to subtype-specific topics finds functional categories that are precisely in agreement with the specific annotations of the subtype. For instance, the first entries of the table, corresponding to the Topic 1 mentioned above, show a strong enrichment for two sets of genes

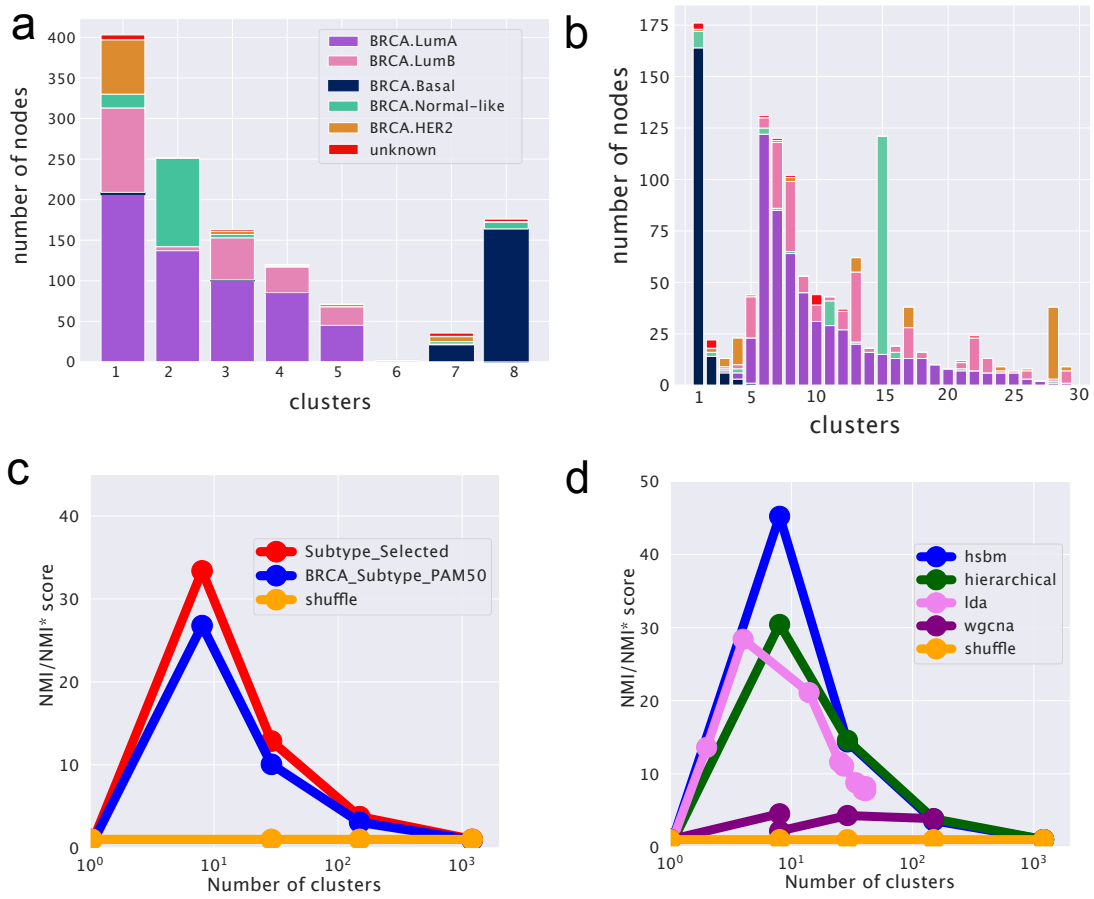


Figure 4.3: **hSBM result for breast cancer analysis.** In (**a**,**b**) it is reported the subtype composition of the clusters obtained via hSBM. Each column is a cluster, each color is a *SubtypeSelected* label from TCGABiolinks. The height of each column is proportional to the number of samples within the cluster. In (**a**) we report the results for the first layer of clustering (8 clusters) and in (**b**) those for the second layer (29 clusters). (**c**) Comparison of scores across hierarchy between TCGABiolinks *SubtypeSelected* labels [89] and TCGA labels from [72]. (**d**) Comparison of scores for different clustering algorithms. In (**c**,**d**) the Normalized Mutual Information (NMI) is scaled to the score obtained with a null model (NMI*). See Methods section for more details.

(labeled, following the GSEA convention as SMID_BREAST_CANCER_LUMINAL_A_UP and SMID_BREAST_CANCER_NORMAL_LIKE_UP) taken from [119] which fully agree with our subtype annotation in the topic space.

It is worth mentioning once again that in our analysis we selected only genes which are generically expressed in breast and not specifically differentially expressed in breast cancer. This makes the above results a non trivial consistency check of our procedure and further supports our idea of a role of the whole gene expression pattern of the cell in driving breast cancer subtype phenotypes.

Predicting breast subtype annotation

One of the advantages of a topic model approach is that it is also a dimensionality reduction process. Topics can be interpreted as new coordinates one can use to visualize and study the data.

We used the topic space as an embedding space to train a neural network model which can then be used as an efficient classifier to associate samples to their specific subtype. Using topics as features and *SubtypeSelected* as labels, our task becomes a simple supervised learning classification problem. The use of the topic space greatly simplifies the data space, and therefore the classifier can be trained much

faster and with fewer parameters. Moreover, we showed that the topics have a non-trivial biological meaning and this can help the classifier in identifying the relevant structures in a possibly noisy data set. We obtained a high accuracy classifier using a 399 dimensional topic space (the lowest level of the hierarchical organization of the topic space) starting from a space with almost 20,000 genes. Figure 4.4 reports in detail the performance of the classifier.

To evaluate the performances of our predictor, we performed the same analysis using K-Nearest Neighbors which is a standard and very popular tool in this context. It turns out that the performances of our predictor model (AUC = 0.98) are greater than those of k-NN (AUC = 0.90) on the same dataset. This tells us that the organization of the samples in the original gene expression space is not trivial, and that the projection into the topics space improves the ability of the predictor to assign the correct labels to test samples. We report the AUC score since it is less influenced by unbalanced classes than, for instance, the accuracy score. We applied K-NN using 5 n_neighbors and using the euclidean metric on the $\log_2(FPKM + 1)$ transformed data.

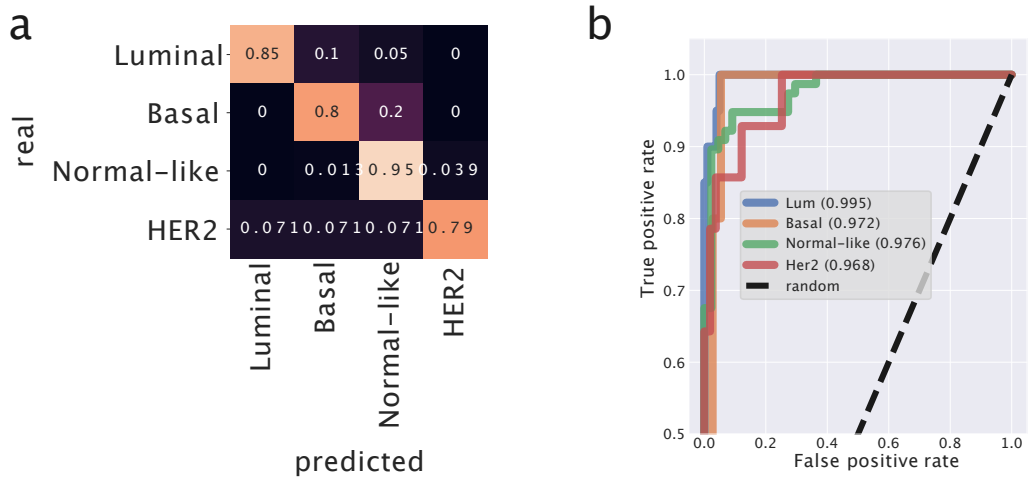


Figure 4.4: **Predictor model for breast cancer.** We built a neural network and trained it on the low-dimensional topic space to classify the different breast cancer subtypes. In (**a**) we report the confusion matrix. In (**b**) the Receiving Operation Characteristic curve and the corresponding Area Under Curve for each subtype estimated using a One-vs.-All strategy four times. The diagonal ($TPR = FPR$) corresponding to random guessing is reported for reference. *Luminal* and *Basal* subtypes are the ones with the lowest fraction of False Positives. *Normal-like* is the subtype with the highest fraction of True Positives.

4.2 Analysis of Non-Small-Cell lung cancer samples

To reveal the potentialities of topic modeling in a different context, we analyzed Non-Small-Cell Lung Cancer data taken again from TCGA. Lung cancer subtypes are currently defined by their pathological characteristics. The two predominant histological phenotypes of Non-Small-Cell Lung Cancer are adenocarcinoma and squamous cell carcinoma [20]. TCGA-LUAD and TCGA-LUSC projects provide transcripts for samples of these two subtypes. In the same way as in the breast analyses, we collected FPKM data with Genomic Data Commons' tools. In this case we selected 3000 highly variable genes (the second of the two options mentioned in the introduction). Results with the other choice, a tissue specific selection of genes, can be found in the Supplementary Material in Figure S4.

The binary choice (LUAD versus LUSC) represents a much easier task for a clustering algorithm and indeed, as we shall see, hSBM is able to correctly separate LUAD from LUSC. TCGA repository on lung cancer data allows for a non-trivial test of clustering algorithms. Indeed, Cline et al. in [23] observed that some samples from TCGA-LUSC have gene expression levels that are more similar to LUAD than LUSC, although their similarity to LUAD is modest. On this basis, they suggested that these samples may be borderline for subtype classification, for example because representing

tumors that are less differentiated and thus difficult to classify by pathology. The list of these samples, labeled as *Discordant LUSC*, is provided [23]. We analyzed how hSBM actually classifies these samples. Finally, in the context of lung cancer, thanks to the recent work of [132], we may perform another non trivial test of our topic modeling approach. We can combine together healthy and cancer tissues and look at the ability of hSBM to separate healthy samples from cancer ones.

Classification of Non-Small-Cell Lung Cancer subtypes: LUAD and LUSC

Running hSBM on the above data we found three different layers of resolution with 2, 8 and 58 clusters and 5, 12 and 42 topics, respectively.

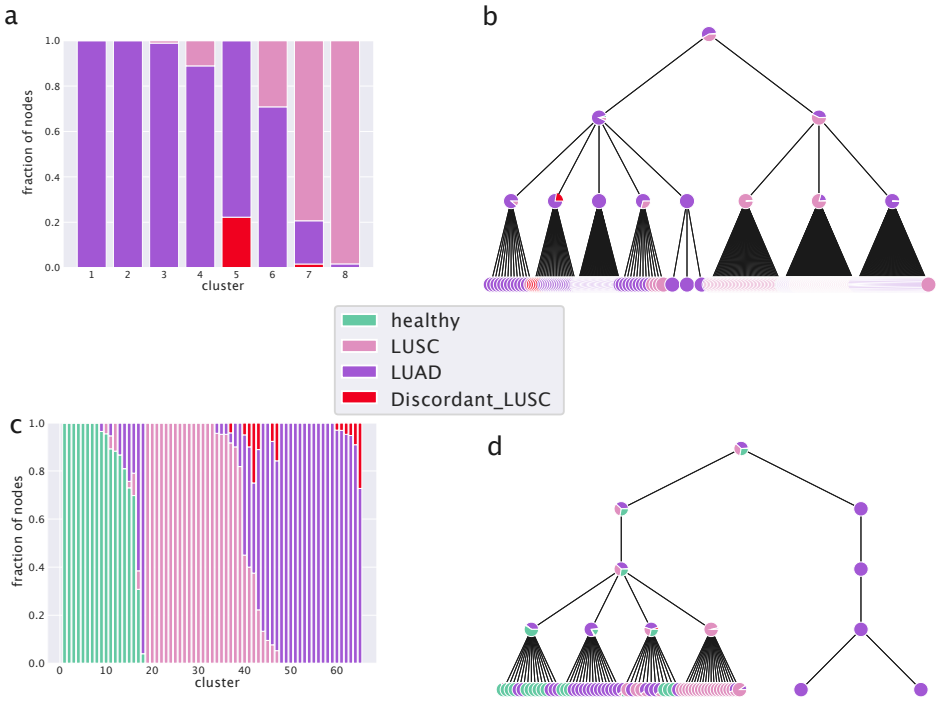


Figure 4.5: **The hSBM classification of Non-Small-Cell Lung Cancer subtypes.** In (**a**) the columns represent clusters and the colors refer to the sample annotations. Columns are normalized to the total number of samples in the cluster so that the height of different portions of the column are proportional to the fraction of LUAD (adenocarcinoma) or LUSC (squamous cell carcinoma) samples in that cluster. (**b**) The hierarchical structure. Already in the first layer, many LUAD samples are separated from LUSC, in the next layers the separation is complete. In (**c**,**d**) we report the results of hSBM analysis including also healthy samples. In both settings *Discordant LUSC* are classified with LUAD.

The results of our analysis are reported in Figures 4.5 and 4.6. In particular, Figure 4.5a shows that hSBM is able to separate well the two subtypes and that indeed most of the *Discordant LUSC* samples are clustered together with the LUAD ones, capturing the fact that they are more similar to LUAD than LUSC. It is instructive to follow the hierarchical organization of clusters (Figure 4.5b). Already in the first layer, many LUAD samples are separated from LUSC, while in the next layers the separation is complete.

Classification of LUAD and LUSC samples versus healthy tissues

We also tested the ability of clustering algorithms and in particular hSBM to separate healthy from cancer tissues. We downloaded data with healthy (taken from the Genotype Tissue Expression [79] GTEx project) and cancers samples provided in a unified framework by [131, 132]. We selected only

samples with valid metadata available from TCGABiolinks or GTEx. Looking at Figure 4.5c we see that also in this case hSBM identifies LUAD and LUSC subtypes and that both are separated from healthy samples. The majority of *Discordant LUSC* samples are clustered with LUAD as discussed before. Even if in principle the task of identifying three categories instead of two is harder, it seems that the inclusion of healthy tissues actually improves the performance of hSBM. Looking at Figure 4.6c, we see that the scores are higher than without healthy tissues.

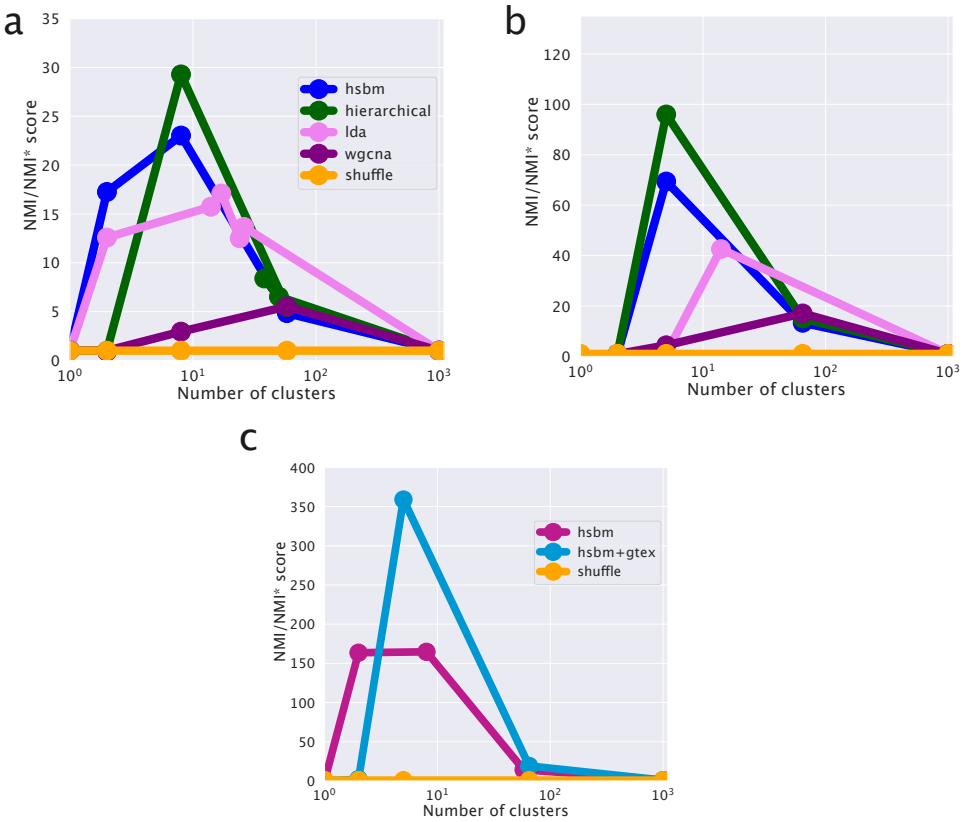


Figure 4.6: **Comparison of different clustering algorithms.** In (**a**) we report the scores for the classification of Non-Small-Cell Lung Cancer subtypes without healthy tissues. In (**b**) the scores in presence of healthy tissues. (**c**) The direct comparison between the case with and without healthy samples shows that their addition improves the cancer classification. Note that the score on the y-axis is normalized with respect to a case-dependent sample reshuffling (as explained in detail in the Methods section). This explains the different ranges of the scores in the panels.

Figure 4.6 shows that in the lung setting hSBM outperforms both LDA and WGCNA and it is compatible with hierarchical clustering.

As discussed in the breast cancer analysis, WGCNA is set to match the hSBM automatically retrieved number of topics and clusters. Note that WGCNA with very relaxed thresholds on the correlations becomes essentially equivalent to hierarchical clustering as it is shown in Figure 4.2. However, the relatively high performance score on the sample cluster structure comes at the cost of predicting a much larger number of topics (90) with respect to the 5, 12 and 42 topics retrieved by hSBM in the three different layers of resolution. A similar warning also holds true for LDA. In order to make a fair comparison with hSBM which has no free parameter, we used LDA with parameters set to their default values. In principle, LDA performances could be improved by suitably fine tuning its parameters, but such a extensive parameter exploration was beyond the scope of the present paper.

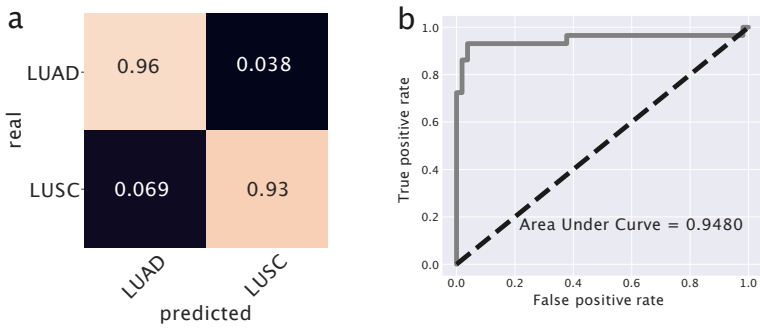


Figure 4.7: **Prediction model for lung cancer.** In (**a**) we used topics as features to train this model and we report the confusion matrix. In (**b**) the Receiving Operation Characteristic curve is reported. The Area Under Curve is reported as a score.

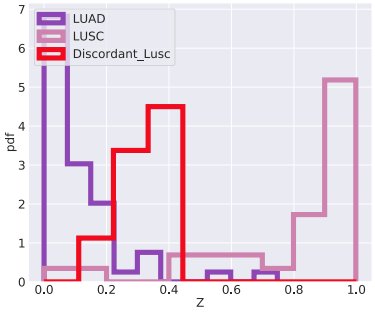


Figure 4.8: **Output of the last layer of the predictor.** Z is the output of the sigmoid function on the last layer $Z = \sigma(z)$; namely, Z is the probability of being LUSC.

Predicting LUAD and LUSC

The topic embedding space can be used to build a predictor analogous to the one we developed for breast cancer. In this case, the goal is to classify correctly LUAD and LUSC.

This predictor is actually a neural network with one hidden layer composed by 20 neurons and an output layer activated by a sigmoid function for the binary classification. We report in Figure 4.7 the output of this model on the test set. LUAD and LUSC are classified with high accuracy (accuracy: 0.9268, AUC: 0.9493). Additionally in this case, we compared our results with a standard K-NN predictor. K-NN achieves slightly better performances (accuracy: 0.9756, AUC: 0.9733).

The classifier we are using is inherently probabilistic since in output gives the probability that a sample belongs to a specific class. When the difference between tumor subtypes is not so clearly defined, but there is instead a continuum of possible cancer types, a careful analysis of the actual classification probabilities can be informative. This is the case for the classification of the *Discordant LUSC* samples mentioned above. Figure 4.8 reports the algorithm output Z, which should be interpreted as the probability of a sample to be of LUSC type. The *Discordant LUSC* samples have a score in the range 0.3–0.4 and interestingly seem to emerge as an intermediate peak in the classification probability distribution. They typically have a probability to be LUSC greater than standard LUAD samples, even if this probability is below the classic 0.5 threshold for LUSC classification.

4.3 Code availability

The codes, notebooks and data to reproduce this work are available on a GitHub repository at https://github.com/fvalle1/topicTCGA.

CHAPTER 5

nSBM: my original branch to the problem

When hSBM was released, its authors [40] demonstrated that topic modeling [57] can be seen as a problem of community detection [34, 35] on bipartite networks. We propose to apply the same principles but in a context where multiple information are available and documents are not longer linked only to words but they are also linked to other sources of information.

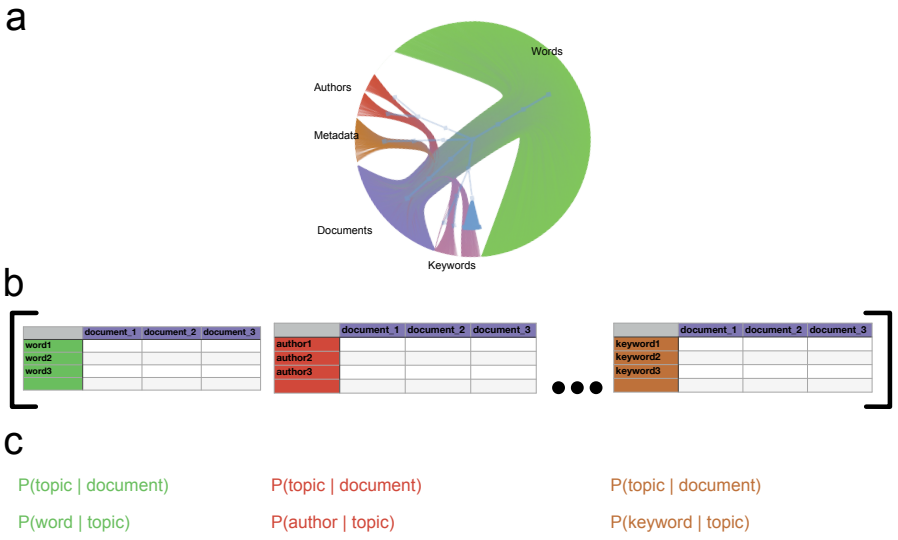


Figure 5.1: **Schematic representation of the software.** (**a**) A n-partite network is build with documents, words and other layers of information. (**b**) This is nothing but the representation of n Bag of Words. (**c**) After running Stochastic Block Modeling simultaneously on the whole network, one has in output we topic model on each layer. In particular $P(topic|document)$ and $P(\{word, keyword, author \ \ldots\}|topic)$.

The idea of adding a layer of metadata to networks or to a graph of drug-drug interaction was proposed by [33, 62] in the context of using SBM to reconstruct links, in this work the idea is not to reconstruct links, but we were focused on finding communities and eventually taking advantage of the fuzziness given by the topic modeling interpretation.

We pictured in Figure 5.1 our approach to the problem: starting from multiple Bag of Word-like matrices, we built a n-partite network. The output consists of $P(topic|sample)$ and $P(word|topic)$ estimated at each branch of the model. Moreover, the clustering is performed concurrently for each kind of data and similar (of the same kind) nodes are grouped into *blocks* we call clusters (for samples) and topics (for every other kind of data). We ran a nested version of the model and so all of this information is available at different layers of resolution.

The fitting procedure is performed taking advantage of hierarchical Stochastic Block Model (hSBM). hSBM, as widely discussed in this thesis, is a kind of generative model that tries to maximise the

probability that the model describes the data

$$P(\theta|\mathcal{A}) \propto P(\mathcal{A}|\theta)P(\theta).$$

Its approach is completely non-parametric and it aims to maximise the posterior probability $P(\theta|\mathcal{A})$. We used the `minimise_nested_blockmodel_dl` function from graph-tool [99] to minimise the Description Length $\Sigma = -\log P(\mathcal{A}|\theta) - \log P(\theta)$ in a nested version of the model [100, 101, 102, 103, 104]. Minimising Σ equals the maximisation of the posterior probability. In our setting $\mathcal{A}$ is nothing but a block matrix in which each block is a Bag of Tokens (i.e. words, keywords, authors ...). It can be seen as a union of two dimensional matrix with entries $\mathcal{A}_{ij}$: $\mathcal{A}_{ij}$ is the number of elements i in document j. We set the model to do a sort of model selection minimising the Description Length Σ multiple (10) times and we choose each time the model with the shortest Description Length.

Different kind of information on real-world datasets

Datasets have metadata. In this work we focused on a bunch of texts datasets of very different kind such as: the papers of the American Physical Society (APS), PubMed Central to Reuters articles, Wikipedia pages and tweets. When the full text was not available we built the Bag of Words [2, 52] from the words in the titles and in the abstracts.

Each dataset has its kind of *metadata* to add information to the document, one of the most common is represented by the keywords entered by the authors themselves. Twitter doesn't provide keywords, but we selected *hashtags* which actually plays a similar role. In the Wikipedia example we considered the so-called *categories* as keywords.

See section A.1 for further details on the datasets utilised here.

We wanted to benchmark our model performance and to do so we used two different approaches described below.

Which model is compressing more the data? was the first question we wanted to answer. If one has to compare the behaviour of two models on the same network the right tool to measure the performances is the Description Length Σ [44, 101, 112] (see also section A.4). Σ represents how much information (i.e. the number of bits) is needed to encode both the network and the model parameters. If we compare the Description length for the same network but different models, this tells us which model is compressing more the data.

In the setting discussed in this work, we are comparing different models (actually SBMs with different topologies and number of parameters) on two different networks (every network is built with $2, 3 \ldots$ branches). In this case, since Σ is a, non linear, increasing function in the number of edges E the networks with more branches had trivially a greater Σ. In order to have a more significant measure we considered the Description Length per Edge ($\frac{\Sigma}{E}$); this is not even ideal since Σ is not linear in E (see Figure 5.7), but it could give us some hints on which model is compressing more the information per edge.

In Figure 5.2a we reported $-\Delta\left(\frac{\Sigma}{E}\right)$ which tells us the improvement on using the tri-partite model with respect to the bi-partite hSBM. Since $\Delta\Sigma$ is the ratio of two posterior probability, this is similar to the estimation of the Bayesian factor [70] if one compares the two models. These $-\Delta\Sigma$s are all positive, this means that the model with three branches compresses better the information per edge and there is a gain in the posterior probability in using this kind of model.

It is possible to estimate $\Lambda = \exp^{-\Delta\Sigma}$ in the TCGA example (the one with the highest $-\Delta\Sigma$) it is $\Lambda \geq \exp 0.5 \geq 10^{0.21} \approx 1.6$. This means that nSBM description of each edge is almost two times more probable than the model without miRNAs. Even if this is not an impressive improvement, we are using information already available and we are also improving the classification. We would like to suggest to use any source of information available that can lead to an improvement of existing models.

Measuring the Description Length is not the only way to benchmark this models.

How good was the partition in clusters? We compared the output of the model with a given ground truth for each dataset. In the databases of research papers we considered the journal as a good

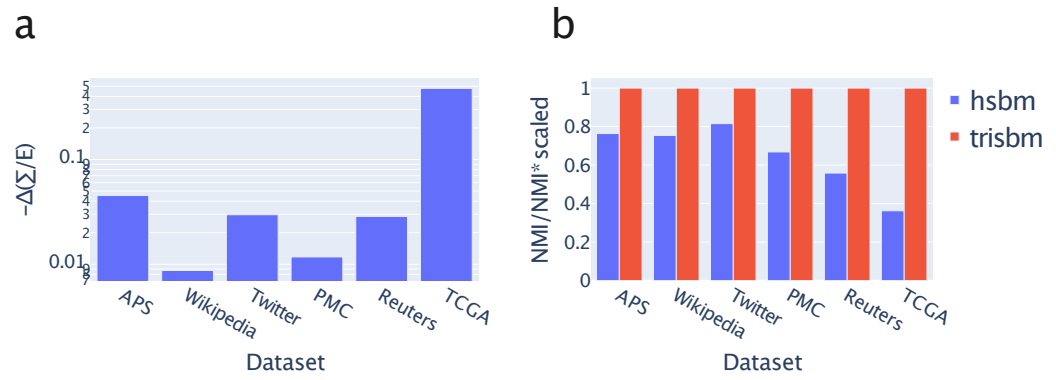


Figure 5.2: **Studying the gain when adding multiple information in different datasets** (**a**) Difference in Description Length per edge E between hSBM and nSBM $\Delta\Sigma$. This is the gain obtained adding other kind of information. (**b**) NMI/NMI^* scaled to nSBM for different datasets.

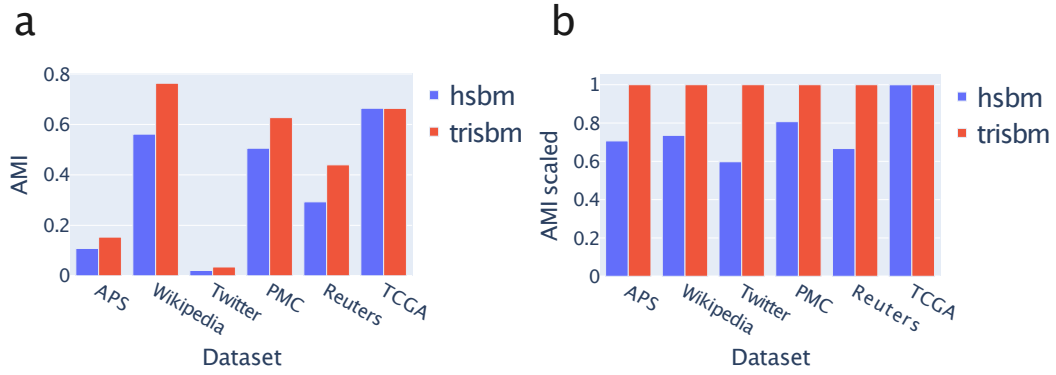


Figure 5.3: **Adjusted Mutual Information score as an alternative to NMI.** (**a**) Adjusted Mutual Information scores for different dataset. (**b**) scores scaled to triSBM to highlight the gain one has appending additional layers of information.

variable to define the topic of the papers. In Reuters and Wikipedia we downloaded the topic of the document and the Twitter dataset come with a label denoting its positive or negative attitude.

In Figure 5.3 we reported the Adjusted Mutual Information between the partition in output by models and the ground truth of each dataset. In the Figure 5.3 we reported also another score the Adjusted Mutual Information (it compares the Mutual information between the model's output and the ground truth with the mutual information between two random partitions) which confirms the results obtained with the NMI.

If we compare the maximum NMI/NMI^* reached by each model, as shown in Figure 5.2b we have another insight that adding layers of information helps the classification of the documents.

5.1 Synthetic datasets

We tested this topic model approach with multiple kind of information also using synthetic data. In this case we have full control over the properties of the data.

Our approach for the generation of synthetic data is based on what was done by [84] when adding a supervision layer to Latent Dirichlet Allocation [10]. For each document a variable θ_d, like in LDA, describes the topic distribution (Ptopic|document)), from this it is sampled, from a Dirichlet

Figure 5.4: **Supervised topic models are better even when network-based.** -Log-Likelihood obtained with different algorithms. Latent Dirichlet Allocation (LDA) and its supervised version (sLDA) are compared to their network-based concurrent. hierachical Stochastic Block Model (hSBM) and nSBM (this work).

distribution, a variable z_{id} which describes the topic assignment of word i in document d. For each document d, a single rating Y_d is drawn taking into account the average topic assignment across words ($\bar{z}$).

In this work we extended this in two ways: first we consider the possibility of adding tokens of different kinds (e.g. keywords, authors ...) representing additional layers of information; moreover, we assume that more than one token can be assigned to a given document at each layer.

We ran hSBM [40] and nSBM (this work) on this kind of data and obtained the results reported in Figure 5.4. The approach of adding different sources of information lead to a gain in the performances of the models, as [84] demonstrated that sLDA can improve LDA [10], here we show that adding branches is an upgrade of the bipartite hSBM. These network-based models, as already demonstrated by [40], are themselves an improvement of the LDA-based algorithms.

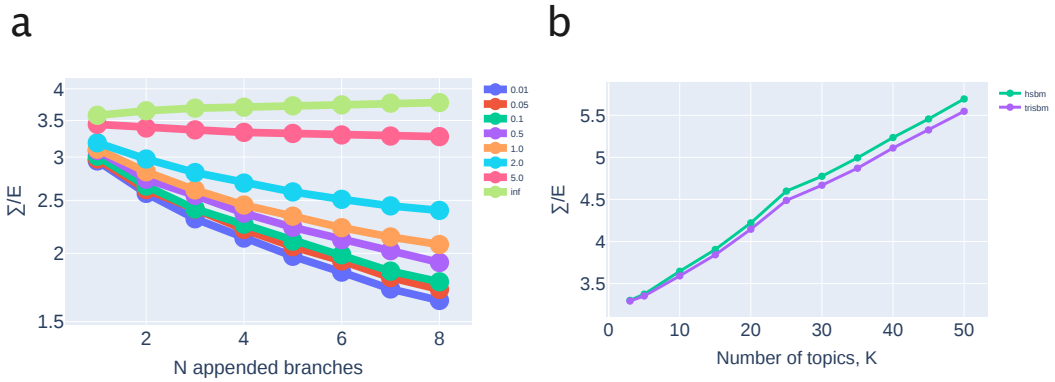


Figure 5.5: **Studying description length Σ appending branches or topics.** ((**a**) Description length (per edge) Σ/E adding more and more synthetic branches. The new branches are sampled with a different parameter σ giving them a different correlation with the average document topic.)The Description Length (per edge) as a function of the number of planted topics K. (**b**) Description length (per edge) Σ/E in configuration varying the number of topics K.

Layers may bring different level of information

We have shown that adding certain layers of information can help topic modeling to find more informative partitions. As pointed out recently by [33] there are regimes in which metadata completely dominates the inference process and situations in which metadata does not play any role.

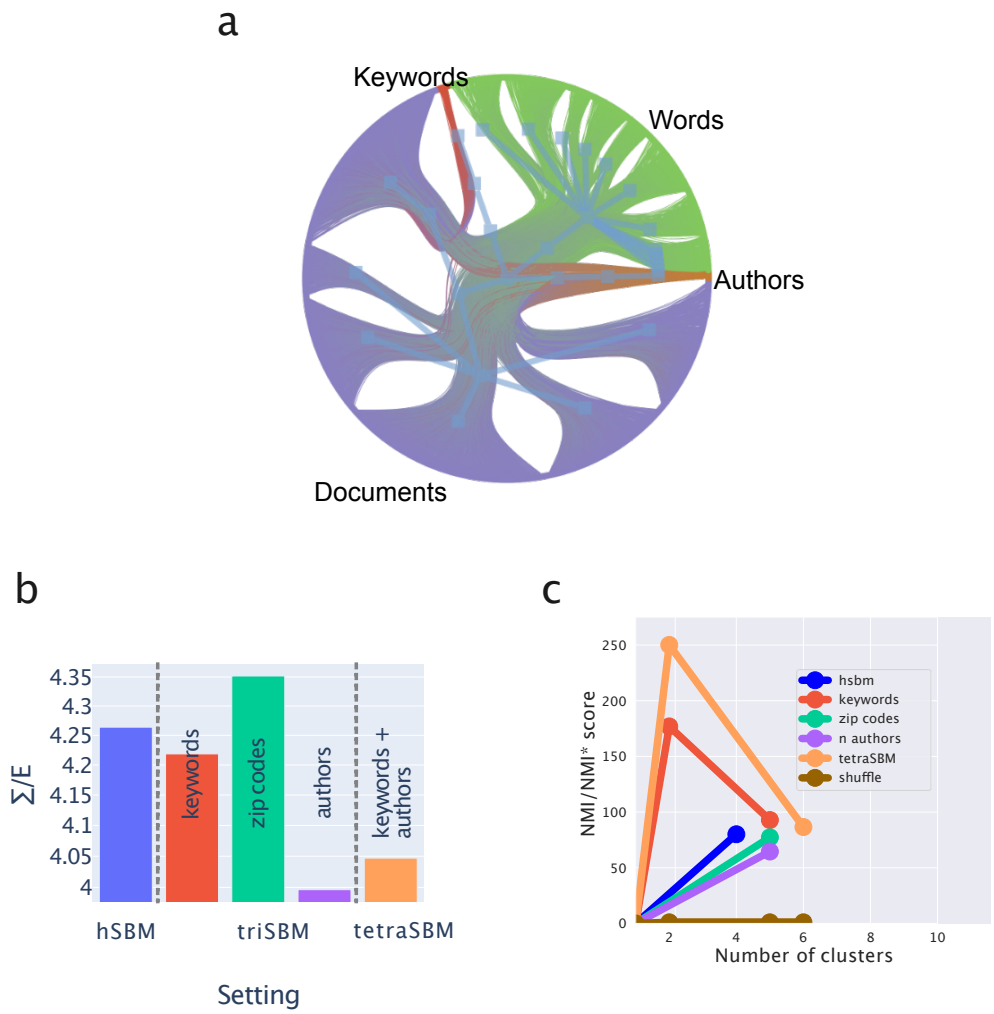


Figure 5.6: **Making bi and tri-partite networks can improve the classification in real data.** (**a**) A real world example of different setting when different sources of information are considered more than the Bag of Words. Tetra-partite network of APS where documents are connected to words, keywords and the number of authors. (**b**) Description length per edge Σ/E in this three different settings. (**c**) NMI/NMI^* score for the different layers in the titles from American Physical Society. We compared a network made with a classical Bag of Words (hSBM), adding a layer with keywords (triSBM), adding a layer with institution's zip codes, authors' number and, finally, adding two layers: one with keywords and one with the number of authors (tetraSBM).

Again, we investigated how new branches can affect the learning process. On a real data example, adding different kind of metadata may have a different impact on the performances of topic model. We show in Figure 5.6b that adding zip codes (of the authors' institution), keywords or the number of authors of a paper have a different impact on the learning process.

We consider the number of authors as a feature despite authors themselves, we expected that only certain fields are characterized by big collaborations, for instance. This was done since we choose a subset of recent papers, and there were few authors with more than a paper published on APS in this short time window.

Moreover we tested the same question on synthetic data whose branches had different grade of noise. In this case we are able to generate branches with tokens related to the planted topics. In fact, $Y_{d,b,j}$ (the j-th token of document d in branch b) is sampled from a normal distribution $N\left(\vec{z}_d\eta, \sigma^2\right)$. If σ^2 is small the tokens on additional branches will carry the exact information about the topic the document belongs to. If σ^2 is big there is a weaker relation between tokens of the new branch and the planted topics, we would expect this kind of new layer to be less informative.

Figure 5.5b reports our results on synthetic data. When branches are added, the Description Length Σ (per edge) decreases. Moreover, if the appended branch is "noisy" ($\sigma^2 \gg 0.1$) there is not a gain in Description Length (*inf* line in Figure 5.5a).

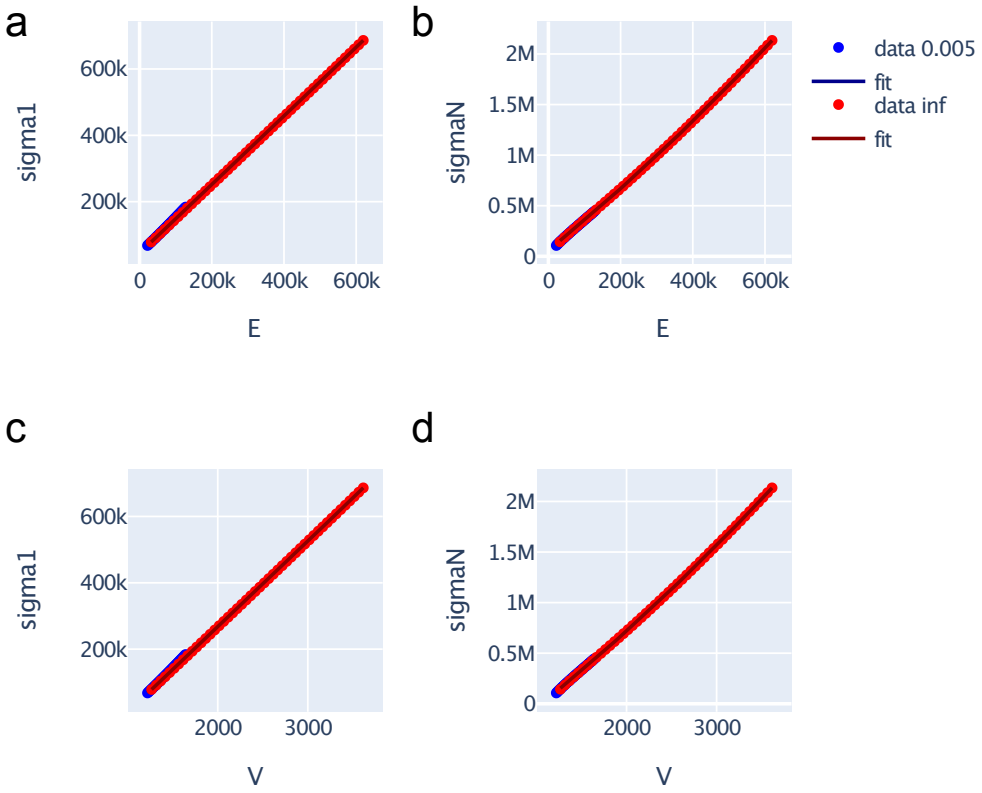


Figure 5.7: **The description length of a trivial network scales more than linearly with E and V.** Scaling of the description length of trivial partitions with the number of vertices V and the number of edges E.

5.2 Description of the software

Together with this work we released a Python package easy to install and useful to perform SBM topic model on n-partite networks.

This inherits the code provided by [40], available from https://github.com/martingerlach/hSBM_Topicmodel, and adds the methods to extend their topic modeling analyses simultaneously to each branch of a n-partite network. We passed the gene expression matrix as a weighted graph to their model.

In this case the graph required to running the model is created using a list of Bag of Words [2] (represented as pandas DataFrames [85]). Each Bag of Words corresponds to a different layer of information as shown in Figure 5.1a.

Once one has installed the package, it is possible to fit, for instance, a tetra-partite network with words, keywords and authors as shown below.

```
from nSBM import nSBM
model = nSBM()
model.make_graph_multiple_df(df, [df_keywords, df_authors])
model.fit()
model.save_data()
```

The output is the same as hSBM; its output-files contain: $P(topic|document)$, $P(word|topic)$, clusters (blocks of documents) and topics (blocks of words) at each layer of the hierarchy. Running nSBM this kind of output is produced for each layer; in this case each branch is processed separately as in the scheme represented in Figure 5.1c.

The package to run nSBM [126] can be downloaded from GitHub (https://github.com/BioPhys-Turin/nSBM) or, alternatively, it can be installed using Anaconda (https://anaconda.org/conda-forge/nsbm) running `conda install nsbm -c conda-forge`.

The code, which itself uses the aforementioned package, to reproduce the analyses in this work is available at https://github.com/fvalle1/multiSBM.

CHAPTER 6

Multi-omics topic modeling

It is, by now, well-established that miRNAs play an important role in several human diseases, particularly in cancer. Accordingly, miRNAs have been proposed as diagnostic biomarkers of human cancers [13, 17, 53]. This is particularly true for breast cancer, for which several studies have highlighted the prognostic role of miRNAs [8].

Following this line of evidence, we integrated miRNA expression levels with protein-coding mRNA levels using a $n = 3$ version of nSBM (which, in the following, we shall denote as triSBM). In this case, the analysis output, besides the clusters of samples and the topics of genes, will also contain a collection of miRNA-topics. In Figure 6.1 we pictured a bi- and a tri-partite network.

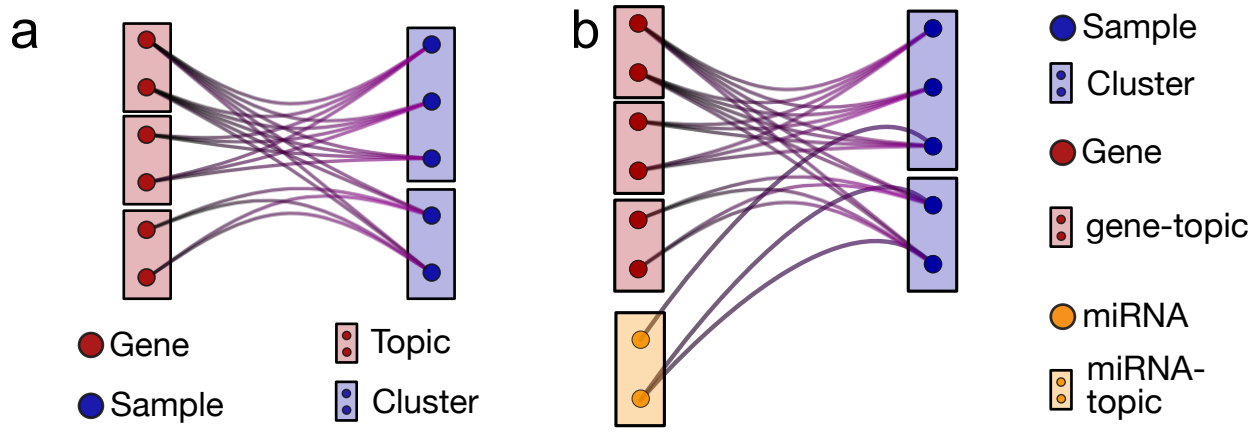


Figure 6.1: **Cartoon of multipartite networks with samples, protein-coding genes, and microRNAs.** (**a**) A bipartite network with a layer of protein-coding genes and a layer of samples. A gene is connected to a sample if it is expressed in that sample and the link weight is proportional to the expression level. (**b**) A tripartite network obtained by adding the miRNA expression layer. The topic model algorithm essentially outputs a block or topic structure in each layer.

6.1 The inclusion of miRNAs in the topic modeling analysis leads to a better separation of healthy and tumor tissues

We tested the ability of the algorithm in recognizing healthy from cancer samples. The hSBM algorithm showed good performances on this task by considering only gene expression dataas in Section 4 (and Ref. [127]), as summarized in Figure 6.2b. we then tested triSBM, in which gene expression levels were considered jointly with miRNA levels in the same set of TCGA samples. The detailed procedure and the algorithm output at different hierarchical levels are described in the Methods section. We found a significant improvement in the performance of the algorithm. In fact, Figure 6.2a clearly shows that normal samples are collected in a single cluster by triSBM, while the separation is less neat in the absence of information on miRNA expression (Figure 6.2b).

The two model settings (hSBM and triSBM) are compared quantitatively in Figure 6.3 using Normalized Mutual Information (NMI) as a score [114, 117]. The NMI score is explained in detail in the Methods section.

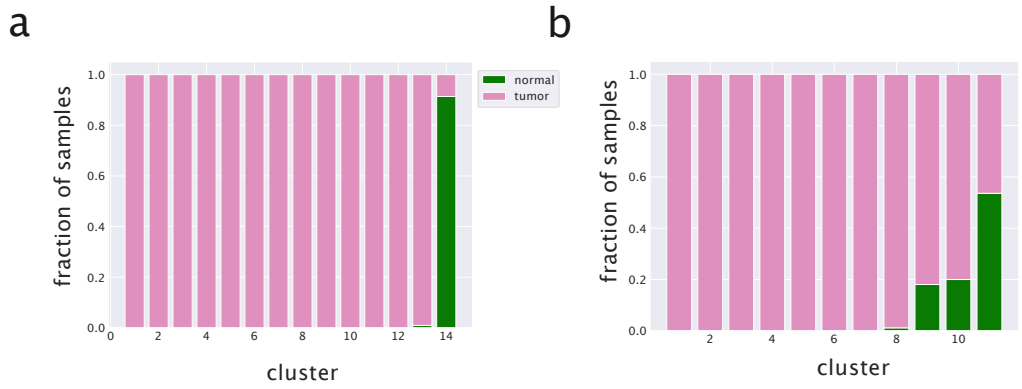


Figure 6.2: **Clustering of breast samples with and without the miRNA branch.** we compare normal and solid tumor tissues from TCGA using (**a**) triSBM and (**b**) hSBM at a similar resolution level.

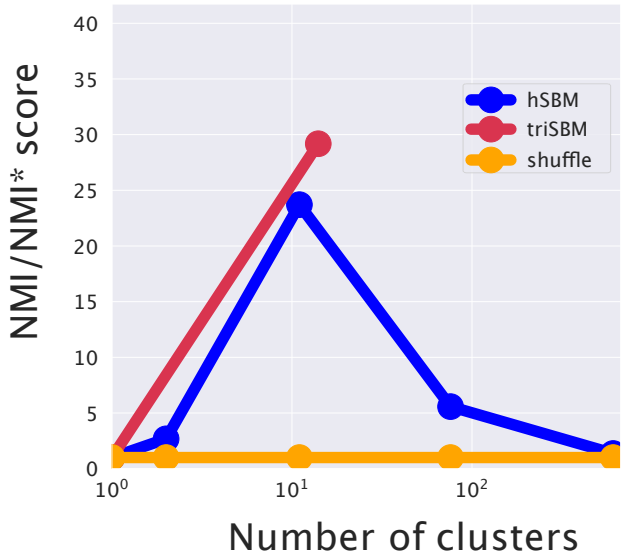


Figure 6.3: **The increase in performance when separating tumor and normal samples by the addition of the miRNA layer.** The NMI is evaluated at different resolution levels (numbers of clusters) using (triSBM) or not using (hSBM) the information of miRNA expression. The normal/tumor annotation from TCGA is used as ground truth.

6.2 Validation on an independent source of data: METABRIC

We applied the same pipeline applied on TCGA to the METABRIC [27] dataset and measured the agreement between our partition on this data and the labels provided by [60]. we confirmed the results obtained on TCGA: the triSBM model has a better agreement (NMI score is reported

in Figure 6.4) with the labels assumed as ground truth with respect to the model without miRNA (hSBM).

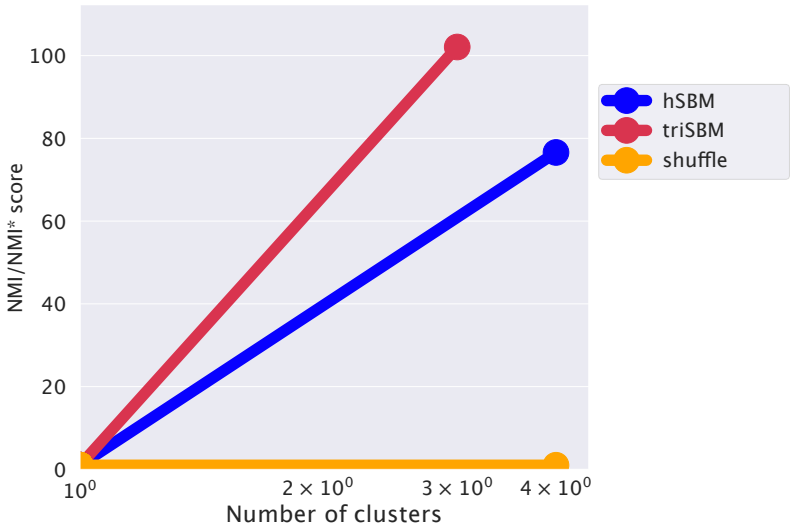


Figure 6.4: **Validation on METABRIC dataset.** we run hSBM bi-partite model and triSBM model including both mRNA and miRNA data on METABRIC dataset. As reported here the improvement given by the introduction of an additional feature can be seen also on this dataset.

6.3 Including regulatory interactions in the TriSBM framework

MiRNAs exert their biological function by regulating target genes at the post-transcriptional level. It is thus of great importance to be able to include this information in the topic modeling analysis. This is not an easy task, since miRNAs act in a combinatorial way: typically, several miRNAs cooperate to regulate a single target gene; at the same time, a single miRNA can regulate hundreds of targets. Moreover, while the standard miRNA–target regulatory interaction is of inhibitory type, it sometimes happens that a miRNA can have a widespread (indirect) activatory role by interfering with a repressed epigenetic pathway. These are the so-called "epi-miRNAs" [94, 111] that have been recently shown to play an important role in cancer development [111]. Keeping track of these interactions can be of crucial importance to correctly decode the information contained in the miRNA expression data. To this end, one can make use of a few specialized databases of miRNA–target interactions. In particular, in the following, we shall use MirDip [124] and TarBase [96], which are among the most popular ones and are somehow complementary in their target selection choices.

To integrate the regulatory information, we made use of the analogy of this problem with inclusion of the citation information among documents in standard topic modeling applications to texts [63]. In our case, the additional links are not between samples (as it would be a citation link or a hyperlink); therefore, for links between branches in particular, We added gene–miRNA links.

We ran the tripartite model as described before; then, in a second moment, we added links gene–miRNA from regulatory network (We tested separately MirDip [124] and TarBase [96]), as shown in Figure 6.5a. On the fitted triSBM model, we ran steps of the fast merge-split implementation of SBM [104] to improve the description length (see Methods for a precise definition) of the data made by the model, taking advantage of the gene-regulation information in a way similar to the citation between documents when they are used to improve the classification ability of hSBM in that context.

We report in Figure 6.5b the Normalized Mutual Information, measuring the ability of the full process (fit triSBM, add links, run merge-split) in identifying the breast subtypes. Remarkably enough, we see that by including the information on miRNA–genes interactions, we reach a higher NMI, i.e., a better agreement of our clusters with the subtype organization. This does not happen when simply running merge-split after triSBM is run.

This shows that it is possible to integrate not only multiple layers of sample-related information, but also knowledge about correlations between different kinds of features. Our results represent a first proof of concept in this direction, and we plan to further pursue this type of analysis in future.

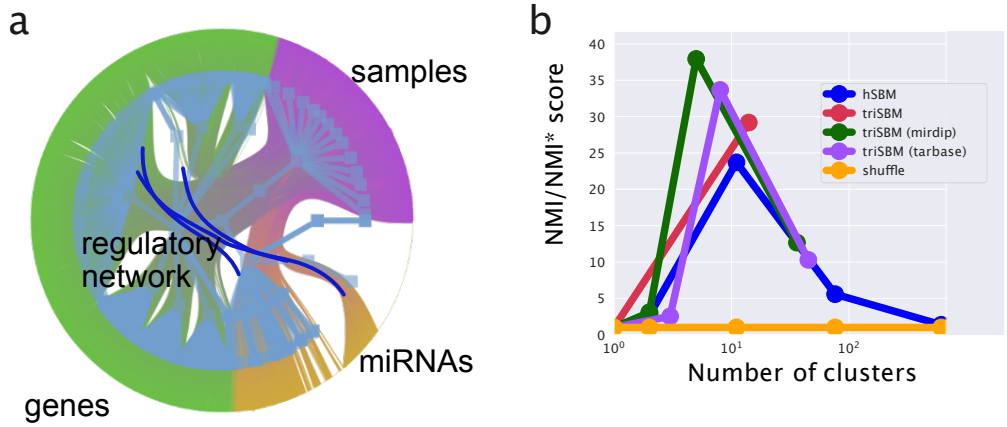


Figure 6.5: **Configuration and scores when adding gene–miRNA links.** (**a**) A graphic of a tripartite network with links gene–miRNA. (**b**) The scores of this new setting using two different (mirDIP [124] and TarBase [96]) regulatory networks separately.

6.4 Adding further layers of information: the case of Copy Number Variation

As we discussed in the introduction, the nSBM algorithm can be extended in principle to any other layer of information on the samples. A natural candidate is Copy Number Variation (CNV). It is well-known that chromosomal aberrations are a hallmark of cancer and that several types of cancer are characterized by a well-defined set of chromosomal loci whose deletion or duplication can drive the onset of that particular type of cancer. we already noticed that, using the information contained in the miRNA branch, we could identify two loci whose alteration were known to be associated with the onset of breast cancer. In TCGA database, we also have the information on the CNV values for all samples. We included this information by adding a fourth branch to our algorithm (accordingly, we shall call it in the following, "tetraSBM") As a preliminary test, we selected only genes with positive CNV (i.e., genes contained in duplicated loci) and that were neglected for the moment deletions.

We performed a gene selection also in this new branch. Highly Copied Genes were selected, keeping the ones with an average (over samples) CNV greater than 3.5. A total 1 353 genes passed our selection. This selection would select genes with at least 2 duplications (CNV = 4) on average.

It is important to stress that, at this stage nodes, which corresponds with the same gene in the gene expression branch and in the CNV branch, are completely uncorrelated and are seen by the algorithm as independent nodes. we shall discuss below how to address this issue.

In our setting, we have 3 000 protein-coding genes in the gene expression branch, 1 353 genes in the CNV branch, and 417 of them are represented by nodes in both branches.

We ran the tetraSBM model on this network with samples, protein-coding genes, miRNAs, and CNV genes and obtained two hierarchical levels. In the first one, the four branches were partitioned into 13 clusters, 7 gene-topics, 5 miRNA-topics, and 5 CNV-topics. In the second one, we found 397 clusters, 49 gene-topics, 14 miRNA-topics, and 31 CNV-topics.

Performing the usual Gene Set Enrichment Analysis we found, with very low values of False Discovery Rate (FDR), a few chromosomal loci that we think represent the complete collection of chromosomal aberration associated with breast cancer and could be used as a robust signature

of this type of tumor. The relevance of this result is supported by the other set of enriched keywords (taken from [92]), and show that for some of these loci, the association with breast cancer is already known and is indeed very strong.

On the other side, if we test the performance of tetraSBM to identify the samples subtype, we see that, including the information on CNV, we have a **decrease** in the NMI value (see Figure 6.6). This is not surprising because within the duplicated (or deleted) loci, besides the few drivers of the cancer, there are hundreds of "hitchhikers" genes that simply add noise to the process of subtype classification performed by the other two layers (genes and miRNAs). The variability of the gene expression values that are associated to the different cancer subtypes (and in fact, are allowed to classify the subtypes in the hSBM and triSBM versions of the algorithm) were completely shadowed by the noise induced by the CNV branch.

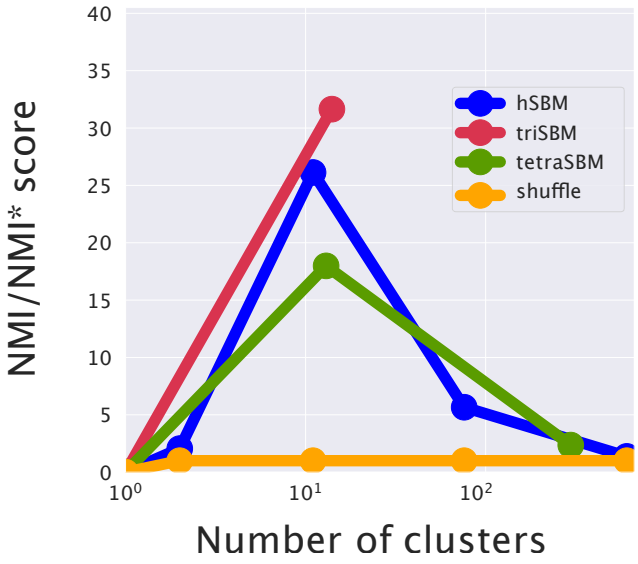


Figure 6.6: **Normalised Mutual Information of models with genes and mRNA (hSBM), plus miRNA (triSBM) and plus both miRNA and CNV (tetraSBM).** Adding CNV introduces noise to the model.

This tells us that adding further layers of information does not automatically improve the quality of clustering. It is always important to perform a careful analysis of the biological information contained in the data and of its possible interference with the other layers. In this particular example, we learned that miRNAs cooperate together to assign coregulated genes to the same gene-topic and samples of the same subtype in the same clusters. This fact becomes particularly clear looking at the probability (see Equation (A.5) in the Methods section and [103] for further details) of moving nodes between groups: when moving a gene between gene-topics, it is more probable to move in a topic where there are genes with many connections to the miRNAs connected to the gene itself. This is confirmed by the fact that, as We discussed in the previous sections, there are miRNA-topics that overlap with clusters of miRNA [18] known to coexpress in breast cancer. On the other hand, the CNV features force samples with the same duplicated loci to be together and this seems not to be correlated with the cancer subtype, at least in TCGA-BRCA data.

This does not mean that the addition of CNV data is useless. It is only by including CNV that we may have, as we have seen, precise information on the chromosomal aberrations involved in breast cancer. It is also interesting to notice that this information is somehow complementary to the one we obtained in the previous section looking at the miRNA clusters. The chromosomal loci that we detected there are not present in this CNV analysis because their CNV value is below the threshold we fixed to include CNVs in the tetraSBM.

6.5 Code and nSBM software package

The Python package to run nSBM [126] can be downloaded from GitHub (https://github.com/BioPhys-Turin/nsbm, accessed on February 10 2022) or, alternatively, can be installed using Anaconda (https://anaconda.org/conda-forge/nsbm, accessed on on February 10 2022) by running `conda install nsbm -c conda-forge`.

We discussed in this paper the application using genomics data; however, the package is written in a way that makes it agnostic with respect to the type of data it receives in input and to the number of branches. One can ideally integrate as many different sources ('omics) of data as needed. Eventually, it can process not only biological data, but every kind of dataset whose input could be represented as a rectangular matrix (Bag of Words) for each feature.

PART III

Contributions to the community

CHAPTER 7

Open Source contributions

During my Ph.D. I took the opportunity to contribute to the open source community multiple times and I developed code that is now available (under open source licenses) via `git` on the GitHub platform. Some of the project take advantage and are reproducible thanks to a containerized approach using Docker [87].

7.1 Contributions to existing projects

hierarchical Stochastic Block Model

I contributed actively to develop the hSBM model: I proposed multiple consistent Pull Requests (#9, #18, #33, #44) and important bug fixing (#29, #30, #45). Thanks to those PRs I became one of the contributors of the hSBM algorithm (Figure 7.1).

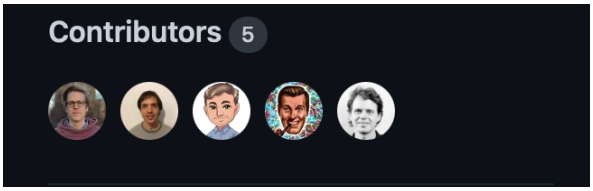


Figure 7.1: **Contributors of hSBM on GitHub**.

GSEApy: a little bug-fixing for a powerful tool

I contributed to GSEApy https://github.com/zqfang/GSEApy/pull/112 in order to have a Python tool to interact with the powerful Gene Set Enrichment Analysis Tool [122].

lda gibbs sampling

I modified the Python lda implementation, in particular I fixed(https://github.com/lda-project/lda/pull/118) the estimation of the Log-Likelihood. This was different than the classical and documented way of estimating Log-Likelihood and therefore it could lead to meaningless results.

I also made some improvements to spatial-LDA (https://github.com/rmsander/spatial_LDA/pull/5).

7.2 Original developments and packages

nSBM: multi branch topic modeling

I developed and released the n-branched Stochastick Block Modeling discussed deeply in Sections 5 and 6. It is available at https://github.com/fvalle1/sbm and it is distributed as a `conda` package https://github.com/conda-forge/staged-recipes/pull/14567. It can be simply installed running `conda install -c conda-forge nsbm`.

```
from nsbm import nsbm

model = nsbm()
model.make_graph_multiple_df(df, df_key_list)

model.fit(n_init=1, B_min=50, verbose=True)
```

Figure 7.2: **Example of use of the nsbm package.**

topicpy: Python meets topic modeling

I developed a Python package named `topicpy`(https://pypi.org/project/topicpy/) available through *pip* `pip install topicpy` to perform all the common analyses and plots described in this work.

7.3 Side projects

Component System

I developed a Tool for the Analysis of COmponent Systems (tacos https://github.com/fvalle1/tacos), written in C++ it can efficently estimate quantities described in Section 2 to study rectangular matrices from a component systems point of view.

```
Running Tacos
threads: 2
Please write some options
0 ---> read mainTable.csv
1 ---> read and extimate correlation mainTable.csv
2 ---> extimate means and variances
3 ---> GenerateNullData
4 ---> read nullTable.csv
5 ---> nullTable.csv read and extimate correlation
6 ---> nullTable.csv extimate means and variances
7 ---> read and make bipartite graph
```

Figure 7.3: **Example of use of the TACOS package.**

LatentTrees

I developed a latent tree model to generate Zipf's law like entries https://github.com/fvalle1/latentrees, it has also a GPU implementation https://github.com/fvalle1/latentrees/independentrees.

```
from latentrees import *
runtime = analyses()
runtime.append_model(distribution = lambda node: rng.integers(node-1-np.sqrt(3)*abs(node), node+1+np.
    sqrt(3)*abs(node)), name="unif")
runtime.append_model(distribution = lambda node: round(rng.normal(node,abs(node))),name="normal")
runtime.append_model(distribution = lambda node: rng.gamma(1, node+1), name="gamma")

runtime.append_model(name="negative_binom")
moi_index = "negative_binom" #model of interest
layers = runtime[moi_index].layers
L = runtime[moi_index].L
nl = runtime[moi_index].nl
cnts = layers[-1].sorted_nodes
```

Figure 7.4: **Example of use of the latentree package** to sample numbers power-law distributed.

The Hopefield network: a modern implementation

I developed hopefield4py https://pypi.org/project/hopfield4py/ in order to run simple simulations using the Hopfield model [59] (https://towardsdatascience.com/the-hopfield-network-67267d0569d2).

This is available at https://github.com/fvalle1/hopfield.

```cpp
// Create a dataset with memories
std::vector<Memory> training_dataset;
training_dataset.emplace_back(Memory(6, 1, 1, 0, 0, 1, 1));
training_dataset.emplace_back(Memory(6, 1, 1, 1, 1, 1, 1));

//Create a model
auto model = Model(training_dataset.size(), training_dataset[0].size());
model.load_memories(training_dataset);

// Build a corrupted memory
spin corrupted_data[] = {1, 1, 1, 1, 1, 0};
Memory corrupted_memory;
std::memmove(corrupted_memory.fData,corrupted_data, corrupted_memory.size_of());

// Train the model. Use can use kNull, kCPU, kGPU, kMultiThread, kOMP
// If the chosen device is not available another one is automatically picked up
model.train(kCPU);
model.reconstruct(corrupted_memory);

cout << "Reconstructed: " << endl;
for (uint8_t i = 0; i < corrupted_memory.size(); i++) cout << corrupted_memory.fData[i] << " ";

cout << model;
```

Figure 7.5: **Example of using my Hopfield tool** to reconstruct trivial memories.

```python
import tensorflow as tf
from hopfield4py import Hopfield

data = tf.convert_to_tensor([[1,1,1,1,1,1],[1,-1,-1,1,1,1]])
model = Hopfield(6)
model.load(data)
corrupted = tf.convert_to_tensor([1,-1,-1,1,1,1], dtype=tf.double)
reconstructed = model.reconstruct(corrupted)
```

Figure 7.6: **Example hopfield4py tool** to reconstruct trivial memories.

PART IV

Conclusions

CHAPTER 8

Conclusions

We demonstrated that topic models are powerful tools to mine information from transcriptomic datasets.

In particular the main goals of this work are:

- the description of RNA-Sequencing data using statistical laws (see section 2) found also in linguistics and use this fact to motivates the use of tools like topic modeling on this kind of data.
- among topic models hSBM presents some advantages, for instance it is non parametric and it avoids to search few markers, as discussed in section 3.
- hSBM has good performances on cancer data from TCGA as discussed in Section 4;
- I developed a new tool `nSBM` that allows scientists to integrate multiple kind of data in a hSBM framework. (Section 5);
- Using the *nSBM* python package, inherited from hSBM [40], ready to install and easily executable on n-partite networks, it will be straightforward to address different types of biological data;
- the integration of multiple sources of data, such as microRNA expression levels and the protein-coding mRNA ones, greatly improves the ability of the algorithm to identify breast cancer subtypes as discussed in the Section 6.

In conclusion, we released a new tool to easily integrate different sources of data and we showed some application in different cases, with some sources of data (mRNA, miRNA, CNV). Indeed, this approach can be applied to other datasets and, more importantly, to any possible sources of data (genomics, proteomics, lncRNA, circRNA. . .).

Appendices

APPENDIX A

Materials and Methods

A.1 Data

TCGA Data

The results published here are in part based upon data generated by The Cancer Genome Atlas (TCGA) managed by the NCI and NHGRI. Information about TCGA can be found at http://cancergenome.nih.gov.

We downloaded data from TCGA using tools provided by Genomic Data Commons (GDC) [43]. We downloaded *Gene Expression Quantification* data type in *transcriptome profiling* category. We choose *RNA - Seq* with *HTSeq - FPKM* as workflow type. We downloaded the 1 222 samples from TCGA-BRCA project and 1 145 from TCGA-LUSC and TCGA-LUAD projects.

We downloaded the subtype annotations using TCGABiolinks GUI [24, 89, 118]. We downloaded the independent Breast Cancer Consensus Subtypes (BCCS) related to the TCGA files provided by the Supplementary files of [60].

During the analysis of breast cancer, we downloaded both the *SubtypePam50* classification labels provided by [72] and the more recent annotations reported in [22]. We discussed the performance of hSBM in classifying the two in [127].

The Genotype Tissue Expression dataset

The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. The data used for the analyses described in this manuscript were obtained from the GTEx Portal (version phs000424.v8.p2) [79]. GTEx data were downloaded in Transcript Per Million (TPM) format. We also downloaded from the GTEx portal https://gtexportal.org the annotations of samples and in particular we focused on the tissue type (the area from which the tissue sample was taken) and its subtypes (*SMTS* and *SMTSD* labels).

METABRIC miRNA Landscape Data

We downloaded METABRIC data from the European Genome-Phenome archive. We downloaded METABRIC miRNA landscape study (*EGAS00000000122*), in particular, Normalized miRNA expression data (*EGAD00010000438*) and Normalized mRNA expression (*EGAD00010000434*).

Unified Dataset

We downloaded data with both healthy and diseased tissues from a dataset prepared by [132]. They processed data from GTEx and TCGA with the same pipeline and successfully corrected for study-specific biases, enabling comparative analysis. We downloaded the second version of their normalized data from figshare [131].

Only samples with a valid annotation from TCGABiolinks or GTEx were considered. We applied a $\log_2(FPKM + 1)$ transformation; this reduced the number of edges E and let the algorithm to be faster even with a large number of nodes N.

Single-cell datasets

The Mouse Cell Atlas (MCA) was selected as the main illustrative dataset for the component system Section 2. In the MCA more than $\sim 4 * 10^5$ single cells were profiled using scRNAseq from all major organs [50]. An advantage of this dataset is the use of Unique Molecular Identifiers (UMI) [64]. This technique allows the identification of the absolute number of unique RNA molecules detected by sequencing, thus eliminating the amplification noise. In the context of single-cell gene expression assays, this method provides a reliable estimate of the number of mRNA detected for coding genes and an estimate of the transcriptome size sampled.

We also analyzed the compendium of Tabula Muris (TM) for comparison. This atlas comprises an analogous number of cells from 20 organs and tissues [123] that were processed with the Smart-seq2 protocol [108], which produces a full-length transcriptome profiling but does not use UMI.

Finally, we analyzed two additional single-cell datasets, relative to a HEK cell line and to mouse fibroblasts, profiled with the Smart-seq3 protocol [49].

Texts data

In the Section 5 of this work, we analysed texts from different sources. In particular we choose.

PMC Text Mining collection from Open Access Subset of articles from March to April 2020 https://www.ncbi.nlm.nih.gov/pmc/tools/textmining/. Keywords available in the fields `kwd-group` made the additional layer. Different journals were considered as classes.

Twitter Subset of tweets available from http://www.nltk.org/nltk_data/ [9]. Positive and negative tweets were considered as classes. The additional layer was made with hashtags.

Wikipedia Data from a Wikipedia texts already used by [40] with the respective Physics/Chemistry assignment. Additional layer is composed by categories obtained trough the English endpoint of the Wikimedia API https://www.mediawiki.org/wiki/API:Main_page

APS made data available to researchers; they are based on their publications for use in research about networks and the social aspects of science at https://journals.aps.org/datasets. We picked a subset of papers from 2020 issues in Physical Review A/B/C/D/E. Additional layers of information were built using keywords from *classificationSchemes* and the number of authors.

Reuters articles were downloaded using nltk python tool [9] described at https://www.nltk.org. A subset of 1 000 articles from the 10 most common categories were used. In this setting the additional layer was build using titles.

Synthetic datasets

As described above we tested our approach on synthetic data. This matrices are sampled with the following generation process: for each topic K, we draw β_K from a Dirichlet distribution $D(\phi)$. For each document d:

1. we draw θ_d from a Dirichlet distribution $D(\alpha)$;
2. for each of the N_d words, w
   a) we draw topic assignment z_{wd} from a multinomial $M(\theta_d)$;
   b) and we draw the i-th word w from a multinomial $M(\phi_{z_{wd}})$;
3. We estimate the average assignment $\bar{z_d} = \frac{1}{N_d} \sum_{i=1}^{N_d} z_{id}$;
4. For each additional branch B;
   a) we draw $N\iota_d$ branch-specific components Y_{dlj} from a normal distribution $N(\bar{z_d}\eta, \sigma^2)$.

If not explicitly defined otherwise we choose the following parameters: the number of topics $K = 10$, the number of documents $D = 250$, the number of words (with repetition) $W = 1000$, $\alpha = \frac{1}{K}$, $\phi = \frac{1}{K}$, $\eta = \frac{10}{K}$, $\sigma = 0.005$ and, finally, the number of tokens of each kind is chosen equal to the number of words $N_{d\prime} = N_d$.

In this way, we extend [84] with B additional branches and, for each branch, a given number of words or tokens is assigned to each document.

A.2 Gene and miRNA Selection

The data provided in the atlas consisted of 1 000 samples associated with almost 20 000 protein-coding genes and 2 000 miRNA entries. Without preprocessing this would have led to an adjacency matrix too big to be handled efficiently by the algorithms.

We performed two kinds of preprocessing to reduce the number of nodes and the number of edges.

In order to reduce the number of nodes, we filtered genes and miRNAs selecting only the highly variable ones. The highly variable are the ones with the highest dispersion (variance over mean) with respect to the genes with the same average expression. This selection was performed using the *scanpy* python package [134]. This analysis was performed separately on genes and microRNAs since they are provided by different experiments and different normalization. We selected in this way 3 000 genes and $\sim 1\,200$ miRNAs.

Furthermore, we applied a standard approach to reduce the weights of the links and applied a $\log(counts + 1)$ transformation to the data before running the topic models. This helped us to reduce by some order of magnitudes the number of edges.

An interesting feature of the SBM type of algorithm is that they are typically robust with respect to gene selection. In the analyses of this paper, we considered only highly variable genes; however, in the Supplementary material of [127], we discussed different types of gene selections showing that they were typically leading to similar performances.

In the analysis of the METABRIC dataset, we utilized the previously selected genes and microRNA.

A.3 Models

Hierarchical Stochastic Block Model

We adapted hierarchical stochastic block model (hSBM) to gene expression data.

Hierarchical stochastic block model is a kind of generative model that tries to maximize the probability that the model θ describe the data $\mathcal{A}$

$$P(\theta|\mathcal{A}) = P(\mathcal{A}|\theta)P(\theta) \tag{A.1}$$

using a non-parametric approach. In the setting described in this paper, $\mathcal{A}$ is the gene expression matrix and the entries $\mathcal{A}_{ij}$ represent the number of counts of gene i in sample j. In other words, $\mathcal{A}$ is the incidence matrix of a bipartite network composed by genes and samples, the edges of this network are weighted by the gene expression. The minimise_nested_blockmodel_dl function from the graph-tool package [99] minimizes the description length $\Sigma = -lnP(\mathcal{A}|\theta) - lnP(\theta)$ of the model. We used the nested version of the model since we expected some sort of hierarchical structure in the data [100, 101, 102, 104].

We set the algorithm to minimize the description length Σ many times and selected the model that obtained the shortest description length.

As output of the model, we find the probability distributions $P(\text{topic}|\text{sample})$ and $P(\text{gene}|\text{topic})$. These probabilities are defined, in terms of entries of the program as follows:

$$P(\text{topic}|\text{sample}) = \frac{\text{number of half-edges on sample coming from topic}}{\text{number of half-edges on sample}} \tag{A.2}$$

and

$$P(\text{gene}|\text{topic}) = \frac{\text{number of half-edges to topic going to gene}}{\text{number of half-edges to topic}}. \tag{A.3}$$

The complexity of hSBM is $O(VLn^2V)$ if the graph is sparse, i.e., if $E \sim O(V)$ [100], where V is the number of vertices (samples and genes) and E the number of edges. For $E >> V$ the complexity increases and the CPU time needed to minimize the description length increases as well. In this case, to reduce the CPU bottleneck, one can apply a log-transformation to the data, which strongly reduces the number of edges E.

Implementation of WCGNA

In this work, some of the analysis required other clustering methods. We tested Weighted Gene Correlation Network Analysis (WGCNA) [76, 136], which outputs *modules* of correlated genes. We used the implamentation in its R package https://cran.r-project.org/package=WGCNA. This was run using default parameters: power was set to the lowest for which the scale-free topology fit index curve flattens out, minModuleSize was set to 5 and mergeCutHeight to 0.2. WGCNA creates *modules* of genes, we considered them as topics. In order to obtain clusters we cut the tree built using modules to estimate distances between samples. WGCNA is a valid option when the number of topics or clusters (e.g., suptypes) is well-known a-priori and, probably, a grid search of the best parameters will lead to even better results, but this is beyond the scope of this work since we wanted to compare it to hSBM which is completely non-parametric.

Implementation of Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation (LDA) [10] is a standard and well-known topic model and we used the implementation provided by scikit-learn [56]. The model was configured using the default setting for the parameters α and β [1]: they were set to $\frac{1}{K}$, being K the number of topics. K was set based on the number of clusters in output from hSBM. When managing LDA output, we selected the *argmax* of $P(\text{topic}|\text{sample})$ to define clusters.

To get a list out of the LDA topic distribution we selected the 20 most distinctive genes of each topic. The distinctiveness was described and used in LDA analyses of GTEx by [29] and is the minimum Kullback-Leibler of $P(\text{topic}|\text{gene})$.

$$D^g[K] = min_{l \neq k} \theta_{kg} \log\left(\frac{\theta_{kg}}{\theta_{lg}}\right) + \theta_{lg} - \theta_{kg} \tag{A.4}$$

being $\theta_{\{k,l\}g}$ the $P(\text{gene}_g|\text{topic}_{\{k,l\}})$.

Topic Mapping

Another algorithm of topic modeling we took into account is Topic Mapping (TM), described by [75], with defaults parameters: the number of runs `-r` was set to 10 and the minimum topic size `-t` was even set to 10.

Topic Mapping requires in input a corpus of text and not a Bag of Words [2, 52] or a design matrix, this is why we built a corpus in which each text was constructed from the 1000th most expressed gene in each sample. In other words each sample were translated in a text composed by the 1000 genes most represented in that sample.

Hierarchical clustering

In the end we tested hierarchical clustering [69] (hierarchical) implemented in *sklearn* and set to use *euclidean* affinity and *complete* `linkage` as done by [29]. Hierarchical clustering is not a topic modeling algorithm, we reported it for comparison and for its simplicity to be run and interpreted.

[1] α and β represent the parameters of the Dirichlet distribution from which the words (of a topic) and the topics (of a document) are sampled

nSBM: A Multibranch Stochastic Block Modeling Algorithm

As deeply discussed in Section 5 we implemented nSBM wich has some advanteges over the bipartite hSBM, in particular:

- the training process is performed simultaneously in all branches of the network: this means that all the types of data contribute to the learning process at the same time, without, in principle, any preference at the beginning.
- nSBM is completely nonparametric [101] and as hSBM it minimizes the Description Length $\Sigma = -\log P(\mathcal{A}|\theta) - \log P(\theta)$. We used the `minimise_nested_blockmodel_dl` function from graph-tool [99]. In our setting, $\mathcal{A}$ is a block matrix in which each block is a "Bag of Features" (i.e., genes, miRNAs, ...). It can be seen as a two-dimensional matrix whose entries w_{ij} are the weights mentioned above. The probability of accepting the move of a node with a neighbor t from group r to group s is [101, 103]

$$P(r \to s|t) = \frac{e_{ts} + \epsilon}{e_t + \epsilon B}, \tag{A.5}$$

  where e_{ts} is the number of edges between groups t and s; e_t is the total number of edges connected to group t. From this, another advantage of a multibranch approach should be clear: different 'omics may have their own normalization. In fact, when moving a sample from r to s, the probability is estimated considering only the branch to which t belongs. If the node t is a gene, e_{ts}/e_t is normalized, taking only into account the mRNA expression values.

As the output of the model, we find the probability distributions P(topic|sample) and P(gene|topic). These probabilities are defined, in terms of entries of the program, as follows:

$$P(\text{topic}|\text{sample}) = \frac{\text{number of half-edges on sample coming from topic}}{\text{number of half-edges on sample}} \tag{A.6}$$

and

$$P(\text{gene}|\text{gene-topic}) = \frac{\text{number of half-edges to gene-topic going to gene}}{\text{number of half-edges to gene-topic}}. \tag{A.7}$$

The same is true for miRNA-topics and for each and every eventual additional layer of features.

We ran the model on a 48-core machine with 768 GB of memory [1].

A.4 Evaluation metrics

Normalized Mutual Information (NMI)

In order to evaluate the agreement between the sample partitions and the annotations, we chose the so-called "Normalized Mutual Information" (NMI) [114], it was proposed in [117] as a new evaluation framework for topic models. Moreover, as discussed in [127][2], it can be shown that NMI is the harmonic average of two metrics that evaluate, respectively, the completeness and the homogeneity of a partition of annotated samples.

Given a set C of labeled samples and a partition K in clusters of these samples, the NMI is defined as the harmonic average of homogeneity h and completeness C:

$$NMI = 2\frac{h * C}{h + C} \quad ,$$

where the homogeneity is defined as

$$h = 1 - \frac{H(C|K)}{H(C)}$$

[2] I wrote a generalistic article about this at https://towardsdatascience.com/v-measure-an-homogeneous-and-complete-clustering-ab5b1823d0ad

and the completeness as

$$C = 1 - \frac{H(K|C)}{H(K)} \quad .$$

In these definitions $H(C)$ and $H(K)$ are the usual Shannon entropies associated to the partitions C and K; $H(C|K)$ and $H(K|C)$ are defined as:

$$H(C|K) = -\Sigma_{c\in C,k\ \in K} \frac{n_{ck}}{N} \log\left(\frac{n_{ck}}{n_k}\right)$$

and

$$H(K|C) = -\Sigma_{c\in C,k\ \in K} \frac{n_{ck}}{N} \log\left(\frac{n_{ck}}{n_c}\right)$$

respectively, where n_c is the number of samples lebelled c, n_k the number of samples in the cluster k and n_{ck} the number of samples lebelled c in the cluster k. With this definition it is easy to understand the meaning of homogeneity and completeness. If all the samples in cluster k belong to the label c then $n_k = n_{ck}$ and $h = 1$. Similarly the completeness C equals 1 if all samples belonging to the label c are in the same cluster k.

The NMI is estimated using Shannon's entropy formula to measure the quantity of information in the partition. The problem of this measure is that even in a random partition, there is a residual entropy and the NMI is not zero; this effect is particularly important in the layers of the models with high resolution (many small clusters/topics). In order to avoid this bias, we evaluated this default NMI by randomizing the ground truth annotations of the samples. This was performed multiple (~ 50) times, each time preserving the number of clusters and the number of samples in every cluster; we call the average NMI on these multiple random assignments NMI^*; this is the residual information on the considered partition. Moreover, we set the normalized score NMI/NMI* to 1 when both NMI and NMI* are zero, which is at the first layer where only one cluster is present. In the results, we reported $NMI/NMI*$, which measures how much information the model learns with respect to a random assignment. It is important to stress that this measure has no absolute value and should not be used to compare performances on different datasets; however, it can be successfully used to compare different algorithm in the same dataset.

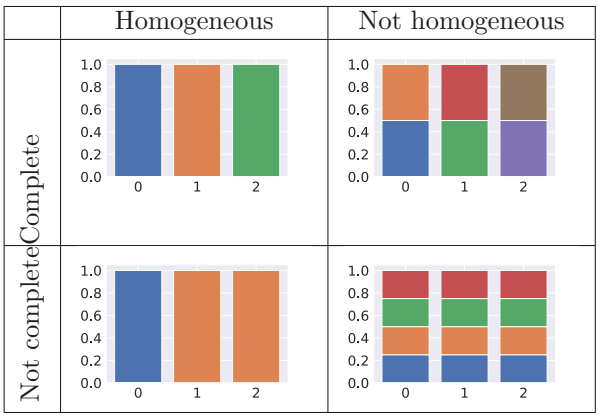


Table A.1: **Examples of homogeneity and completeness. Homogeneous clusters contain all nodes with the same label.** A label is complete if it is fully represented by a single cluster. In this image some examples of these definitions. The NMI score discussed in this work is nothing but the geometric average of completeness and homogeneity [114].

Description Length Σ measures how well a model describes the data

In addition to the NMI, it is also possible to compare different classes of topic modeling algorithms on their ability to compress the data [103, 135]. This can be addressed measuring the description length

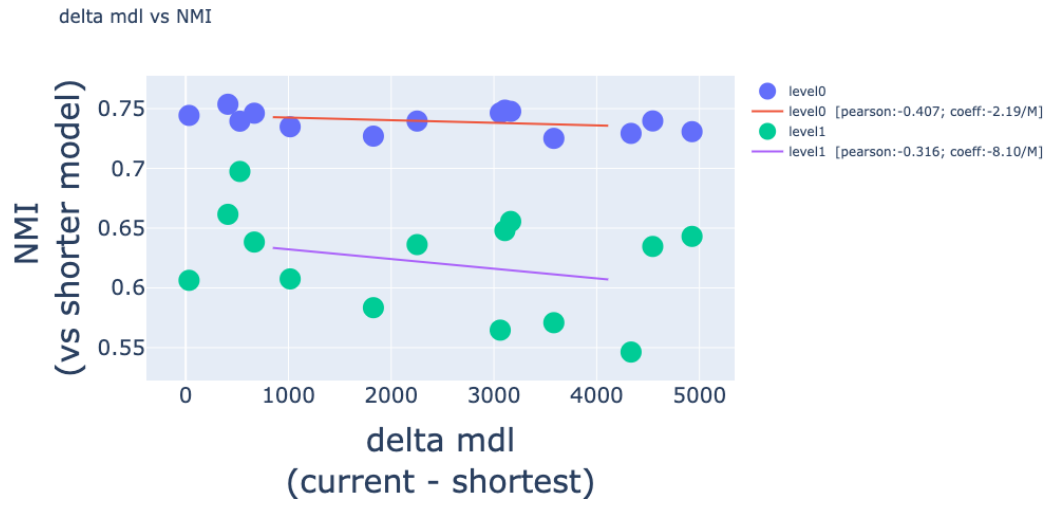


Figure A.1: **Correlation between the description length and the Normalized mutual information**. The two measures are anti-correlated: the shorter the description length the higher the normalised mutual information.

Σ of a model, which represents, in nat units, the number of bits a model requires to describe the data network. Unlike NMI, it has the advantage not to rely on any ground truth. Using $\frac{\Sigma}{E}$ (where E is the total number of edges), it is possible to measure the quantity of information that the model requires to describe an edge. In a typical run on cancer data, hSBM requires a $\frac{\Sigma}{E} \sim 6,26$, which is greater than the $1,4$ units required by triSBM. One can estimate the difference of the two $\Delta\left(\frac{\Sigma}{E}\right) \simeq 4,9$; this can be related to the Bayes factor [70] (being the posterior $P = \exp -\Sigma)$ $\Lambda = \exp \Delta\Sigma \simeq e^{4,9} \simeq 10^{2,1}$, meaning that the model with miRNA is a ~ 100 times more probable description of the data network links.

We tested in a setting like the one described in section 5 if it exists a correlation between the NMI and the Description Length score. This is not trivial at all since the DL is a *unsupervised* score and NMI is a *supervised* one instead. The Figure A.1 represents this anti-correlation. The fact that with this kind of data the DL and the NMI are related further justifies the use of hSBM-like models that tries to minimise the DL in a setting in which one would like to obtain the maxima Normalised Mutual Information with the considered ground truth.

A.5 Investigate the enrichment of the topics

A topic is nothing but a list of genes, it can be investigated using hypergeometric tests. The results shown in this paper are computed using the GSEA [122] tool.

A.6 Box topic and gene ontologies

Once an algorithm outputs the distribution over topics, we centered the $P(\text{topic}|\text{sample})$. The new probability distribution centered was $\overline{P}(\text{topic}|\text{sample}) = P(\text{topic}|\text{sample}) - \frac{1}{R}\Sigma_{s\in\text{samples}}P(\text{topic}|\text{s})$, being R the total number of samples. This centered $P(\text{topic}|\text{sample})$ can be studied, after grouping samples by their tissue.

Topics are nothing but lists of genes, this naturally brought to perform gene ontology tests. We searched for enrichment using GSEA [122].

A.7 Predictor on latent space

We built a Neural Network predictor using topics as features, doing this we could train a simple model on a low-dimensional space instead of a complex model on the original $\sim 20\,000$ dimensional space.

We fit hierarchical Stochastic Block Model using this subset. In output we obtained the topic distribution of samples.

At this point we considered topics as features and $P(\text{topic}|\text{sample})$ as entries of the design matrix X. The entry X_{ij} is the $P(\text{topic}_j|\text{sample}_i)$. The matrix needed a further normalisation to be fitted using Stochastic Gradient Descent later. The normalised matrix $\overline{X}$ was obtained subtracting features' means and dividing by its range each feature. The entries of this new matrix are $\overline{X_{ij}} = \frac{X_{ij} - \langle X_{ij\prime} \rangle_{j\prime}}{0.5 * (\max_{j\prime} X_{ij\prime} - \min_{j\prime} X_{ij\prime})}$.

The dataset was split into a training and a test set; the training set contains the 95% of the samples. This is quite unbalanced, but we were not going to use it to really evaluate the performance of the model. Moreover, 25% of the training set was used as validation set. The model consisted of a neural net with an hidden layer with 100 neurons activated by *ReLU*, the optimiser was Stochastic Gradient Descent (SGD). Finally, the output layer uses a *softmax* activation to classify tissues.

We wanted to evaluate the performance on completely new samples, never retraining neither topic modeling neither the neural net. From the original dataset we selected all the samples of the tissues considered, not involved in topic modeling. These are $\sim$ 5 000 new samples never fitted by hSBM or the neural net. We project this *unseen* samples into the topic space. To do this, firstly we selected the genes involved in topic modeling, the ones the passed the highly variable filter we imposed. The $P(\text{topic}|\text{sample})$ for the new samples were simply $P(\text{topic}|\text{sample}) = \Sigma_{\text{gene}} P(\text{topic}|\text{gene}) * P(\text{gene}|\text{sample})$, in other words each expression array is multiplied to the matrix *genes*$\times$*topics*. We now evaluate the Neural Net performances on this new dataset.

The confusion matrix and the Receiving Operating Characteristic curves were estimated using scikit learn [98]. The ROC curve represents the True Positive Rate or sensitivity ($\frac{TP}{TP+FN}$) versus the False Positive Rate or $1 -$ specificity ($\frac{FP}{FP+TN}$).

The whole predictor model was implemented using keras [21].

A.8 A MNIST-like dataset with LaTeX symbols

I wrote the code to create images in the same format as the MNIST dataset [78] using data from the detexify project using handwrittten LaTeX symbols https://detexify.kirelabs.org/classify.html, the code is freerly available at https://github.com/fvalle1/detexifyasMNIST.

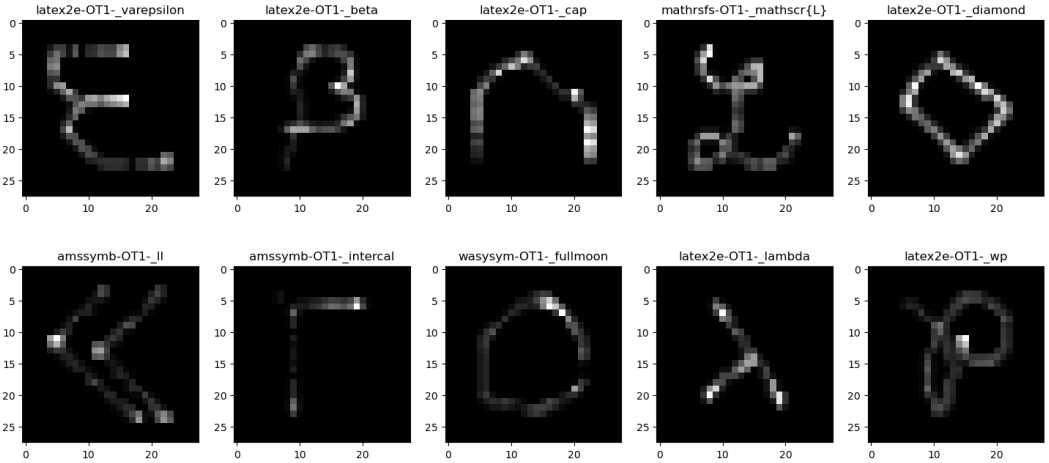


Figure A.2: **Examples of images in the MNIST format from detexify.**

APPENDIX B

Figure Legends

Figure B.1: **Organ color legend for figures in the main text.**

Bibliography

[1] Aldinucci, M. et al. 'OCCAM: a flexible, multi-purpose and extendable HPC cluster'. In: *Journal of Physics: Conference Series* vol. 898, no. 8 (2017), p. 082039.

[2] Aldous, D. J., Ibragimov, I. A. and Jacod, J. *Ecole d'Ete de Probabilites de Saint-Flour XIII, 1983*. en. Ed. by Hennequin, P.-L. École d'Été de Probabilités de Saint-Flour. Berlin Heidelberg: Springer-Verlag, 1985. DOI: 10.1007/BFb0099420.

[3] Almaas, E. and Barabási, A.-L. 'Power Laws in Biological Networks'. In: *Power Laws, Scale-Free Networks and Genome Biology*. Boston, MA: Springer US, 2006, pp. 1–11. DOI: 10.1007/0-387-33916-7_1.

[4] Altmann, E. G. and Gerlach, M. 'Statistical Laws in Linguistics'. In: *Creativity and Universality in Language*. Ed. by Degli Esposti, M., Altmann, E. G. and Pachet, F. Cham: Springer International Publishing, 2016, pp. 7–26. DOI: 10.1007/978-3-319-24403-7_2.

[5] Andrzejewski, D., Zhu, X. and Craven, M. 'Incorporating domain knowledge into topic modeling via Dirichlet Forest priors'. In: *Proceedings of the 26th Annual International Conference on Machine Learning*. ICML '09. Montreal, Quebec, Canada: Association for Computing Machinery, June 2009, pp. 25–32. DOI: 10.1145/1553374.1553378.

[6] Ashley, E. A. 'Towards precision medicine'. In: *Nature Reviews Genetics* vol. 17, no. 9 (Sept. 2016), pp. 507–522. DOI: 10.1038/nrg.2016.86.

[7] Berger, A. C. et al. 'A Comprehensive Pan-Cancer Molecular Study of Gynecologic and Breast Cancers'. English. In: *Cancer Cell* vol. 33, no. 4 (Apr. 2018), 690–705.e9. DOI: 10.1016/j.ccell.2018.03.014.

[8] Bertoli, G., Cava, C. and Castiglioni, I. 'MicroRNAs: New Biomarkers for Diagnosis, Prognosis, Therapy Prediction and Therapeutic Tools for Breast Cancer'. eng. In: *Theranostics* vol. 5, no. 10 (2015), pp. 1122–1143. DOI: 10.7150/thno.11543.

[9] Bird, S., Klein, E. and Loper, E. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. en. Google-Books-ID: KGIbfiiP1i4C. O'Reilly Media, Inc., June 2009.

[10] Blei, D. M., Ng, A. Y. and Jordan, M. I. 'Latent dirichlet allocation'. In: *Journal of Machine Learning Research* vol. 3, no. Jan (2003), pp. 993–1022. DOI: 10.5555/944919.944937.

[11] Bosetti, C. et al. 'Cancer mortality in Europe, 2005–2009, and an overview of trends since 1980'. en. In: *Annals of Oncology* vol. 24, no. 10 (Oct. 2013), pp. 2657–2671. DOI: 10.1093/annonc/mdt301.

[12] Brunet, J. P. et al. 'Metagenes and molecular pattern discovery using matrix factorization'. In: *Proceedings of the National Academy of Sciences* vol. 101, no. 12 (Mar. 2004), pp. 4164–4169. DOI: 10.1073/pnas.0308531101.

[13] Calin, G. A. et al. 'Human microRNA genes are frequently located at fragile sites and genomic regions involved in cancers'. en. In: *Proceedings of the National Academy of Sciences* vol. 101, no. 9 (Mar. 2004), pp. 2999–3004. DOI: 10.1073/pnas.0307323101.

[14] Cantini, L. and Caselle, M. 'Hope4Genes: a Hopfield-like class prediction algorithm for transcriptomic data'. In: *Scientific Reports* vol. 9 (Jan. 2019). DOI: 10.1038/s41598-018-36744-y.

[15] Cantini, L. et al. 'Detection of gene communities in multi-networks reveals cancer drivers'. In: *Scientific Reports* vol. 5 (2015). Publisher: Nature Publishing Group, p. 17386.

[16] Cantini, L. et al. 'MicroRNA-mRNA interactions underlying colorectal cancer molecular subtypes'. In: *Nature Communications* vol. 6 (Nov. 2015). DOI: 10.1038/ncomms9878.

[17] Cantini, L. et al. 'A review of computational approaches detecting microRNAs involved in cancer'. In: *Frontiers in Bioscience-Landmark* vol. 22 (June 2017), pp. 1774–1791. DOI: 10.2741/4571.

[18] Cantini, L. et al. 'Identification of microRNA clusters cooperatively acting on epithelial to mesenchymal transition in triple negative breast cancer'. en. In: *Nucleic Acids Research* vol. 47, no. 5 (Mar. 2019), pp. 2205–2215. DOI: 10.1093/nar/gkz016.

[19] Chang, K., Creighton, C., Davis, C. et al. 'The cancer genome atlas pan-cancer analysis project'. In: *Nature Genetics* vol. 45, no. 10 (2013). Publisher: Nature Publishing Group, p. 1113.

[20] Chen, Z. et al. 'Non-small-cell lung cancers: a heterogeneous set of diseases'. In: *Nature Reviews Cancer* vol. 14, no. 8 (Aug. 2014), pp. 535–546. DOI: 10.1038/nrc3775.

[21] Chollet, F. et al. *Keras*. 2015.

[22] Ciriello, G. et al. 'Comprehensive molecular portraits of invasive lobular breast cancer'. In: *Cell* vol. 163, no. 2 (2015). Publisher: Elsevier, pp. 506–519.

[23] Cline, M. S. et al. 'Exploring TCGA pan-cancer data at the UCSC cancer genomics browser'. In: *Scientific Reports* vol. 3 (2013). Publisher: Nature Publishing Group, p. 2652. DOI: 10.1038/srep02652.

[24] Colaprico, A. et al. 'TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data'. In: *Nucleic Acids Research* vol. 44, no. 8 (2016). Publisher: Oxford University Press, e71–e71. DOI: 10.1093/nar/gkv1507.

[25] Cosentino Lagomarsino, M. et al. 'Universal features in the genome-level evolution of protein domains'. In: *Genome Bioliology* vol. 10, no. 1 (2009). Publisher: BioMed Central, R12.

[26] CP, W., E, W. and BW, S. *World Cancer Report: Cancer Research for Cancer Prevention*. en. 2020.

[27] Curtis, C. et al. 'The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups'. eng. In: *Nature* vol. 486, no. 7403 (Apr. 2012), pp. 346–352. DOI: 10.1038/nature10983.

[28] D'haeseleer, P. 'How does gene expression clustering work?' In: *Nature Biotechnology* vol. 23, no. 12 (2005). Publisher: Nature Publishing Group, pp. 1499–1501.

[29] Dey, K. K., Hsiao, C. J. and Stephens, M. 'Visualizing the structure of RNA-seq expression data using grade of membership models'. en. In: *PLOS Genetics* vol. 13, no. 3 (Mar. 2017). Ed. by Kundaje, A., e1006599. DOI: 10.1371/journal.pgen.1006599.

[30] Eisen, M. B. et al. 'Cluster analysis and display of genome-wide expression patterns'. In: *Proceedings of the National Academy of Sciences* vol. 95, no. 25 (Dec. 1998), pp. 14863–14863.

[31] Eisler, Z., Bartos, I. and Kertész, J. 'Fluctuation scaling in complex systems: Taylor's law and beyond'. In: *Advances in Physics* vol. 57, no. 1 (2008). Publisher: Taylor & Francis, pp. 89–142.

[32] Eshima, S., Imai, K. and Sasaki, T. 'Keyword Assisted Topic Models'. In: *arXiv:2004.05964 [cs, stat]* (Apr. 2020). arXiv: 2004.05964.

[33] Fajardo-Fontiveros, O., Guimerà, R. and Sales-Pardo, M. 'Node Metadata Can Produce Predictability Crossovers in Network Inference Problems'. In: *Physical Review X* vol. 12, no. 1 (Jan. 2022), p. 011010. DOI: 10.1103/PhysRevX.12.011010.

[34] Fortunato, S. 'Community detection in graphs'. en. In: *Physics Reports* vol. 486, no. 3-5 (Feb. 2010), pp. 75–174. DOI: 10.1016/j.physrep.2009.11.002.

[35] Fortunato, S. and Hric, D. 'Community detection in networks: A user guide'. In: *Physics Reports* vol. 659 (Nov. 2016), pp. 1–44. DOI: 10.1016/j.physrep.2016.09.002.

[36] Furusawa, C. and Kaneko, K. 'Zipf's Law in Gene Expression'. In: *Physical Review Letters* vol. 90, no. 8 (Feb. 2003). Publisher: American Physical Society, p. 088102. DOI: 10.1103/PhysRevLett.90.088102.

[37] Gabriele, M. et al. 'Comprehensive analysis of long non-coding RNAs in breast cancer using topic modeling'. In: (Sept. 2022). DOI: 10.1101/2022.09.13.507779.

[38] Gerlach, M. and Altmann, E. G. 'Scaling laws and fluctuations in the statistics of word frequencies'. In: *New Journal of Physics* vol. 16, no. 11 (Nov. 2014), p. 113010. DOI: 10.1088/1367-2630/16/11/113010.

[39] Gerlach, M. and Altmann, E. G. 'Stochastic Model for the Vocabulary Growth in Natural Languages'. en. In: *Physical Review X* vol. 3, no. 2 (May 2013), p. 021006. DOI: 10.1103/PhysRevX.3.021006.

[40] Gerlach, M., Peixoto, T. P. and Altmann, E. G. 'A network approach to topic models'. In: *Science Advances* vol. 4, no. 7 (2018). Publisher: American Association for the Advancement of Science, eaaq1360–eaaq1360. DOI: 10.1126/sciadv.aaq1360.

[41] Godoy-Lorite, A., Guimerà, R. and Sales-Pardo, M. 'Network-Based Models for Social Recommender Systems'. en. In: *Business and Consumer Analytics: New Ideas*. Ed. by Moscato, P. and Vries, N. J. de. Cham: Springer International Publishing, 2019, pp. 491–512. DOI: 10.1007/978-3-030-06222-4_11.

[42] Grilli, J. 'Macroecological laws describe variation and diversity in microbial communities.' In: *Nature Communications* vol. 11, no. 1 (2020), p. 4743. DOI: 10.1038/s41467-020-18529-y.

[43] Grossman, R. L. et al. 'Toward a shared vision for cancer genomic data'. In: *New England Journal of Medicine* vol. 375, no. 12 (2016). Publisher: Mass Medical Soc, pp. 1109–1112.

[44] Grunwald, P. D. and Grunwald, A. *The minimum description length principle*. MIT Press, 2007.

[45] GTEx Consortium and others. 'The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans'. In: *Science* vol. 348, no. 6235 (2015). Publisher: American Association for the Advancement of Science, pp. 648–660.

[46] Guimerà, R. 'One model to rule them all in network science?' en. In: *Proceedings of the National Academy of Sciences* vol. 117, no. 41 (Oct. 2020), pp. 25195–25197. DOI: 10.1073/pnas.2017807117.

[47] Guimerà, R. et al. 'Predicting Human Preferences Using the Block Structure of Complex Social Networks'. en. In: *PLoS ONE* vol. 7, no. 9 (Sept. 2012), e44620. DOI: 10.1371/journal.pone.0044620.

[48] Hagemann-Jensen, M. et al. 'Single-cell RNA counting at allele and isoform resolution using Smart-seq3'. In: *Nature Biotechnology* vol. 38, no. 6 (2020), pp. 708–714.

[49] Hagemann-Jensen, M. et al. 'Single-cell RNA counting at allele and isoform resolution using Smart-seq3'. en. In: *Nature Biotechnology* vol. 38, no. 6 (June 2020), pp. 708–714. DOI: 10.1038/s41587-020-0497-0.

[50] Han, X. et al. 'Mapping the mouse cell atlas by microwell-seq'. In: *Cell* vol. 172, no. 5 (2018). Publisher: Elsevier, pp. 1091–1107.

[51] Harbeck Nadia, G. M. 'Breast cancer'. In: *Lancet* vol. 389(10074) (2017), pp. 1134–1150. DOI: 10.1016/S0140-6736(16)31891-8.

[52] Harris, Z. S. 'Distributional Structure'. en. In: *WORD* vol. 10, no. 2-3 (Aug. 1954), pp. 146–162. DOI: 10.1080/00437956.1954.11659520.

[53] He, K. et al. 'Regulatory network reconstruction of five essential microRNAs for survival analysis in breast cancer by integrating miRNA and mRNA expression datasets'. en. In: *Functional & Integrative Genomics* vol. 19, no. 4 (July 2019), pp. 645–658. DOI: 10.1007/s10142-019-00670-7.

[54] Heaps, H. S. *Information retrieval, computational and theoretical aspects*. Academic Press, 1978.

[55] Hebenstreit, D. et al. 'RNA sequencing reveals two major classes of gene expression levels in metazoan cells'. In: *Molecular Systems Biology* vol. 7, no. 1 (2011). Publisher: John Wiley & Sons, Ltd Chichester, UK, p. 497.

[56] Hoffman, M., Bach, F. R. and Blei, D. M. 'Online Learning for Latent Dirichlet Allocation'. In: *Advances in Neural Information Processing Systems 23*. Ed. by Lafferty, J. D. et al. Curran Associates, Inc., 2010, pp. 856–864. DOI: 10.5555/2997189.2997285.

[57] Hofmann, T. 'Probabilistic latent semantic indexing'. In: *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*. SIGIR '99. Berkeley, California, USA: Association for Computing Machinery, Aug. 1999, pp. 50–57. DOI: 10.1145/312624.312649.

[58] Holland, P., Laskey, K. B. and Leinhardt, S. 'Stochastic blockmodels: First steps'. In: (1983). DOI: 10.1016/0378-8733(83)90021-7.

[59] Hopfield, J. J. 'Neural networks and physical systems with emergent collective computational abilities.' In: *Proceedings of the National Academy of Sciences* vol. 79, no. 8 (Apr. 1982), pp. 2554–2558.

[60] Horr, C. and Buechler, S. A. 'Breast Cancer Consensus Subtypes: A system for subtyping breast cancer tumors based on gene expression'. eng. In: *NPJ breast cancer* vol. 7, no. 1 (Oct. 2021), p. 136. DOI: 10.1038/s41523-021-00345-2.

[61] Hoshida, Y. 'Nearest Template Prediction: A Single-Sample-Based Flexible Class Prediction with Confidence Assessment'. In: *PLoS ONE* vol. 5, no. 11 (Nov. 2010). DOI: 10.1371/journal.pone.0015543.

[62] Hric, D., Peixoto, T. P. and Fortunato, S. 'Network Structure, Metadata, and the Prediction of Missing Nodes and Annotations'. en. In: *Physical Review X* vol. 6, no. 3 (Sept. 2016), p. 031038. DOI: 10.1103/PhysRevX.6.031038.

[63] Hyland, C. C. et al. 'Multilayer networks for text analysis with multiple data types'. en. In: *EPJ Data Science* vol. 10, no. 1 (Dec. 2021), pp. 1–16. DOI: 10.1140/epjds/s13688-021-00288-5.

[64] Islam, S. et al. 'Quantitative single-cell RNA-seq with unique molecular identifiers'. In: *Nature Methods* vol. 11, no. 2 (2014). Publisher: Nature Publishing Group, p. 163.

[65] Jagadeesh, J., III, H. D. é. and Raghavendra, U. 'Incorporating Lexical Priors into Topic Models'. In.

[66] Jiang, D., Tang, C. and Zhang, A. 'Cluster analysis for gene expression data: a survey'. In: *IEEE Transactions on Knowledge and Data Engineering* vol. 16, no. 11 (2004). Publisher: IEEE, pp. 1370–1386. DOI: 10.1038/nbt1205-1499.

[67] Johnson, I., Gerlach, M. and Sáez-Trumper, D. 'Language-agnostic Topic Classification for Wikipedia'. In: *arXiv:2103.00068 [cs]* (Feb. 2021). arXiv: 2103.00068.

[68] Joulin, A. et al. 'Bag of Tricks for Efficient Text Classification'. In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. Valencia, Spain: Association for Computational Linguistics, Apr. 2017, pp. 427–431.

[69] Jr, J. H. W. 'Hierarchical Grouping to Optimize an Objective Function'. In: *Journal of the American Statistical Association* vol. 58, no. 301 (1963). Publisher: Taylor & Francis _eprint: https://amstat.tandfonline.com/doi/pdf/10.1080/01621459.1963.10500845, pp. 236–244. DOI: 10.1080/01621459.1963.10500845.

[70] Kass, R. E. and Raftery, A. E. 'Bayes Factors'. In: *Journal of the American Statistical Association* vol. 90, no. 430 (June 1995), pp. 773–795. DOI: 10.1080/01621459.1995.10476572.

[71] Kiselev, V. Y., Andrews, T. S. and Hemberg, M. 'Challenges in unsupervised clustering of single-cell RNA-seq data'. In: *Nature Reviews Genetics* vol. 20, no. 5 (2019). Publisher: Nature Publishing Group, pp. 273–282.

[72] Koboldt, D., Fulton, R., McLellan, M. et al. 'Comprehensive molecular portraits of human breast tumours'. In: *Nature* vol. 490, no. 7418 (2012). Publisher: Nature Publishing Group, p. 61. DOI: 10.1038/nature11412.

[73] Kohane, I. S. 'The twin questions of personalized medicine: who are you and whom do you most resemble?' In: *Genome Medicine* vol. 1, no. 1 (2009), pp. 4–4. DOI: 10.1186/gm4.

[74] La Vecchia, C. et al. 'Cancer mortality in Europe, 2000–2004, and an overview of trends since 1975'. en. In: *Annals of Oncology* vol. 21, no. 6 (June 2010), pp. 1323–1360. DOI: 10.1093/annonc/mdp530.

[75] Lancichinetti, A. et al. 'High-Reproducibility and High-Accuracy Method for Automated Topic Classification'. In: *Physical Review X* vol. 5, no. 1 (Jan. 2015). DOI: 10.1103/PhysRevX.5.011007.

[76] Langfelder, P. and Horvath, S. 'WGCNA: an R package for weighted correlation network analysis'. In: *BMC Bioinformatics* vol. 9, no. 1 (2008). Publisher: Springer, p. 559. DOI: 10.1186/1471-2105-9-559.

[77] Lazzardi, S. et al. 'Emergent Statistical Laws in Single-Cell Transcriptomic Data'. en. In: *bioRxiv* (June 2021), p. 2021.06.16.448706. DOI: 10.1101/2021.06.16.448706.

[78] Li Deng. 'The MNIST Database of Handwritten Digit Images for Machine Learning Research [Best of the Web]'. In: *IEEE Signal Processing Magazine* vol. 29, no. 6 (Nov. 2012), pp. 141–142. DOI: 10.1109/MSP.2012.2211477.

[79] Lonsdale, J. et al. 'The genotype-tissue expression (GTEx) project'. In: *Nature Genetics* vol. 45, no. 6 (2013). Publisher: Nature Publishing Group, pp. 580–585. DOI: 10.1038/ng.2653.

[80] Marguerat, S. et al. 'Quantitative analysis of fission yeast transcriptomes and proteomes in proliferating and quiescent cells'. In: *Cell* vol. 151, no. 3 (2012). Publisher: Elsevier, pp. 671–683. DOI: 10.1016/j.cell.2012.09.019.

[81] Mazzolini, A. et al. 'Statistics of Shared Components in Complex Component Systems'. In: *Physical Review X* vol. 8, no. 2 (Apr. 2018). Publisher: American Physical Society, p. 021023. DOI: 10.1103/PhysRevX.8.021023.

[82] Mazzolini, A. et al. 'Heaps' law, statistics of shared components, and temporal patterns from a sample-space-reducing process'. In: *Physical Review E* vol. 98, no. 5 (Nov. 2018). Publisher: American Physical Society (APS). DOI: 10.1103/physreve.98.052139.

[83] Mazzolini, A. et al. 'Zipf and Heaps laws from dependency structures in component systems'. In: *Physical Review E* vol. 98, no. 1 (July 2018). Publisher: American Physical Society, p. 012315. DOI: 10.1103/PhysRevE.98.012315.

[84] Mcauliffe, J. and Blei, D. 'Supervised Topic Models'. en. In: *Advances in Neural Information Processing Systems* vol. 20 (2007), pp. 121–128.

[85] McKinney, W. et al. 'Data structures for statistical computing in python'. In: vol. 445. 2010, pp. 51–56.

[86] Melé, M. et al. 'The human transcriptome across tissues and individuals'. In: *Science* vol. 348, no. 6235 (2015). Publisher: American Association for the Advancement of Science, pp. 660–665. DOI: 10.1126/science.aaa0355.

[87] Merkel, D. 'Docker: lightweight linux containers for consistent development and deployment'. In: *Linux Journal* vol. 2014, no. 239 (2014), p. 2.

[88] Morelli, L., Giansanti, V. and Cittaro, D. 'Nested Stochastic Block Models applied to the analysis of single cell data'. In: *BMC Bioinformatics* vol. 22, no. 1 (Nov. 2021), p. 576. DOI: 10.1186/s12859-021-04489-7.

[89] Mounir, M. et al. 'New functionalities in the TCGAbiolinks package for the study and integration of cancer data from GDC and GTEx'. In: *PLoS Computational Biology* vol. 15, no. 3 (2019). Publisher: Public Library of Science, e1006701. DOI: 10.1371/journal.pcbi.1006701.

[90] Newman, M. E. J. and Clauset, A. 'Structure and inference in annotated networks'. en. In: *Nature Communications* vol. 7, no. 1 (June 2016), p. 11863. DOI: 10.1038/ncomms11863.

[91] Newman, M. E. 'Power laws, Pareto distributions and Zipf's law'. In: *Contemporary Physics* vol. 46, no. 5 (2005). Publisher: Taylor & Francis, pp. 323–351.

[92] Nikolsky, Y. et al. 'Genome-wide functional synergy between amplified and mutated genes in human breast cancer'. eng. In: *Cancer Research* vol. 68, no. 22 (Nov. 2008), pp. 9532–9540. DOI: 10.1158/0008-5472.CAN-08-3082.

[93] Nimwegen, E. van. 'Scaling laws in the functional content of genomes'. In: *Power Laws, Scale-Free Networks and Genome Biology* (2006). Publisher: Springer, pp. 236–253.

[94] Osella, M. et al. 'Interplay of microRNA and epigenetic regulation in the human regulatory network'. eng. In: *Frontiers in Genetics* vol. 5 (2014), p. 345. DOI: 10.3389/fgene.2014.00345.

[95] Pang, T. Y. and Maslov, S. 'Universal distribution of component frequencies in biological and technological systems'. In: *Proceedings of the National Academy of Sciences* vol. 110, no. 15 (Apr. 2013). Publisher: National Academy of Sciences, pp. 6235–6239. DOI: 10.1073/pnas.1217795110.

[96] Papadopoulos, G. L. et al. 'The database of experimentally supported targets: a functional update of TarBase'. eng. In: *Nucleic Acids Research* vol. 37, no. Database issue (Jan. 2009), pp. D155–158. DOI: 10.1093/nar/gkn809.

[97] Parker, J. S. et al. 'Supervised Risk Predictor of Breast Cancer Based on Intrinsic Subtypes'. In: *Journal of Clinical Oncology* vol. 27, no. 8 (Mar. 2009), pp. 1160–1167. DOI: 10.1200/JCO.2008.18.1370.

[98] Pedregosa, F. et al. 'Scikit-learn: Machine Learning in Python'. In: *Journal of Machine Learning Research* vol. 12, no. 85 (2011), pp. 2825–2830.

[99] Peixoto, T. P. 'The graph-tool python library'. In: *figshare* (2014). DOI: 10.6084/m9.figshare.1164194.

[100] Peixoto, T. P. 'Hierarchical block structures and high-resolution model selection in large networks'. In: *Physical Review X* vol. 4, no. 1 (2014). DOI: 10.1103/PhysRevX.4.011047.

[101] Peixoto, T. P. 'Nonparametric Bayesian inference of the microcanonical stochastic block model'. In: *Physical Review E* vol. 95, no. 1 (2017). Publisher: APS, pp. 12317–12317.

[102] Peixoto, T. P. 'Efficient Monte Carlo and greedy heuristic for the inference of stochastic block models'. In: *Physical Review E* vol. 89, no. 1 (Jan. 2014), p. 012804. DOI: 10.1103/PhysRevE.89.012804.

[103] Peixoto, T. P. 'Model Selection and Hypothesis Testing for Large-Scale Network Models with Overlapping Groups'. In: *Physical Review X* vol. 5, no. 1 (Mar. 2015). Publisher: American Physical Society, p. 011033. DOI: 10.1103/PhysRevX.5.011033.

[104] Peixoto, T. P. 'Merge-split Markov chain Monte Carlo for community detection'. In: *Physical Review E* vol. 102, no. 1 (July 2020), p. 012305. DOI: 10.1103/PhysRevE.102.012305.

[105] Peixoto, T. P. 'Revealing Consensus and Dissensus between Network Partitions'. In: *Physical Review X* vol. 11, no. 2 (Apr. 2021), p. 021003. DOI: 10.1103/PhysRevX.11.021003.

[106] Perou, C. et al. 'Molecular portraits of human breast tumours'. In: *Nature* vol. 406, no. 6797 (Aug. 2000), pp. 747–752. DOI: 10.1038/35021093.

[107] Piccardi, T. and West, R. 'Crosslingual Topic Modeling with WikiPDA'. In: *arXiv:2009.11207 [cs]* (Feb. 2021). arXiv: 2009.11207. DOI: 10.1145/3442381.3449805.

[108] Picelli, S. et al. 'Full-length RNA-seq from single cells using Smart-seq2'. In: *Nature Protocols* vol. 9, no. 1 (2014). Publisher: Nature Publishing Group, pp. 171–181.

[109] Prat, A. et al. 'PAM50 assay and the three-gene model for identifying the major and clinically relevant molecular subtypes of breast cancer'. In: *Breast Cancer Research and Treatment* vol. 135, no. 1 (2012). Publisher: Springer, pp. 301–306.

[110] Prat, A. and Perou, C. M. 'Deconstructing the molecular portraits of breast cancer'. In: *Molecular Oncology* vol. 5, no. 1 (2011). Publisher: Elsevier, pp. 5–23.

[111] Reale, E. et al. 'Investigating the epi-miRNome: identification of epi-miRNAs using transfection experiments'. eng. In: *Epigenomics* vol. 11, no. 14 (Nov. 2019), pp. 1581–1599. DOI: 10.2217/epi-2019-0050.

[112] Rissanen, J. 'Modeling by shortest data description'. en. In: *Automatica* vol. 14, no. 5 (Sept. 1978), pp. 465–471. DOI: 10.1016/0005-1098(78)90005-5.

[113] Ronde, J. J. de et al. 'Concordance of clinical and molecular breast cancer subtyping in the context of preoperative chemotherapy response'. In: *Breast Cancer Research and Treatment* vol. 119, no. 1 (Jan. 2010), pp. 119–126. DOI: 10.1007/s10549-009-0499-6.

[114] Rosenberg, A. and Hirschberg, J. 'V-measure: A conditional entropy-based external cluster evaluation measure'. In: *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. 2007, pp. 410–420.

[115] Schumacher, C., Vose, M. D. and Whitley, L. D. 'The No Free Lunch and Problem Description Length'. In: *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO-2001*. Morgan Kaufmann, 2001, pp. 565–570.

[116] Segura, M. L. R. T. et al. 'A 3D transcriptomics atlas of the mouse nose sheds light on the anatomical logic of smell'. English. In: *Cell Reports* vol. 38, no. 12 (Mar. 2022). DOI: 10.1016/j.celrep.2022.110547.

[117] Shi, H. et al. 'A new evaluation framework for topic modeling algorithms based on synthetic corpora'. In: *Proceedings of Machine Learning Research*. Ed. by Chaudhuri, K. and Sugiyama, M. Vol. 89. Proceedings of Machine Learning Research. PMLR, Apr. 2019, pp. 816–826.

[118] Silva, T. C. et al. 'TCGAbiolinksGUI: A graphical user interface to analyze cancer molecular and clinical data'. In: *F1000Research* vol. 7, no. 439 (2018). Publisher: F1000 Research Limited, p. 439. DOI: 10.12688/f1000research.14197.1.

[119] Smid, M. et al. 'Subtypes of breast cancer show preferential site of relapse'. In: *Cancer Research* vol. 68, no. 9 (May 2008), pp. 3108–3114. DOI: 10.1158/0008-5472.CAN-07-5644.

[120] Sorlie, T. et al. 'Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications'. In: *Proceedings of the National Academy of Sciences* vol. 98, no. 19 (Sept. 2001), pp. 10869–10874. DOI: 10.1073/pnas.191367098.

[121] Stoeger, T. et al. 'Large-scale investigation of the reasons why potentially important genes are ignored'. In: *PLoS Biology* vol. 16, no. 9 (Sept. 2018). Publisher: Public Library of Science, e2006643.

[122] Subramanian, A. et al. 'Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles'. In: *Proceedings of the National Academy of Sciences* vol. 102, no. 43 (2005). Publisher: National Academy of Sciences _eprint: https://www.pnas.org/content/102/43/15545.full.pdf, pp. 15545–15550. DOI: 10.1073/pnas.0506580102.

[123] Tabula Muris Consortium and others. 'Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris.' In: *Nature* vol. 562, no. 7727 (2018), p. 367.

[124] Tokar, T. et al. 'mirDIP 4.1—integrative database of human microRNA target predictions'. In: *Nucleic Acids Research* vol. 46, no. D1 (Jan. 2018), pp. D360–D370. DOI: 10.1093/nar/gkx1144.

[125] Ueda, H. R. et al. 'Universality and flexibility in gene expression from bacteria to human'. In: *Proceedings of the National Academy of Sciences* vol. 101, no. 11 (2004). Publisher: National Academy of Sciences _eprint: https://www.pnas.org/content/101/11/3765.full.pdf, pp. 3765–3769. DOI: 10.1073/pnas.0306244101.

[126] Valle, F. *nSBM: multi branch topic modeling*. June 2021. DOI: 10.5281/zenodo.5045446.

[127] Valle, F., Osella, M. and Caselle, M. 'A Topic Modeling Analysis of TCGA Breast and Lung Cancer Transcriptomic Data'. en. In: *Cancers* vol. 12, no. 12 (Dec. 2020), p. 3799. DOI: 10.3390/cancers12123799.

[128] Valle, F., Osella, M. and Caselle, M. 'Multiomics Topic Modeling for Breast Cancer Classification'. en. In: *Cancers* vol. 14, no. 5 (Feb. 2022), p. 1150. DOI: 10.3390/cancers14051150.

[129] Vallès-Català, T. et al. 'Multilayer Stochastic Block Models Reveal the Multilayer Structure of Complex Networks'. In: *Physical Review X* vol. 6, no. 1 (Mar. 2016), p. 011036. DOI: 10.1103/PhysRevX.6.011036.

[130] Van der Maaten, L. and Hinton, G. 'Visualizing data using t-SNE.' In: *Journal of machine learning research* vol. 9, no. 11 (2008).

[131] Wang, Q., Gao, J. and Schultz, N. 'Unified RNA-seq datasets in human cancers and normal tissues - normalized data'. In: (2017). DOI: 10.6084/m9.figshare.5330593.v2.

[132] Wang, Q. et al. 'Unifying cancer and normal RNA sequencing data from different sources'. In: *Scientific Data* vol. 5 (2018). Publisher: Nature Publishing Group, p. 180061. DOI: 10.1038/sdata.2018.61.

[133] Wang, X. et al. 'Direct Comparative Analyses of 10X Genomics Chromium and Smart-seq2'. In: *Genomics, Proteomics & Bioinformatics* vol. 19, no. 2 (Apr. 2021), pp. 253–266. DOI: 10.1016/j.gpb.2020.02.005.

[134] Wolf, F. A., Angerer, P. and Theis, F. J. 'SCANPY: large-scale single-cell gene expression data analysis'. In: *Genome Biology* vol. 19, no. 1 (2018). Publisher: BioMed Central, p. 15. DOI: 10.1186/s13059-017-1382-0.

[135] Yen, T.-C. and Larremore, D. B. 'Community detection in bipartite networks with stochastic block models'. In: *Physical Review E* vol. 102, no. 3 (Sept. 2020), p. 032309. DOI: 10.1103/PhysRevE.102.032309.

[136] Zhang, B. and Horvath, S. 'A general framework for weighted gene co-expression network analysis'. In: *Statistical Applications in Genetics and Molecular Biology* vol. 4 (2005). DOI: 10.2202/1544-6115.1128.

[137] Zhou, W. et al. 'An overview of topic modeling and its current applications in bioinformatics'. In: *SpringerPlus* (2016). DOI: 10.1186/s40064-016-3252-8.

[138] Zipf, G. K. *Human behavior and the principle of least effort: An introduction to human ecology.* Ravenio Books, 2016.

Acknowledgements

Non posso che iniziare gli acknoledgments ringraziando Anna, fidanzata con cui ho condiviso infinite avventure, traslochi, proposte e organizzazione di grandi salti durante il dottorato.

Un grazie a tutti i miei parenti: Lucia, Giorgio, Paola, Antonella e Manlio, Armando, Guido. I miei nonnə Lina e Battista. Mario, Antonella e Nicolò. Tutti voi mi siete, in un modo o nell'altro stati vicino e grazie per aver continuato a supportarmi durante questi anni. I miei (quasi) suocerə Ricky e Manu e i tutti rami di parentela che si aggiungeranno presto!

Un saluto e ringraziamento ai miei amici e amiche senza i qualə questi anni sarebbero stati sicuramente più faticosi: Simone, il primo che ho incontrato a Fisica e la cui amicizia sarà sempre forte e resisterà alla sfide più difficili; Lele amico sempre disponibile a fare due chiacchere, una cena o una passeggiata a Belmonte. Flavio, Marco, Tiziano, Elisa, Laura ed Elan con cui ho condiviso inenarrabili viaggi sulla canavesana durante Fisica e altretttanto inenarrabili cene durante il dottorato.

Un saluto speciale a Gabriele amico[1] dentro e fuori Fisica ed ad Enrica, amica scoperta in questi anni.

Un pensiero alle persone (vicine di casa o "carcerati") che hanno reso Alessandria un posto un po' meno grigio da cui pendolare.

Infine ringrazio l'intero gruppo ByoPhys
http://personalpages.to.infn.it/~caselle/BioPhys/BioPhys.html[2]: Michele Caselle per avermi instradato nella tesi prima e nel dottorato poi e per avermi supportato anche molto oltre le sue aree di competenza. Un grazie a Matteo, supervisor in cui ho trovato un amico, il cui atteggiamento critico e scientifico[3] mi sarà di insegnamento a lungo. Tutte le persone con cui ho avuto modo di collaborare in questi tre anni, in particolare Marco Gherardi, Loredana Martignetti e Antonio Scialdone. Un pensiero ai vari tesistə i cui progetti si sono intersecati ai miei: mentre vi aiutavo con qualche riga di codice o a fare funzionare OCCAM mi avete insegnato molto e avete arricchito il mio percorso: Elisa, Riccardo, Stefano, Alessio, Andreina, Gabriele, grazie! Marta, Silvia, Letizia e Antonio, è stato bello condividere group meeting e l'ufficio con voi!

[1] co-relatore, cotitolare di un marchio, testimone, testimoniato e CEO
[2] durante il postdoc fatemi rifare il sito
[3] e i plot con i pallini grandi