Spectral and deep learning approaches to Hi-C data analysis

SISSA
Scuola Internazionale Superiore di Studi Avanzati

Physics Area - PhD course in
Physics and Chemistry of Biological systems

Spectral and deep learning approaches to Hi-C data analysis

Candidate:
Stefano Franzini

Advisor:
Cristian Micheletti

Academic Year 2020-21

Contents

Introduction

The discovery of chromatin in the nucleus of cells in 1879 by Walther Flemming sparked two centuries of scientific interest in nuclear organization [1]. The chromatin fiber, composed by DNA wound around protein complexes called histones, was soon linked to genetic heredity and evolution: even before the discovery of the double helix structure of DNA by Watson and Crick, August Weismann was able to see the connection between nuclei replication, in particular meiosis, and traits inheritance. However, the sequence of nucleotides in the DNA of an organism is not the sole determinant of cell type.

In fact all cells in one organism are able to specialize in a variety of very different tissues, while all sharing the same genome. The information allowing this striking differentiation cannot be written in the shared sequence of DNA nucleotides: it must then be encoded in some epigenetic trait, shared among cells with the same task [2].

Recent research has brought to light the connection between epigenetic markers, molecules which regulate gene-expression and cell fate, and genome folding[2]: while DNA packing is a necessity to allow a two meters long genome to fit inside of a nucleus 10 μm wide [3], the three dimensional spatial organization of chromosomes has a much larger role to play in determining cell types.

Chromosome conformation capture (3C) experimental techniques have been determinant in shedding light on this aspect of genome folding [4]. These experiments probe interactions between intra- and inter-chromosomal regions of the genome: in particular, Hi-C experiments [5], one of the latest incarnations of 3C techniques, have allowed for genome-wide sampling and can be regarded as fingerprints of the three-dimensional configurations of chromosomes.

They have revealed a hierarchy of structural patterns unique to each cell-type, such as point-like contacts corresponding to loops [2]; strongly interacting blocks of loci (TADs), around 1 million base pairs in size (1 Mb) [2]; compartments which span the whole chromosome (typically around 100 Mb long in mammalians) and show the tendency of chromatin to separate into different phases [2]. All of these patterns have been found to be correlated to the presence of epigenetic markers [6, 7, 8], and have been experimentally shown to influence the development of illnesses and malformations [9, 10].

Interestingly, however, averaging through different chromosomes also reveals the presence of common, or aspecific, interaction patterns which can be ascribed to polymer physics:

for instance the decay of interaction counts according to the genomic distance between loci is a shared attribute of different cell-types [2].

The abundance of Hi-C experiments has allowed the focus to widen from the analysis of single maps to that of pairs of experiments in large datasets [11, 12, 13]. In particular efforts have been made to obtain quality metrics that guarantee the compatibility of different experiments on the same cell-type, and to quantify the genuine differences occurring between different lines.

This analysis is important for a number of reasons: to verify the reproducibility of interaction patterns among biological replicates of one cell-type; to compile compatible experiments into one and obtain maps of better quality; to locate significant differences which suggest biological motivations [12, 13].

In order to do this, however, one must beware the peculiarities of these interaction maps: biases such as the distance dependence of interaction counts [14], as well as the presence of noise [11, 12, 13], must be carefully taken into account for any analysis to be successful.

An under-utilized tool is that of spectral analysis: since the inception of Hi-C, results have shown that the spectrum of these matrices is rich of biological and structural information [15, 7]. For example the first principal component of properly normalized matrices has been interpreted as the mark of genomic compartments, and linked to chromatin types characterized by measurable epigenetic markers [15]. More recently, spectral properties have been employed in reproducibility analysis to compare different matrices [12, 13]. However, these promising studies were limited only to a few eigenspaces: the rest of the spectrum of Hi-C matrices remains a mostly unknown quantity.

The first part of this work ventures in these uncharted territories to obtain a more systematic view of the spectral properties of Hi-C maps and how to employ them.

In chapter 1 I will provide a brief introduction to chromosome conformation capture (3C) techniques [4, 16, 5], from the original method to the more advanced ones, focusing especially on Hi-C experiments [5]. I will discuss in detail some of the patterns of interaction found in Hi-C maps, such as TADs and compartments, and introduce the standard techniques used to detect or define them [2]. At the end of the chapter I will introduce the problem of Hi-C maps comparisons and how spectral methods have been previously used to tackle this kind analyses [12, 13].

In chapter 2 I compare the spectral properties of Hi-C maps to those of random matrices. I show that a large part of the spectra of (properly normalized) Hi-C maps closely resembles that of random matrices: these eigenspaces can be considered aspecific, in the sense that they provide a common background which has the same universal properties in all Hi-C matrices. Only the top eigenspaces, ordered according to the absolute values of the associated eigenvalue, display significant discrepancies with respect to the null model provided by random matrices. I dub this part of the spectrum "essential" in reference to the concept of essential spaces in the analysis of elastic networks [17]: as suggested by previous studies [15, 7], these should contain the majority of the structural and biologi-

cal information encoded in Hi-C maps. Can they be used to enhance these signals? Are comparisons between essential components easier to interpret?

In order to answer these questions, Chapter 3 delves deeper in the analysis of the essential component of Hi-C maps: by truncating the spectral sum over the eigenspaces, one can obtain an essential matrix which does not contain aspecific interactions. I show that this procedure leads to sharper interaction patterns which correlate better with those found in non processed high quality experiments. Specifically, I find that TADs are better reproduced among biological replicates. Comparing the essential components of a large number of experiments also allows to discriminate between cell-types, with results on this task which are competitive with other published methods [12, 13].

At the end of the chapter I also adapt essential component analysis to single cell Hi-C maps, each accounting for the genome structure of a single cell [18], instead of ensemble averages, observing an improvement with respect to the baseline on classification tasks: however I do not reach the same quality as for bulk matrices, showing that the challenges offered by single cell Hi-C maps must be tackled with different strategies.

These results show that it is possible to encode the information contained in Hi-C matrices in a small number of degrees of freedom, not only making them more manageable in computational applications, but also enhancing structural and biological features that are difficult to capture in high resolution maps. This may be of particular importance for single cell Hi-C maps, since, because of their sparseness, they may benefit more from a dimensional reduction scheme, both in terms of compression and analysis improvement.

Motivated by this observation, in chapter 4 I look for a dimensional reduction algorithm specifically designed to compress Hi-C data. The variational autoencoder is a neural network architecture that provides such a tool [19]: the first half of the network, the encoder, operates by compressing input data to a lower dimensional vector in a regularized latent space, while a decoder restores the original data in the second half. By training the network to match the output to input data, one can teach autoencoders the most parsimonious low-dimensional representation of the dataset [19].

Local interaction patterns (spanning only a small portion of the whole matrix) are shared across different cell-types and chromosomes [20]: one can then use the autoencoder to learn lower-dimensional representations of these patterns and obtain compressed versions of Hi-C maps. I compute the intrinsic dimension [21] for square 50×50 cut-outs sampled from Hi-C matrices, and show that one can obtain a 25-fold compression of Hi-C maps with reconstruction errors significantly smaller with respect to linear methods such as PCA. This could open the door to analyzing large number of HiC maps at a very high resolution with reasonable memory usage. Moreover, by giving a more parsimonious description of Hi-C data, this dimensional reduction scheme is also able to enhance the information present in latent space matrices and obtain competitive results in a number of tasks, such as classification and TAD calling.

I also apply the autoencoder to single cell Hi-C maps, and find that in this case compression is not able to improve analyses, suggesting that more advanced algorithms or

expert knowledge is needed when dealing with these kind of data.

Finally, at the end of the chapter, I compare the action of the autoencoder on the spectral properties of Hi-C matrices. Interestingly I discover that widely different techniques act in a similar manner on the spectrum of Hi-C maps, by enhancing top eigenspaces with respect to the rest.

The first part of the work (chapters 2 and 3) is mostly based on the following published paper:

- S. Franzini, M. di Stefano, C. Micheletti
  "essHi-C: essential component analysis of Hi-C matrices"
  Bioinformatics **37**, Pages 2088–2094, 2021

Chapter 1

Chromosome Conformation Capture: an overview

The problem of genome folding and organization has captured the interest of researchers for a long time. After all, the details of this subject are quite puzzling. The human genome contains more than 3 billion base-pairs divided into 23 pairs of chromosomes: if these were to be aligned end to end, they would span roughly 2 meters, more than the height of an adult man. Moreover the DNA double helix is quite rigid because of electromagnetic interactions, which makes folding even more difficult. Yet each cell of our body is able to contain these long polymers inside its nucleus, an organelle which has a diameter of roughly 10 μm [2]. It is only natural to wonder at such a feat of dense packing.

While it was always clear that the genome is not a simple one-dimensional rigid polymer, but rather folds into a complex three dimensional structure, the past has seen a long debate about the specificity of this organization in the interphase, i.e. the portion of the cell cycle that occurs between consecutive cell divisions.

Some experiments can show chromatin, the fiber made up of DNA and histones, as highly structured [22]: supercoiling forms thicker fibers, which then loop and fold into a hierarchy of ever more complex structures. But other techniques detect very little structure in what can be seen as a soup of disorganized nucleosomes [23]. Fluorescence in situ hybridization experiments, which are able to locate specific loci in single-cell assays, reveal enormous variability between each sample [24]: the distance between any pair of loci, as well as their sub-nuclear position, can change enormously cell to cell.

However these experiments also point to specific trends in organization. Chromosomes occupy well defined territories [25], with limited intermingling in their strands at their borders [26]. While their specific position in the nucleus can have a large degree of cell-to-cell variability, many chromosomes prefer to shift either towards the periphery of the nucleus or towards its center [27, 28, 25]. Additionally, chromatin displays the typical behavior of a copolymer, i.e. a polymer whose components have different properties, with alternating

sequences of euchromatin and heterochromatin which interact preferentially with sequences of the same variety, which leads to the formation of phase separated compartments [2, 29]. Finally, chromosomes also undergo a series of dramatic conformational changes as the cell goes through its life cycle: as the moment of mitosis approaches, chromatin is wound up into the classic rod-shaped structures observed in every biology textbook, which have to unfold again after cell replication.

All of these observations give a vision of a nuclear organization that, while stochastic on the one hand, is guided by some mechanisms which also relate to gene expression and suppression on the other.

Enormous steps forward have been made in our comprehension of chromosomes spatial organization in the last decades, and how their folding helps determine many of their functions, such as gene regulation, or DNA maintenance and replication. This has been in part possible thanks to the development and improvement of imaging techniques which work at the single cell level, but also due to the birth of chromosome conformation capture (3C) technology [4]: this chapter will outline the basics of how these experiments work, how they were refined since their original creation, and what discoveries about the spatial organization of chromosomes they allowed. Afterwards, the chapter will present some of the open questions in the field and the state of the art methods which were devised in order to tackle them.

1.1 A brief history of the genome

Before delving into 3C techniques, it is interesting to briefly review the history of the study of the genome, from its discovery to the present.

In 1879 Walter Flemming was able to observe filamentous structures inside the cell nucleus by using aniline dyes, and dubbed these fibers chromatin and connected it to heredity. Since then scientist have studied its properties, both physical and chemical, in the pursuit of a better understanding of the genome inner workings.

During the last two centuries many discoveries contributed to create a fascinating and complex picture of the machines that are our cells, and how they work: histones were discovered in 1884, not much later than chromatin itself; in 1924 Emil Heitz coined the terms euchromatin and heterochromatin, to refer to structural differences within the fiber; Conrad Waddington proposed epigenetic landscapes in 1942.

Of course one of the most momentous findings in the field was the discovery of the double helix structure in 1953 by Watson and Crick, which allowed a more in depth description of DNA. Nevertheless this was not to be the end-all be-all of structural research on the genome, not by a long shot: rather, it was the stepping stone towards answering even more difficult questions.

Another fundamental step in understanding the genome structure was the discovery, in 1984, of chromosome territories in many eukaryotes cells during their interphase (yeast *S.*

cerevisiae being an exception), thanks to fluorescence labeling techniques.

The motivation for such a relentless interest in the search of a comprehensive description of the structural and dynamical properties of chromosome is not only driven by scientific curiosity, but also by the urge of having a better understanding of gene regulation, which may be the door to many medical applications.

It is in this context that chromosome conformation capture (3C) techniques have been introduced by Job Dekker and coworkers in 2002, with the objective of studying the three-dimensional spatial organization of chromosomes.

1.2 Chromosome Conformation Capture

Chromosome conformation capture techniques were introduced by Dekker's lab in 2002 [4] to study chromosome structure by mapping in vivo interactions between chromatin loci, and has since achieved high resolution at genome-wide scale. The main idea behind the development of 3C methods is that a matrix containing many or all contact frequencies between loci along the chromosome would enable scientists to infer the three-dimensional organization of the chromosome.

3C is used to detect the ensemble frequency of interaction of any pair of genomic loci. Combined with deep DNA sequencing can generate genome-wide interaction frequency matrices, which have indeed been used as the starting point of many models of the spatial organization of the genome.

The 3C procedure is outlined by the following steps, illustrated in figure 1.1: first cells are fixed with formaldehyde to fix the spatial arrangement of chromosome in place. Restriction enzymes are then used to digest chromatin into fragments, while it is still frozen in place, so that spatially proximal segments stick together. DNA ends of these segments are then re-ligated into a loop, and the DNA is subsequently purified.

This assay produces a large set of unique DNA ligation products, each one of which represent a single spatial co-location event in one cell of the population ensemble. DNA molecules can be then easily identified by PCR or DNA sequencing, so 3C based techniques significantly lower the difficulty of determining relative spatial positions of loci inside cells, by tackling the simpler process of DNA sequence analysis instead.

Many of the molecular steps used by 3C techniques were previously developed in a different context, and they were used separately. For instance, proximity ligation had been used to detect bending and looping of DNA due to protein interaction, both in vivo and in vitro [30, 31, 32]. 3C main innovation lies in the fact that it was designed as an unbiased method, able to detect any spatial proximity irrespective of the mechanism that brought the strands of chromatin together.

PCR was initially used to read and count individual ligation products in order to ascertain whether different loci would interact with each other more frequently than expected. The original 3C paper [4] validated the approach by comparing its reads to the previ-

ously known interactions between specific loci (e.g. homologous chromosomes and yeast centromeres or telomeres); furthermore, it presented a (sparse) matrix of the interaction frequencies for a complete chromosome, which was used to infer the population average conformation of its folding.

1.3 3C evolution: towards genome-scale and high-resolution chromatin interaction matrices

The human genome is more than 3 billion base pairs in length: while other organism may have smaller chromosomes, the number of possible chromatin interactions makes using PCR to analyze 3C difficult[2]. This explains why, since its birth, many adaptations were introduced to allow mapping of chromatin loci interactions at the genome scale and higher resolutions, both in cell populations and single cell experiments.

All of these variants make use of the same basic steps of the 3C protocol: chromatin crosslinking, DNA digestion through the usage of enzymes, purification, and re-ligation of DNA strands ends in close spatial proximity. The differences arise in the method used for ligation product detection [33].

4C, also called 4c-seq, was published in 2006 [34, 6] and is the first 3C variant: it focuses on a single region of interest and uses inverse PCR to amplify all loci that interact with it. Deep sequencing analyses are then able to recover the one-vs-all interaction profile relative to that single genomic element of interest. While this method is less suited for chromosome folding inference, since it does not produce a complete matrix of the interaction frequencies, 4C profiles still provide a wide range of information about interesting aspects of genomic folding: in fact, they are primarily used to identify long-range DNA contacts, especially between regulatory DNA modules, such as the loops that can be formed by enhancers and promoters, or architectural chromatin loops between cohesin- and CTCF-associated domain boundaries. Moreover, 4C-seq contact profiles can help reveal the boundaries of contact domains and can identify the structural domains that co-occupy the same nuclear compartment.

5C (chromosome conformation capture carbon copy [16]), published on the same year as 4C, is used to detect all interactions frequencies between pairs of loci within a certain region, which can be up to several Mb in size, thus it can be thought of as a "many-to-many" approach. This method uses a ligation-mediated amplification followed by detection of the standard 3C library. While it follows the same usual 3C protocol steps, the 3C products are incubated with a complex mix of primers designed to anneal exactly at one of the restriction sites of the genomic region of interest. If the digestion of the DNA and its subsequent ligation work efficiently, the two primers end up facing each other at the ligation junction. They can then be ligated by using Taq DNA ligase. Finally, PCR is used to amplify the ligation product and obtain a map of the interaction frequencies in the

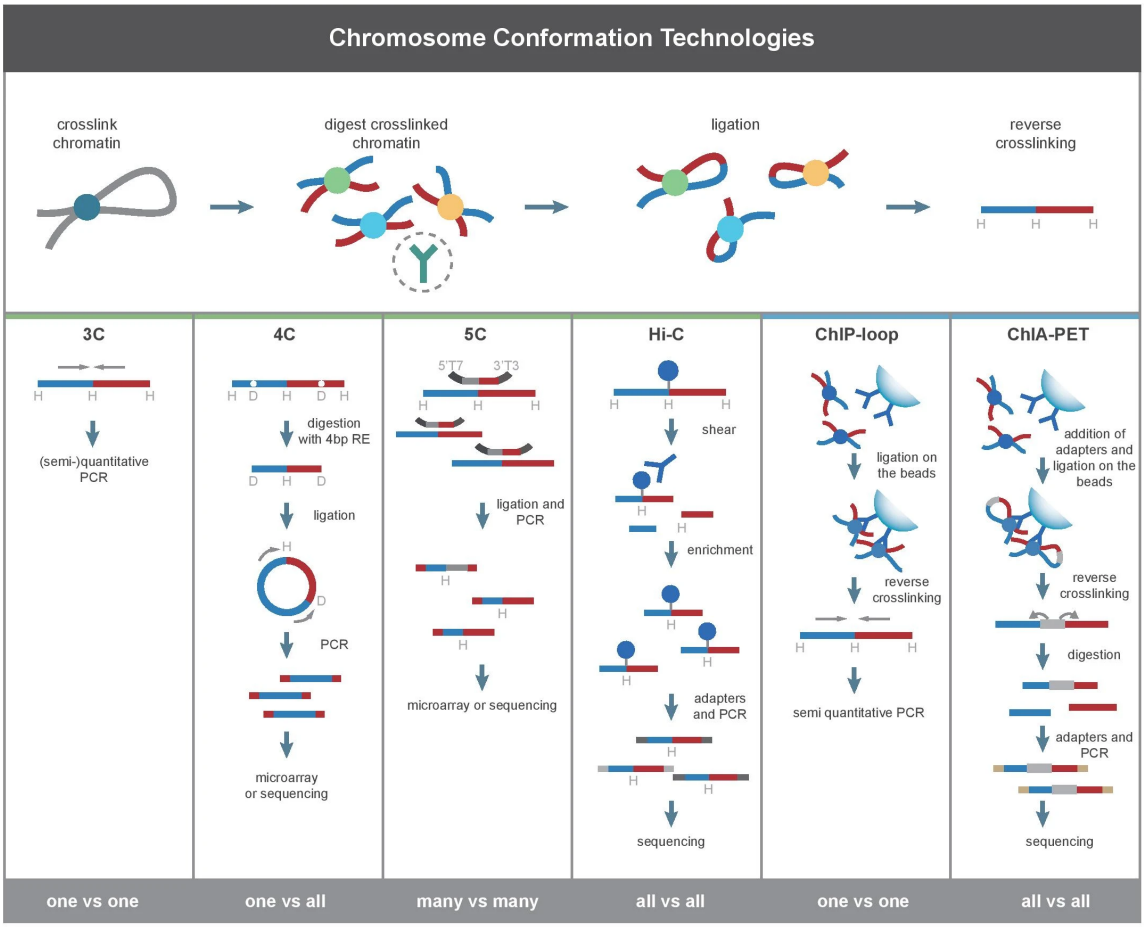

Figure 1.1: **reproduced from G. Li et Al.** The figure recapitulates the fundamental steps in 3C experiments: chromatin proximity ligation through formaldehyde, digestion of the crosslinked fragments through restriction enzymes, purification and re-ligation of the fragments. The steps following this basic outline depend on the kind of assay one is performing, as shown in the panels below.

region of interest. However, 5C has relatively low coverage (the number of samples collected in the experiment is low) and the approach is unable to be extended over a genome-wide sampling of the contacts, as that would require millions of 5C primers to be used.

The Hi-C variant, developed in 2009 by Job Dekker and Liebermann-Aiden [15], enables the unbiased detection of all pair-wise interactions in the genome: in this protocol the DNA ends, formed after chromatin digestion by restriction enzymes, are labeled with biotinylated nucleotides before being ligated again. After purification, this allows one to retrieve sequences that should represent two different restriction loci, that were ligated together based on their proximity. The pair of sequences can be individually aligned to the genome, so that one can determine the fragments that were in contact. Hence, all possible pair-wise interactions throughout the genome are tested [15].

Given the length of the human genome, the number of possible pair-wise interactions is extraordinarily large (counting $\sim 10^{14}$ possible interactions for the genome digested into 250 base pairs fragments[33]): this means that extremely deep sequencing is required to obtain well resolved maps. However even the most deeply sequenced datasets only contain enough interaction counts to be binned with a stride between 1-10 Kb [8, 10, 33]. As I will elaborate on in the following sections, this has given rise to a demand for methods that allow to validate and compile the reads of different experiments relative to the same cell strains, in order to achieve deeper sequencing.

Being able to map all-vs-all interactions, Hi-C maps have enabled researchers to uncover many of the folding principles of complete genomes, and I will discuss the hierarchy of patterns discovered in Hi-C maps in the next subsection. However Hi-C experiments are very costly to perform and, even the most deeply sequenced ones, do not capture the complete interaction landscape that arises from chromosome folding: a need is still present for targeted methods such as 4C or 5C, that efficiently reveal information about loci or regions of interest [33].

This is why further developments in the field has seen the emergence of newer targeted techniques, such as ChIA-PET [35] and HiChIP [36], which only focus on a specific subset of all the possible genomic loci, obtaining a much richer coverage of these regions.

Another branch of inquiry that has seen growing attention in recent years is that relative to single-cell genome conformation analysis [37, 38, 39]: there are now several methods that allow to sample configurations of the folded genome at the single-cell level, as opposed to population-wide assays, revealing a riveting variability in patterns of interactions [29]. This seem to point towards the presence of both dynamic and stochastic processes which may shape the folding conformations of the genome. Nevertheless, the interaction matrices obtained by these methods tend to be rather sparse, making analysis all the much harder.

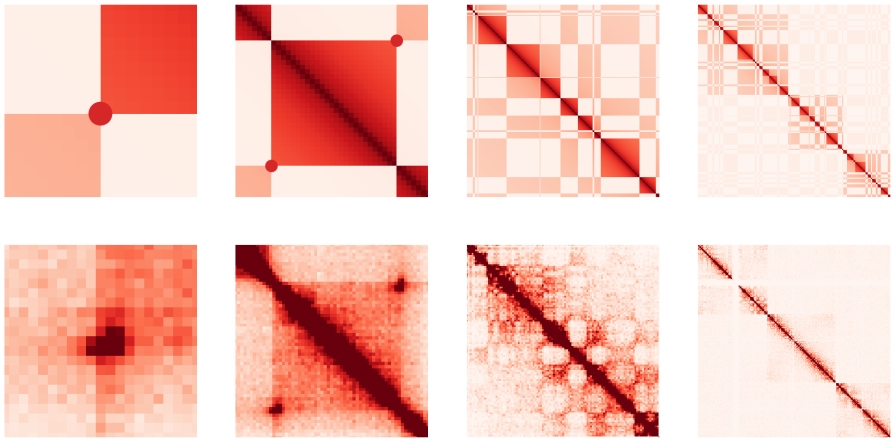

Figure 1.2: Some of the features that can be encountered in Hi-C maps. From left to right: Point-wise interactions, such as loops, appear as bright spots; TADs are blocks of strongly interacting loci located along the diagonal, a few hundreds of Kb in size; A/B compartment form a characteristic checkerboard pattern at the chromosome scale; the cis/trans ratio between average interaction frequencies within the same chromosome and different chromosomes, a signature of chromosomal territories. Moreover, one can observe the distance dependence of interaction frequencies, where the contact probability between loci within the same chromosome decreases with their genomic distance.

1.4 Interaction patterns in Hi-C matrices

After the collection of Hi-C interaction reads, the procedure usually involves mapping the data to a matrix, as well as filtering and bias correction steps [2]. After this is done, one obtains a genome-wide interaction matrix, where each bin contains the interaction frequency between two genomic loci. Extraction of biological relevant information from these matrices is complicated by a number of factors: the fraction of cells where each interaction occurs cannot be directly recovered from the maps; moreover one has to differentiate between the actual biological signal and noise, as well as identify and interpret the pattern emerging from the analyses.

The fact that matrices represent an ensemble average of the interactions occurring over a vast cell population is a critical complication in interpreting signals [2]: one cannot distinguish whether interactions involving multiple pairs of loci co-occur in each cell of the population or if they are mutually exclusive. In fact, if one were to directly translate interaction frequencies to distances, one would quickly discover that some pairs cannot be in contact at the same time. This also opens up the question of ergodicity: are the measured interactions frozen in place in each separate cell, or are they dynamic, cycling through different contact and open states? Even observing a smooth matrix, apparently devoid of location-specific interactions, is not a guarantee of the absence of structure in the underlying chromosome conformations. It only means that these structures, if they exist, may be specific only of a single cell, instead of the whole population: then averaging on the ensemble can hide their presence.

Secondly, current analyses do not rely on an explicit formal definition of what a specific interaction pattern looks like to search for it in the matrices; rather they define these patterns implicitly as the outputs of some method. The result is that comparing methods aimed at identifying the same type of interaction pattern is often difficult in the absence of a common gold standard for what the correct result would be.

Moreover, different interaction patterns overlap and coexist with each other, and it can be difficult, or impossible, to isolate the signature of each structure. In practice, one usually makes the simplifying assumption that patterns are independent, i.e. they can be disentangled either because the effect of other patterns on the structure of interest is negligible, or because they can be normalized out through some procedure.

Nevertheless, several different interaction patterns were observed in Hi-C matrices: they vary in scale, from looping interactions involving only a couple of loci at small genomic scales, to large-scale patterns spanning a whole chromosome, or even the whole genome. Some of these patterns can be recognized in a variety of different cell-types, and even across different organisms, while others are specific of each strain and condition. In the next subsections I will briefly introduce and describe some of the most relevant of these patterns.

1.4.1 Point interactions

At the smallest scale, point interactions (see figure 1.2) involve only a pair of strongly interacting loci, often separated by a large genomic distance (from a few Kb to several Mb) and hinting at a biological function: this is the case of enhancer-promoter interactions [40, 41, 42, 43, 44], which are localized genomic elements up to 1 Kb in size, which activate the expression of a gene by coming together in a looping event. It is worth to notice that loop formation is thought to be mediated by proteins, and that evidence points towards a dynamic, rather than equilibrium, mechanism of loop extrusion (although the debate is currently still open)[45].

Such point interactions are expected to appear on the Hi-C map as a local enrichment of contact probabilities [40, 41, 46]. Current approaches to find these patterns do not provide explicit models for what point interactions look like. Rather, they focus on locating outliers with higher interaction frequencies than expected for a background model which consists of overlaps of other larger scale interaction patterns. The chief among these is the distance-decay function which exponentially suppresses interactions between distal genomic elements, and arises from chain connectivity; however also other elements can be incorporated. Notice that without an explicit definition of what a point interaction looks like, it can be difficult to distinguish between *bona fide* signals and experimental noise. As such it is important to test also biological replicates of the interaction map for consistency, and to validate outliers through other independent methods informed by the knowledge of underlying biological processes.

Finally, one should beware that not every biologically relevant interaction needs to be an outlier: interactions between genomic loci close by along the chromatin fiber can have important regulatory and biological functions without being more likely than average for that distance; on the other hand, since interactions between distal elements are suppressed, even a small bump in the binned frequencies can suggest the presence of a biologically important contact event, which, however, only occurs in a few cells.

1.4.2 Topologically Associating Domains, TADs

At larger scales with respect to point interactions, Hi-C maps reveal the existence of sub-Mb structures that have been dubbed topologically associating domains or TADs (see figure 1.2 [47, 48, 49, 50]. These are contiguous regions where loci tend to display much higher interaction frequencies between each other than with loci outside the domain: they show up on the interaction matrix as darker blocks along the diagonal.

While the detection of such structures may seem simple through this intuitive definition, in practice TADs tend to be more complicated [51, 49, 48, 52, 53]. First, it easy to observe a hierarchy of nested TADs at different scales in most Hi-C matrices [54], making it difficult to actually define the boundaries of any one structure; moreover TADs may overlap, further complicating the task of detecting boundaries. As such, methods for seeking TADs, lacking

an explicit definition, tend to define them implicitly as their output.

One method I will use in the following chapters is the Insulation Score (IS) [55]: one can assign a score to each binned locus of the matrix by computing the average interactions occurring in the matrix in a 500 Kb $\times$ 500 Kb square centered on the chosen bin. The minima of the one-dimensional profile of the score across the matrix correspond to TAD boundaries, where the signal is small with respect to well-connected adjacent areas.

Nonetheless, given a definition of how to detect TADs, they have been shown to be associated with biological functions, although the connection is still not fully understood. TAD-like structures have been detected in multiple organisms, from mammals to bacteria [47, 49, 56, 50, 48], and their disruption has been linked to a wide range of diseases, among which cancer, a variety of limb malformations, and a number of brain disorders. One hypothesis describes TADs as micro-environment which lock enhancer-promoter interactions and prevent them from happening between different domains [29, 57, **?**]. In fact, some types of genes (such as transfer RNA genes and housekeeping genes) appear near TAD boundaries more often than would be expected by chance. As such, it seems that TADs are strongly involved in gene regulation. However some studies have been able to uncouple genome topology and gene expression, casting doubts about this aspect of TADs function [58].

The folding mechanism of these structures is another problem around which the debate is still going on [59, 56, 60, 45]. Their boundaries have been found to be enriched for CTCF and cohesin binding sites [48, 61], which may be a hint to understand the way they form. One mechanism that has been proposed is active loop extrusion by cohesin, and computer simulations have shown that transcription induced supercoiling causes loop growth with a reasonable speed and in the correct direction[62, 63]. However proof for DNA loop-extrusion is so far limited only to condensin (a protein complex similar to cohesin).

It is important to note that TADs have been found to be relatively conserved between different cell types, and even between different organisms, in some specific cases [64].

1.4.3 Genomic Compartments

Another interaction pattern that encompasses the whole chromosome is given by genomic compartments [15], which were first discovered in early Hi-C studies. Compartments show up on contact maps as a "checkerboard"-like motifs of alternating blocks of enriched and depleted interactions which span 1 to 10 Mb in size (see figure 1.2.

The explanation for the emergence of the checkerboard pattern is straightforward as it can be modeled by assuming the chromosome to contain two distinct genomic regions that alternate along the length of the chromatin fiber, as in block-copolymers: attractive interactions between regions of the same type and repulsive interactions between those of different types lead to phase separation of the genome, producing the characteristic signature pattern on the contact maps. The two regions are referred to as A and B compartments, although more thorough investigations have lead to the definition of at

least five sub-compartments, two active (A1, A2) and three inactive (B1, B2, B3).

Despite this definition being intuitive, the main method to recover compartments does not rely directly on it, but rather uses an indirect route by computing the first principal component of the correlation map computed from the interaction map[15, 2]: in order to do this one first assigns a Pearson correlation coefficient to each pair of loci by computing the correlations between the respective contact map columns; the matrix obtained through this method contains positive correlations between loci in the same compartment, and negative correlations between different compartments. By using principal component analysis to isolate the first principal component one is able to find the optimal grouping of the loci into A/B compartments according to the sign of the component. Alternatively, standard clustering procedures is also able to uncover the groupings.

Genomic compartments strongly correlate with a number of epigenetic markers and with chromatin state: DNA accessibility, gene density, GC content, and histone marks are all good indicators of the compartment type of a region. Euchromatin makes up A-type compartments and contains gene-dense, lightly packed regions of accessible DNA, while heterochromatin B-type compartments contain tightly packed inaccessible regions of the genome.

Compartments have been found to have high variability between different cell-types and biological conditions, correlating with large scale conformational and regulatory changes in the genome[15].

1.4.4 Distance-dependent Interaction Frequency

Contact frequencies inside chromosomes decrease, on average, with the genomic distance between the involved loci. This can be seen in the interaction matrix as a decreasing gradient of interaction frequencies the further one moves away from the diagonal.

Measurements can either rely on the discrete binning of the matrix, by taking the average interaction frequency at a fixed genomic distance, or by interpolation of a continuous function to the data. This usually reveals a dependency of the interaction frequencies from the distance which is compatible with a power-law decay, with exponent $alpha \simeq -2$ [2], so log-log plots are used to better highlight the dependency.

This peculiar pattern can be explained by the properties of chain connectivity [65, 66]: on average, loci which have a small contour distance along the chain will come into contact more frequently than those that are farther away because of the random fluctuations of the chromosome conformation. In fact various polymer models, such as the ideal chain model, can recover this interaction pattern without invoking specific loci-dependent interactions.

The distance-induced bias can disrupt some analyses, for example those used to determine chromosomal compartments, so in many cases it is useful to normalize the matrix in order to remove the distance-dependence. The product of this procedure is called an observed-over-expected matrix, as it identifies contacts that are enriched or depleted with respect to their background.

There are cases, however, in which the measured data deviate dramatically from the ones expected from simple polymer models, which may point towards interesting biological phenomena underlying chromosome folding. One such case is encountered when studying the evolution of interaction matrices throughout a cell life-cycle [67, 68, 69]: as the cell approaches mitosis, the chromosomes undergo striking structural changes in order to prepare for cell-division, taking on the iconic cylinder shape that can be found in any biology textbook. This is reflected in contact maps as a further depletion of contacts beyond a certain distance threshold ($\sim$ 10 Mb); moreover, recent experiments have pointed out the appearance of a secondary diagonal, parallel to the primary one, of enriched contacts, which can be interpreted as a complex chromosomal structure where dense chromatin loops surround a central helical scaffold [69]. In such cases more refined polymer models are needed to describe the chromosome structure, and analysis cannot make away with the distance-dependence of the interaction frequencies.

1.4.5 Cis/Trans Interaction Ratio

At the scale of the whole genome, the largest interaction pattern is given by the ratio between contact frequencies of loci within the same chromosome and between different chromosomes[15]. This ratio consistently shows, across different cell-types and species, that contacts occur preferentially between loci in the same chromosome (i.e. cis), rather than between different chromosomes (i.e. trans): the Cis/Trans ratio is interpreted as a signature of the presence of chromosome territories, meaning that, even during interphase, chromosomes are physically separated and occupy a distinct region of the nucleus, only intermingling at their interface[2].

Typical values for the cis/trans interaction ratio in the human genome range between 40 and 60 for high quality experiments. However, since noise can affect similarly both intra-chromosome and inter-chromosome interactions, and because of the universality of this ratio across different cell-types, a lower than expected ratio may indicate lower quality experiments [2].

Nevertheless there are situations in which measured trans-interaction frequencies between two regions of distinct chromosomes can become as large as those expected between regions of the same chromosome[2]: this is usually a signature of chromosomal translocation, a phenomenon often associated with cancer in which one observes unusual rearrangements of chromosomes, i.e. detached fragments of one chromosome are swapped with those of another, and vice-versa. Because of the way contact reads are sequenced, two regions that are physically part of the same chromosome, due to this rearrangement, can be attributed to different chromosome, explaining the presence of enriched trans-interaction frequencies and depleted cis-interaction frequencies in the Hi-C maps.

1.5 Hi-C maps comparisons

Over the past few years the growing quantity of Hi-C experiments has stratified into large datasets of contact maps available to be analyzed in search of significant biological meaning. Many methods have been developed to measure their properties at the single matrix level and locate the recurring interaction patterns detailed above: there are now multiple tools to discover loops, TADs, and compartments [70, 71, 72, 73, 74].

However, with the new abundance of data, part of the attention has shifted towards pair-wise comparison of matrices. The objective of such assays is twofold. First, it seeks to validate the quality of experiments coming from different sources. Second, comparative studies serve to establish which features of genome folding are varied across different cell-types, and which are conserved [13].

The first problem is one of reproducibility[75, 13, 12, 11]: when two experiments sampling the contact maps of biological replicates (i.e. cell populations coming from the same strain) are available, a common practice is to pool together their interaction counts in order to obtain a better sampled matrix for downstream analysis. However, this cannot be done without first introducing some quality measurement tool to investigate whether the two experiments are concordant or present significant differences, since the latter scenario would introduce undesirable biases in all subsequent analyses, tainting the results.

The other side of the coin are the occasions in which one wants to compare matrices obtained from different cell populations: in this scenario the question becomes whether known (or presumed) biological differences also have an effect on genome folding, and then to locate and quantify significant differences that may have biological explanations, important to understand the interplay between genomic structure and gene expression. In this case comparative tools can be used to establish configurational variations across different stages of cell development and cell fate, or due to differences in gene transcription, in disease-related phenotypic alterations, and in cancer.

Strategies previously devised to analyze 1D genomic assays, such as ChIP-seq, DNA methylation, and RNA sequencing, cannot be reliably applied out of the box to Hi-C matrices, because these experiments present new challenges [11, 76].

Hi-C maps encode the signature interaction patterns from many complex multi-scale 3D structures, such as TADs and A/B compartments, as well as others described in the previous section, so they present a rich and unique phenomenology that is qualitatively different from that of 1D assays. Moreover, the resolution of a contact map (i.e. the size of a bin in terms of base pairs) can be thought of as a free parameter, that can only be determined heuristically based on the sequencing depth of the experiment. Finally, the ligation noise of contact frequencies estimates is also linked to sequencing depth [13].

With so many peculiarities, it is no surprise that naively computing simple correlation measures, frequently used as a reproducibility score, fails to capture all the complexities of Hi-C data. This is partly due to the fact that these kind of measures consider each bin of the contact matrices on its own, disregarding the strong interdependence with contacts

formed with other loci. For this reason a number of more sophisticated reproducibility measures have been introduced in the past years: while there are many examples of such methods (such as HIC-rep[11]), the class I will focus on is that of spectral methods, which use spectral properties of Hi-C matrices (eigenvalues and eigenvectors) to compare them.

Two methods of this class have been developed to quantify the reproducibility of two Hi-C experiments: the first one, named HiC-Spector [12], computes the laplacians of intra-chromosomal matrices and compares the first 20 ordered eigenvectors in order to obtain a quality score; the second, GenomeDISCO [13], uses random walks on the graph defined by the interaction maps in order to smooth them, and then performs a bin by bin comparison of the smoothed matrices. While the latter may not use spectral properties directly, the smoothing procedure involves taking the n-th power of the transition matrix obtained from the Hi-C map, with n being the number of steps of the random walk, which increases the relative importance of the top eigenspaces with respect to the rest of the spectrum. A more detailed description of the two methods is offered in Appendix A.

My strategy is based on two main observations: one is that one must take into account the distance-decay when comparing matrices [11], which I will do through a normalization of the Hi-C maps; the second is that most eigenspaces of normalized matrices behave as the eigenspaces of random matrices. These form an aspecific background which can be disregarded in order to enhance the essential features of each Hi-C map and make them easier to sort according to their cell-type.

1.6 Summary and conclusion

A great deal of effort was poured into understanding the organization of the genome over the years, both because of the interest in the basic physical and biological principles driving it, and because of the practical medical implications which stem from this research. In the last two decades, 3C-based techniques provided a fundamental step forward in the study of chromosome folding and its relationship to gene regulation: the idea of sampling *in-vivo* contacts between genomic loci as a proxy of the 3D folding conformation has allowed, with the advent of the Hi-C method, the simultaneous detection of genome-wide interaction patterns, shedding light on a plethora of hierarchically organized structures which span all scales, from point-wise interactions, to TADs hundreds of Kb in length, to chromosome-wide compartments, and chromosome territories.

With this wealth of information, the real challenge has become building a solid understanding of the features appearing in Hi-C contact maps: a long list of computational tools has emerged to provide unbiasing normalization of binned contact frequencies, detect different kinds of patterns, and distinguish between statistically relevant interactions and noise. A particular problem that has become more and more relevant in the last few years, with the increase in the number of the Hi-C datasets, is the concern about the reproducibility of experiments: are the same interaction patterns shared between Hi-C matrices

obtained independently by experiments performed on cell-populations of the same strain? And on the flip-side of the coin: what are the statistically significant difference between different cell-types?

To answer these questions, naive methods, such as bin-by-bin comparisons, have proven fruitless[11]: the complexity of Hi-C maps and their inherent noisiness makes it all but impossible to perform statically significant analyses on the basis of just local information. Rather, specialized tools that have emerged to tackle the reproducibility problem, deal with these challenges by incorporating information about the surroundings of each locus involved in a given interaction, and do so by different routes[11, 12, 13, 77].

One class of methods analyzes the spectral properties of the interaction matrices[12, 13, 77]. By doing so, global properties of the maps and the hierarchy of features they contain emerge naturally as a consequence of the spectral decomposition, allowing one to easily isolate and compare statistically significant interactions. However, the two methods which have taken this route, while being very successful at their endeavor, do not propose a systematic treatment of the spectral properties of the matrix.

One can then ask whether there's an optimal way to isolate the essential spectral properties of Hi-C matrices using physical intuition. Developing such methods would help enhance the signal contained in Hi-C interaction maps without increasing the coverage through costly experiments, allowing for more reliable analyses even on low quality matrices.

As an answer to these needs, in the next chapters I will explore the connection between random matrix theory and the a-specific part of the spectrum of Hi-C maps, i.e. the featureless part of the spectrum whose properties are repeated irrespective of the cell-type; as a consequence I will propose, essHi-C, a method to enucleate the specific part of the spectrum and employ it to obtain more robust contact maps and a metric to compare them.

Chapter 2

Random Matrix Theory applied to Hi-C matrices

In this chapter, which will be mostly based on published work[77], I will use Random Matrix Theory (RMT) as a frame of reference to understand the spectral properties of Hi-C matrices, in a novel approach which allows one to separate noise-like features from signal.

Statistical mechanics forgoes the knowledge of deterministic trajectories in order to describe thermodynamic properties averaged over an ensemble of microstates, which are sampled from an underlying probability distribution. Random Matrix Theory can be thought of in the same terms: instead of describing the properties of a single specific matrix, RMT aims to study the ensemble properties of matrices whose elements are sampled from some distribution.

The context of the first application of this mathematical theory was nuclear physics, where Eugene Wigner introduced random matrices to model the nuclei of heavy atoms. There random matrices were used to replace the complicated Hamiltonian of the quantum system under study and calculate averages. It was the year 1957 and the success of Wigner's approach had just opened a new venue of analysis, which would prove extremely effective in tackling a wide range of problems.

In fact Random Matrix Theory was widely adopted in physics: it is used in quantum chromodynamics, in two dimensional quantum gravity, in the description of the fractional Hall effect, of quantum dots, of Anderson localization, and of superconductors. Of course, the theory has found applications outside of physics too: the distribution of the zeros of the Riemann Zeta function can be modeled by the eigenvalues of certain random matrices; random matrices have been used to describe computation errors in operations such as matrix multiplication; in neuroscience they can describe the network of synaptic connections between neurons in the brain.

However the potential of the application of RMT to Hi-C matrices has not been explored

yet. In this chapter I will compare the spectra of Hi-C matrices to random matrices and show the similarities they share.

The chapter is organized in the following manner: first I introduce the dataset of Hi-C matrices I will analyze, as well as their normalization. Next I will briefly run through some of the basic concepts and findings in RMT, which I will then use as a term of comparison for the spectral properties of Hi-C matrices. Based on the results, I will argue that the information content of Hi-C maps is mainly encoded in their top eigenspaces, while the rest of the spectrum describes aspecific noisy interactions

2.1 Hi-C dataset

In this part of the study, I will use a dataset consisting of intra-chromosome Hi-C matrices from 78 experiments performed on 9 different human cell lines, including five with normal karyotype (GM12878, IMR90, NHEK, HMEC, hESC) and four with cancerous karyotypes (T47D, K565, KBM7, SKBR3), the same used in the published article this chapter is based on.

I will mostly use consider matrices binned at 100 Kb, but some analyses in the following chapters will be performed also on higher resolution maps to give a more complete view.

The sra-toolkit was used to fetch the Hi-C datasets from the public sequence read archive (SRA) and to convert them to FASTQ format after validation. The TADbit pipeline was used to (i) check the quality of the FASTQ files; (ii) maps the paired-end reads to the Homo sapiens reference genome (release GRCh38/hg38) using GEM accounting for restriction enzymes cut-sites; (iii) remove non informative reads using the default TADbit filtering options; (iv) merge datasets within each experiments when appropriate; (v) normalize each experiment using the OneD method at 100 kb resolution.

2.1.1 Normalization of Hi-C maps

The presence of the contact frequency decay, dependent on the genomic distance between interacting loci, exponentially inflates the importance of the diagonal. In some applications, such as the comparison between different matrices which is one of our aims, this is not ideal: in fact it has been shown that most of the variability between different cell-types is given by A/B compartments, away from the diagonal[78, 79]. On the other hand TADs, which span interactions between loci close along the chain, have been shown to be conserved in a wide array of cell-types[80, 81, 82, 83]

In order to avoid this problem we apply a normalization scheme in which each entry $A_{i,j}$ of the Hi-C matrix A is divided by the average of the interaction frequencies between loci at the given genomic distance $s = |i - j|$. Hence we obtain a normalized matrix whose entries are $B_{i,j} = \frac{A_{i,j}}{I(s)}$.

This is called the Observed over Expected normalization as it removes the genomic distance dependency of the interaction frequencies, allowing to easily determine which

contacts are enriched with respect to the expected background interaction, which often only contains information about the chain-connectivity of the chromatin filament.

It is important to point out that this is not the only normalization scheme one can use. For example one could obtain the expected interaction rate as a function of the genomic distance from an ensemble average (as opposed to obtaining it at the single matrix level), or from a polymer model[46].

Moreover, there are specific cases in which this scheme fails to consider the biological importance of the distance dependence: one glaring case is found in mitotic chromosomes, where a secondary diagonal is observed along with the central one, pointing to an enrichment of contacts between loci 3 Mb apart along the chromatin chain. In such cases an *ad hoc* treatment may be more indicated.

2.2 Random Matrices

Random matrices are matrices whose elements are randomly sampled from a distribution. Depending on the needs, this simple definition can be further enriched by incorporating other required properties, such as symmetry, or hermiticity, obtaining several sub-classes of random matrices, which take the name of ensembles. In fact, one is not usually interested in the properties of a single random matrix: rather, the focus is placed on averages and common features of these ensembles of random matrices.

The most studied among all ensembles are the Gaussian ensembles, where matrices are characterized by having elements distributed according to Gaussians. They are divided into different classes according to additional properties one imposes.

One can require the matrix to have real values and to be symmetric, for example by taking a random matrix M, whose elements have been independently sampled from the same distribution, and defining a new matrix $M' = \frac{1}{2}(M + M^T)$, where the operation $(\cdot)^T$ denotes matrix transposition. The eigenvalues of such a matrix M' are all real. This defines the first of the Gaussian ensembles, and it's called the Orthogonal Gaussian ensemble (GOE). The name refers to the fact that matrices of this ensemble are invariant under orthogonal transformations, such as rotation.

One can also make the entries of the matrix complex or quaternionic, but in order to have real eigenvalues additional symmetry requirements are needed: matrices with complex entries need to be self-adjointed, i.e. hermitian, and matrices with quaternionic elements must be self-dual. By considering such matrices one obtains the Gaussian Unitary (GUE) and Gaussian Symplectic ensembles (GSE) respectively[84].

Since Hi-C matrices have strictly real entries, I will only consider random matrices of the Gaussian Orthogonal ensemble.

In the next subsection I will detail some of the spectral properties of this particular ensemble.

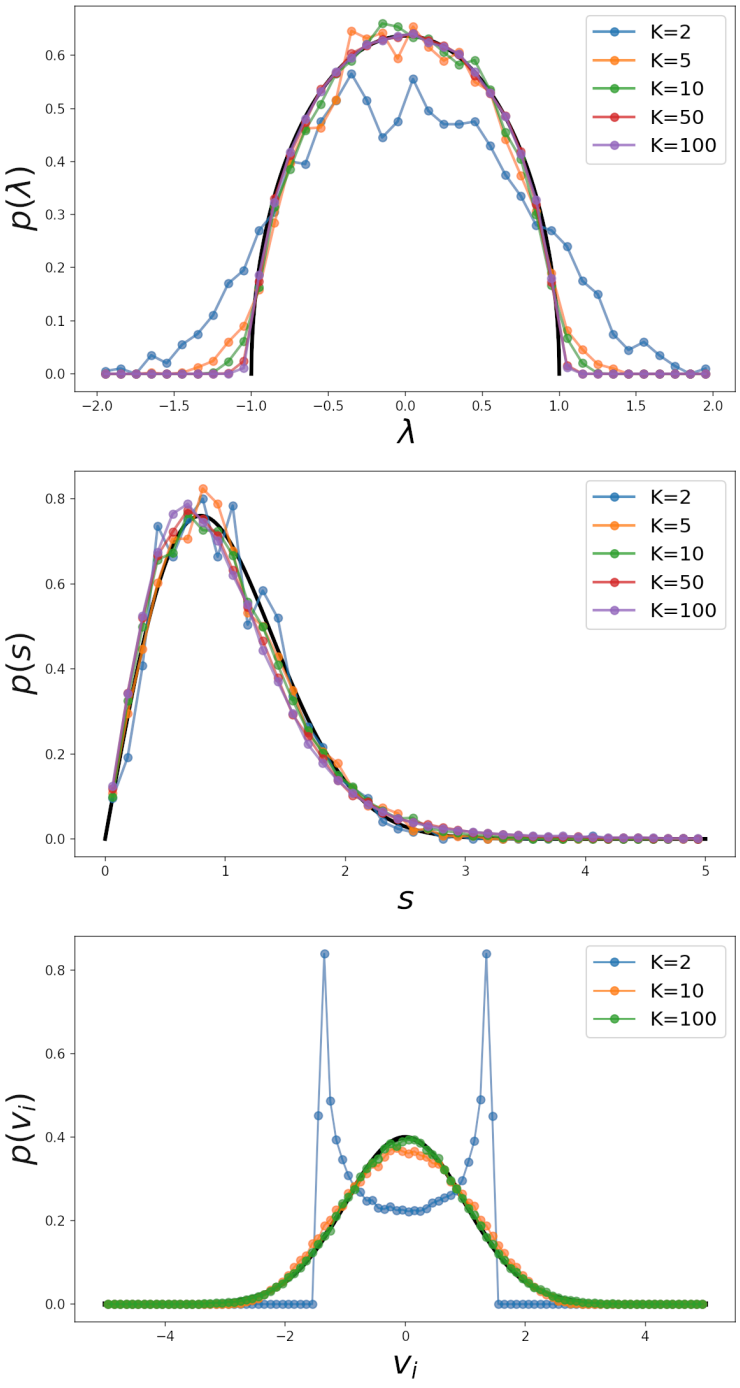

Figure 2.1: Spectral properties of random matrices of the Gaussian Orthogonal ensemble. The top image shows the distribution of eigenvalues, compared to the semi-circle expected distribution; the middle image shows the distribution of the spacings, compared the the Wigner surmise distribution; the bottom image shows the distribution of the components of an eigenvector of a random matrix, compared to the expected Gaussian distribution.

2.2.1 The spectrum of random matrices

Here I consider random matrices whose elements are distributed according to a Gaussian distribution $N \exp[-x^2/(2\sigma^2)]$, with N being a normalization factor. Hence these matrices belong to the Gaussian orthogonal ensemble.

If one considers matrices of this ensemble having infinite size, one can observe that their eigenvalues are distributed according to a semi-circle distribution [84]. However, even spectra of finite size matrices converge to the same expected distribution as the size increases.

Given matrices of linear size K, the first observation one can make on their spectra is that they are concentrated (i.e. significantly nonzero) over an interval of the real axis enclosed by $\pm\sqrt{2K}\sigma$. This does not mean that there are no eigenvalues greater than $\sqrt{2K}\sigma$ or smaller than $-\sqrt{2K}\sigma$; rather it means that regions outside this interval become more and more depleted as K increases.

It is useful to normalize the eigenvalues by the factor $\sqrt{2K}\sigma$ so that the interval in which they are concentrated is $[-1, 1]$ for any choice of σ and K.

For large values of K the expected distribution of the eigenvalues is a semi-circle[84], and this observation takes the name of Wigner's Semi-circle law. Thanks to the normalization one can use a single explicit formula for the expected distribution, given by

$$p(\lambda) = \frac{2}{\pi}\sqrt{1-\lambda^2} \tag{2.1}$$

A mathematical proof of Wigner's Semi-circle law is beyond the scope of the present work, however it is interesting to show empirically how well the measured distributions of eigenvalues conform to the expected behavior. Figure 2.1 shows the histograms of such distributions for various values of K (each one containing the eigenvalues for 1000 random matrices sampled from the GOE ensemble). As one can observe, for very small values of K the differences between the measured distribution and the expected semi-circle are notable, and eigenvalues tend to spill far beyond the $[-1, 1]$ interval. However, increasing K rapidly leads to a depletion of the eigenvalues beyond the radius of the semi-circle and to better and better approximation of the expected distribution: qualitatively, at $K = 100$ it is difficult to spot deviations from the expected behavior.

Another property of the spectrum is that, while eigenvalues can be thought of as random variables, they are not *independent* random variables[84]. In fact one can observe a repulsion between eigenvalues which would not be present were they independent. Indeed, this is the property that Wigner used to describe the level spacings in the nuclei of heavy atoms. The so called Wigner surmise tells us that given the spacings $s = (\lambda_{n+1} - \lambda_n)/ < \lambda_{n+1} - \lambda_n >$ between contiguous eigenvalues, the distribution of s is given by

$$p(s) = \frac{\pi}{2}\exp\big[-\frac{\pi}{4}s^2\big]. \tag{2.2}$$

This holds for matrices of the Gaussian Orthogonal ensemble, but analogous results are available also for the other Gaussian ensembles.

This result leads to an important corollary: the probability of sampling two very close eigenvalues (i.e. $s \to 0$) is very small. So the eigenvalues of random matrices have a tendency to avoid each other.

Finally, another aspect of the spectral properties of random matrices which merits attention is the distribution of the components of their eigenvectors.

Thanks to the rotational invariance of the Gaussian Orthogonal ensemble, the eigenvectors of a random matrix of this ensemble uniformly sample the surface of a unit sphere in K dimensions[85]. This is an exact result, but in the limit of large K it also has an important implication about the distribution of the components of said eigenvectors: they can be considered approximately distributed according to a Gaussian.

To obtain the variance of this Gaussian, one can consider the components v_i of the eigenvector $\mathbf{v}$ as approximately independent, then

$$< \sum_i v_i^2 > \approx \sum_i < v_i^2 > = K < v_i^2 > = 1, \tag{2.3}$$

where the last equality is given by the normalization constraint stating that $||\mathbf{v}|| = 1$. This is only strictly true for $K \to \infty$, however it is a good approximation also for large K, which is the case we are going to consider, given that Hi-C matrices are usually hundreds of bins wide at 100 Kb (from $K = 2500$ for chromosome 1 to chromosome 21 having $K = 480$ at this resolution).

Figure 2.1 shows the deviations of the distribution of the eigenvectors components of random matrices from the expected Gaussian, as K increases. One can immediately see that while for $K = 2$ the deviations are obvious even from a qualitative point of view, with the distribution being bimodal, they rapidly decrease so that at $K = 10$ the distribution already closely resembles the expected Gaussian. At $K = 100$ differences are small enough to be inconsequential.

In the next section I will employ random matrices of the Gaussian Orthogonal ensemble as a null model for aspecific features of the Hi-C matrices, shared across the dataset irrespective of cell-types.

2.3 Hi-C vs random matrices: comparison of the spectral properties

In this section I proceed to compare the spectral properties of Hi-C matrices, both with and without the OoE normalization, to those of random matrices sampled from the Gaussian Orthogonal ensemble.

I start by comparing the distributions of the eigenvalues: Figure 2.3 compares the spectrum of the OoE normalized intra-chromosomal Hi-C matrix relative to chromosome 17 of

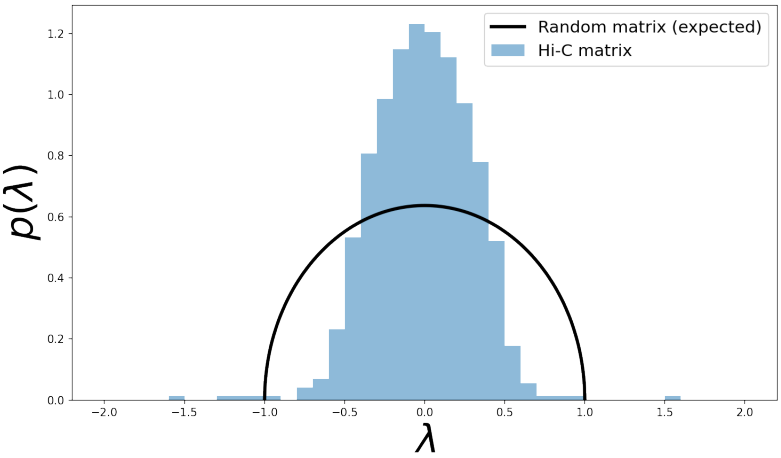

Figure 2.2: Properties of the spectrum of Hi-C maps: comparison to the semi-circular distribution expected for random matrices of the Gaussian Orthogonal ensemble. The histogram shows the distribution of the eigenvalues of the experiment HIC001 relative to its intra-chromosomal Hi-C matrix of chromosome 17 (OoE normalized).

experiment HIC001 to the expected random matrix distribution, i.e. Wigner's semicircular law. To do this I normalize the eigenvalues of the Hi-C map in the same way as those of random matrices of the GOE, by a factor $\sqrt{2K}\sigma$, where σ is the standard deviation of the eigenvalues.

When I do this for the whole spectrum of the Hi-C matrix, one observes that the histogram does not match the null-model provided by the random matrices. This is not surprising: one would not expect Hi-C matrices to entirely behave like random matrices, unless genome folding was a completely random process. Since this is not the case, the spectrum displays glaring differences with that of random matrices.

However, one can notice the presence of outliers in the distribution of Hi-C eigenvalues: whereas most of the spectrum is concentrated in a central zone, some eigenvalues appear to have much larger absolute values, and inhabit the fringes of the histogram. One can order the eigenvalues according to their absolute values, making the outliers correspond to the top part of the spectrum.

This observation is interesting because one can wonder what would happen if one removed the top eigenvalues *before* computing the normalization. In figure 2.3 I consider different values of a threshold n^*, and compute the standard deviation σ of the spectrum only for those eigenvalues with orders larger than n^*.

One observes that, as soon as the first $n^* = 10$ eigenspaces are removed, the spectrum distribution of OoE matrices starts to closely resemble that expected for random matrices. This becomes even more evident when one considers larger thresholds, removing the top

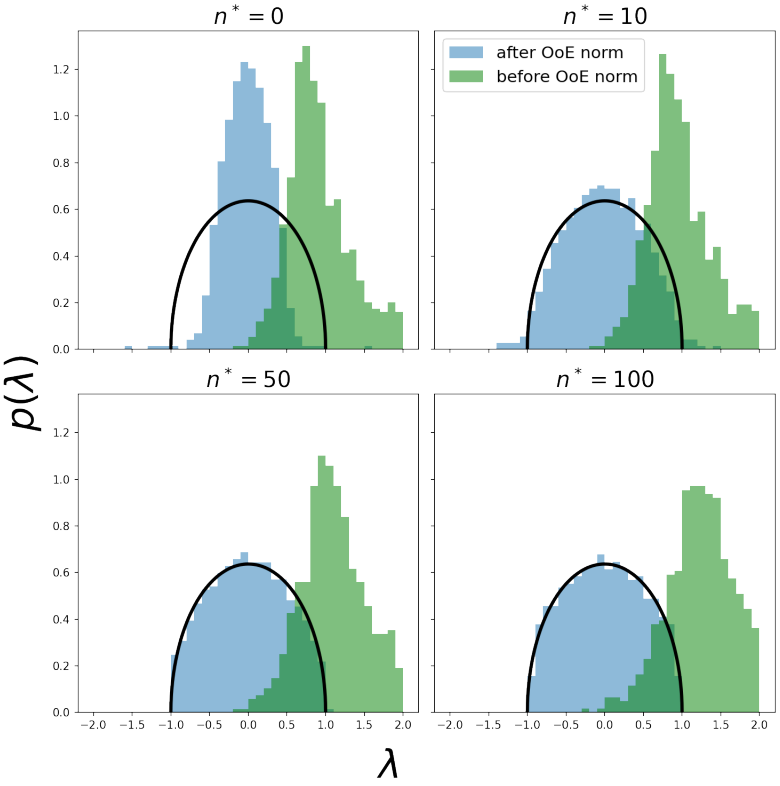

Figure 2.3: Properties of the spectrum of Hi-C maps: comparison to the semi-circular distribution expected for random matrices of the Gaussian Orthogonal ensemble. Each panel shows the results for different values of n^*, corresponding to the quantity of non-random eigenspaces which are excluded from the analysis. The histograms show the distribution of the eigenvalues of the experiment HIC001 relative to its intra-chromosomal Hi-C matrix of chromosome 17, as computed before and after the observed-over-expected normalization (green and blue coloring, respectively).

50 or even 100 eigenvalues.

A simple interpretation, could be that interaction frequencies found in Hi-C matrices are the result of two superimposed signals: one contains relevant biological information, while the other displays noise like features, it fluctuates rapidly and is not specific of any particular cell-type, having the same properties of random matrices. I call the former the *essential* component of the matrix, in reference to essential space analysis performed on the spectra of elastic networks[17]. In the model I am presenting, the essential component is encoded into the top eigenspaces of Hi-C matrices, while the aspecific part is confined to the rest of the spectrum. The two can be separated by tuning a threshold n^* according to some criterion.

As such, an element of a Hi-C matrix A of linear size K can be decomposed into an essential component $A^{ess}_{i,j}$ and an aspecific one $A^{asp}_{i,j}$ as

$$A_{i,j} = \sum_{n=0}^{K} \lambda_n a_i^{(n)} a_j^{(n)} = \sum_{n=0}^{n^*-1} \lambda_n a_i^{(n)} a_j^{(n)} + \sum_{n=n^*}^{K} \lambda_n a_i^{(n)} a_j^{(n)} = A^{ess}_{i,j} + A^{asp}_{i,j}, \tag{2.4}$$

where λ_n is the n-th eigenvalue, $a_i^{(n)}$ is the i-th component of the n-th eigenvector.

The results of figure 2.3 can be reproduced in different chromosomes and different Hi-C maps across the 9 cell-types of the dataset[77]. For this reason this range of values for n^*, between 10 and 100, already provides an estimate of the number of essential eigenspaces of Hi-C maps.

Figure 2.3 also shows the spectral distributions of the original matrices, where the observed-over-expected normalization is not applied, so that the interaction frequencies still depend on the genomic distance between couples of interacting loci. One can observe that, as opposed to those of normalized matrices, these spectra do not conform to the distribution expected for random matrices. This means that the observed-over-expected normalization is important to decouple the essential and aspecific signals present in the matrix.

The second aspect that can be investigated is the distribution of level spacings, the distance between consecutive eigenvalues, which in random matrices follows the Wigner surmise.

As for the previous analysis, different values of n^* lead to different distributions because of the definition of the spacings, which are normalized according to the average difference between consecutive eigenvalues $< \lambda_{n+1} - \lambda_n >$. Figure 2.4 shows, again, that the presence of large eigenvalues with large gaps between them in the initial region of the spectrum leads to a squeezed distribution, but removing this part of the spectrum allows to obtain a better agreement with the expected one. Spacing distributions obtained from an actual random matrix sampled from the Gaussian Orthogonal ensemble are shown alongside those from Hi-C matrices: one can notice that the histograms obtained from the spectrum of a single

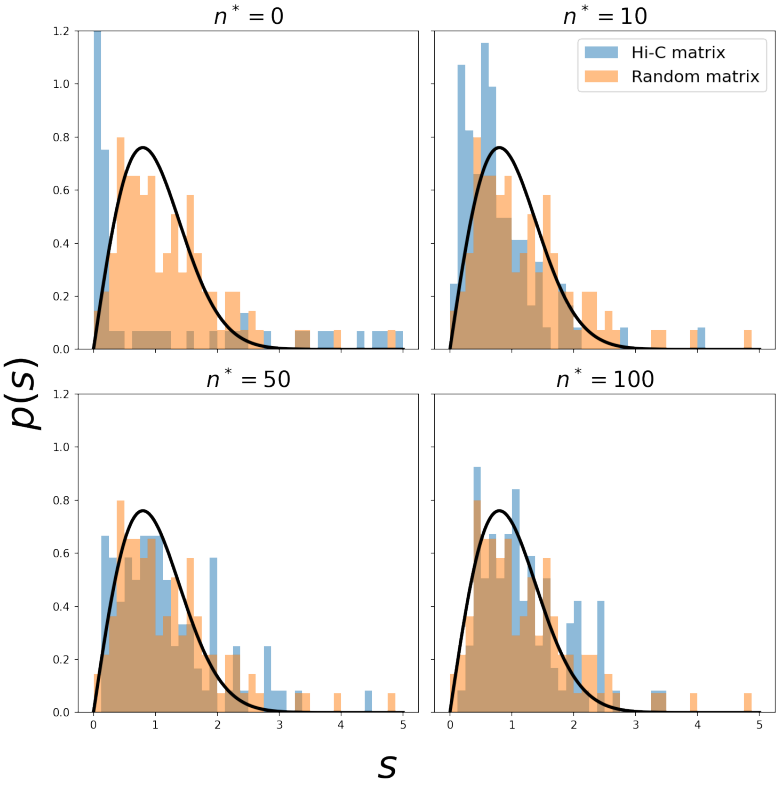

Figure 2.4: Properties of the spacings between eigenvalues of Hi-C maps: comparison to the Wigner's surmise distribution expected for random matrices of the Gaussian Orthogonal ensemble. Each panel shows the results for different values of n^*, corresponding to the quantity of non-random eigenspaces which are excluded from the analysis. The histograms show the distributions of the spacings of the experiment HIC001 relative to its intra-chromosomal Hi-C matrix of chromosome 17 (blue), and of the eigenvalues of a single random matrix (orange).

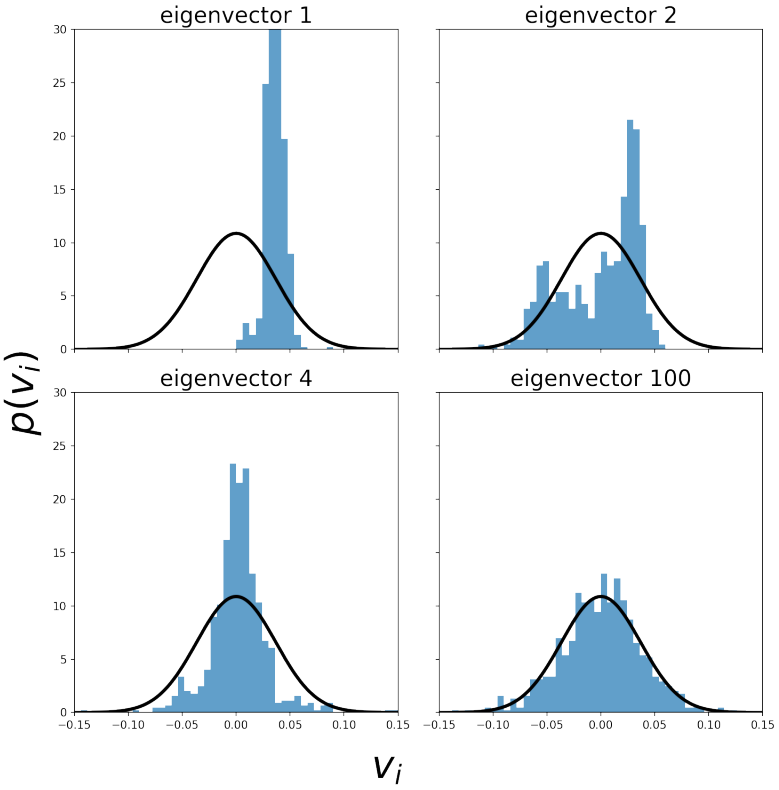

Figure 2.5: Properties of the eigenvector components of Hi-C maps: comparison to the normal distribution expected for random matrices of the Gaussian Orthogonal ensemble. Each panel shows the results for eigenvectors of different orders. The histograms show the distributions of the eigenvector components for the experiment HIC001 relative to its intra-chromosomal Hi-C matrix of chromosome 17.

matrix, random or otherwise, are very noisy with respect to those presented in Figure 2.4, which contained contributions from the spectra of many random matrices.

The final property I am going to analyze is the distribution of the components of eigenvectors, which is expected to be a Gaussian for the eigenvectors of a random matrix. In this case one does not need to separate the essential part of the spectrum from the aspecific one *a priori* in order to carry out the analysis, because one can consider a single eigenvector at a time.

A qualitative comparison between the measured distribution of the components with the expected Gaussian is shown in Figure 2.5 for eigenvectors of different orders.

The first few eigenvectors display glaring discrepancies from the Gaussian distribution. The components of the first eigenvector are concentrated in the positive region of the axis,

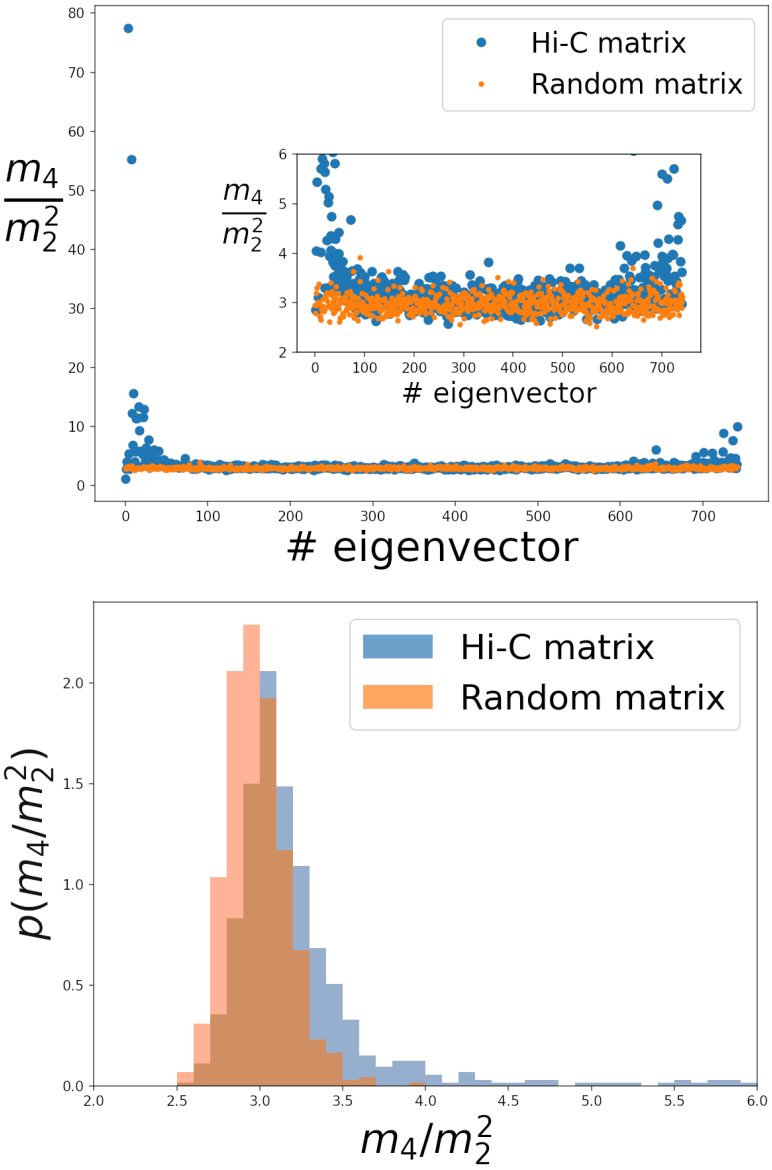

Figure 2.6: Analysis of the kurtosis of Hi-C maps eigenvectors, compared to those of random matrices of the Gaussian Orthogonal ensemble. The upper panel shows the ratio between the moments of each eigenvector, ordered decreasingly according to the absolute value of the associated eigenvalue, of the experiment HIC001 relative to intra-chromosomal Hi-C matrix of chromosome 17, along with the same measurement for eigenvectors of a random matrix. The inset in the same panel shows a zoom-in around the value expected for normal distributions, which is 3. The lower panel shows the distributions of the points plotted above, both for the Hi-C matrix and the random matrix.

suggesting all loci participate in this component of the interaction; those of the second eigenvector, on the other hand, present a bimodal behavior, which shows a separation of the loci into two well defined clusters. These observation can be repeated for different chromosomes and experiments, and seem robust features of Hi-C maps: in the next chapter I will show their connection to some of the patterns previously introduced. Here, however, it is interesting to note that these deviations, even without further observations suggested by the nature of the matrices, point towards a highly non trivial behavior which cannot be explained by the random matrix null model.

On the other hand, if one looks at the components of the 100-th eigenvector, and beyond, one finds that they qualitatively adhere to the expected Gaussian, suggesting that the corresponding eigenspaces only contain aspecific noise, comparable to that of a random matrix.

It is also informative consider a more quantitative analysis: one can take the second and fourth moment of the measured distributions, their variance m_2 and kurtosis m_4 respectively, as a function of the degree of the corresponding eigenvector. As the ratio $\frac{m_4}{m_2^2}$ is expected to be 3 for Gaussian distributions, it can be used as a quantitative estimator of normality.

Figure 2.6 shows the results of this analysis: the initial region of the spectrum shows a highly non-Gaussian behavior, displaying large deviations from the expected ratio, with a peak around $n^* = 10$. After the peak, the following eigenvectors slowly transit towards normality: significant deviations are still visible in a large portion of the spectrum, but eigenvectors of degree 100 or higher tend to display a Gaussian behavior consistent with the one measured for components of the eigenvectors of a random matrix. Other smaller deviations are found in the final tail of the spectrum, corresponding to eigenvalues close to 0.

2.4 Summary and conclusion

Random matrices provide compelling models for many physical phenomena where averaged ensemble properties can be more informative with respect to specific details of a single realization, especially when disorder and noise are heavily involved. In fact, many properties of such matrices, in particular those linked to their spectra, are universal. They do not strongly depend on specific aspects of the matrices, but rather on a small set of rules involving their symmetries. This is true in the regime of large matrices, which is effectively achieved for linear sizes of order 100 or greater[84].

For matrices of the Gaussian Orthogonal ensemble, symmetric real matrices whose entries are normally distributed random variables, one finds that eigenvalues are distributed according to a semi-circle and display spacings following Wigner's surmise. Moreover, the components of their eigenvectors approximately behave as independent Gaussian numbers.

These matrices provide a null model for the stochastic component of Hi-C matrices,

albeit one that has not been explored before. In fact a large part of the spectrum of OoE normalized Hi-C matrices is consistent with random matrices. On the other hand, the top eigenspaces, ordered according to the absolute value of their eigenvalues, display large deviations with respect to the null model.

Hence one can enucleate an essential component which only contains the top eigenspaces by setting a threshold spectral order n^*. Quantitative analysis of the distributions of eigenvectors components suggests that an optimal value for n^* should not be larger than 100, and is probably closer to 10.

The remainder of the spectrum has properties that only depend on the size K of the matrix, and exhibits noise like behavior. While these eigenspaces can still carry a residual part of the biological signal contained in Hi-C matrices, their sum can be regarded as an aspecific component, common to all matrices of the same size.

Possessing the properties of random matrices, the aspecific part of the spectrum displays similar properties in different chromosomes and across cell-types. The next chapter will be devoted to understanding whether this part of the spectrum can be disregarded in order to carry out some analyses and what benefices can be reaped by limiting oneself to study the essential component of the spectrum instead.

Chapter 3

Analysis of the essential spaces of Hi-C matrices

In the previous chapter the spectral properties of intra-chromosomal Hi-C matrices were compared to those of matrices of the Gaussian Orthogonal ensemble: strikingly, a large part of the spectrum of Hi-C matrices displays properties which are compatible with random matrices, suggesting that only, or mostly, aspecific signals are encoded in these eigenspaces. More interestingly, though, the top few eigenspaces, ordered according to the absolute value of the associated eigenvalue, seem to deviate significantly from the random behavior of aspecific eigenspaces. I call this part of the matrix *essential*, as a nod to the analysis of essential spaces in elastic network [17], making the rest of the spectrum *non-essential*.

This chapter, based on published work [77], delves deeper in the properties of the essential part of the matrix.

I start by comparing the full matrices to their essential part. At the single matrix level I will show that removing the aspecific component improves concordance with experiments having deeper sequencing (i.e. better sampling) of the same cell-type and helps recover specific features, such as TADs, with a higher degree of fidelity.

At the dataset level, I will show that pairs of essential matrices are easier to sort between biological replicates (experiments of the same cell-type) and non-replicates: I will compare the result with those of other published spectral methods [12, 13]

Finally, I will apply an ad-hoc version of this analysis to single-cell Hi-C matrices in order to cluster them according to their position along the cell cycle.

3.1 The essential spaces

As seen in the previous chapter, eigenspaces circa up to the 100th display remarkably non-random properties, which may be a hint that they contain most of the biologically significant signal carried by Hi-C matrices. One can then ask whether removing the aspecific

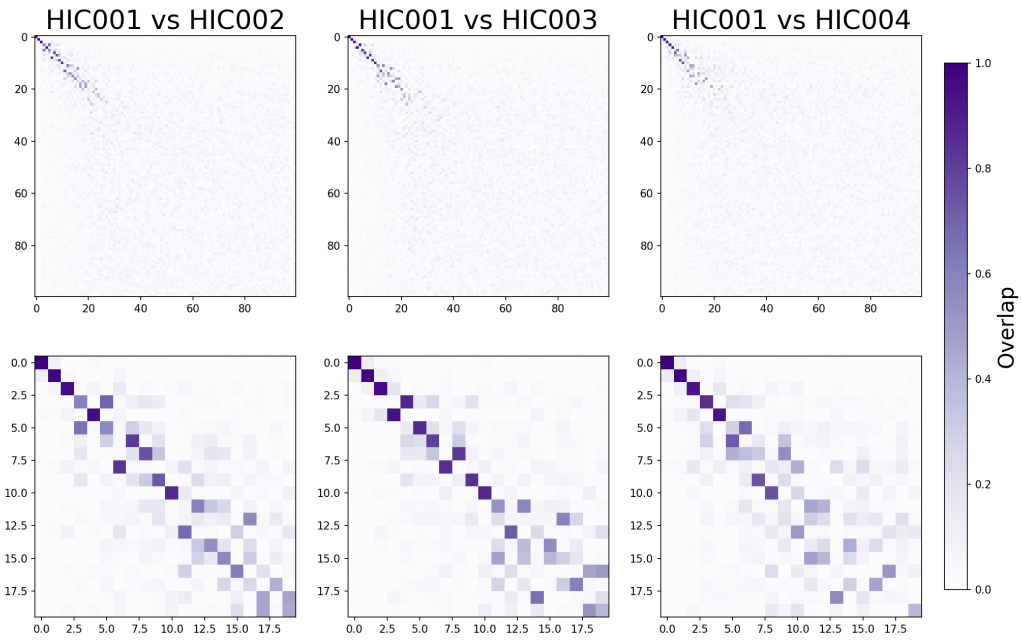

Figure 3.1: Absolute value of the overlap of the top eigenvectors of HIC001 with those of HIC002, HIC003, and HIC004 (chromosome 17). All of these experiments share the same cell-type (GM12878). The top panels show the first 100 eigenvectors, while the bottom ones zoom on the first 20.

part of the matrix can help with different tasks where noise can be a problem, for example reproducibility measurement or detection of certain patterns [14, 11, 12, 13].

In order to answer these questions I must first define a threshold n^* for the number of essential eigenspaces. For simplicity, here I fix $n^* = 10$, which is more stringent with respect to the previous observation that eigenspaces up to 100 display some non-random behaviors. This choice is suggested by the observation of the overlaps between eigenvectors of matrices of the same cell-type (GM12878), defined as their inner product $v_A \cdot v_B$ of the eigenvectors of two matrices A and B, and plotted in figure 3.1. One can see that while the top ~ 10 eigenvectors tend to preserve their identity between different experiments, the remaining eigenspaces become more and more intermixed.

This qualitative argument for taking $n^* = 10$ is supported *a posteriori* by the observation that analyses on essential matrices reach optimal results around this value for the threshold. Nevertheless the result are very robust with respect to larger choices for n^*.

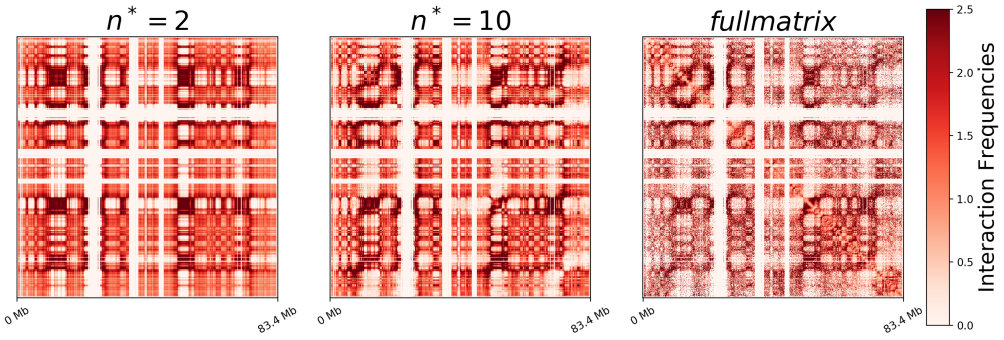

Figure 3.2: Essential matrices (for $n^* = 2$ and $n^* = 10$) computed from experiment HIC001, chromosome 17. The results are shown along the full matrix.

Hence, following a principle of parsimony, I only consider the top $n^* = 10$ eigenspaces as essential.

Having fixed the value for n^*, the resulting essential matrix is given by a sum over the top 10 projectors of the essential eigenspaces, weighted by the corresponding eigenvalues

$$A^{ess}_{i,j} = \sum_{n=1}^{n^*} \lambda_n v_i^{(n)} v_j^{(n)} = \sum_{n=0}^{n^*} \lambda_n P^{(n)}_{i,j} \tag{3.1}$$

where $P^{(n)}$ is the projector over the n-th eigenspace. An example of the resulting essential matrix is plotted in figure 3.2, along with the original matrix.

It can be shown that this is the optimal approximation of rank n^* of the original matrix with respect to the Frobenius norm (see Appendix B).

3.1.1 Physical interpretation of the eigenspaces

As seen above, in figure 3.1, the top 10 eigenspaces are the most conserved in different experiments: it is then interesting to ask whether they can be given some physical interpretation in terms of the patterns observed in Hi-C matrices.

Previous studies [7] also attempted an interpretation of eigenvectors of Hi-C maps in a different context, by first applying an iterative correction scheme to the matrices and then comparing the resulting eigenvectors to known structural and biological signals. Although in that case the analysis was only carried out on the top three eigenvectors, I will also follow the same route by computing the correlations with two one dimensional signals.

The first signal I consider is the column-wise sum of the elements of matrix A:

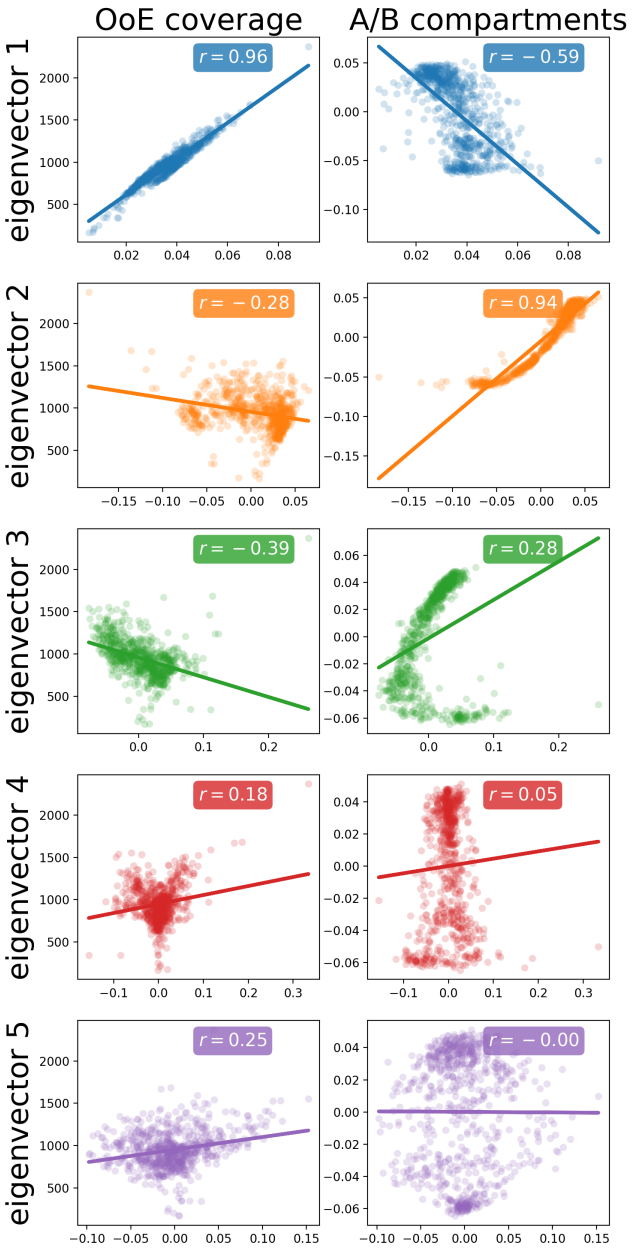

Figure 3.3: Top eigenvectors of HIC001 chromosome 17 to two known biological signals: the Observed over Expected coverage and the first principal component of the Hi-C matrix, corresponding to A/B compartments. Each panel contains the Pearson correlation coefficient r.

$$c_i = \sum_j A_{i,j} \tag{3.2}$$

In the case of matrices which did not undergo the observed over expected normalization procedure, this quantity would be interpreted as the coverage of the matrix. In the case of OoE-normalized matrices, the interaction between any pair of loci is normalized to be 1 on average, so that the expected coverage is simply $\langle c_i \rangle = N$, where N is the linear size of the matrix. Hence, naturally, the quantity c_i can be interpreted as the observed over expected coverage.

Next I consider the A/B compartments signal, σ_i, as computed using PCA using the method detailed in van Berkum et Al. [86] and Miura et Al. [87].

Figure 3.3 shows the linear fit of the top 5 eigenvectors components to these two signals.

One immediately sees that the first eigenvector is strongly correlated (Pearson $r = 0.96$) with the coverage: this vector is mostly positive (with the constraint that its first component be positive) and encodes the interaction propensity of each node. On the other hand, the second eigenvector is correlated with A/B compartments ($r = 0.94$) and oscillates between positives and negatives values which represent loci with a prevalence of either euchromatin or heterochromatin. These results are consistent with what is observed in previous studies on the topic [7].

Simply adding together the projectors associated to these two eigenspaces is enough to obtain a recognizable interaction matrix, albeit many features get lost, as can be seen in figure 3.2.

These results are encouraging, but one has to ask whether a certain feature is always associated with the same eigenvector in different matrices, or whether eigenvectors can change places or get mixed together. Figure 3.1 shows that while the first two or three eigenvectors always maintain their position and do not mix much between matrices, the others are more prone to getting swapped (i.e. there is a mismatch in the index of two highly overlapped eigenvectors) or mixed (i.e. a certain eigenvector of the first matrix A significantly overlaps with multiple eigenvectors of the other matrix B).

This means that there is no strong evidence that eigenvectors of higher order can be identified with a single well defined Hi-C interaction patterns.

3.2 Enhancement of specificity

While the essential part of a Hi-C matrix is the best approximation of rank n^* of that matrix, with respect to the Frobenius norm, the objective of this work is not to reproduce Hi-C matrices, which also contain aspecific patterns: rather I want to enucleate the salient interaction patterns.

In this section I will show that essential matrices built from the first 10 eigenspaces of Hi-C matrices have more consistent properties within experiments of the same cell-type,

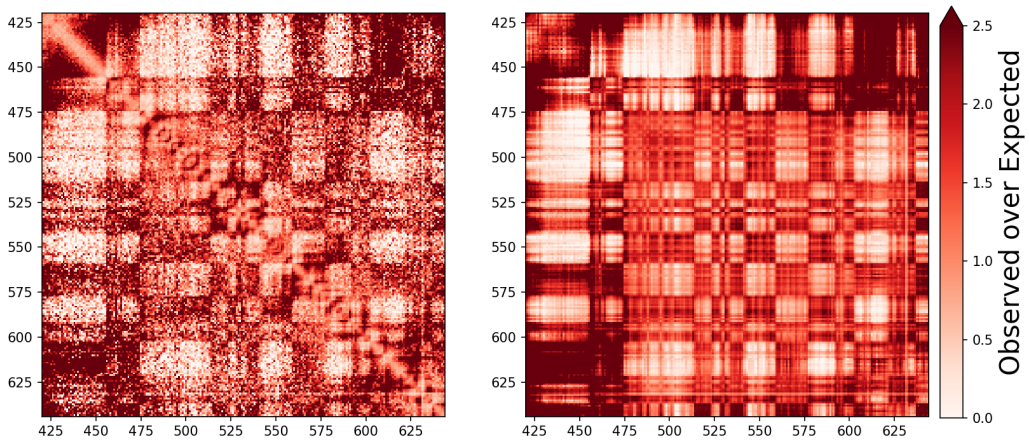

Figure 3.4: Visual comparison of a zoomed portion of experiment HIC001 chromosome 17 at 100 Kb resolution, before (left) and after (right) applying the spectral filter in order to extract the essential component.

and display stronger correlation coefficients with high resolution maps within the dataset. They can also be analyzed to identify known spatial patterns, such as TADs, to a higher degree of robustness.

3.2.1 Visual inspection

First, I start with a qualitative comparison of full Hi-C maps with the essential ones: figure 3.4 shows a close up of the map of chromosome 17 from experiment HIC001, before and after applying the spectral filter in order to extract the essential component. A visual inspection reveals that features in the essential matrix are more uniform and present sharp boundaries, although some of the more fine-grained patterns, especially along the diagonal, are not retained.

The next few subsections will deal with more quantitative analyses which will further clarify the advantages of using essential matrices and whether or not they reflect the same patterns contained in the full Hi-C maps.

3.2.2 Signal to Noise Ratio

One can define a ratio γ between the signal and noise contents of a matrix [88, 89]. To do so, one can assume that a matrix A can be written as a sum of some unknown true

interaction pattern Σ and a noise η:

$$A = \Sigma + \eta \tag{3.3}$$

The former has the property that it is shared across all maps of experiments on the same cell-type, while the distribution of the latter is assumed for simplicity to have the following characteristics:

$$\begin{aligned} \langle \eta_{i,j} \rangle &= 0 \\ \langle \eta_{i,j}^2 \rangle &= \sigma^2, \end{aligned} \tag{3.4}$$

where the average is carried out on different realization. Notice that, at $100Kb$, a matrix for a single chromosome typically contains $\sim 10^5$ bins: hence averaging the noise on the matrix elements yields results similar to the actual statistical average.

Notice that by this definition, the noise is simply the variation between elements of the same dataset, rather than an intrinsic property of a single matrix. With this being the case one can then obtain the following quantities even without knowing Σ and η explicitly, by introducing a second matrix of the same cell-type $B = \Sigma + \eta'$, where η' is some other independent realization of the noise, having the same properties of η:

$$\begin{aligned} \langle A + B \rangle &= \langle 2\Sigma \rangle = 2\bar{\Sigma} \\ \langle (A - B)^2 \rangle &= \langle 2\eta^2 \rangle = 2\sigma^2, \end{aligned} \tag{3.5}$$

where I used the fact that the signal Σ is the same in the both matrices, and that the realizations of the noise are independent. I can finally define the signal to noise ratio γ as

$$\gamma = \frac{\bar{\Sigma}}{\sigma} \tag{3.6}$$

This quantity should be close to 1 if the signal and noise contained in the matrices have the same amplitude, but should be much higher in datasets where matrices are more robust.

For all chromosomes of experiments from the GM12878 cell-type, I compute γ on all pairs of matrices and plot the results in figure 3.5. The figure shows that in all cases essential matrices display larger values of γ on average, with increases between 5 and 10 times the γ values of original matrices. This shows that essential spaces contain highly specific features and there is little variance between essential matrices extracted from experiments of the same cell-type. This dramatic boost shows little bias with respect to chromosome length.

3.2.3 Correlation with high resolution Hi-C maps

The previous analysis shows that the removal of non essential spaces increases the homogeneity of the dataset. However this analysis does not prove that all the relevant features

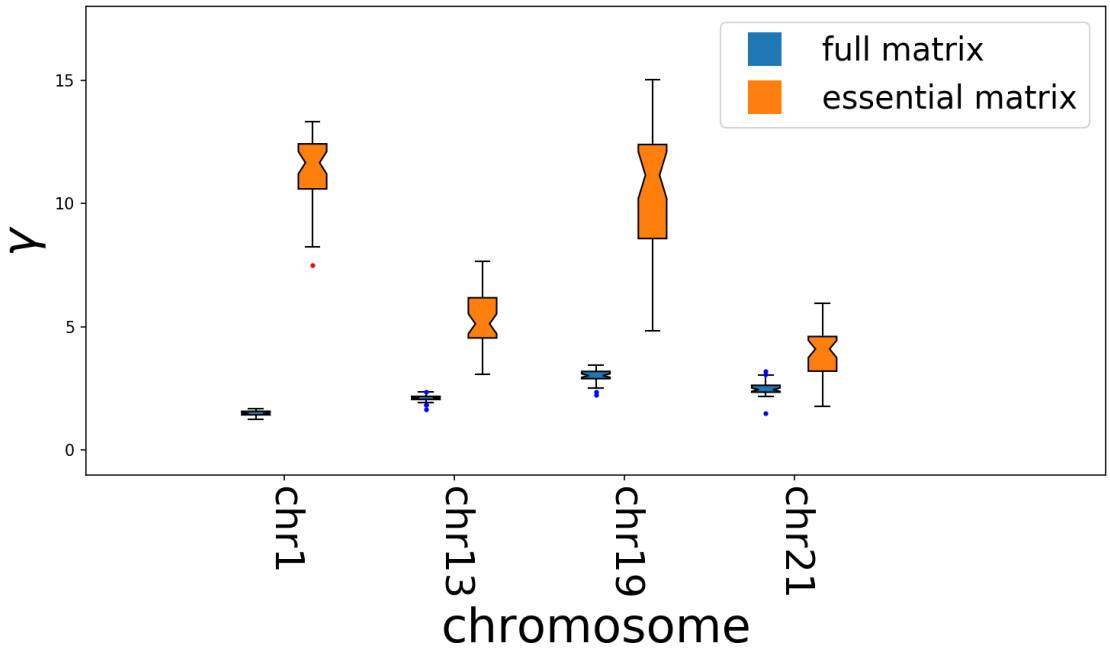

Figure 3.5: Signal to noise ratio γ as boxplots for 4 chromosomes, selected for length and gene richness. Other chromosomes are shown in the appendix. Boxplots show: central line, median; box limits, 75th and 25th percentiles; whiskers, 1.5 times the interquartile range; outliers beyond this range are shown as individual points.

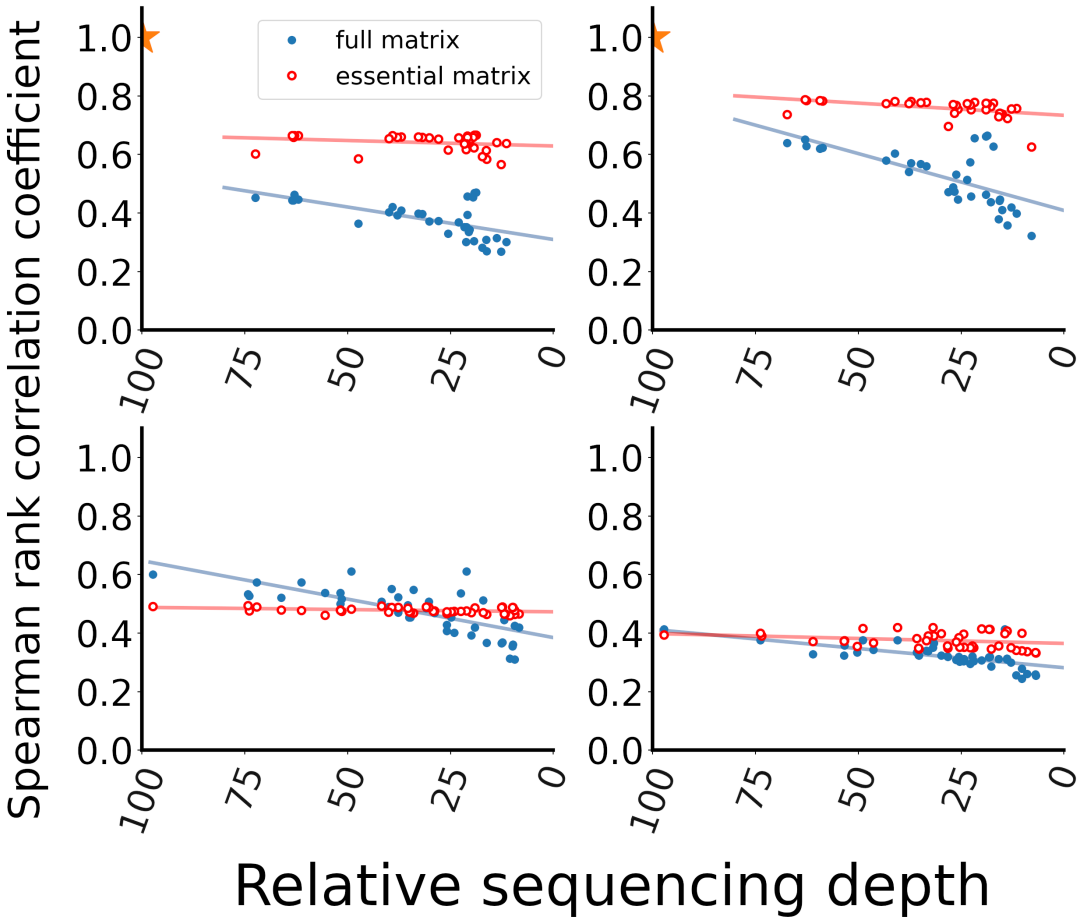

Figure 3.6: Spearman Rank correlation coefficient between a representative matrix (HIC003) and other experiments with lower sequencing depth of the same cell-type (GM12878, upper panels) and different cell-type (IMR90, lower panels). The panels on the left refer to chromosome 1, those on the right to chromosome 13.

are maintained in the passage between a Hi-C matrix and its essential counterpart: the removal of aspecific spaces could lead to an unwanted loss of signal.

To show that this is not the case, this subsection will compare Hi-C maps of the GM12878 cell-type and their essential components to the gold standard given by the highest resolution map of that cell-type in the dataset (i.e. the one with the largest sequencing depth, HIC003). I stress that the gold standard is given by the full matrix of experiment HIC003, even when it is compared to essential components.

To carry out the comparison I will use the Spearman rank correlation, which is a standard tool to compare matrices, and compute the correlation coefficients ρ between the gold standard and other matrices. Notice that I will compare the essential components to the original Hi-C matrix of experiment HIC003.

Figure 3.6 shows the results of the analysis: the correlation coefficient is consistently higher for the essential matrices than it is for their original Hi-C counterparts, meaning that discounting the aspecific part improves the robustness of high resolution features. Moreover, while the original full Hi-C maps show a visible dependence from the sequencing depth, this is visibly not the case for their essential components: in this case the dependence is much milder, with even the lowest sequenced matrices achieving as good a correlation as those with higher sampling.

As a control, I repeat this test also between HIC003, which is kept as the gold standard, and experiments from the IMR90 cell-type: in this case one expects correlations to remain low even after the extraction of the essential component of the Hi-C matrices. Figure 3.6 shows that this is the case, as both full matrices and essential ones display lower correlation coefficients than experiments from the same cell-type as HIC003, and comparable to each other. Essential matrices also display, again, milder dependence on sequencing depth, with ρ being more or less constant over the whole range of values taken by it.

These results show that essential spaces hold the most genuine features, i.e. those encountered in high resolution matrices, and are mostly unaffected by a reduction of the sequencing depth in the matrix from which they are extracted (at least on the range considered here, which accounts for a ten-fold variation in the sequencing depth with respect to the gold standard). While some of the signal may still be lost by removing non-essential spaces, most of it is contained in the essential ones.

Notice also that these observations are general, in the sense that they hold irrespective of the chromosome considered, with only slight variations in the details[77].

3.2.4 Application to TAD detection

Essential matrices can also help identifying structural features from Hi-C maps, such as topological associating domains, or TADs.

I used a standard measure of local contacts insulation [55], described in section 1.4.2, to obtain the boundaries of the TADs in both the original full Hi-C matrices and in their essential components, and compared the results to the gold standard given by the TADs

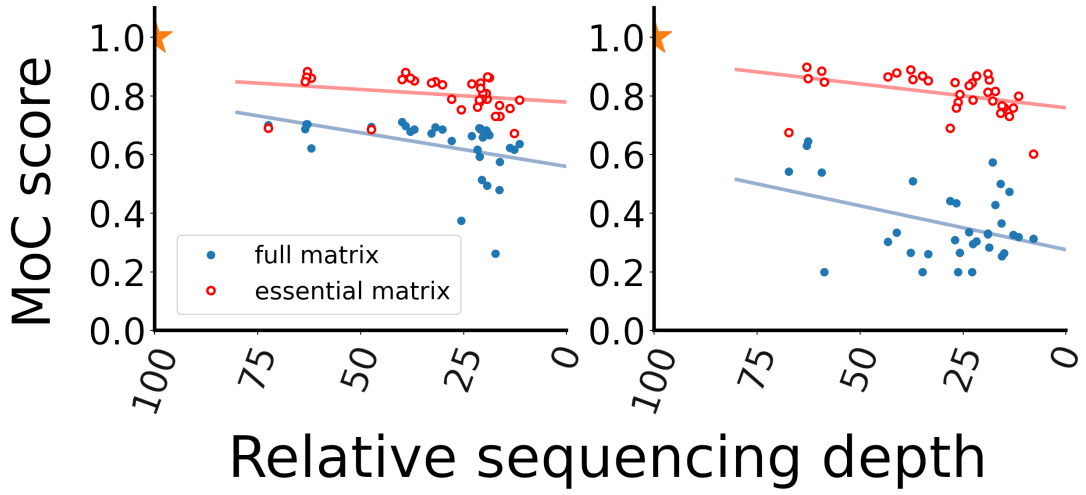

Figure 3.7: MoC score computed by comparing TADs partitions obtained by applying the Insulation Score to a representative matrix (HIC003) and other experiments with lower sequencing depth of the same cell-type (GM12878). The panel on the left refers to chromosome 1, the one on the right to chromosome 13.

boundaries of the highest sequencing depth full matrix in the GM12878 cell-type (HIC003).

Next I used the MoC score[90, 91] to measure the similarity of such TAD segmentations with respect to the gold standard, which is a standard analysis to assess TAD reproducibility.

Given two sets of TADs, MoC asseses the overlap between each pair of TADs, measured in number of bins and considering the overall size of both TADs. MoC ranges from 0, for complete lack of concordance, to 1, perfect concordance. It has the desirable property of being symmetric and has been used in the same context in a previous work[91].

The results are shown for chromosomes 1 and 13 in figure 3.7 as a function of the sequencing depth of the full matrices. At all sequencing depths, one can observe that essential matrices improve significantly TAD detection over full ones.

More in general, one can expect essential component analysis to enhance contact patterns or structural features at lengthscales comparable with those that are exclusively, or mostly, covered by the retained essential eigenvectors. Such lengthscales can be obtained through Fourier component analysis of the eigenvectors.

The characteristic lengthscale is computed from the real Fourier transform of the eigenvector's components, F_k, and is set equal to $N/\langle k \rangle$, where N is the linear size of the matrix, and

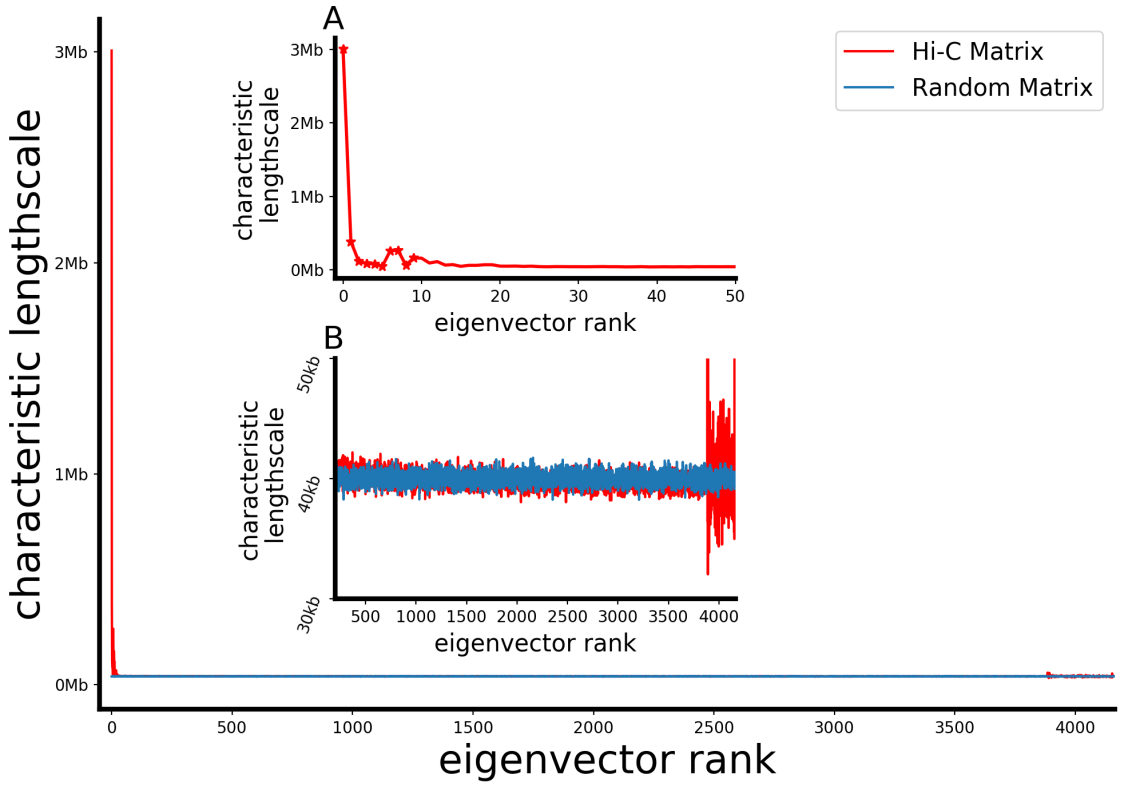

Figure 3.8: Characteristic lengthscales of eigenvectors determined through Fourier analysis. Inset A displays the characteristic lengthscale of the first 50 eigenvectors of the Hi-C matrix. Data for the first 10 eigenvectors (star symbols) range from 3Mb for the very first eigenvector to 40kb. The lengthscale of lower raking eigenvectors is shown in inset B, showing superposable features to random matrices. The final part of the Hi-C spectrum, which displays larger fluctuations, corresponds to null eigenspaces, and carried no meaningful information.

$$\langle k \rangle = \frac{\sum_k k F_k^2}{\sum_k F_k^2} \tag{3.7}$$

is the average wave-vector.

Figure 3.8 shows that top ranking eigenvectors cover an atypically broad spectrum of lengthscales, from tens of Kb to Mb, compared to the remainder of the spectrum, whose characteristic lengthscales are close to the matrix resolution and to values observed for random matrices.

One may then expect large-scale features to be well captured by essential component analysis, while more local ones, such as loops, may project outside of the essential spaces.

By Fourier analysis one can thus estimate *a priori* the lengthscale of features that ought to be well captured by the essential components of a Hi-C matrix.

3.3 Comparison of Hi-C matrices

The abundance of Hi-C data available has recently shifted the focus from the analysis of the single matrices to the comparison of features present in different experiments.

The first objective of comparisons is to assess the reproducibility of experiments carried out on independent cell cultures. One can ask whether two experiments on the same cell-type lead to similar interaction patterns in the respective Hi-C matrices, to what degree they differ and whether different experimental procedures, such as using different restriction enzymes, can lead to different results.

The compatibility of experiments from the same cell-type has already been assumed in the previous sections, but only a thorough quantitative comparison can assure that this assumption really holds true.

Reproducibility is not only a theoretical question about the stability of the measured interactions, but also has important practical consequences: when two experiments are statistically compatible their interaction counts can be compiled to obtain a better sampled, higher resolution matrix [13].

The same tools used to assess the reproducibility of Hi-C matrices can be adopted to reveal interesting differences between two interaction maps, suggesting large conformational rearrangements in otherwise identical genomes.

Beyond reproducibility analysis, which is already challenging for complex data such as Hi-C matrices, lies the question of the structure of datasets: what are the relationships between Hi-C maps from different cell-types? Are they organized into clusters? And if so, are all clusters equally distant from each other or is a hierarchical structure present? Can Hi-C maps be represented on a lower dimensional manifold?

The next few subsections will address some of these questions.

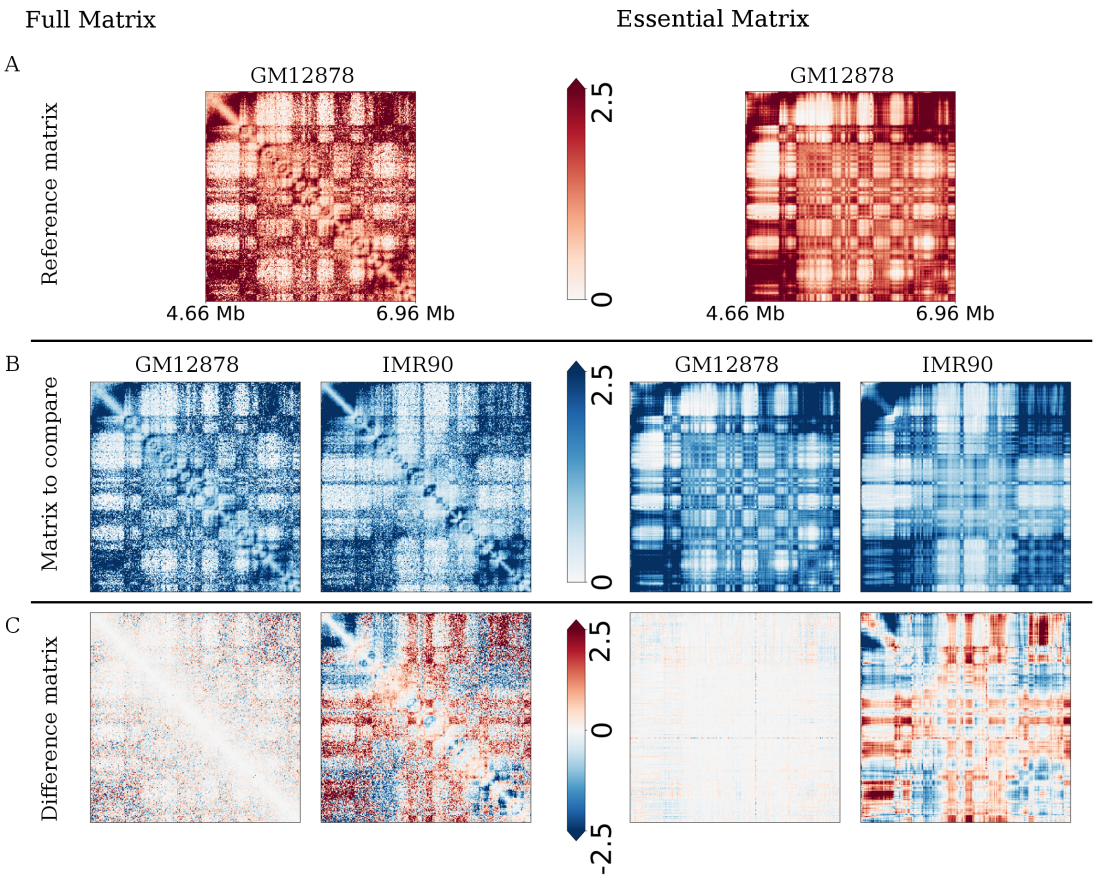

Figure 3.9: Visual comparison between various Hi-C maps of chromosome 17 (zoomed in for clarity). HIC001 (GM12878) is used as reference and compared with a biological replicate of the same cell-type, HIC002 (GM12878), and a non replicate, HIC050 (IMR90). The elementwise difference between the reference and the other matrices is shown in the bottom panels, both for the full matrices and for their essential components.

3.3.1 Visual inspection

Again, the first step in the analysis pipeline is to visually inspect the matrices, and differences between them, before and after the spectral filter is applied to obtain their essential components.

Figure 3.9 shows a zoom in of the comparison of HIC001 (chromosome 17, GM12878 cell-type) to biological replicate (HIC002, GM12878 cell-type) and a non replicate (HIC050, IMR90 cell-type), both for the full matrices and the respective essential maps.

The entry by entry subtractions of the pairs displayed in the bottom panel reveal sharper and more deeply marked differences between non replicates essential matrices. More importantly, the difference matrix between the two GM12878 replicates shows a uniform background in essential matrices, while speckled patterns are clearly visible for full matrices.

This is because of the statistical drop-out noise, a phenomenon in which some of the bins do not contain any interaction counts because the sampling depth of the experiment is not enough to reveal rarely occurring contacts between two loci. This is often the case for pairs of loci with a large genomic distance, so that long range interactions are more likely to incur this particular noise. Because the drop outs are different in each matrix, an entry-by-entry subtraction between the two matrices leads to the presence of large differences in unsampled bins: even if the difference in a certain area is zero on average, the measured local differences sum up to large discrepancies between the two matrices.

Since essential eigenvectors present slower fluctuations with respect to non-essential ones (as shown in figure 3.8), drop-out noise is less of an issue when comparing essential matrices. This leads to a higher degree of similarity between replicates, as seen in figure 3.9.

3.3.2 Metric distance

To quantify the difference between two experiments, I introduce a metric distance. First, by looking at the single chromosome matrices, one can use the Euclidean distance defined as

$$d(A,B) = \sqrt{\sum_{i,j}(A_{i,j} - B_{i,j})^2}. \tag{3.8}$$

This can be rewritten explicitly in terms of the eigenvectors and eigenvalues of the two matrices:

$$d^2(A,B) = \sum_{i,j}(A_{i,j}^2 + B_{i,j}^2 - 2A_{i,j}B_{i,j}), \tag{3.9}$$

where

$$\sum_{i,j} A_{i,j}^2 = \sum_{i,j} \left(\sum_k \lambda_k a_i^{(k)} a_j^{(k)} \right)^2 = \sum_{k,k'} \lambda_k \lambda_{k'} a_i^{(k)} a_i^{(k')} a_j^{(k)} a_j^{(k')}, \tag{3.10}$$

with $a_i^{(k)}$ being the i-th component of the k-th eigenvector of A, $\mathbf{a}^{(k)}$, and λ_k being the corresponding eigenvalue. Here, by using the fact that

$$\sum_i a_i^{(k)} a_i^{(k')} = \mathbf{a}^{(k)} \cdot \mathbf{a}^{(k')} = \delta_{k,k'} \tag{3.11}$$

one obtains

$$\sum_{i,j} A_{i,j}^2 = \sum_k \lambda_k^2. \tag{3.12}$$

Moreover, if $\mathbf{b}^k$ is the k-th eigenvector of B and μ_k its corresponding eigenvalue, one can write

$$\sum_{i,j} A_{i,j} B_{i,j} = \sum_{i,j} \sum_{k,k'} \lambda_k \mu_{k'} a_i^{(k)} b_i^{(k')} a_j^{(k)} b_j^{(k')} = \sum_{k,k'} \lambda_k \mu_{k'} \left(\mathbf{a}^{(k)} \cdot \mathbf{b}^{(k')} \right)^2. \tag{3.13}$$

Thus one can finally write down the original distance as

$$d^2(A,B) = \sum_k \left(\lambda_k^2 + \mu_k^2 - 2 \sum_{k'} \lambda_k \mu_{k'} \left(\mathbf{a}^{(k)} \cdot \mathbf{b}^{(k')} \right)^2 \right) \tag{3.14}$$

It is easy to see that by truncating the spectral sums over the eigenspaces at n^* one obtains the metric distance between the two essential matrices computed from A and B.

However, further analyses, which I will present in section 3.3.5 in this chapter, suggests to introduce a scaling factor on the spectrum in order to make the distance more robust with respect to n^*:

$$\bar{\lambda_k} = \frac{\lambda_k}{\sum^{n<n^*} |\lambda_n|}, \bar{\mu_k} = \frac{\mu_k}{\sum^{n<n^*} |\mu_n|} \tag{3.15}$$

so that the final metric distance between essential matrices is given by

$$d_{ess}^2(A,B) = \sum_k^{n^*} \left(\bar{\lambda}_k^2 + \bar{\mu}_k^2 - 2 \sum_{k'}^{n^*} \bar{\lambda}_k \bar{\mu}_{k'} \left(\mathbf{a}^{(k)} \cdot \mathbf{b}^{(k')} \right)^2 \right). \tag{3.16}$$

Once one can compute distances between intra-chromosomal interaction matrices, they can be combined in order to obtain the genome-wide distance between two experiments α and β

$$d_{ess}(\alpha,\beta) = \left[\sum_{n} d^2_{ess}(M_{\alpha,n}M_{\beta,n})\right] \tag{3.17}$$

where the subscript n indicates a sum over the chromosomes, and $M_{\alpha,n}$ is the intra-chromosomal matrices for chromosome n and experiment α.

3.3.3 Clustering

I compute the genome-wide essential distance between each pair of the 79 experiments which make up the dataset. In order to visualize the structure of the dataset, figure 3.10 shows a $3D$ multi-dimensional scaling (MDS) projection of the experiments. In this dimensional reduction scheme, each point, representing an experiment, is constrained to approximate as well as possible the distances measured with respect to all other experiments. Here I use MDS only to offer a spatial representation of the dataset: one can immediately see that experiments of with different cell-types, colored differently, form groups separated by gaps.

I compile the pair-wise distances in a distance matrix, presented in figure 3.11 both for full matrices and essential ones. A visual inspection reveals, again, that, when experiments are ordered according to their cell-type, blocks of similar experiments with small distances between each other appear on the essential distance matrix. These blocks are separated from each other by larger distances, suggesting natural groupings of experiments.

However this is not the case when looking at distances between full matrices, which do not seem to contain any reliable grouping.

The emergent groups seen in the matrix of essential distances can be examined through more quantitative analyses: here I employ a hierarchical clustering scheme with the Ward method, which not only allows to separate experiments in different groups, but also to study the hierarchical relationships between different clusters.

Along with the distance matrices, Figure 3.11 shows dendrograms which are the result of this clustering procedure: each leaf represents an experiment and is colored according to its cell-type; at each successive grouping, clusters are hierarchically linked to each other by minimizing the Ward distance between them, represented by the height of their connection in the dendrogram; by setting a threshold Ward score on the y-axis of the dendrogram, one can toggle the number of clusters. Natural grouping are easily detectable by looking for large jumps between two groups, which indicate a gap between them.

The dendrograms presented for full and essential matrices display striking differences.

First, although dendrograms are drawn using the same unit for the Ward score on the y-axis, the length of the branches for the essential matrices are more than twice as long as the ones for full matrices (2.5 maximum Ward score). This fact indicated that essential matrices account for more definite clusters compared to the full matrices.

The contents of the clusters are also very different. In full matrices, one can resolve the breast cancer cell lines (T47D), whose chromosomal aberrations reflect in large-scale

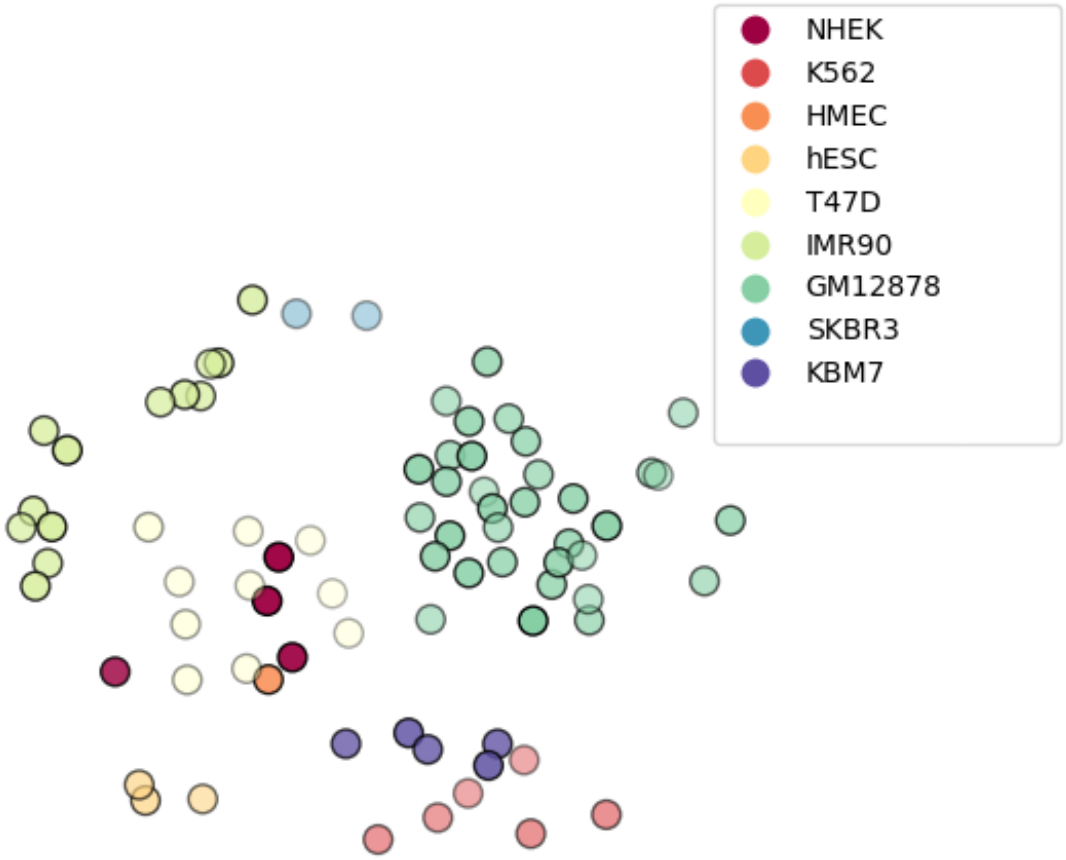

Figure 3.10: A $3D$ MDS projection of all experiments for visualization purposes: it shows an approximation of the spatial configuration taken by the experiments with respect to each other, according to the genome-wide distances computed on their essential components.

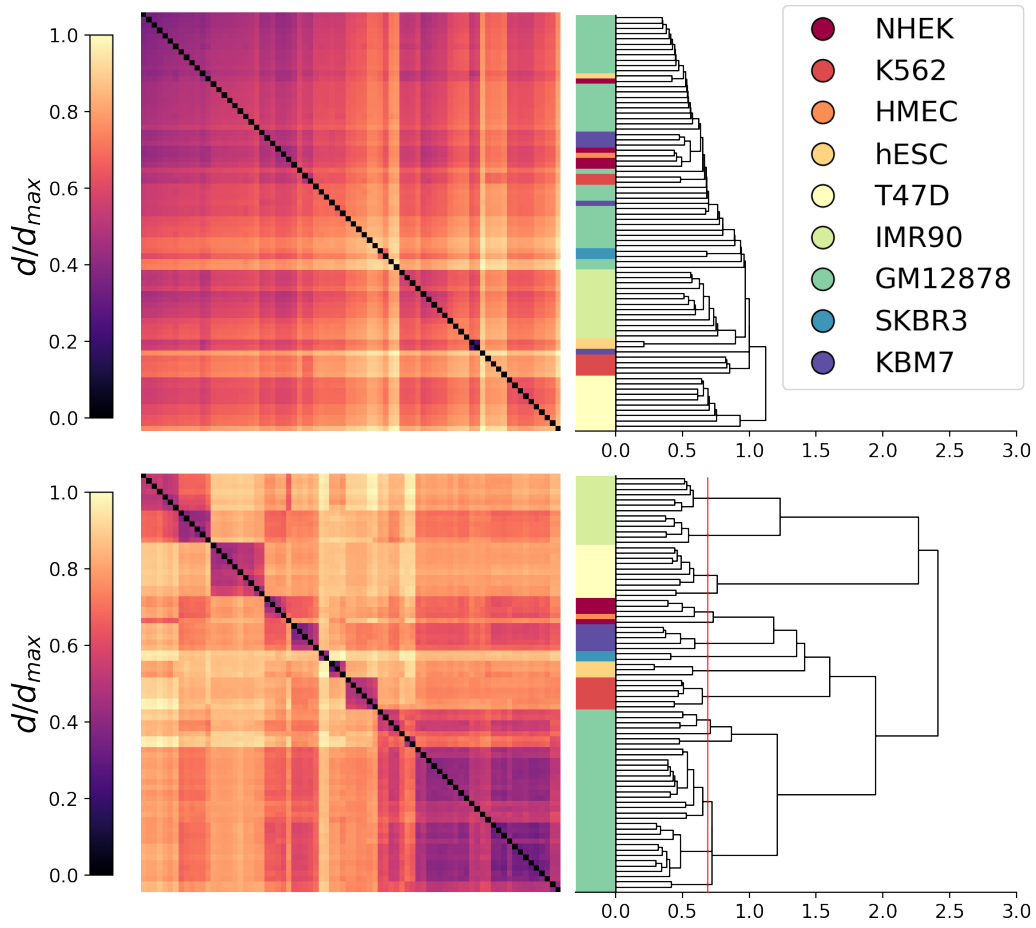

Figure 3.11: Clustering of different cell lines: pairwise genome-wide distance matrices between experiments for full and essential Hi-C maps are shown along with the respective Ward dendrograms. The pairwise distance matrices are normalized to the maximum. The red line refers to the optimal number of subdivisions decreed by the Dunn score.

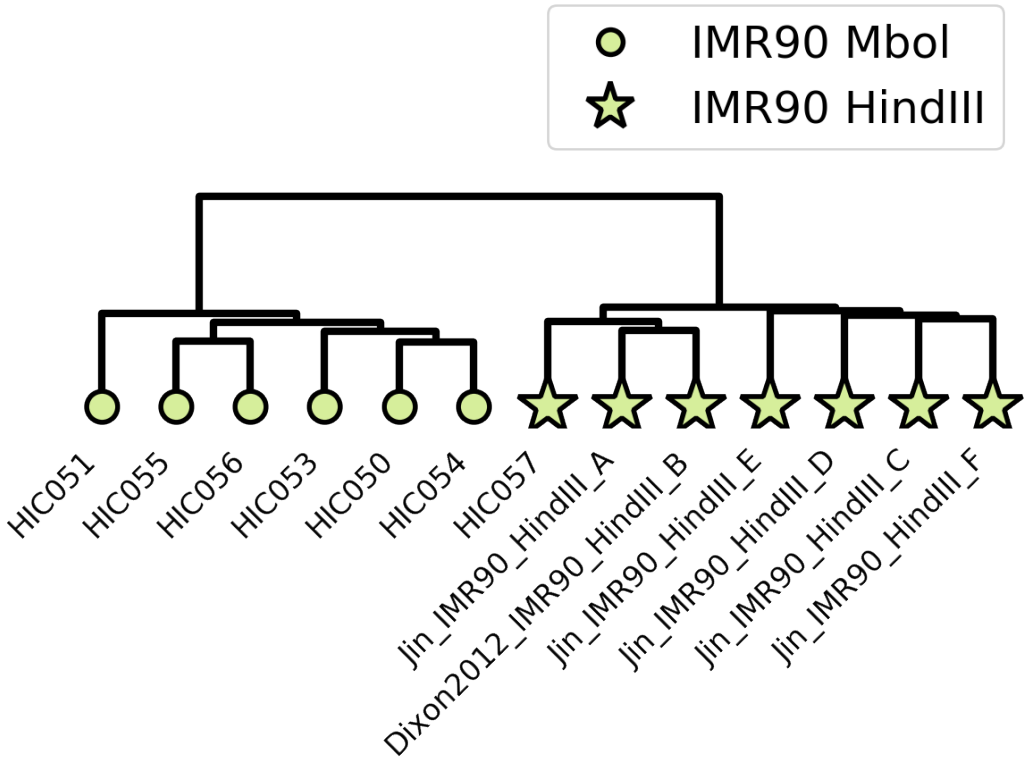

Figure 3.12: Section of the dendrogram in Figure 3.11 regarding the IMR90 cell line, which shows the correlation with the used restriction enzymes

conformational changes [92], also due to anomalous karyotype. However, other cell lines are poorly resolved, including the most numerous ones of GM12878 and IMR90.

On the other hand essential matrices return sharp subdivisions between different cell-types already from the first separations of the dendrogram hierarchy, suggesting deep gaps between the clusters.

The quality of these groupings can be quantitatively assessed by using the Dunn index, which computes the ratio between the intra-cluster distances and the inter-cluster ones. By applying this analysis for different groupings, one discovers that the optimal number of clusters in this dataset is 13, which is larger than the number of cell lines (9).

All clusters except one contain one cell-type only, however, while all experiments of cell-types K562, hESC, SKBR3, and KBM7 are contained within a single cluster, other cell-types ensembles display a richer internal structure: this is apparent especially for IMR90 experiments, which are sharply divided into two clusters, shown in Figure 3.12. Upon

closer inspection, one finds that this division is not an arbitrary one, but corresponds to different restriction enzymes and experimental techniques: one cluster contains only In-situ experiments using Mbol, while the other only dilution Hi-C experiments using HindIII.

Other subdivisions within cell-lines are not as clear-cut and cannot be explained only in terms of different experimental methodologies. However one interesting case is given by the only mixed cluster, containing NHEK and HMEC experiments: both cell lines are epithelial samples, which may explain their similarity, moreover all experiments within the cluster share the same methodology (In-situ, Mbol). On the other hand, the single isolated NHEK experiment is a dilution Hi-C and uses the HindIII reduction enzyme.

This shows that analysis of the essential components not only allows identifying different cell-types, but also hierarchically highlights differences due to experimental procedures within clusters of biological replicates.

3.3.4 Comparison to other methods

Biological replicates are expected to be close to each other, and further apart from non replicates: in the analysis above I showed that this is in fact the case, at least for essential matrices, and that clusters emerge naturally from the dataset once Hi-C matrices are cleaned from non-essential components.

Since the cell-types of the experiments are known, one can further quantify the discriminatory performance of metric distances and similarities scores by introducing receiving operating characteristic (ROC) curves.

These are obtained by considering all pairs of experiments, ordered according to their distances, and labeling them as biological replicates or non replicates depending on their cell-types. Then, by varying a threshold distance, one can count the amount of biological replicates (true positives) and non replicates (false positives) below that threshold. By computing the rates of these two quantities as the threshold increases and plotting them on the y and x-axis respectively one obtains a ROC curve which quantifies the ability of the employed method to discriminate between the two cases.

One expects, for a method with good discriminatory power, that the curve display a rapid increase where most, if not all, true positives are found before encountering any false positive, followed then by a plateau. On the other hand, a completely random discriminator, unable to categorize true and false positives, would trace a straight line bisecting the x-y plane. Notice that, since one is considering rates, rather than absolute numbers, the numerosity of True Positives with respect to False Positives does not matter.

Moreover, while a single number cannot fully capture the behavior of ROC curves, usually the quality of the result is summarized by the Area Under the Curve (AUC) score, which, as per its name, simply measures the surface of the plane which falls under the ROC curve. This is expected to be 0.5 for a completely random discriminator and 1 for a perfect one. This score can also be interpreted as the probability that an individual randomly chosen pair of biological replicates has a lower distance than a randomly chosen pair of non

replicates.

In Figure 3.13 I plot these ROC curves not only for essential and full matrices, but also for other methods.

The higher discriminatory power of essential matrices is conveyed by the ROC curves shown there. Full matrices yield an AUC parameter of 0.6, which only marginally improves with respect to the random reference (AUC=0.5). On the other hand, essential matrices reach a nearly optimal performance, with $AUC = 0.98$.

As further terms of reference Figure 3.13 shows ROC curves obtained by applying other methods to Hi-C matrices: a Gaussian filter, Hi-C Spector, and GenomeDISCO.

The first one computes euclidean distances between Hi-C matrices smoothed through a Gaussian filter, averaging neighboring bins in order to curb noise. This is a standard method used in other comparative studies in order to provide a measuring stick against which to pit more refined algorithms. Here I follow a previous study [93] in using a unitary standard deviation for the Gaussian kernel, corresponding to 100 Kb in terms of the genomic distances spanned by a single bin.

The other two methods, Hi-C Spector [12] and GenomeDISCO[13], are of interest here because they are also based on the use of spectral properties to compare Hi-C matrices.

The AUC values for the Gaussian filter, Hi-C Spector and GenomeDISCO are 0.90, 0.91 and 0.82, respectively, which are all significant. Some of these methods, including spectral ones, were purposely devised towards the comparative analysis of Hi-C matrices. The about optimal performance by the essential matrices is thus appealing as it is natively formulated as an enhancement method of individual matrices which can be adopted in comparative contexts too, as shown here.

3.3.5 Robustness of the AUC score

The discussion up to this point has not dealt with the problem of the robustness of the essential components analysis. One may ask what effect changing the value for n^* would have on the results presented thus far, or whether this analysis is able to capture the most important features at the same level if the resolution of the dataset is altered.

This section about the comparison of Hi-C maps offers a natural metric to study such issues: the AUC score presented above can in fact be used to summarize how well the essential component analysis fares in different situations when the parameter n^* is changed.

Figure 3.14 shows how the AUC relative to genome-wide distances performs as n^* is changed from 1 to 100 (on a logarithmic x-axis): while using only one eigenvector gives a worse than random AUC, as soon as one considers 2 or more eigenspaces the results become essentially stationary, reaching the optimality at $n^* = 5$ and experiencing only mild variation over the range considered here.

By only considering distances computed on chromosomes of similar lengths, one can extend the range of n^*, which would otherwise be limited by the presence of short chromosomes. Figure 3.15 shows that beyond $n^* = 100$ the AUC slowly decline, but still maintains

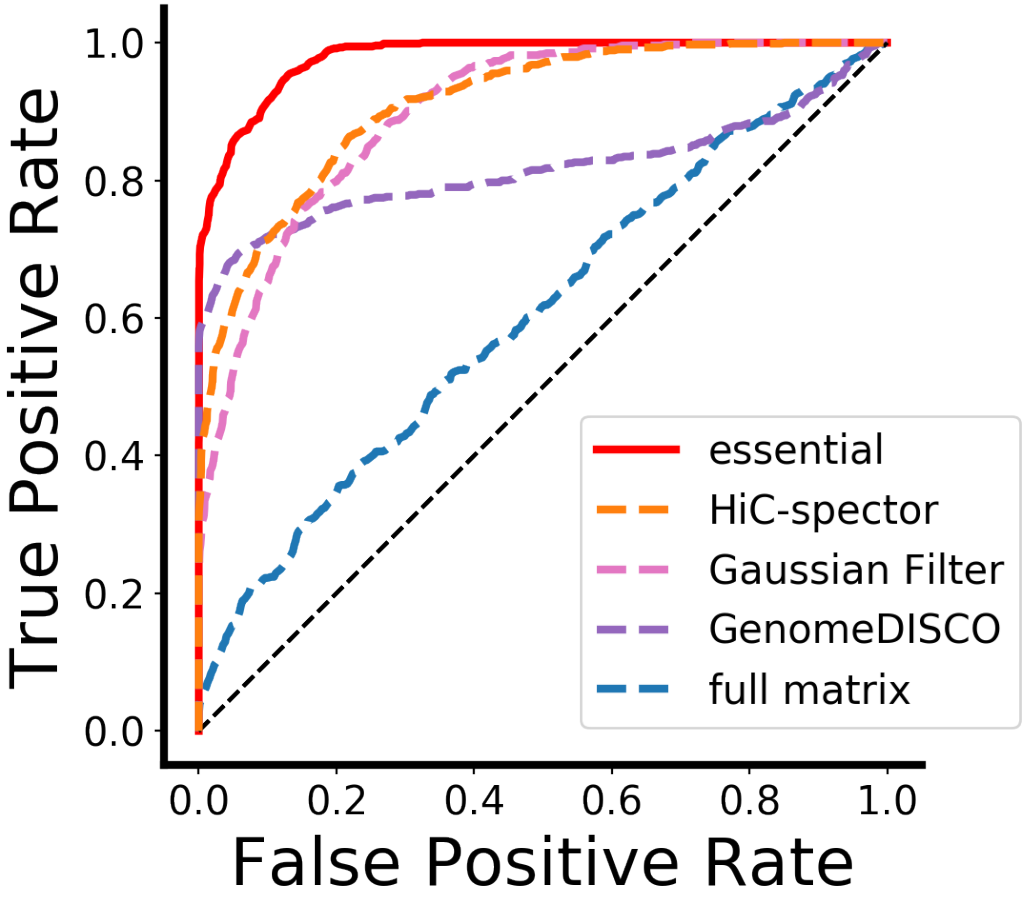

Figure 3.13: ROC curves for the full and essential matrices, along with other methods. The black dashed line indicates the random discriminator reference.

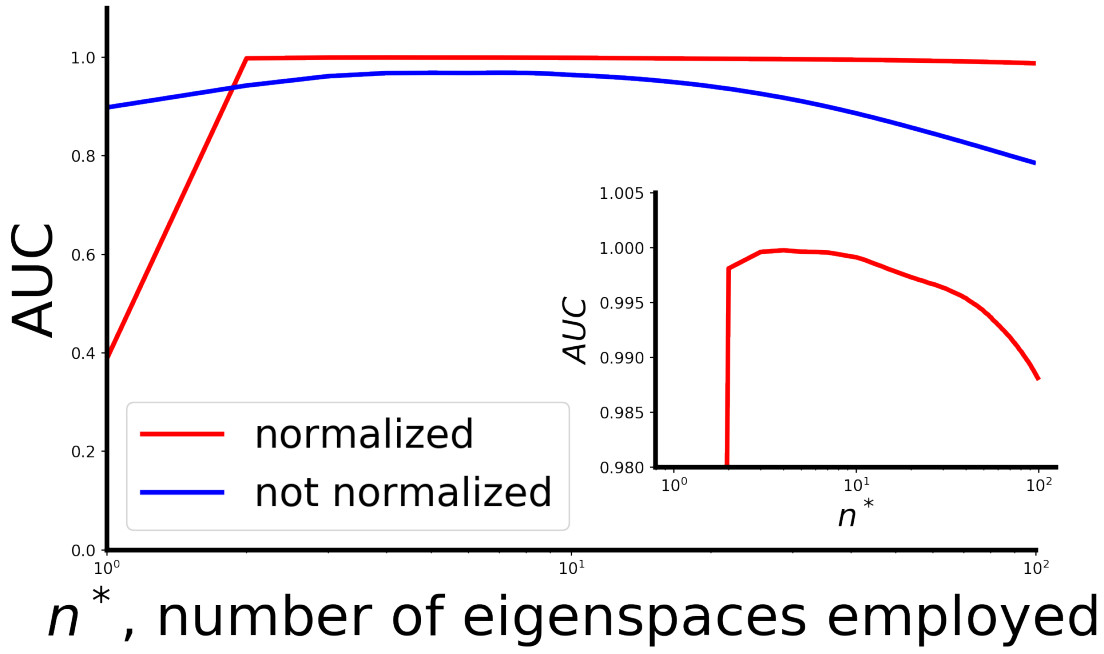

Figure 3.14: Area under the curve (AUC) for ROC curves at different values of n^*, based on the same analysis of Figure 3.13. The plot illustrates the improved n^*-dependent stability using the distance of eq. (3.17) with the eigenvalues rescaling compared to not using the rescaling (i.e. by setting $\bar{\lambda}_n = \lambda_n$. The inset shows a zoom of the ROC curve for the normalized case.

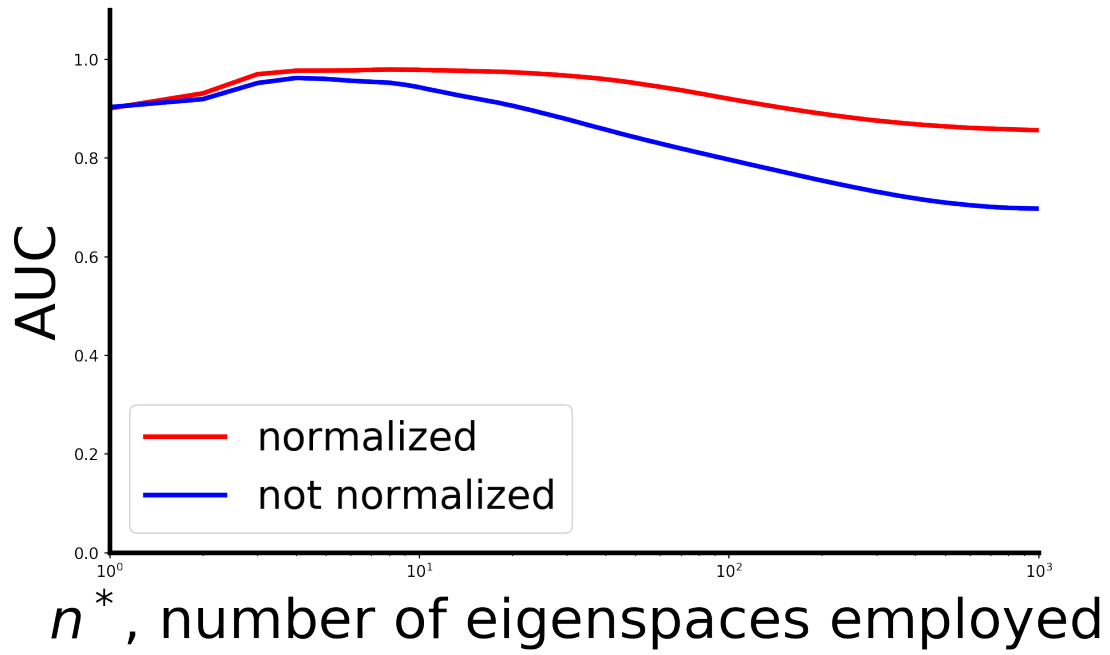

Figure 3.15: Area under the curve (AUC) for ROC curves at different values of n^*, based on the same analysis of Figure 3.13. The plot illustrates the improved n^*-dependent stability using the distance of eq. (3.17) with the eigenvalues rescaling compared to not using the rescaling (i.e. by setting $\bar{\lambda}_n = \lambda_n$. The inset shows a zoom of the ROC curve for the normalized case. In order to consider larger values of n^* only chromosomes 9 to 15, which have similar sizes, are considered.

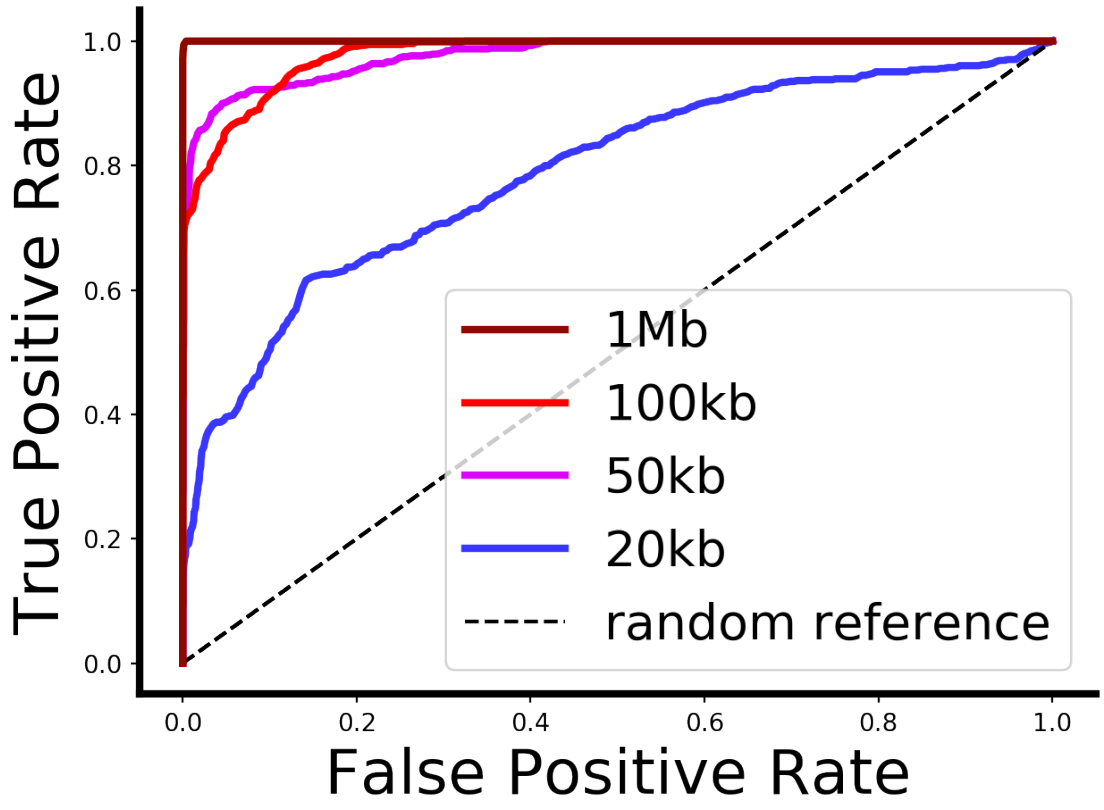

Figure 3.16: ROC curves for the essential matrices at different resolutions. The black dashed line indicates the random discriminator reference.

a significant value, larger than 0.9.

The same plots also shows the importance of the spectral normalization of eq. (3.15) in stabilizing the results, as comparing essential matrices obtained without using the normalized eigenvalues (blue lines) results in visibly faster decaying values of the AUC.

Another aspect of robustness that must be investigated is the effect of changing the resolution at which Hi-C matrices are analyzed. While the resolution in itself, i.e. the size of the matrix binning, can be seen as an independent variable, it is in fact linked to the sequencing depth, the number of interactions sampled in the experiment: the sizes of the patterns that can be reliably resolved can be lowered only if the matrix is well sampled. In a heterogeneous dataset it is especially difficult to obtain meaningful comparisons, as some matrices only allow for low resolution analyses, while others could offer much more information at higher resolutions.

Figure 3.16 shows the ROC curves for different resolutions: $1Mb$, $100Kb$, $50Kb$, and $20Kb$. As the resolution increases, i.e. when using smaller binning sizes, the performance decreases: $1Mb$ comparisons are able to distinguish between different cell-types almost perfectly, while $100Kb$ and $50Kb$ matrices still retain a significant discriminatory power. On the other hand, a visible drop happens at $20Kb$, which is nevertheless a significant improvement with respect to the element by element comparison.

3.4 Single-cell Hi-C

An entirely distinct context is that of single-cell Hi-C analysis, which deals not with matrices whose interaction counts are compiled from all cells in a sample, but rather come from a single cell.

This allows one to obtain thousands of interaction maps, each depicting a single configuration frozen in time, instead of the ensemble average of multiple configurations one would get for bulk Hi-C maps. Moreover, even within the same cell population, single-cell maps can be divided into groups depending on the phase of the cell-cycle each individual is in: G1, S, G2, and mitosis phases (see figure 3.17) are all characterized by large conformational changes of the genome. This is most striking near mitosis, when chromosomes coalesce into the characteristic rod-like shape before replication.

Of course this has a huge effect on the patterns found in Hi-C maps at different stages of the cell-life. Contact maps of mitotic chromosomes, for instance, do not display the usual checkerboard patterns. Instead, the interactions become dominated by their dependence on the genomic distances: along with the main diagonal, describing interactions between neighboring loci along the chain, a secondary diagonal appears, located at a distance of $\sim 3Mb$ [69].

Methods that perform assays at the single-cell level have helped improve the understanding of many aspects of genome organization, but they also offer additional challenges for the analysis of the results: for example, as seen in figure 3.22, single-cell matrices are

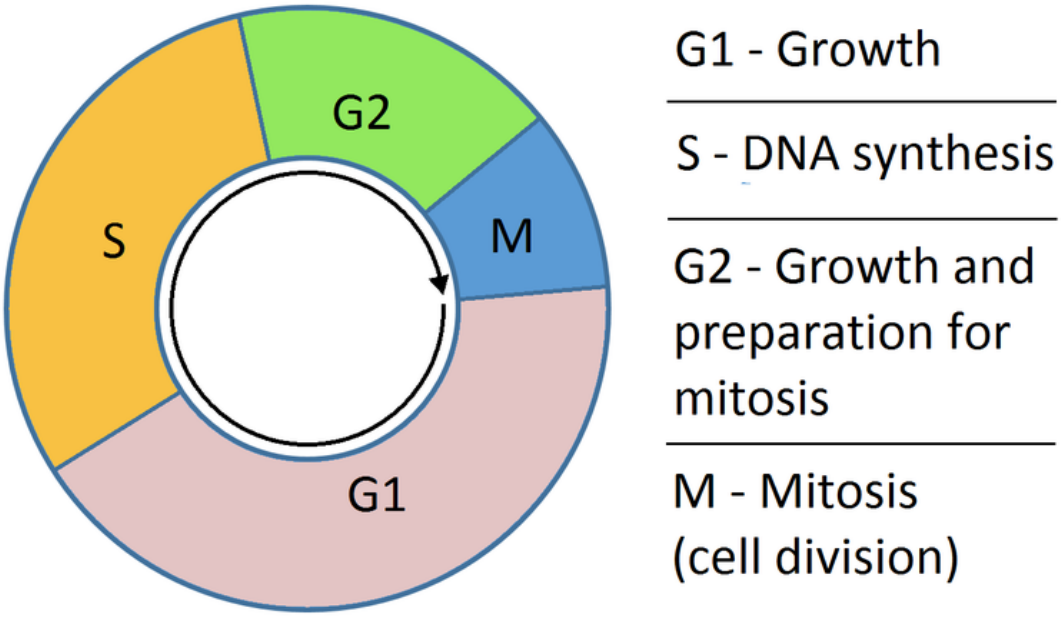

Figure 3.17: The four phases of the cell cycle. G1 - the intitial growth phase. S - the phase in which DNA is synthesised. G2 - the second growth phase in preparation for cell division. M - mitosis; where the cell divides to produce two daughter cells that continue the cell cycle. Credits to Simon Caulton, CC BY-SA 3.0 https://creativecommons.org/licenses/by-sa/3.0, via Wikimedia Commons

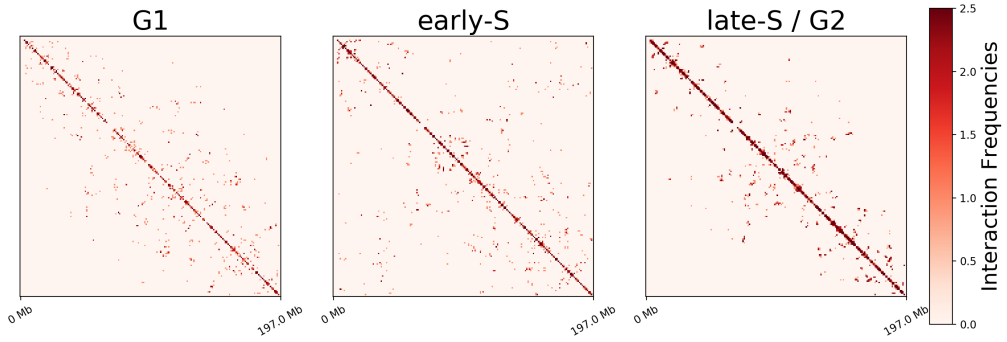

Figure 3.18: Three single-cell Hi-C matrices relative to chromosome 1 of experiments NXT-972 (G1 phase), NXT-2244 (early-S phase), and NXT-2126 (late-S / G2 phase). The matrices are not normalized.

much sparser than bulk matrices.

In this section I will apply essential component analysis to the dataset obtained by Nagano et Al. [18], which covers different cell-cycle stages of the mESC mouse embryonic cell line.

In the original paper, Nagano et Al. [18] time-order the dataset *a posteriori* through an elegant dimensional reduction procedure, necessary to extract meaning from the inherently sparse matrices. Here I will show that clustering obtained from the essential component analysis can recover the same groupings of the original paper.

3.4.1 Time ordering in the original paper

Here I briefly recount the method used by Nagano et Al. [18] to order cells along the cell-cycle.

To devise their method, they proceed from the informed observation of the behavior of chromosomes during the cell-cycle. Mammalian mitotic chromosomes are rod-shaped and previous analyses of the contact maps obtained from M phase cells has revealed that intra-chromosomal contacts are enriched for genomic distances which ranges between 2 Mb and 12 Mb [67]. Contacts in this region are dubbed *mitotic* contacts.

A comparison of the percentage of mitotic contacts and short-range contacts (those which happen between loci at genomic distances smaller than 2 Mb) shows a circular pattern which reflects the position of each cell along the cell-cycle. On the basis of this observation, Nagano et Al. argue that a gradual remodeling of the chromosomal conformations takes place during the cell life. Hence chromosomal conformations can be used to phase cells at various stages of their life.

Moreover they observe that early- and late-replicating topological domains are defined by their copy number dynamics during S-phase. Normalized TAD coverage across the single-cell dataset reflects a strong correlation between domains previously annotated as earlier late-replicating in mouse ES cells [94]. They define a *repli-score* based on the copy-number ratio of early-replicating regions to total coverage for each cell. This score is expected to have a low value for G1 and G2 cells approaching mitosis, with a peak for cells in the S-phase.

Combining repli-score with the circular pattern exhibited by the frequencies of short-range and mitotic contacts they are able to retrieve exactly the expected trajectory. Together, the mitotic signatures and time of replication analysis define two major anchors to support phasing of the entire single-cell Hi-C dataset along the cell cycle [18].

Figure 3.19, taken from the original article by Nagano et Al., shows the details of their work, from the experimental procedure to the time ordering of experiments, and finally to pooling Hi-C maps of cells in the same phase.

Their time ordering serves as the gold standard for my subsequent analysis.

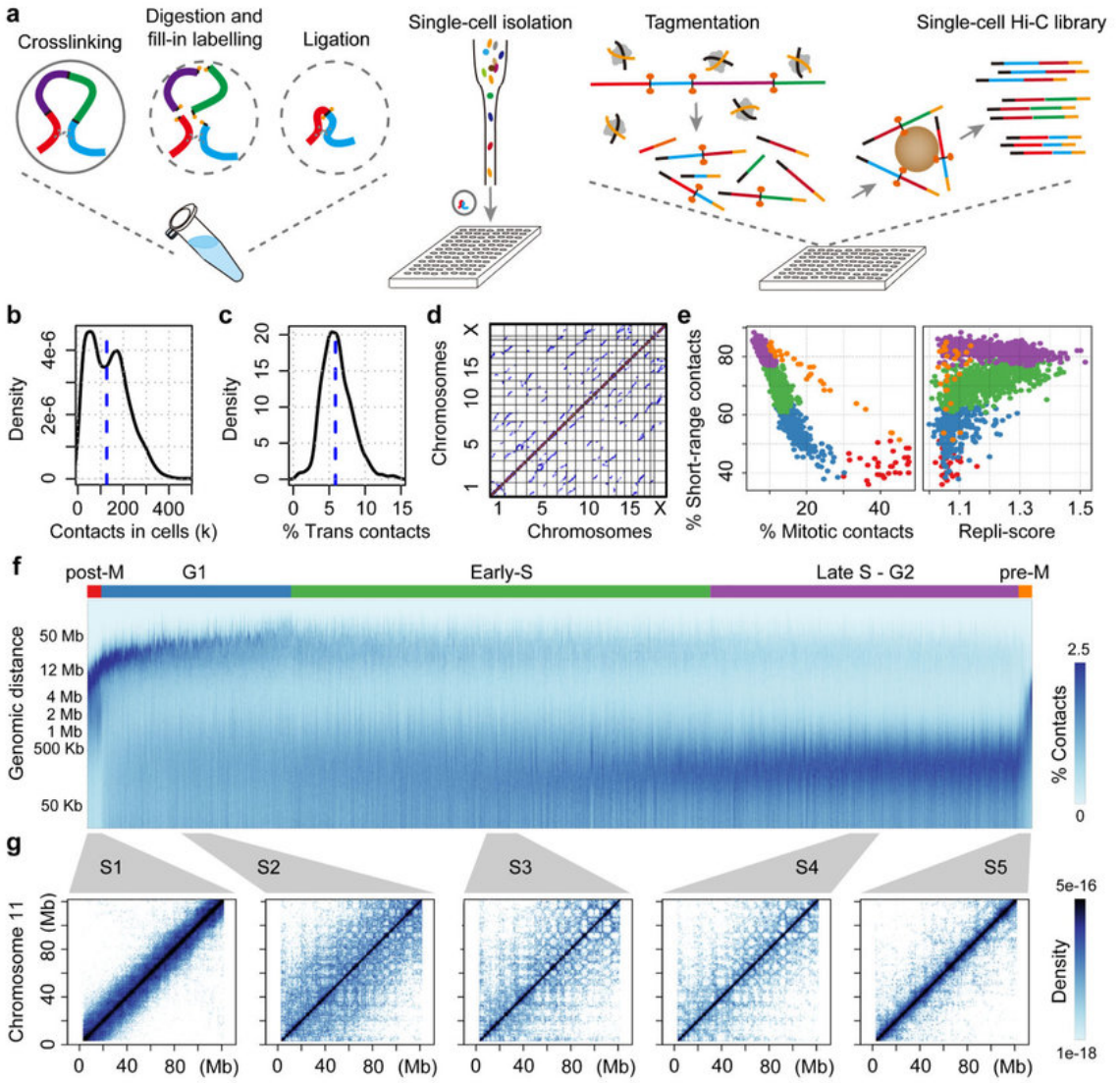

Figure 3.19: **Taken from Nagano et Al.** [18]. **a**, Single-cell Hi-C schematic. **b**, Number of informative contacts retrieved per cell that passed the quality control filter. Median 127,233 (dashed line). **c**, Percentage of trans-chromosomal contacts per cell that passed the quality control filter. Median 5.87% (dashed line). **d**, Genome-wide contact map of a representative mitotic cell (1CDX4_242). **e**, Percentage of short-range ($\leq$ 2Mb) versus mitotic band (2–12Mb) contacts per cell (left), and repli-score (right). Cells are grouped by percentage short-range and percentage mitotic contacts and coloured by group. **f**, Single-cell contact decay profiles ordered by in silico inferred cell-cycle phasing, with approximate cell-cycle phases shown on top. Each column represents a single cell. **g**, Selected phased and pooled contact maps

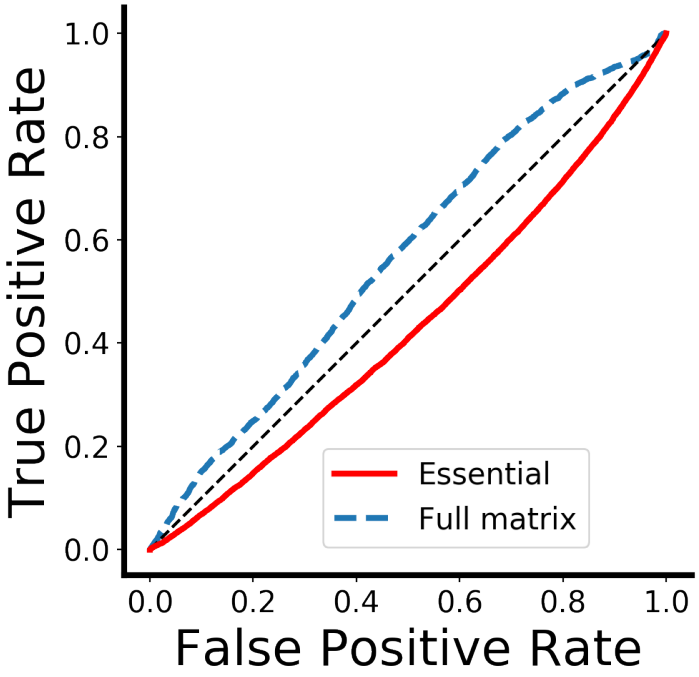

Figure 3.20: The upper panel shows Roc curves for full and essential single-cell Hi-C maps, the dashed line represents the random discriminator reference. The lower panel shows the correspondence of the clustering (3 groups) to the time-ordering performed by Nagano et Al. [18].

3.4.2 Application of essential component analysis

Here I apply the essential component analysis to the single cell dataset and compute distances between experiments. I adopt Nagano et Al. labeling of experiments into G1, early-S, late-S/G2 phases as the gold standard (disregarding pre-M and post-M phases, as each of these only contains one viable member).

For each raw matrix in the dataset I compute the observed over expected normalization in the same way I did for bulk matrices, and from there I extract the first 10 essential spaces. I compile the genome-wide pair-wise distances into a distance matrix from which, using the labels defined above, I compute the ROC curve. The result is shown in figure 3.20: in the case of single cell Hi-C maps, the discriminator used for bulk matrices is completely ineffective, scoring an $AUC = 0.43$, which is not only lower than the result for full matrices ($AUC = 0.55$), but also worse than the random case.

Why is this the case?

Not only single cell Hi-C matrices are much sparser with respect to their bulk counter-

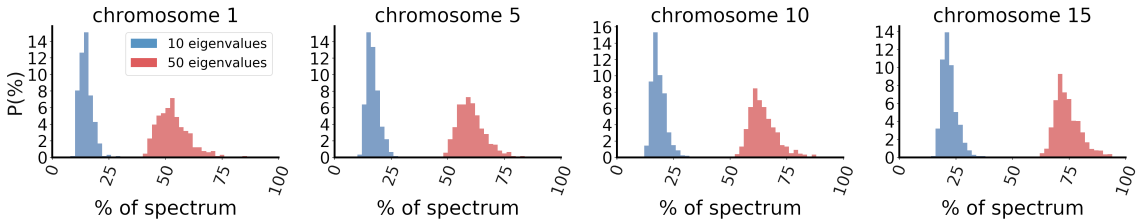

Figure 3.21: Percentage of the spectrum covered by the first 10 (blue) or 50 (red) eigenvalues. The histograms display the distribution over the population of single cell experiments contained in the dataset.

parts. The different stages of the cell-cycle observable in single-cell Hi-C are characterized by a variation in the dependence of the interaction counts on the genomic distance [18, 69]. This means, for example, that mitotic chromosomes are characterized by the presence of a secondary diagonal of enhanced contacts and a stronger depletion of long distance contacts with respect to interphase chromosomes [69]. Because it carries relevant biological information, one cannot discount the dependence on the genomic distance when analyzing single-cell Hi-C data.

Hence, I adapted the original essential component analysis in two ways: first, I apply it to raw single-cell matrices, without applying the usual OoE normalization. This makes it more difficult to capture the relevant information with just 10 eigenspaces, so the second adaptation is to perform the analysis with $n^* = 50$. This allows one to capture most of the trace of the sparse and non-normalized single-cell Hi-C matrices, as shown in figure 3.21.

Re-computing the distances with these adaptation allows me to obtain the ROC curves in figure 3.22. The set of full single cell Hi-C matrices cannot be clustered in a meaningful time-ordered way, as shown by the near-diagonal trend (blue line) in the ROC plot of Figure 3.22 (with an AUC of 0.55). On the other hand, the essential matrices display a noticeable and significant improvement, $AUC = 0.68$. Indeed, the same metrics and clustering procedure shown in figure 3.11, adopted for the bulk dataset, returns primary partitions that are in very good accord with the time-ordered cellular stages proposed by Nagano et Al. [18].

This shows that, with some adaptations, the essential component analysis is able to capture the partitions obtained by Nagano et Al. [18] by directly comparing matrices.

3.5 Summary and Conclusion

Essential component analysis is different from approaches addressing data noise in Hi-C matrices at the local or bin-wise level. The latter are in fact designed to remove biological biases [95, 75] or numerical imbalance [7, 96] from Hi-C matrices at a local scale, while the essential component analysis presented in this work discounts the non-specific component

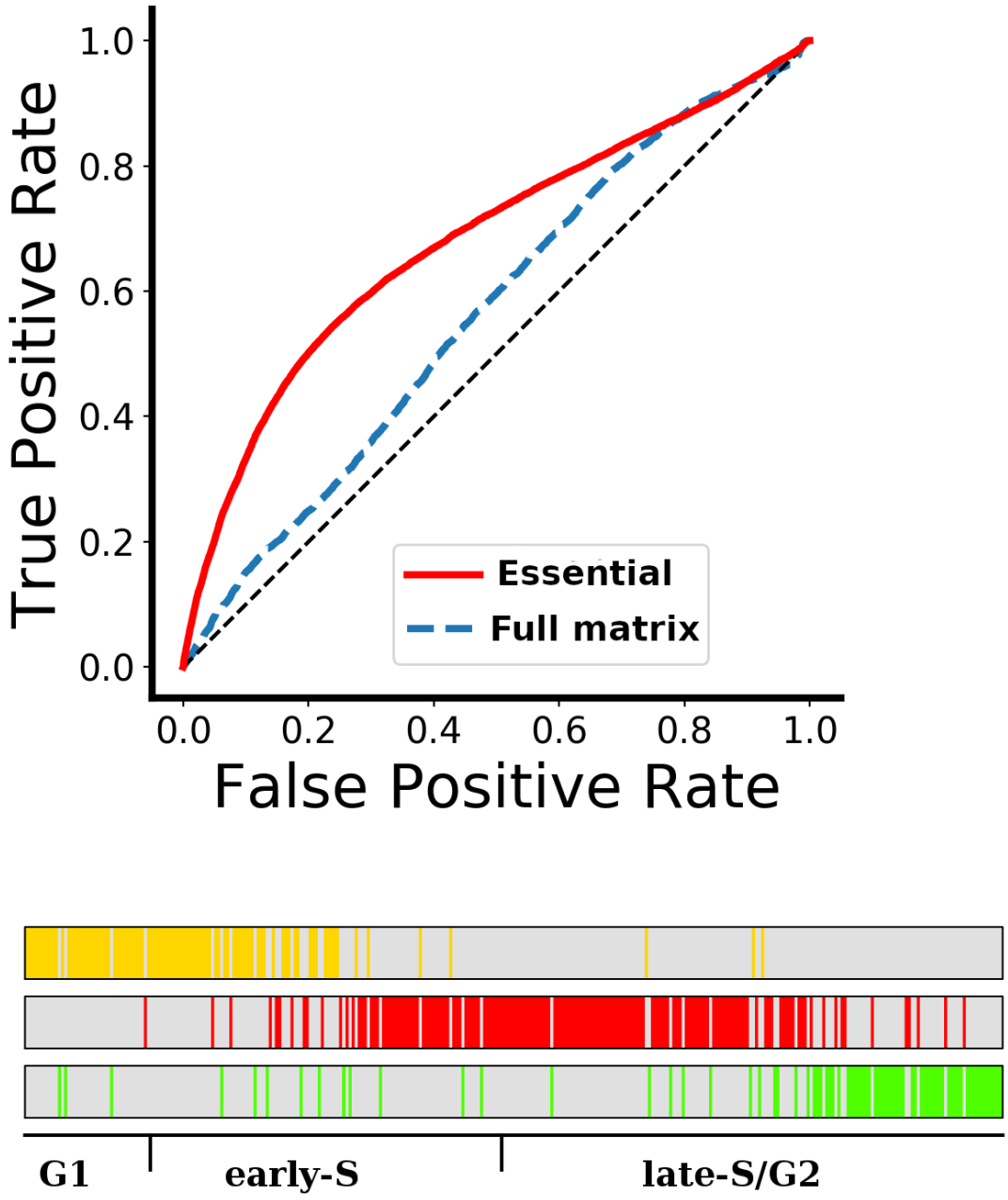

Figure 3.22: The upper panel shows Roc curves for full and essential single-cell Hi-C maps, the dashed line represents the random discriminator reference. The lower panel shows the correspondence of the clustering (3 groups) to the time-ordering performed by Nagano et Al. [18].

by isolating the spectral, hence generally global, properties that differ from those of random matrices.

The resulting enhanced specific content of the essential matrices is illustrated by the clearer and sharper features which emerge once the aspecific part is removed, and, more quantitatively, by the comparison of different instances of Hi-C matrices from biological replicates or different cell-types. The subtraction of matrices of the same cell line is noticeably more uniform and less noisy for the essential matrices compared to the full ones, with the distribution of entries being closer to the expected value of zero. In addition, the subtraction of the essential matrices of different cell lines provides a clearer highlighting of the different features, which are instead convolved with noise in full matrices.

I presented two applications of the essential component analysis, chosen for their relevance and challenging nature. Firstly I compared full Hi-C matrices obtained at high sequencing depth, with matrices at lower depth, both in the full and essential forms. The comparison demonstrated that essential matrices can significantly boost the correlation with the highest depth reference matrix and allow for much more consistent identification of TADs. In fact, essential matrices can retain roughly the same level of accordance with the high resolution matrix (in both applications) despite the decreasing sequencing depth. These results provide a striking illustration of the significant potential that the essential component analysis holds for isolating specific interaction features that would require a major increase in sequencing depth to be discerned in full matrices.

Secondly, I carried out the unsupervised clustering of a heterogeneous ensemble of Hi-C matrices covering several cell lines. Good correspondence of cell lines and the subdivisions obtained from the hierarchical clustering are observed only for essential matrices, with little evidence of structure found in full matrices. Furthermore, subdivisions based on essential matrices of the IMR90 cell lines correlate with different restriction enzymes used in the Hi-C assays for the two subsets: this unexpected result shows that different experimental conditions can reflect in contact probabilities which are sufficiently distinct to be picked up by the analysis of the essential components, despite being subtle enough to bypass more naive methods, such as the bin by bin comparison of full matrices.

Overall, the results show that essential matrices are better suited than full ones to isolate significant contact patterns, which ought to be useful also in contexts where contact propensities are used for chromosome modeling both to generate mean-field genome structures [97, 98] or to highlight the cell-to-cell variability [59, 99].

All the above results are obtained with a rule-of-thumb value for the number of essential spaces to consider, setting the parameter $n^* = 10$, but further analysis shows that, in part thanks to the normalization of the spectrum, these findings are very stable, allowing one to more confidently use this tool in different contexts (different chromosomes, or different datasets) without worrying too much about fine tuning.

Finally, to illustrate the perspective potential of the essential component analysis I discussed a preliminary application to single-cell matrices, focusing on the dataset obtained from Nagano et Al. [18]. The ROC curves show that the time ordering presented in the

original paper cannot be recovered from full scHi-C matrices. This is consistent with the fact that a dimensional-reduction of scHi-C matrices was needed in order to obviate to the sparsity of the full matrices and establish their time-ordering.

Therefore, it is significant and appealing that, once the aspecific parts of the matrices are discounted by the essential component analysis, a clear correlation with the time ordering of Nagano et Al. emerges, and the main cellular phases are recovered. This suggests that essential component analysis may be a beneficial tool to use in conjunction with more fine-tuned ones, like the dimensional reduction scheme presented in the original paper, in order to obtain information about the dataset which would not be available otherwise (e.g. the bin-by-bin comparison of matrices is not possible once they pass through the dimensional reduction tool, which only considers the average interaction rate as a function of the genomic distance).

More in general, these results emphasize the advantages offered by the essential component analysis across very different contexts.

The results open numerous perspectives for using essential component analysis to optimally isolate biologically and physically-relevant information from Hi-C matrices. Beyond the applications considered here, one can expect this tool to be useful in comparative contexts where variations of chromosome compartmentalization could be picked up with enhanced reliability and hence better related to epigenomics changes [100] or cell differentiation [10, 101, 102].

In addition this approach could be useful when simultaneously analyzing large ensemble of matrices, where their encoding into a limited number of essential spaces can serve as a lossy data compression scheme of retaining the most important features.

The tool is freely available for academic use as the essHi-C software package and can be accessed at https://github.com/stefanofranzini/essHIC

Chapter 4

Dimensional Reduction and Lossy Compression of Hi-C Matrices

In the last chapter I presented the results of applying a spectral filter to Hi-C matrices in order to remove their aspecific component: not only do essential components behave as an high quality matrix, they also allow for better comparison between different experiments in order to assess reproducibility and relevant differences. However, one compelling aspect I did not investigate is the ability of essential component analysis to describe each Hi-C map using only a limited amount of degrees of freedom.

Essential component analysis obtains a reduced representation by retaining a limited amount of eigenspaces. However the eigenspaces of each matrix are different: hence one may ask whether it is possible to obtain a dimensional reduction based on common properties shared by all Hi-C maps. Aside from the scientific curiosity about recurrent patterns of Hi-C maps and the optimal number of dimensions at which they can be represented, answering this question may also lead to operative advantages: a method to encode Hi-C matrices with a few degrees of freedom, while also offering a way of restoring the original data, would serve as a lossy compression algorithm tailored to Hi-C maps. This would potentially allow to increase the volume of data that can be handled during analysis.

Such endeavor is predicated on the premise that, at the local scale, patterns contained in one Hi-C map are similar to those found in different chromosomes or cell-types [20]: this means that only a small set of building blocks can be arranged in order to obtain all the variability found in interaction maps at a global level. If this assumption is true, once an algorithm is able to encode and decode the local patterns obtained from a small set of matrices, it can be applied to other datasets without further tuning.

In order to obtain such an algorithm, I turn my attention to a set of tools that are complementary to the spectral methods (which are based on linear algebra): machine learning techniques offer a number of network architectures which are singularly apt for manifold learning and have already been used with a high degree of success in the context

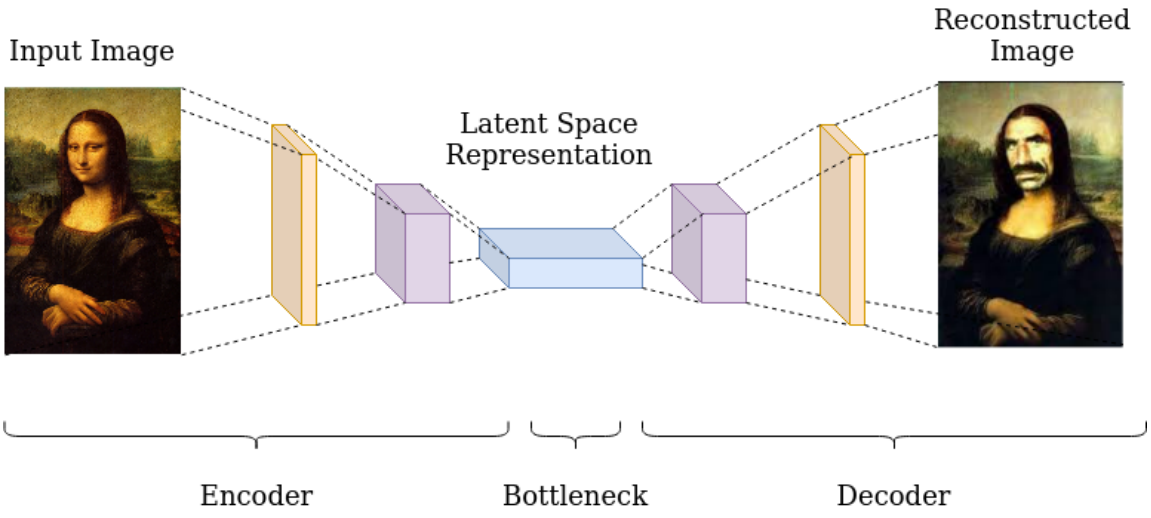

Figure 4.1: Scheme of an autoencoder: a network which passes data through a bottleneck layer at a reduced dimensionality and then tries to reconstruct it on the other side. Reconstructions may present some minor distortions.

of Hi-C matrices [93, 103, 104], but with different objectives with respect to the one of dimensional reduction and data compression. Nevertheless the performance of networks on tasks such as quality enhancement shows that they have the ability of learning the probability distribution underlying repeated patterns present in Hi-C matrices [93, 103, 104].

An architecture which qualifies for the job is that of the autoencoder[105], a network which passes data through a bottleneck layer at a reduced dimensionality and then tries to reconstruct it on the other side (see figure 4.1 for a sketch). This procedure ensures that the encoded data only retains the most important information about the original matrix. Here I present a version of this architecture tweaked to better address the problems found in Hi-C maps and test both its ability to retain information about the original matrix and the quality of the decoded one.

The idea is to split matrices into small cut-outs, containing local patterns, that can be passed through an autoencoder in order to learn a low dimensional representation and an inverse transformation. Doing so will allow to generalize the results obtained on a limited dataset to other cell-types or chromosomes on which the neural network has not been trained.

The chapter is organized as follows: first I will address the problem of dimensional reduction from a theoretical perspective by examining the clustering of small cut-outs sampled from Hi-C matrices of different chromosomes, containing local portions of their interaction patterns. This is done both to investigate the optimal dimensionality of the

dataset and to make sure that patterns from different sources are compatible at the local level, so that the compression algorithm can be generalized. Next I will revise the methods that can be used for this task and detail the architecture of the autoencoder employed in this work. The following sections are devoted to presenting the results of the algorithm and examining the structures found in the encoded and decoded data, with a focus on the ability of the algorithm to retain information and the consequences of the lossy compression. Finally, I will compare matrices reconstructed from the encoded space through the compression algorithm and the essential components of the original matrices.

4.1 Dimensional Reduction of Local Patterns

An implicit assumption often used when looking at Hi-C matrices is that recognizable local patterns get repeated in different contexts[20, 2]. Every chromosome and cell-type has its own typical global structure, leading to different interaction maps that can be compared and distinguished from each other. However, if one looks at small scales, some structures, such as loops, or TADs, or the boundaries between compartments, can be found in different positions in many Hi-C matrices sampled from different cell populations. Hence, these local patterns can be thought of as building blocks that can be rearranged to obtain all global combinations encountered in interaction maps.

This is, of course, a plausible assumption. Many methods have been used to define and then spot specific patterns on Hi-C maps in order to describe their structural properties [2]. Here however the concern is less on the definition of some specific local pattern and more on the possibility of finding a compact description of repeated motifs in order to perform dimensionality reduction at the matrix level.

In this section I will explore the properties of local patterns found in Hi-C matrices. First I will ask whether a dataset made up of these local patterns (in the sense that I will discuss below) contains any regularities that allow for an efficient dimensionality reduction. I will then perform an analysis to obtain an estimate for the intrinsic dimension of the dataset, i.e. the dimensionality of the manifold upon which the dataset rests, which is the optimal number of degrees of freedom that describe the data points. Furthermore, I will compute clusters and verify that patterns sampled from different chromosomes and cell-types are not separated by gaps, so that the results obtained from a limited dataset can be generalized to other matrices.

4.1.1 The Dataset

Before proceeding I need to define what exactly constitutes the dataset of local patterns. Here I simply consider square cut-outs sampled randomly from Hi-C matrices, KK bins wide at 100 kb resolution, with $K = 50$. For each of cut-out present, its transpose is also added, in order to equilibrate the dataset.

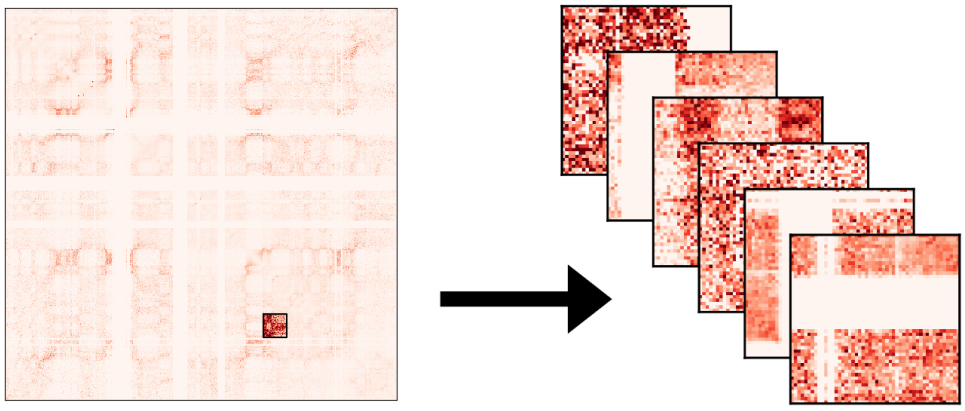

Figure 4.2: Scheme of the sampling of local pattern: a 50×50 square block is selected randomly from a matrix and added to the dataset. Examples of the sampled patterns are plotted on the right.

Figure 4.2 shows a scheme of the sampling procedure and examples of the look of the members of the dataset.

The matrices used to build the dataset are the same bulk Hi-C maps considered in the previous chapters, containing experiments from 9 different types of human cells. Since one of the objectives is data compression, only the raw matrices are considered, as normalization, unbiasing [7] and other analyses can be performed downstream.

Finally, only chromosomes 1, 2, 3, 18, 19, and 20 are used to obtain the cut-outs of the dataset, in order to be able to perform later analyses on other chromosomes and verify whether the methods developed in this work can be successfully generalized.

By sampling 40 cut-outs per matrix, the dataset contains (not counting the transpose cut-outs) 18720 data points.

4.1.2 Intrinsic Dimension

Dimensionality reduction is an operation which transforms data from a higher dimensional space into a lower dimensional one while preserving some meaningful properties of the original dataset, such as distances, or qualitative groupings. Generally, dimensionality reduction is meaningful if the features of the data points display some correlations, often non-linear, so that the dataset effectively lies on some manifold embedded in the

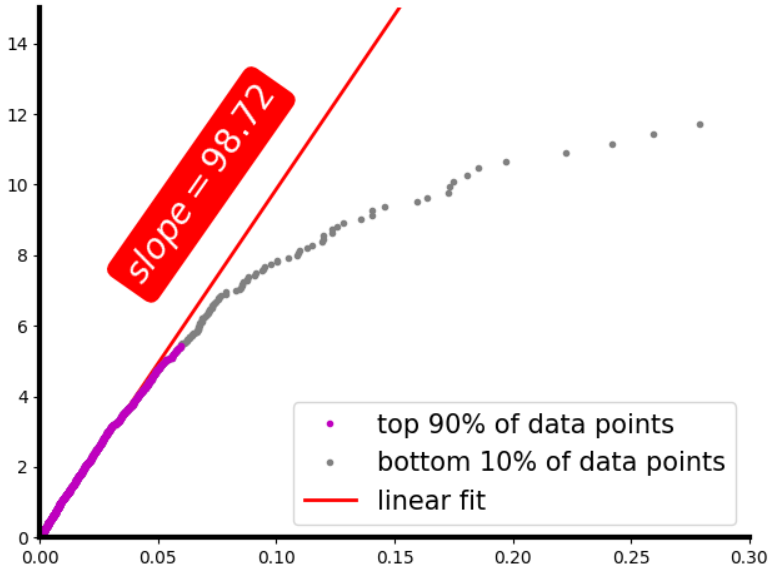

Figure 4.3: Intrinsic dimension analysis of the cut-outs dataset: the linear fit is performed on the top 90% of the data points, while the remaining 10% tail is discarded.

high-dimensional space. Then the problem of dimensionality reduction becomes that of describing this manifold and finding a representation of the original data-points in terms of the natural coordinates of that manifold.

For example one could consider a dataset of points lying on the surface of a sphere in three dimensions: while each point is described by three coordinates, one can also use two angles in spherical coordinates to obtain the same information with fewer degrees of freedom. This is also an example of a manifold with a non-trivial topology, as the sphere is only locally homeomorphic to a plane, but not globally.

For now my objective is not to obtain a low-dimensional representation of the data points (this will be done later by applying the autoencoder network), but simply to find out whether dimensionality reduction is possible on the dataset and to what degree. The intrinsic dimension can be thought of as the minimal number of features (or coordinates) needed to represent the data: this is akin to asking the dimensionality of the manifold upon which the data points lie. In the above example, the intrinsic dimension is 2, the number of coordinates needed to describe a point on the surface of a sphere.

Ideally, the intrinsic dimension represents and estimate of the optimal efficiency of a compression: using fewer features to describe the dataset may end up removing some important information about its structure.

To estimate the intrinsic dimension I use the method described by Facco et Al. [21], which requires only for the density of the data points to be locally (within the distance of

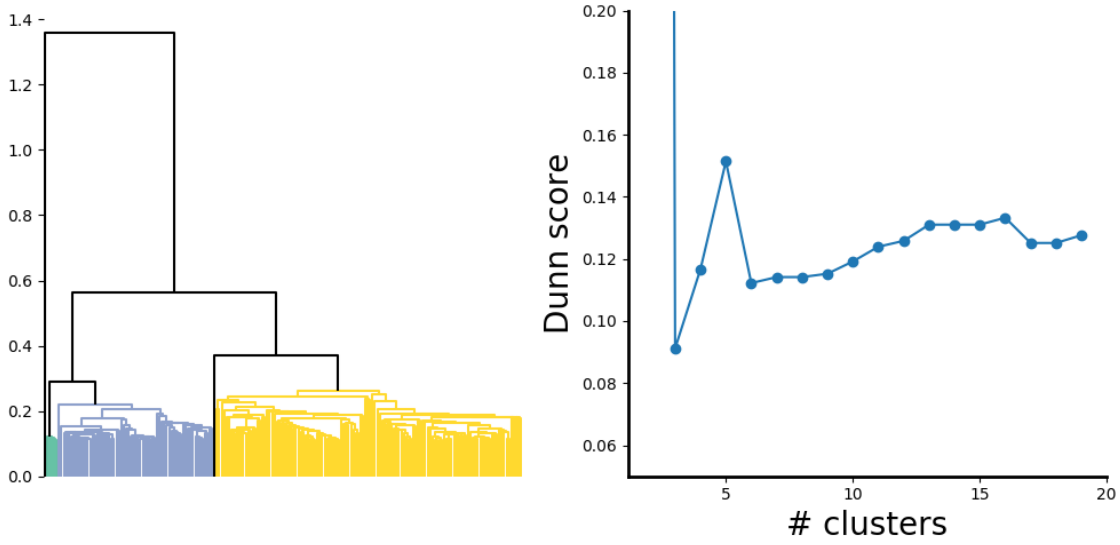

Figure 4.4: The left panel contains the dendrogram of the training dataset, as determined through the Ward linkage method. Colors refer to the 5 clusters subdivision, with two isolated data points shown as black lines that reach the leaves of the dendrogram. The right panel shows the Dunn score of the same dataset.

two nearest neighbors) homogeneous. I start by computing the euclidean distances between each pair of data points in the dataset, which I then use as inputs for the method.

The results of the test are shown in figure 4.3: the analysis reveals that, disregarding a tail of outliers, the intrinsic dimension of the dataset is 98.72, close to 100, so that one can obtain a 25-fold reduction with respect to the original 50×50 degrees of freedom.

4.1.3 Clustering

Another interesting aspect that can be investigated is the emergence of clusters in the dataset of local patterns. I am interested in whether patterns form clusters according to their cell-type or chromosome to understand if generalizations are possible to matrices not present in the training set.

As in the previous section, euclidean distances are computed between each pair of cut-outs and used in order to apply the Ward hierarchical clustering I adopted for Hi-C experiments in chapter 3. Figure 4.4 shows the dendrogram resulting from this analysis (clipped to allow visualization): the major gaps between branches of the dendrogram suggest a subdivision into 5 clusters. In order to confirm this quantitatively I also compute the Dunn score for different numbers of clusters: the result, shown alongside the dendrogram, shows that while the most robust separation is the one with only two clusters, with one

containing a single outlier data point, the second higher Dunn score is obtained for the division into 5 clusters.

In order to understand the structural meaning behind these clusters I first plot their most representative members: for each cluster (having more than one data point) I order its members according to the sum of the distances to other members of the same cluster, i.e. the quantity:

$$s_a = \sum_{b \epsilon C} d_{a,b} \tag{4.1}$$

where a and b are both members of the cluster C, and $d_{a,b}$ is their distance. The representatives are those with the smallest values of s_a, as they live in the most densely populated neighborhoods. Figure 4.5 shows that while the first two clusters contain isolated outliers, the third one contains the empty areas corresponding to the centromeres (which appear as void stripes on the matrices). The fourth and the fifth clusters contain a wider variety of patterns, so an identification of a common thread connecting all the representatives is more difficult to find.

To further investigate the differences between these two clusters and their significance, I focus my attention on the position of their members on Hi-C matrices: given an interaction map, one can obtain a $K \times K$ cut-out for each bin of coordinates (i, j), with $K = 50$. This cut-out can be compared with the points in the dataset and assigned to a cluster with the KNN algorithm, with $k = 100$: first one finds its first $k = 100$ nearest neighbors and counts how many belong to each cluster, then the assignment is given to the cluster with the most members among the nearest neighbors. Moreover, for each cluster one can assign a probability of belonging to that cluster by computing the quantity

$$p_C(i, j) = \left. \frac{\#(nn\,\epsilon\,C)}{k} \right|_{k=100} \tag{4.2}$$

where $\#(nn\,\epsilon\,C)$ is the number of nearest neighbors belonging to cluster C.

This gives a map of the probabilities for each cluster, which helps contextualize the positions in each matrix where patterns belonging to that cluster are more likely to appear. These maps can also be overlapped by showing, for each bin, the color corresponding to the most probable cluster assignment. Figure 4.6 shows the results of this analysis for chromosome 17 of experiment HIC001 (only considering clusters 1, 2, and 3, i.e. removing outliers): cluster 1, containing empty squares, is most probable alongside masked regions, such as centromeres, which are not sampled in experiments, but is never the most likely assignment. Cluster 2 occupies the boundaries delimited by centromeres and some spots along the diagonal, while the rest of the matrix is occupied by cluster 3, which covers most of the Hi-C map.

More interesting groupings may emerge when applying more refined techniques to compute the distances (or similarities) between the cut-outs. In fact, as shown in the previous

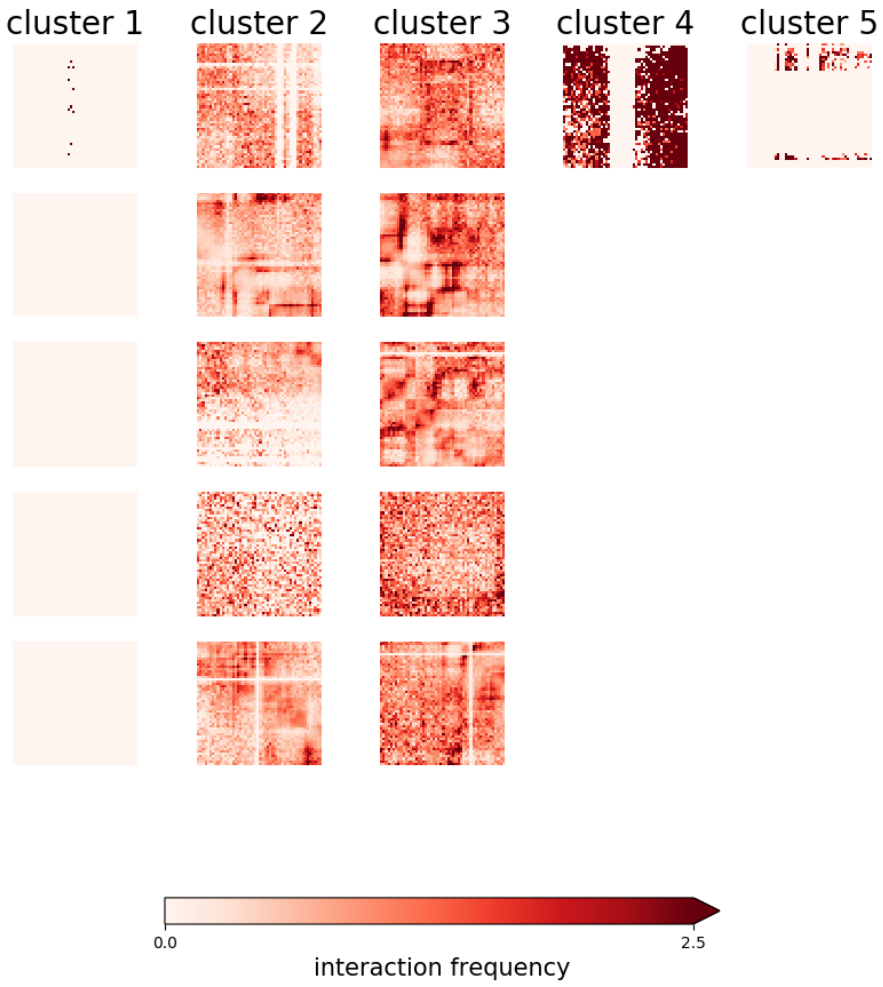

Figure 4.5: Top 5 representative cut-outs for each cluster (except for cluster 4 and 5, which only contain one element each).

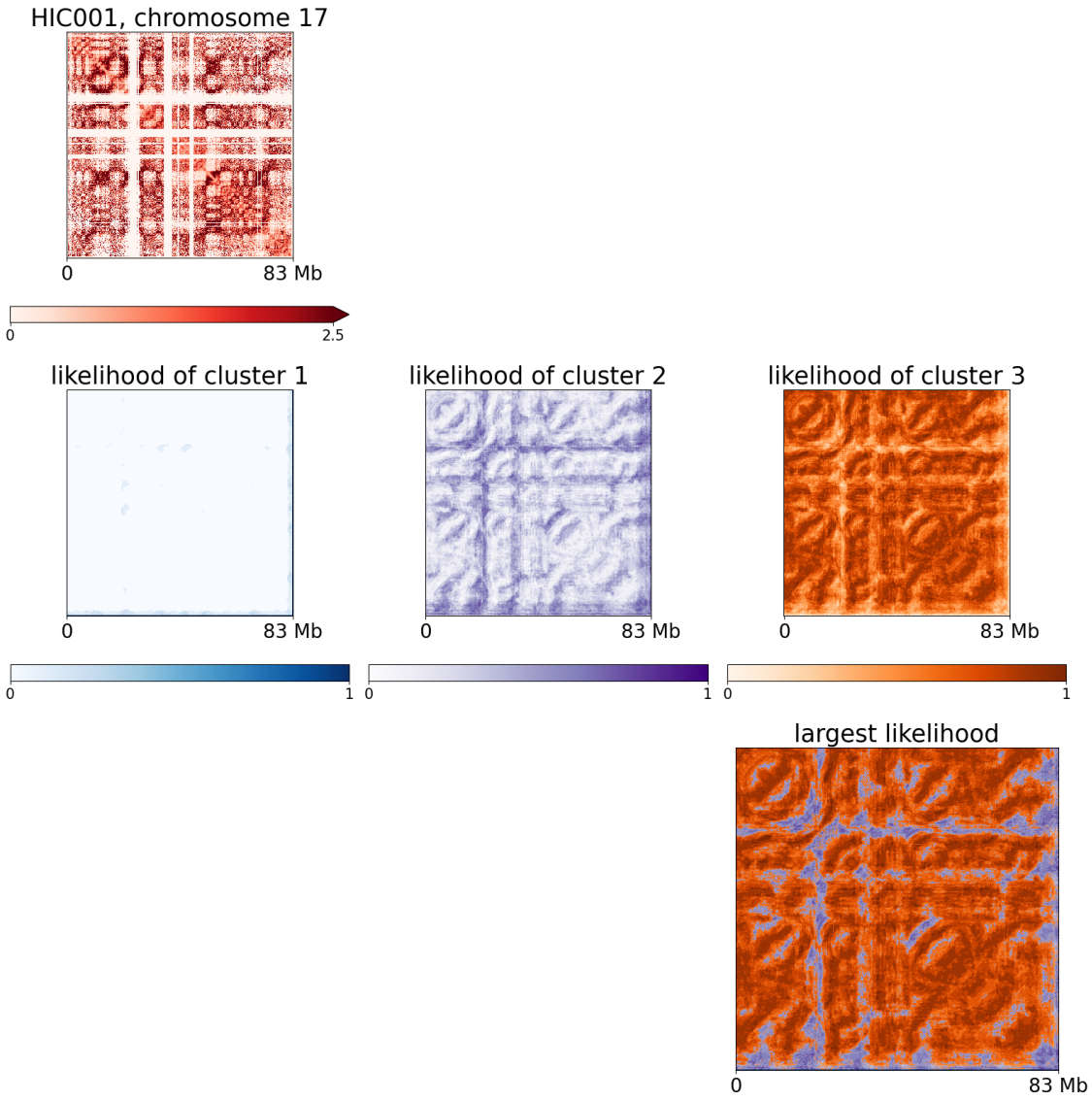

Figure 4.6: Interaction matrix of experiment HIC001, chromosome 17, colored according to the likelihood of each bin belonging to either of the first three clusters obtained by the Ward hierarchical clustering. The last panel shows the largest likelihood among those computed.

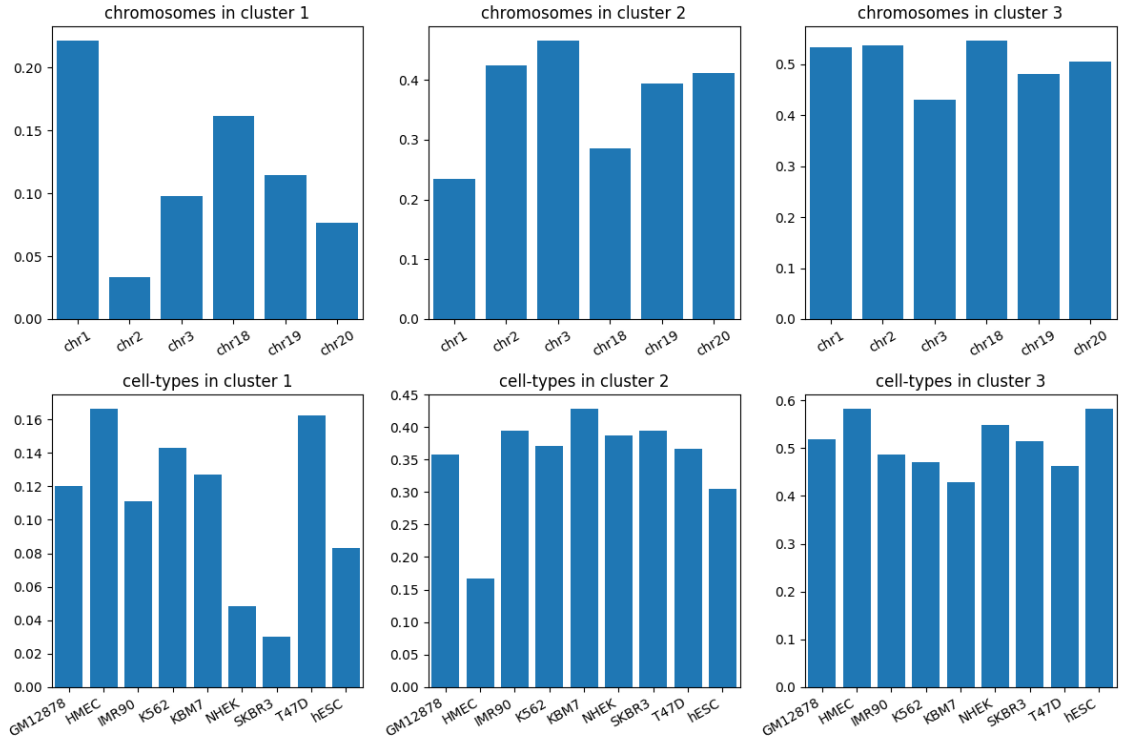

Figure 4.7: Distribution of chromosomes and cell-types for each cluster (except clusters 4 and 5, which only contain one member each). Notice that for cell-types, the bars are normalized by the total number of each cell-type contained in the dataset.

chapter, essential component analysis greatly enhances the ability to distinguish between different cell-types. Here it could be applied to enhance structural similarities and differences in local patterns, that have not emerged due to the usage of a naive Euclidean distance. However here I am mainly interested in raw local patterns as they are, without enhancements, because they form the dataset that need to pass through the dimensionality reduction scheme.

Returning to the original purpose of this section, I consider the segmentations of the three main clusters (i.e. excluding the two outliers) according to the chromosomes and cell-types from which the cut-outs have been sampled. The first panel of figure 4.7 shows, for each cluster, the distributions of cut-outs sampled from each chromosome. The second row of figure 4.7 shows the same for the cell-types. In this case, since the numerosity of each class is different, bars have been normalized by the number of cut-outs of the corresponding cell-type present in the dataset. In both cases fluctuations are present, but the clusters are well mixed, which points towards the presence of shared local patterns:

the same basic patterns can be found in each chromosome and cell-type independently, and thus the analyses of the following sections can be extended outside the training set considered here.

4.2 The Autoencoder Architecture

An autoencoder is a type of neural network architecture that can be used to learn an efficient encoding of unlabeled data[105]. In order to learn and validate the encodings, the network is trained to use them to reproduce the input. As such, the autoencoder contains two parts that can be thought as distinct: the first one, called encoder, usually contains convolutional layers which progressively reduce the degrees of freedom of the input to the dimension of the latent space, i.e. the space where the encodings live. The second part, the decoder, inverts the process by progressively upsampling the encodings in order to produce a point in the same dimensionality of the original input. The reconstructed signal is then compared to the original one in order to compute the loss function used to train the network.

By creating a dimensional bottleneck and attempting to reconstruct the original signal, autoencoders are able to find an efficient representation for a dataset, typically for the purpose of dimensionality reduction, by training to ignore contingent perturbation and focusing on patterns shared throughout the dataset.

While the first application that comes to mind for this type of network is dimensionality reduction, which is what I am mainly interested in, autoencoders find other uses as denoisers and in anomaly detection [19]. Moreover variants which tweak the loss function, such as regularized autoencoders and variational autoencoders, can also be used in different contexts, such as conditional generative models [19].

In practice, autoencoders have already seen some applications to Hi-C maps, in particular in the context of super-resolution [106, 107]: there the task is to map a low resolution matrix (usually artifically down-sampled from some real experiment) to its high-resolution version. In that case, however, the latent space was not studied, and the usage of the autoencoder as a compression tool was not explored.

Here I will use a modified autoencoder architecture optimized on the dataset of local patterns sampled from Hi-C matrices. Figure 4.8 shows the basic architecture of the autoencoder: one starts with a $K \times K$ (with $K = 50$) cut-out sampled from an Hi-C map, which I call A, and passes it through the encoder in order to obtain $e(A)$, an array of k elements in the latent space, with k being a parameter of the model. The square shape of the encoded vector is not strictly necessary, but allows for a more intuitive visualization of the results into a matricial form by placing the latent vectors encoded from neighboring cut-outs next to each other. The encoded vector is then passed through a decoder to obtain $A' = d(e(A))$, a $K \times K$ cut-out to be compared to the original one ($K = 50$ is fixed).

Let us delve deeper into the details of the architecture.

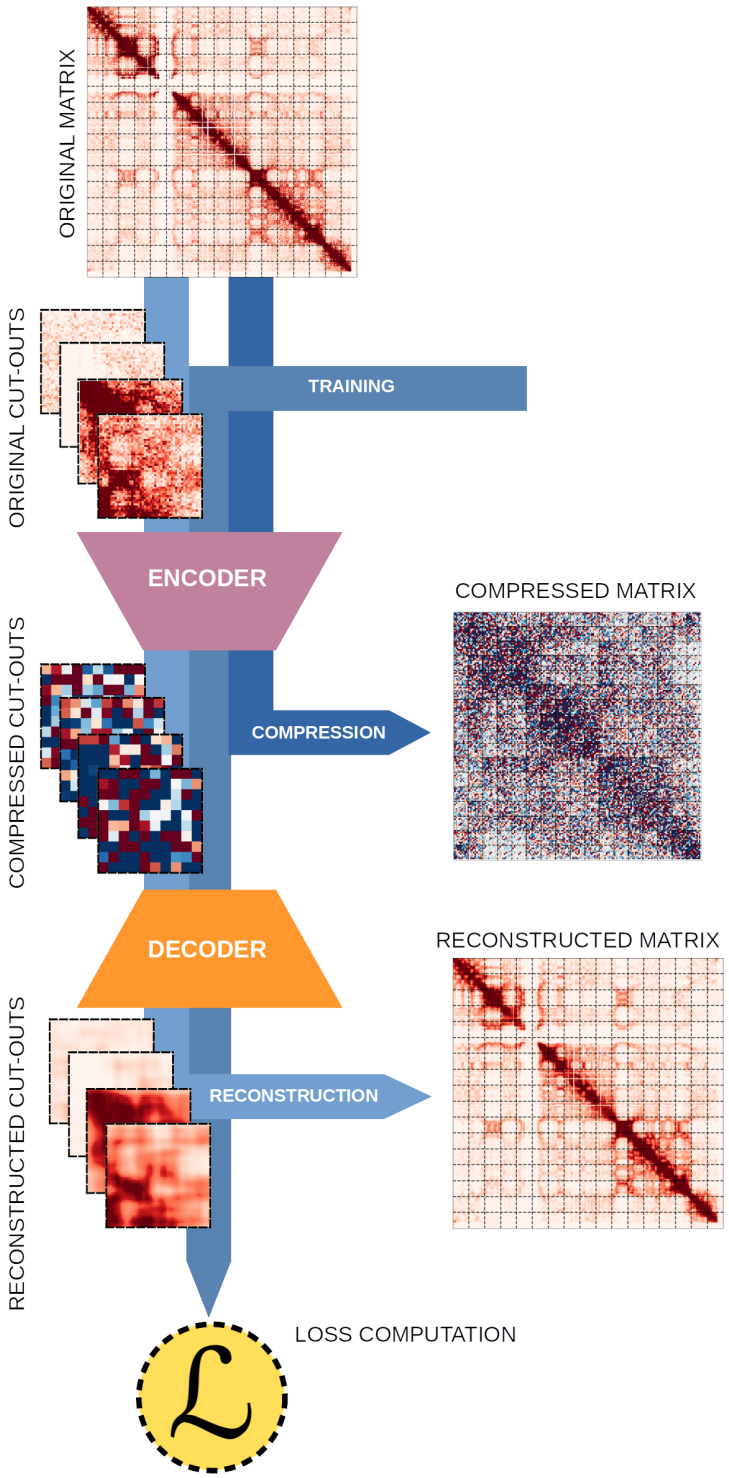

Figure 4.8: Scheme of the autoencoder: cut-outs extracted from the training dataset or a matrix are passed through the encoder to obtain their latent space representations, which in turn can be fed to the decoder to retrieve their reconstructions. Reconstructed and compressed cut-outs can be used to compute the loss function, or stitched back together in order to obtain the reconstructed and compressed, respectively, representations of the original matrix.

The basic loss function of an autoencoder is given by the mean square error (MSE) between the original and the reconstructed signal:

$$\mathcal{L} = MSE(A, A') = \frac{1}{K \times K} \sum_{i,j} \left(A_{i,j} - A'_{i,j} \right)^2 \tag{4.3}$$

where A is the original cut-out, A' is the cut-out reconstructed by the network, and $K = 50$ is their linear size.

Since the only requirement this autoencoder has to satisfy is that the input and output be similar to each other, its latent space is known to be prone to severe over-fitting [105]: the autoencoder will try to retain as much information as possible through the dimensional bottleneck, but at the same time it will not form a proper representation for points outside the training set. This means that when the decoder is applied to these points, the resulting reconstruction does not resemble any real pattern encountered in Hi-C matrices. Such a behavior may be detrimental when trying to generalize the procedure outside the original dataset, as the autoencoder may not understand input patterns even when they resemble each other.

In order to avoid this, the latent space must be regularized. Variational autoencoders [105] do this by exchanging points in the latent space with distributions, so that the whole space is covered, and over-fitting is avoided. In order to perform this change, the encoder is changed to return not a single point, but two vectors (with the dimensionality of the latent space) containing the means and variances of the distribution along each dimension. These are restrained by the presence of an additional term in the loss function:

$$\mathcal{L}_0 = MSE(A, A') + D_{KL}(N(\mu, \sigma), N(0, 1)) \tag{4.4}$$

where the second term D_{KL} is the Kullback Leibler divergence between the observed normal distribution $N(\mu, \sigma)$ with mean vector $\mu = \{\mu_i\}_{i=1}^{i<k}$ and variance vector $\sigma = \{\sigma_i\}_{i=0}^{i<k}$, and the target normal distribution $N(0, 1)$:

$$D_{KL}(N(\mu, \sigma), N(0, 1)) = \frac{1}{2} \sum_{i=0}^{i<k} \left(\sigma_i^2 + \mu_i^2 - 1 - 2 \ln(\sigma_i) \right) \tag{4.5}$$

Notice that the covariance matrix of the generated distribution is constrained to be diagonal.

Finally, one can observe that Hi-C maps are symmetric with respect to the diagonal. This property should be carried over in the latent space in order to obtain symmetric results when one inputs symmetric cut-outs. This is enforced by constraining the transpose encoded vector $e(A)^T$ to be equal to the encoded vector of the transposed input $e(A^T)$:

$$\mathcal{L}_{SYM} = MSE(e(A)^T, e(A^T)) = \frac{1}{k^2} \sum_{i,j} \left(e(A)_{j,i} - e(A^T)_{i,j} \right)^2 \tag{4.6}$$

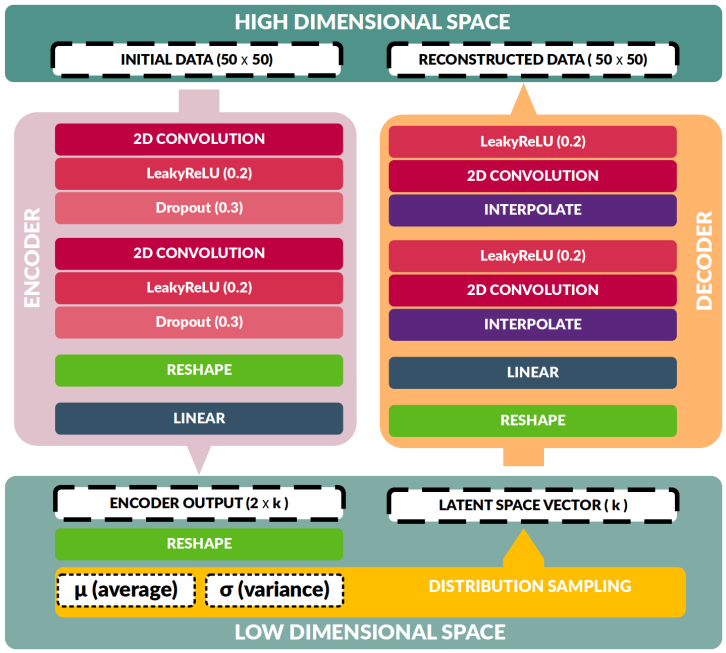

Figure 4.9: In depth scheme of the variational autoencoder.

this is then plugged into the general loss function to obtain

$$\mathcal{L} = MSE(A, A') + D_{KL}(N(\mu, \sigma), N(0, 1)) + \mathcal{L}_{SYM} \tag{4.7}$$

Figure 4.9 shows the details of the architecture of the variational autoencoder.

The encoder contains three layers: the first two are standard Convolutional layers followed by LeakyReLU activations and noisy dropout. The third contains a reshaping of the resulting square tensor into a linear shape, followed by two Linear combination layers. The parameters of these layers are chosen in order to progressively reduce the linear dimensionality of the matrices from $K = 50$ to k: however in the variational autoencoder scheme, the encoder output is given in the form of two vectors containing the averages μ and the variances σ of the underlying distribution. These are used to sample a single point which is then passed to the decoder.

The decoder does not use standard Deconvolutional layers, as they are prone to output artifacts: usually these present themselves in the form of noticeable periodic perturbations in the generated tensors[108]. Rather, I implement a layer formed by a Rescaling followed by a standard Convlution layer (resize-convolution layer [108]), which has been shown to be more efficient and less prone to create artifacts[108]. Again, the parameters are chosen in order to upsample vectors of k elements to $K \times K$ matrices.

4.3 Analysis of the reconstructed space

In this section I will analyze the reconstructed matrices obtained by passing Hi-C interaction maps through the autoencoder. Various sizes k of the latent space, and thus different variations of the same basic architecture presented above, have been used in order to compare results.

First I will show visual examples of the original and reconstructed matrices, as well as differences between the two; then I will proceed to more quantitative analyses. In doing so I will keep two objectives in mind: the first is to reconstruct the original matrix through the autoencoder with minimal differences, the second is to check that inevitable differences between the two do not harm further analyses one could perform on the matrices downstream of compression and restoration.

The former objective, the consistency between the original and reconstructed matrices, can be checked by computing the MSE score, which is also defined above in the context of the loss function. Notice that while this is indeed a term of the loss function $\mathcal{L}$, it is not the only one: the objective of the autoencoder is to optimize its parameters with respect to a more complex loss function which also contains constraints about the latent space. This makes the comparison of the original and constructed matrices through the MSE score function non trivial. In order to contextualize the results, they will be compared to other methods.

The latter objective is more broad in its ramifications: what loss of information can be deemed acceptable when dealing with Hi-C matrices? While the ideal result would be a lossless compression, it is my opinion that the most important parameter in the real scenario of lossy compression is the ability of restored Hi-C matrices to retain biological and structural information. I will perform two analyses that broadly capture the spirit of this objective: first I will compare TADs obtained from original and reconstructed matrices. Then I will perform comparisons between pairs of matrices and test the degree to which a simple euclidean distance can discriminate between biological replicates and non-replicates.

Every analysis will be performed both on Hi-C matrices of chromosomes from which the cut-outs were sampled $(1, 2, 3, 18, 19, 20)$ and other chromosomes, which are not part of the learning process of the method.

4.3.1 Visual inspection

First I pass Hi-C matrices through the autoencoder to obtain reconstructed ones. To do this, I start by cutting the matrix into squares $K \times K$ in size ($K = 50$); zero padding is added to the matrix in order to obtain a size commensurable with the squares. Afterwards, each cut-out is passed separately through the autoencoder; the resulting reconstructed blocks are collected and positioned next to each others in order to reproduce the original matrix.

Figure 4.10 shows the results: the original matrix of experiment HIC001, chromosome 3, is displayed along with its reconstructions produced by the autoencoder at different

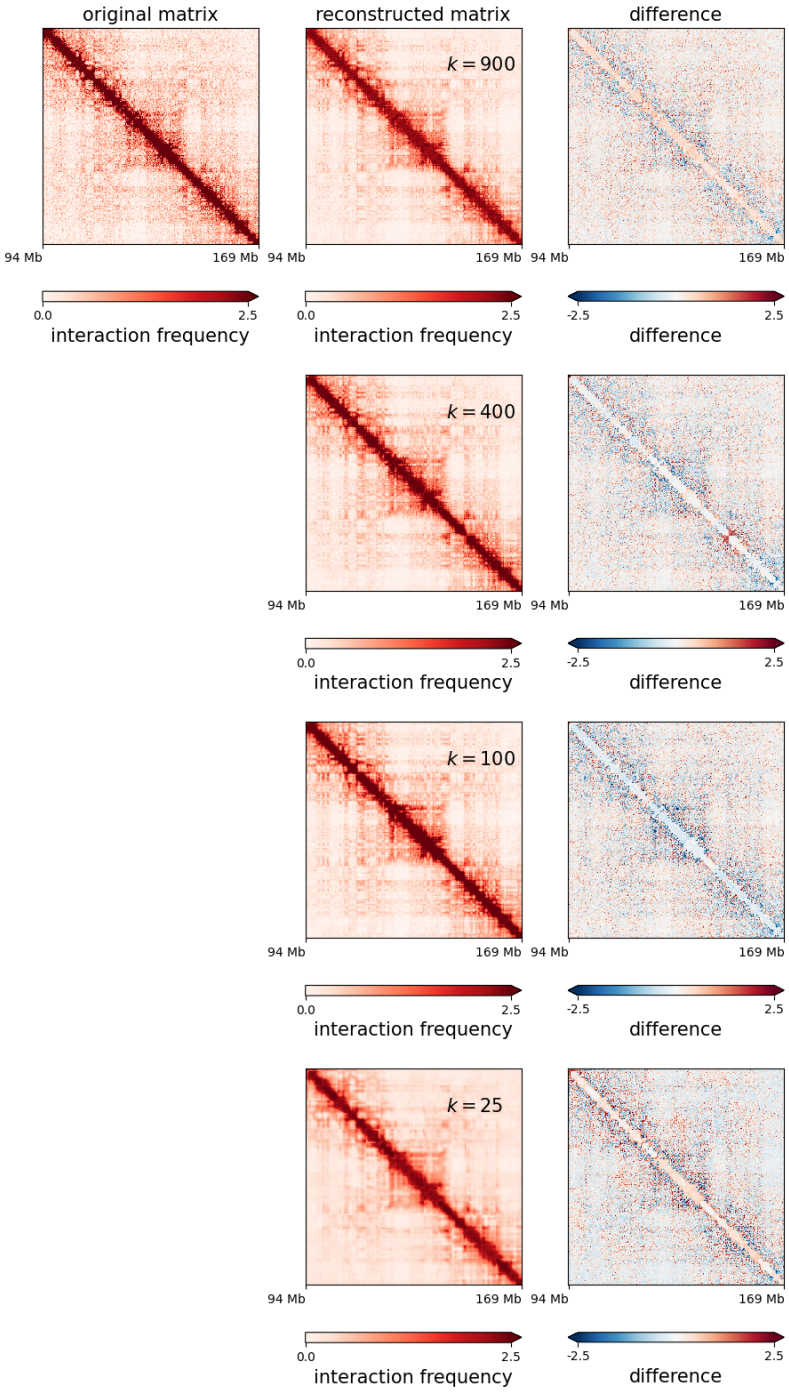

Figure 4.10: Visual inspection of the reconstruction capabilities of the autoencoder for different sizes of its latent space, on HIC001, chromosome 3.

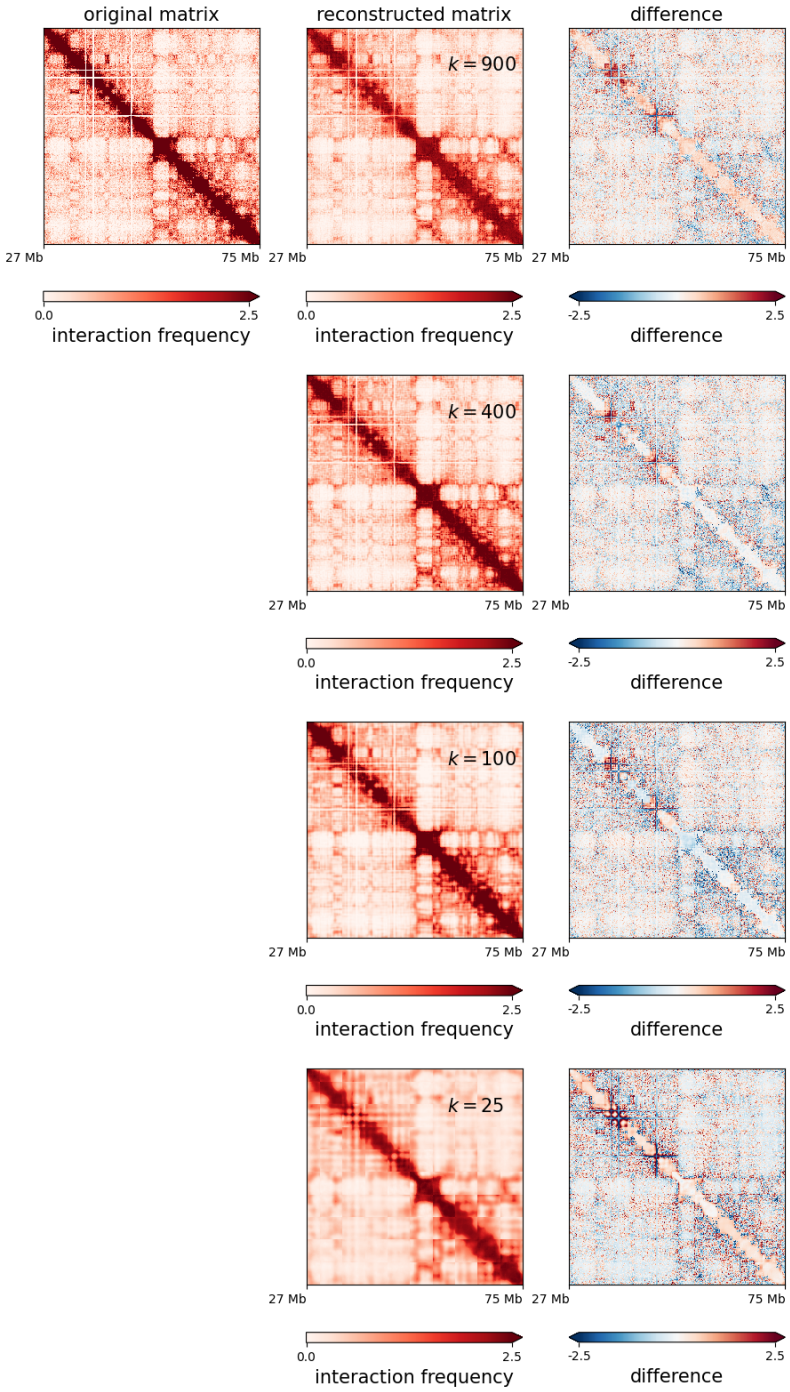

Figure 4.11: Visual inspection of the reconstruction capabilities of the autoencoder for different sizes of its latent space, on HIC001, chromosome 17.

values of the linear size k of the latent space. Differences between the original and its reconstructed matrices are also plotted.

Figure 4.11 plots the same for experiment HIC001, chromosome 17, which was not used during training.

The first observation is that reconstructed matrices qualitatively resemble the original ones both for chromosomes used during training and those that the neural network has not seen before. This is a sign that the scheme used to compress and restore matrices can be successfully generalized to other dataset without dramatic drops in performance: more quantitative analyses are offered below.

Secondly, one can qualitatively observe that larger values of k generally correspond to sharper matrices, with more defined boundaries between domains. However to better quantify the quality of the reconstruction an analysis of the MSE reconstruction loss is need.

4.3.2 Reconstruction fidelity

In order to quantify the ability of the autoencoder to restore matrices after compression, without significant loss of information, I analyze the MSE between original and decoded matrices over different chromosomes.

To contextualize the result, I compare them to two alternative models that act on the matrices.

The first, a Gaussian filter with unit variance, acts by averaging neighboring bins and is used to smooth matrices, at the cost of loosing some of the sharpness of the interaction patterns. While this is not a compression algorithm, its action on Hi-C maps is qualitatively similar to what one obtains by passing the maps through a narrow dimensional bottleneck during the compression: as such it makes sense to employ this method as a baseline of the baseline result.

A second term of reference is provided by a principal component analysis trained on the cut-outs dataset and applied to each separate block, as was done for the autoencoder. PCA is closely related to autoencoders: in fact it can be proven that only using linear layers to compose the autoencoder, without any non-linear activations between them, leads to the same results as PCA [109]. This means that PCA can be interpreted as the linear counterpart of non-linear autoencoders, such as the one devised here. Its ability to provide a linear dimensional reduction as well as restoring the original matrices from the low dimensional latent space makes it a good candidate to compare the results with.

Figure 4.12 shows a visual comparison of matrices obtained by applying the three techniques to experiment HIC001, chromosome 17: the Gaussian filter visibly blurs the patterns of the original matrix, and while this may attenuate the effect of the noise, it also degrades the quality of the boundaries between domains. PCA obtains better reconstructions, which are clearly better than the one offered by the Gaussian filter. The autoencoder is also able to achieve good reconstruction quality, and it is not immediately clear whether its results

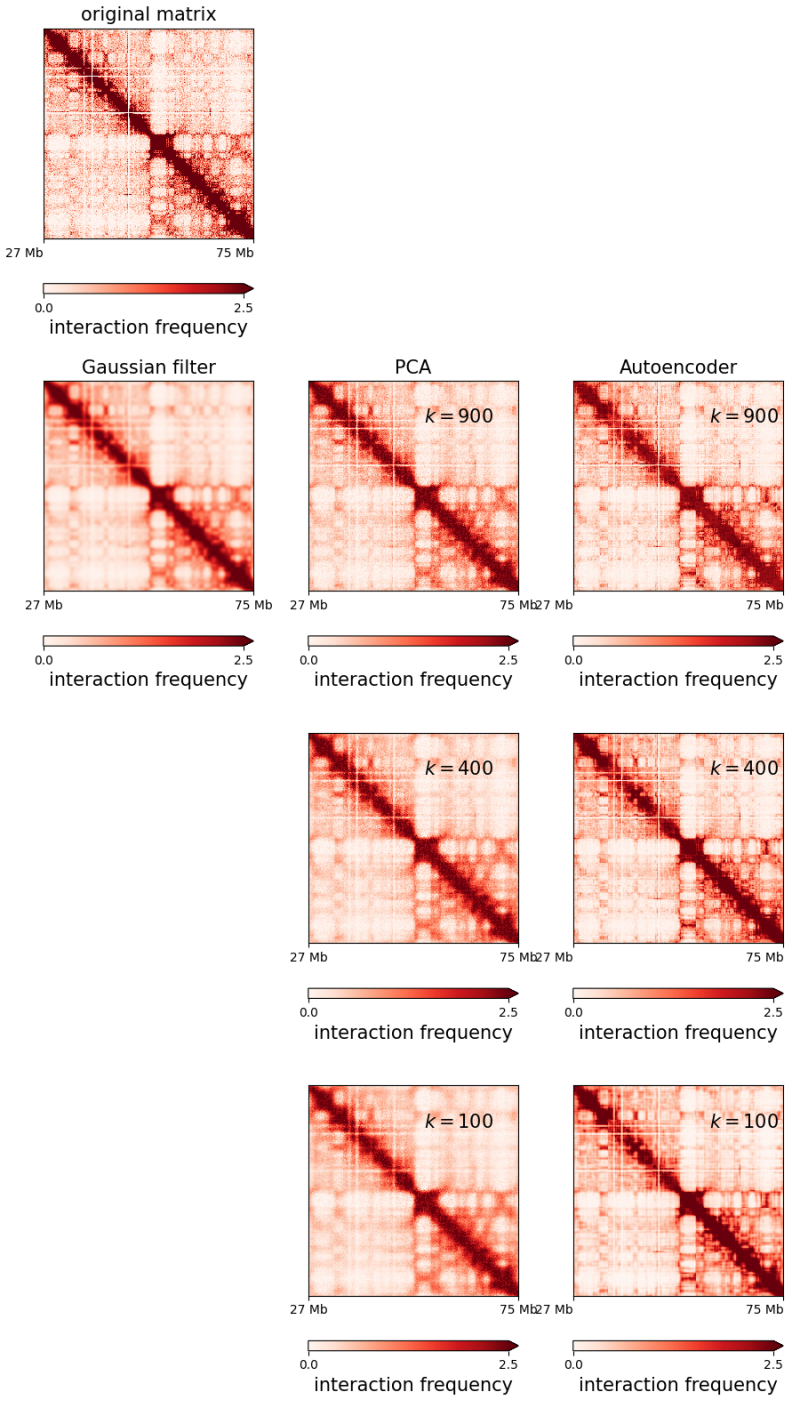

Figure 4.12: Reconstruction of HIC001, chromosome 17, using three different methods: a Gaussian filter, Principal Component Analysis, and the autoencoder.

are better than those achieved by PCA. By computing the MSE reconstruction loss, one can compare the two methods from a more objective standpoint.

Figure 4.13 shows the results for different chromosomes and different sizes of the latent space: autoencoders with larger k perform slightly better, although the value of the MSE is roughly stable for any value above $k \simeq 100$, which also corresponds to the intrinsic dimension of the dataset. Below this threshold the results quickly degrade. Nevertheless even the worst performance of the autoencoder (at $k = 4$) is an order of magnitude better than results obtained by PCA: while this may not seem immediately clear from visual inspection of the reconstructed matrices, this comparison quantitatively shows the superiority of the autoencoder. The Gaussian filter is vastly outperformed by both methods, displaying errors which are four orders of magnitude larger.

It is interesting to notice that in all methods the MSE scores increase as the size of the chromosomes decreases. This may be connected to the presence of larger errors along the diagonals, which hold more weight in small chromosomes: it suggests that the results of the autoencoder could be further improved by explicitly designing the loss function to counterbalance this aspect of Hi-C maps by paying more attention to the reconstruction of cut-outs close to the diagonal.

4.3.3 Preservation of structural details

Here I look at the preservation of TADs in reconstructed matrices.

I apply the MOC score I first used in chapter 3: I compare the TADs obtained from the insulation score computed on original matrices to those of reconstructed matrices. As a term of reference I employ the same score to compare TADs of biological replicates and I use the average as the baseline.

Figure 4.14 shows the results for chromosome 1 and 17, as a function of the size of the latent space. In both cases the MOC score is roughly constant for $k \simeq 100$ and above, but rapidly decreases for smaller latent spaces. The stable portion of the MOC score for reconstructed matrices is comparable to the results of the biological replicates comparison. This is not ideal, as it shows that some structural properties are lost during the decompression of matrices, however this effect might be mitigated by improving the reconstruction of the diagonal with an *ad hoc* loss function.

4.3.4 Preservation of biological information

Here I check whether reconstruction preserves the biological information about the cell-types of each experiment.

This can be done by looking at the ability of the euclidean distance between experiments to discriminate between biological replicates and non replicates. Like in the previous chapter, ROC curves are used to quantify the results and an overall score is provided by the Area Under the Curve (AUC).

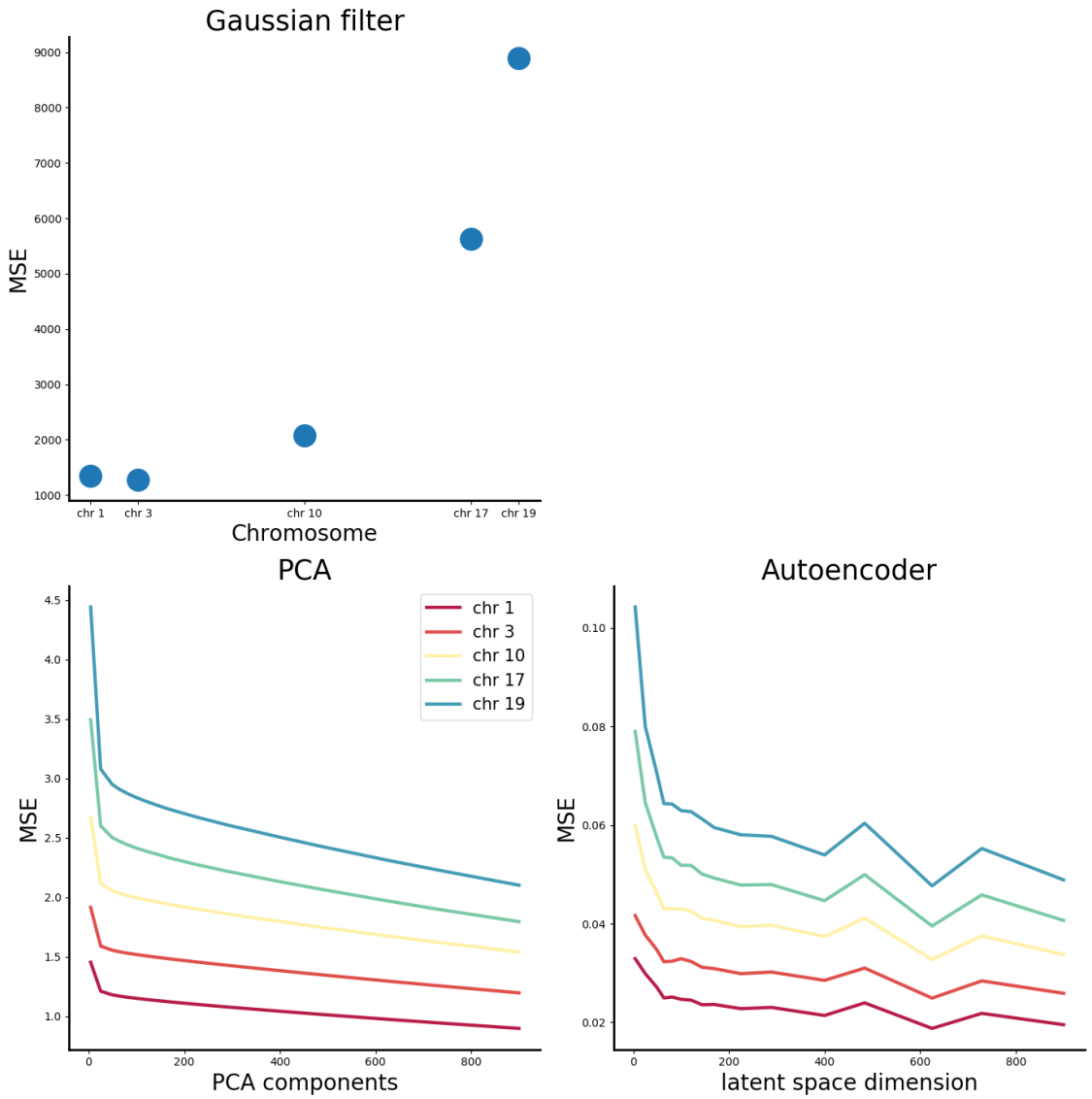

Figure 4.13: Plots of the MSE scores for the three methods. The score obtained by the Gaussian filter is shown for different chromosomes; in the case of PCA and the autoencoder, the MSE scores are plotted as a function of the number of principal components or dimensions of the latent space respectively.

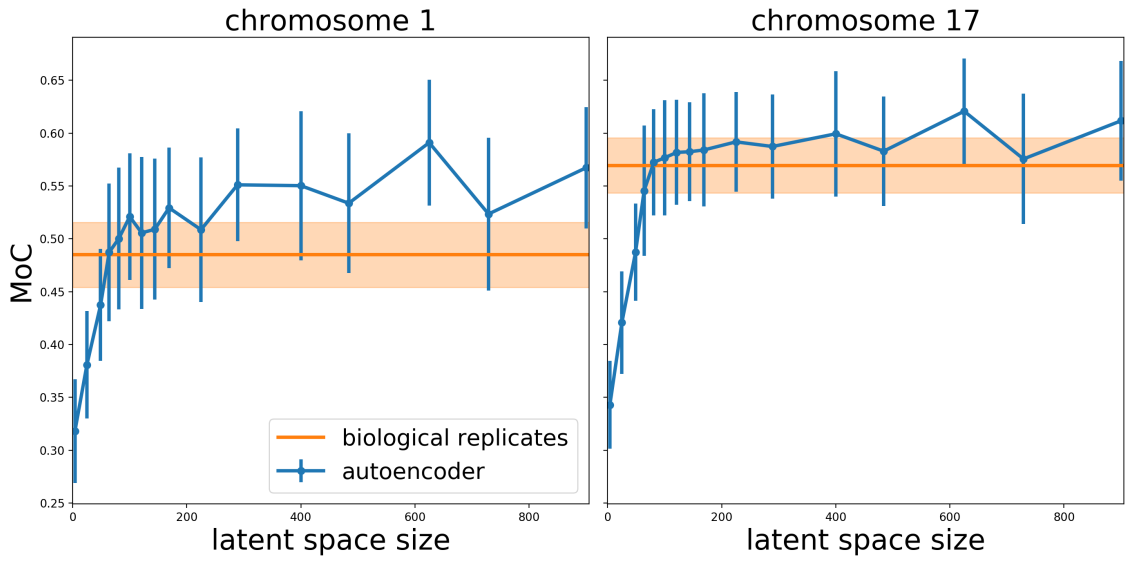

Figure 4.14: MOC scores for TADs computed on original and reconstructed matrices as a function of the latent space size. The error bars indicate one standard deviation from the dataset average. The MOC scores computed on original matrices of biological replicates are also shown as a term of reference, and the shaded area indicates one standard deviation from the dataset average.

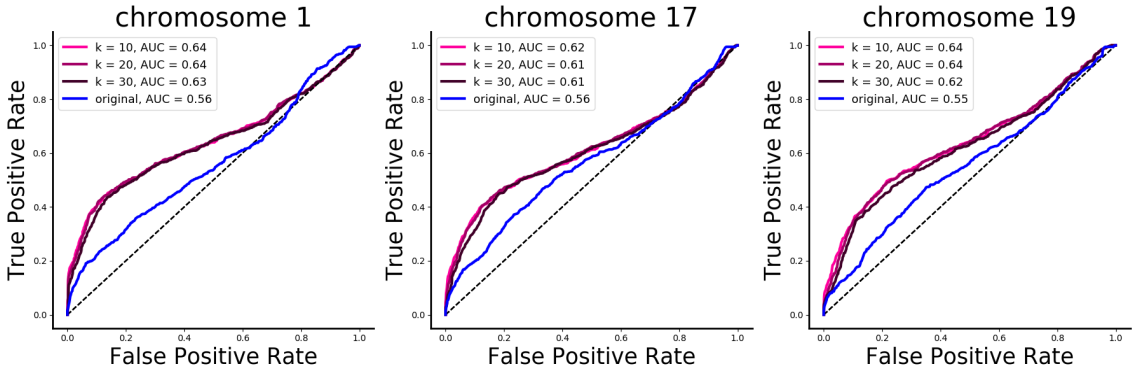

Figure 4.15: ROC curves computed from reconstructed matrices at different sizes of the latent space, in different chromosomes. The ROC curves computed from the original matrices are shown for comparison, and AUC scores are provided for each curve.

The results are plotted in figure 4.15 for different chromosomes (both those employed in training the autoencoder and the others), and different values of the latent space dimension k: the ROC curves are mostly overlapped in both sets of chromosomes, showing that the passage through the autoencoder and its dimensional bottleneck does not compromise the structure of the dataset. In fact, the reconstructed dataset displays slightly larger AUC scores in most cases, which is a sign that the part of the signal which did not pass through the dimensional sieve is not biologically relevant.

The quality of the results shown here is not comparable to that of essential component analysis presented in the previous chapter, but this is to be expected. Using a simple euclidean distance, with a bin to bin comparison of the Hi-C maps, has already been shown to be ineffectual. Moreover, I did not account for the genomic distance bias, which makes differences along the diagonal orders of magnitude more important than those between long distance regions. In reality, structures along the diagonal have been shown to be conserved in different cell-types, with much structural variability being contained in those long distance regions which this naive analysis does not focus on.

Of course the objective of the autoencoder is exactly to *reproduce* the information present in untreated matrices, not to enhance it. This allows for more refined analyses to be applied downstream of the compression and restoration scheme. However, this result begs the question of whether applying the autoencoder in a more favorable context would allow it to also enhance the results beyond those obtained on the original dataset. I tackle this problem in the last section of this chapter, where I compare the autoencoder to essential component analysis.

4.4 Analysis of the latent space

The previous section dealt with the results of passing matrices through the autoencoder in order to compress and then reconstruct them. In this section I will delve into the latent vectors that encode the compressed information about Hi-C maps.

I will first present a visual inspection of the matrices constructed by placing latent vectors of neighboring cut-outs near to each other. These objects can be thought of as the compressed matrices and part of the section will be devoted to finding out whether the dimensional reduction reveals any new structural information about the dataset.

In fact, as shown in the previous section, the autoencoder preserves (to a good degree) structural and biological features, and is able to nicely reproduce the original matrices. Compressed representations of local patterns contain fewer redundancies and may be easier to analyze in search of emergent groupings in the latent space.

As such, I repeat the clustering analysis on the compressed cut-outs obtained from the training dataset and check whether the previous dendrogram structure of figure 4.4 can be confirmed, or new clusters emerge. Moreover I also repeat the ROC curve analysis using the latent matrices described above in order to check whether discrimination between

biological replicates and non replicates improves in this space.

4.4.1 Visual inspection

As usual the first step in analyzing the results is to look at them from a qualitative standpoint: figure 4.16 shows the latent space representations of the original Hi-C maps of an array of different chromosomes and cell-types, represented in matricial form.

To obtain these, I start by cutting up the original matrix into blocks and pass them through the encoder (the first part of the autoencoder). The output is the latent space representation of the starting cut-out, which one can pass through the decoding portion of the network in order to restore it to its original form. However in this case I collect the latent space vectors, which are reshaped as $k \times k$ matrices, and place them side by side in the same order of the matrix they come from.

The resulting matrices are the compressed versions of the original ones: one can start by noticing that, thanks to the symmetry term of the loss function, they appear symmetric with respect to the diagonal, preserving the same property of the full matrices. In the patterns presented in figure 4.16 one can recognize some of the features contained in the original matrix, such as a darker diagonal. Moreover different cell-types display visibly different patterns, which may allow one to more easily differentiate between groups.

4.4.2 Clustering

Using latent space representations of high dimensional data has been shown to improve clustering, allowing better separation between groups of related local patterns. For this reason, it is interesting to study the structure of the representations of the training dataset in the low dimensional latent space and apply clustering.

I start by computing euclidean distances between the vectors obtained by applying the encoder to the dataset and obtain a dataset wide distance matrix. From this, I can employ the usual hierarchical Ward clustering to obtain an arbitrary number of subdivisions.

In order to choose an optimal grouping I employ the Dunn score, which computes the ratio between cluster distances and their sizes. The plot in figure 4.17 shows that the Dunn score is always smaller than the one observed for the original cut-outs in figure 4.4: this signals that any gap between groupings in the latent space will be small with respect to their diameter. This is a feature of the regularization performed by the KL divergence of the autoencoder loss, that guarantees that no gaps are present in the latent space [105]. However meaningful clusters may still emerge: figure 4.17 shows two possible main subdivisions, into 2 or 8 clusters, and signaled by two isolated peaks.

In order to visualize the groupings, I plot a dendrogram of the hierarchical clustering in figure 4.18: one can immediately see that the dataset splits neatly into different groups with large jumps of the Ward score, signaled by longer arms connecting them. Again, this points to the subdivisions suggested by the Dunn score analysis: a two cluster subdivision

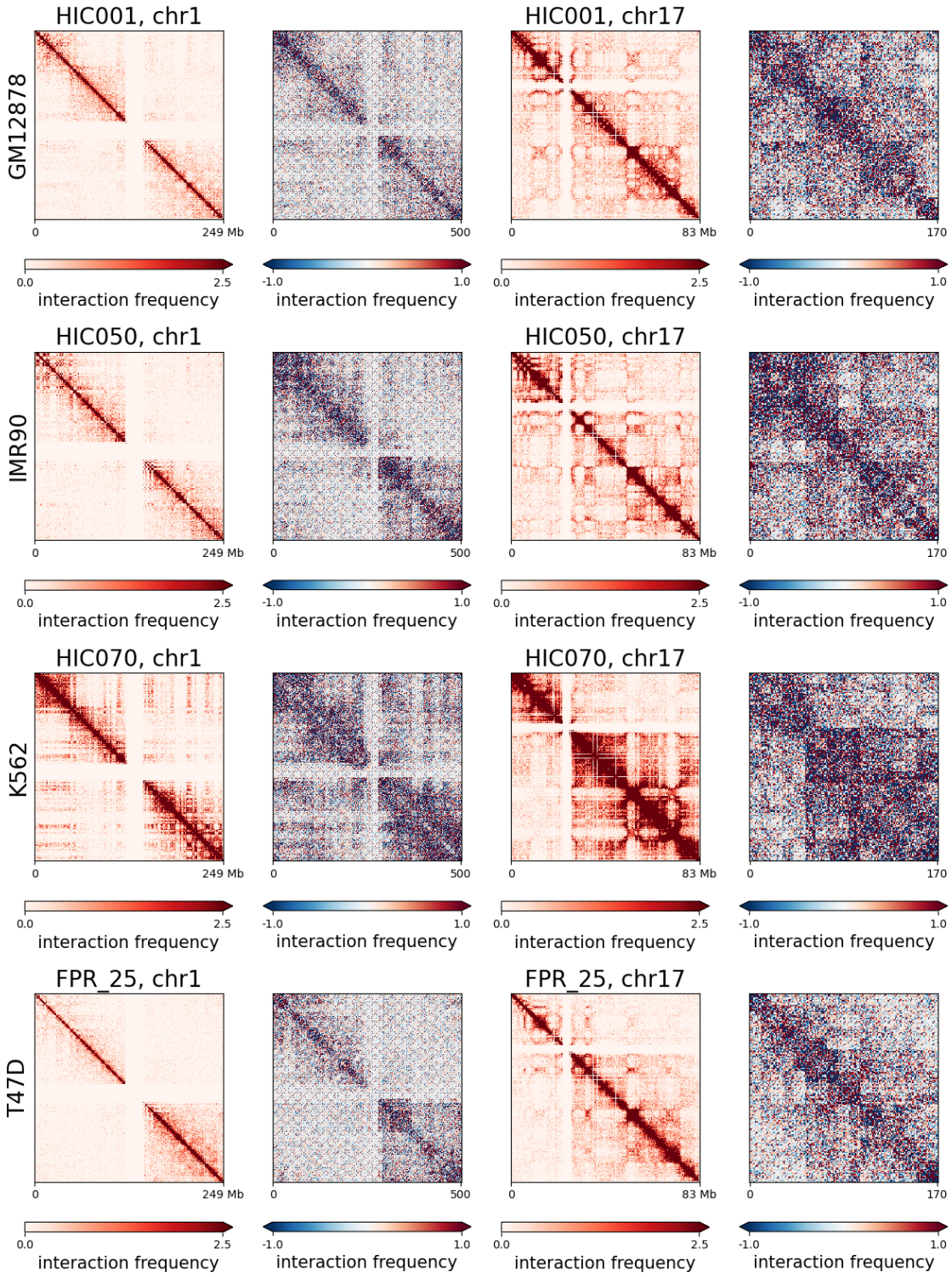

Figure 4.16: Visualization of latent space representations for two different chromosomes (1,17) and four cell-types (GM12878, IMR90, K562, T47D). The size of the latent space of the autoencoder is 10×10.

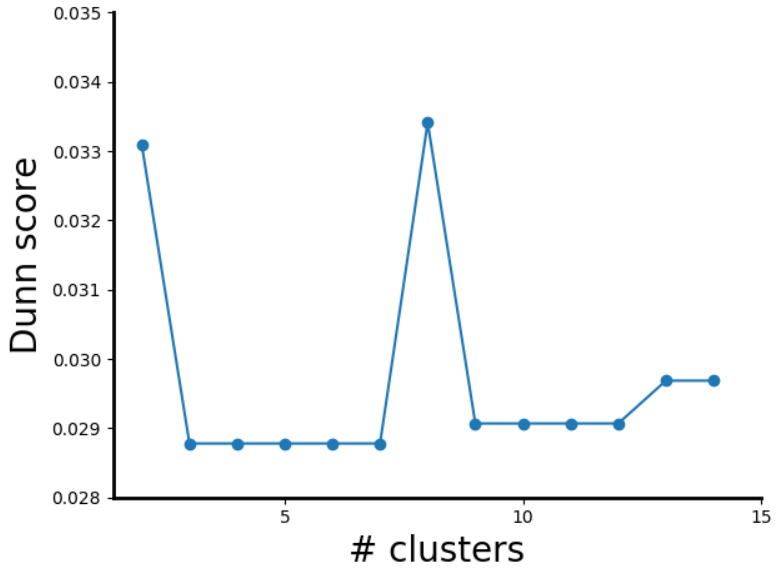

Figure 4.17: Dunn score computed on latent representations of the training set data points.

correspond to the first and largest gap, while each of these groups can still be subdivided into finer partitions, obtaining eight clusters.

Notice that the situation is different from what is observed for the original high dimensional dataset of local patterns: there (figure 4.4) the Dunn score suggested a subdivision in either two or five clusters, but the second grouping was not as clear as the one found in the latent space. Moreover two of the clusters found in that case contained only a single cut-out, whereas here the smallest cluster counts three members, with all the others having large populations. An inspection of the dendrograms immediately confirms that the subdivision found in the latent space is much more robust with respect to the original one.

The identity of the clusters can be understood by looking at their representatives, plotted in figure 4.19: following the main division into two groups, clusters one to three are centered on the masked centrosomes and contain a large quantity of white spaces; the others are characterized by an array of diverse patterns. Visual inspection suggests that subdivisions in the latter group are mainly due to the differences in the average interaction: a more quantitative analysis confirms the presence of a division based on the strength of the average interactions found in the cut-outs, as seen in figure 4.20.

It is also interesting to ask whether a correspondence is present between the clusters found in the original dataset (original clusters, for simplicity) and those found in its latent space representations (latent clusters). In order to answer this question and probe to what degree this correspondence extends, I plot in figure 4.21 the histograms for each original cluster, showing how many of their members are contained in each of the latent clusters.

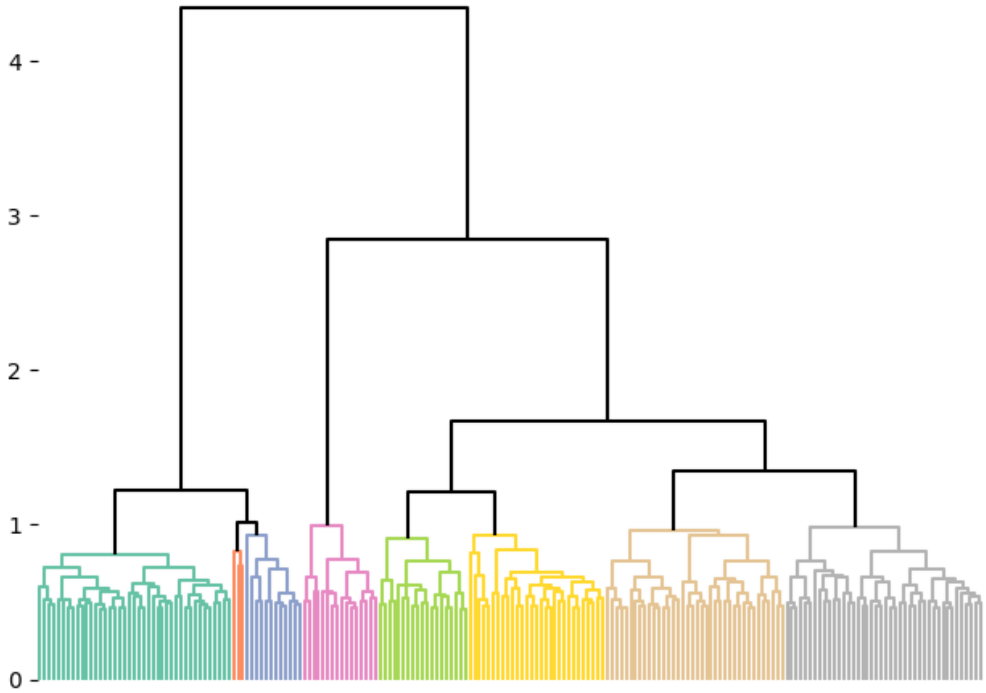

Figure 4.18: Dendrogram of the latent space representations of the training set data points, as determined through the Ward linkage method. Colors refer to the 8 clusters subdivision.

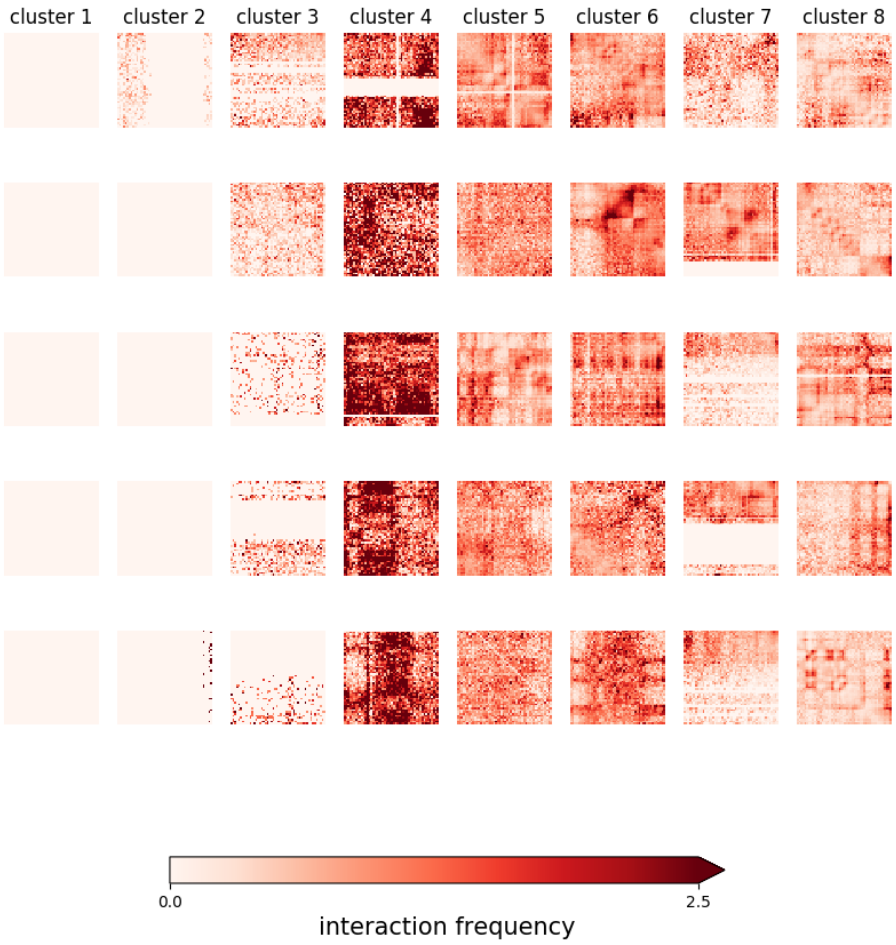

Figure 4.19: Top 5 representative cut-outs for each cluster in latent space.

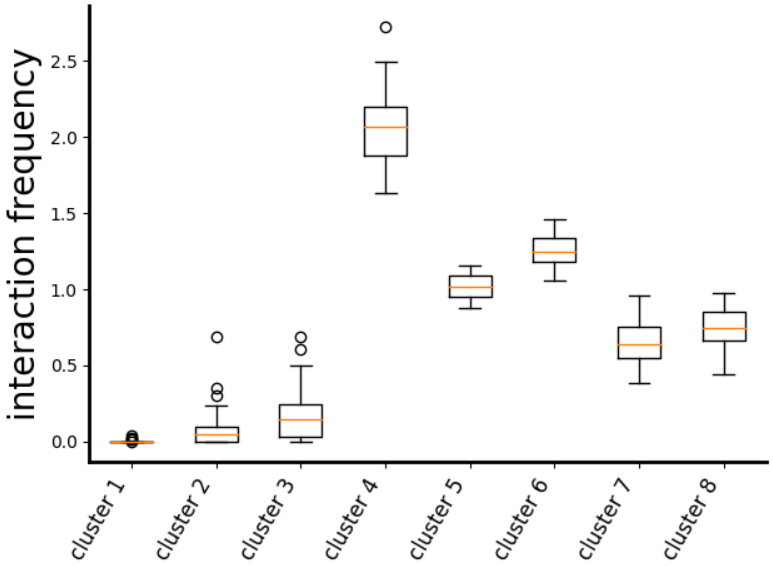

Figure 4.20: Interaction frequency measured in different clusters as boxplots. Boxplots show: central line, median; box limits, 75th and 25th percentiles; whiskers, 1.5 times the interquartile range; outliers beyond this range are shown as individual points.

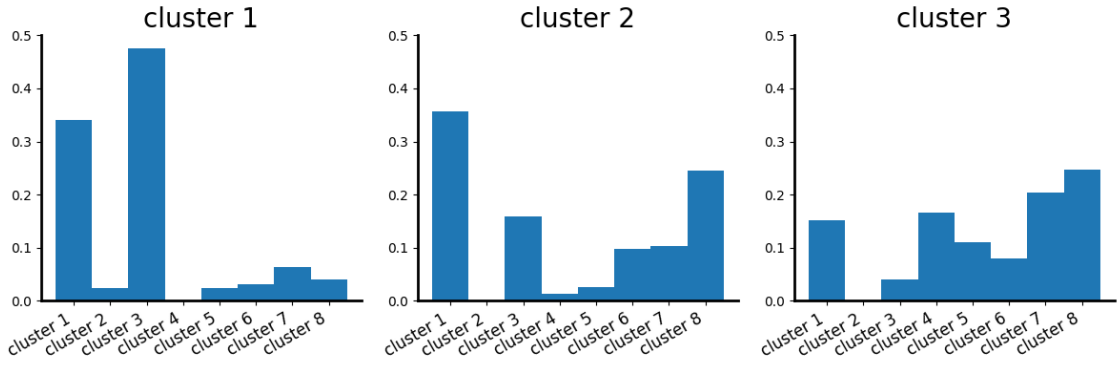

Figure 4.21: Subdivision of the original clusters (computed from cut-outs in the high-dimensional space) into latent clusters (computed from latent space representations of the same cut-outs), given as histograms.

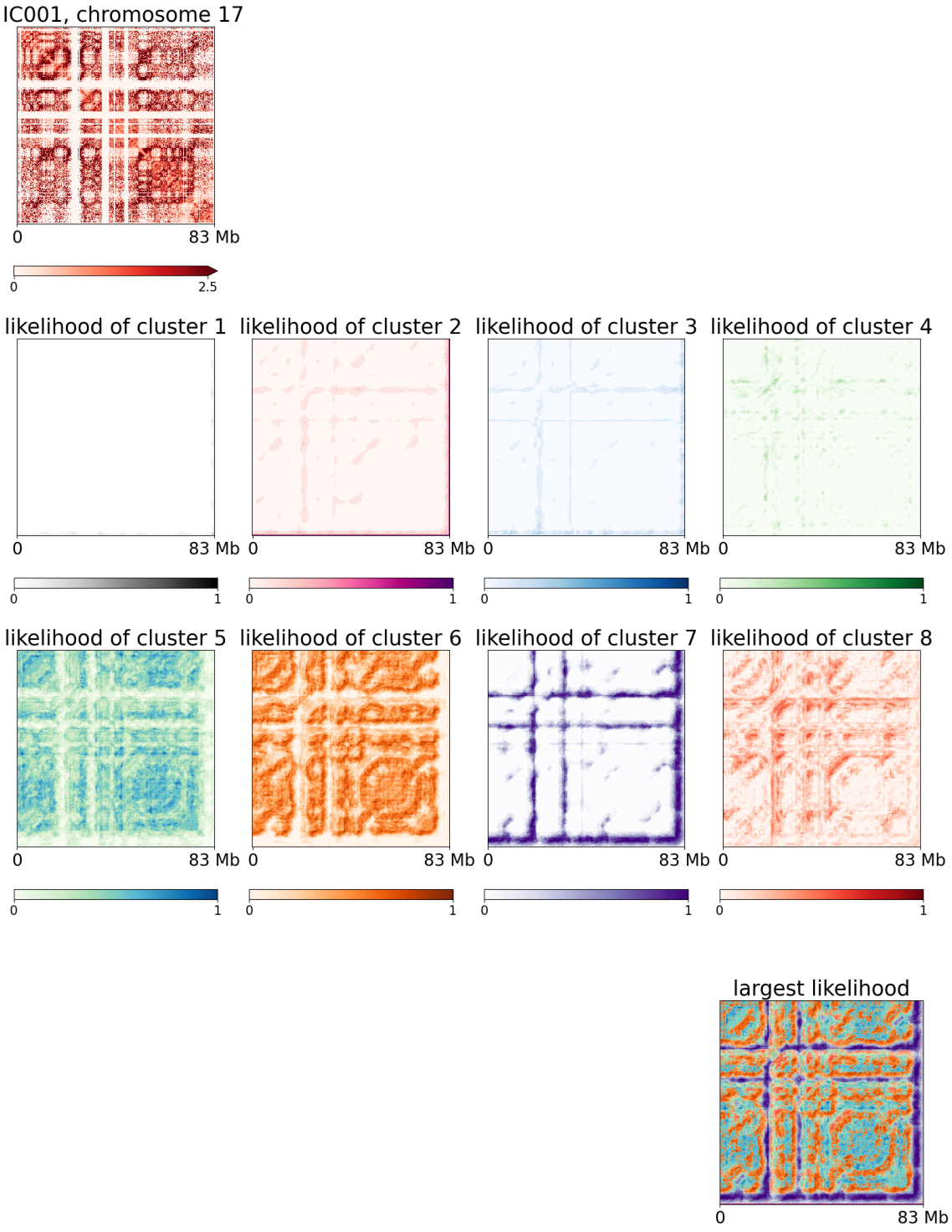

Figure 4.22: Interaction matrix of experiment HIC001, chromosome 17, colored according to the likelihood of each bin belonging to one of the clusters obtained by the Ward hierarchical clustering on the latent space representations of the training dataset. The last panel shows the largest likelihood among those computed.

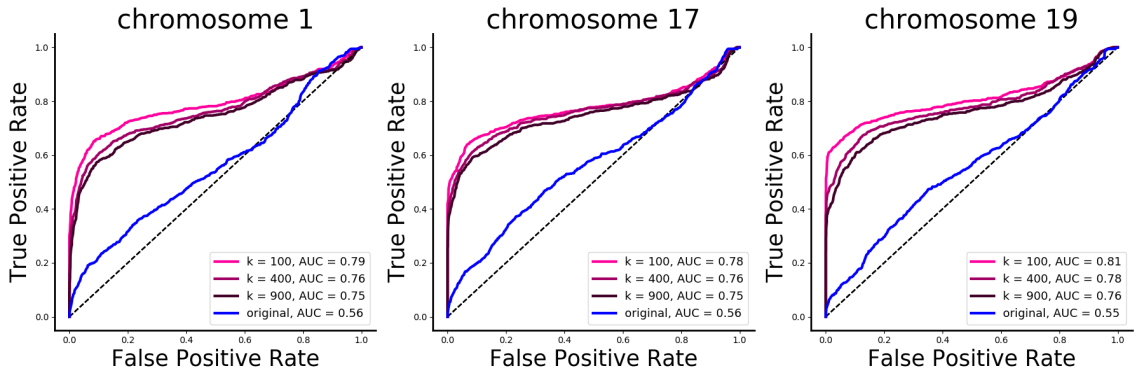

Figure 4.23: ROC curves computed from latent matrices at different sizes of the latent space, in different chromosomes. The ROC curves computed from the original matrices are shown for comparison, and AUC scores are provided for each curve.

The first original cluster, corresponding to white cut-outs sampled from the masked centrosome regions, predominantly contains patterns from the first and third latent clusters.Many members of the second cluster, identifying boundary regions near the centrosomes, also belong to the first and third latent space clusters, but a stronger presence of the other latent clusters can also be observed. Finally, the third original cluster is mainly composed of members belonging to latent clusters four to eight. While a complete identification between groups of latent clusters and the original ones is not possible, these histograms show that the main division in the latent space, between cluster one to three, and four to eight, corresponds to the division between patterns near the centrosomes and others.

Finally, an analysis of the location of these patterns on a full Hi-C matrix is shown in figure 4.22, repeating the analysis in section 4.1.3: different colors indicate different clusters, and the intensity is proportional to the probability of the cut-out at those coordinates to belong to a certain cluster. The figure shows a very low overall likelihood for most clusters on this particular matrix (HIC001, chromosome 17), with the exception of clusters one, three, five, and seven. Of these, the latter is never the most likely. The others alternate in patterns similar to those of the original clusters, and the most salient feature is the presence of bright spots belonging to cluster one at the intersections between masked areas of the matrix.

4.4.3 Discrimination of biological replicates in latent space

The dimensional reduction of Hi-C maps may enhance the discriminative power of bin to bin comparisons, making the distinction between biological replicates and non-replicates easier. This is thanks to the dimensional bottleneck provided by the autoencoder, that prevents patterns of interaction which are not shared across the dataset, such as noise,

from being encoded.

In order to test whether this is the case I start by computing, for each matrix in the bulk Hi-C dataset, its latent space representation using the encoder part of the autoencoder. For each pair of latent vectors I can obtain the euclidean distance by taking the differences between corresponding bins. By doing so one obtains a distance matrix which can then be used in a number of analysis, as shown in the previous chapter.

Here I limit myself to just computing ROC curves based on the ability of the euclidean distance to distinguish between biological replicates and non replicates. Figure 4.23 shows the results, also compared to those obtained by applying the same analysis of the original matrices.

While the reconstructed matrices did not improve much the results obtained on the original matrices by this analysis, using the corresponding latent vectors leads to a visible and much larger increase of the quality ROC curve, also reflected in a quantitative improvement of the AUC score.

This improvement suggests the application of this analysis also to the tougher problem of determining biological replicates and non replicates in the context of single cell Hi-C maps, which will be the theme of the next section.

4.5 Application to single cell Hi-C

Single cell Hi-C maps possess a few features that make dividing them into different cell-types challenging [110]: their sparseness means that only a few isolated interactions are sampled for each map. On top of this, one has to face the problem of the inherent variability of the structures within the same cell-type at the single cell level.

In the previous chapter I have shown how one can employ the essential component analysis to improve ROC curves (where the reference classes have been determined beforehand through an *ad hoc* analysis).

In this section I want to explore whether a different tactic can be employed for the classification task: I will pass the single cell Hi-C matrices contained in the dataset of reference [37] through the encoder in order to obtain their low dimensional representations, and compare these latent vectors as in the previous section to obtain a distance matrix. Finally I will compute ROC curves and compare the result to the ones obtained by other methods.

4.5.1 Autoencoder training

Since the matrices to which the autoencoder must be applied is radically different (both because of their intrinsic nature, both because they are represented at 1 Mb resolution), it makes sense to re-train the autoencoder from scratch. Nevertheless, in order to try out different combinations, I employ three different trainings: the first one is simply the one already used for bulk matrices at 100 Kb; the second retrains the autoencoder, with the

same architecture, with cut-outs sampled from matrices at 1 Mb, and the same chromosomes chosen above $(1, 2, 3, 18, 19, 20$; finally, the third trains the autoencoder directly on cut-outs taken from single cell Hi-C maps (chromosomes $1, 2, 3, 4, 5$).

4.5.2 Visual inspection of reconstructed matrices

Figure 4.24 shows, side by side, the reconstruction of the same single cell matrix by the autoencoders trained in three different ways.

Reconstruction using the original training, performed on cut-outs from matrices at 100 Kb, is able to reproduce most of the contacts present in the original matrix, even between distant loci, although it smudges otherwise sharp point-like interactions. Moreover, even more importantly, it produces very visible artifacts in the shape of darker squares along the diagonal.

Training the autoencoder at 1 Mb also does not lead to good reconstructions: in this case the artifacts are less visible, but still present, and patterns are more blurred than the previous attempt.

Predictably, only the autoencoder trained directly on single cell Hi-C maps is able to faithfully reconstruct the original matrix: only in this case is the autoencoder able to reproduce minute point-wise details present in the original matrix without blurring them and without adding noticeable artifacts.

This behavior is explained by the diversity between the bulk dataset and the single cells one: greater sparsity and a greater variety in the local patterns make these Hi-C maps too different from the bulk ones to be represented through the same dimensional reduction.

4.5.3 Classification results

Distances between matrices for the three models detailed above are computed by comparing latent space vectors. As usual I obtained these by dividing the original raw matrix into cut-outs and passing them through the encoder. Then, their latent representations, in matricial shape, are juxtaposed in order to obtain the full latent vector corresponding to the initial matrix.

Figure 4.25 shows the results for the three models, as well as those obtained by taking the euclidean distance between the original raw matrices, and the one computed in the previous chapter using the spectral components analysis.

Despite efforts to train these models, the results show that computing distances between matrices using their latent space representations worsens the quality of the ROC curves, instead of improving it. In fact all three discriminators based on the models yield ROC curves which are compatible with the one expected for a random classification.

This means that, contrary to what happens for bulk Hi-C maps, dimensional reduction is not able to single out biological information in the case of single cell Hi-C. While the autoencoder explicitly trained on datasets extracted from single cell Hi-C maps is able

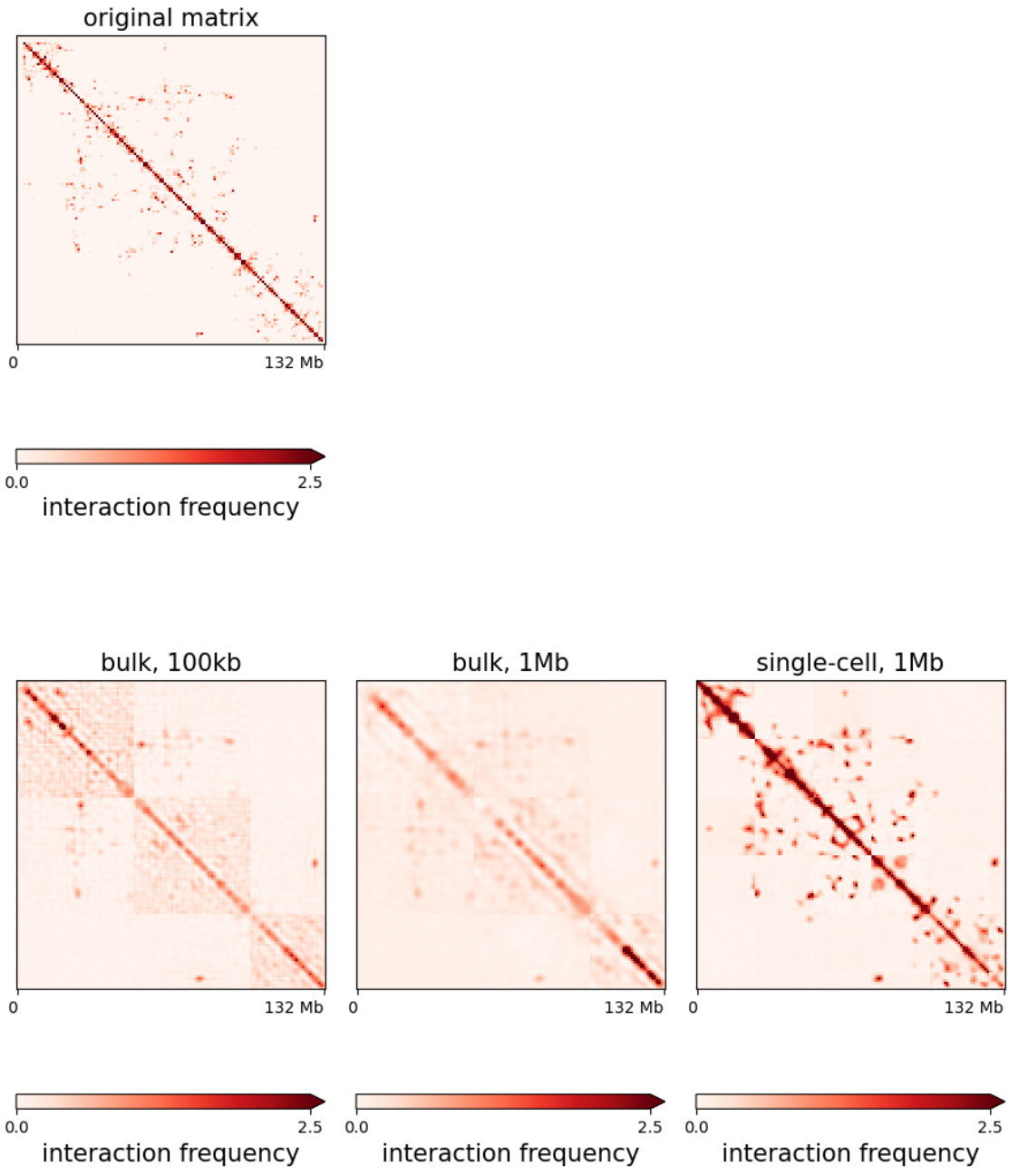

Figure 4.24: Visualization of the reconstruction of a single cell Hi-C matrix by three autoencoders trained on different datasets.

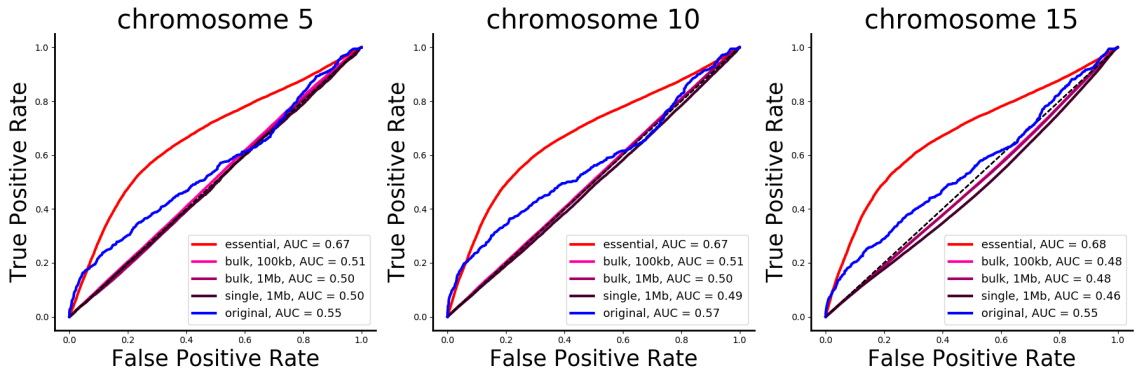

Figure 4.25: ROC curves computed from latent matrices produced by the three autoencoders trained on different datasets. The ROC curves computed from the original matrices and those obtained by essential component analysis are shown as comparison, and AUC scores are provided for each curve.

to correctly reconstruct the initial patterns, it still cannot solve problems linked to the intrinsic variability of the patterns contained in single cells, even among cells which are at the same stage of their life cycle, and the sparseness of the matrices.

It is possible that a deeper architecture, able to reduce the dimensionality even more, would be able to better understand the structure underlying these data, but this goes beyond the scope of the present work.

4.6 Comparison with essential component analysis

Previous sections have shown that the action of the autoencoder on Hi-C maps is remarkably similar, at least from a qualitative point of view, to that of the essential component analysis presented in chapters 2 and 3. Even if their objectives are different, both enucleate a small set of features from the matrices, which are then used to reconstruct the map. In the case of the essential component analysis, these features are the top eigenvalues and eigenvectors, and the operation of truncating the spectral sum which defines the Hi-C map is akin to reducing its dimensionality and then trying to restore the original Hi-C map, cleaned of the aspecific part of the spectrum. On the other hand, the autoencoder applies a *divide and impera* strategy in which small cut-outs containing local patterns extracted from an Hi-C maps are passed through non-linear layers of neurons in order to obtain a low dimensional representation, which is then expanded back to the original size: thanks to the learning procedure, carried out on a large dataset of such local patterns, this dimensional bottleneck is able to stop atypical features, such as noise, from being reproduced on the

other side.

The results are reconstructed matrices (called essential matrices in the case of essential component analysis) characterized by sharper and less noisy patterns in both cases. And in both cases this improves classification of biological replicates and non replicates.

The question arises: are these diverse algorithms doing, essentially, the same thing? What are the main differences between the two approaches? How do their results compare when applied in a common setting?

To answer these questions, one first has to define the playing field on which the two models can be compared. In this section I will consider the dataset of bulk Hi-C matrices used throughout this chapter and the previous ones. Since the natural context of application of essential component analysis is on OoE normalized matrices, here I will consider such matrices. Because the autoencoder is trained on local patterns, where the influence of the distance dependence is limited, I will not retrain it. The results will show that retraining is in fact not necessary.

First I will start by comparing the spectral properties of Hi-C matrices before and after having passed through the autoencoder. Next the two models, autoencoder and essential component analysis, will be used to classify pairs of experiments as biological replicates or non replicates, computing distances and ROC curves, which I will use to quantify the results. In the case of the autoencoder, distances will be computed both between reconstructed matrices and their latent space representations, as two separate models.

4.6.1 How does the spectrum changes?

Here I explore the action of the autoencoder on the spectrum of Hi-C maps: in order to do so, I apply it to a matrix so as to obtain its reconstruction and compute their spectral properties, i.e. their eigenvalues and eigenvectors.

Figure 4.26 shows part of the spectrum of experiment HIC001, chromosome 17, before and after the application of the autoencoder, as well as their ratio, for different sizes of the latent space. The effect of the autoencoder is, in general, to suppress the higher order eigenvalues, with the size of its latent space determining where the detachment from the original spectrum occurs: the smaller the dimension of the latent space size, the sooner visible deviations start to appear. On closer inspection of the ratio between reconstructed and original eigenvalues, one finds that the top ~ 20 eigenvalues are the most conserved ones, while those of larger order become more and more suppressed. In particular, one can notice that, among the autoencoders, the one with $k = 25$ is the most dissimilar, as it does not completely preserve its top eigenvalues.

This seems to confirm the fact that, similarly to the essential component analysis, the autoencoder also privileges the first eigenspaces in order to reproduce the salient structural features of the matrix. However, the autoencoder does not remove non-essential eigenspaces, although their weight is diminished with respect to the original matrix.

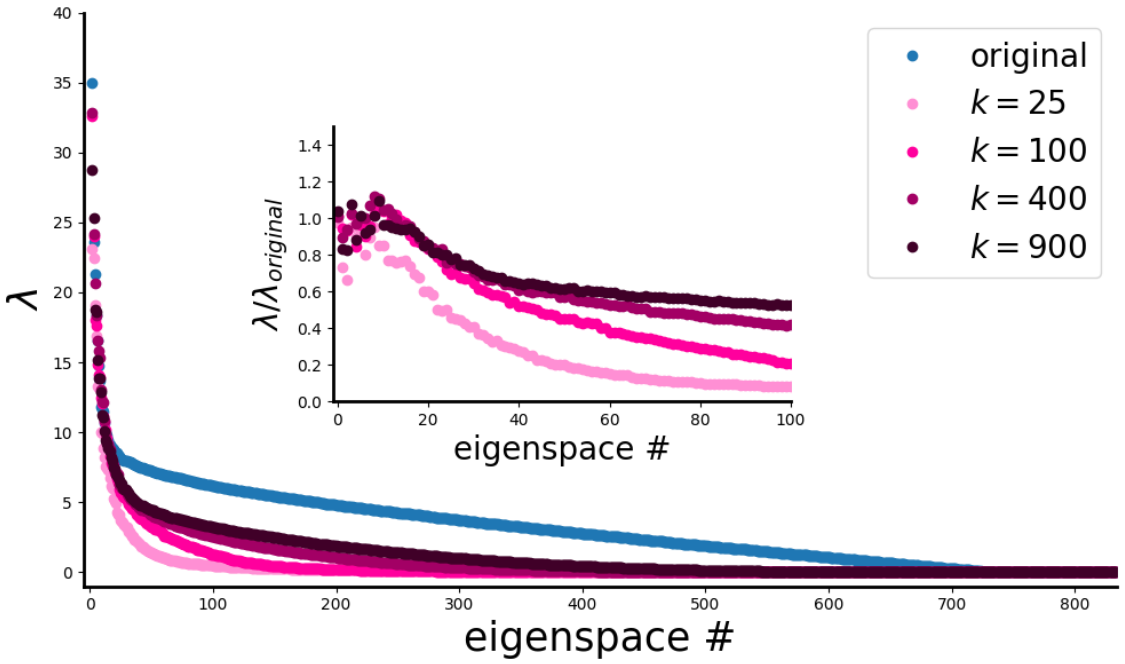

Figure 4.26: Comparison of the spectra of HIC001, chromosome 17 (normalized by its distance dependence), before and after applying the autoencoder, for different values of the latent space size. The inset shows the ratio between the eigenvalues obtained after reconstruction and those of the original matrix.

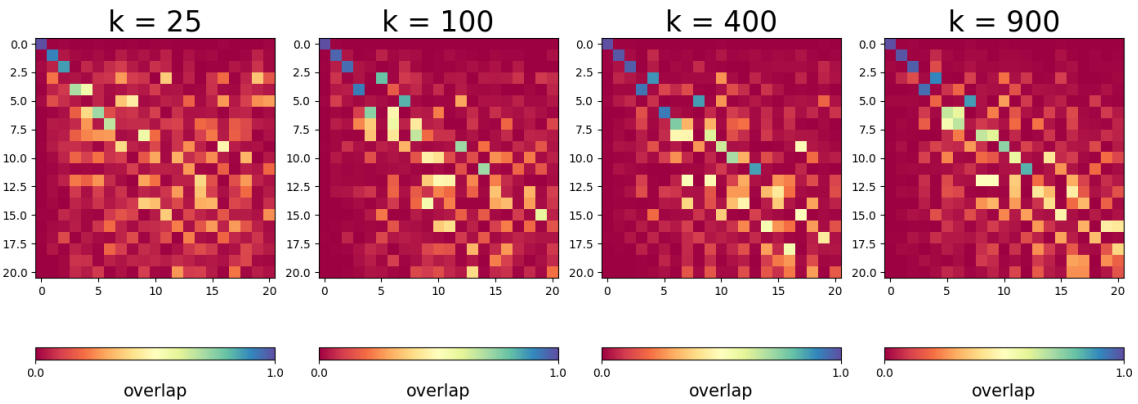

Figure 4.27: Overlaps between the top eigenvectors of HIC001, chromosome 17, before and after passing through the autoencoder for different sizes of its latent space.

One also has to verify whether the eigenspaces maintain their identity while going through the autoencoder: are the eigenvectors before the dimensional reduction similar to the reconstructed ones? Or are the reconstructed eigenvectors mixtures of the original ones? Do swaps in the order of the eigenspaces occur?

In order to answer these question, figure 4.27 shows the matrix of the overlaps between top eigenvectors of HIC001, chromosome 17, before and after the action of the autoencoder, for different sizes of its latent space. Only the top three eigenvectors completely conserve their identity, while others become mixtures of an increasingly larger number of the eigenvectors of the original matrix. The effect is more accentuated as the size of the latent space of the autoencoder decreases. This confirms, again, the importance of the structural features encoded in the top eigenvalues.

In order to be more quantitative in my assessment, I introduce the root mean-square inner product (RMSIP) [17], given by

$$RMSIP = \sqrt{\frac{1}{10}\sum_{i,j}^{10} v_i \cdot w_j} \tag{4.8}$$

where v_i and w_j are the i-th eigenvector of the original matrix and the j-th eigenvector of the reconstructed matrix, respectively. This is a standard score used to quantify the similarity between the essential spaces of two matrices, i.e. their top 10 eigenspaces. It has the important property of being rotationally invariant, so that trivial symmetries of the two essential spaces are taken into account. This property, however, means that there exist some set of vectors $\{v'\}$ and $\{w'\}$, given by linear combinations of the eigenvectors

$\{v\}$ and $\{w\}$, for which the following properties hold true [17]:

- A basis vector of one set is orthogonal to all basis vectors of the other except the one with the same index
$$v'_i \cdot w'_j = \delta_{i,j}\alpha_i \tag{4.9}$$
- the index provides a natural ordering of the basis vectors in terms of decreasing mutual consistency;
$$if\ i < j \Rightarrow \alpha_i > \alpha_i \tag{4.10}$$

These optimal basis can be used to characterize the consistency of two matrices at a finer level. This allows to monitor how consistent the optimally mixed spaces are at the level of the single pair of vectors, rather than giving a global measure.

Figure 4.28 shows the overlaps of optimally mixed eigenvectors for experiment HIC001, chromosome 17, for different sizes of the latent space: while for $k \geq 100$ the results are similar, with most optimally mixed vectors being consistent before a drop of the overlap on the last one, if one uses an autoencoder with $k = 25$, one observes a gradual decline of the overlaps as the index progresses, revealing a wholly different behavior. Moreover as the latent space is shrunk from $k = 900$ to $k = 100$, the RMSIP score decreases from ~ 0.89 to ~ 0.87, while further reducing its dimension to $k = 25$ leads to a drop in the RMSIP to ~ 0.77. It also shows the overlaps between optimally mixed vectors as k changes.

These observations further corroborates the idea that at least 100 degrees of freedom (corresponding to $k = 100$) are necessary to describe the dataset of the local patterns, and the quality of the reconstruction rapidly deteriorates as one descends to lower values.

4.6.2 Classification of biological replicates in OoE normalized matrices

Carrying out classification tasks on latent space representations of matrice has been shown to improve the quality of the ROC curves, even if the results do not reach the same AUC scores as tools explicitly designed to tackle this problem. However, it is noteworthy to notice that the autoencoder was applied to raw matrices, which are usually treated before applying classification tools: not only are they normalized to remove sampling biases, but also to explicitly take into account the genomic distance dependence of the interaction frequencies [14].

Hence it is interesting to apply the autoencoder to OoE normalized matrices, which were also used for the essential component analysis in the previous chapter, and compare the results on classification tasks. I stress that the autoencoder is not re-trained on normalized matrices.

For each matrix in the dataset I compute its latent space representation for $k = 100$. I compute the distance between each pair of experiments and use the resulting distance matrix to obtain ROC curves.

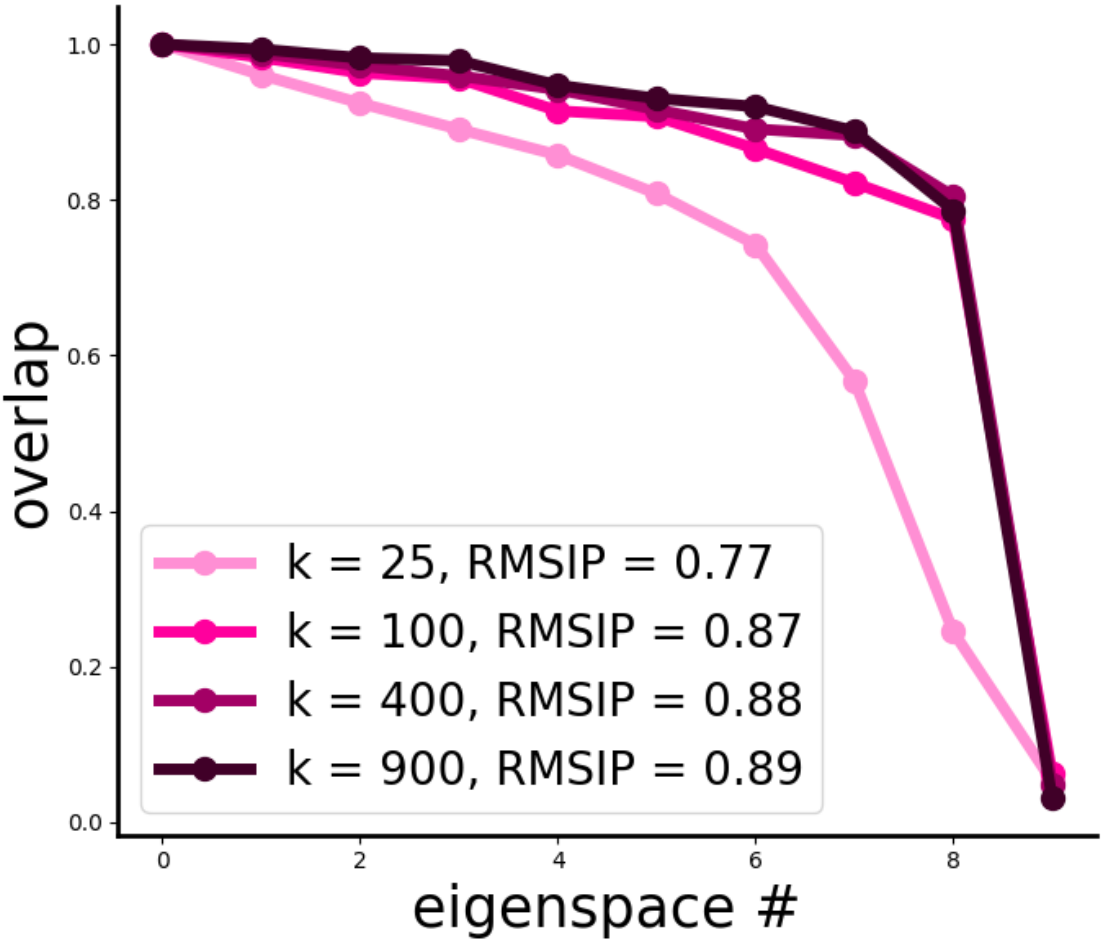

Figure 4.28: Overlaps between optimally mixed vectors obtained from the top 10 spaces, for different sizes of the latent spaces of the autoencoders. The RMSIP score is given in the legend.

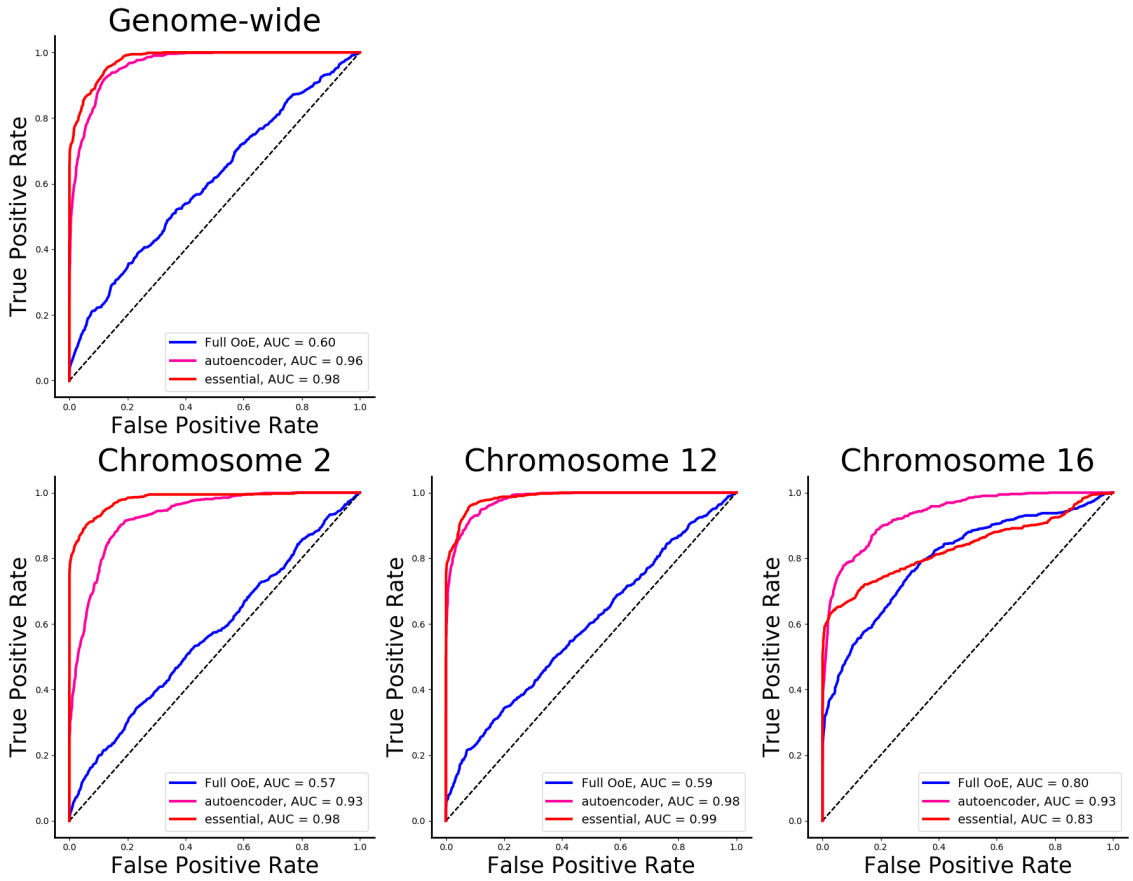

Figure 4.29: Comparison of ROC curves obtained by computing distances between full OoE matrices, their latent space representations ($k = 100$), and their essential components.

Figure 4.29 shows the results for different chromosomes, both contained in the training dataset and outside, as well as the genome-wide results obtained by combining the single chromosome distances: the ROC curves for latent space representations are plotted, alongside the results obtained in the previous chapter from essential component analysis and the baseline euclidean distances between full OoE matrices.

The genome-wide results reveal that, as expected, the quality of the ROC curves improves with respect to those plotted in figure 4.23. They greatly outperform ROC curves relative to full matrices, and this holds true even when one singles out each individual chromosome. Most interestingly, the autoencoder is only slightly outperformed by essential component analysis in the genome-wide case, and on some chromosomes the relationship is reversed.

These results are competitive with respect to those obtained by other published methods (such as those plotted in 3.13 [12, 13, 77]) and confirm the fact that dimensional reduction, in one shape or another, is key to successfully understanding the spatial features encountered in Hi-C maps. Classification tasks also provide an interesting example of an application of this procedure in which using low dimensional latent space representations can be more advantageous than carrying out the analysis in the original high dimensional space.

It is also important to consider, however, the number of degrees of freedom used to obtain these results by the autoencoder and by the essential component analysis: even without an explicit count for each chromosome, one can immediately see that the number L of degrees of freedom of the former model scales like $L \propto N^2$, where N is the linear size of the matrix, while the latter scales as $L \propto N$, because the number of components is fixed. As such, the autoencoder, even with the dimensional reduction, uses a much larger number of degrees of freedom to achieve its results, and the most efficient choice for classification remains the essential component analysis.

4.7 Summary and Conclusion

In this chapter I tackled the problem of dimensionality reduction and compression of Hi-C maps: it is known that local patterns found in interaction matrices present similarities across different chromosomes, cell-types, and even organisms[2]. They can be regarded as the common building blocks of Hi-C maps, and these regularities can be exploited to encode a low dimensional representation of the matrices, useful both for storage purposes and to enhance certain analyses.

The problem is then how to learn this encoding of the local patterns and how to restore the original information. Autoencoders are a family of neural networks architectures that provide both an encoding part[105], which derives a low dimensional latent space representation of the input, and a decoding part, which restores data to the original size. Deep learning techniques, and specifically autoencoders, have been previously applied to

Hi-C maps in different contexts, that of super-resolution, and provide a sound solution for the problem at hand.

Specifically, the model I devised involves sampling small cut-outs from Hi-C maps, the local patterns, to build a training dataset for the autoencoder: only some chromosomes are used in collecting the samples, so that the others can be employed to test over-fitting. An analysis of the training dataset reveals that its intrinsic dimension is around 100, and that some clusters are present: however these subdivision do not represent a local variance between chromosomes or cell types, but rather structural properties of the matrices, linked to the regions masked by the presence of centromeres. The network base architecture is a variational autoencoder, specifically suited for the purpose of generating images from a regularized latent space, and is modified to take into account the symmetry of the matrices. Different sizes of the latent space have been used in order to test its limits.

The first test of this model is about its ability to restore latent space representation to the original size, reproducing the original matrix. I compared the results with PCA, that can be regarded as a linear autoencoder, and Gaussian smoothing, finding that the autoencoder outperforms both (as measured by the MSE reconstruction score). Moreover, reconstructed matrices preserve structural features such as TADs at the same level as biological replicates. The biological information about cell-types is also preserved, as proven by the performance of classification tasks (measured by the AUC score of ROC curves).

Then I explored the latent space of the autoencoder. The low dimensional representations of the cut-outs used in the training set reveal a richer structure than what was observable in the original high dimensional space: the dendrogram of figure 4.18 shows a larger number of clusters separated by clear gaps according to the Ward score. However an inspection of the members of these clusters reveals that they are compatible with those found in the original space, and they represent their subdivisions. The latent space representations of the matrices can also be compared for classification purposes and achieve an higher degree of accuracy with respect to the original matrices, as proven by the ROC curves obtained.

This suggests that the autoencoder may also be able to help in the more difficult problem of classifying single cell Hi-C maps. However, despite trying different training sets (the original one, one sampled from bulk matrices at 1Mb resolution, and one sampled directly from single cell Hi-C maps) in order to better adapt the autoencoder to the different context, the discriminator was not able to distinguish between biological replicates and non replicates. This is probably due to the sparseness and inherent variability of single cell matrices, although a deeper autoencoder might be able to successfully operate even in this environment.

Finally, I compared the action of the autoencoder on the matrices to that of essential component analysis detailed in the previous chapter. To do so I applied the autoencoder to matrices where the genomic distance dependence of the interactions had been normalized out. Interestingly, the autoencoder naturally enhances the top eigenspaces and preserves their structure, with minimal mixing between their eigenvectors. On the other hand, the

rest of the eigenvectors become randomly mixed after the application of the autoencoder.

The autoencoder is also able to compete with essential component analysis in classification tasks, although it does so by using an overwhelmingly larger number of degrees of freedom. In particular, the comparison of latent space vectors yields higher quality discrimination of biological replicates and non replicates, showing that not only some analyses can be performed in the latent space, but they actually benefit from it.

In conclusion, this work is mean to be a proof of concept for compression algorithms based on deep learning that exploit the local regularities of patterns found in Hi-C maps. Lossy compression allows for a 25-fold reduction in the number of degrees of freedom of a matrix with almost no change in the reconstruction performance. The autoencoder used for this task was also able to nicely generalize to instances outside the training set, without any significant increase in the reconstruction error.

Interestingly, not only compression would allow to more easily handle large Hi-C datasets, but it also improves the results of certain analyses. In particular, comparisons of the latent space representations of Hi-C maps score better than the originals in classification tasks, and when the autoencoder is applied to OoE matrices, it obtains results comparable to other published methods.

Summary and Conclusions

This work presents advanced spectral and deep learning methods to analyze Hi-C contact matrices. The common ground is the ability of these tools to improve the performance of a number of tasks such as TAD calling and classification by considering a lower-dimensional representation of the original data.

The first part investigates the spectral properties of Hi-C matrices: while the importance of the first non-trivial eigenspace had been known since the inception of Hi-C experiments [15, 7], a systematic study of the whole spectrum was lacking.

I find that a large portion of the eigenspaces display the same properties of random matrices, and can be explained by a background of aspecific interactions which do not depend on cell-type. Removing the aspecific part by way of a spectral filter reveals the essential component of the Hi-C matrix, containing only the top eigenspaces and characterized by sharper interaction patterns strongly correlated with high quality (non processed) experiments.

Employing the essential components significantly enhances structural and biological analyses, as shown by the performance in TAD calling and cell-type classification tasks.

In the second part I exploit the fact that local patterns are shared across chromosomes and cell-types to learn a lower-dimensional representation of Hi-C matrices by way of a variational autoencoder [105]. Autoencoders and other deep neural networks have been employed in various analyses on Hi-C maps, but the dimensional reduction capabilities of such tools and their effects on matrices still remain unexplored.

Not only does the autoencoder presented here offer a 25-fold compression of Hi-C data, opening the way for higher performance analyses of high resolution matrices and large dataset: low dimensional representations of matrices better capture the biological information about cell-types, and their pair-wise comparisons allow for better classification results.

This shows a clear pattern in which reducing the dimensionality of Hi-C data, for example through spectral or deep learning methods, is a needed step in obtaining high quality reproducible results in tasks assessing structural and biological properties.

In perspective, one area which has presented significant challenges is that of single cell Hi-C maps: their sparseness and peculiar characteristics make the tougher to tackle with standard tools. Spectral methods, with *ad hoc* modifications, are able to improve

the baseline results of classification tasks, but not to the same degree as for bulk matrices; dimensional reduction through the autoencoder is not able to correctly capture the properties of single cell maps and fail to correctly classify them.

However single-cells, given their sparseness, could greatly benefit from data compression schemes able to isolate robust biological and structural features from random variability. Developing specific tools to address the unique challenge they pose could be the next step building on the results presented in this work.

Appendix A

Spectral Methods

A.1 Hic-Spector

HiC-Spector [12] was the first method to offer a metric to quantify the reproducibility of two Hi-C experiments based on the spectral properties of their contact maps. To do so, each contact matrix W is first converted into a laplacian L:

$$L = D - W, \tag{A.1}$$

where D is a diagonal matrix whose non-null elements are defined as $D_{i,i} = \sum_j W_{i,j}$, which, in the context of Hi-C maps, is the experimental coverage of the bins. Furthermore, the laplacian matrix thus obtained is normalized again by applying the transformation

$$\mathcal{L} = D^{-\frac{1}{2}} L D^{-\frac{1}{2}}. \tag{A.2}$$

This ensures that 0 is an eigenvalue of $\mathcal{L}$, so that the set of eigenvalues of $\mathcal{L}$ given by $\{0 \leq \lambda_1 \leq ... \leq \lambda_{n-1}\}$, where n is the linear size of the matrix, is called the spectrum of $\mathcal{L}$. To each eigenvalue one can associate the corresponding eigenvector, so that one obtains the ordered set $\{v_0 \leq v_1 \leq ... \leq v_{n-1}\}$ of $\mathcal{L}$ eigenvectors.

Given two experiments with contact maps W^A and W^B respectively, HiC-Spector seeks to quantify the similarity between them by decomposing the laplacian corresponding to each matrix (i.e. Lap^A and Lap^B respectively) into its spectral components, and compare the ordered eigenvectors thus obtained. This is done via a distance metric

$$S_d(A, B) = \sum_{i=0}^{r-1} ||v_i^A - v_i^B||, \tag{A.3}$$

where $||\bullet||$ represents the Euclidean norm, and r is a parameter of the model which sets the number of leading eigenvectors of $\mathcal{L}^A$ and $\mathcal{L}^B$ that one needs to compare. The original paper notices that the distance between higher-order eigenvectors is essentially identical

to the one between unit vectors whose components are randomly sampled from a gaussian distribution, so they can be discarded. It then sets $r = 20$ as a rule of thumb that works for practical purposes.

The mathematical intuition behind this method is that the normalized laplacian $\mathcal{L}$ is closely related to a random-walk process happening on the underlying graph W, so that lower-order eigenvectors of $\mathcal{L}$ can be thought of as the steady state distribution and the slowly decaying modes of diffusion on said graph.

By comparing the spectral properties of the matrices, Hi-C Spector was able to capture global features of the Hi-C maps: as a results it can better separate biological replicates, i.e. experiments on the same cell-type, and non replicates.

However Hi-C Spector does not allow one to directly compare the underlying patterns found in Hi-C maps for differential analysis, as it only offers a score of the reproducibility of the two maps. Nor does it allow to obtain an explicit relationship between the lower-order, more important eigenvectors, and statically significant patterns found in the underlying Hi-C maps.

Moreover, the method uses a strong underlying assumption that the eigenvectors of two different matrices W^A and W^B are always ordered in the same way, so that one only needs to compare eigenvectors of the same order. This would imply that a certain feature of Hi-C maps is always encoded in the same eigenvector, but this is not necessarily true: since the eigenvalues of the matrix spectrum are often similar to each other, small perturbations of a matrix can lead to switching between two or more eigenvectors. It is easy to find examples where eigenvectors of different orders are strongly correlated, while little or none similarity is present between those of the same order. Furthermore, there is no assurance that eigenvectors preserve their identity in different matrices: a certain feature can be encoded by just one eigenvector in a matrix, and be spread between two or more eigenvectors in another.

Nevertheless the method, despite these simplifying assumptions, is able to obtain the objective that it set out to: finding a metric distance to compare Hi-C experiments and assess the degree to which they are reproduced. Then, the fact that some aspects can be improved on makes one hopeful to gain even better insight with methods founded on the same intuition.

A.2 Genome DISCO

Genome DISCO [13] represents the contact maps of Hi-C experiments as a network of interacting loci, with adjacency matrix A, where each node i is a genomic locus of a specified resolution and size in terms of nucleotides. The weight of each edge between pairs of loci-nodes is then given by the normalized, experimental contact frequency between them. This matrix is then converted into a transition transition probability matrix, so that all rows sum to 1.

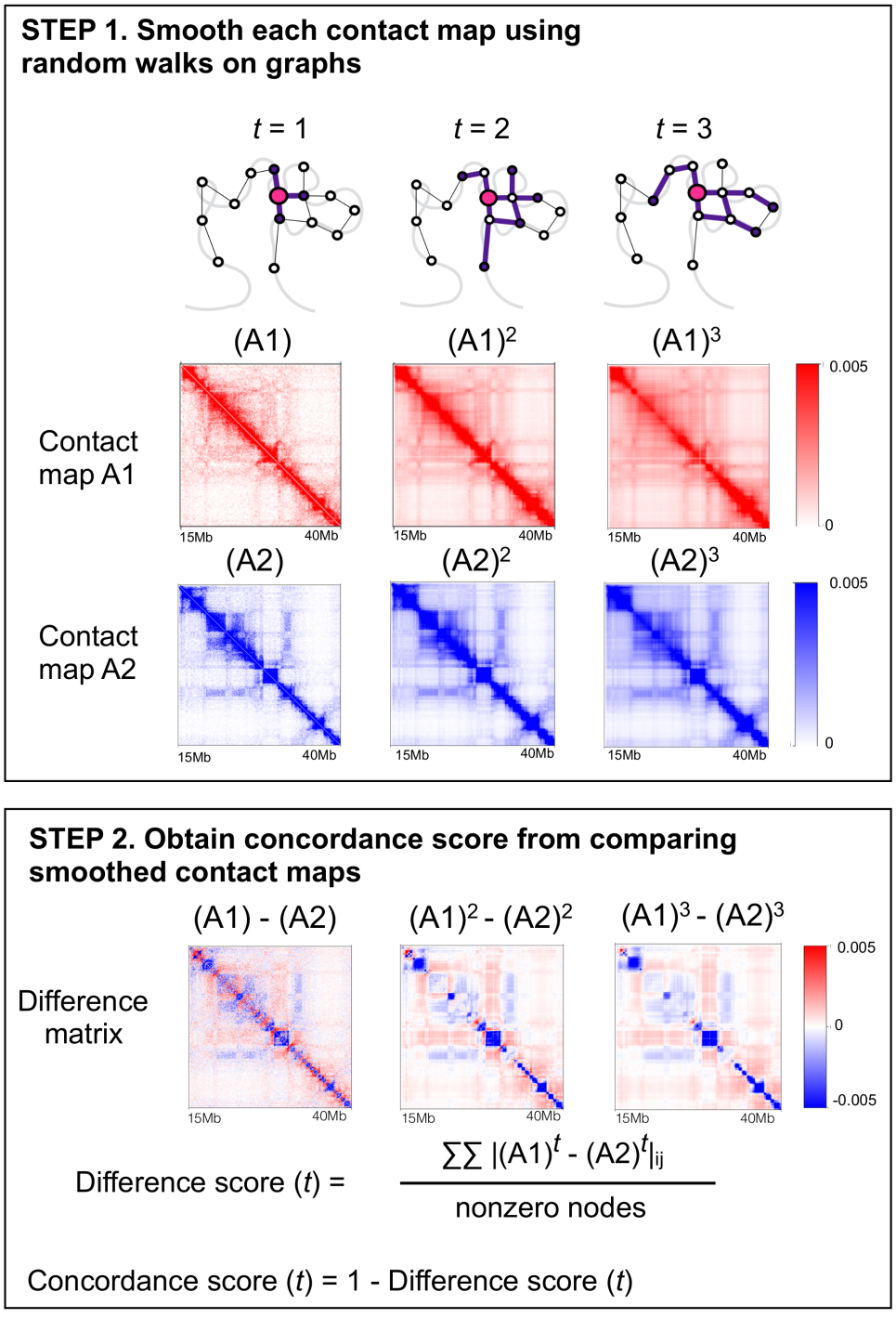

Figure A.1: **reproduced from Ursu et Al.** An overview of GenomeDISCO: The first step to compare matrices is smoothing them through a random walk, the second one is to take the element by element difference between them and compute the concordance score. The results depend on the level of smoothing, governed by the parameter t, i.e. the number of steps of the random walk.

To estimate a concordance score between Hi-C experiments, Genome DISCO independently denoises each contact map by using random walks. For each node i in the contact map, one computes the probability of reaching another node j with a t-steps random walk. The more well connected the two nodes are, i.e. the more high-probability paths in the network connect them, the higher the confidence that the two loci are connected. On the other hand, if only low probability paths exist between two loci, the probability of their contact being due to background ligation noise increases, and their interactions can be disregarded with respect to those between more well-connected loci.

Then, in order to obtain the contact probability between i and j after t steps of the random walk, one only needs to consider the (i, j)-th entry of the t power of the transition matrix A^t, i.e. $(A^t)_{i,j}$.

This procedure allows one to remove much of the experimental noise present in the matrix, so that pairwise comparisons of the matrix elements become more significant. In fact one could argue that $(A^t)_{i,j}$ does not only contain the local information about the contact frequency of the two genomic loci i and j, but is also informed by the behavior of neighboring loci in the 3D structure of the chromosome. Following the work by Ursu et Al., one can then compute the distance between two Hi-C experiments as the L^1 distance between the transition probability matrices

$$d_t(A, B) = \frac{\sum_{i,j} |(A^t)_{i,j} - (B^t)_{i,j}|}{N_{nonzero} = \frac{1}{2}(|\{A_i| \sum_j A_{i,j} > 0\}| + |\{B_i| \sum_j B_{i,j} > 0\}|)}. \tag{A.4}$$

The authors did not explicitly address the link with spectral methods, focusing instead on the more physical concept of random walks on networks. However one can notice that, mathematically, the procedure proposed by Ursu et Al. is equivalent to suppressing the higher-ranked eigenvalues of the transition probability matrix spectrum. In fact, if one diagonalizes the probability transition matrix A, obtaining a matrix whose non-null diagonal elements are given by the eigenvalues $\{1 \geq \lambda_1 \geq ... \geq \lambda_{n-1}\}$. Then, rising this matrix to the t-th power means rising each eigenvalue to that same power, so that the non-null elements become $\{1 \geq \lambda_1^t \geq ... \geq \lambda_{n-1}^t\}$, which, on account of the original eigenvalues being smaller than 1, will become smaller and smaller as t increases, except for the first one, with the lower-ranked eigenvalues decreasing more slowly than higher-ranked eigenvalues. If one goes back to the original matrix, this means that the importance of the higher-ranked eigenvectors in determining the observed patterns rapidly decreases with t: just as with Hic-Spector a large part of the spectrum is effectively cut off from the analysis.

Appendix B

Optimality of the essential component with respect to the Frobenius norm

If one considers the Frobenius norm [111, 112] defined by

$$||A|| = \sqrt{\sum_{i,j} A_{i,j}^2} \tag{B.1}$$

one can show that A^{ess} is indeed the best approximation of rank n^* of A, meaning that $\varepsilon = ||A - A^{ess}||$ is minimum.

To show this, one can use the fact that $||\cdot||$ is invariant with respect to unitary transformations to decompose A into three matrices

$$A = U\Sigma U^\dagger, \tag{B.2}$$

where Σ is the diagonal matrix. It is easy to see then that $||A|| = \sqrt{\sum_k \lambda_k^2} = \sqrt{Trace(A^2)}$. Then, if we consider the essential matrix of rank r we obtain the decomposition

$$A^{ess} = U\Sigma_r^{ess}U^\dagger, \tag{B.3}$$

where Σ_r^{ess} is the diagonalized essential matrix having the r highest eigenvalues of A along its diagonal, followed by $K - r$ zeros. The Frobenius norm of their difference is given by

$$\begin{aligned} \varepsilon_r = ||A - A^{ess}|| = \\ = ||U(\Sigma - \Sigma_r^{ess})U^\dagger|| = \\ = ||\Sigma - \Sigma_r^{ess}|| = \sqrt{\sum_{k=r+1} \lambda_k^2}. \end{aligned} \tag{B.4}$$

Consider an arbitrary matrix B of rank r and ε_B, i.e. the norm of $(A - B)$: I need to prove ε_r is the minimum value of ε_B, based on the minimization of $||A - B||$. To do so, assume that B is already the matrix of rank r giving the minimum value of ε_B, and its decomposition can be written as

$$B = V\Sigma_B V^\dagger = V \begin{bmatrix} \hat{\Sigma}_B & 0 \\ 0 & 0 \end{bmatrix} V^\dagger, \tag{B.5}$$

where $\hat{Sigma_B}$ is a diagonal matrix holding the r non-zero eigenvalues of B. I then define a new matrix C given by

$$C = V^\dagger A V = \begin{bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{bmatrix}, \tag{B.6}$$

where C_{11} is a $r \times r$ matrix. Then one has

$$\begin{aligned} \varepsilon_B = ||A - B|| = \\ = ||V^\dagger(A - B)V|| = \\ = ||C - \Sigma_B|| = ||C_{11} - \hat{\Sigma}_B|| + ||C_{12}|| + ||C_{21}|| + ||C_{22}||. \end{aligned} \tag{B.7}$$

Since I assumed that B already minimizes ε_B, one must have $C_{12} = 0$. Otherwise I would be able to build a new rank r matrix $\hat{B}$ given by

$$\hat{B} = V \begin{bmatrix} \hat{\Sigma}_B & C_{12} \\ 0 & 0 \end{bmatrix} V^\dagger, \tag{B.8}$$

so that the new Frobenius norm

$$\begin{aligned} \varepsilon_{\hat{B}=||A-\hat{B}||=} \\ = ||C_{11} - \hat{\Sigma}_B|| + ||C_{21}|| + ||C_{22}||. \end{aligned} \tag{B.9}$$

would be smaller than ε_B. With the same spirit one needs to have $C_{21} = 0$ and $C_{11} = \hat{\Sigma}_r$. Then

$$C = V^\dagger A V = \begin{bmatrix} \hat{\Sigma}_B & 0 \\ 0 & C_{22} \end{bmatrix} \tag{B.10}$$

Since $\hat{\Sigma}_B$ is diagonal, it contains the top r eigenvalues of A. In fact one has

$$||A - B|| = ||C - \Sigma_B|| = ||C_2 2|| \tag{B.11}$$

but since V is a unitary transformation one gets

$$||A||^2 = ||C||^2 = ||\hat{\Sigma}_B||^2 + ||C_2 2||^2, \tag{B.12}$$

so that

$$||C_{22}||^2 = ||A||^2 - ||\hat{\Sigma}_B||^2 = \sum_k \lambda_k^2 - ||\hat{\Sigma}_B||^2 \tag{B.13}$$

For obvious reasons, this reaches the minimum when $\hat{\Sigma}_B$ holds the top r eigenvalues of A. Then

$$\epsilon_B^2 = ||C_{22}||^2 = \sum_{k=r+1} \lambda_k^2 = \epsilon_r \tag{B.14}$$

Therefore the essential matrix A^{ess} is the best rank r approximation to A based on the Frobenius norm.

Bibliography

[1] F. B. Churchill, "August weismann and a break from tradition," vol. 1, no. 1, pp. 91–112, 1968.

[2] B. R. Lajoie, J. Dekker, and N. Kaplan, "The hitchhiker's guide to hi-c analysis: Practical guidelines," *Methods*, vol. 72, pp. 65–75, jan 2015.

[3] K. E. van Holde, *Chromatin*. Springer New York, 1989.

[4] J. Dekker, "Capturing chromosome conformation," *Science*, vol. 295, pp. 1306–1311, feb 2002.

[5] J.-M. Belton, R. P. McCord, J. H. Gibcus, N. Naumova, Y. Zhan, and J. Dekker, "Hi–c: A comprehensive technique to capture the conformation of genomes," *Methods*, vol. 58, pp. 268–276, nov 2012.

[6] Z. Zhao, G. Tavoosidana, M. Sjölinder, A. Göndör, P. Mariano, S. Wang, C. Kanduri, M. Lezcano, K. S. Sandhu, U. Singh, V. Pant, V. Tiwari, S. Kurukuti, and R. Ohlsson, "Circular chromosome conformation capture (4c) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions," *Nat Genet*, vol. 38, pp. 1341–1347, oct 2006.

[7] M. Imakaev, G. Fudenberg, R. P. McCord, N. Naumova, A. Goloborodko, B. R. Lajoie, J. Dekker, and L. A. Mirny, "Iterative correction of hi-c data reveals hallmarks of chromosome organization," *Nat Methods*, vol. 9, pp. 999–1003, sep 2012.

[8] S. S. Rao, M. H. Huntley, N. C. Durand, E. K. Stamenova, I. D. Bochkov, J. T. Robinson, A. L. Sanborn, I. Machol, A. D. Omer, E. S. Lander, and E. L. Aiden, "A 3d map of the human genome at kilobase resolution reveals principles of chromatin looping," *Cell*, vol. 159, pp. 1665–1680, dec 2014.

[9] M. Franke, D. M. Ibrahim, G. Andrey, W. Schwarzer, V. Heinrich, R. Schöpflin, K. Kraft, R. Kempfer, I. Jerković, W.-L. Chan, M. Spielmann, B. Timmermann, L. Wittler, I. Kurth, P. Cambiaso, O. Zuffardi, G. Houge, L. Lambie, F. Brancati, A. Pombo, M. Vingron, F. Spitz, and S. Mundlos, "Formation of new chromatin

domains determines pathogenicity of genomic duplications," vol. 538, pp. 265–269, oct 2016.

[10] B. Bonev, N. M. Cohen, Q. Szabo, L. Fritsch, G. L. Papadopoulos, Y. Lubling, X. Xu, X. Lv, J.-P. Hugnot, A. Tanay, and G. Cavalli, "Multiscale 3d genome rewiring during mouse neural development," *Cell*, vol. 171, pp. 557–572.e24, oct 2017.

[11] T. Yang, F. Zhang, G. G. Yardımcı, F. Song, R. C. Hardison, W. S. Noble, F. Yue, and Q. Li, "HiCRep: assessing the reproducibility of hi-c data using a stratum-adjusted correlation coefficient," *Genome Res.*, vol. 27, pp. 1939–1949, aug 2017.

[12] K.-K. Yan, G. G. Yardımcı, C. Yan, W. S. Noble, and M. Gerstein, "HiC-spector: a matrix library for spectral and reproducibility analysis of hi-c contact maps," *Bioinformatics*, vol. 33, pp. 2199–2201, mar 2017.

[13] O. Ursu, N. Boley, M. Taranova, Y. X. R. Wang, G. G. Yardimci, W. S. Noble, and A. Kundaje, "GenomeDISCO: a concordance score for chromosome conformation capture experiments using random walks on contact map graphs," *Bioinformatics*, vol. 34, pp. 2701–2707, mar 2018.

[14] E.-W. Yang and T. Jiang, "GDNorm: An improved poisson regression model for reducing biases in hi-c data," in *Lecture Notes in Computer Science*, pp. 263–280, Springer Berlin Heidelberg, 2014.

[15] E. Lieberman-Aiden, N. L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner, R. Sandstrom, B. Bernstein, M. A. Bender, M. Groudine, A. Gnirke, J. Stamatoyannopoulos, L. A. Mirny, E. S. Lander, and J. Dekker, "Comprehensive mapping of long-range interactions reveals folding principles of the human genome," *Science*, vol. 326, pp. 289–293, oct 2009.

[16] J. Dostie, T. A. Richmond, R. A. Arnaout, R. R. Selzer, W. L. Lee, T. A. Honan, E. D. Rubio, A. Krumm, J. Lamb, C. Nusbaum, R. D. Green, and J. Dekker, "Chromosome conformation capture carbon copy (5c): A massively parallel solution for mapping interactions between genomic elements," *Genome Research*, vol. 16, pp. 1299–1309, oct 2006.

[17] F. Pontiggia, A. Zen, and C. Micheletti, "Small- and large-scale conformational changes of adenylate kinase: A molecular dynamics study of the subdomain motion and mechanics," *Biophysical Journal*, vol. 95, pp. 5901–5912, dec 2008.

[18] T. Nagano, Y. Lubling, C. Várnai, C. Dudley, W. Leung, Y. Baran, N. M. Cohen, S. Wingett, P. Fraser, and A. Tanay, "Cell-cycle dynamics of chromosomal organization at single-cell resolution," vol. 547, pp. 61–67, jul 2017.

[19] G. Dong, G. Liao, H. Liu, and G. Kuang, "A review of the autoencoder and its variants: A comparative perspective from target recognition in synthetic-aperture radar images," vol. 6, pp. 44–68, sep 2018.

[20] C. Matthey-Doret, L. Baudry, A. Breuer, R. Montagne, N. Guiglielmoni, V. Scolari, E. Jean, A. Campeas, P. H. Chanut, E. Oriol, A. Méot, L. Politis, A. Vigouroux, P. Moreau, R. Koszul, and A. Cournac, "Computer vision for pattern detection in chromosome contact maps," vol. 11, nov 2020.

[21] E. Facco, M. d'Errico, A. Rodriguez, and A. Laio, "Estimating the intrinsic dimension of datasets by a minimal neighborhood information," vol. 7, sep 2017.

[22] A. S. Belmont and K. Bruce, "Visualization of g1 chromosomes: a folded, twisted, supercoiled chromonema model of interphase chromatid structure.," *Journal of Cell Biology*, vol. 127, pp. 287–302, oct 1994.

[23] M. Eltsov, K. M. MacLellan, K. Maeshima, A. S. Frangakis, and J. Dubochet, "Analysis of cryo-electron microscopy images does not support the existence of 30-nm chromatin fibers in mitotic chromosomes in situ," *Proceedings of the National Academy of Sciences*, vol. 105, pp. 19732–19737, dec 2008.

[24] I. Williamson, S. Berlivet, R. Eskeland, S. Boyle, R. S. Illingworth, D. Paquette, J. Dostie, and W. A. Bickmore, "Spatial genome organization: contrasting views from chromosome conformation capture and fluorescence in situ hybridization," *Genes Dev.*, vol. 28, pp. 2778–2791, dec 2014.

[25] M. Cremer, J. von Hase, T. Volm, A. Brero, G. Kreth, J. Walter, C. Fischer, I. Solovei, C. Cremer, and T. Cremer *Chromosome Research*, vol. 9, no. 7, pp. 541–567, 2001.

[26] M. R. Branco and A. Pombo, "Intermingling of chromosome territories in interphase suggests role in translocations and transcription-dependent associations," *PLoS Biol*, vol. 4, p. e138, apr 2006.

[27] S. Boyle, "The spatial organization of human chromosomes within the nuclei of normal and emerin-mutant cells," *Human Molecular Genetics*, vol. 10, pp. 211–219, feb 2001.

[28] W. A. Bickmore, "The spatial organization of the human genome," *Annu. Rev. Genom. Hum. Genet.*, vol. 14, pp. 67–84, aug 2013.

[29] J. H. Gibcus and J. Dekker, "The hierarchy of the 3d genome," *Molecular Cell*, vol. 49, pp. 773–782, mar 2013.

[30] K. Cullen, M. Kladde, and M. Seyfred, "Interaction between transcription regulatory regions of prolactin chromatin," *Science*, vol. 261, pp. 203–206, jul 1993.

[31] L. Ulanovsky, M. Bodner, E. N. Trifonov, and M. Choder, "Curved DNA: design, synthesis, and circularization.," *Proceedings of the National Academy of Sciences*, vol. 83, pp. 862–866, feb 1986.

[32] D. Kotlarz, A. Fritsch, and H. Buc, "Variations of intramolecular ligation rates allow the detection of protein-induced bends in DNA.," *The EMBO Journal*, vol. 5, pp. 799–803, apr 1986.

[33] J. Dekker, "Chromosome folding: Contributions of chromosome conformation capture and polymer physics," in *Modeling the 3D Conformation of Genomes*, pp. 1–18, CRC Press, jan 2019.

[34] M. Simonis, P. Klous, E. Splinter, Y. Moshkin, R. Willemsen, E. de Wit, B. van Steensel, and W. de Laat, "Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture–on-chip (4c)," *Nat Genet*, vol. 38, pp. 1348–1354, oct 2006.

[35] M. J. Fullwood, M. H. Liu, Y. F. Pan, J. Liu, H. Xu, Y. B. Mohamed, Y. L. Orlov, S. Velkov, A. Ho, P. H. Mei, E. G. Y. Chew, P. Y. H. Huang, W.-J. Welboren, Y. Han, H. S. Ooi, P. N. Ariyaratne, V. B. Vega, Y. Luo, P. Y. Tan, P. Y. Choy, K. D. S. A. Wansa, B. Zhao, K. S. Lim, S. C. Leow, J. S. Yow, R. Joseph, H. Li, K. V. Desai, J. S. Thomsen, Y. K. Lee, R. K. M. Karuturi, T. Herve, G. Bourque, H. G. Stunnenberg, X. Ruan, V. Cacheux-Rataboul, W.-K. Sung, E. T. Liu, C.-L. Wei, E. Cheung, and Y. Ruan, "An oestrogen-receptor--bound human chromatin interactome," *Nature*, vol. 462, pp. 58–64, nov 2009.

[36] M. R. Mumbach, A. J. Rubin, R. A. Flynn, C. Dai, P. A. Khavari, W. J. Greenleaf, and H. Y. Chang, "HiChIP: efficient and sensitive analysis of protein-directed genome architecture," *Nat Methods*, vol. 13, pp. 919–922, sep 2016.

[37] T. Nagano, Y. Lubling, T. J. Stevens, S. Schoenfelder, E. Yaffe, W. Dean, E. D. Laue, A. Tanay, and P. Fraser, "Single-cell hi-c reveals cell-to-cell variability in chromosome structure," *Nature*, vol. 502, pp. 59–64, sep 2013.

[38] V. Ramani, X. Deng, R. Qiu, K. L. Gunderson, F. J. Steemers, C. M. Disteche, W. S. Noble, Z. Duan, and J. Shendure, "Massively multiplex single-cell hi-c," *Nat Methods*, vol. 14, pp. 263–266, jan 2017.

[39] I. M. Flyamer, J. Gassler, M. Imakaev, H. B. Brandão, S. V. Ulianov, N. Abdennur, S. V. Razin, L. A. Mirny, and K. Tachibana-Konwalski, "Single-nucleus hi-c reveals unique chromatin reorganization at oocyte-to-zygote transition," *Nature*, vol. 544, pp. 110–114, mar 2017.

[40] F. Jin, Y. Li, J. R. Dixon, S. Selvaraj, Z. Ye, A. Y. Lee, C.-A. Yen, A. D. Schmitt, C. A. Espinoza, and B. Ren, "A high-resolution map of the three-dimensional chromatin interactome in human cells," *Nature*, vol. 503, pp. 290–294, oct 2013.

[41] A. Sanyal, B. R. Lajoie, G. Jain, and J. Dekker, "The long-range interaction landscape of gene promoters," *Nature*, vol. 489, pp. 109–113, sep 2012.

[42] W. Deng, J. Lee, H. Wang, J. Miller, A. Reik, P. D. Gregory, A. Dean, and G. A. Blobel, "Controlling long-range genomic interactions at a native locus by targeted tethering of a looping factor," *Cell*, vol. 149, pp. 1233–1244, jun 2012.

[43] I. Krivega and A. Dean, "Enhancer and promoter interactions—long distance calls," *Current Opinion in Genetics Development*, vol. 22, pp. 79–85, apr 2012.

[44] S. V. Razin, A. A. Gavrilov, E. S. Ioudinkova, and O. V. Iarovaia, "Communication of genome regulatory elements in a folded chromosome," *FEBS Letters*, vol. 587, pp. 1840–1847, may 2013.

[45] J. Nuebler, G. Fudenberg, M. Imakaev, N. Abdennur, and L. A. Mirny, "Chromatin organization by an interplay of loop extrusion and compartmental segregation," vol. 115, pp. E6697–E6706, jul 2018.

[46] F. Ay, T. L. Bailey, and W. S. Noble, "Statistical confidence estimation for hi-c data reveals regulatory chromatin contacts," *Genome Research*, vol. 24, pp. 999–1011, feb 2014.

[47] E. P. Nora, B. R. Lajoie, E. G. Schulz, L. Giorgetti, I. Okamoto, N. Servant, T. Piolot, N. L. van Berkum, J. Meisig, J. Sedat, J. Gribnau, E. Barillot, N. Blüthgen, J. Dekker, and E. Heard, "Spatial partitioning of the regulatory landscape of the x-inactivation centre," *Nature*, vol. 485, pp. 381–385, apr 2012.

[48] J. R. Dixon, S. Selvaraj, F. Yue, A. Kim, Y. Li, Y. Shen, M. Hu, J. S. Liu, and B. Ren, "Topological domains in mammalian genomes identified by analysis of chromatin interactions," *Nature*, vol. 485, pp. 376–380, apr 2012.

[49] T. Sexton, E. Yaffe, E. Kenigsberg, F. Bantignies, B. Leblanc, M. Hoichman, H. Parrinello, A. Tanay, and G. Cavalli, "Three-dimensional folding and functional organization principles of the drosophila genome," *Cell*, vol. 148, pp. 458–472, feb 2012.

[50] C. Hou, L. Li, Z. S. Qin, and V. G. Corces, "Gene density, transcription, and insulators contribute to the partition of the drosophila genome into physical domains," *Molecular Cell*, vol. 48, pp. 471–484, nov 2012.

[51] D. Filippova, R. Patro, G. Duggal, and C. Kingsford, "Identification of alternative topological domains in chromatin," *Algorithms Mol Biol*, vol. 9, no. 1, p. 14, 2014.

[52] N. Naumova, E. M. Smith, Y. Zhan, and J. Dekker, "Analysis of long-range chromatin interactions using chromosome conformation capture," *Methods*, vol. 58, pp. 192–203, nov 2012.

[53] T. Mizuguchi, G. Fudenberg, S. Mehta, J.-M. Belton, N. Taneja, H. D. Folco, P. FitzGerald, J. Dekker, L. Mirny, J. Barrowman, and S. I. S. Grewal, "Cohesin-dependent globules and heterochromatin shape 3d genome architecture in s. pombe," *Nature*, vol. 516, pp. 432–435, oct 2014.

[54] J. Fraser, C. Ferrai, A. M. Chiariello, M. Schueler, T. Rito, G. Laudanno, M. Barbieri, B. L. Moore, D. C. Kraemer, S. Aitken, S. Q. Xie, K. J. Morris, M. Itoh, H. Kawaji, I. Jaeger, Y. Hayashizaki, P. Carninci, A. R. Forrest, C. A. Semple, J. Dostie, A. Pombo, and M. N. and, "Hierarchical folding and reorganization of chromosomes are linked to transcriptional changes in cellular differentiation," *Mol Syst Biol*, vol. 11, p. 852, dec 2015.

[55] E. Crane, Q. Bian, R. P. McCord, B. R. Lajoie, B. S. Wheeler, E. J. Ralston, S. Uzawa, J. Dekker, and B. J. Meyer, "Condensin-driven remodelling of x chromosome topology during dosage compensation," *Nature*, vol. 523, pp. 240–244, jun 2015.

[56] T. B. K. Le, M. V. Imakaev, L. A. Mirny, and M. T. Laub, "High-resolution mapping of the spatial organization of a bacterial chromosome," *Science*, vol. 342, pp. 731–734, oct 2013.

[57] Y. Shen, F. Yue, D. F. McCleary, Z. Ye, L. Edsall, S. Kuan, U. Wagner, J. Dixon, L. Lee, V. V. Lobanenkov, and B. Ren, "A map of the cis-regulatory sequences in the mouse genome," *Nature*, vol. 488, pp. 116–120, jul 2012.

[58] Y. Ghavi-Helm, A. Jankowski, S. Meiers, R. R. Viales, J. O. Korbel, and E. E. M. Furlong, "Highly rearranged chromosomes reveal uncoupling between genome topology and gene expression," *Nat Genet*, vol. 51, pp. 1272–1282, jul 2019.

[59] L. Giorgetti, R. Galupa, E. P. Nora, T. Piolot, F. Lam, J. Dekker, G. Tiana, and E. Heard, "Predictive polymer modeling reveals coupled fluctuations in chromosome conformation and transcription," *Cell*, vol. 157, pp. 950–963, may 2014.

[60] F. Benedetti, A. Japaridze, J. Dorier, D. Racko, R. Kwapich, Y. Burnier, G. Dietler, and A. Stasiak, "Effects of physiological self-crowding of DNA on shape and biological properties of DNA molecules with various levels of supercoiling," *Nucleic Acids Research*, vol. 43, pp. 2390–2399, feb 2015.

[61] K. V. Bortle, M. H. Nichols, L. Li, C.-T. Ong, N. Takenaka, Z. S. Qin, and V. G. Corces, "Insulator function and topological domain border strength scale with architectural protein occupancy," *Genome Biol*, vol. 15, no. 5, p. R82, 2014.

[62] D. Racko, F. Benedetti, J. Dorier, and A. Stasiak, "Transcription-induced supercoiling as the driving force of chromatin loop extrusion during formation of TADs in interphase chromosomes," vol. 46, pp. 1648–1660, nov 2017.

[63] D. Racko, F. Benedetti, J. Dorier, and A. Stasiak, "Are TADs supercoiled?," vol. 47, pp. 521–532, nov 2018.

[64] M. Yu and B. Ren, "The three-dimensional organization of mammalian genomes," *Annu. Rev. Cell Dev. Biol.*, vol. 33, pp. 265–289, oct 2017.

[65] P. de Gennes, *Scaling concepts in polymer physics*. Ithaca [u.a.]: Cornell Univ. Pr., 1979.

[66] G. Fudenberg and L. A. Mirny, "Higher-order chromatin structure: bridging physics and biology," *Current Opinion in Genetics Development*, vol. 22, pp. 115–124, apr 2012.

[67] N. Naumova, M. Imakaev, G. Fudenberg, Y. Zhan, B. R. Lajoie, L. A. Mirny, and J. Dekker, "Organization of the mitotic chromosome," *Science*, vol. 342, pp. 948–953, nov 2013.

[68] S. A. Schalbetter, A. Goloborodko, G. Fudenberg, J.-M. Belton, C. Miles, M. Yu, J. Dekker, L. Mirny, and J. Baxter, "SMC complexes differentially compact mitotic chromosomes according to genomic context," *Nat Cell Biol*, vol. 19, pp. 1071–1080, aug 2017.

[69] J. H. Gibcus, K. Samejima, A. Goloborodko, I. Samejima, N. Naumova, J. Nuebler, M. T. Kanemaki, L. Xie, J. R. Paulson, W. C. Earnshaw, L. A. Mirny, and J. Dekker, "A pathway for mitotic chromosome formation," *Science*, vol. 359, p. eaao6135, jan 2018.

[70] F. Serra, D. Baù, M. Goodstadt, D. Castillo, G. J. Filion, and M. A. Marti-Renom, "Automatic analysis and 3d-modelling of hi-c data using TADbit reveals structural features of the fly chromatin colors," *PLoS Comput Biol*, vol. 13, p. e1005665, jul 2017.

[71] C. Lazaris, S. Kelly, P. Ntziachristos, I. Aifantis, and A. Tsirigos, "HiC-bench: comprehensive and reproducible hi-c data analysis designed for parameter exploration and benchmarking," *BMC Genomics*, vol. 18, jan 2017.

[72] J. Wolff, R. Backofen, and B. Grüning, "Loop detection using hi-c data with HiCExplorer," mar 2020.

[73] P. Soler-Vila, P. Cuscó, I. Farabella, M. D. Stefano, and M. A. Marti-Renom, "Hierarchical chromatin organization detected by TADpole," *Nucleic Acids Research*, vol. 48, pp. e39–e39, feb 2020.

[74] J. Wang, A. Chakraborty, and F. Ay, "dcHiC: differential compartment analysis of hi-c datasets," feb 2021.

[75] E. Vidal, F. le Dily, J. Quilez, R. Stadhouders, Y. Cuartero, T. Graf, M. A. Marti-Renom, M. Beato, and G. J. Filion, "OneD: increasing reproducibility of hi-c samples with abnormal karyotypes," *Nucleic Acids Research*, vol. 46, pp. e49–e49, jan 2018.

[76] G. G. Yardımcı, H. Ozadam, M. E. G. Sauria, O. Ursu, K.-K. Yan, T. Yang, A. Chakraborty, A. Kaul, B. R. Lajoie, F. Song, Y. Zhan, F. Ay, M. Gerstein, A. Kundaje, Q. Li, J. Taylor, F. Yue, J. Dekker, and W. S. Noble, "Measuring the reproducibility and quality of hi-c data," *Genome Biol*, vol. 20, mar 2019.

[77] S. Franzini, M. D. Stefano, and C. Micheletti, "essHi-c: essential component analysis of hi-c matrices," *Bioinformatics*, feb 2021.

[78] Y. Xu, T. Shen, and R. P. McCord, "3d genome structure variation across cell types captured by integrating multi-omics," sep 2019.

[79] H.-J. Kim, G. G. Yardımcı, G. Bonora, V. Ramani, J. Liu, R. Qiu, C. Lee, J. Hesson, C. B. Ware, J. Shendure, Z. Duan, and W. S. Noble, "Capturing cell type-specific chromatin compartment patterns by applying topic modeling to single-cell hi-c data," vol. 16, p. e1008173, sep 2020.

[80] T. Cremer and M. Cremer, "Chromosome territories," vol. 2, pp. a003889–a003889, feb 2010.

[81] J. Dekker and E. Heard, "Structural and functional diversity of topologically associating domains," vol. 589, pp. 2877–2884, sep 2015.

[82] T. Sexton and G. Cavalli, "The role of chromosome domains in shaping the functional genome," vol. 160, pp. 1049–1059, mar 2015.

[83] J. R. Dixon, D. U. Gorkin, and B. Ren, "Chromatin domains: The unit of chromosome organization," vol. 62, pp. 668–680, jun 2016.

[84] G. Livan, M. Novaes, and P. Vivo, *Introduction to Random Matrices*. Springer International Publishing, 2018.

[85] S. O'Rourke, V. Vu, and K. Wang, "Eigenvectors of random matrices: A survey," *Journal of Combinatorial Theory, Series A*, vol. 144, pp. 361–442, nov 2016.

[86] N. L. van Berkum, E. Lieberman-Aiden, L. Williams, M. Imakaev, A. Gnirke, L. A. Mirny, J. Dekker, and E. S. Lander, "Hi-c: A method to study the three-dimensional architecture of genomes.," *JoVE*, may 2010.

[87] H. Miura, R. Poonperm, S. Takahashi, and I. Hiratani, *Practical Analysis of Hi-C Data: Generating A/B Compartment Profiles*, pp. 221–245. New York, NY: Springer New York, 2018.

[88] *Astronomical Optics*. Elsevier, 2000.

[89] M. Mahesh, "The essential physics of medical imaging, third edition.," vol. 40, p. 077301, jun 2013.

[90] D. Pfitzner, R. Leibbrandt, and D. Powers, "Characterization and evaluation of similarity measures for pairs of clusterings," vol. 19, pp. 361–394, jul 2008.

[91] M. Zufferey, D. Tavernari, E. Oricchio, and G. Ciriello, "Comparison of computational methods for the identification of topologically associating domains," vol. 19, dec 2018.

[92] M. Rondón-Lagos, L. V. D. Cantogno, C. Marchiò, N. Rangel, C. Payan-Gomez, P. Gugliotta, C. Botta, G. Bussolati, S. R. Ramírez-Clavijo, B. Pasini, and A. Sapino, "Differences and homologies of chromosomal alterations within and between breast cancer cell lines: a clustering analysis," *Mol Cytogenet*, vol. 7, no. 1, p. 8, 2014.

[93] Y. Zhang, L. An, J. Xu, B. Zhang, W. J. Zheng, M. Hu, J. Tang, and F. Yue, "Enhancing hi-c data resolution with deep convolutional neural network HiCPlus," vol. 9, feb 2018.

[94] and John A Stamatoyannopoulos, M. Snyder, R. Hardison, B. Ren, T. Gingeras, D. M. Gilbert, M. Groudine, M. Bender, R. Kaul, T. Canfield, E. Giste, A. Johnson, M. Zhang, G. Balasundaram, R. Byron, V. Roach, P. J. Sabo, R. Sandstrom, A. S. Stehling, R. E. Thurman, S. M. Weissman, P. Cayting, M. Hariharan, J. Lian, Y. Cheng, S. G. Landt, Z. Ma, B. J. Wold, J. Dekker, G. E. Crawford, C. A. Keller, W. Wu, C. Morrissey, S. A. Kumar, T. Mishra, D. Jain, M. Byrska-Bishop, D. Blankenberg, B. R. Lajoie, G. Jain, A. Sanyal, K.-B. Chen, O. Denas, J. Taylor, G. A. Blobel, M. J. Weiss, M. Pimkin, W. Deng, G. K. Marinov, B. A. Williams, K. I. Fisher-Aylor, G. Desalvo, A. Kiralusha, D. Trout, H. Amrhein, A. Mortazavi, L. Edsall, D. McCleary, S. Kuan, Y. Shen, F. Yue, Z. Ye, C. A. Davis, C. Zaleski, S. Jha, C. Xue, A. Dobin, W. Lin, M. Fastuca, H. Wang, R. Guigo, S. Djebali, J. Lagarde, T. Ryba, T. Sasaki, V. S. Malladi, M. S. Cline, V. M. Kirkup, K. Learned, K. R. Rosenbloom, W. J. Kent, E. A. Feingold, P. J. Good, M. Pazin, R. F. Lowdon, and L. B. Adams, "An encyclopedia of mouse DNA elements (mouse ENCODE)," vol. 13, no. 8, p. 418, 2012.

[95] E. Yaffe and A. Tanay, "Probabilistic modeling of hi-c contact maps eliminates systematic biases to characterize global chromosomal architecture," *Nat Genet*, vol. 43, pp. 1059–1065, oct 2011.

[96] P. A. Knight and D. Ruiz, "A fast algorithm for matrix balancing," *IMA Journal of Numerical Analysis*, vol. 33, pp. 1029–1047, oct 2012.

[97] F. Serra, M. D. Stefano, Y. G. Spill, Y. Cuartero, M. Goodstadt, D. Baù, and M. A. Marti-Renom, "Restraint-based three-dimensional modeling of genomes and genomic domains," *FEBS Letters*, vol. 589, pp. 2987–2995, may 2015.

[98] M. Trussart, F. Serra, D. Baù, I. Junier, L. Serrano, and M. A. Marti-Renom, "Assessing the limits of restraint-based 3d modeling of genomes and genomic domains," *Nucleic Acids Research*, vol. 43, pp. 3465–3477, mar 2015.

[99] H. Tjong, W. Li, R. Kalhor, C. Dai, S. Hao, K. Gong, Y. Zhou, H. Li, X. J. Zhou, M. A. L. Gros, C. A. Larabell, L. Chen, and F. Alber, "Population-based 3d genome structure analysis reveals driving forces in spatial genome organization," *Proc Natl Acad Sci USA*, vol. 113, pp. E1663–E1672, mar 2016.

[100] R. Vilarrasa-Blasi, P. Soler-Vila, N. Verdaguer-Dot, N. Russiñol, M. D. Stefano, V. Chapaprieta, G. Clot, I. Farabella, P. Cuscó, M. Kulis, X. Agirre, F. Prosper, R. Beekman, S. Beà, D. Colomer, H. G. Stunnenberg, I. Gut, E. Campo, M. A. Marti-Renom, and J. I. Martin-Subero, "Dynamics of genome architecture and chromatin function during human b cell differentiation and neoplastic transformation," *Nat Commun*, vol. 12, jan 2021.

[101] R. Stadhouders, E. Vidal, F. Serra, B. D. Stefano, F. L. Dily, J. Quilez, A. Gomez, S. Collombet, C. Berenguer, Y. Cuartero, J. Hecht, G. J. Filion, M. Beato, M. A. Marti-Renom, and T. Graf, "Transcription factors orchestrate dynamic interplay between genome topology and gene regulation during cell reprogramming," *Nat Genet*, vol. 50, pp. 238–249, jan 2018.

[102] J. Paulsen, T. M. L. Ali, M. Nekrasov, E. Delbarre, M.-O. Baudement, S. Kurscheid, D. Tremethick, and P. Collas, "Long-range interactions between topologically associating domains shape the four-dimensional genome during differentiation," *Nat Genet*, vol. 51, pp. 835–843, apr 2019.

[103] Q. Liu, H. Lv, and R. Jiang, "hicGAN infers super resolution hi-c data with generative adversarial networks," vol. 35, pp. i99–i107, jul 2019.

[104] H. Hong, S. Jiang, H. Li, G. Du, Y. Sun, H. Tao, C. Quan, C. Zhao, R. Li, W. Li, X. Yin, Y. Huang, C. Li, H. Chen, and X. Bo, "DeepHiC: A generative adversarial network for enhancing hi-c data resolution," vol. 16, p. e1007287, feb 2020.

[105] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

[106] M. Highsmith, O. Oluwadare, and J. Cheng, "Deep learning for denoising hi-c chromosomal contact data," jul 2019.

[107] M. Highsmith and J. Cheng, "VEHiCLE: a variationally encoded hi-c loss enhancement algorithm for improving and generating hi-c data," vol. 11, apr 2021.

[108] A. P. Aitken, C. Ledig, L. Theis, J. Caballero, Z. Wang, and W. Shi, "Checkerboard artifact free sub-pixel convolution: A note on sub-pixel convolution, resize convolution and convolution resize," *ArXiv*, vol. abs/1707.02937, 2017.

[109] E. Plaut, "From principal subspaces to principal components with linear autoencoders," 2018.

[110] J. Zhou, J. Ma, Y. Chen, C. Cheng, B. Bao, J. Peng, T. J. Sejnowski, J. R. Dixon, and J. R. Ecker, "Robust single-cell hi-c clustering by convolution- and random-walk–based imputation," vol. 116, pp. 14011–14018, jun 2019.

[111] G. W. Stewart, "On the early history of the singular value decomposition," *SIAM Rev.*, vol. 35, pp. 551–566, dec 1993.

[112] W. Liu and S. Weiss, *Wideband Beamforming*. John Wiley & Sons, Ltd, mar 2010.