UNIVERSITY OF CAMBRIDGE

Interpreting Deep Learning for cell differentiation

Supervised and Unsupervised models viewed through the lens of information and perturbation theory

Helena Andrés Terré

Newnham College

This dissertation is submitted on November, 2019 for the degree of Doctor of Philosophy

Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text. It is not substantially the same as any that I have submitted, or am concurrently submitting, for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my dissertation has already been submitted, or is being concurrently submitted, for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. This dissertation has 35,789 words, therefore does not exceed the prescribed limit.

Helena Andrés Terré
November, 2019

Abstract

"Predicting the future isn't magic, it's artificial intelligence" Dave Waters.

In the last decades there has been an unprecedented growth in the field of machine learning, and particularly within deep learning models. The combination of big data and computational power has nurtured the evolution of a variety of new methods to predict and interpret future scenarios. These data centric models can achieve exceptional performances on specific tasks, with their prediction boundaries continuously expanding towards new and more complex challenges.

However, the model complexity often translates into a lack of interpretability from a scientific perspective, it is not trivial to identify the factors involved in final outcomes. Explainability may not always be a requirement for some machine learning tasks, specially when it comes in detriment of performance power. But for some applications, such as biological discoveries or medical diagnostics, understanding the output and determining factors that influence decisions is essential.

In this thesis we develop both a supervised and unsupervised approach to map from genotype to phenotype. We emphasise the importance of interpretability and feature extraction from the models, by identifying relevant genes for cell differentiation. We then continue to explore the rules and mechanisms behind the models from a theoretical perspective. Using information theory to explain the learning process and applying perturbation theory to transform the results into a generalisable representation.

We start by building a supervised approach to mapping cell profiles from genotype to phenotype, using single cell RNA-Seq data. We leverage non-linearities among gene expressions to identify cellular levels of differentiation. The ambiguity and even absence of labels in most biological studies instigated the development of novel unsupervised techniques, leading to a new general and biologically interpretable framework based on Variational Autoencoders.

The application and validation of the methods has proven to be successful, but questions regarding the learning process and generative nature of the results remained unanswered. I use information theory to define a new approach to interpret training and the converged solutions of our models.

The variational and generative nature of Autoencoders provides a platform to develop

general models. Their results should extrapolate and allow generalisation beyond the boundaries of the observed data. To this extent, we introduce for the first time a new interpretation of the embedded generative functions through Perturbation Theory. The embedding multiplicity is addressed by transforming the distributions into a new set of generalisable functions, while characterising their energy spectrum under a particular energy landscape.

We outline the combination of theoretical and machine learning based methods, for moving towards interpretable and generalisable models. Developing a theoretical framework to map from genotype to phenotype, we provide both supervised and unsupervised tools to operate over single cell RNA-Seq. data. We have generated a pipeline to identify relevant genes and cell types through Variational Autoencoders (VAEs), validating reconstructed gene expressions to prove the generative performance of the embeddings. The new interpretation of the information learned and extracted by the models defines a label independent evaluation, particularly useful for unsupervised learning. Lastly, we introduce a novel transformation of the generative embeddings based on quantum and perturbation theory.

Our contributions can and have been extended to new datasets, according to the nature of the tasks being explored. For instance, the combination of unsupervised learning and information theory can be applied to a variety of biological or medical data. We have trained several VAE models with additional cancer and metabolic data, proving to extract meaningful representations of the data. The perturbation theory transformation of the embedding can also lead to future research on the generative potential of Variational Autoencoders through a physics perspective, combining statistical and quantum mechanics.

We believe that machine learning will only continue its fast expansion and growth through the development of more generalisable more interpretable models.

"Prediction is very difficult, especially if it's about the future" Niels Bohr

Acknowledgements

This thesis is the result of the most challenging and fruitful time on my career, where I have been able to grow as a scientist and interact with many inspiring and brilliant people.

First I would like to thank my supervisor, Pietro Lió, for giving me this opportunity and all the support in the past years. After many creative ideas and inspiring conversations, his trust encouragement have been essential for my development as a scientist. From him I have learned to see the exciting and fun side of research, and the benefits of always keeping an open mind.

This thesis has been shaped in many ways through the interactions with many brilliant collaborators. A special acknowledgment to Nikola, Zohreh, Ioana, Ramon, Ben, Paul, Ifrah and Cristian, who among others have contributed in many ways on forging the science developed during my PhD. I would also like to thank all the past and present members of the AI group at the Computer Laboratory in the University of Cambridge, for their constant support and inspiration.

I will forever be grateful for all the friends, new and old, that have been with me in this journey. To Dan, Eva, Rebecca, Adriana, Julia, Emilia, Ana, and all the people I have had the privilege to meet during my time in Cambridge. Because every coffee and conversation had, has shaped me into the person I am now. To my frisbee family, responsible for most of my fun and cheerful times in the last years. Tessa, Tom, and all the Cambridge contingent for uncountable laughs and many wet and cold training sessions. To all the SYC ladies, for being a team full of inspiring women and athletes. To team Spain and the community, for giving me the strength and warmth needed to get through difficult times. And to Marina, Julia, Anna and my friends back in Barcelona, for being there from the beginning and driving along this journey with me. I am also thankful and proud of being a part of the strong and inspiring community of women at Newnham college.

Finally, it would not have been possible to reach this point on my career with the infinite love and affection given by my family. My parents, who have always believed in me and supported any path I have decided to take. To my sister Marta, for always being an inspiration and challenging me to go one step further. I will be forever grateful for their time, dedication and unconditional love.

Contents

CHAPTER 1

Introduction

Understanding, quantifying and predicting natural world phenomena are some of the main purposes of science. Their aim is to create a universal body of empirical, theoretical and practical knowledge, that changes and expands while we learn from our surroundings.

Technology has grown along and broadly benefited from scientific advancements. But is has also played an essential role in many discoveries and scientific accomplishments. Both science and technology have jointly moved forward during history in different fields, sharing common goals but often approaching them from different perspectives.

For instance, from the furthest star to the tiniest atom, optics laid the foundations to then develop powerful imaging techniques that provided new evidences to a range of other scientific fields.

All branches of science are interconnected, and leveraging the overlap between them is what ultimately has allowed some of the greatest advances in human history.

In recent years, the computing power and resources have increased exponentially [66], reaching unexpected levels and providing a powerful platform for testing hypothesis and proving theories. That is the case of machine learning and artificial intelligence, the fundamental theories of which are believed to have their origin in the 19th century with the introduction of Bayes's theory (1812) and Markov Chains (1913). It wasn't until the 1950s that Alan Turing proposed the "learning machine" which would contribute to the development of genetic algorithms. The first Neural Network Machine and the Perceptron were also incepted in that decade, and 20 years later the foundations of back-propagation were published under the name of "Automatic Differentiation". But due to a lack of computational power and insufficient technological resources, the resurgence and application of machine learning didn't take place until the late 80s. It was then when, mostly guided by a data-driven approach, scientists started creating computer programs to analyse large amounts of data. Techniques such as Support Vector Machines (SVMs), Recurrent Neural Networks (RNNs), and later on deep, reinforcement and unsupervised learning became

more widespread. Currently, the integration of deep learning techniques with some of the most advanced algorithms and powerful computers has lead to programs able to beat humans in performance on some specific tasks [43, 90–92], and image recognition techniques with a direct application for medical or commercial purposes among many others.

Although these techniques display incredibly good performances for specific tasks, they haven't been followed by a strong theoretical framework to explain such results. The fast technological development has been based on a data-driven approach, without a proper statistical or mathematical interpretation, coining the term "black box" among the scientific community when referring to some of these models.

High dimensionality and complexity are some of the challenges of deep learning that still need to be tackled in order to gain interpretability of the results. From an analytical perspective, understanding how the learning process occurs, or obtaining an explicit mapping between inputs and outputs of the model could aid this purpose.

As it has happened before in the history of science and technology, the solution to some of these questions may have its origin in a non directly related field. Figure 1.1 highlights the importance of developing a solid core knowledge for machine learning, and its impact over many applications. A multidisciplinary approach and a different perspective, followed by a rigorous mathematical formulation, may have the key towards a more explainable and interpretable AI.

It was in fact the collaboration between a physicist and a biologist that lead to one of the ground breaking scientific findings of the last decade. Since the discovery of DNA by James Watson and Francis Crick in 1953, scientists have been intrigued and tried to decipher the genetic code embedded in every single cell of our body. A great amount of data and the extremely complex mechanisms involved in any biological system, makes these tasks incredibly difficult to tackle by a simple human brain without the help of some technology. From microscopes to the most advanced genome-reading techniques, biology has been evolving at a very high pace during the last decades, reaching some precision levels and amounts of data never achieved before.

Large sized and multi-modal datasets have launched the development of new technologies, drugs and a better understanding of the highly complex systems that are living organisms. The techniques used for analysis are diverse, and adaptable to data properties such as sparsity or linearity, all fitted to achieve the final goals.

One of the most recent and ground-breaking experimental advances in the last decade has been single cell analysis. With the introduction of single cell RNA-Seq, we have now the opportunity to access genetic information stored at individual cell level. With greater granularity, a next generation of genome level studies has been reached, so new challenges on the analysis and interpretation side have been introduced. The methods developed

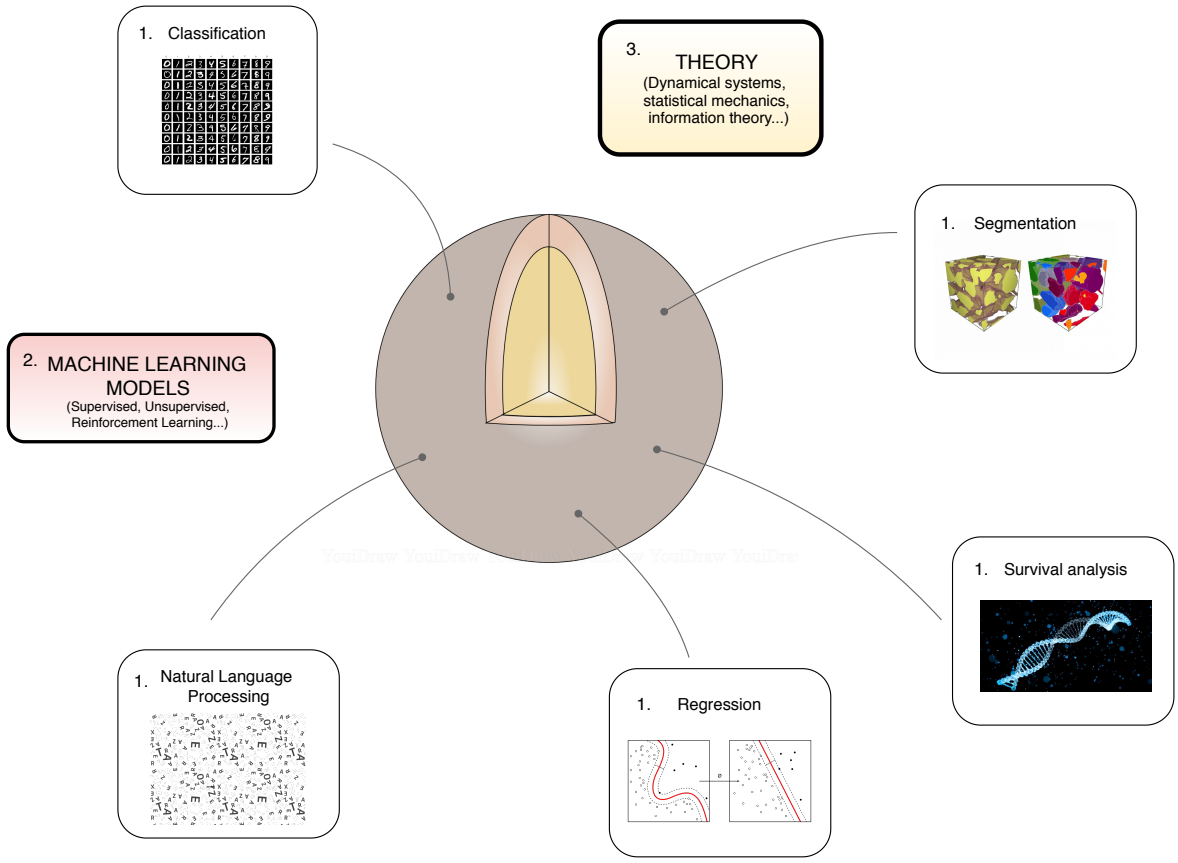

Figure 1.1: ***Multidisciplinary approach to machine learning.*** *Machine learning and its applications can be understood as a layered sphere.* ***(1)*** *The external layer contains all the practical problems, from classification to segmentation, together with non-linear regressions, survival analysis or natural language processing.* ***(2)*** *Data driven models are situated on the first inner layer, where some of them can be applied to solve several practical problems. For instance, both segmentation and classification can be tackled using convolutional neural networks.* ***(3)*** *In the core we find the theoretical interpretation to machine learning. Theories from dynamical systems or information theory are used to describe the evolution and results of models. Ideally, they must be generalisable and universal for all machine learning techniques.*

for bulk genome analysis are not prepared to deal with single cell data and exploit all its capabilities. High dimensionality and sparsity of the data are some of the challenging features of these datasets. Some advanced statistical and machine learning methods have been shown successful for processing and dealing with these properties.

Biological data is known to be multi-scale and complex, with different processes and levels of information encoded on top of each other. Extracting the relevant inputs for a particular target has proven to be possible using machine learning when the amount and quality of data is enough, but in order to interpret such results and build a model that can reproduce and explain such findings there is a need for more robust and generalisable methods.

When dealing with biological data, special attention needs to be given to the inter-

pretability of the results. As important as building models that can correctly identify and predict specific outcomes, it is often needed to understand how are such results achieved. The lack of transparency of some machine learning models has become a major drawback when applied to some areas of biological and health sciences.

1.1 Motivation and Research Questions

This thesis combines a theoretical and practical approach to machine learning, while building a framework for single cell genome analysis. Results are interpreted from both mathematical and biological perspectives. The combination of theory and application, aimed towards more transparent and accessible models, provides a deeper understanding of the outcomes and predictions of such models. It also enables the analysis and selection of the best approach for each application. We have borrowed techniques from computer science and physics, such as information theory or perturbation analysis, to explore in depth the learning process and results. That allows for a better understanding of the outputs and helps optimising models for data extraction.

We introduce supervised and unsupervised learning approaches to tackle biological datasets for specific tasks. Supervised techniques with genome data, used for classification, yield very positive results and have proved to be very successful when prior knowledge and labels are available. Unfortunately, that is not always the case in biology, and we decided to extended the problems so they can be solved within an unsupervised framework. The lack of specificity in the tasks aimed to be fulfilled from an unsupervised approach adds a new layer of difficulty in the evaluation front. Using labels or prior knowledge can be seen as a contradiction to their unsupervised nature, as we try to avoid sacrificing generality. Therefore, there is a need to define a framework where both application and fundamental or theoretical results could be understood, and for that we saw information theory as a good resource. Entropies and information loss can be used to describe the learning process, and analyse the quality of the results.

Finally, we also studied some of the learned outputs and developed a general approach to generative functions and energy landscapes. Our method is inspired by perturbation theory from quantum physics, where we define a set of perturbed and unperturbed wave-functions associated to the different energy states of the original system.

Given the general aim and the multidisciplinary nature of this thesis, the research questions and contributions fall into one or more of these three main categories: biological (B), technical or machine learning based (ML) and theoretical (T).

1. Map from single cell genotype data to phenotype B ML

We introduce a supervised and unsupervised learning approach to analyse single cell RNA-Seq data. The final goal of both methodologies is very similar, to extract relevant attributes from genetic data in order to identify cell types and phenotypical traits. But the means are different, and therefore their evaluation is not the same.

In Chapter 3 the models are optimised to maximise classification accuracy, based on a set of labels already provided. Chapter 4 gives a less restricted mapping between gene expression and phenotype spaces.

We developed an unsupervised pipeline to create an embedded generative space directly from the genotype, in which one can successfully identify common traits and features from the data.

2. Identifying relevant genes and markers from the B

A second mapping, between the lower dimensional embedding and a classification or hierarchy among samples, can be validated by combining feature extraction and biological knowledge extracted from the literature. We use the results obtained from the unsupervised learned models to detect relevant genes for each group of phenotypically similar cells. It is possible to then use such genes to validate the solutions, and compare to further experimental observations.

3. Use information theory to analyse the learning process and optimise models ML T

We develop an information theory approach to tackle the explainability gap. Both supervised and unsupervised methods are evaluated using entropy based metrics. Information flow between the input and output layers, and model convergence are some of the features studied during training. We also analyse the relation between information and classification accuracy for supervised learning, together with their relation to disentanglement of the embedding in VAEs.

4. Build a framework to study VAEs based on perturbation theory from physics. T

Ideally all models should converge to the same solution, but the final trained networks always show different distributions, even though they are often able to maintain their accuracy levels. The fact that many independently trained models yield the same outcome can be understood under the umbrella of reproducibility in science. In particular, each VAE produces a set of generative functions that characterises the same system. Each label or cluster of the system can be seen as an energy level. When each energy level can be associated to more than one conformation or function, it generates a particular problem known as degeneration in quantum physics. In order to solve this problem, perturbation theory has been used to identify a set of eigen-functions and eigen-values that characterise the system. We used this approach to generalise the results of unsupervised learning, testing for perturbed energy potentials, and finding their corresponding energy spectrum and wave-functions.

1.2 Thesis overview

The thesis is structured as follows, starting from a data driven perspective and moving towards a theoretical approach. A visual representation of the thesis structure is displayed in figure 1.2.

Chapter 2 - Background This chapter is divided into three main sections; biology, machine learning and physics content. It introduces the main concepts and ideas needed to understand the goals and scope of this work. In the biology section there is information about the experimental design, data acquisition and processing. The biological processes involved and current hypothesis or models are also described, together with some of the computational techniques and methodologies developed to study single cell RNA-Seq and haematopoiesis. The machine learning section introduces some of the models used for supervised and unsupervised learning. It gives an historical and methodological overview, that will then be further developed in the following chapters according to their particular application. The last section presents some of the intersections between physics and machine learning, explaining some of the studies based on information theory such as the information bottleneck, and some applications on quantum physics and statistical mechanics.

Chapter 3 - Supervised Learning This chapter displays the work and results obtained from a primarily data driven approach on single cell RNA-Seq data, to study haematopoiesis. Given a set of cells or samples, characterised in a highly dimensional space defined by their individual gene expressions, we map them to a lower dimensional space that corresponds to their phenotype. Provided that we have their labels, we use a deep neural network classifier to predict cell types. More specifically, we were interested in identifying the stem cells and analyse their level of differentiation, based solely on genome data. We developed a "stemness" measure, optimising the architecture and models to obtain the maximum classification accuracy. The results are compared and discussed against the ones obtained using other computational methods.

Chapter 4 - Unsupervised Learning This chapter develops and studies a more general perspective on the analysis of single cell RNA-Seq. We use Variational Auto-Encoders to reduce the dimensionality and find a generative embedding based on gene expression data. In the first section, we present a general pipeline to map from genotype to phenotype from a completely unsupervised approach using the latent space. It describes and evaluates each stage based on classification accuracy and feature extraction, comparing the clusters detected and their corresponding marker genes to those reported in the literature.

In the second section, we use the Turing Test to prove that the reconstructed gene expression decoded from the VAE embedding is equivalent to the input data, and successfully preserves all of its properties and structure.

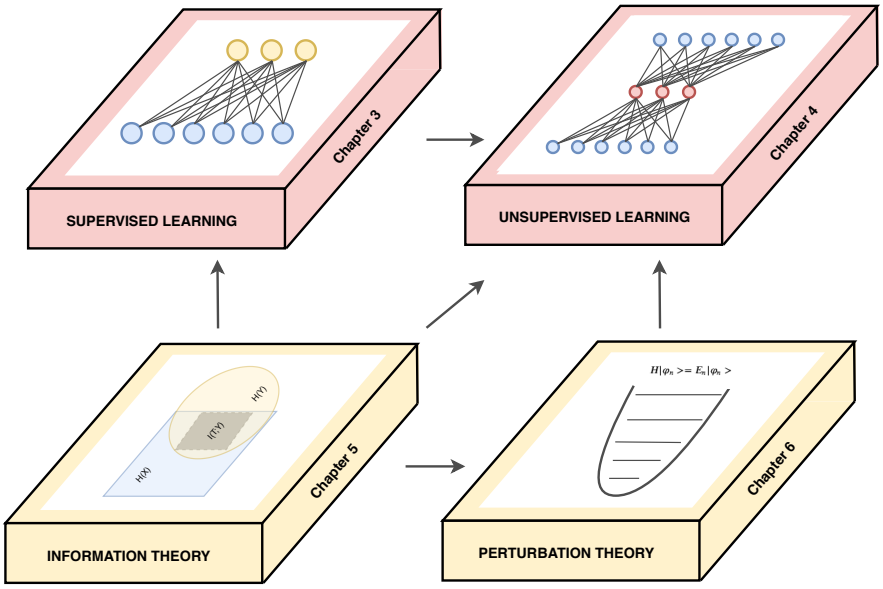

Figure 1.2: ***Structure of the thesis.*** *Chapter 2 introduces the fundamental ideas from biology, machine learning and physics needed in the thesis. Chapter 3 and 4 are data driven, exploring the mapping between genotype and phenotype from a supervised and unsupervised perspectives. Chapter 5 provides a different angle to model evaluation, using information theory, and is applied to the techniques introduced in both previous chapters. Chapter 6 presents a novel interpretation to the solutions generated from unsupervised learning, tackling degeneracy using perturbation theory, to achieve general embeddings.*

Chapter 5 - Information Theory This chapter is based on the need of an evaluation framework not based on a set of pre-defined labels. From an unsupervised perspective, using classification accuracy to evaluate the performance of our models can be seen as redundant, and even constrain the outcome of the analysis. We introduce the different measures and analysis based on an information theory approach, mainly applied to; study the information flow and loss during training for supervised and unsupervised learning, feature extraction and disentanglement among embedding components for unsupervised learning.

Chapter 6 - Perturbation Theory This chapter considers the problem of degeneracy between VAE embeddings for a particular system. It provides a theoretical interpretation of the latent space, and aims to achieve a universal lower dimensional and generative representation of the data. It re-formulates the generative embedding of VAEs, and provides a transformation to a set of characteristic wave-functions that are only dependent

on the system structure. Based on perturbation theory from quantum physics, it develops a method to identify eigen-functions and eigen-values related to the wave-functions and energy spectrum of the system. Several energy potentials are tested and the optimal ones are used to derive the final solution.

1.3 Related publications

From the research developed in this thesis, we have written and presented several articles. The most relevant ones related to this particular work are:

a. Athanasiadis, E. I., Botthof, J. G., Andres, H., Ferreira, L., Lio, P., Cvejic, A. (2017). ***Single-cell RNA-sequencing uncovers transcriptional states and fate decisions in haematopoiesis.*** Nature communications, 8(1), 2045.

b. Bica, I. , Andres-Terre, H., Cvejic, A., Lio, P. (2019). ***Unsupervised generative and graph representation learning for modelling cell differentiation.*** bioRxiv. Scientific Reports (accepted with major revision)

c. Simidjievski, N. , Bodnar, C. , Tariq, I. , Scherer, P. , Andres-Terre, H., Shams, Z., Jamnik, M. , Lio, P. (2019). ***Variational Autoencoders for Cancer Data Integration: Design Principles and Computational Practice.*** Frontiers in Genetics (accepted).

d. Barsacchi, M., Andres-Terre, H., Li, P. (2018). ***GEESE: Metabolically driven latent space learning for gene expression data.*** bioRxiv, 365643.

e. Andres-Terre, H. Lio, P. (2019). ***Information theory of deep learning for mapping from single cell genome.*** Submitted to MLCB2019

f. Andres-Terre, H. Lio, P. (2019). ***Perturbation theory approach to study the latent space degeneracy of Variational Autoencoders.*** arXiv, preprint arXiv:1907.05267

g. Webb, E., Day, B., Andres-Terre, H., Li, P. (2019). ***Factorised Neural Relational Inference for Multi-Interaction Systems.*** arXiv, preprint arXiv:1905.08721. ICML 2019 Workshop on Learning and Reasoning with Graph-Structured Representations

Publication (a) was the first collaboration where we used single cell RNA-Seq data, where we analysed the level of differentiation of individual cells. It includes the work presented in Chapter 3 on Supervised Learning.

Preprint (b) has been submmitted to Scientific Reports and it has been accepted with major revision. It contains the pipeline developed for single cell RNA-Seq analysis described in Chapter 4. This work has been previously presented at the International Conference of Complex Systems (2018) and the WiML Workshop (2019) at NeurIPS.

Paper (c) has been accepted by Frontiers in Genetics. We used the techniques developed in Chapter 4 to analyse and evaluate different integration approaches for cancer data.

In preprint (d) we explored the latent space generated by Variational Autoencoders when applied to metabolic datasets, and their combination with Flux Balance Analysis.

The content of (e) has part of the results obtained from to the Information Theory approach introduced in Chapter 5. It has been submitted to the Machine Learning in Computational Biology conference and is currently under review.

Manuscript (f) contributes to the Perturbation Theory approach, which is explained in Chapter 6.

The work developed in Chapter 5 has allowed and nourished some contributions with different applications, such as the ones presented in (d) and (g).

CHAPTER 2

Background

The research developed in the chapters of this thesis has been designed as a general framework, such that it can be applied to a wide range of datasets and different scientific questions.

However, most results are tested and validated using the particular biological dataset on haematopoiesis that inspired our first supervised approach. Therefore, in this chapter we give an overview of the biological background and context in which our methods are based. We describe the main questions tackled, characterise the data and resources available nowadays.

We also introduce basic concepts on the theoretical and computational side, explaining some of the techniques used together with the fundamentals on information and perturbation theory.

2.1 Introduction to Single Cell

An average human body is estimated to have approximately 37.2 trillion cells [10]. Even though the majority of them share the exact same genetic code, stored in their nuclei as DNA, they can have significantly different functions and shapes. The activity of these genes is what determines their fate and individually characterises each cell. Transcriptome information obtained through gene sequencing was originally developed to analyse the expression of a subset or the entire genotype for bulk populations, assuming that cells with common features also express a similar transcriptome. Even though that is sometimes the case, evidence has shown that there is also a non-negligible level of heterogeneity among ensembles that can reflect cell type composition and even trigger cell fate decisions [24, 40, 56, 59, 87].

The analysis of entire transcriptomes at a single-cell level was only first introduced over two decades ago by James Eberwine et al. [22], and Iscove and colleagues [15]. But it

wasn't until 2009 that the first single-cell transcriptome analysis based on a next-generation sequencing platform was made public [98].

Being able to characterise cells at an individual level adds an extra layer of complexity when mapping from genotypes to phenotypes. But it also has the potential to address new biological questions that would otherwise be impossible to answer, such as identifying rare populations or outliers for drug resistance and treatment relapse [85], or deconvolute diverse immune cell populations in healthy and diseased states [86].

In the last decade there has been a blooming interest in analysing and monitoring heterogeneity at a single-cell level on a global scale. As the experimental techniques have rapidly improved, computational power and tools have also been developed to store and handle large amounts of data.

Next generation sequencing platforms have revolutionised genomic research, by performing sequencing of millions of small fragments of DNA in parallel. This allows a much faster and accurate aquisition of data, while providing an insight to the DNA variation due to multiple sequencing of gene bases [8].

RNA-seq uses next-generation sequencing to account for the quantity of RNA in a biological sample. Before single cell granularity, transcriptome analysis was performed by using DNA-microarrays, which measure the expression levels of large numbers of genes simultaneously by using a collection of microscopic DNA spots attached to a solid surface.

As opposed to DNA-microarray analysis, RNA-seq performs a direct sequencing and doesn't rely on pre-defined sequences, avoiding related biases introduced during hybridisation of microarrays. It has a high dynamic range, and significantly reduces the variability of lowly and highly expressed genes compared to microarrays. Based on their technology and acquisition method, RNA-seq is particularly useful to analyse changes in gene expression over time, or differences among groups or treatments, as it provides higher sensitivity, a wider range of expression and is not bounded to existing genomic sequencing information [117].

The experimental procedure to perform single cell RNA-Seq always follow the same pipeline, although some may differ on the particular techniques and tools used in each step. Figure 2.1 displays the general workflow for single cell RNA-Seq experiments.

The first or initial step is to dissociate and isolate cells from tissue samples. This can be done in several ways, including micro-dissection and manipulation, flow cytrometric cell-sorting, microfluidic platforms and droplet-based methods. Then the isolated cells are lysed in order to preserve and capture mRNA molecules. The captured molecules will then be converted into cDNA by using Reverse Transcription. This step is critical as the efficiency of this reaction determines how much and what part of the RNA population will be sequenced.

The small amounts of generated cDNA are then amplified either by PCR or in vitro

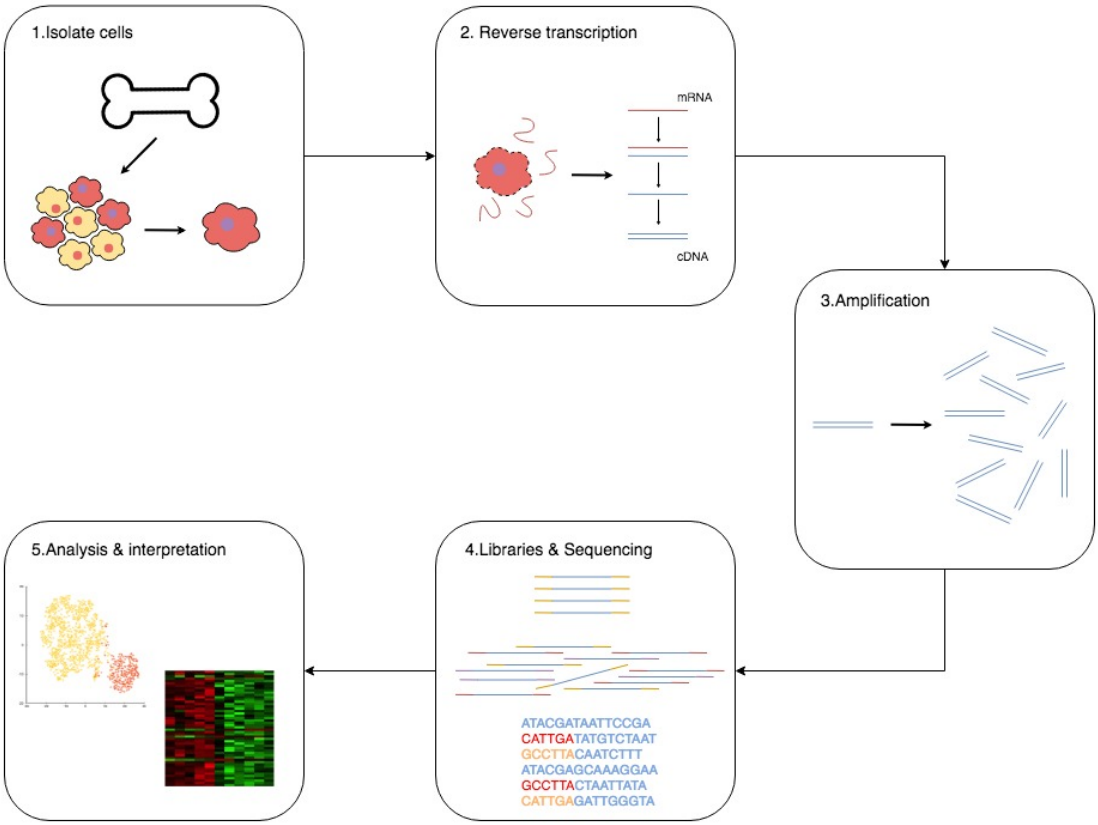

Figure 2.1: ***Single cell RNA-Seq standard experimental workflow.*** *The procedures implemented to extract and analyse genomic data from individual cells is different from those of microarrays and bulk data. 1, 2. The first step is to isolate cells from tissues, so they can be lysed and individually capture their mRNA molecules. 3, 4. The mRNA is converted into cDNA using Reverse transcription, to then be amplified and pooled using sequencing libraries. The final output has the expression level, displayed as the number of counts, of each gene for every single cell. 5. The final bioinformatics analysis often includes assessing for gene variability, and performing quality control over cells.*

transcription. After obtaining the amplified cDNA, it is processed and pooled using cDNA sequencing libraries, which are often part of the next-generation sequencing family and similar to those used for bulk samples.

Finally, bioinformatic methods are used to perform quality control and assess sample and gene variability, to then proceed towards computational and statistical approaches to interpret robust data biologically.

The number of reads obtained directly from sequencing need to be pre-processed in order to account for technical variations such as batch effects, cell-specific capture efficiency, amplification biases or dropouts. After alignment and de-duplication are performed, the result is an initial gene expression profile matrix that will then be normalized and inspected

to discard low-quality cells. Normalisation is essential to remove cell-specific bias, while the estimate of confounding factors is critical to identify and remove biological variation such as cell-cycle, and technical noise [42].

Even after pre-processing, the datasets obtained from single cell RNA-Seq analysis have some challenging attributes. Transcriptome data is inherently noisy, and therefore often technical noise is not easy to distinguish from biological variability. However, this stochasticity can also provide a platform to detect putative regulatory relationships among genes, which are often non-linear. High dimensionality of the samples presents another obstacle for data analysis, as it suffers from the commonly known "curse of dimensionality" [9]. It implies that when measuring distances in a high dimensional space, the differences tend to be small and therefore non-reliable to distinguish among samples.

In recent years, many advances have been made in order to tackle the limitations imposed by experimental design, and the datasets generated are becoming more extensive and informative than ever. The computational challenges can also be a burden in order to analyse and interpret all this data, mainly due to a few broadly shared attributes. The research potential of single-cell transcriptomes covers a broad range of applications, often following either a gene or a cell level path.

2.1.1 Cell differentiation - hematopoiesis

Mammalian blood formation is the most extensively studied system of stem cell biology, with its final goal being to obtain a better understanding of the molecular mechanisms controlling fate-determining events. A particular cell type, the haematopoietic stem cell (HSC), is responsible for generating more than 10 different blood cell types during the lifetime of an organism [69]. Formation and existance of blood cells is essential for any organism's survival, as they are in charge of carrying oxygen, promoting organ development and protecting organs against different pathological conditions [31]. Defects in the roots or along the differentiation process can lead to severe problems such as anemia or other haematological disorders including leukemia.

Haematopoiesis is the complex process through which blood cellular components are formed. It is constantly taking place in human bodies, daily producing approximately $10^{11} - 10^{12}$ new blood cells [96]. It involves a large variety of signalling pathways, and molecular mechanisms that are shared among most of the higher vertebrates. Access to genetic and other experimental data play an important role when choosing a model system to study haematopoiesis. Zebrafish have been used and led to many novel insights in the study of haematopoiesis due to their unique features such as external fertilization, optical transparency, genome editing and easy high-resolution optical imaging in live animals [31]. Malfunctions and defects in zebrafish blood cell production often mimic those of the human system, therefore exploration and modelling of such mechanisms contribute to a

better understanding of the general processes involved in haematopoiesis.

Zebrafish presents primarily two major waves of blood cell's formation [70]. Primitive haematopoiesis takes place during early embryonic development, giving rise to primitive erythroid and myeloid cell populations. Definitive haematopoiesis takes place later in development, where stem and progenitor cells give raise to all the rest of adult blood cells. The anatomical sites differ from those of mammals, but molecular mechanisms have been proved to be conserved [19, 70].

Modelling the process of cell differentiation has been done through two main approaches; deterministic and stochastic theories. The first one assumes that cells follow specific paths of differentiation, guided by certain stimuli and micro-enviroment factors. Stochastic theories instead, model haematopoiesis as a collection of random processes leading to a certain output, where variability plays a crucial role and can be understood as a continuous process. Figure 2.2 illustrates the fundamental models developed to explain cell differentiation.

Originally, differentiation experiments were carried out over populations of cells, isolated and characterised according to a set of cell-surface markers. The definition of cell types and differentiation stages is commonly generalised by using a set of selected molecular markers, which often present significant transcriptional and functional heterogeneity [34, 44, 73, 77, 109]. This approach assumes a stepwise set of binary choices with early and irreversible segregative pathways. However, recent studies have shown that some lineages may exhibit a certain flexibility, and not all steps are direct nor irreversible [1, 67, 112].

With the introduction of single-cell techniques, it has been possible to characterise cellular states and their transitions at a genome level for individual cells. It exposes data heterogeneity and elucidates cell fate decision mechanisms in greater detail. The limitation imposed by a restricted number of surface markers to identify cellular states can be overcomed by computational technique, where each cell is projected on the reconstructed differentiation path giving an insight to the state transitions occurring during differentiation [4].

The hierarchical paradigm classically used to explain haematopoiesis has been shifting towards a continuous stochastic approach, where the roadmap presents a certain flexibility and adapts to different environments or pathological conditions [116]. The classical discrete approach to describe the lineages or paths between HSCs and their progenies was proposed nearly 20 years ago, and has the form of a tree-like branched roadmap. According to those first models, differentiation takes place as a stepwise process from multilineage to oligo- or lineage restricted, and then to unipotent and mature blood cells. At the beginning of the 2000s the haematopoietic hierarchy was revised, due to the identification of new cell types and the complexity of lineage differentiation [1, 107, 108].

The introduction of multi-omics at a single-cell level has elucidated the heterogeneity

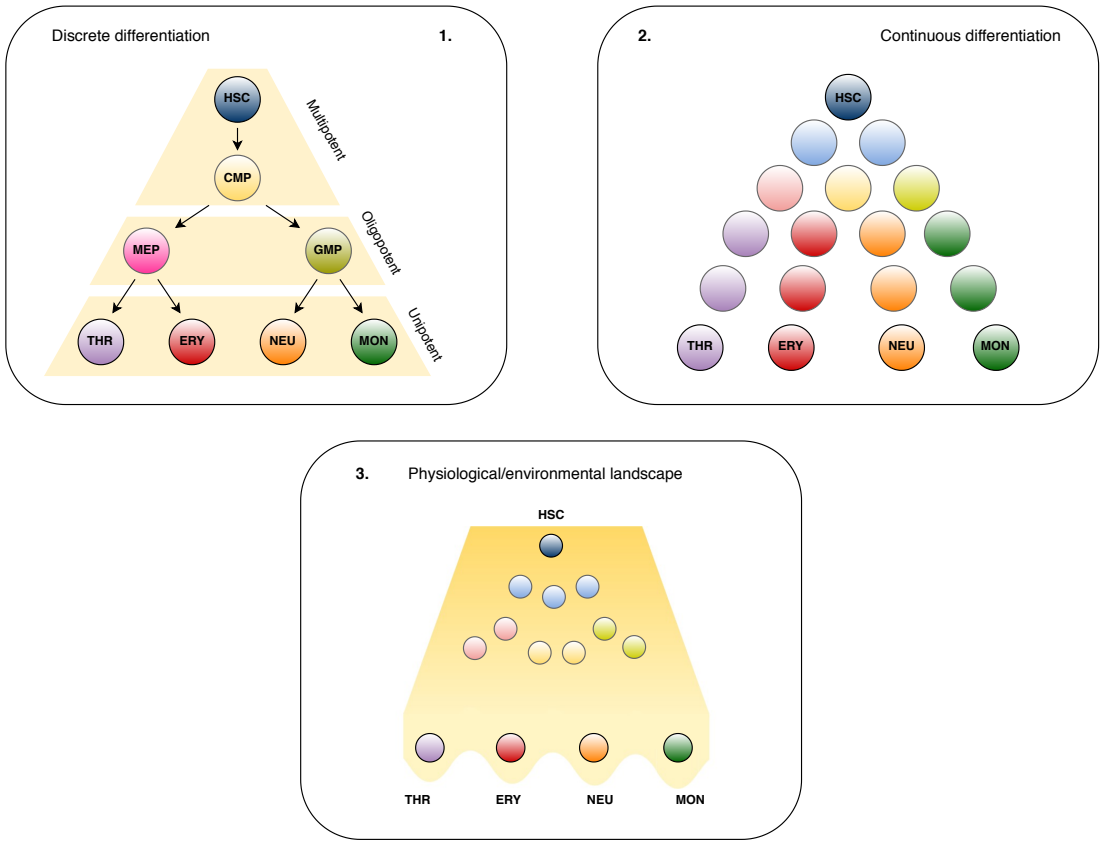

Figure 2.2: ***Different approaches to model hematopoiesis.*** *1. Discrete approach. The roadmap between multipotent haematopoietic stem cells HSCs and adult blood cells consists on a discrete number of well defined states. From HSCs the first step towards commitment are common myeloid progenitors (CMPs), which still preserve multipotency. Then megakaryocyte-erythroid progenitors (MEPs) and granulocyte-macrophage progenitors (GMPs) are the main lineage restricted states. Finally, the fully committed and unipotent blood cells are Thrombocytes (THRs), Erythrocytes (ERY), Neutrophils (NEU) and Monocytes (MON) . 2. Continuous or stochastic model of differentiation. The path towards mature blood cells is stochastic and the transitions take place on a continuous highly dimensional space. 3. Mixture model for continuous differentiation. The trajectories are guided by an energy landscape defined by physiological or environmental conditions. The cells are not fully free to transition towards different state. Their trajectories are influenced by multiple variables, defined by a combination of genetic and external conditions.*

among cells during the entire differentiation process. It has made possible to formulate transitions and paths between progenitors and adult cells in the continuum. With the help of computational tools, the entire transcriptome can be analysed and individual cells are ordered according to an artificially generated pseudo-time, based on the similarity of their genetic profiles. The newly defined continuous models are more flexible in terms of lineage segregation and cell decision making. They allow higher levels of variability during the process, and also open the door to new scientific questions and challenges at

computational and biological levels.

2.1.2 Computational and bioinformatics tools

Cells live in a dynamic environment, engaging in a number of processes on a multi-scale level. Single cell RNA-Seq captures a static picture of a cellular collective at a certain point in time, where individual cells may be undergoing and representing different stages of several dynamical processes.

A few computational tools have been developed to recognise these states directly from genomic data at a single cell level. Since the first single-cell RNA-Seq experiment published in 2009, many techniques have been extended and new ones introduced in order to tackle the challenges brought in by the properties of these datasets.

In general, all methods have a common basic pipeline as depicted in Figure 2.3. Starting with dimensionality reduction, they usually combine it with distance measures and a clustering algorithms to recognise similarities among samples. The final targets of the analysis may differ among experiments, depending on their research goals. Some of the most common applications are cell hierarchy reconstruction and regulatory network inference.

Cell hierarchy reconstruction is mostly used for cell differentiation and response to stimuli studies. Samples are often the main object of research, and characterising their states or transitions is the primary objective. Trajectory analysis and pseudotime are some of the techniques used by most studies for this purpose.

Regulatory networks can be inferred using single cell by accounting for transcriptional bursting and stochastic gene expression. Modules of co-regulated genes can be identified, and combined with derived pseudo-time and clustering techniques to infer relations between genes and generate the corresponding networks.

Since our interest is mostly on differentiation and characterisation of cell types, we will focus on the techniques developed for cell hierarchy reconstruction.

The general pipeline for single cell RNA-Seq analysis after data pre-processing starts with feature selection and dimensionality reduction. When analysing distances and comparing between samples that lay on a highly dimensional space, which in our case is defined by the number of genes, there is an additional hazard added by the "curse of dimensionality". It occurs when the number dimensions is significantly high, and distances between samples become very small and non characteristic of the separable properties. To avoid its effect, one should increase the number of samples, to an extent in which it often becomes impossible due to experimental conditions. Another solution is to decrease the number of variables, either by using feature selection or applying one of the many dimensionality reduction methods that are currently available.

Feature selection involves identifying the most important genes among those obtained

from the experimental setting, which is commonly achieved by using those with highest variance.

On the other hand, several algorithms can be used to perform dimensionality reduction. They map from a dataset with a large number of variables, to a lower dimensional representation with a certain number of abstract components. The most commonly used among biologists and bioinformaticians are Principal Component Analysis (PCA), locally linear embedding (LLE), t-distributed stochastic neighbour embedding (tSNE) and Isomap [5, 82, 99]. The choice of dimensionality reduction technique can play a crucial role downstream in the analysis, as it may lead to the loss of important biological information.

Distances between cells are then calculated on the lower dimensional space, by using one of the many distance measures available. For instance, the most common measures are the Euclidean distance, cosine similarity and Pearson or Spearman correlations. When choosing the metrics, one needs to account for the topology of the generated spaces as well as scale variability.

To identify patterns and similarities among samples one can perform clustering over the data. Once the distance measures have been defined, the next step is to decide which clustering algorithms need to be applied.

One of the most popular and well known is k-means, which iteratively identifies cluster centres (centroids) based on a greedy algorithm, and assigns the label of the closest centroid to each sample.

Hierarchical clustering is a different technique that sequentially combines cells into larger clusters (agglomerative) or divides them into smaller groups (divisive). It generates a tree with a hierarchy of cells that can reveal further substructures within the data, depending on where the cutting threshold is defined.

Gaussian mixture models are statistical methods that fit a set of normal distributions to the data. Each of the curves can be interpreted as a different cluster of points.

Community detection algorithms have also been used to identify groups of cells. They are network based methods, often constructed over k-nearest neighbours graphs, that detect groups of densely connected nodes. The algorithms can be applied to detect tightly connected communities in a graph, without having to specify the number of partitions reached in the solution. One of the most widely used is the Louvain algorithm [13], a greedy optimisation method that aims to maximise for modularity, while preserving scalability and speed when applied to larger datasets [53, 110].

2.1.2.1 Dimensionality reduction

Principal Component Analysis Principal Component Analysis (PCA) is a dimensionality reduction technique, based on a statistical procedure that transforms an initial set of variables into a reduced set of linearly uncorrelated components (principal components).

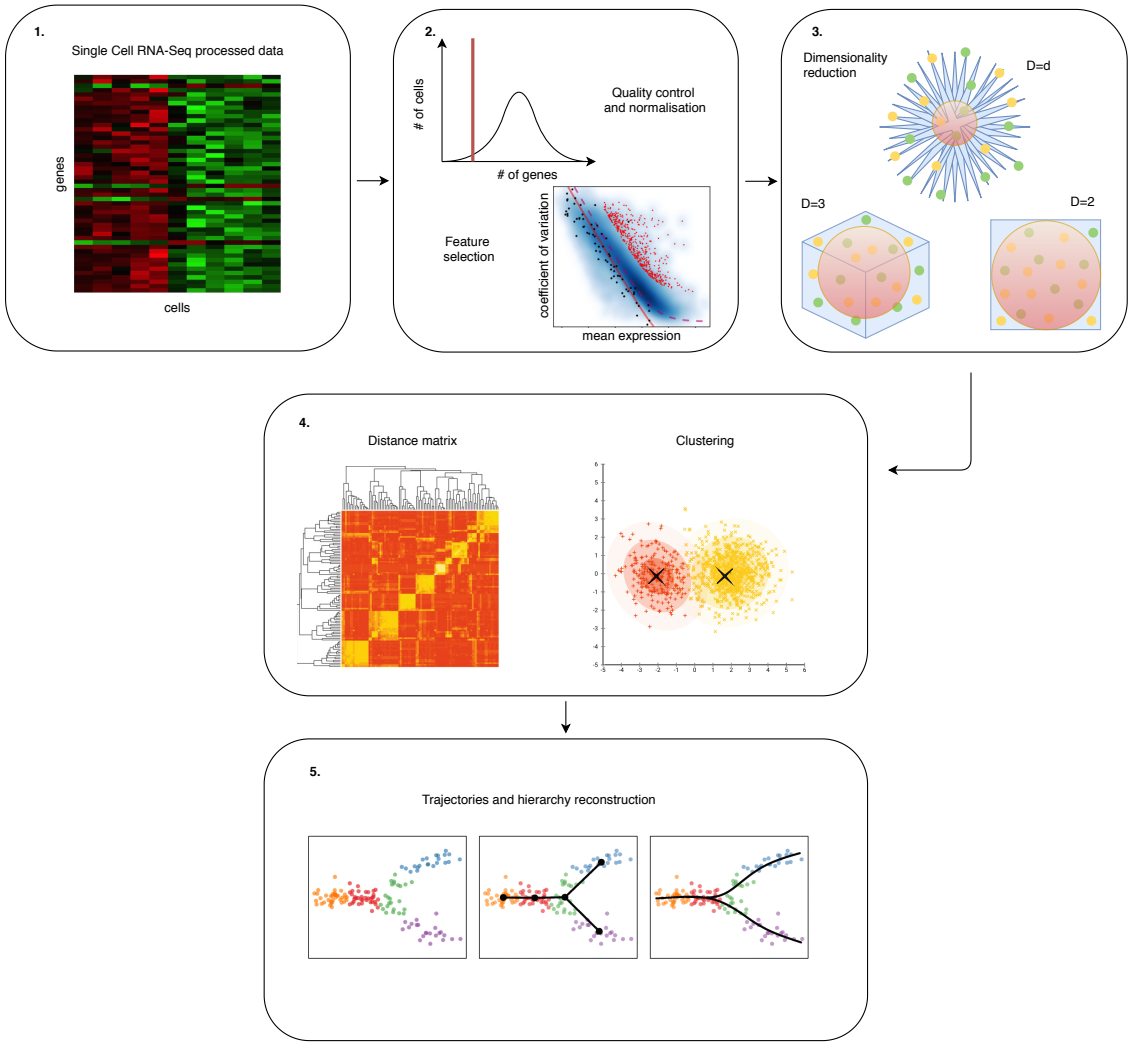

Figure 2.3: ***Pipeline for single cell computational analysis and modelling.*** *1. The raw data obtained as result of the experimental single cell RNA-Seq. procedure contains the expression level of each gene for every individual cell. 2. Data pre-processing consists of normalising and identifying early outliers, together with optional feature selection to eliminate those genes with very low variability. 3. Even after the initial feature selection, the number of genes or dimensions of the data is still very large. High dimensional spaces suffer from the 'curse of dimensionality', for which the sparse distribution of samples over the space doesn't allow the correct measure of distances. Therefore, dimensionality reduction techniques are used to apply more complex algorithms over the data. 4. Distances among cells are calculated, and clustering is used to identify similar samples. 5. Further assumptions and different modelling approaches are implemented to identify trajectories and reconstruct cellular processes.*

To generate the orthogonal projection, a first component is found by defining a vector with the largest variance possible, which accounts for the maximum variability in the data.The second principal component will then be found by maximising the variance but under the constraint of being orthogonal to the first component. Successively, a new set of components is found by applying the same rule, so the final result is an uncorrelated

set of orthogonal vectors. The requirement of no correlation means that the maximum number of PCs possible is limited by the number of initial features and the total amount of samples.

PCA is performed by eigenvalue decomposition of a data covariance matrix. The new basis transforms the matrix into its diagonal form, with the diagonal elements representing the variance of each axis.

In order to map a vector x_i from its original space with m variables to the new space defined by m orthogonal basis, one can define a transformation $R = XW$. Where X and R are its original and transformed representations of the data respectively, and W is the transformation matrix. To reduce the space dimensionality we can define a truncated transformation by choosing the $\hat{m}$ first eigenvalues obtained from principal component analysis, and applying $R_{\hat{m}} = XW_{\hat{m}}$.

PCA can also be associated to singular value decomposition (SVD). SVD is another transformation method based on matrix factorisation, where the original matrix can be expressed as a product $X = U^T$. With Σ being a rectangular diagonal matrix with the singular values $\sigma_{(k)}$ of X, and U and W containing the left and right singular vectors of X. The transformation is defined by $R^{SVD} = XW = U\Sigma$, where in order to truncate the transformation the $\hat{m}$ largest singular values and their singular vectors are used $R^{SVD}_{\hat{m}} = U_{\hat{m}}\Sigma_d = XW_{\hat{m}}$.

PCA is widely used and has proved to be successful for dimensionality reduction, denoising and visualisation or pattern detection among data. But it also has shown some limitations that need to be considered when analysing the outputs. The fact that it is a linearly based statistical technique, means that it follows the assumption that the underlying structure of the data is also linear. Patterns that are highly correlated may not be resolved due to all the PCs being uncorrelated, and the fact that its ultimate goal is to maximise variance may dismiss some additional potential on clustering detection [55].

tSNE T-distributed Stochastic Neighbour Embedding (tSNE) is a non-linear dimensionality reduction technique particularly well-suited for visualisation purposes. Developed by Laurens van der Maaten and Geoffrey Hinton [60], it models each high-dimensional sample as a two or three dimensional point in a newly defined embedding. Similar objects are therefore mapped to nearby spaces, while dissimilar objects will be positioned far from each other.

The algorithm behind tSNE starts with the definition of probability distributions over pairs of high-dimensional objects. The chance of selecting two similar objects will be higher according to their joint probability distribution, while dissimilar objects will have very low probability of being picked. A new posterior distribution is then generated among the points over the embedding space, such that the Kullback-Leibler divergence with respect to the prior is minimised. The algorithm's cost function is non convex, implying that

different initialisations often lead to dissimilar results.

Due to it's lack of stability, tSNE is often combined with other dimensionality reductions techniques, such as PCA or SVD. When the initial number of features is very high, a first reduction of the number of dimensions suppresses some of the noise, and speeds up the computation of pairwise distances between samples.

2.1.2.2 Monocle

Monocle is a toolkit for analysing single-cell gene expression experiments. It performs differential expression analysis, learns the trajectories and order of cells according to a generated pseudo-time, based on a particular biological process. Monocle is also able to identify genes that are dynamically regulated during that process.

It was initially proposed as an extension of a prior algorithm for temporally ordering bulk microarray samples [61], upgraded to account for the variability introduced by single cell data. By using an unsupervised approach without any prior knowledge on marker genes, it orders single-cells using "pseudo-time". This artificially generated order establishes a quantitative measure of progress through a biological process, based on their gene expression profiles.

The first step of the algorithm consists of representing each cell on a high-dimensional Euclidean space, defined by the number of genes. The number of dimensions are then reduced using Independent Component Analysis (ICA), while preserving the fundamental distances between major cell populations.

The algorithm subsequently builds a Minimum Spanning Tree (MST) among all cells, to extract the longest path. This path corresponds to the longest sequence of transcriptionally similar cells.

Finally, this will be used as trajectory for individual cells progressing through differentiation. Monocle also examines diverging paths and alternative trajectories to find substructures, identifying branched biological processes.

When applied to differentiation, it is able detect genes activated or repressed in early differentiation, together with potential upstream regulators. It can help identifying previously unprescribed transcription factors and key genes involved in cell differentiation. It can even detect potential subtypes of cells, and compare between different states.

2.2 Introduction to Unsupervised Learning

The second half of this thesis has a major methodological and theoretical orientation. Therefore, in this section we will be introducing the main concepts in machine learning and physics developed in the subsequent chapters.

It covers the history and fundamental ideas on Autoencoders, focusing about Variational Autoencoders. We introduce information theory and its application to Artificial Intelligence. It finishes with a brief summary of the intersection between physics and AI, giving context to the perturbation theory approach explained in the last chapter.

2.2.1 Autoencoders and VAEs

The main idea behind Autoencoder is to train a neural network within an unsupervised approach, learning to reconstruct an initial input from a particular embedding defined in one of its hidden layers.

The concept of using a reconstructed input to train a network is not new, in fact it has been present among the machine learning community for decades. It was first introduced in the 1980s by Hinton and the PDP group, as a solution or alternative to backpropagation [83]. Recirculation is an alternative training technique that differs from general feedforward networks, where the activations of the original input are compared to the ones of the reconstructed input. It is rarely used for machine learning applications, but has given rise to the latter scheme of Autoencoders in unsupervised learning. Consequently, some interest was risen in relation to their theoretical interpretation, with the work of Bourlard and Kamp [14] related to singular value decomposition, or Baldi and Hornik [6] on principal component analysis. Despite the initial interest, not much research was devoted to the mathematical side of Autoencoders, but more emphasis was given to the application of deep learning architectures and their different uses.

Autoencoders were initially designed and integrated as denoising techniques, for dimensionality reduction combined with feature extraction. One of the first reported models was introduced in 1987 by LeCun, Gallinari et al. [29] as a denoising alternative to Hopfield models [39]. Recently, a number of variants of the original autoencoder have been developed for a variety of applications. Their relation to latent variable models and the variational extension of the original interpretation has placed them at the forefront of generative modelling.

Among the variations that have been developed from the original Autoencoder, some of the most popular ones are denoising, sparse, contractive or variational autoencoders [50, 80, 105, 115]. The complexity and particular architecture of the networks is flexible, often tightly related to the data structure and the model's learning objective. For instance, convolutional autoencoders are commonly used for denoising purposes with two dimensional data (images) [20]. Ladder or variational VAEs have been used for clustering detection with gene expression or unidimensional data [95]. Hierarchical or X-VAEs have been recently introduced for data integration, with multiple data sources [64, 97].

2.2.1.1 Variational Autoencoder

The Variational Autoencoder (VAE), proposed by Kingma and Welling [50], uses stochastic inference to approximate the latent variables z as probability distributions. A graphic example of a simple VAE architecture is shown in Figure 2.4. The distributions z can reconstruct the original input from the latent space and capture relevant features from the data. VAEs are scalable to large datasets, and can manage intractable posterior distributions by fitting an approximate recognition model, using a reparametrised variational lower bound estimator.

They have been broadly tested and used for data compression or dimensionality reduction. Their adaptability and potential to handle non-linear behaviours has made them particularly well fitted to work with complex data.

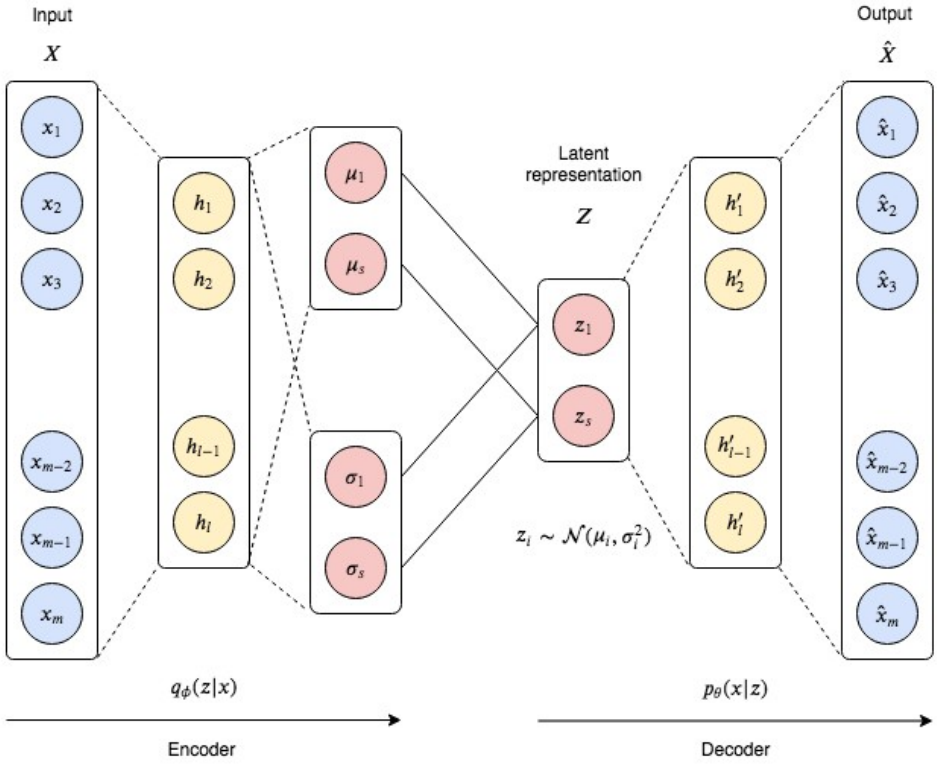

Figure 2.4: ***Variational Autoencoder.*** *The combination of the Encoder (inference model) and the Decoder (generative model) constitutes the general idea behind Variational Autoencoders. The embedding layer learns a set of probability distributions z_i. They are then used by the reparametrisation trick, to map and reconstruct the original input from the new latent representation of the data.*

Variational Autoencoders are built upon a probabilistic framework where the high dimensional data or input $X = x^{(i)}{}_{i=1}^{N}$ is drawn from a continuous random variable x with distribution $p_{data}(x)$. It assumes that the natural data X lies in a lower dimensional space, that can be characterised by an unobserved continuous random variable z and parameters θ .

In the Bayesian approach, the prior $p_\theta(z)$ and conditional or likelihood $p_\theta(x|z)$ come from a family of parametric distributions, with Probability Density Functions (PDFs) differentiable almost everywhere with respect to both θ and z. The true parameters θ^* and the values of the latent variables z^i are unknown to us, but the VAE approximates the often intractable true posterior $p_\theta(z|x)$ by using a recognition model $q_\phi(z|x)$ and the learned parameters ϕ represented by the weights of a neural network.

The Variational Autoencoder builds an inference or recognition model $q_\phi(z|x)$, where given a datapoint x it produces a distribution over the latent values z from where it could have been drawn. This is also called a probabilistic encoder.

A probabilistic decoder will then, given a certain value of z, produce a distribution over the possible corresponding values of x, therefore constructing the likelihood $p_\theta(x|z)$. The decoder is also a generative model, since the likelihood $p_\theta(x|z)$ can be used to map from the latent to the original space and learn to reconstruct the inputs.

The Variational Autoencoder model assumes latent variables to be the centered isotropic multivariate Gaussian $p_\theta(z) = N(z; 0, I)$, and then let $p_\theta(x|z)$ be multivariate Gaussian or Bernoulli with parameters approximated by using a fully connected neural network. The true posterior $p_\theta(z|x)$ is intractable, but we will assume it takes the form of a Gaussian with an approximately diagonal covariance. In this case, the variational approximate posterior will also need to be a multivariate Gaussian with diagonal covariate structure:

$$\log q_\phi(z|x^{(i)}) = \log N(z; \mu^{(i)}, \sigma^{2(i)} I)$$

Where the mean $\mu^{(i)}$ and standard deviation $\sigma^{(i)}$ are outputs of the encoding neural network.

It uses a reparametrisation trick to sample from the posterior $z^{(i,l)} \approx q_\phi(z|x^{(i)})$, with a variational inference approach where the random variable z can be expressed as a deterministic variable $z^{(i,l)} = g_\phi(x^{(i)}, \epsilon^{(l)} = \mu^{(i)} + (\sigma^{(i)} * \epsilon^{(l)})$. And $\epsilon^{(l)} \approx N(0, I)$ is an auxiliary variable with independent marginal $p(\epsilon)$.

Since $p_\theta(z)$ and $q_\phi(z|x^{(i)})$ are Gaussian, we can directly compute and differentiate the KL divergence without estimation. The resulting likelihood for this model on datapoint $x^{(i)}$ is:

$$L(\theta, \phi, x^{(i)}) \approx \frac{1}{2} \sum_{j=1}^{J} (1 + \log((\sigma_j^{(i)})^2) - (\mu_j^{(i)})^2 - (\sigma_j^{(i)})^2) + \frac{1}{L} \sum_{l=1}^{L} \log p_\theta(x^{(i)}|z^{(i,l)})$$

The marginal likelihood can be obtained as a sum of all the individual datapoints, such that:

$$\log p_\theta(x^{(1)}, ..., x^{(N)}) = \sum_{i=1}^{N} \log p_\theta(x^{(i)})$$

which can be re-written as

$$\log p_\theta(x^{(i)}) = D_{KL}(q_\phi(z|x^{(i)})||p_\theta(z)) + L(\theta, \phi; x^{(i)})$$

The first term is the KL divergence between the approximate and true posterior, and the second term is the variational lower bound. It can be re-written as

$$L(\theta, \phi; x^{(i)}) = -D_{KL}(q_\phi(z|x^{(i)})||p_\theta(z)) + E_{q_\phi(z|x^{(i)}[\log q_\theta(x^{(i)}|z}$$

also known as the evidence lower bound or $ELBO(\theta, \phi)$.

The VAE is therefore trained to optimise this function with respect to the variational and generative parameters ϕ and θ. Ideally, the training objective should optimise both reconstruction (generative model) and the difference between the inferred and true posterior distribution.

We have used VAEs for dimensionality reduction and the extraction of relevant information for mapping from genotype to phenotype, under an unsupervised approach. We assume that cell differentiation processes follow a set of rules that can be derived from a reduced number of parameters, which is smaller than the initial dimensions of the genotype. Therefore, we use VAEs to approximate these parameters and find a generative set of functions that describe the biological system being studied.

2.3 Introduction to the Physics of Machine Learning

2.3.1 Information theory and AI

The field of information theory is believed to be originally formulated by mathematician and electrical engineer Claude Shannon in 1948, in his paper 'A Mathematical Theory of Communication' [88]. Its final goal is to quantify the role of information, and its dynamics, by combining topics from mathematics, probability, statistics, computer science and physics.

The first decades of the 21st century have been named as the Information Age (also known as the Computer or Digital Age). It has been considered an historic period following the Industrial Revolution, where the economy is mainly based on information technology. Some of the main features of such economy are controlled by ownership and trading of large amounts of data, with privacy and communication processes being some of the core drivers of social evolution. As opposed to prior economy drivers, information is an abstract concept without an evident quantitative measure. The need for a solid mathematical

framework to analyse information flow and retrieval are only some of the incentives that boosted the development of information theory.

It is not a coincidence that a large amount of the notions used in machine learning and Artificial Intelligence are derived either directly or indirectly from this field. For instance, the popular cross-entropy loss function, maximum information gain, the Viterbi algorithm or the idea of encoding-decoding data are all under the scope of Information Theory.

According to Shannon's first approach, the information content of an entity is not so much related to its form or meaning, but defined in terms of a probability distribution and its uncertainty. These probabilities can be associated to random variables, and then used to define some of the basic quantities of information. Entropy uses such distributions to measure the information stored in a single variable, while Mutual Information combines them to outline the shared content amongst two or more variables.

Entropy is one of the building blocks of information theory, widely used in most derivations and principles. It gives a measure of uncertainty for a certain distribution, often associated to a dataset or set of observations. Given a random variable X, the entropy is defined as

$$H(X) = -\sum_{i=1}^{n} p(x_i) log_2(p(x_i))$$

where $p(x_i)$ is the probability of the i'th outcome of X.

The applications of entropy measures range from automatic decision tree construction, with feature selection driven by entropy criteria, to model selection based on the Principle of Maximum Entropy [27, 30, 120].

The Cross-Entropy function or loss is one of the most common usages of entropy in machine learning. It measures the similarity between two distributions p and q over the same set of outcomes x

$$H(p, q) = -\sum_{x} p(x) \log q(x)$$

It is widely used in classification problems, where the loss $H(p, q)$ increases when the predictions differ from the true outputs.

The information shared between two different random variables with probability distributions X and Y can also be quantified using measures such as the joint or conditional entropy. Mutual information unveils the relation between X and Y, while capturing the amount of information that can be obtained from one random variable by observing another. It is defined as

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x, y) \log(\frac{p(x, y)}{p(x)p(y)})$$

where $p(x, y)$ is the joint probability distribution between X and Y.

The non-linear nature of mutual information is particularly useful to analyse dependencies among random variables, therefore being frequently used for feature selection. It is also often used in Bayesian Networks to establish the strength and structure among different variables.

Similar to mutual information, the Kullback-Leibler divergence (KL) also assess the similarity between random variables. It is an asymmetric measure of divergence, capturing the loss of information between distributions. It is defined by

$$D_{KL}(P||Q) = \sum_i P(i) \log(\frac{P(i)}{Q(i)})$$

where P is a certain probability distribution and Q is its approximation.

The KL-divergence is often used as a distance metric. However, it is not strictly a distance due to it's non symmetric nature, and the fact that it doesn't satisfy the triangle inequality. It has been used to measure randomness in continuous time-series, and to evaluate information gain when comparing statistical models of inference. In machine learning, one of its most famous applications is to measure and penalise entanglement in Variational Autoencoders, as explained in 2.2.1.1

The intersection between information theory and Neural Networks is imperatively highlighted through the Information Bottleneck (IB) theory. It was first introduced in 1999 by Naftali Tishby, Fernando C. Pereira and William Bialek [100]. Multiple extensions and additions have been made to the original theory, where the machine learning and physics communities have adopted its terms to explain some of their new models and results obtained.

Under its theoretical framework, the flow and loss of information are characterised along the different layers of a network. It quantifies the tradeoff between distortion and complexity of the representations during and after training. These results have lead to many discussions around the theories of learning in Deep Neural Networks [89] [84].

The IB method provides a mathematical interpretation of the information lost between variables X and Y, when approximated from a compressed representation T. Also known as the rate-distortion problem, it is used to optimise with respect to distortion and compression of the latent representations, while analysing the position of the distribution $p(X, Y)$ on the information plane.

The relation between the bottleneck theory and autoencoders emerges almost naturally. Both models evolve around learning the optimal embedding from a particular dataset. The general idea is to compress the initial input while preserving the essential information contained in the data. The comparison between the input and the reconstruction of an autoencoder gives an idea of the information lost through the embedded representation S. The IB approach also allows to characterise the flow and loss of information in each

layer of the Autoencoder. This analysis provides a theoretical support for hyper-parameter selection, while characterising the learning dynamics of the models. Its application to real datasets provides a lower bound on the Information Bottleneck, and establishes a sufficient statistics boundary to the Information Curve.

The application of the IB in a variational setting, such as in VAEs, was first explored in 2017 [2]. They leveraged the reparametrisation trick to optimise the embedding representation as a balance between compression and distortion. Their results outperformed those that trained with other forms of regularisation, improving generalisation and robustness to adversarial attack.

The computation of mutual information can be particularly challenging, as it requires approximating distributions from data. However, for some distributions the calculations can be derived almost explicitly, these being when X, Y and T are discrete or jointly Gaussian. The latter case are those being used in the embedding layer of Variational Autoencoders. We use information theory for parameter optimisation and to interpret the results obtained by supervised and unsupervised learning models in this thesis.

2.3.2 Perturbation theory and AI

Generative functions, energy landscapes and degeneracy feature frequently both in quantum physics and machine learning problems. The challenge of a many-body problem in quantum physics originates when trying to understand and extrapolate wave functions to describe highly complex multi-particle systems.

A wave function ψ is the entity that can be used to describe from very simple to extremely complex entities. For instance, from a simple particle to highly intrincate molecules, the wave function characterises the different energies and quantum states associated to these bodies.

The derivation of such functions can be obtained explicitly or numerically by solving the many-body Schrödinger equation,

$$\mathscr{H} |\psi\rangle = i\frac{d}{dt} |\psi\rangle$$

where $\mathscr{H}$ is the Hamiltionian operator.

Even though an exponential amount of information is needed to encode a generic many-body quantum state, quantum entanglement and a finite number of relevant configurations allow an efficient compression that can encode for a certain number of states.

Neural Networks have been implemented to find ground states and describe the unitary time evolution of complex interacting quantum systems, by using reinforcement learning combined with variational and stochastic techniques [17].

These models successfully encode and represent the ground states of many body

prototypical spin models in one or two dimensions. They are given the Hamiltonian operator and its spectrum of energies E given by $\mathscr{H}|\psi\rangle = E|\psi\rangle$, and their time-dependent representation defined by the Schrödinger equation. Stochastic estimates of the energy gradient are obtained and used to train the models, in order to converge to a general solution.

In quantum mechanics, it is often observed that particular energy levels correspond to one or more different measurable states of a quantum system. This problem is known as quantum degeneracy.

It is mathematically represented by a Hamiltonian $\mathscr{H}$ with more than one linearly independent eigenstates with the same energy eigenvalue. The origin of degeneracy in quantum-mechanical systems often arises from the presence of symmetries and a group of symmetrical transformations over the operator $\mathscr{H}$.

One of the most well studied and extensively used solutions to degenerate systems consists of breaking those symmetries, through applying small external perturbations. This approximation scheme is named *Perturbation Theory*.

It consists of the introduction of perturbed potentials, used to find the solutions to the eigenvalue equation of the perturbed system $\mathscr{H}^*$, given the unperturbed representation $\mathscr{H}^0$.

By introducing an additional weak disturbance to the system's Hamiltonian, the solutions or corrections of the perturbed system can be calculated. They are used to describe a complicated unsolved system through a simple, solved one.

The problem of degeneracy is not only found in quantum physics, in fact it has been recalled by several studies as the reason behind some of the Neural Networks training difficulties [68, 75]. They argue that some singularities present during training mostly due to symmetries and overlaps between nodes, lead to degenerate manifolds in the loss landscape that slow down training.

Other studies have analysed the effect of degeneracy over the correlation between input and the corresponding latent codes in VAEs [119]. In order to mitigate the effect of degeneration and preserve information, they introduce a new method based on skipping connections without increasing the model complexity.

In Chapter 6 we introduce a new approach to interpret the VAE embeddings as a set of degenerate generative distributions. The energy levels or classes of the embedded system often have multiple representations, due to symmetries and overly simplified priors. We developed a framework based on Perturbation Analysis to leverage the generative nature of VAEs, and transform the embedding functions into a general set of unperturbed wave-functions that characterise the system.

CHAPTER 3

Supervised Learning

Single cell RNA-Seq datasets are collections of high dimensional arrays $\mathscr{X} = \{x^1 ... x^n\}$, where x^i contains the levels of gene expression for cell i. In order to obtain such information, the experimental techniques employed are very invasive, often involving the lysis or breakdown of cellular integrity. Therefore, although we have access to very detailed and valuable information about the genotype from individual cells, but we often lack a direct relation to their phenotype or cellular properties.

Gene expression data is known for being highly dimensional, often sparse and with moderate levels of noise among the samples. To generate a mapping to a lower dimensional space representing the phenotype traits, one needs to identify the relevant information and eliminate the excess stored in the dataset.

Several computational methods have been developed to process and perform such tasks. Most of them are based on dimensionality reduction methods such as PCA or tSNE, followed by clustering algorithms to capture similarities between cells in the embedded spaces. Although dimensionality reduction is able to significantly reduce the amount of noise, it often comes at the cost of information loss. Thus, it is not straightforward to establish a two directional relation between genetic profiles and phenotypical properties. The most popular unsupervised clustering approaches used are hierarchical [35], k-means [51, 113] and graph-based clustering [33, 111].

Some techniques have been developed in order to tackle particular biological problems, such as cell differentiation, through single cell data. When the samples being analysed are part of a dynamical process, ideally one would like to monitor their expression levels over time. However that is not possible due to the cell lysing required by the sequencing protocols. Instead, the datasets provide a static picture of a cellular culture, with samples undergoing one or several dynamical transitions simultaneously. The computational methods applied to these datasets often construct and make use of pseudo-time, an imaginary time construction used to place individual cells along developmental trajectories.

The most popular tools used are Monocle, TSCAN, diffusion maps or destiny, SLICER and Ouija [3, 16, 45, 78, 79, 103, 106]. They are based on either linear or non-linear dimensionality reduction techniques, combined with several biological and geometrical assumptions to reconstruct differentiation trajectories.

The assumptions imposed by these models often establish their own constraints that lower their predicting power and domain of analysis. We have designed a data driven approach, based on machine learning techniques, to identify cell types and transitioning states.

Cell labels or phenotypical properties are sometimes available for single cell data, through particular genes acting as cell surface markers. A direct mapping can then be constructed from genotype to phenotype via supervised learning, for example using a Multilayer Perceptron or Neural Network classifier. The complexity of the model and its performance will be determined by the amount of data available and its intrinsic properties such as the number of relevant components within the input dimensions and their interactions.

A direct mapping to the phenotype by using labels requires noise removal and information extraction, in order to constrain the results to the k classes given.

We designed and trained a Multi-Layer-Perceptron to classify cells according to their type, based solely on genomic information. The classifier was trained with adult cells, and used to predict the fate of stem cells or those in the process of differentiation. Moreover, we define a quantitative measure of differentiation or "Stemness" to be assigned to the each sample and determine their level of differentiation.

3.1 From genotype to phenotype - direct mapping

Single Cell RNA-Seq. data measures the number of counts or level of genetic expression for individual cells. For a given experiment with m analysed genes and n samples or cells, one can define a matrix $\mathscr{X}$ with dimensions $m \cdot n$, outlining all the experimental information. The observations are represented as vectors or events in the Genetic space, defined as follows:

Definition Genetic Space *$\mathscr{G}$: Let $\mathscr{X} = \{x^1 \dots x^n\}$ be a set of n cells, each defined by an m -dimensional array $x^i = [x^i_1 \dots x^i_m]$. The Genetic space $\mathscr{G}$ is a topological space defined by the set of points in $\mathscr{X}$ and their neighbourhoods. Two cells will have the same genetic profile $x^i = x^j$ in $\mathscr{G}$ if and only if $i = j$.*

The biological interpretation of each event in the hyperspace often involves a classification and/or segmentation challenge. Each of these cells belongs to a phenotype subclass well defined in the phenotypical space $\mathscr{P}$, with the following definition:

Definition Phenotype Space $\mathscr{P}$*: Let* k *be the number of differentiable phenotypical properties among the elements of* $\mathscr{X} = \{x^1, ..., x^n\}$*. We can assign a subclass* y^i *to each point in* $\mathscr{X}$ *such that a new set* $\mathscr{Y} = \{y^1 \dots y^n\}$ *is defined. The Phenotype space* $\mathscr{P}$ *is a topological space defined by the set of points in* $\mathscr{Y}$ *and their neighbourhoods.*

Two different cells with x^i and x^j are associated to the same point $y^i = y^j = k$ in $\mathscr{P}$, if their phenotypical attributes correspond to the same subclass. There is a direct mapping or transformation ρ between $\mathscr{G}$ and $\mathscr{P}$, such that $\rho(x^i) = \rho(x^j) = k$.

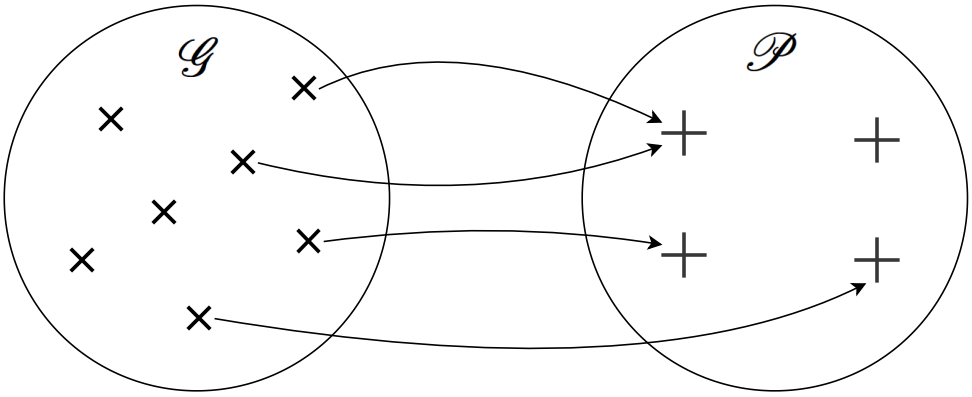

Figure 3.1: ***Mapping from the genotype*** $\mathscr{G}$ ***to phenotype*** $\mathscr{P}$***.*** *Different events in* $\mathscr{G}$ *can map to the same phenotype* $\mathscr{G}$*, due to genetic variability. However, with direct mapping it is not possible to reconstruct the exact inverse mapping from phenotypes to genetic profiles.*

When the cell classes are known, one can use Supervised Learning to infer the transformation $\rho(\mathscr{X})$. A Neural Network can be designed and trained to learn the classification or mapping between $\mathscr{G}$ and $\mathscr{P}$, as depicted in Figure 3.1 . Unfortunately, prior biological knowledge is often limited and can lead to poorly assigned segmentation labels. That adds a high level of extrinsic uncertainty to the data, which can translate into non-generalisable transformations $P(X)$.

Although the Genetic space $\mathscr{G}$ contains the full amount of information obtained experimentally, it becomes impossible to study the properties of the entire space due to its high dimensionality and a limited amount of samples. But since not all the information is relevant in order to constrain the cells to a discrete number of phenotypical states, it is possible to map them into a lower dimensional space $\mathscr{P}$, with well defined geometry and separable sub-classes.

The major phenotypical groups among cells can usually be segmented by applying linear transformations over the genotypical space. But further sub-structures and smaller clusters defined by non-linear relations among components in $\mathscr{G}$ are often missed by most linear methods. We have shown that non-linear methods such as Multi-Layer-Perceptrons are able to capture those sub-classes, together with all the other major groups among samples.

3.1.1 Stemness measure

The analysis of gene expression data on differentiation processes aims to identify cell types based on their genotypes. Adult or mature cells exhibit some characteristic genetic profiles, which are usually separable even by linear classifiers over $\mathscr{G}$. However, those cells in the process of differentiation exhibit transitioning genotypes, with high levels of variability among their gene expressions. Those profiles are unable to be separated solely by basic linear techniques, as they are often part of sub-classes determined by non-linear transitioning states. The transitioning genotypes still have a tendency or probability to commit towards any of the final mature classes, some of them may have already started committing towards a particular type. Pure Stem Cells are those with lowest level of commitment, they are able to equally transition to any of the adult phenotypes. Thus, they are very valuable to analyse due to their differentiation potential, but have also been proved very difficult to identify only from their genetic representation.

We used the outputs of non-linear classifiers to derive a quantitative measure of commitment or Stemness value. For each of the stem or semi-differentiated cells, we extract a value of their state of differentiation based on their genetic profile and their classification probabilities, to identify those cells with lowest commitment. Such measure is defined by:

$$S^i = \sum_k p(y_k^i|x^i) \log \frac{p(y_k^i|x^i)}{p(y_k^i)}$$

where $p(y^i = j)$ is the output or probability of cell i to belong to class j.

Previous models such as Monocle are able to construct trajectories and identify the differentiation branches related to the mature cell types. However, the assumptions imposed by the algorithm suppress the non-linearities at a very initial stage, denying the detection of potential sub-branches. By using a non-linear mapping, we allow alternative transitions to be considered, and generate a non-binding probability vector from where Stem Cells and other sub-types can be singled out.

3.2 Implementation

We have tested our approach over a dataset on single cell gene expression data, characterising the process of haematopoiesis on Zebrafish [4]. It is represented by a set $\mathscr{X} = \{x^{(i)}\}_{i=1}^n$ where $x^{(i)} = [x_1^{(i)}\, x_2^{(i)}\, \dots x_m^{(i)}]$ expresses the genetic profile of cell i. It contains $m = 1871$ gene expression measurements from $n = 1724$ cells. The expression levels are measured by the number of experimental counts obtained for each gene, and pre-processed using log-normalisation.

The labels were extracted from Athanasiadis *et al.* [4], where they computationally

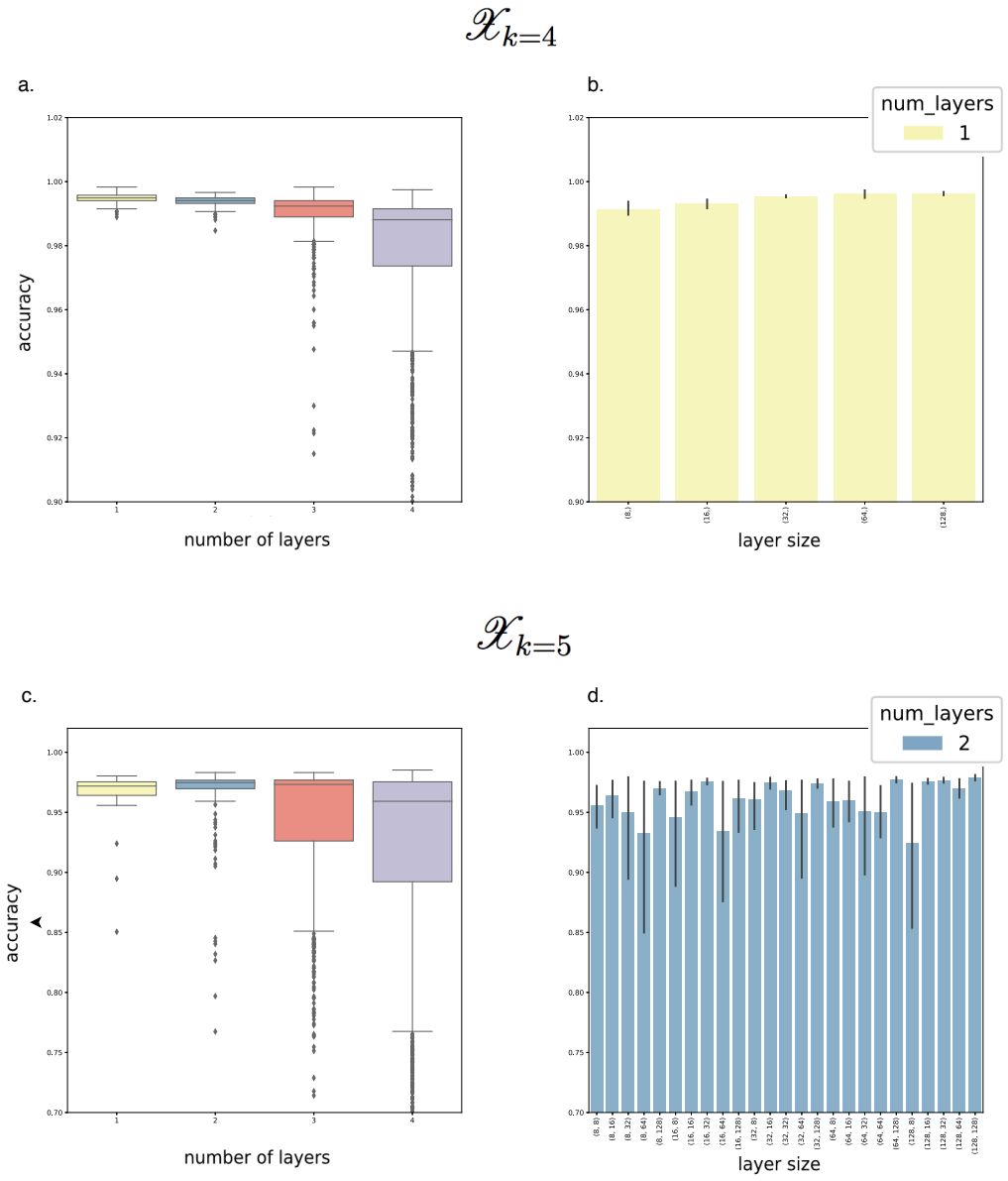

Figure 3.2: ***Neural Networks hyperparameter tuning.*** *We have trained 2340 models to perform classification over the dataset, in order to select the optimal hyperparameters. a, b. For the subset $\mathscr{X}_{k=4}$ of only mature cells one of the least complex architectures, with one hidden layer and 128 nodes, has proved to achieve the highest classification accuracy. c, d. For $\mathscr{X}_{k=5}$, two hidden layers with 128 and 32 nodes respectively have given the optimal results.*

reconstruct the differentiation trajectories in vitro. Five cell states were found in the dataset using the Monocle2 algorithm [79]: Monocytes, Neutrophils, Erythrocytes, Thromobocytes and HSPCs (Hematopoietic Stem and Progenitor Cells). We then showed how a data driven approach using non-linear methods provides a more detailed insight into the states

of differentiation of such cells.

To be able to use the gene expression measurements as part of a machine learning framework, we pre-processed and normalised the data. Gene expressions are normalised using Min-Max scaling such that the number of counts are scaled to the range $[0, 1]$. We use this normalisation to optimise the performance of our Neural Networks, without altering the original shape of the distributions. Through this transformation, we are able to model the gene expression for each cell as a multivariate Bernoulli distribution.

We want to recognise and prove the importance of non-linearities when mapping from genotype $\mathscr{G}$ to phenotype $\mathscr{P}$. We test this hypothesis by evaluating the performance of different classifiers over two subsets of cells $\{\mathscr{X}_{k=4}, \mathscr{X}_{k=5}\}$. The first one contains only mature samples, those that are identified as of the 4 adult hematopoietic types. While the second one contains all cells, both mature and transitioning ones, including HSPCs.

Both subsets are analysed using linear regression (LR), Decision Trees (DT), Random Forest (RF), Support Vector Machines (SVM) and Multilayer Perceptron (MLP).

The results are evaluated with k-fold Cross-Validation, using classification accuracy as measure of performance.

Linear regression was built using the PyTorch library [71] and trained to convergence, using Adam optimiser [49]. Decision Trees and Random Forest were generated using the scikit-learn package [74] and set without a maximum depth, with the second one using 100 trees per estimate. The Support Vector Machine analysis was also performed by using scikit-learn package [74], with a penalty parameter of $C = 1.0$. The hyperparameters of Multilayer Perceptrons were optimised according to their performance in terms of accuracy, and are shown in Figure 3.2.

Architectures for $\mathscr{X}_{k=4}$ and $\mathscr{X}_{k=5}$ are shown in Figure 3.3. Both models are trained until reaching loss convergence (100 epochs) and implemented an Adam optimiser [49]. For the subset $\mathscr{X}_{k=4}$ we used a single fully connected hidden layer of size 128 with ReLU activation, and Batch Normalization, with an output layer that returns the probability of each class using a Softmax function. For the subset $\mathscr{X}_{k=5}$ the optimal network is deeper, with two fully connected layers of sizes 128 and 32 with ReLU activations, and Batch Normalization. The output layer also returns the probability of each class using a Softmax function.

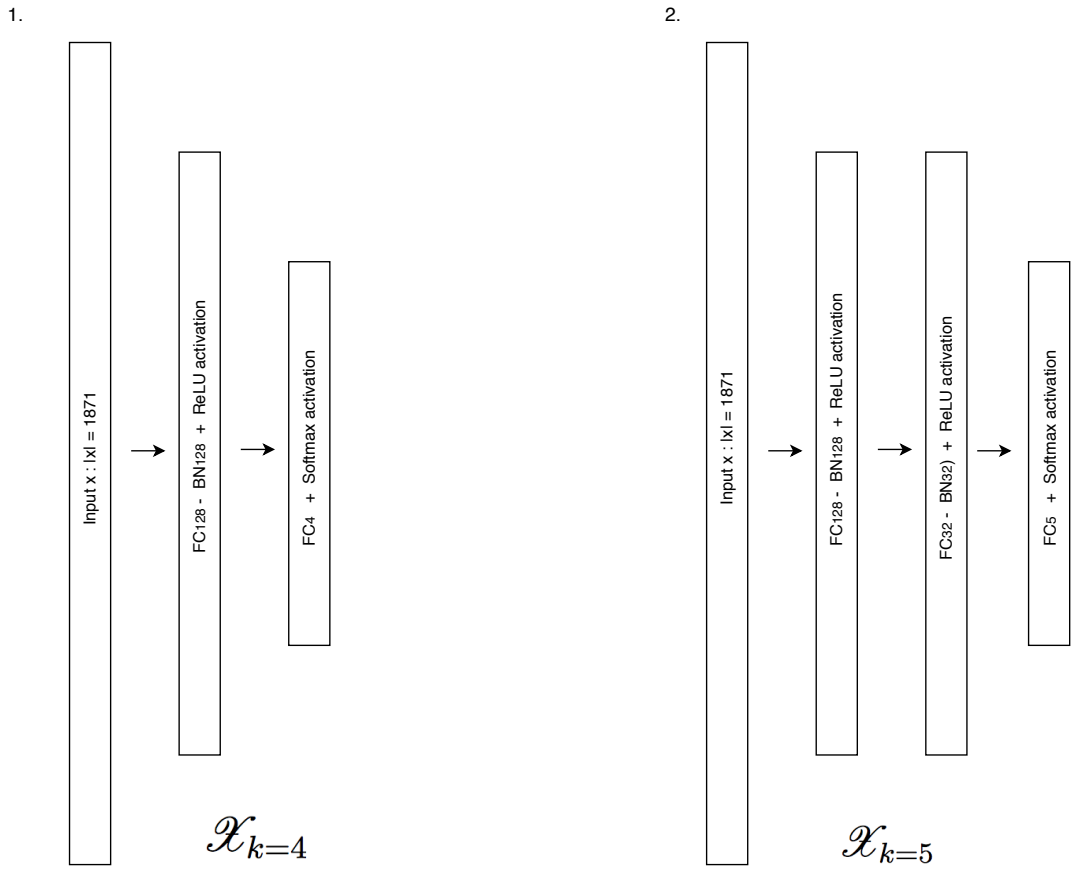

Figure 3.3: ***MLP classifier architectures.*** *The final architectures selected for the MLP classifiers according to hyperparameter optimisation. 1. A single layer with 128 nodes achieves high classification accuracy over the subset of mature cells $\mathscr{X}_{k=4}$. 2. A deeper network with two layers and {128,32} nodes displays the highest accuracy over $\mathscr{X}_{k=5}$, most likely due to an increase in the data complexity and non-linearities related to the classification task.*

3.3 Discussion

Defining a relation between genotype to phenotype through the use of predefined labels can be seen as a classification problem. The dimensions in $\mathscr{G}$ are input variables, while their mapping in $\mathscr{P}$ are the targets or classes.

Classification is one of the most recurrent tasks amongst supervised learning. A variety of techniques are available to solve these problems, both using Artificial Intelligence and also through classic statistical methods. The choice is often made based on the data properties and nature of the classification.

For instance, when classes are linearly separable one can use Linear Regression (LR) or Support Vector Machines (SVM) classifiers.

With Decision Trees (DT) each prediction is learned as a set of simple decision rules inferred from the data features. Random Forest (RF) are built upon Decision Trees, with a number of classifiers fitted on various sub-samples of the dataset. The final decision is made on the average of all the inferred trees, aiming to improve the predictive accuracy and control over-fitting.

Non-linearly separable samples can be classified using methods such as a Multilayer Perceptron (MLP), or Deep Neural Networks. They have proved to be very accurate within non-linear classification tasks, and can provide a probabilistic output that allows a more detailed insight into the predicted results.

Classifier	$\mathscr{X}_{k=4}$	$\mathscr{X}_{k=5}$
LR	0.97 ± 0.01	0.89 ± 0.03
DT	0.92 ± 0.03	0.88 ± 0.05
RF	0.95 ± 0.03	0.92 ± 0.04
SVM	**0.98 ± 0.01**	0.93 ± 0.04
MLP	0.94 ± 0.03	**0.94 ± 0.02**

Table 3.1: ***Classification accuracy of linear and non-linear classifiers.*** *Performance assessment between the different techniques used to predict cell types based on their genotype. For adult cells the mapping can be done by linear regression, while the classification of all samples including stem cells is optimal when performed by non-linear methods such as a Multilayer Perceptron.*

We have tested the performance of different classifiers in terms of the accuracy of their predictions, for subsets $\mathscr{X}_{k=4}$ and $\mathscr{X}_{k=5}$. This allowed us to compare the classifier accuracies over only mature $\mathscr{X}_{k=4}$, and mature together with stem cells $\mathscr{X}_{k=5}$.

$\mathscr{X}_{k=4}$ contains only mature cells, which we prove that can be detected by most classification algorithms due to their linearly separable genotypes. Subset $\mathscr{X}_{k=5}$ has a mixture of adult and transitioning cells, with potential non-linearly separable sub-classes. We show that linear and non-linear classifiers reach very high accuracies over the dataset of mature cells. However, when the dataset contains a mixture of mature and stem or HSPC cells, only non-linear models are able to maintain the same classification accuracy levels, as it can be seen through the Multi-Layer Perceptron results on Table 3.1.

Adult or mature cells have stable and well defined genetic profiles. The combination of over and under expressed genes allows particular functions and properties linked to their phenotype. Often, the set of genes or variables in $\mathscr{G}$ that characterise these classes can be

reduced to a few, reducing the classification problem to only a few input variables. That is one of the reasons most of the classifiers we trained and evaluated only with adult cells show a high prediction rates, even with high dimensional data.

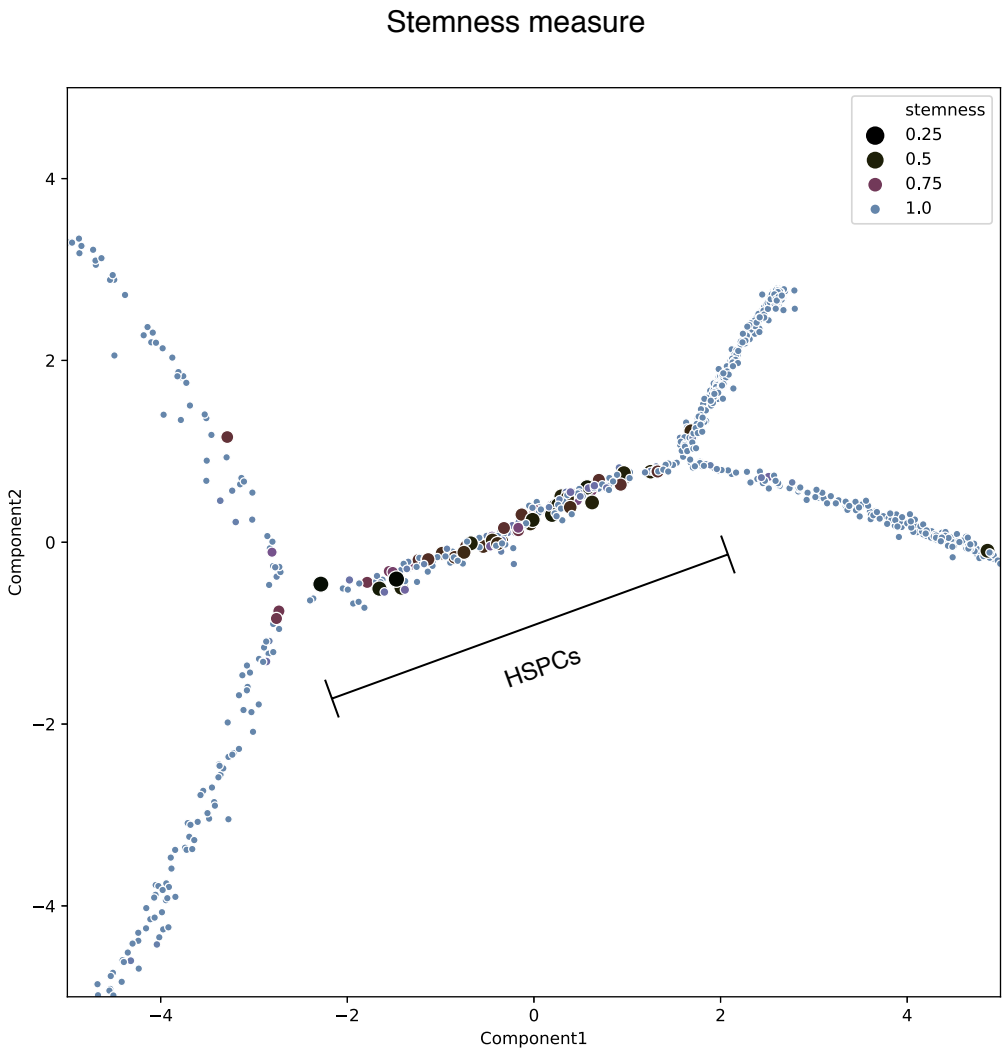

Figure 3.4: ***Stemness values.*** *Visual representation of the individual cells using the coordinates extracted from Monocle [103]. The position of a sample corresponds to its location along the differentiation trajectory. Stem and transitioning cells are in the central region of the plot, while the branches correspond to each mature cell type. By using our measure of Stemness we are able to identify those with lowest probability to commit to any of the mature phentoypes, and therefore most likely to be original Stem Cells.*

The results are different when classifying all cells among five groups. Four being mature ones, while the fifth contains those that are stem or in process of differentiation. The spectrum of differentiation states, and variability among samples translates to a poorer performance for most of the linear classifiers. Instead, non-linear Neural Networks are able to maintain their prediction rate. The addition of transitioning cells introduces a new level of variability among genetic profiles, with more of the features in $\mathscr{G}$ relevant for characterising differentiation stages, and a potential non-linearly separable distribution among samples.

Another advantage of MLPs over other classifiers is their probabilistic nature. Instead of just providing a boolean or deterministic classification, as the one for instance obtained with SVMs or RFs, the Neural Network classifier returns for each sample the probability of belonging to every class. We leverage the probabilistic approach to define the *stemness measure* or level of differentiation of individual cells.

To explore the diversity among HSPCs, the combination of non-mature and stem cells, we used the Stemness value introduced in 3.1.1. We use the prediction probabilities obtained from the MLP classifier to compute a quantitative measure of differentiation of cellular commitment for each sample. In Figure 3.4 the cells are plotted using the coordinates from the branching distribution obtained through Monocle [103], and the sizes corresponding to their Stemness values. Monocle is an algorithm based on linear dimensionality reduction, combined with a nearest neighbour approach. It defines a pseudo-time to characterise the main differentiation trajectories among samples, providing a visual interpretation of the branching process. However, due to its main initial assumptions, the order in which cells lay on the differentiation branches is highly dependent on the algorithm initialisation. Although the main branches or trajectories remain unaltered among different initialisations, the algorithm is unable to provide a higher insight into their level of commitment. In other words, the coordinates of those cells in the middle area (HSPCs) of Figure 3.4 do not correlate with their probability to differentiate towards one of the mature branches. That is most likely caused by the loss of non-linear information during the first dimensionality reduction. Some of the non-linear variables or genes related to differentiation are disregarded when analysed together with those relevant for the mature types.

The Stemness value is able to provide a new level of granularity for the analysis of single cell RNA-Seq data. We provide a quantitative measure of differentiation for those transitioning cells, while identifying some of the potential stem cells among all the samples.

3.4 Summary

We have found a direct mapping between genotype $\mathscr{G}$ and phenotype $\mathscr{P}$, and remarked the importance of accounting for non-linearities among data points. If the dataset contains stem cells or cells going through differentiation processes, supervised learning classifiers such as Multi-Layer Perceptrons or Neural Networks will provide more reliable results. Their probabilistic nature also allow to define a measure of commitment for the non mature cells, and identify those that are pure stem cells.

CHAPTER 4

Unsupervised Learning

Single cell RNA-Seq has a large potential to provide new insights in order to tackle a variety of biological questions. It gives access to data with a greater level of granularity, unveiling a new layer of genomic detail, and exposing a larger variability among samples.

Former methods used to analyse genomic data are generally fitted for microarray or bulk genome datasets, where sample heterogeneity is hidden among populations or cellular cultures. Linear dimensionality reduction methods such as PCA or LinearSVC, combined with clustering analysis, have lead to many of the biologically relevant results in recent times. But the assumptions imposed by these approaches may be missing some essential properties captured by cellular variability and non-linearities among samples.

The invasive nature of the experimental techniques used translates into the loss of phenotypical information about individual samples. The metabolic or structural relevance of some genes allowed to establish some local connections between them and particular phenotypes, always based on pure experimental analysis and literature. However, on a larger scale there are still significant gaps in mapping from the genetic space $\mathscr{G}$ to phenotype $\mathscr{P}$.

We want to leverage heterogeneity and the non-linearies nature of Single Cell data to improve that mapping, without having to sacrifice information due to prior assumptions or scale.

Therefore, in this Chapter we introduce an unsupervised approach to mapping from genotype to phenotype. By using VAEs, we define an encoded lower dimensional representation of the gene expression profiles in a new generated space $\mathscr{S}$. This encoding reduces noise and non-relevant information stored in $\mathscr{G}$, while compressing the data. It can then be used to identify clusters k which correspond to groups of cells with similar or the same phenotypical properties $\mathscr{P}$. The mapping between clusters in $\mathscr{S}$ and clusters in $\mathscr{P}$ is done through feature importance, by identifying relevant genes for each different cluster and validating these results using biological literature.

4.1 Extraction of the latent space

4.1.1 From genotype to phenotype - indirect mapping

A regular Single Cell RNA-Seq dataset contains m genes and n samples or cells, so one can define a set of vectors $\mathscr{X}$ with dimensions $m \times n$. Let $x^i = [x_1^i \dots x_m^i]$ be the m-dimensional vector that characterises cell i in the genotype space $\mathscr{G}$, and $y^i = [y_1^i \dots y_k^i]$ its phenotypical profile in $\mathscr{P}$. In order to map from $\mathscr{X}$ to $\mathscr{Y}$ we need to define a deterministic function $y^i = \rho(x^i)$.

When the number of classes and labels for each sample are known, the function can be approximated using a classifier or supervised learning, as we showed in Chapter 3. However, most single cell datasets don't own this information. Therefore, one needs to find an alternative path to detect phenotypical patterns based only on the observed genotypes.

We introduce a two step approach to mapping from genotype $\mathscr{G}$ to phenotype $\mathscr{P}$, depicted in Figure 4.1. We use a lower dimensional space $\mathscr{S}$ defined by the embedding of genetic profiles generated via Variational Autoencoders.

Definition Latent Space *$\mathscr{S}$: Let s be the number of latent dimensions needed to encode all the relevant information from $\mathscr{X}$ as events in $\mathscr{G}$. The latent space $\mathscr{S}$ with dimensions $s < m$ is such that for any set of observations or events $\mathscr{X}$ in $\mathscr{G}$, we can find another set $S = \{s^i\}$ with $s^i = [s_1^i \dots s_s^i]$ in the latent space with direct and inverse mapping to $\mathscr{X}$. The mapping $s^i = \psi(x^i)$ and its inverse $x^i = \psi^{-1}(s^i)$ is unique, such that $\psi(x^i) = \psi(x^j)$ if and only if $i = j$.*

When using a VAE to define the Latent Space $\mathscr{S}$, the reverse mapping from S to $\mathscr{X}$ also needs to hold. The recovery condition is then true and the reconstructed events can be located in the original genetic space.

Definition Recovery Condition *$\psi^{-1}(s^i)$: Let $\hat{\mathscr{X}}$ be the reconstructed set of events, mapped from the latent space $\mathscr{S}$ back to the original space $\mathscr{G}$. The reverse mapping $\hat{x^i} = \psi^{-1}(s^i)$ needs to ensure that $\rho(x^i) = \rho(\hat{x^i}) = y^i$.*

In other words, the Latent Space $\mathscr{S}$ and its mapping need to ensure the conservation of phenotypical properties.

The transformation $\psi(x^i)$ is a bijective function of variational nature. The transformation needs to guarantee that groups of cells with different phenotype y_t^i and y_{t+1}^j are always separable in $\mathscr{S}$.

The new space $\mathscr{S}$ is also required to be ϵ-stable, meaning that for any two input data points x^i and x^j in $\mathscr{G}$, the following inequation holds $(1-\epsilon)||x^i - x^j||^2 \leq ||\hat{x^i} - \hat{x^j}||^2 \leq (1+\epsilon)||x^i - x^j||^2$ [7]. Intuitively, it implies that Euclidean distances in the original input space are conserved throughout the transformation and in the newly generated output space $\mathscr{S}$.

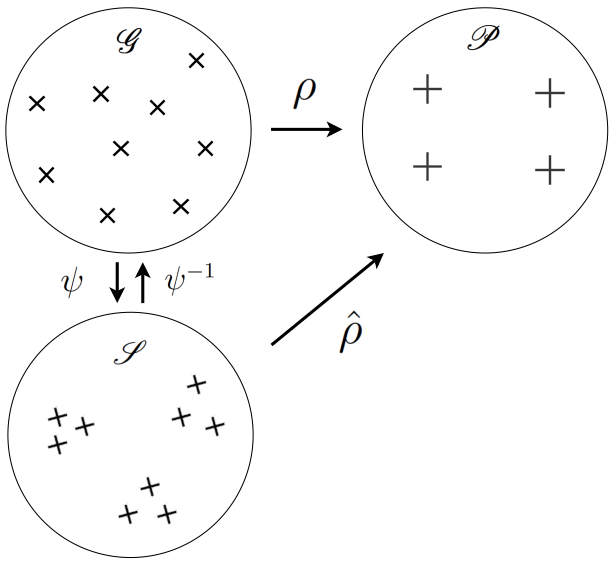

Figure 4.1: ***Indirect mapping from genotype $\mathscr{G}$ to phenotype $\mathscr{P}$, through $\mathscr{S}$.*** *Events in $\mathscr{G}$ are mapped to a lower dimensional representation $\mathscr{S}$, with a unique and phenotypically separable embedding. The direct function between events in $\mathscr{G}$ or $\mathscr{S}$, and the phenotype $\mathscr{P}$ are non-reversible and represented by ρ and $\hat{\rho}$.*

The operator $\psi(x^i)$ that maps from $\mathscr{G} \rightarrow \mathscr{S}$ needs to be invertible in order to generate the reverse mapping $\psi^{-1}(s^i)$. Ideally, the product of $\psi(x^i)$ and $\psi^{-1}(s^i)$ should be an orthogonal projection on $\mathscr{G}$, such that the product of the projection and its inverse is the identity function.

The operators $\psi(x^i)$ and $\psi^{-1}(s^i)$ are approximated by two symmetric Neural Networks, learned simultaneously from the data using an Auto-encoder. Specific details on such approximation are given in the following Section 4.1.2.

We use a particular type of Auto-encoders named Variational Auto-encoders, which account for variability or stochasticity among the input space. The latent or embedding components are a set of learned Gaussian distributions, used as generative functions to reconstruct the original input and learn both $\psi(x^i)$ and $\psi^{-1}(s^i)$ simultaniously.

Mappings $\mathscr{G} \rightarrow \mathscr{P}$ and $\mathscr{S} \rightarrow \mathscr{P}$ are not bijective, therefore the operators ρ and $\hat{\rho}$ are non-invertible. Their composition is also non-invertible $\hat{\rho}(s^i) = \hat{\rho}(\rho(x^i))$, so it is not possible to define a direct and deterministic inverse mapping from the phenotype $\mathscr{P}$ to the original or embedding spaces.

4.1.2 Methodology

We propose using a generative probabilistic framework [50] to model biological processes that lead to the changes in the gene expression along the different stages of differentiation. The pipeline developed consists of several steps, starting from a variational projection to a generative embedding space. We then identify structures among the data by applying

clustering techniques. The mapping to the phenotype space is achieved by feature extraction and in depth analysis of the results, which are supported by experimental results from the literature.

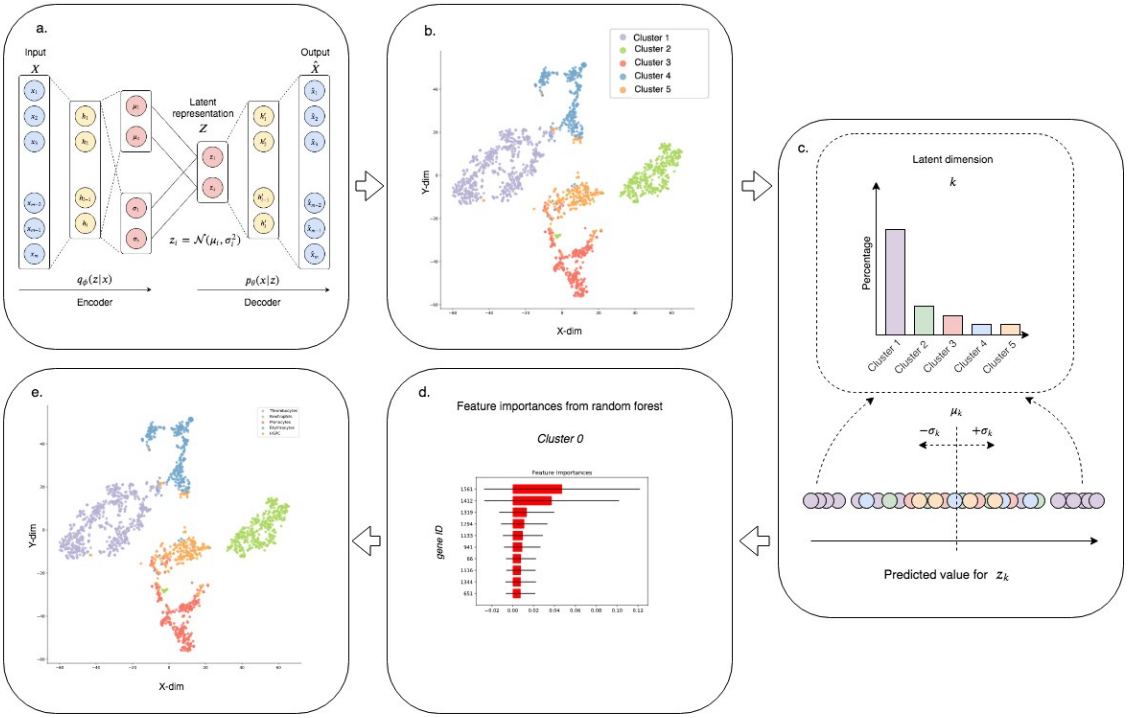

Figure 4.2: ***Pipeline to identify cell types using DiffVAE.*** *a. Train the VAE and define the latent space $\mathscr{S}$ from the generative embedding learned S. b. Use tSNE to visualise the distribution of cells and identify clusters. c. Analyse the relation between latent components and cell groups. d. Rank the genes according to their feature importance, select the most important ones for each individual cluster. e. Use the highest ranked genes or markers to interpret the biological relevance of the embedding, mapping from clusters to phenotypical profiles or cell types.*

4.1.2.1 Variational Autoencoder

Let $\mathscr{X} = \{x^i\}_{i=1}^n$ be a high-dimensional single-cell RNA-seq dataset consisting of the gene expression of n independent and identically distributed cells. Each gene expression vector x^i is an observation from a continuous random variable x, having distribution $p_{\text{data}}(x)$. Gene expression data is assumed to be generated by a semi-random process, modelled by an unobserved continuous random variable z with parametrised prior distribution $p_\theta(z)$.

The marginal likelihood $p_\theta(x)$, also known as evidence, is computed by integrating over the possible latent representations:

$$p_\theta(x) = \int_{z \in \mathscr{S}} p_\theta(x, z) \mathrm{d}z = \int_{z \in \mathscr{S}} p_\theta(x|z) p_\theta(z) \mathrm{d}z. \tag{4.1}$$

Computing the integral involves spanning the space of values for z which is often intractable. Its inference implies the computation of the posterior $p_\theta(z|x) = (p_\theta(x|z)p_\theta(z))/p_\theta(x)$, which is also intractable as it requires the marginal likelihood $p_\theta(x)$.

To learn in such a framework we use variational inference and approximate the posterior using the variational distribution $q_\phi(z|x)$. We thus build a variational autoencoder model [50], and use a multivariate Gaussian $N(z; \mu, \text{diag}(\sigma^2))$ distribution with mean μ and variance σ^2 to approximate $q_\phi(z|x)$.

The first section of the VAE is a neural network trained to estimate $q_\phi(z|x)$, the encoder. In addition, a isotropic multivariate Gaussian prior is assigned to the latent representation: $p_\theta(z) = N(z; 0, I)$.

The decoder is another neural network trained to reconstruct the gene expression data from the latent representation and thus estimate $p_\theta(x|z)$ via the variational generative embedding. See Figure 1.a. for a graphical illustration of the model.

The training objective of the standard autoencoder model [50] penalises the mutual information between the input and the latent representation [101]. There is also no need to encourage disentanglement in the latent representation [118], as the standard autoencoders do not have a generative nor variational nature.

Disentanglement is desirable within VAEs because ideally, the latent representation S should find a set of uncorrelated components able to separate the biological factors leading to the different cell types.

We introduce DiffVAE, a variational autoencoder that can be used to model and study the differentiation of cells using gene expression data. DiffVAE is part of the MMD-VAE family of autoencoders [118] and is trained to maximize the following objective:

$$L_{\text{DiffVAE}}(\theta, \phi; x) = \mathbb{E}_{q_{\phi(z|x)}}[\log p_\theta(x|z)] - \text{MMD}(q_\phi(z) \| p_\theta(z)), \tag{4.1}$$

where $\mathbb{E}_{q_{\phi(z|x)}}[\log p_\theta(x|z)]$ represents the reconstruction accuracy and the maximum mean discrepancy (MMD) [21, 32, 57]. The divergence between $q_\phi(z)$ and $p_\theta(z)$ measures how different the moments of two probability distributions are. The intuition behind MMD divergence is given by the fact that two probability distributions are identical if and only if their moments match.

Zhao *et al.* [118] proved that this training objective maximises mutual information between the input and the latent representation. Moreover, minimising the divergence $MMD(q_\phi(z) \| p_\theta(z))$, will encourage q_ϕ to be similar to the prior $p_\theta(z) = N(z; 0, I)$ with diagonal covariance matrix, which will lead to disentanglement in the latent dimension.

4.1.2.2 Clustering methods

DiffVAE is trained to map gene expression data from single cells to an embedding space $\mathscr{S}$, while preserving enough information to reconstruct the original profiles. The size of the

embedding s depends on the complexity of the data being encoded, and the architecture of the networks.

To visually identify the different groups of cells we used t-Distributed Stochastic Neighbour Embedding (t-SNE) [60]. It provides a 2-dimensional embedding for each cell based on the local distances to its nearest neighbours. t-SNE is only used for visualisation purposes, as the abstract embedding generated by this technique doesn't preserve global distances, and is highly dependant on its hyper-parameters and initialisation. But its neighbouring based optimisation approach is very powerful to characterise some of the relations between samples, and in particular when the number of original dimensions is not large.

K-means clustering is applied to the raw and t-SNE embedding, to obtain 5 cell clusters. Each of the identified clusters will then be analised and mapped to a particular phenotypical group.

4.1.2.3 Mapping from latent dimensions to cell types

DiffVAE was designed to model the data generating process giving rise to the observations in our dataset $\mathscr{X}$. Thus, this method should be able to identify the biological mechanisms that result in the observed gene expression value for our cells. Consider the analysis of a latent dimension t for any of the models. Let $z_t = [z_t^{(1)}\, z_t^{(2)}\, \dots\, z_t^{(N)}]^T$ be the predicted value of the encoder for z_t across all of the cells in the dataset. Let μ_t and σ_t be the mean and standard deviation of z_t. We define:

$$\mathscr{X}_t = \{x^i \in \mathscr{X} \,|\, z_t^{(i)} \geq \mu_t + \sigma_t \vee z_t^i \leq \mu_t - \sigma_t\} \tag{4.1}$$

as the set of cells at least a standard deviation from the mean in latent dimension t.

By computing the percentage distribution of the cells in $\mathscr{X}_t$ across the distinct cell clusters found in the dataset, we can evaluate how well the latent dimension is encoding the differentiation of the cells in a particular cluster. See the third section of Figure 4.2 for a graphic representation. Thus, for each cluster k we compute the percentage of cells from cluster k in each of $\mathscr{X}_t$, with $t \in \{1, 2, \dots, s\}$. The latent dimensions relevant for the differentiation of cells in cluster k will be the ones with the highest percentage of cells from cluster k in $\mathscr{X}_t$.

4.1.2.4 Identifying genes

Extracting a biological interpretation of the clusters identified requires external information, from the literature and experimental evidences. We use feature analysis over the results to identify those input variables or genes related to each subgroup computed via clustering. We have done this through two different methodologies; network weights and feature

importance ranking from random forest.

Using the network weights requires the identification of those latent components that are relevant for each cluster or cell type. This can be done directly by extracting and using the weights from the model once it has converged. Once the latent components are identified, they can be used to trace back their connection to the original dimensions or genes.

We select the latent dimensions optimal to separate the cells of each, and then compute the highest weights. High weight connections are obtained using the weight matrices from the decoder, a fully connected Neural Network between the embedding and reconstructed expressions. Let $z \in \mathbb{R}^m, h^{(1)} \in \mathbb{R}^{n_1}, h^{(2)} \in \mathbb{R}^{n_2}, \hat{X} \in \mathbb{R}^n$, be the sequence of layer activations in the decoder, where the embedding Z represents the input, $h^{(1)}, h^{(2)}$ are hidden layers and $\hat{X}$ is the output. Weight matrices for the connections between layers in the decoder are extracted from matrices $W^{(0)} \in R^{m \times n_1}, W^{(1)} \in R^{n_1 \times n_2}$.

Let $W \in R^{m \times n}$ be the weight matrix for all connections between the latent dimensions and output. W can be computed as a product of the weight matrices between all individual fully connected layers $W = W^{(0)} \cdot W^{(1)}$. Each matrix element W_{ij} indicates the connection strength between latent component i and gene j. For each component, genes are sorted by their absolute weight value. Top ranked genes are referred to as the highest weighted genes, as shown in the fourth section of Figure 4.2.

However, the list of genes extracted through this method depends strongly on the network's convergence. The ranking is sensitive to small changes on the model architecture and training hyper-parameters. We developed and compared the results with a parallel approach based on random forest and feature importance ranking. A random forest is a collection of decision trees, where each node is conditioned on a unique feature. The locally optimal condition of such trees is chosen based on their impurity measure, related to the entropy or information gain when used for classification tasks. From training the trees, one can compute the relevance of each feature in terms of how much they influence the weighted impurity of every individual tree. For a forest, the average for each feature can be computed and used to rank their importance in relation to the final outcome.

As a result of the comparison between the two approaches to rank feature importance, we established random forest feature ranking as our principal resource. This decision was made on the basis of generality and independence from modelling and network properties. The results from feature ranking, and use of external knowledge from biomedical literature, allows the mapping from clusters to cell types. The haematopoietic gene markers identified from our models are supported by external experimental evidence, and used as major indicators for phenotypical profiles.

4.1.3 Implementation

The pipeline developed is used to analyse single-cell gene expression data from haematopoietic stem, transitioning and mature cells in Zebrafish [4] .

Let the complete dataset be denoted as $\mathscr{X} = \{x^i\}_{i=1}^N$, where $x^i = [x_1^i\, x_2^i\, \dots x_k^i]$ contains all the transcriptomic data for cell i. The Zebrafish dataset analysed has $m = 1871$ gene expression measurements from $n = 1724$ cells. We used the 1871 genes with highest variability among the 1724 zebrafish single cells [4]. Considering the cell states or labels as initially unknown, we followed our pipeline to identify the different phenotypical classes from an unsupervised approach.

To use the transciptome data as input for our DiffVAE, additional data normalisation is applied through Min-Max scaling. Expression values for all genes are scaled to the range $[0, 1]$, providing more stability to the network training. In our probabilistic framework, gene expression is modelled for each cell as a multivariate Bernoulli distribution.

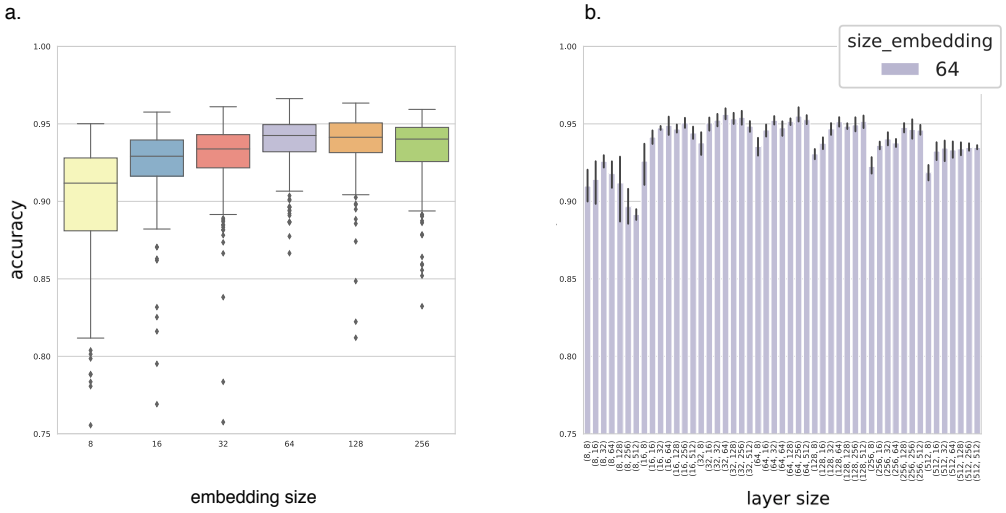

Figure 4.3: ***VAE hyperparameter tuning.*** *We have performed hyper-parameter selection to determine the size of the embedding, with a total of 882 VAEs trained. They are evaluated in terms of classification accuracy, and the final selection was made according to the highest accuracy and stability among models. The optimal embedding has a size of $s = 64$, and two hidden layers with 64 and 256 nodes respectively. a. Comparison between models with different number of latent components. b. Accuracy values of VAEs with embedding size $s = 64$, and different number of nodes in the hidden layers.*

We performed hyperparameter optimisation based on classification accuracy for both, the embedding and reconstructed gene expression, exploring different architectures and number of latent dimensions. The different performances in terms of classification accuracy are shown in Figure 4.3.

The final selected model consists of an embedding with size $s = 64$, two fully connected hidden layers with ReLU activation and batch normalisation, as it is shown in Figure 4.4.

VAEs are trained until convergence, minimising the ELBO loss with MMD divergence introduced in Section 4.1.2. All models were built and trained using the PyTorch library [71] and trained to convergence, using Adam optimiser [49].

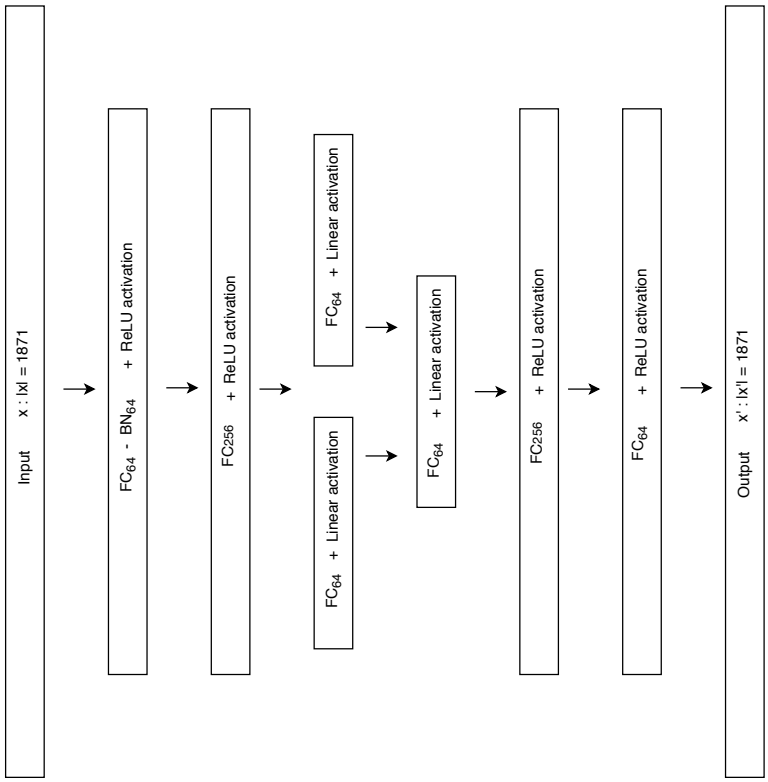

Figure 4.4: ***VAE architecture.*** *Final architecture selected for the VAEs according to the hyperparameter selection based on classification accuracy. A two layer encoder with 64 and 256 nodes respectively, fully connected to an embedding with 65 generative components. The decoder mirrors the encoding network, with two layers of 256 and 64 nodes. Its output contains the same number of variables as the encoder input, and will correspond to the learned reconstruction of the Gene Expression.*

The embeddings obtained from VAEs are visualised using t-Distributed Stochastic Neighbour Embedding [104], which highlights the structures and emphasises the local distances among cells. The algorithm was implemented by using the *sklearn.manifold.TSNE()* class [74].

K-means was used to perform clustering over the VAE embedding, and also the two dimensional representation obtained from tSNE for comparison. K-means was implemented using the class *sklearn.cluster.KMeans()* from the scikit-learn package [74].

Clustering results are evaluated by computing the Adjusted Rand Index (ARI) [41]

between the outputs and the true labels or cell types. It measures the similarity between the two by comparing all pairs of samples, and counting the pairs that are assigned to the same clusters both for predicted $\hat{Y}$ and true labels Y.

The overlap between $\hat{Y}$ and Y can be summarised by a contingency table $[n_{ij}]$, where each entry denotes the number of objects in common between $\hat{Y}_i$ and Y_j: $n_{ij} = |\hat{Y}_i \cap Y_j|$, $a_j = \sum_i n_{ij}$ $b_i = \sum_j n_{ij}$. The Adjusted Rand Index or score is then calculated using equation (4.1.3).

$$ARI = \frac{\sum_{ij} \binom{n_{ij}}{2} - [\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}]/\binom{n}{2}}{\frac{1}{2}[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}] - [\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}]/\binom{n}{2}} \tag{4.1}$$

This metric was implemented by using the class *sklearn.metrics.adjusted_rand_score()*.

We implement our own method to identify the optimal latent components for separating particular classes from the rest of samples. As explained in Section 4.1.2, means and standard deviations are used to calculate the percentage of outlier cells for each embedding component, represented by Gaussian distributions. This allows a further analysis over the quality of the embedding, enabling a measure of the direct relation between latent components and clustering or data structure.

To identify the genes that are relevant for each cluster in $\mathscr{S}$, we use feature importance ranking computed via random forest. We train the forest through scikit-learn [74], set without a maximum depth and with 100 trees per estimate. The class *sklearn.ensemble.RandomForestClassifier()* has a built in attribute to compute the feature ranking which was used to determine the scores for all input features, and identify those that were relevant for each particular cluster.

Once the most important genes for each cluster are identified, knowledge from biomedical literature and experts advice was used to align clusters with cellular classes or phenotypical attributes.

4.1.4 Discussion

The results obtained from the characterisation of different cell types using our pipeline, correlate with the trajectories computationally reconstructed using Monocle2 [78] over the Zebrafish haematopoietic dataset [4]. In particular, there is 89.9% overlap between the cell types identified using DiffVAE and the labels obtained by [4].

4.1.4.1 Embedding

After performing an extensive hyperparameter exploration and selecting the optimal architecture, the trained model is used to generate the embedded representation of cellular profiles. t-SNE is then applied to the $m = 65$ dimensional representation of the samples,

as well as to the original and reconstructed gene expression data, to obtain a qualitative evaluation of each space.

The generative embedding obtained from VAEs achieves a greater separation between clusters, and succeeds in segmenting the phenotypical groups according to the different cell types.

Data representation	***Classification accuracy***
Gene Expression ($\mathscr{X}$)	0.93 ± 0.03
Embedding (S)	$\mathbf{0.95 \pm 0.03}$
Reconstructed Expression ($\hat{\mathscr{X}}$)	$\mathbf{0.95 \pm 0.02}$

Table 4.1: ***Classification accuracy****. We compare the classification accuracy for different data representations; genotype, embedding and reconstructed genotype. The classifier implemented was a Support Vector Machine*

After dimensionality reduction, clustering is performed over the embedded representations [58]. Using the Zebrafish cell types found by Athanasiadis et al. [4] as true labels, we compare DiffVAE with the classification performance obtained from the original and reconstructed Gene Expression data. Their performance is quantitatively compared using support vector machine analysis (SVM), due to its simplicity in terms of hyperparameters and robustness over different sample dimensionalities. The results are shown in table 4.1, where the embedded and reconstructed representations present similar or higher accuracy than the original input. The generated representation is therefore successful in encoding and achieving a lower dimensional representation of the original data, where the topology of the newly defined space is able to separate among sample classes.

4.1.4.2 Clustering

We measured clustering performance using the Adjusted Rand Index between the true labels and the cluster labels was computed, as shown in equation (4.1.3),

The results reported in Table 4.2 represent the mean ARI obtained over clustering on the different data representations. The embedding built by DiffVAE gives the best clustering performance overall. In addition, computing the t-SNE embedding on top of the latent representation tends to increase the overlap between true and predicted labels. This is due to a greater separation among clusters given by the stochastic neighbouring embedding, but it also implies the loss of Euclidean properties and generative potential of the latent space $\mathscr{S}$.

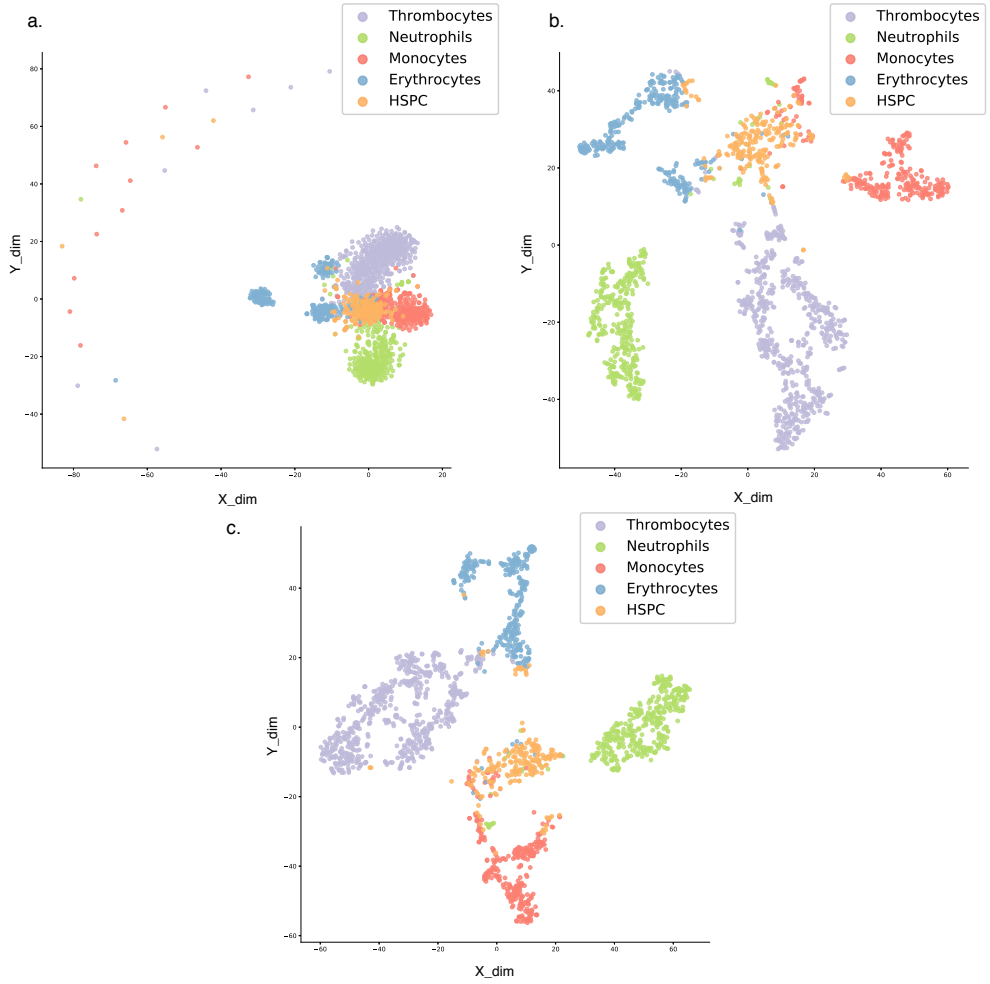

Figure 4.5: ***tSNE visualisation of cells over the original and learned spaces.*** *We use tSNE to visualise the data structure and distribution of cells over their high dimensional representations $\{\mathscr{X}\ S,\ \hat{\mathscr{X}}\}$. a. tSNE plot over the Gene Expression space, with 1871 initial dimensions. b. tSNE plot over the generative embedding space, with $s = 64$. c. tSNE plot over the reconstructed gene expression space, with the same dimensionality as the original data but a more separable representation.*

4.1.4.3 Analysis of the latent components

We performed a further analysis over the latent components defining the embedding generated by VAEs. Many of the components represented by Gaussian distributions are capable to separate only one or two of the clusters identified at a time. We developed and executed a full analysis of the latent dimensions in relation to the labels determined by the clustering, the results are shown in Figure 4.6.

A mixture of all the embedding components is what constitutes the final representation. However, individual components can be studied to establish their relation to particular labels, for instance to perform feature extraction using the weights of the network or to

Data	***Original***	***tSNE***
Gene Expression ($\mathscr{X}$)	0.51	0.52
PCA	0.63	0.72
Embedding (S)	**0.70**	**0.85**
Reconstructed Expression ($\hat{\mathscr{X}}$)	0.63	0.76

Table 4.2: ***Adjusted Rand Index between clusters and phenotype.*** *We measured the overlap between the clusters identified over different data representations and cell types. The values correspond to the ARI index* (4.1.3)*, and show that the VAE embedding has the highest values both for the original latent space with $s = 64$ dimensions, and its 2-dimensional tSNE representation.*

simply analyse the separability of the different cell types. In Figure 4.6 we can observe that some components are particularly good at separating one or two cell types from the rest. For instance, latent component number 24 is able to separate mainly Monocytes from all the other cells, while component number 16 is very good at identifying HSPCs.

This approach provides a more detailed exploration into the relation between the data structure and original genetic space $\mathscr{G}$. The variability over different models and information overlap between latent components due to entanglement effects can generate multiple embeddings related to a singular dataset. Further analysis and potential solutions to this problem are introduced in the following chapters of this thesis.

4.1.4.4 Identifying relevant genes for each cluster

Once the clusters have been identified and the sample labels predicted, they need to be mapped to their corresponding phenotypical classes, or cell types. This task is particularly challenging, as the biological definition of cell type is not always well determined. For simplicity, we will consider all cells with similar function and phenotypical properties as part of the same type. These cells also share common genetic features that characterise their profiles. Some of them have been identified as gene markers through experimental evidence, and have been proved to be relevant for specific metabolic or structural functions.

Our aim was to identify those genes that are relevant for each cluster, measure their overlap and analyse their biological relevance in order to interpret the results. We obtained a list of genes with higher relevance for each specific group of samples by performing feature analysis using random forest. Among the highest ranked genes, a few are selected as markers based on literature references and expert knowledge. They can then be used as

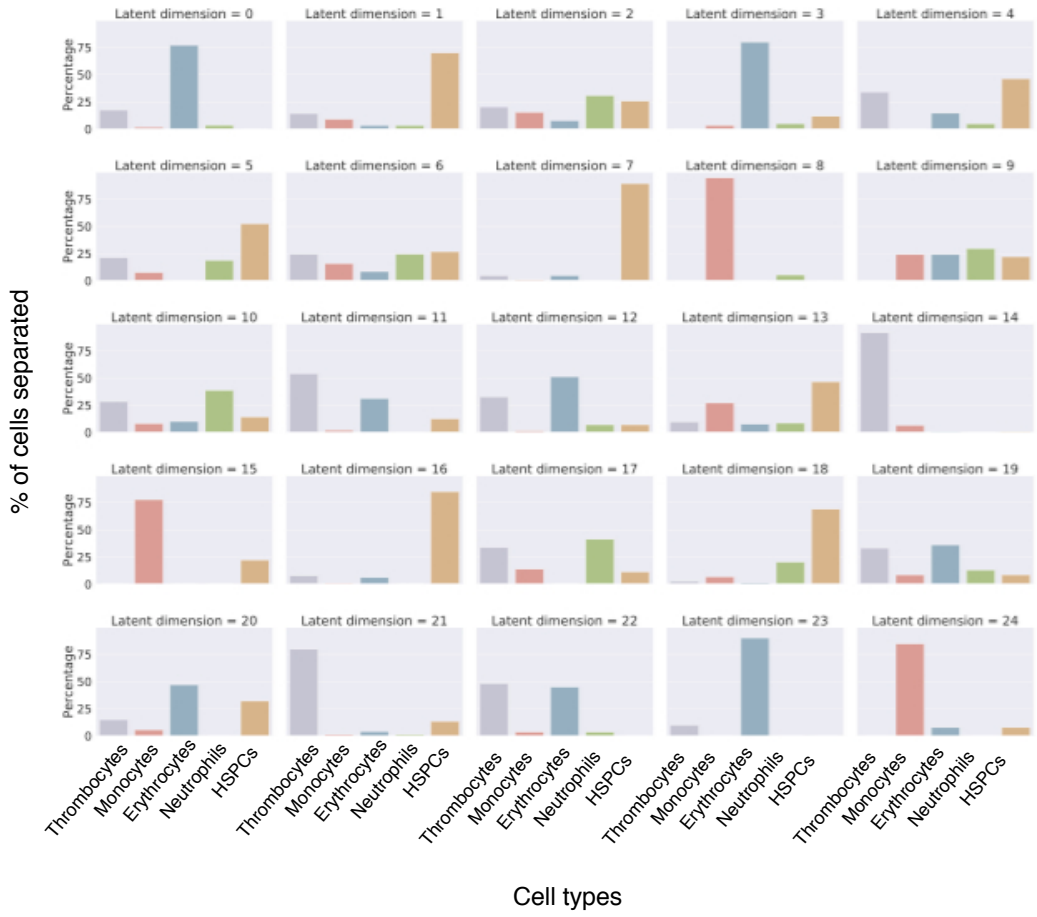

Figure 4.6: ***Proportion of cells separated by each of the latent components****. We studied the relation between individual components of the embedding and the clusters detected. The extended exploration of $\mathscr{S}$ shows that the characterisation of particular cell types through the encoding Gaussians is not uniform. Some cell types are frequently separated by a number of components, while others can only be detected by a few. This pattern is repeated within models with both the same and different number of embedding dimensions, and indicates a relation with the distribution of information and entanglement among the latent components.*

discriminative features to assist the mapping to phenotypic labels.

In Table 4.3 the relevant genes are identified for each cluster, so that they can be used to relate or map each cluster to a cell type. These lists are obtained after an extended literature review, combined with the technical knowledge from our collaborators and experts in haematopoietic differentiation. The relevant genes we identified for each cluster are compared and used to map from $\mathscr{S}$ to $\mathscr{P}$.

Cell types				
Thrombocytes	**Monocytes**	**Erythrocytes**	**Neutrophils**	**HSPCs**
si:ch211-161c3.6 [4], *cad* [4, 54], *pcna*[54]	*illr4* [4], *ponzr6*, *npsn* [72], *abcb9* [28], *lyz*[38]	*lgals2a* [4], *c1qc*, *c1qa* [4], *s100a10b*, *mafbb*[47, 102]	*alas2* [4], *ba1l* [4], *aqp1a.1*[4], *hbaa1*[4], *slc4a1a* [65, 76], *ba1* [4], *si:xx-by187g17.1*	*fn1b*[4], *itga2b*[4, 48], *bmp6*, *thbs1b* [4], *fhl1a*, *ctgfa*, *apln*

Table 4.3: ***Relevant Genes identified for each cell type.*** *We use feature ranking from random forest to identify a set of markers for each cluster. The relevance of the genes has been validated through multiple experimental evidences, and contrasted with expert knowledge. We use the list of markers to extract a biological interpretation of the clusters detected, and recognise the different phenotypical groups.*

4.1.5 Summary

We have presented an unsupervised pipeline to map from genotype to phenotype, with an interpretable approach capable to extract relevant genes and explain the final mapping between the latent embedding and cell types. This methodology is a good solution for those studies on single cell RNA-Seq based on cell differentiation, or attempting to analyse gene expression data of cells with significant non-linear behaviour. The non-linear behaviour of the encoder and decoder allow the identification of stem cells, while capturing other relevant information without any label restriction. The learned embeddings are then used to extract relevant inputs or genes for each cluster detected, allowing the potential characterisation of new genes related to the main cell types.

4.2 Reconstructed Gene Expression

Assessing the quality of artificially generated Gene Expression is not a simple task. There is a lack of realistic gold standards and we do not have an intuitive understanding of high dimensional expression data. In order to perfectly quantify the degree of realism of simulated data we would necessarily require a full knowledge of the real network of regulatory interaction together with other layers of genetic regulation. Since this information is often not available, we use a range of alternative qualitative and quantitative statistical methods, based on bioinformatics and clustering tools, to obtain a reliable set of measures that compare between the original and reconstructed data.

Maier et al. [62] derived specific histograms from quality measures and standard gene expression analysis in order to compare different properties between datasets. Overlap scores are used to quantify the discrepancies between two histograms, and are defined as:

$$O(a,b) = \frac{\sum_{i=1}^{n} min(a_i, b_i)}{\sum_{i=1}^{n} max(a_i, b_i)}$$

where a and b are two histograms corresponding to the property being analysed.

In their study, Maier et al. [62] used expression data extracted from four microarray compendia obtained for Escherichia coli and Saccharomyces cerevisiae from the M3D Database [26] and from the DREAM5 competition [63]. Experimental conditions represent the combination of drug, environmental and gene perturbations, and each of them may contain multiple replicates. Therefore, in microarray data the observations or samples are m dimensional vectors with gene expressions for multiple cells under the same experimental conditions. They also used predefined Gene Regulatory Networks for the systems studied, and performed further analysis to determine the preservation of such networks.

In our study, we modified some of these measures in order to apply them to single cell RNA-Seq data, which naturally exhibits different statistical properties. The properties characterised are:

- **Intensity histogram**. All the measurements on expression levels from a dataset are combined and summarised in one histogram. Their distribution is useful to evaluate the effectiveness of data normalisation.

- **Range of gene expression**. It measures the overall difference between the minimum and maximum expression, with a 99.5th intensity percentile. A histogram of gene ranges is constructed from the values computed for each gene. The range of the artificially generated data can scale very differently, due to a certain level of randomness in their generation. Meier et al. [62] multiply the artificial ranges by a factor so the median of the histograms match.

- **Silhouette coefficient**. To evaluate cluster consistency and quality, agglomerative hierarchical clustering [23] is used to group genes and samples separately. Silhouette coefficients [81] are computed to compare the average distances between and within clusters, providing measures of separation and compactness of the clusters. Hierarchical clustering illustrates the emergence of patterns among genes or samples, relevant features to be preserved when generating new data.

4.2.1 Implementation

The overlap score proposed by Maier et al. [62] is sensitive to symmetries, which may affect its accuracy as discrepancy metric. In collaboration with Vinyals, we propose alternative measures to achieve higher accuracies when evaluating statistical properties, while capturing additional properties not covered by the aforementioned measures.

Synthetic data should always aim to be consistent with the original datasets in terms of the structure and relations among groups of genes. These are often unique from the systems studied, encoding relevant information about the data on different scales. Therefore, any artificially generated data, claiming to replicate such systems, should exhibit the same properties.

The novel measures developed by Vinyals are based on the correlation coefficient between gene distance matrices.

Definition Gamma coefficient γ*: Let* X *and* $\hat{X}$ *be two* $n \times n$ *symmetric matrices holding the pairwise distances between all genes. One can define a coefficient to measure how faithfully a matrix preserves pairwise distances with respect to the other, such that:*

$$\gamma(X, \hat{X}) = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{i=i+1}^{n} \left(\frac{X_{i,j} - \mu_X}{\psi_X}\right)\left(\frac{\hat{X}_{i,j} - \mu_{\hat{X}}}{\psi_{\hat{X}}}\right)$$

$$\mu_X = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{i=i+1}^{n} X_{i,j}$$

$$\psi_X = \sqrt{\frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{i=i+1}^{n} (X_{i,j} - \mu_X)^2}$$

Based on the correlation and Gamma coefficients, we define and use a set of measures related to the silhouette coefficients. They are used to assess the quality of the reconstructed gene expression in comparison to the original data, focusing on the preservation of distances and clustering or structural properties. The following measures are implemented:

- **Distance between real and artificial distance matrices**. Let D^X and $D^{\hat{X}}$ be

two distance matrices based on the distance function d.

$$D^X_{i,j} = d(X_{:,i}, X_{:,j})$$

$$D^{\hat{X}}_{i,j} = d(\hat{X}_{:,i}, \hat{X}_{:,j})$$

The coefficient $\gamma(D^X, D^{\hat{X}})$ measures the correlation between the pairwise distances among genes from real and synthetic data.

- **Distance between real and artificial data dendograms**. Let $\mathscr{C} : \mathbb{R}^{n\times n} \to \mathbb{R}^{n\times n}$ be a function that performs agglomerative hierarchical clustering according to a given linkage function. It takes an $n \times n$ distance matrix as input and returns the $n \times n$ distance matrix of the resulting dendrogram. Let T^X and $T^{\hat{X}}$ be the real and artificial dendogrammatic distances.

$$T^X = \mathscr{C}(D^X)$$

$$T^{\hat{X}} = \mathscr{C}(D^{\hat{X}})$$

The coefficient $\gamma(T^X, T^{\hat{X}})$ measures the structural similarity between the dendograms. Note that $\gamma(D^X, D^{\hat{X}})$ does not necessarily correlate with $\gamma(T^X, T^{\hat{X}})$.

- **Squared difference between cophenetic correlation coefficient**.

The cophenetic correlation coefficient $\gamma(D^X, \mathscr{C}(D^X))$ measures how faithfully a dendogram preserves the original distance matrix [94]. It quantifies the loss of information taking place when performing hierarchical clustering with respect to the original distance matrix.

The artificially generated data based on a gene expression dataset, should have similar cophenetic coefficients to the original ones in order to be replicate the properties of the real dataset.

4.2.2 Discussion

We used the metrics introduced to measure the similarity between the original Gene Expression Data and the reconstructed version generated by VAEs. We also compared both the original and synthetic datasets to a randomly generated set of samples.

VAEs are trained to minimise the reconstruction error, optimising the distance between original and reconstructed data. It is essential that, in addition to the reconstruction error, statistical properties and structure of the data are also consistent. We trained and implemented VAEs to generate and evaluate new samples or synthetic cells.

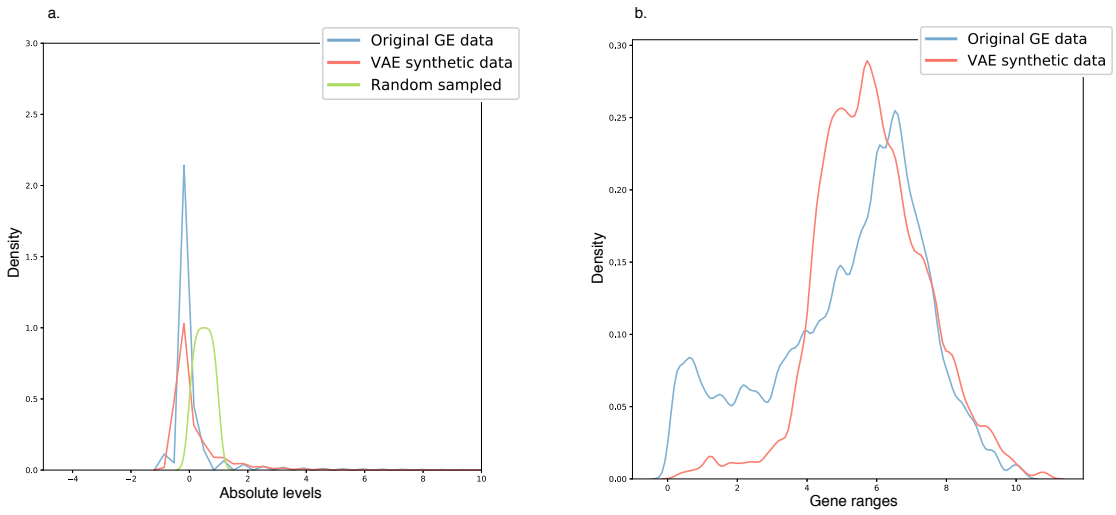

Figure 4.7: ***Qualitative analysis of the Reconstructed Data*** $\hat{X}$***.*** *We compare the distributions of the reconstructed gene expression* $\hat{X}$ *obtained from the VAE, and the original expression data* X*. a. Intensity distributions of the original, reconstructed and random generated data. The original and reconstructed data have multiple similarities, both distributions marking a high level of sparsity and long tails. b.Distribution of gene ranges for the original and reconstructed expressions. Synthetic data generated by the VAE has similar average to the original distribution, but there are fewer genes with low variability among the reconstructed expressions.*

In Figure 4.7 a results show the qualitative comparison of intensity distributions. Gene expression intensities of both datasets have similar distributions, with an expected peak of intensity centred at zero given the level of sparsity of Single Cell RNA-Seq. data.

The long tailed distributions are common for both original and artificial data, meaning that there are only a few genes with large values of intensity.

Figure 4.7 b shows the distribution of gene ranges. Synthetic data tends to have a more evenly distributed range of expressions, with fewer genes having the same intensity levels among all samples. This is due to the variational approach of the generator, which gets rid of the outliers and produces a more homogenous set of reconstructed expressions.

We also computed the quantitative coefficients designed to measure the similarity among datasets. Performance of VAEs and the quality of their reconstructed gene expression $\hat{X}$ are evaluated. Upper and lower bounds are obtained from the original data X and a randomly generated set X_{rand}.

We used the following coefficients:

$$S_{dist} = \gamma(D^X, D^Z)$$

$$S_{dend} = \gamma(T^X, T^Z)$$

$$\gamma(D^X, T^X) - \gamma(D^Z, T^Z)$$

Table 4.4 shows all three measures, concluding that the reconstructed gene expression has similar values to those of the original data. Even though the scores are not identical to those of X (upper bound), they are all significantly higher than the random case (lower bound) X_{rand}. All the coefficients are between the upper and lower boundaries, meaning that most distances are preserved.

coefficients	X	$\hat{X}$	X_{rand}
S_{dist}	0.62	0.40	0.00
S_{dend}	0.20	0.17	0.00
S_{sdcc}	0.00	0.03	0.10

Table 4.4: ***Quantitative analysis of the Reconstructed Data*** $\hat{X}$***.*** *We use* $\{S_{dist}$*,* S_{dend}*,* $S_{sdcc}\}$ *to measure the properties of synthetic expressions generated by VAEs. The coefficients are used to assess the data legitimacy in terms of distances and clustering preservation. The values obtained for the reconstructed data* $\hat{X}$ *are all significantly higher than the random case, and below the real data scenario. This means that the reconstruction is able to capture the main properties of the original data, for each one of the metrics computed.*

4.2.3 Summary

We have analysed and proved that the reconstructed gene expression generated by the VAEs preserves the properties of the original data. This is relevant both to prove that the embedding learned by the encoder is meaningful and able to encode most of the relevant information contained in the Gene Expression, but also reinforces the generative nature of our approach. In future studies VAEs can be used to produce additional data samples. In this Chapter we have proved that the synthetic or artificial cells generated will have the same or very similar characteristics to the real ones, and can be used for further analysis.

CHAPTER 5

Information theory

The development of machine learning techniques in biology, and particularly of unsupervised learning, has highlighted the need for a validation framework independent of experimental or pre-settled labels. We have developed a theoretical approach to interpret the performance of supervised and unsupervised models when learning from single cell RNA-Seq data.

Among some of the challenges presented by the analysis of gene expression data, the increased dimensionality both in the genetic and phenotypical space convolutes the corresponding mapping of samples. The identity of each cell, and its relation to the others from a biological perspective, is one of the analytical goals often contended by noise and lack of label specificity.

By analysing the flow and loss of information in supervised and unsupervised models we provide a better understanding of the network requirements for good learning. We used entropy and the shared information between layers to build a general approach and evaluate the results. Without the need for a specific solution, it extends well to different branches of machine learning.

The performance of different models can be interpreted as a trade-off between loss of information from the original space of events, and learning about targets. It can be measured using entropies and mutual information, as it is shown by the Information Bottleneck (IB) method. We have tested it both with supervised and unsupervised learning, characterising and evaluating models according to these terms. This provides an interpretation of the results that can lead for instance to a more universal hyperparameter optimisation, and overall a better understanding of the learning behaviour.

We believe that by developing a theoretical interpretation of the network performance, it will be possible to develop flexible models without prior assumptions or bounding labels. The benefits of such approach are both of biological and computational interest. It promotes the discovery of new groups or labels with biological relevance, and helps understanding the learning process of networks to avoid over-fitting and improve generalisation of models.

5.1 Genotype and phenotype projections

Dimensionality reduction and feature extraction using Deep Neural Networks can be explained from the point of view of information flow. Under the theoretical framework of the Information Bottleneck (IB) [100] one can quantify the amount of information lost between layers, as well as obtain generalisation bounds or the optimal information limits.

Given a dataset with m analysed genes and n samples, the dataset $\mathscr{X}$ is a high dimensional representation of the data in the genetic space $\mathscr{G}$. The phenotypical properties of the cells are summarised in $\mathscr{Y}$, where each cell has a label assigned from one of the k possible classes. The amount of information contained in $\mathscr{X}$ is significantly greater than the one contained in $\mathscr{Y}$, and this encodes the entropy or measure of disorder of each representation. However, not all the information in $\mathscr{X}$ is informative or related to phenotypical profiles.

The major phenotypical groups among cells can usually be segmented by applying linear transformations over $\mathscr{G}$. But further sub-structures and smaller clusters defined by non-linear relations among components in $\mathscr{G}$ are often dismissed by linear methods. When data can be represented over a space defined by linearly independent components, hyperplanes and linearly segmentation methods can be used. Nonetheless, the variables or genes that define $\mathscr{G}$ are part of a great complex network. In fact, they not only have non-linear dependencies, but their relations often contribute to the identification of different groups in $\mathscr{P}$.

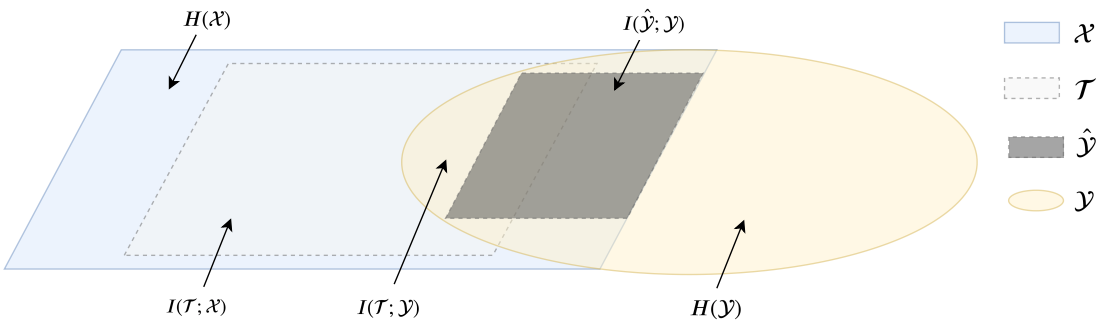

Figure 5.1: ***Entropies and mutual information projection - direct mapping.*** *The genotype and phenotype projections represent the overall amount of entropy contained in each representation $\{H(\mathscr{X}), H(\mathscr{Y})\}$. For instance, genotype space encodes a variety of structural, metabolic and functional properties, not necessarily related to the particular mapping targeted by the data. The phenotype space of events also has many different representations. Direct mapping extracts the overlap between the two entropy projections, or mutual information. Our classifier will learn the optimal estimate of the intersection between genotype and phenotype $I(\hat{\mathscr{Y}};\mathscr{Y})$, while eliminating the excess of information from $\mathscr{X}$ through its hidden layers $I(\mathscr{T};\mathscr{X})$.*

Figure 5.1 shows the projection of information from the input $\mathscr{X}$ and targets $\mathscr{Y}$, in the supervised learning case. Inputs are a high dimensional and noisy representation of the data, while targets or labels have low dimensions and entropy, as the samples collapse to one or several categorical features. The overlap between the input and output information projections is what is extracted and learned from the data. Neural Networks aim to approximate and cover the intersection $I(\hat{\mathscr{Y}};\mathscr{Y})$.

The information projection under the unsupervised learning scheme is portrayed in Figure 5.2. Within VAEs, the input and target representations are meant to be identical. The amount of information shared between $\mathscr{X}$ and the embedding S should be only the necessary information to reconstruct the original space $\hat{\mathscr{X}}$ while compressing the data. Although compression leads to a general loss of information, latent embeddings often capture part of the full overlap between $\mathscr{X}$ and $\mathscr{Y}$.

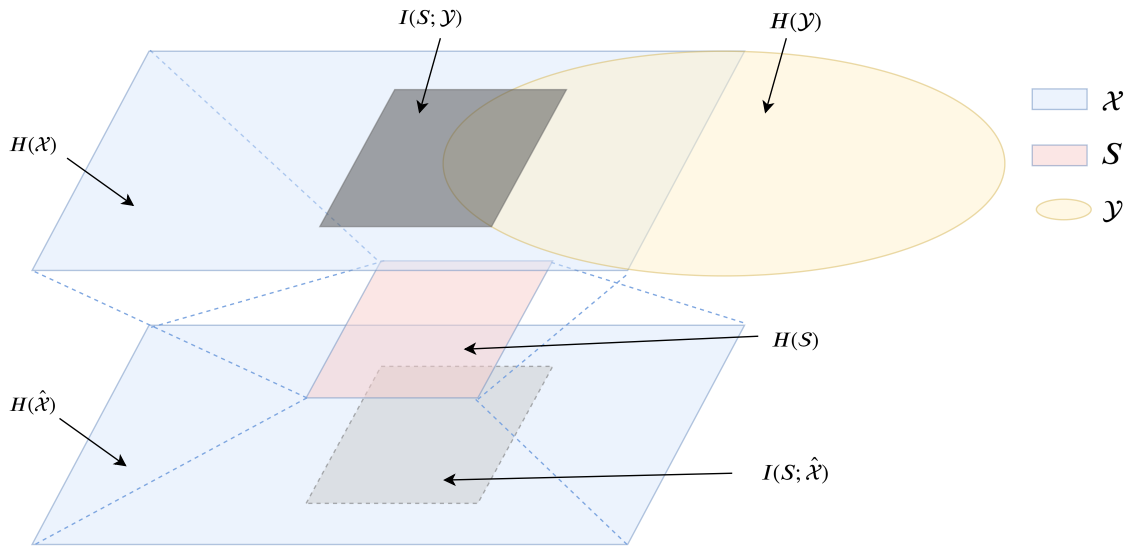

Figure 5.2: ***Entropies and mutual information projection - indirect mapping.*** *The learning objective for unsupervised learning doesn't maximise the information contained in the intersection between genotype and phenotype. Instead, it learns a compressed representation S only with the minimal amount of information to reconstruct the original data $\{\mathscr{X}, \hat{\mathscr{X}}\}$. The compressed representation is not required to maximise the information content for a certain $\mathscr{Y}$. However, the removal of non relevant noise from $\mathscr{X}$ can unveil new phenotypical properties when analysing $I(S, \mathscr{Y})$.*

5.2 Information limits

Deep neural networks allow the construction of high level distributed representations by using a sequential processing of the data, so they can learn useful hierarchical representations of the data. The information bottleneck (IB) was proposed as a theoretical method to analyse the information flow through the network, starting with an input $\mathscr{X}$ processed towards the output $\mathscr{Y}$.

Given their joint distribution $p(\mathscr{X};\mathscr{Y})$ and assuming statistical dependency between $\mathscr{X}$ and $\mathscr{Y}$, one can measure their mutual information $I(\mathscr{X};\mathscr{Y})$ as,

$$I(\mathscr{X},\mathscr{Y}) = \int_{\mathscr{X}}\int_{\mathscr{Y}} p(x,y)\log\left(\frac{p(x,y)}{p(x)p(y)}\right)dxdy$$

Let $\hat{S}$ be the minimal sufficient statistics of $\mathscr{X}$ with respect $\mathscr{Y}$. $\hat{S}$ then captures only the necessary features to characterise $\mathscr{Y}$. We can formulate the mapping of $\mathscr{X}$ as a Markov Chain $\mathscr{Y} \rightarrow \mathscr{X} \rightarrow \hat{S}$, following the data processing inequality (DPI) [**?**], and find an optimal representation by minimising the Lagrangian:

$$L[p(\hat{x}|x)] = I(\mathscr{X};\hat{S}) - \beta I(\hat{S};\mathscr{Y})$$

where the parameter β regulates the tradeoff between the complexity of the representation $R = I(\mathscr{X};\hat{S})$ and the amount of information learned $D_{IB} = I(S;\mathscr{Y})$. The optimal solutions for the IB variational problem will depend on the residual information between $\mathscr{X}$ and $\mathscr{Y}$.

For some distributions $p(\mathscr{X};\mathscr{Y})$, the exact minimal sufficient statistics may not exist. Therefore one can only reach a certain amount of compression before the information loss or distortion starts.

The optimal tradeoff is defined by a rate-distortion curve over the information plane, as depicted in Figure 5.3. The IB limit indicates an exponential growth of distortion levels or loss of information, with respect of an increment on the compression of the data representation. Over the optimal achievable limit, when the data is represented by its minimal sufficient statistics, any data compression will lead to the loss of information. However, when operating outside the minimal statistics, not all the information contained in $\mathscr{X}$ is related to $\mathscr{Y}$, and suboptimal bifurcations of the information curve arise.

Critical values of β generate bifurcation points, leading to sub-optimal curves or pseudo-stable trajectories. These bifurcations are purely related to the joint probability $p(\mathscr{X},\mathscr{Y})$, independent of any modelling assumptions. They may correspond, for instance, to the structure and topology of the $\mathscr{X}$ representations. The suboptimal bifurcations points indicate the maximum levels of compression allowed before a critical increase on data distortion.

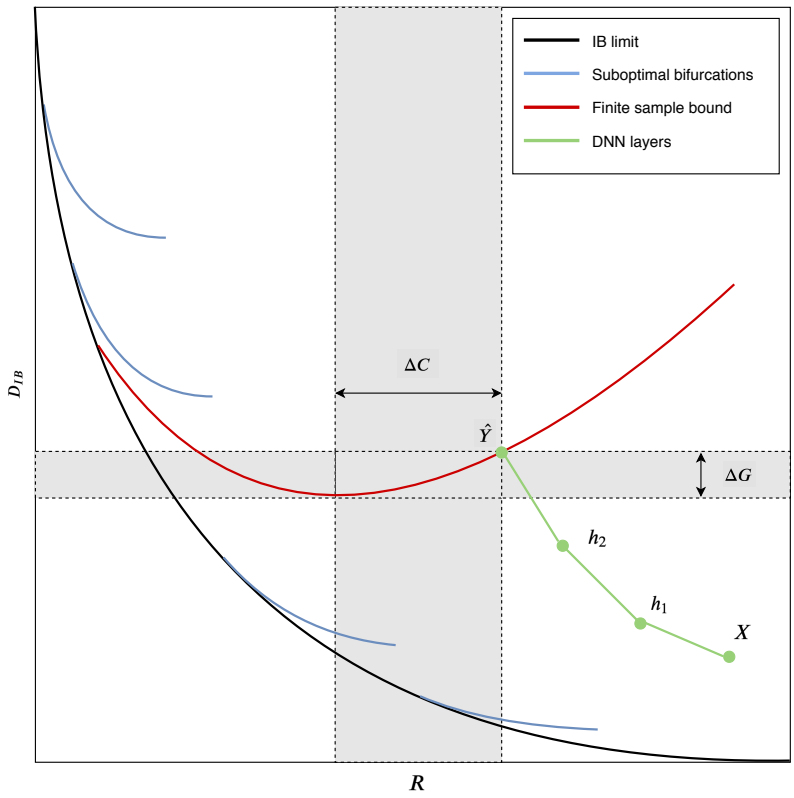

Figure 5.3: ***Information plane.*** *The information plane, introduced by Tishby et al. [100] portrays the relation between compression and distortion of information. For a given sample in the IB limit, with zero noise in its representation, any compressed depiction derived from the original leads to a major information loss. Instead, if such representation does contain non-informative features, one can achieve the optimal level of compression to reach minimal sufficient statistics. These are suboptimal bifurcation limits, with a critical balance between compression and distortion. In the finite sample bound regime, increasing values of noise in the representation leads to higher distortion. Therefore, it is possible to find the optimal value of compression and minimise distortion, following the generalisation bounds* $\{\Delta C, \Delta G\}$*. These bounds are only related to the number of samples and complexity of the output, and represent the difference in terms of compression and distortion from the minimal sufficient statistics. The information plane can also be used to study information dynamics through Deep Neural Networks (DNNs). The input* X *always display the lowest levels of compression and distortion. While moving through the network layers* $\{h_1, h_2\}$*, one can observe an increase of both parameters until the final output predicitions* $\hat{Y}$ *are reached.*

Deep Neural Networks aim to extract the most informative representation, therefore approximate minimal sufficient statistics, while using the least complex architecture.

One of the main limitations of using information theory in machine learning is the generalisation bounds and finite samples. The limited access to input and output spaces set by a finite number of samples, can have an effect on the distribution estimates and their extended implementations. However, the bounds of representational complexity does not depend directly on the dimensionality of X, but on the amount of information and its internal structure. For a given number of samples N, the minimal sufficient statistics is

able to compress all this information in $|\hat{S}| = K$ dimensions. It is possible to use the IB principle even with a finite number of samples, as long as the representational complexity is bounded to a certain number of dimensions K [101].

The generalisation bounds only depend on the number of samples and representation K. Their independence from the high dimensionality of X enables an estimation of the minimum or optimal limit over the finite sample bound curve, given the approximations.

$$I(\hat{S}; \mathscr{Y}) \leq \hat{I}(\hat{S}; \mathscr{Y}) + O(\frac{K|\mathscr{Y}|}{\sqrt{N}})$$

$$I(\mathscr{X}; \hat{S}) \leq \hat{I}(\mathscr{X}; \hat{S}) + O(\frac{K}{\sqrt{N}})$$

This gives a worst case upper bound on the optimal achievable curve, such that for a given number of samples one can find the optimal trade-off between complexity and accuracy of the representation.

5.2.1 Variational Autoencoders

The IB approach can be used to understand the learning process of Variational Autoencoders. We consider the input $\mathscr{X}$ and output $\hat{\mathscr{X}}$ as two ends of the data processing inequality (DPI), while our embedding encompasses the minimal sufficient statistics $\hat{S}$.

$$\mathscr{X} \rightarrow \hat{S} \rightarrow \hat{\mathscr{X}}$$

The model will be trained to learn the optimal compression of $\mathscr{X}$ given a certain cardinality $K = |S|$, such that the reconstructed data $\hat{\mathscr{X}}$ preserves the maximum amount of information given a sample size N.

In this particular case, the two information parameters correspond to the distortion $D_{IB} = I(\mathscr{X}; \hat{\mathscr{X}}|S)$ and compression $R = I(\mathscr{X}; S)$. VAEs aim to generate a reconstructed data representation $\hat{\mathscr{X}}$ identical to the original $\mathscr{X}$, through a compressed embedding S. Compression leads to an entropy loss between $\mathscr{X}$ and S, but ideally that loss does not increase when mapping to the reconstructed space. Therefore, the overall distortion of the VAE should be equal to the information lost on the encoder. We want S to achieve the highest compression possible, without critically increasing its levels of general distortion.

Due to the unsupervised and generative nature of VAEs, the quality of the embeddings can be evaluated from various perspectives. As a generative method, the main goal is to optimise for distortion $D_{IB} = I(\mathscr{X}; \hat{\mathscr{X}}|S)$ such that the reconstructed or synthetic data remains as similar as possible to the original. In the case of using VAEs for dimensionality reduction, the objective is to optimise for compression $R = I(\mathscr{X}; S)$, to approximate the minimal sufficient statistics while keeping a low distortion rate.

The embeddings can also be evaluated in terms of the quality of their information for a particular classification task. In this case, the goal is to assess the overlap between the information contained in S and the labels $\mathscr{Y}$. Since the VAEs achieve compression in a completely unsupervised manner, it is not possible to ensure that the embedding space will encode the information also contained in $\mathscr{X}$ and $\mathscr{Y}$. The compression achieved by S needs to be constrained to the amount of information captured between the original data $\mathscr{X}$ and the target $\mathscr{Y}$. Large compression values lead to high rates of distortion between $\mathscr{X} \rightarrow S$, which involves loss of information that can be detrimental to $S \rightarrow \mathscr{Y}$. The ultimate goal is therefore to maximise compression without compromising on the information shared between $\mathscr{X}$ and $\mathscr{Y}$.

5.2.1.1 Disentanglement

The objective function used to train VAEs is a tradeoff between reconstruction accuracy and disentanglement between the embedding components. Disentanglement is a penalty added as the divergence between the learned multivariate Gaussian, and its prior $p_\theta(z) = N(z; 0, I)$.

VAEs therefore optimise for both generative and inference models. The reconstruction loss will ensure low distortion rates, while disentanglement provides a minimal set of latent components able to compress the data. It has been shown that the ELBO function can favour reconstruction over performing the correct inference, most times due to a big difference of dimensionality between $\mathscr{X}$ and S. This is related to the loss of information through encoding layers, penalising the mutual information between the input and the latent representation [101].

Zhao et al. proposed a new class of objectives [118] that improves the quality of variational posteriors regardless of the decoding distribution, finding a more effective set of latent features. sThey derived an extended version that weights and counter-acts the discrepancy between $\mathscr{X}$ and S, as well as adding a term to minimise information loss through the encoder.

$$L_{InfoVAE} = -\lambda D(q_\phi(z)||p_\theta(z)) + E_{q(z)}[D(q_\phi(x^{(i)}|z)||p_\theta(x|z)] + \alpha I_q(x; z)$$

Notice that this particular loss is not restricted to the use of D_{KL} as a measure of divergence between $q_\phi(z)$ and $p(z)$, but instead allows other divergence families. Following the results shown in [118], we used Maximum Mean Discrepancy (MMD) [21, 32]. MMD quantifies the distance between two probability distributions by comparing all of their moments.

We show that flexible disentanglement penalties are particularly beneficial for smaller embeddings. Forcing the latent components to be disentangled can often lead to information

loss. We explore the relation between disentanglement, information and classification accuracy to understand and define a universal evaluation for unsupervised embeddings.

5.3 Implementation

The implementation of information measures to simultaneously study the compression and distortion rates introduces several challenges. In order to estimate entropies and the joint information, one needs to estimate the probability distributions associated with each layer. The quality of these approximation is usually bounded by the number of samples and layer dimensionality.

We have adapted and developed the approach presented and first used in [114] on the implementation of the 2017 Schwartz-Ziv paper [89]. The entropies and mutual information are calculated during training, with an estimation of the layer output distributions using a binning strategy. For a particular layer, the outputs of all neurons are divided and located into 100 bins, in order to estimate its distribution and entropy value. The joint information is measured between all layers, and normalised by the maximum entropy of the input, so the results can be compared independently of the original entropy values.

The analysis is performed over the original dataset of single cell RNA-Seq on Zebrafish hematopoietic cells [4], used as well in Chapter 3 and 4.

The supervised learning experiments are performed over 3900 models, with up to 4 hidden layers and sizes $|h_l| = \{8, 16, 32, 64, 128, 256, 512\}$. Each architecture is independently trained and evaluated five times. This allows the information curve estimate for the system through compression and distortion rates for each model.

For unsupervised learning, we analyse 1470 models and explore the following embedding sizes $|S| = \{8, 16, 32, 64, 128, 256\}$. Symmetric encoder and decoder networks are used, with two layers and a range of hidden layer sizes $|h_l| = \{8, 16, 32, 64, 128, 256, 512\}$. Each architecture is independently trained and evaluated five times.

The optimal architectures are then used to characterise the flow of information through the network during training.

Disentanglement between components in the embedding of the VAEs is estimated using the Kullback-Leibler divergence between Gaussian distributions, as defined by

$$D(z_1, z_2) = log(\frac{z_{\sigma_2}}{z_{\sigma_1}}) + \frac{z_{\sigma_1}^2 + (z_{\mu_1} - z_{\mu_2})^2}{2z_{\sigma_2}^2} - \frac{1}{2}$$

All the experiments were ran using a TITAN Xp GPU with 12196MiB Graphic Card frame buffer memory.

5.4 Discussion

5.4.1 Supervised learning

We explore the relation between information loss through network learning and the classification accuracy. The results are bounded by the number of samples and model complexity. We use a similar approach to the one in [101], where the information bottleneck shows the limits imposed by a finite number of samples, and their effect over classification performance. Generalisation bounds are set by the dimensionality of data, complexity of the model, and a finite number of samples. We show how the network learning involves a balance between distortion $D_{IB} = I(\mathscr{X}; \hat{\mathscr{X}}|S)$ and compression $R = I(\mathscr{X}; S)$.

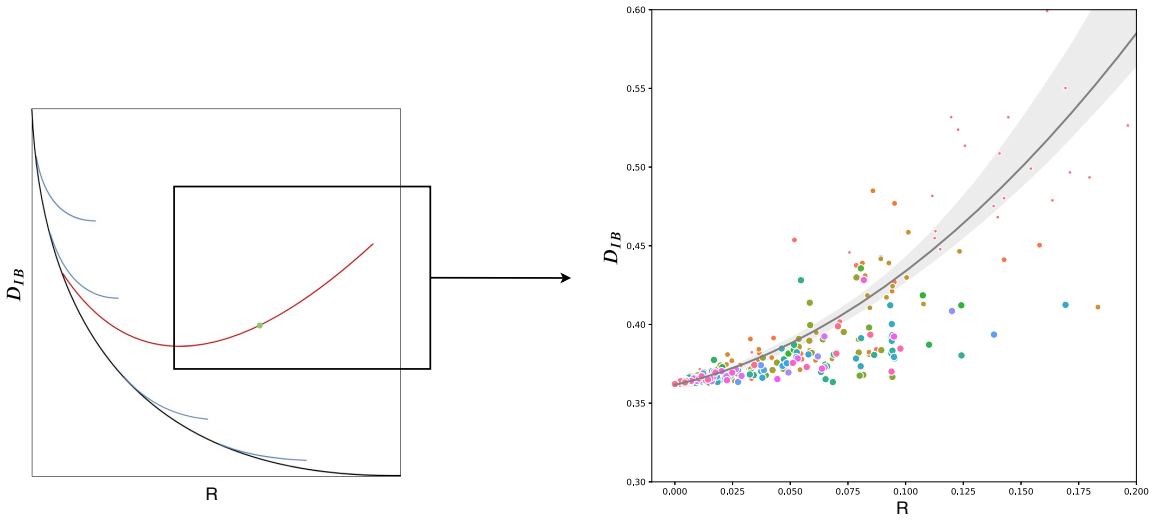

Figure 5.4: ***Finite sample bound.*** *We reproduce the finite sample bound experimentally, by measuring the compression and distortion levels of the classifiers output layer. Each point corresponds to a trained model, with colours representing particular architectures and sizes related to the number of layers. The optimal model is chosen by means of compression and distortion. A smaller value of R corresponds to a larger compression, while large values of D_{IB} represent higher distortion. Therefore, the optimal models will be those that minimize both compression R and distortion D_{IB}, on the lower left corner of the computationally generated figure.*

We extracted information measures from 3900 Neural Networks, to explore the information plane and determine the bounds imposed by sample size and system complexity. Figure 5.4 shows the finite sample bound version of the IB curve generated from the supervised learning analysis.

We observe that when approaching the maximum levels of compression, there is a stabilisation in terms of distortion. That is due to the fact that we are approaching minimal sufficient statistics, where the compression is maximal but the distortion is constant. This

optimal regime is reached when the output layer $\hat{\mathscr{Y}}$ contains all the possible information shared between the gene expression and phenotypical spaces $I(\mathscr{X}, \mathscr{Y}|\hat{\mathscr{Y}})$. The best model will be the one in the range of stable distortion with maximal compression.

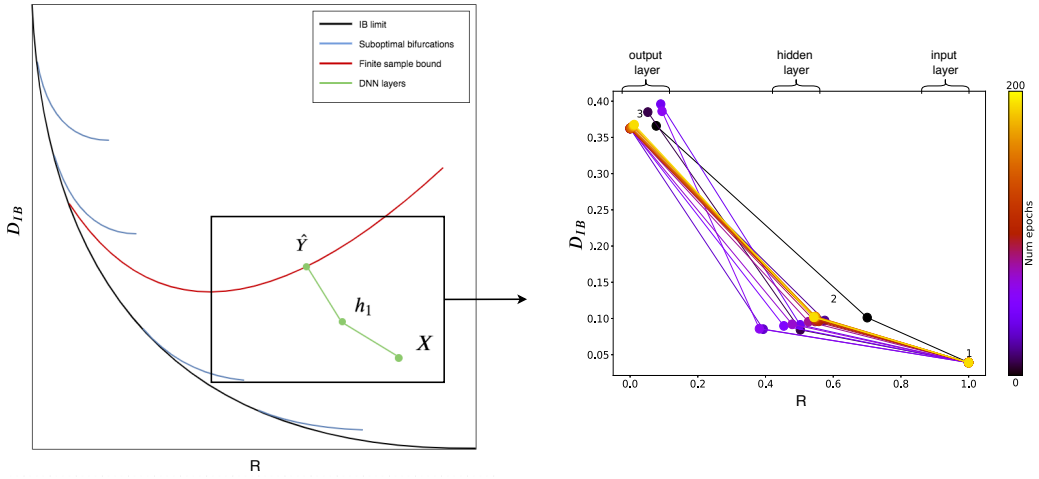

Figure 5.5: ***Information dynamics.*** *We study the dynamics of information during network training. The evolution of compression and distortion levels is analysed for different layers of the classifier. The final values achieved after network convergence to the trained solution, show stability in terms of information. Compression and distortion achieve their greatest values over the output layer, as expected.*

Once the hyper-parameters and architecture of the Neural Network are selected, one can analyse the information dynamics through the network. Figure 5.5 shows the evolution of information for each layer during training. The first layer or input has always the lowest compression and distortion values, as it contains all the information available from the original space $\mathscr{X}$. As we move forward through the network, each consecutive layer shows an increase in the level of compression, as well as higher distortion. The last layer has both high compression and distortion, as it successfully learns only the relevant information from the input $\mathscr{X}$ that overlaps with the output $\mathscr{Y}$, such that it fulfils the classification task.

The classification accuracy surface over the information plane, is portrayed in Figure 5.6. Those models with low compression and low distortion, where most of the information from $\mathscr{X}$ is still preserved, show on average lower accuracies. However, most of them converge to similar performance outputs. Models with high compression and distortion rates present a larger variability in terms of classification accuracy, but in general achieve the maximal performances. Some of the models surpass the limit of distortion and display a significant accuracy drop, due to a potential loss of relevant information for the mapping $\mathscr{X} \rightarrow \mathscr{Y}$.

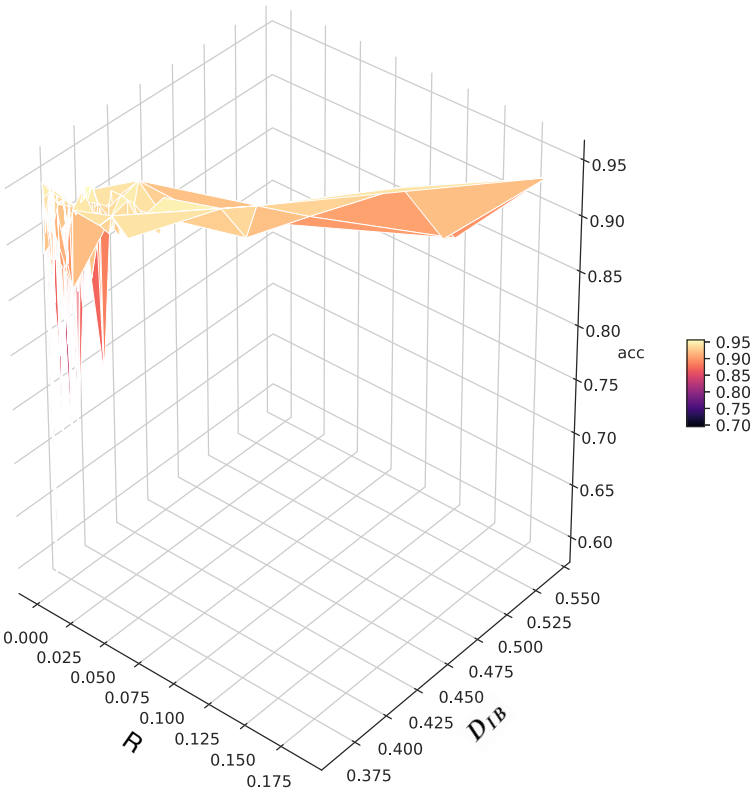

Figure 5.6: ***Accuracy surface over the information plane.*** *Accuracy distribution over the two dimensional information plane. We analysed the relation between classification accuracy with the compression and distortion values achieved by the output layers. High levels of compression lead to greater accuracies, although when combined with a large distortion the prediction power of the networks becomes more erratic. This is due to a potential over-compression of the data, involving an excess of distortion and information loss. Lower levels of compression lead to less accurate but more stable classification performances.*

5.4.2 Unsupervised learning

Using information theory to evaluate unsupervised learning provides a general interpretation of the training process, and a measure of performance independent of any bounded targets. Whether the models are used for dimensionality reduction, generative modelling or classification tasks, information measures are used to assess levels of compression and

distortion of the data, characterising information loss. We also addressed the effect of disentanglement among latent components, and its relation to the quality of information preserved.

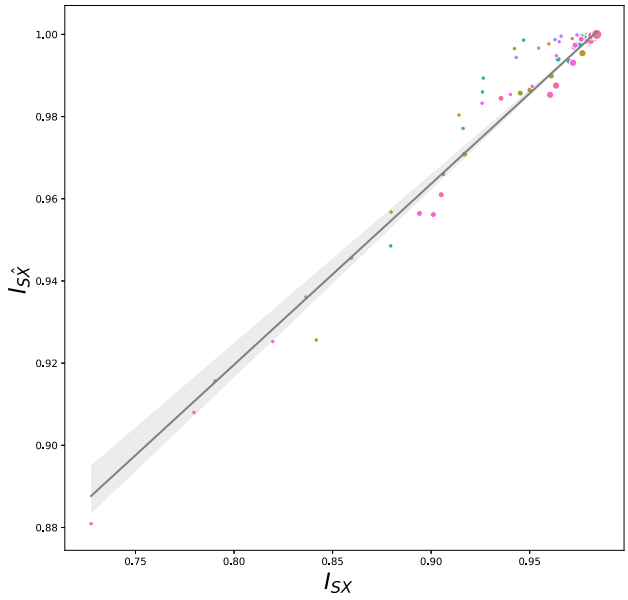

Figure 5.7: ***Mutual information of the embedding.*** *One of the main VAE objectives is to achieve compression while minimising distortion, therefore the embedding should present a relative amount of compression when compared to the original data. We measured the information shared between the original data and its embedded representation $I(S, \mathscr{X})$, together with the one shared between the embedding and VAE reconstructed data $I(S, \hat{\mathscr{X}})$. Each point corresponds to a trained model, with colours representing particular architectures and sizes related to the number of latent components. Ideally, the decoder preserves all the information encoded in S, and therefore doesn't distort the reconstructed data. The information shared by S and $\hat{\mathscr{X}}$ should be equal or greater to the one shared with $\mathscr{X}$. The optimal models will be those located over the regression line, and it's shaded standard deviation. Particularly, those with lower values of $I_{S\hat{X}}$ and therefore larger compression.*

The analysis of the information plane and curves defined by VAEs is different from the supervised case. VAEs aim to reconstruct the original input, therefore minimise distortion while maximising compression. When the models are trained and compression takes place in the embedding, the information contained in the reconstructed data can only achieve equal or lower values of that encoded by the latent representation. This can be observed in Figure 5.7, where the mutual information between the input and the embedding has a linear correlation with that of the embedding and output $I(\mathscr{X}; S) = A \times I(\mathscr{X}; \hat{\mathscr{X}}|S)$.

The optimal models are those that achieve a significant compression while minimising distortion. From Figure 5.7, the optimal architectures are the ones above the regression, where the decoder distortion is minimal for a given compression achieved through the

encoder.

The information dynamics is analysed separately through the layers of the encoder and decoder during training, as shown in Figure 5.8. We selected one of the optimal architectures from the information analysis to perform the dynamical study.

From the results we observe in Figure 5.8, when encoding the data the embedding layers are mostly in charge of the compression. Layers 3 and 4 of our model correspond to the embedding, which show low values for both compression and distortion. In the mean time, hidden layers barely contribute to compression but add significant amounts of distortion. Due to the lack of true labels as results of the unsupervised approach, the distortion levels of the encoding layer are minimal. This is the consequence of using the embedding distributions as final outcomes for measuring distortion.

For the decoder, compression levels are significant, but distortion remains minimal for most layers as it tries to preserve the information captured by the embedding. The input layer of the decoder z shows high distortion values in the first training epochs. This is due to the embedding distributions being randomly initialised, so the implementation of re-sampling over those distributions doesn't converge to the reconstructed expression, as it does for meaningful embeddings.

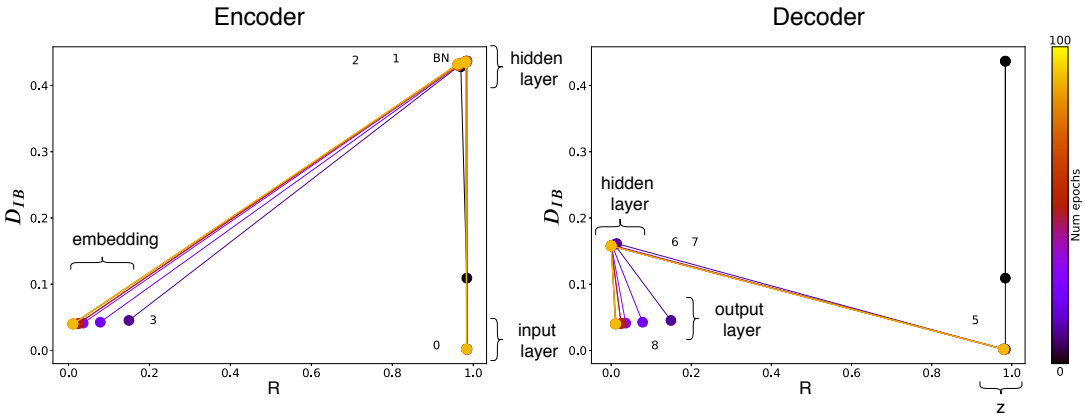

Figure 5.8: ***Information dynamics of VAEs.*** *We evaluate information dynamics for the encoder and decoder networks during training. The evolution of compression and distortion levels is analysed for different layers.* ***(left)*** *The VAE encoder achieves large values of compression in the embedding, while minimising distortion when compared to the other hidden layers.* ***(right)*** *For the decoder, distortion values are kept lower than the encoder ones, while compression remains the same as it is equivalent when mapping back to the reconstructed data* $I(S, \mathcal{X}) = -I(\hat{\mathcal{X}}, S)$.

We conducted a series of tests to analyse the relation between the amount of information compressed in the latent representation, and the embedding potential when used for classification tasks. We also study the effect of the latent components entanglement over classification performance, and its relation to compression and distortion.

From Figure 5.9 we observe that when the amount of information shared between the

input and the embedding is high, the embedding classification performance covers the entire spectrum of accuracies. In other words, for low values of compression, the embeddings can either be very accurate or very bad predictors for a particular classification task. Instead, when compression increases the majority of models present high classification accuracy. This is explained by a potential approximation of the true minimal statistics, where the excess of information contained by $\mathscr{X}$ is eliminated and only relevant information is preserved.

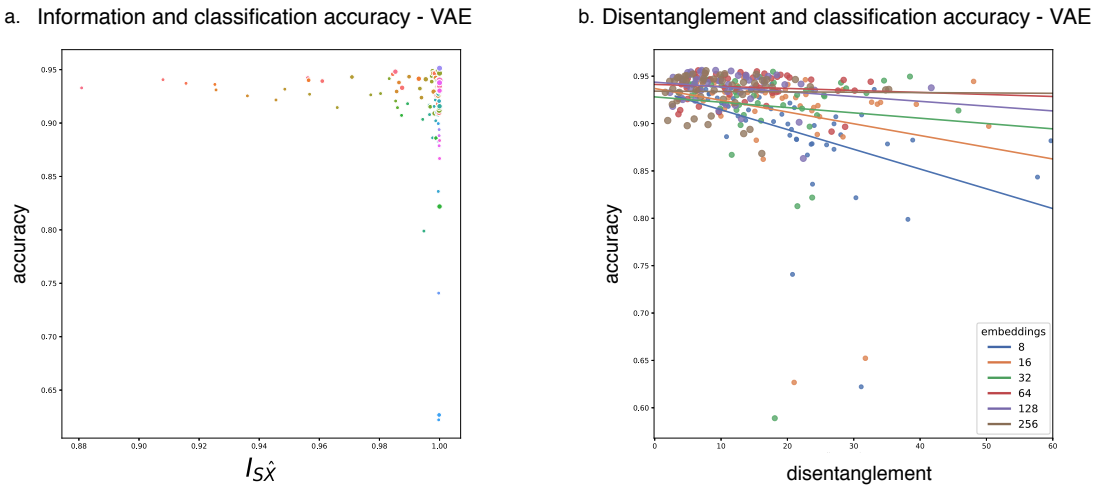

Figure 5.9: ***Information, disentanglement and classification accuracy. (left)*** *1.We evaluate the relation between classification accuracy of the embeddings and the amount of information preserved. Each data point represents the embedding of a trained VAE, with colours corresponding to the embedding dimension and size related to the network architecture. For highly compressed embeddings, the classification accuracy is higher and homogeneous among models.* ***(right)*** *2. The entanglement among latent components of the embedding has an effect over classification accuracy. For smaller embeddings with a low number of components, imposing strong disentanglement via the training loss can be detrimental in terms of classification power, as relevant information may be lost during the learning process.*

The relation between accuracy and disentanglement of the embedding latent components is shown in Figure 5.9. The disentanglement levels are relevant in terms of accuracy when the number of embedding dimensions are low. In lower dimensional embeddings, the imposition of strong disentanglement among components can be detrimental for classification accuracy. This supports the idea that a less strict divergence penalty during training allows more informative embeddings, as the covariance among components can by itself encode additional information. From these results one can extract that while constraining the learned embeddings towards disentangled distributions can improve generalisation, forcing the converged solutions towards isotropic multivariates may induce information loss. If the information lost by the disentanglement enforcement is relevant for classification tasks, this translates into a drop of accuracy values. This effect is noticeably

greater on lower dimensional embeddings, as additional latent components may be able to dissipate those effects by using the new distributions to encode the information.

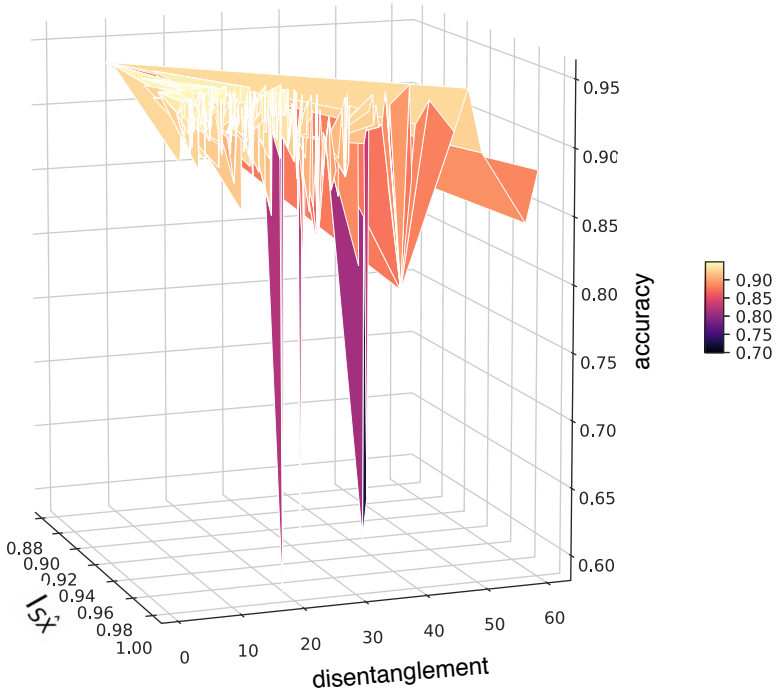

Figure 5.10: ***Accuracy surface over information and disentanglement.*** *Accuracy surface distribution as a function of information and disentanglement. We visualise the prediction power for the embedding, in relation to the amount of information preserved and the entanglement among latent components. Classification accuracy increases with higher compression, when values of $I_{S\hat{X}}$ are smaller, as the embedding succeeds on removing the excess of information contained in the original data. However, strong levels of disentanglement lead to a performance drop when imposed to the majority of embeddings.*

The analysis of information, disentanglement and classification accuracy in VAEs, has also led to some particular observations. From Figure 5.10 we conclude that embeddings that achieve a significant information compression often lead to more accurate representations of the data. Those with higher compression levels, need lower disentanglement values

in order to achieve a good performance in classification tasks. Strong disentanglement leads to lower accuracies in most cases, supporting the hypothesis that only moderate levels of entanglement are needed in order to extract informative embeddings.

5.4.3 Summary

We developed an alternative method to evaluate the performance of supervised and unsupervised models based on information theory. This approach can be used to identify optimal models and network architectures independently from the targets or data labels. It is particularly useful when such labels have large amounts of noise or uncertainty associated, such as the ones analysed in this thesis related to cell phenotypes. It is also well suited for unsupervised learning evaluation, as it doesn't depend on any prior knowledge of the data.

From the Supervised learning analysis we have established a relation between information compression, distortion and the classification accuracy. Those models with higher values of compression and minimal distortion have shown optimal classification accuracies. For our particular dataset, we have characterised the learning dynamics in terms of the entropies of our model's layers, introducing an alternative interpretation of the learning process.

The study of Unsupervised learning, and in particular VAEs, has lead to several conclusions. Since there is no definite target space, we analysed both the encoder and decoder networks in terms of compression and distortion. The optimal models are selected given a maximal compression in the encoder, and minimal distortion in the decoder. The other relevant finding of our approach has been the relation between disentanglement among encoding distributions and information compression. Given an encoder with a high level of compression, forcing a strong disentanglement between components lead to the loss of information. The classification capability of the embedding is reduced, and we observe a drop on accuracy values. Therefore a small embedding with large values of compression should impose strong disentanglement among components, as it risks the loss of information.

CHAPTER 6

Perturbation theory

The use of Variational Autoencoders for different Machine Learning tasks has drastically increased in the last years. They have been developed as denoising, clustering and generative tools, highlighting a large potential in a wide range of fields.Their embeddings are able to extract relevant information from highly dimensional inputs.

However, the results from different models trained separately can fluctuate significantly, leading to obscure interpretations and a lack of generality towards new unobserved data. Having a range of embeddings or functions representing the same original space can be seen as a degeneration problem. Such problems are very frequent in particle and quantum physics, where more than one function can be used to describe the same energy state from a particular system. We leverage the relation between theoretical physics and machine learning to explain and solve this challenge, by introducing a new approach to correct degeneration using perturbation theory.

Our objective is to derive a transformation of the VAE embedding, moving towards generalisable functions for data representation. The new functions are unique and map to all the embeddings generated by one or multiple VAEs.

VAEs provide a lower dimensional representation of the data, as a set of generative functions capable to reconstruct the original data space. Their architectures, being a combination deep neural networks, makes them very prone to suffer from model degeneracy [46, 119]. Degeneracy is closely related to model instability, often appearing as a result of placing probabilities over a reduced portion of the sample space. In order to build general and stable models, corrections need to be implemented over the generated results.

Inspired by perturbation analysis in quantum physics, we have developed a novel approach to unveil structures and the energy spectrum encoded in the data, by correcting generative functions extracted from a VAE.

The new functions represent the system and associate every class or state to a particular energy value. The energy spectrum can be derived and analysed, providing a new

unsupervised interpretation of the data.

Perturbation theory is based in the implementation of additional energy potentials on the derivation of energy functions and quantum states. Our approach has the potential to further explore different energy landscapes or perturbations, and their effect over the converged solutions.

Our methodology has been tested with both artificially generated data, and the real dataset on RNA-Seq for haematopoietic cell differentiation [4]. We prove that the new functions can be associated to the different clusters, and the energy spectrum can be related to the data structure. With promising experimental results, further research is needed in order to expand and exploit the potential of our approach to additional fields such as engineering or deep learning interpretability.

6.1 Methodology and problem definition

The VAE embedding constitutes a set of generative functions that represent our data, as multivariate Gaussians $\{S_i\}$

$$S_i = \frac{exp(-\frac{1}{2}(z_i^* - z_{\mu_i})^T \Sigma_i^{-1}(z_i^* - z_{\mu_i}))}{\sqrt{(2\pi)^k |\Sigma_i|}}$$

each sample is an observation of a particular state defined by the generative function S_i, where z_μ and Σ are the mean and covariance s-dimensional vectors of the multivariate Gaussian. Embedding layers therefore learn generative parameters for each sample, and use them in the reparametrisation trick to generate new observations z_i^*. These are decoded and mapped to the reconstructed space by the decoder. Samples in the same sub-group should be generated by the same function.

VAEs are optimised to achieve a positive definite covariance matrix Σ, where its latent components are all disentangled. However, despite the models convergence to a meaningful solution, full disentanglement is often not achieved. The entanglement among components implies that Σ is not full rank, which defines a degenerate multivariate normal distribution that can not have a density function.

The interpretation and use of the embeddings as generative functions for further modelling becomes problematic, as we don't have a unique function describing each state. Instead, a collection of similar embeddings is assigned to each sub-group.

Singular value decomposition is an extension of the polar decomposition, by which one can factorise and transform any real or complex matrix. When applied over Σ a subset of coordinates can be selected in order to transform the matrix and become positive definite. However, the choice of new coordinates only aims to bound the functions into a $rank(\Sigma)$-dimensional affine subspace, which can support Gaussian distributions. The

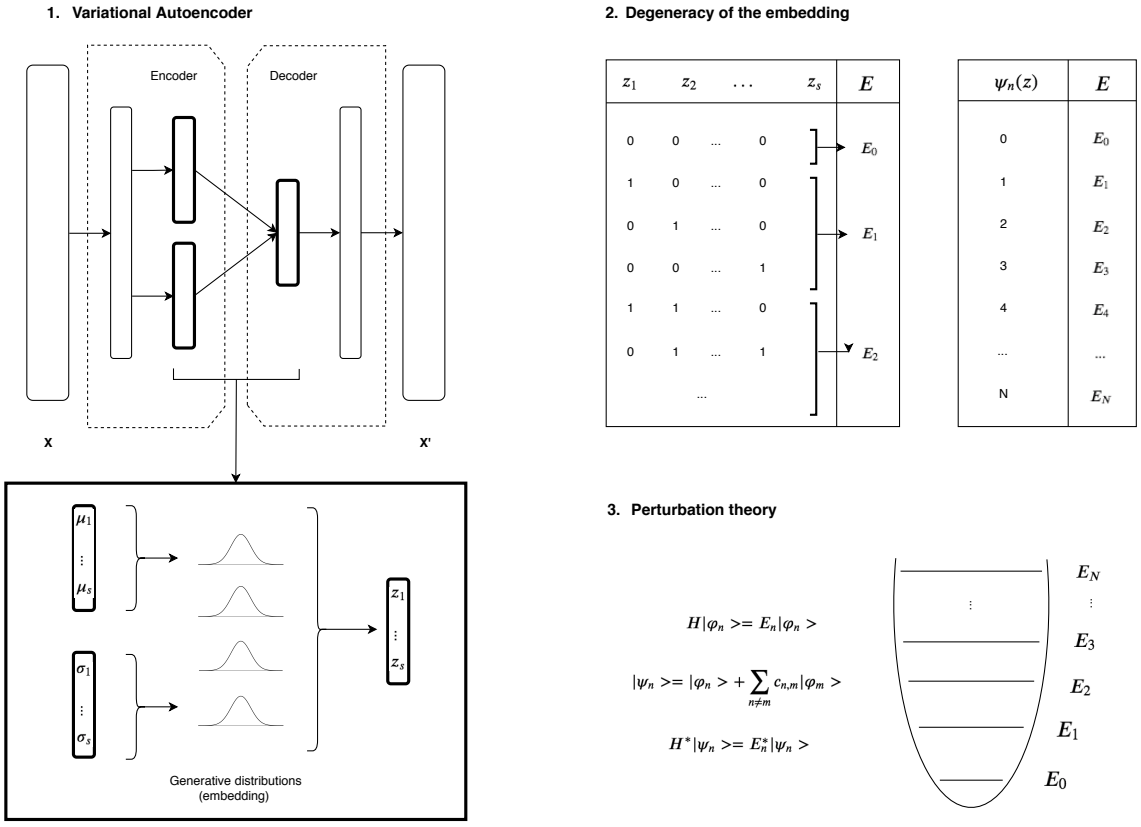

Figure 6.1: ***Diagram of the perturbation analysis approach to solve degeneracy on VAEs.***
(1) *VAEs embedding multi-dimensional generative function, where each component z is fitted to a Gaussian distribution. The latent space provides a lower dimensional representation of the data, encoding relevant features and properties among samples.*
(2) *The multi-dimensional nature and components entanglement lead to degenerate embeddings. Unique generative functions should be assigned to each sample's label. However, the embeddings learned after VAEs convergence assign a range of multivariate Gaussians to samples of the same class, due to modelling symmetries and information overlap among latent components. We want to obtain the truly generalisable generative functions ψ_n with a unique characterisation of the system states.*
(3) *Perturbation theory is used to unveil a spectrum of energies that can be associated to the samples, building a more general approach to interpret the latent space learned through VAEs. By applying a perturbed Hamiltonian over the embeddings, we define the new functions and energy spectrum that represents our system. Perturbed functions are a linear combination of the unperturbed ones, where the coefficients $c_{n,m}$ characterise their separability.*

objective of such transformation is only to define a positive definite covariance matrix, without accounting for any loss of information. We have experimentally shown that singular value decomposition is incapable of defining a new set of generative functions based on VAEs embeddings, as it induces the loss of a significant amount of relevant

information. Therefore, the new set of functions obtained via singular value decomposition are unable to separate between the different data sub-groups or states.

For this reason, we developed a Multilayer Perceptron that learns the functions $\psi_n(S)$, accounting for degeneracy in S. We are able to define a unique set of generative functions ψ_n while encoding relevant features from the data. An overiview of our approach is summarized in Figure 6.1.

The transformation is based on a perturbation theory approach, broadly used in atomic physics, condensed matter and particle physics to solve quantum mechanical problems. It uses a scheme of successive corrections to the zero-field values of energy levels and wave-functions, to identify eigen-states and eigen-energies associated with the system.

In our case, eigen-states ψ_n and eigen-energies E_n correspond to the different data clusters or sub-groups, defined by a quantum number n. Our approach assigns particular energies to each eigen-state, based on the solutions of the learned functions ψ_n.

Given the Hamiltonian operator H of a certain energy landscape, with known eigen-states and eigenvalues $H\left|\psi_n\right\rangle = E_n\left|\psi_n\right\rangle$, one can study how these eigen-states and eigen-energies change when small perturbations are added.

$$H\left|\psi_n\right\rangle = E_n\left|\psi_n\right\rangle$$
$$(H^0 + H^1)\left|\psi_n\right\rangle \approx (E_n^0 + E_n^1)\left|\psi_n\right\rangle$$

Using a first order approximation of the expansions $H = \sum_i^m H^i$ and $E_n = \sum_i^m \lambda^i E_n^i$. The eigen-states or functions $|\psi_n\rangle$ directly derived from VAEs embeddings can be seen as a combination of ground states or a general truth, and a perturbation term or bias added by each particular VAE solution.

$$\left|\psi_n\right\rangle = \left|\phi_n\right\rangle + \sum_{m\neq n} c_{m,n}\left|\phi_m\right\rangle$$

where $|\phi_n\rangle$ are the states theoretically derived from the unperturbed problem with solvable Hamiltonian H^0.

We used as ground truth the unperturbed Hamiltonian from the "particle in a box" problem in quantum physics. In this scenario, H^0 corresponds to the kinetic energy of such particle, and the eigen-states and eigen-energies can be derived by solving the Schrödinger equation. The solutions of $H^0 = -A\frac{d^2}{dz^2}$ are given by:

$$E_n^0 = (\frac{\Pi}{L})^2 An^2$$
$$|\phi_n\rangle = \sqrt{\frac{2}{L}} sin(\frac{\Pi n}{L} z)$$

We have chosen a first order perturbation with t evenly distributed minima on the z domain $H^1 = V(z) = sin(2\pi tz)$. The perturbation potential used was selected for its uniform and universal nature, adding simplicity when generalised to systems with different number of clusters or structure. However, this approach is general and the perturbation implemented can be changed according to the needs and requirements of each problem.

Given the set $\{E^0, E^1, |\phi_n\rangle, |\psi_n\rangle\}$ of all perturbed and unperturbed energies and wave functions, one can extract the particular subset of energies and wave-functions that uniquely characterise our system.

The corresponding perturbed energies E_n^1 and wave functions for our particular perturbation are derived from:

$$E_n^1 = \langle\phi_n| H^1 |\phi_n\rangle$$
$$|\psi_n\rangle = |\phi_n\rangle + \sum_{m\neq n} c_{m,n} |\phi_m\rangle$$
$$c_{m,n} = \langle\phi_m| H^1 |\phi_n\rangle$$

6.2 Implementation

We tested our approach on artificially generated and clustered data, using *sklearn* toolkit *random sample generator* [74]. The sample generator creates a multi-class dataset by allocating one or multiple normally-distributed clusters of points to each class. It also adds noise in the form of correlated, redundant and uninformative features.

The synthetic datasets generated $\mathscr{X}^* = \{x^{*(i)}\}_{i=1}^{n^*}$ have $x^{*(i)} = [x^{*(i)}_1 x; *(i)_2 \dots x^{*(i)}_{m^*}]$ and are used to train a VAE, with an embedding S with s components that eliminates noise and preserves relevant information from the original input. We used $n^* = 2000$ and $m^* = 1000$ to create $\mathscr{X}^*$ with $k = 3$ clusters, encoded in $s = 10$ dimensional VAE embedding.

We also implemented our method to the single cell RNA-Seq haematopoietic dataset used in previous chapters $\mathscr{X} = \{x^{(i)}\}_{i=1}^{n}$ and $x^{(i)} = [x_1^{(i)} x_2^{(i)} \dots x_m^{(i)}]$, with $n = 1724$ cells and $m = 1871$ genes. We adopted the same $s = 64$ dimensional VAE embedding presented previously in this thesis.

We use the learned embeddings to find a unique solution for the system by solving the

Schrödinger equation with a particular perturbed potential [18, 36]. The eigen-function observations $|\psi_n\rangle$ are obtained from the outputs of a Multi-Layer Perceptron, trained to optimise the energy values of our given energy potential.

The Neural Network trained is a simple Multi-Layer Perceptron (MLP), with an $|S| = s$ dimensional input, a single hidden layer with $h_l = \frac{s}{2}$ nodes and ReLU activation, fully connected to a linear output that returns the values of ψ_n. The objective or likelihood function used has the following form:

$$L(\psi, z) = E^0 + E^1$$
$$E^0 = \alpha n^2$$
$$E^1 = (-1)^t \frac{cos(\pi t)[4n^2 + t^2 cos(2\pi n) - N^2] + 2n[tsin(2\pi n)sin(\pi t) - 2n]}{2\pi t[t^2 - 4n^2]}$$

where α weights the influence of the unperturbed or kinetic energy, and t is a hyperparameter defined by the number of minima in the perturbation potential $V(z)$. We used $\alpha = 1$ and $\{t^* = 3, t = 5\}$. The value of n, classically known as quantum number, is derived from its respective periodic wave-function, which in our case is:

$$n = \frac{L}{\pi} \frac{arcsin(\frac{L\psi^2}{2})}{z}$$

.

Pytorch [71] version 0.4.1. was used to build and train the VAEs and Neural Networks.

6.3 Discussion

The Gaussian distributions associated to components of the VAE embedding, for the artificially generated dataset $\mathscr{X}^*$, are shown in Figure 6.2.

Entanglement among components induces degeneracy over the new data representation, where several functions may map to the same cluster of data. Therefore, we highlight the need for generalisable generative functions that can be uniquely associated with the data structure, so they can correctly generate and be used in further modelling studies.

As shown in the previous sections, perturbed energy states can be decomposed as the combination of ground energy states and additional perturbation terms.

$$|\psi_n\rangle = |\phi_n\rangle + \sum_{m \neq n} c_{m,n} |\phi_m\rangle$$

The observations of $|\psi_n\rangle$ obtained from the transformation of VAEs embedding S are portrayed in Figure 6.3. The tSNE representation of the embedding shows the $k = 3$

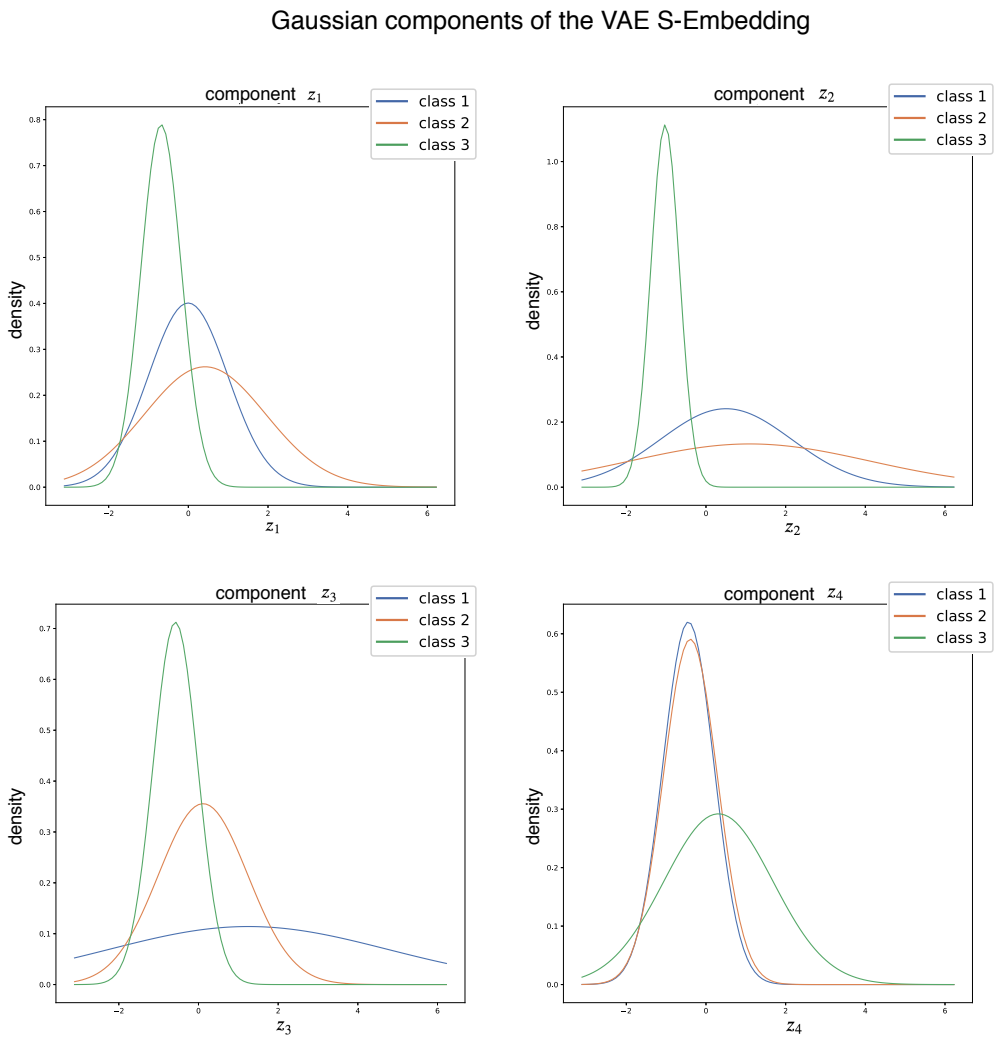

Figure 6.2: ***Gaussian components of the VAE embedding.*** *Latent components of the Gaussian multivariate embedding obtained from VAEs, implemented over an artificially generated dataset with $k = 3$ clusters and $|S| = 10$ embedding. We visualise the normal distributions associated to a few of the embedding components learned by the VAE, for samples from the 3 different clusters. Clockwise order of the figures correspond to components z_i with $i = \{0, 1, 2, 3\}$. Some components successfully learn to characterise particular classes, while others display a significant overlap that diminishes separability.*

clusters artificially generated in $\mathscr{X}^*$. Each cluster has its own generative function and quantum number $|\psi_n\rangle$, which can be identified by using the perturbation coefficients $c_{m,n}$. These coefficients are estimated by analytically solving the integral:

$$c_{m,n} = \langle \phi_m | \, H^1 \, | \phi_n \rangle = \int \phi_m^T V(z) \phi_n dz$$

and consequently associated to each cluster or group of observations.

The implementation of our perturbation approach to derive alternative generative functions from the single cell RNA-Seq. haematopoietic dataset is shown in 6.4. In this case, the number of clusters $k = 5$ is higher, as it is the noise level in the original dataset $\mathscr{X}$. However, the VAE embedding is able to compress the information and eliminate the excess, generating a lower dimensional representation as a set of multivariate Gaussians with $s = 64$.

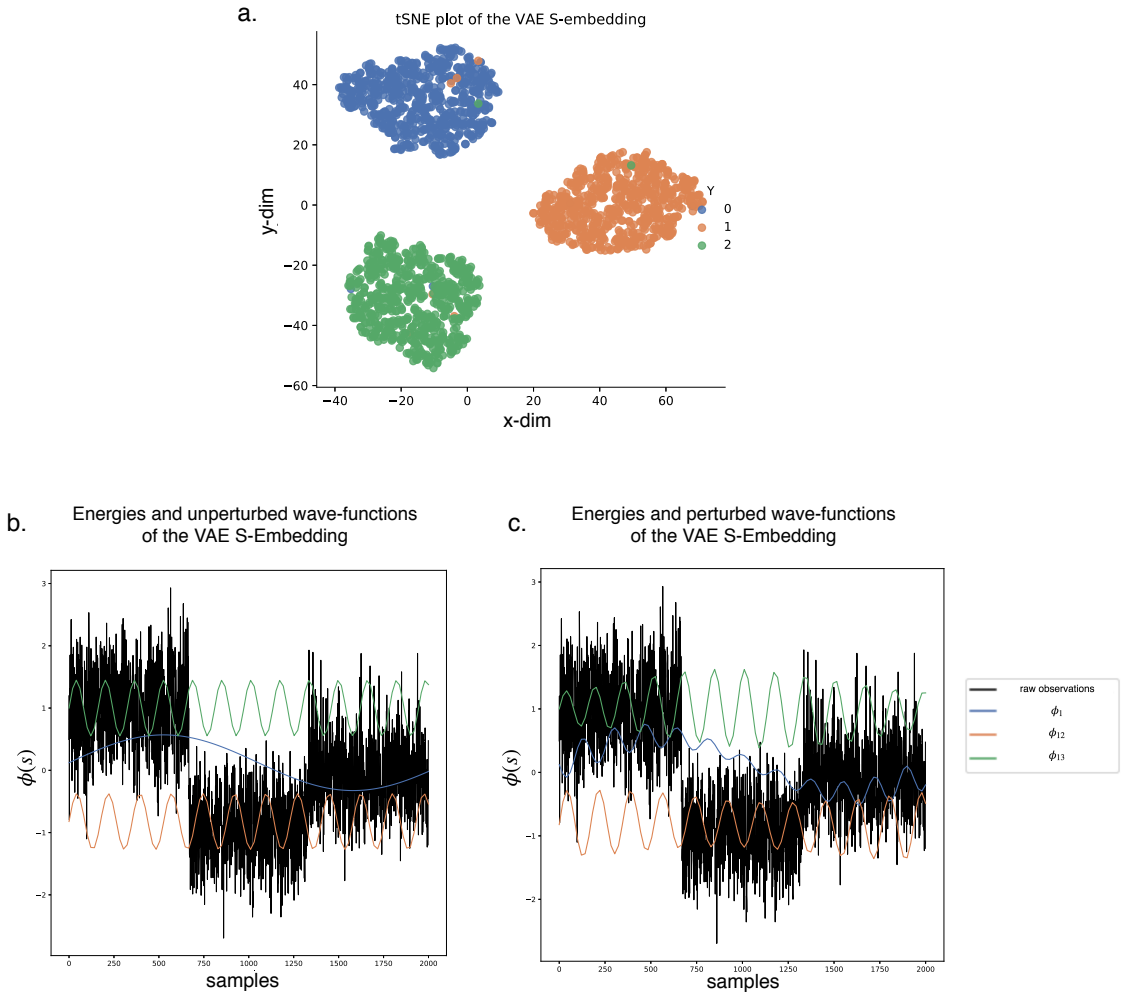

Figure 6.3: ***Functions*** $\psi_n(S)$ ***and spectrum of energies for the clustered synthetic dataset*** $\mathscr{X}^*$. ***(a)*** *tSNE visualisation of the VAE learned embeddings, proving their ability to separate samples according to their true cluster labels.* ***(b,c)*** *In black there are the observations of new generative functions* ψ_n*, as a transformation of the embeddings* S *through an MLP with perturbed objective function. The quantum numbers and energy states associated to each cluster are depicted in different colours by the unperturbed (left) and perturbed (right) unique generative functions.*

The output of our perturbation analysis MLP provides a set of observations from multiple potential energy states $|\psi_n\rangle$, some of them with a relative overlap. However,

using the labels to sort samples allows the observation of generative patterns among the data. Clusters or groups of samples can be associated to different generative functions with unique quantum numbers n, and consequently derive the energy spectrum E_n.

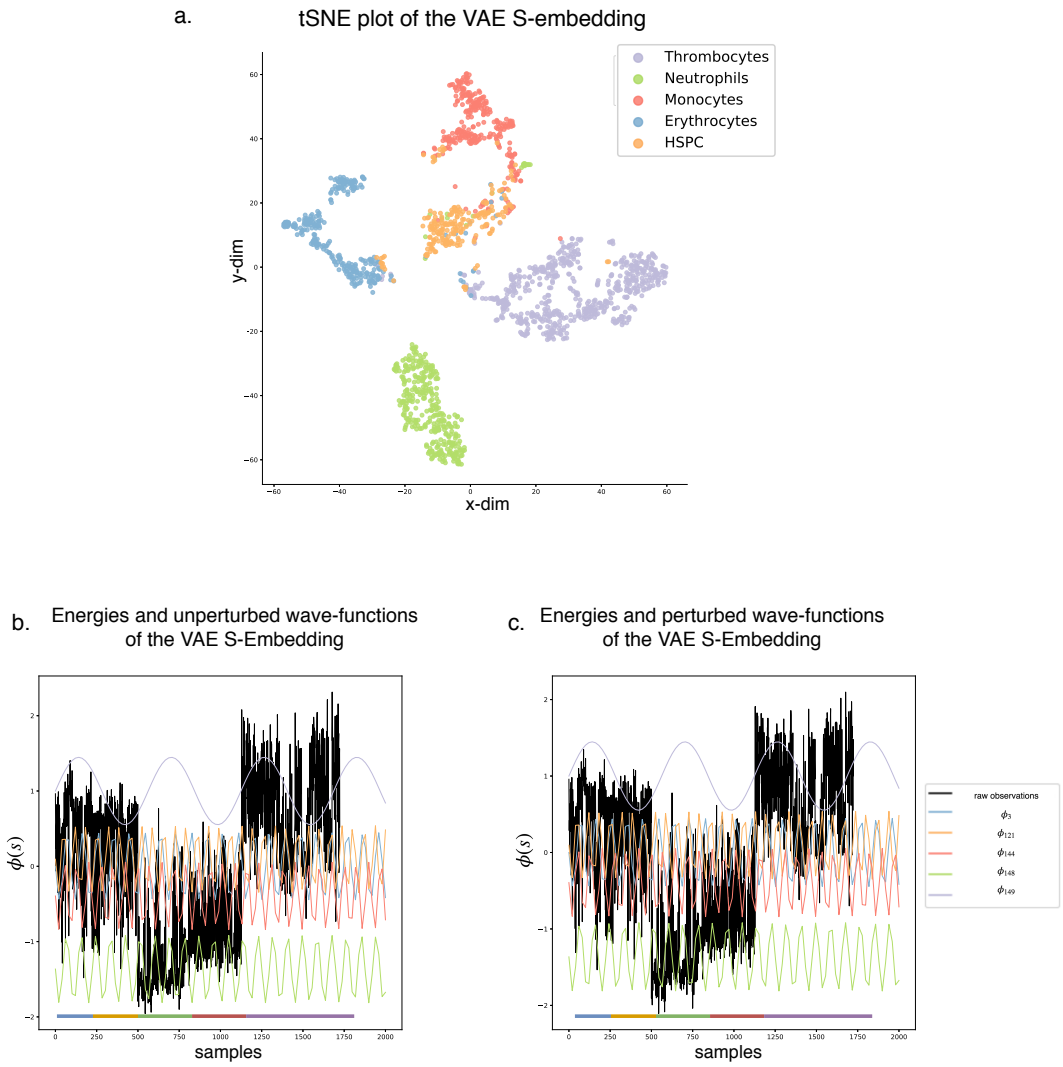

Figure 6.4: ***Functions $\psi_n(S)$ and spectrum of energies for the single cell RNA-Seq dataset $\mathscr{X}$.***
(a) *tSNE visualisation of the VAE learned embeddings, proving their ability to separate samples according to their true cluster labels. Genetic data has an additional layer of complexity, and some of the clusters are not entirely separable in S.* ***(b, c)*** *In black there are the observations of new generative functions ψ_n, as a transformation of the embeddings S through an MLP with perturbed objective function. Samples are sorted according to cell types, and several energy levels can be distinguished among observations, despite a substantial amount of added noise. Quantum numbers and energy states associated to each cell type are depicted in different colours by the unperturbed (left) and perturbed (right) unique generative functions. Since the quantum numbers associated to the cell states are high, the difference between perturbed and unperturbed functions is not strongly pronounced. However, this behaviour was expected due to the high level of noise and similarity among the mature cell clusters.*

One of the particular observations that arise from the quantum numbers and energies obtained for the cellular differentiation data, is the relation with clusters similarities. Clusters with differentiable profiles, from mature cells such as Thrombocytes or Erythrocytes, are associated to the highest quantum numbers and energies. While stem or transitioning cells are believed to be generated by functions with lower energy values.

The results obtained from this approach provide a new interpretation of the VAE embedding functions, with a general representation of the systems. In the future, further research can be developed by exploring different perturbations or new energy landscapes, while analysing their effect over different datasets and their structural representation. The encoding functions could also be used to expand the information boundaries imposed by the intersection of X and Y, as they can be used to generate new data and therefore explore new potential conformations.

6.3.1 Summary

We have introduced a new transformation over the generative embedding of VAEs, based on the analogy with degenerate systems in quantum physics. By using perturbation theory, we are able to transform and define a new set of unique functions that correspond to each class or data-label, from a completely unsupervised approach. Therefore, we are able to identify a set of energies or quantum states, together with their corresponding functions, that are characteristic to the system states instead of the samples being studied (as it happens with the embedding functions obtained directly from VAEs). The transformation succesfully reduces complexity of the data representation and provides a high level intuition over the system states.

This is beneficial in terms of generalisation, as the new functions should be able to extend their generative potential and be applied to new datasets. This transformation can also be leveraged by further mathematical modelling of systems and processes, such as cell differentiation. The unique functions can be used as a unique representation of each system class, to generate new samples for variational modelling of differentiation processes or study the similarity between cell types. It can be used to characterise variability and robustness of the different states.

In this Chapter we have introduced the idea of a transformation based on the similarity of VAEs embeddings with degenerate systems. We believe that with further analysis, valuable results can be achieved in terms of generalisation and specific applications of this approach. The hyperparameters and architecture of the multi-layer perceptron and energy minimization have been optimised based on the two generative examples presented. Therefore a new grid-search and optimisation is recommended when implemented to new datasets and systems.

CHAPTER 7

Conclusion and future directions

The research presented in this dissertation is the result of combining knowledge from the fields of computer science, mathematics, physics and biology. We have introduced novel deep learning approaches to analyse single cell genomic data and interpret its biological and mathematical representation.

In ***Chapter 3*** we developed a successful direct mapping between genotype and phenotype spaces $\mathscr{G} \rightarrow \mathscr{P}$, when the labels in the phenotypical space are available. We show that linear methods are useful to discriminate among adult cell types, but not accurate at separating cells in process of differentiation or stem cells. Classification accuracy was then used to compare and identify optimal neural network architectures to perform the mapping between genome and cell types. Finally, we introduce a new measure of maturity for individual cells, to characterise the cells stage in the process of differentiation and identify those that are in the initial or stem cell states. This measure is based on the outputs or predictions of the non-linear classifier, and can be validated by further interpretation of the input features or biological markers relevant for such results.

Future directions of research arising from these results could lead into the design of experimental protocols to validate the computational results obtained. New markers can be extracted from the input features of the classifier, and their relevance towards stemness measures can be tested experimentally in the fields of synthetic biology and genetic studies.

Chapter 4 presents an unsupervised learning framework, with a lower dimensional representation that captures relevant properties of the data with the potential to generate synthetic samples. We designed a pipeline to map from genotype to phenotype without the need for cell labels, by combining deep unsupervised and generative models (VAEs) with feature extraction. Although biological knowledge is still needed to analyse the extracted features and define an appropriate classification, we improved the interpretability of the method by introducing feature extraction, and identifying the relevant genes for each

cluster in the embedding. Based on classification accuracy of the embedding we selected optimal architectures and training hyperparameters for VAEs. The clustering accuracy of VAEs embeddings is greater than other linear dimensionality reduction methods, such as PCA, and also the original and reconstructed gene expression spaces.

The designed pipeline has been tested and implemented on additional datasets, such as human pancreatic and haematopoietic cells, as shown in Bica et al. [12]. An extensive analysis of the features extracted, both experimentally and computational, can be done to identify new potential cell markers. The combination of feature extraction with gene ontologies can lead to a new spectrum of interpretable models, leveraging previous knowledge and structural information of the data. For instance, Graph Neural Networks and relational inference techniques can be developed to combine the neural unsupervised potential with networks associated to particular systems, driving the learning towards explainable solutions. VAEs have also been exploited for data integration, building cross-modal informative embeddings. We developed and compared different VAE architectures for this purpose in Simidjievski et al. [93] with gene expression, clinical and copy number alteration data for cancer applications. We are also working on the comparison of synthetic data obtained from the VAE reconstructed genetic space $\mathscr{X}^*$, with the results obtained from Generative Adversarial Networks (GANs).

Chapter 5 implements a new approach to evaluate the training and final outputs f Neural Networks based on information theory, within the Information Bottleneck theory. We define it's interpretation in terms of information sharing between inputs and target spaces, and its effect over accuracy values, for supervised and unsupervised learning. The information curve is generated for a Genotype-Phenotype classifier, detecting the optimal architectures for compression and distortion, while studying their relation to classification accuracy. The learning process is characterised by the flow of information and its dynamics along network layers during training.

The IB approach has also been implemented over VAE models, in order to provide a non-supervised evaluation of the embeddings. Optimal architectures are identified based on compression and distortion terms. The flow of information is also analysed for the encoder and decoder layers, showing significant compression in the encoder and minimal distortion through the decoder. We have studied the relation between entanglement among latent components and classification accuracy. The results show that disentanglement is relevant to ensure generalisation, but for low dimensional embeddings the extreme enforcement of disentanglement is detrimental for classification accuracy, due to information loss.

The benefits of information theory as a new validation approach for deep learning are particularly interesting for unsupervised models. It can be implemented for new clinical and biological applications, as an alternative evaluation of the models and results obtained.

Furthermore, it can contribute to the development of pure exploratory research based on unsupervised learning. Information dynamics add a new perspective to network training, with potential crossings with network pruning and optimisation.

The introduction and construction of a novel approach to generalisable embeddings is presented in ***Chapter 6***. A new family of general generative functions is derived from VAE embeddings by implementing perturbation theory. We identify the presence of degeneracy among VAEs, induced by entanglement and symmetries among its generative functions. Perturbation theory is then used to transform S into a new set of unique generative functions ψ_n that characterise the system, and extract the energy spectrum and quantum numbers associated to each cluster. We tested this approach with artificial and real data, proving that we can provide an unsupervised alternative lower dimensional and generative interpretation of the data, solving VAE degeneracy.

We have shown that the intersection of physics and machine learning covers a wide range of applications. For instance, network optimisation and model convergence are tightly linked to the loss energy landscape, which is related to stochastic dynamics and transitions in dynamical systems, long studied by physicists and chemists [25, 37, 52]. The relation of such landscapes to the system structure and complexity is still to be explored. The multi-layered structure of Deep Neural Networks can also be used to derive and explain multiple processes that arise during network training [11], as well as some of the converged solutions of the final models. We believe that the combination of theoretical physics for high dimensional problems and dynamical systems, with the fast development of models based on Deep Neural Networks, will be at the forefront of machine learning research.

Bibliography

[1] J. Adolfsson, R. Månsson, N. Buza-Vidas, A. Hultquist, K. Liuba, C. T. Jensen, D. Bryder, L. Yang, O.-J. Borge, L. A. Thoren, et al., *Identification of flt3+ lympho-myeloid stem cells lacking erythro-megakaryocytic potential: a revised road map for adult blood lineage commitment*, Cell, 121 (2005), pp. 295–306. Elsevier.

[2] A. A. Alemi, I. Fischer, J. V. Dillon, and K. Murphy, *Deep variational information bottleneck*, *arXiv preprint arXiv:1612.00410, (2016).

[3] P. Angerer, L. Haghverdi, M. Büttner, F. J. Theis, C. Marr, and F. Buettner, *destiny: diffusion maps for large-scale single-cell data in r*, Bioinformatics, 32 (2015), pp. 1241–1243. Oxford University Press.

[4] E. I. Athanasiadis, J. G. Botthof, H. Andres, L. Ferreira, P. Lio, and A. Cvejic, *Single-cell RNA-sequencing uncovers transcriptional states and fate decisions in haematopoiesis*, vol. 8, 2017. Nature Publishing Group.

[5] F. Attneave, *Dimensions of similarity*, The American journal of psychology, 63 (1950), pp. 516–556. JSTOR.

[6] P. Baldi and K. Hornik, *Neural networks and principal component analysis: Learning from examples without local minima*, Neural networks, 2 (1989), pp. 53–58. Elsevier.

[7] R. G. Baraniuk, V. Cevher, and M. B. Wakin, *Low-dimensional models for dimensionality reduction and signal recovery: A geometric perspective*, Proceedings of the IEEE, 98 (2010), pp. 959–971. IEEE.

[8] S. Behjati and P. S. Tarpey, *What is next generation sequencing?*, Archives of Disease in Childhood-Education and Practice, 98 (2013), pp. 236–238. Royal College of Paediatrics and Child Health.

[9] R. Bellman, *Dynamic programming*, New York, NY, 707 (2013). Courier Corporation.

[10] E. Bianconi, A. Piovesan, F. Facchin, A. Beraudi, R. Casadei, F. Frabetti, L. Vitale, M. C. Pelleri, S. Tassani, F. Piva, et al., *An estimation of the number of cells in the human body*, Annals of human biology, 40 (2013), pp. 463–471. Taylor & Francis.

[11] G. Bianconi, *Multilayer Networks: Structure and Function*, Oxford university press, 2018.

[12] I. Bica, H. Andrés-Terré, A. Cvejic, and P. Liò, *Unsupervised generative and graph representation learning for modelling cell differentiation*, *BioRxiv, (2019), p. 801605. Cold Spring Harbor Laboratory.

[13] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, *Fast unfolding of communities in large networks*, Journal of statistical mechanics: theory and experiment, 2008 (2008), p. P10008. IOP Publishing.

[14] H. Bourlard and Y. Kamp, *Auto-association by multilayer perceptrons and singular value decomposition*, Biological cybernetics, 59 (1988), pp. 291–294. Springer.

[15] G. Brady, M. Barbara, and N. N. Iscove, *Representative in vitro cdna amplification from individual hemopoietic cells and colonies*, Methods Mol Cell Biol, 2 (1990), pp. 17–25.

[16] K. R. Campbell and C. Yau, *A descriptive marker gene approach to single-cell pseudotime inference*, Bioinformatics, 35 (2018), pp. 28–35. Oxford University Press.

[17] G. Carleo and M. Troyer, *Solving the quantum many-body problem with artificial neural networks*, Science, 355 (2017), pp. 602–606. American Association for the Advancement of Science.

[18] ——, *Solving the quantum many-body problem with artificial neural networks*, Science, 355 (2017), pp. 602–606. American Association for the Advancement of Science.

[19] A. Ciau-Uitz, R. Monteiro, A. Kirmizitas, and R. Patient, *Developmental hematopoiesis: ontogeny, genetic programming and conservation*, Experimental Hematology, 42 (2014), pp. 669–683. Elsevier.

[20] B. Du, W. Xiong, J. Wu, L. Zhang, L. Zhang, and D. Tao, *Stacked convolutional denoising auto-encoders for feature representation*, IEEE transactions on cybernetics, 47 (2016), pp. 1017–1027. IEEE.

[21] G. K. Dziugaite, D. M. Roy, and Z. Ghahramani, *Training generative neural networks via maximum mean discrepancy optimization*, *arXiv preprint arXiv:1505.03906, (2015).

[22] J. Eberwine, H. Yeh, K. Miyashiro, Y. Cao, S. Nair, R. Finnell, M. Zettel, and P. Coleman, *Analysis of gene expression in single live neurons*, Proceedings of the National Academy of Sciences, 89 (1992), pp. 3010–3014. National Acad Sciences.

[23] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein, *Cluster analysis and display of genome-wide expression patterns*, Proceedings of the National Academy of Sciences, 95 (1998), pp. 14863–14868. National Acad Sciences.

[24] A. Eldar and M. B. Elowitz, *Functional roles for noise in genetic circuits*, Nature, 467 (2010), p. 167. Nature Publishing Group.

[25] H. Eyring, *The activated complex in chemical reactions*, The Journal of Chemical Physics, 3 (1935), pp. 107–115. AIP.

[26] J. J. Faith, M. E. Driscoll, V. A. Fusaro, E. J. Cosgrove, B. Hayete, F. S. Juhn, S. J. Schneider, and T. S. Gardner, *Many microbe microarrays database: uniformly normalized affymetrix compendia with structured experimental metadata*, Nucleic acids research, 36 (2007), pp. D866–D870. Oxford University Press.

[27] F. Fleuret, *Fast binary feature selection with conditional mutual information*, Journal of Machine learning research, 5 (2004), pp. 1531–1555.

[28] M. J. Foulkes, K. M. Henry, J. Rougeot, E. Hooper-Greenhill, C. A. Loynes, P. Jeffrey, A. Fleming, C. O. Savage, A. H. Meijer, S. Jones, et al., *Expression and regulation of drug transporters in vertebrate neutrophils*, Scientific reports, 7 (2017), p. 4967. Nature Publishing Group.

[29] P. Gallinari, Y. Lecun, S. Thiria, and F. F. Soulie, *Mémoires associatives distribuées: une comparaison (distributed associative memories: a comparison)*, in Proceedings of COGNITIVA 87, Paris, La Villette, May 1987, Cesta-Afcet, 1987.

[30] R. Gilad-Bachrach, A. Navot, and N. Tishby, *Margin based feature selection-theory and algorithms*, ACM, Proceedings of the twenty-first international conference on Machine learning, 2004.

[31] A. V. Gore, L. M. Pillay, M. Venero Galanternik, and B. M. Weinstein, *The zebrafish: A fintastic model for hematopoietic development and disease*, Wiley Interdisciplinary Reviews: Developmental Biology, 7 (2018), p. e312. Wiley Online Library.

[32] A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, and A. J. Smola, *A kernel method for the two-sample-problem*, Advances in neural information processing systems, 2007.

[33] A. Gribov, M. Sill, S. Lück, F. Rücker, K. Döhner, L. Bullinger, A. Benner, and A. Unwin, *Seurat: visual analytics for the integrated analysis of microarray data*, BMC medical genomics, 3 (2010), p. 21. BioMed Central.

[34] G. Guo, S. Luc, E. Marco, T.-W. Lin, C. Peng, M. A. Kerenyi, S. Beyaz, W. Kim, J. Xu, P. P. Das, et al., *Mapping cellular hierarchy by single-cell analysis of the cell surface repertoire*, Cell stem cell, 13 (2013), pp. 492–505. Elsevier.

[35] M. Guo, H. Wang, S. S. Potter, J. A. Whitsett, and Y. Xu, *Sincera: a pipeline for single-cell rna-seq profiling analysis*, PLoS computational biology, 11 (2015). Public Library of Science.

[36] J. Han, L. Zhang, et al., *Solving many-electron schrödinger equation using deep neural networks*, *arXiv preprint arXiv:1807.07014, (2018).

[37] P. Hänggi, P. Talkner, and M. Borkovec, *Reaction-rate theory: fifty years after kramers*, Reviews of modern physics, 62 (1990), p. 251. APS.

[38] E. A. Harvie and A. Huttenlocher, *Neutrophils in host defense: new insights from zebrafish*, Journal of leukocyte biology, 98 (2015), pp. 523–537. Wiley Online Library.

[39] J. J. Hopfield, *Neural networks and physical systems with emergent collective computational abilities*, Proceedings of the national academy of sciences, 79 (1982), pp. 2554–2558. National Acad Sciences.

[40] S. Huang, *Non-genetic heterogeneity of cells in development: more than just noise*, Development, 136 (2009), pp. 3853–3862. Oxford University Press for The Company of Biologists Limited.

[41] L. Hubert and P. Arabie, *Comparing partitions*, Journal of classification, 2 (1985), pp. 193–218. Springer.

[42] B. Hwang, J. H. Lee, and D. Bang, *Single-cell rna sequencing technologies and bioinformatics pipelines*, Experimental & molecular medicine, 50 (2018), p. 96. Nature Publishing Group.

[43] M. Jaderberg, W. M. Czarnecki, I. Dunning, L. Marris, G. Lever, A. G. Castaneda, C. Beattie, N. C. Rabinowitz, A. S. Morcos, A. Ruderman,

et al., *Human-level performance in first-person multiplayer games with population-based deep reinforcement learning*, *arXiv preprint arXiv:1807.01281.

[44] D. A. Jaitin, E. Kenigsberg, H. Keren-Shaul, N. Elefant, F. Paul, I. Zaretsky, A. Mildner, N. Cohen, S. Jung, A. Tanay, et al., *Massively parallel single-cell rna-seq for marker-free decomposition of tissues into cell types*, Science, 343 (2014), pp. 776–779. American Association for the Advancement of Science.

[45] Z. Ji and H. Ji, *Tscan: Pseudo-time reconstruction and evaluation in single-cell rna-seq analysis*, Nucleic acids research, 44 (2016), pp. e117–e117. Oxford University Press.

[46] A. Kaplan, D. Nordman, and S. Vardeman, *On the instability and degeneracy of deep learning models*, *arXiv preprint arXiv:1612.01159, (2016).

[47] L. M. Kelly, U. Englmeier, I. Lafon, M. H. Sieweke, and T. Graf, *Mafb is an inducer of monocytic differentiation*, The EMBO journal, 19 (2000), pp. 1987–1997. John Wiley & Sons, Ltd.

[48] G. Khandekar, S. Kim, and P. Jagadeeswaran, *Zebrafish thrombocytes: functions and origins*, Advances in hematology, 2012 (2012). Hindawi.

[49] D. P. Kingma and J. Ba, *Adam: A method for stochastic optimization*, *arXiv preprint arXiv:1412.6980, (2014).

[50] D. P. Kingma and M. Welling, *Auto-encoding variational bayes*, *arXiv preprint arXiv:1312.6114, (2013).

[51] V. Y. Kiselev, K. Kirschner, M. T. Schaub, T. Andrews, A. Yiu, T. Chandra, K. N. Natarajan, W. Reik, M. Barahona, A. R. Green, et al., *Sc3: consensus clustering of single-cell rna-seq data*, Nature methods, 14 (2017), p. 483. Nature Publishing Group.

[52] H. A. Kramers, *Brownian motion in a field of force and the diffusion model of chemical reactions*, Physica, 7 (1940), pp. 284–304. Elsevier.

[53] A. Lancichinetti and S. Fortunato, *Erratum: Community detection algorithms: A comparative analysis [phys. rev. e 80, 056117 (2009)]*, Physical Review E, 89 (2014), p. 049902. APS.

[54] A. Y. Leung, J. C. Leung, L. Y. Chan, E. S. Ma, T. T. Kwan, K. Lai, A. Meng, and R. Liang, *Proliferating cell nuclear antigen (pcna) as a proliferative*

marker during embryonic and adult zebrafish hematopoiesis, Histochemistry and cell biology, 124 (2005), pp. 105–111. Springer.

[55] J. Lever, M. Krzywinski, and N. Altman, *Points of significance: Principal component analysis*, 2017. Nature Publishing Group.

[56] L. Li and H. Clevers, *Coexistence of quiescent and active adult stem cells in mammals*, science, 327 (2010), pp. 542–545. American Association for the Advancement of Science.

[57] Y. Li, K. Swersky, and R. Zemel, *Generative moment matching networks*, International Conference on Machine Learning, 2015.

[58] M. D. Luecken and F. J. Theis, *Current best practices in single-cell rna-seq analysis: a tutorial*, Molecular systems biology, 15 (2019). John Wiley & Sons, Ltd.

[59] H. Maamar, A. Raj, and D. Dubnau, *Noise in gene expression determines cell fate in bacillus subtilis*, Science, 317 (2007), pp. 526–529. American Association for the Advancement of Science.

[60] L. v. d. Maaten and G. Hinton, *Visualizing data using t-sne*, Journal of machine learning research, 9 (2008), pp. 2579–2605.

[61] P. M. Magwene, P. Lizardi, and J. Kim, *Reconstructing the temporal ordering of biological samples using microarray data*, Bioinformatics, 19 (2003), pp. 842–850. Oxford University Press.

[62] R. Maier, R. Zimmer, and R. Küffner, *A turing test for artificial expression data*, Bioinformatics, 29 (2013), pp. 2603–2609. Oxford University Press.

[63] D. Marbach, J. C. Costello, R. Küffner, N. M. Vega, R. J. Prill, D. M. Camacho, K. R. Allison, A. Aderhold, R. Bonneau, Y. Chen, et al., *Wisdom of crowds for robust gene network inference*, Nature methods, 9 (2012), p. 796. Nature Publishing Group.

[64] J. Masci, U. Meier, D. Cireşan, and J. Schmidhuber, *Stacked convolutional auto-encoders for hierarchical feature extraction*, Springer, International Conference on Artificial Neural Networks, 2011.

[65] F. E. Moore, E. G. Garcia, R. Lobbardi, E. Jain, Q. Tang, J. C. Moore, M. Cortes, A. Molodtsov, M. Kasheta, C. C. Luo, et al., *Single-cell transcriptional analysis of normal, aberrant, and malignant hematopoiesis in zebrafish*, Journal of Experimental Medicine, 213 (2016), pp. 979–992. Rockefeller University Press.

[66] G. E. Moore, *Cramming more components onto integrated circuits, reprinted from electronics, volume 38, number 8, april 19, 1965, pp. 114 ff.*, IEEE solid-state circuits society newsletter, 11 (2006), pp. 33–35. IEEE.

[67] F. Notta, S. Zandi, N. Takayama, S. Dobson, O. I. Gan, G. Wilson, K. B. Kaufmann, J. McLeod, E. Laurenti, C. F. Dunant, et al., *Distinct routes of lineage development reshape the human blood hierarchy across ontogeny*, Science, 351 (2016), p. aab2116. American Association for the Advancement of Science.

[68] A. E. Orhan and X. Pitkow, *Skip connections eliminate singularities*, *arXiv preprint arXiv:1701.09175, (2017).

[69] S. H. Orkin and L. I. Zon, *Hematopoiesis: an evolving paradigm for stem cell biology*, Cell, 132 (2008), pp. 631–644. Elsevier.

[70] E. J. Paik and L. I. Zon, *Hematopoietic development in the zebrafish*, International Journal of Developmental Biology, 54 (2010), pp. 1127–1137. UPV/EHU Press.

[71] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, *Automatic differentiation in pytorch*, NIPS 2017 Workshop Autodiff.

[72] P. Patil, T. Uechi, and N. Kenmochi, *Incomplete splicing of neutrophil-specific genes affects neutrophil development in a zebrafish model of poikiloderma with neutropenia*, RNA biology, 12 (2015), pp. 426–434. Taylor & Francis.

[73] F. Paul, Y. Arkin, A. Giladi, D. A. Jaitin, E. Kenigsberg, H. Keren-Shaul, D. Winter, D. Lara-Astiaso, M. Gury, A. Weiner, et al., *Transcriptional heterogeneity and lineage commitment in myeloid progenitors*, Cell, 163 (2015), pp. 1663–1677. Elsevier.

[74] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, *Scikit-learn: Machine learning in Python*, Journal of Machine Learning Research, 12 (2011), pp. 2825–2830. Packt.

[75] J. Pennington, S. Schoenholz, and S. Ganguli, *Resurrecting the sigmoid in deep learning through dynamical isometry: theory and practice*, Advances in neural information processing systems, 2017.

[76] W. Pimtong, M. Datta, A. M. Ulrich, and J. Rhodes, *Drl. 3 governs primitive hematopoiesis in zebrafish*, Scientific reports, 4 (2014), p. 5791. Nature Publishing Group.

[77] B. Psaila, N. Barkas, D. Iskander, A. Roy, S. Anderson, N. Ashley, V. S. Caputo, J. Lichtenberg, S. Loaiza, D. M. Bodine, et al., *Single-cell profiling of human megakaryocyte-erythroid progenitors identifies distinct megakaryocyte and erythroid differentiation pathways*, Genome biology, 17 (2016), p. 83. BioMed Central.

[78] X. Qiu, A. Hill, J. Packer, D. Lin, Y.-A. Ma, and C. Trapnell, *Single-cell mrna quantification and differential analysis with census*, Nature methods, 14 (2017), p. 309. Nature Publishing Group.

[79] X. Qiu, Q. Mao, Y. Tang, L. Wang, R. Chawla, H. A. Pliner, and C. Trapnell, *Reversed graph embedding resolves complex single-cell trajectories*, Nature methods, 14 (2017), p. 979. Nature Publishing Group.

[80] S. Rifai, G. Mesnil, P. Vincent, X. Muller, Y. Bengio, Y. Dauphin, and X. Glorot, *Higher order contractive auto-encoder*, Springer, Joint European Conference on Machine Learning and Knowledge Discovery in Databases, 2011.

[81] P. J. Rousseeuw, *Silhouettes: a graphical aid to the interpretation and validation of cluster analysis*, Journal of computational and applied mathematics, 20 (1987), pp. 53–65. Elsevier.

[82] S. T. Roweis and L. K. Saul, *Nonlinear dimensionality reduction by locally linear embedding*, science, 290 (2000), pp. 2323–2326. American Association for the Advancement of Science.

[83] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, *Learning internal representations by error propagation*, tech. rep., California Univ San Diego La Jolla Inst for Cognitive Science, 1985.

[84] A. M. Saxe, Y. Bansal, J. Dapello, M. Advani, A. Kolchinsky, B. D. Tracey, and D. D. Cox, *On the information bottleneck theory of deep learning*, Journal of Statistical Mechanics: Theory and Experiment, 2019 (2019), p. 124020. IOP Publishing.

[85] S. M. Shaffer, M. C. Dunagin, S. R. Torborg, E. A. Torre, B. Emert, C. Krepler, M. Beqiri, K. Sproesser, P. A. Brafford, M. Xiao, et al., *Rare cell variability and drug-induced reprogramming as a mode of cancer drug resistance*, Nature, 546 (2017), p. 431. Nature Publishing Group.

[86] A. K. Shalek, R. Satija, X. Adiconis, R. S. Gertner, J. T. Gaublomme, R. Raychowdhury, S. Schwartz, N. Yosef, C. Malboeuf, D. Lu, et al., *Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells*, Nature, 498 (2013), p. 236. Nature Publishing Group.

[87] A. K. Shalek, R. Satija, J. Shuga, J. J. Trombetta, D. Gennert, D. Lu, P. Chen, R. S. Gertner, J. T. Gaublomme, N. Yosef, et al., *Single-cell rna-seq reveals dynamic paracrine control of cellular variation*, Nature, 510 (2014), p. 363. Nature Publishing Group.

[88] C. E. Shannon, *A mathematical theory of communication*, Bell system technical journal, 27 (1948), pp. 379–423. Wiley Online Library.

[89] R. Shwartz-Ziv and N. Tishby, *Opening the black box of deep neural networks via information*, *arXiv preprint arXiv:1703.00810, (2017).

[90] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, et al., *Mastering the game of Go with deep neural networks and tree search*, vol. 529, Nature Publishing Group, 2016.

[91] D. Silver, T. Hubert, J. Schrittwieser, I. Antonoglou, M. Lai, A. Guez, M. Lanctot, L. Sifre, D. Kumaran, T. Graepel, et al., *Mastering chess and shogi by self-play with a general reinforcement learning algorithm*, *arXiv preprint arXiv:1712.01815, (2017).

[92] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, et al., *Mastering the game of go without human knowledge*, Nature, 550 (2017), p. 354. Nature Publishing Group.

[93] N. Simidjievski, C. Bodnar, I. Tariq, P. Scherer, H. A. Terre, Z. Shams, M. Jamnik, and P. Liò, *Variational autoencoders for cancer data integration: design principles and computational practice*, Frontiers in Genetics, 10 (2019). Frontiers Media SA.

[94] R. R. Sokal and F. J. Rohlf, *The comparison of dendrograms by objective methods*, Taxon, 11 (1962), pp. 33–40. JSTOR.

[95] C. K. Sønderby, T. Raiko, L. Maaløe, S. K. Sønderby, and O. Winther, *Ladder variational autoencoders*, Advances in neural information processing systems, 2016.

[96] D. P. Stites, A. I. Terr, T. G. Parslow, and M. PhD, *Medical immunology*, Appleton & Lange Stamford, CT, 1997.

[97] J. Tan, J. H. Hammond, D. A. Hogan, and C. S. Greene, *Adage-based integration of publicly available pseudomonas aeruginosa gene expression data with*

denoising autoencoders illuminates microbe-host interactions, MSystems, 1 (2016), pp. 25–15. Am Soc Microbiol.

[98] F. Tang, C. Barbacioru, Y. Wang, E. Nordman, C. Lee, N. Xu, X. Wang, J. Bodeau, B. B. Tuch, A. Siddiqui, et al., *mrna-seq whole-transcriptome analysis of a single cell*, Nature methods, 6 (2009), p. 377. Nature Publishing Group.

[99] J. B. Tenenbaum, V. De Silva, and J. C. Langford, *A global geometric framework for nonlinear dimensionality reduction*, Science, 290 (2000), pp. 2319–2323. American Association for the Advancement of Science.

[100] N. Tishby, F. C. Pereira, and W. Bialek, *The information bottleneck method*, *arXiv preprint physics/0004057, (2000).

[101] N. Tishby and N. Zaslavsky, *Deep learning and the information bottleneck principle*, IEEE, 2015 IEEE Information Theory Workshop (ITW), 2015.

[102] M. T. N. Tran, M. Hamada, H. Jeon, R. Shiraishi, K. Asano, M. Hattori, M. Nakamura, Y. Imamura, Y. Tsunakawa, R. Fujii, et al., *Mafb is a critical regulator of complement component c1q*, Nature communications, 8 (2017), p. 1700. Nature Publishing Group.

[103] C. Trapnell, D. Cacchiarelli, J. Grimsby, P. Pokharel, S. Li, M. Morse, N. J. Lennon, K. J. Livak, T. S. Mikkelsen, and J. L. Rinn, *The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells*, Nature biotechnology, 32 (2014), p. 381. Nature Publishing Group.

[104] L. J. van der Maaten and G. E. Hinton, *Visualizing high-dimensional data using t-sne*, Journal of machine learning research, 9 (2008), pp. 2579–2605. Microtome Publishing.

[105] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol, *Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion*, Journal of machine learning research, 11 (2010), pp. 3371–3408.

[106] J. D. Welch, A. J. Hartemink, and J. F. Prins, *Slicer: inferring branched, nonlinear cellular trajectories from single cell rna-seq data*, Genome biology, 17 (2016), p. 106. BioMed Central.

[107] A. Wilson, E. Laurenti, G. Oser, R. C. van der Wath, W. Blanco-Bose, M. Jaworski, S. Offner, C. F. Dunant, L. Eshkind, E. Bockamp, et al., *Hematopoietic stem cells reversibly switch from dormancy to self-renewal during homeostasis and repair*, Cell, 135 (2008), pp. 1118–1129. Elsevier.

[108] A. Wilson, G. M. Oser, M. Jaworski, W. E. BLANCO-BOSE, E. Laurenti, C. Adolphe, M. A. Essers, H. R. MacDonald, and A. Trumpp, *Dormant and self-renewing hematopoietic stem cells and their niches*, Annals of the New York Academy of Sciences, 1106 (2007), pp. 64–75. Wiley Online Library.

[109] N. K. Wilson, D. G. Kent, F. Buettner, M. Shehata, I. C. Macaulay, F. J. Calero-Nieto, M. S. Castillo, C. A. Oedekoven, E. Diamanti, R. Schulte, et al., *Combined single-cell functional and gene expression analysis resolves heterogeneity within stem cell populations*, Cell stem cell, 16 (2015), pp. 712–724. Elsevier.

[110] J. Xie, S. Kelley, and B. K. Szymanski, *Overlapping community detection in networks: The state-of-the-art and comparative study*, Acm computing surveys (csur), 45 (2013), p. 43. ACM.

[111] C. Xu and Z. Su, *Identification of cell types from single-cell transcriptomes using a novel clustering method*, Bioinformatics, 31 (2015), pp. 1974–1980. Oxford University Press.

[112] R. Yamamoto, Y. Morita, J. Ooehara, S. Hamanaka, M. Onodera, K. L. Rudolph, H. Ema, and H. Nakauchi, *Clonal analysis unveils self-renewing lineage-restricted progenitors generated directly from hematopoietic stem cells*, Cell, 154 (2013), pp. 1112–1126. Elsevier.

[113] C. Yau et al., *pcareduce: hierarchical clustering of single cell transcriptional profiles*, BMC bioinformatics, 17 (2016), p. 140. BioMed Central.

[114] A. P. A. R. M. T. Yury Gabuev, Kirill Mazur, *information bottleneck pytorch*. url: `https://github.com/makezur/information_bottleneck_pytorch`, Oct. 2018. GitHub.

[115] C. Zhang, X. Cheng, J. Liu, J. He, and G. Liu, *Deep sparse autoencoder for feature extraction and diagnosis of locomotive adhesion status*, Journal of Control Science and Engineering, 2018 (2018). Hindawi.

[116] Y. Zhang, S. Gao, J. Xia, and F. Liu, *Hematopoietic hierarchy–an updated roadmap*, Trends in cell biology, 28 (2018), pp. 976–986. Elsevier.

[117] S. Zhao, W.-P. Fung-Leung, A. Bittner, K. Ngo, and X. Liu, *Comparison of rna-seq and microarray in transcriptome profiling of activated t cells*, PloS one, 9 (2014). Public Library of Science.

[118] S. Zhao, J. Song, and S. Ermon, *Infovae: Information maximizing variational autoencoders*, *arXiv preprint arXiv:1706.02262, (2017).

[119] H. Zheng, J. Yao, Y. Zhang, and I. W. Tsang, *Degeneration in vae: in the light of fisher information loss*, *arXiv preprint arXiv:1802.06677, (2018).

[120] B. D. Ziebart, A. Maas, J. A. Bagnell, and A. K. Dey, *Maximum entropy inverse reinforcement learning*, (2008). figshare.