University of Warsaw
Faculty of Mathematics, Informatics and Mechanics

Krzysztof Łukasz
Student no. 467102

Sequencing budget allocation for inferring gene regulatory networks from single-cell data

Master's thesis
in BIOINFORMATICS AND SYSTEMS BIOLOGY

Supervisor:
dr Aleksander Jankowski
Institute of Informatics

Warsaw, December 2025

Abstract

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Keywords

blabaliza różnicowa, fetory σ-ρ, fooizm, blarbarucja, blaba, fetoryka, baleronik

Thesis domain (Socrates-Erasmus subject area codes)

11.3 Informatics, Computer Science

Subject classification

Applied computing
Life and medical sciences
Bioinformatics

Tytuł pracy w języku polskim

Alokacja budżetu sekwencjonowania przy odtwarzaniu sieci regulacji genów na podstawie danych z pojedynczych komórek

Contents

Chapter 1

Introduction

The fact that gene regulation is a highly complex and effective process is beyond dispute. It should be clear to anyone just by looking at their own body, a testament to nature's unparalleled engineering. From the photoreceptors in the retina that enable vision, neurons that give rise to consciousness, to the myocytes in our legs responsible for movement - every tissue and organ showcases an astonishing degree of specialization. Yet, every cell in our body originates from a solitary fertilised egg, a zygote, that carries genetic material inherited from both parents. This single cell orchestrates the development of an entire organism, and its genetic blueprint shapes a great portion of our biological identity and development.

Naturally, this raises the question of how is such cellular diversity achieved, given a shared genome across all cells? The quest to answer this began with Watson and Crick's formulation of the central dogma of molecular biology, which laid the foundation for understanding how genetic information flows from DNA through RNA to protein. Once the genetic code was deciphered, some attention turned to the complex mechanisms that govern the ifs, whens, and wheres of gene expression.

The discovery of the transcriptional machinery along with transcription factors, enhancers, silencers, and other regulatory elements offered a glimpse into the intricate systems that regulate gene activity. It became clear that gene expression is not a simple on-off switch, but rather a nuanced, multilayered process influenced by signals which continue to be identified.

To model this complexity, the concept of **gene regulatory networks** (GRNs) emerged. These networks model the interactions between regulatory elements and target genes, forming circuits that control cellular behaviour. Like all models, they are, in the words of statistician George E.P. Box, 'inherently wrong but can be useful' nonetheless. GRNs provide insight into how, during embryonic development, specific transcription factors can direct a cell to become a hepatocyte or how immune cells respond dynamically to external stimuli, orchestrating the expression of hundreds of genes required for defence. These networks not only help explain how stable cell identities are established and maintained, but also offer a framework for identifying dysregulated pathways in disease, holding potential for identifying novel therapeutic targets.

However, reconstruction of these networks remains a challenge. Until recently, merely outperforming random prediction served as baseline for success when evaluating inferred networks against experimental evidence. The inherent complexity of the system, the high dimensionality of gene expression data, the presence of noise, the dynamic nature of gene regulation, and limited biological knowledge all pose obstacles on the way to accurate prediction. Hope was brought by advances in sequencing technologies beyond traditional gene expression profiling. Techniques such as Hi-C, which captures chromatin conformation, ATAC-seq, which measures chromatin accessibility, and ChIP-seq, which identifies protein-DNA interactions,

provided additional layers of regulatory information that could be used to improve GRN inference.

The abundant influx of new data, along with the rapid development of sequencing techniques and GRN inference methods, served as the primary motivation for this work. The central objective is to determine how to most effectively allocate limited resources when designing sequencing studies for GRN inference. This includes answering practical questions such as: How deeply should we sequence each cell? How many cells should we profile? Should we invest in additional modalities like ATAC-seq alongside RNA-seq? By addressing these trade-offs, this work aims to guide more informed and cost-effective experimental strategies for reconstructing GRNs.

This work is divided into four chapters. In chapter 2 we present fundamental biological concepts and mechanisms of gene regulation and gene regulatory networks that are essential for understanding the context of this work. Chapter 3 provides a detailed description of the data and methods used throughout the study. In chapter 4, we present the results concerning the accuracy of gene regulatory network reconstruction under various conditions. Finally, chapter 5 offers a summary of our findings, including the key questions we aimed to answer and those that remain open for future research.

Chapter 2

Background and motivation

DNA makes RNA, RNA makes proteins, and proteins make us

Francis Crick, 1957

Understanding the principles of human genetics is fundamental to modern biology and medicine. The human genome, composed of over three billion base pairs, encodes the instructions required to build and maintain the body's diverse array of cells, tissues, and organs. Despite the fact that nearly every cell contains the same genetic information, differences in gene activity give rise to the remarkable complexity of human development and physiology. These differences are achieved through complex regulatory mechanisms that determine which genes are expressed, when, and to what extent. The ability to decipher and model these regulatory processes is essential for advancing our understanding of cell identity, development, and disease.

Inside nearly every cell there is a nucleus, the core which contains our genetic information: deoxyribonucleic acid (DNA). DNA is a four-letter code made up of nucleotides (bases): adenine (A), guanine (G), cytosine (C) and thymine (T). Human genome consists of around 3.1 billion base pairs. A **gene**, originally defined as a basic unit of heredity, is a functional segment of DNA that contains the instructions for synthesizing a particular RNA molecule. Each gene typically includes a **promoter** region, which signals the start of transcription, **exons**, which are coding sequences, **introns**, which are non-coding segments removed during RNA processing, and regulatory regions, which influence the gene activity. Our genes are stored in groups of several thousand on 23 pairs of chromosomes in the nucleus. Despite the genome containing ∼20,000 protein-coding genes, only about 15 percent is transcribed, as intergenic regions make up the majority of the genome's sequence and only around 1-2 percent is translated into proteins. When a particular gene is expressed, a temporary copy of the sequence is generated in the form of ribonucleic acid (RNA). This copy contains all the information required to make a protein, a process called translation.[1] The concept describing this flow of information was proposed by Francis Crick in 1957 [2] and is known as the Central Dogma of Molecular Biology.

2.1. Regulation of gene expression

2.1.1. Genome organisation

In a typical diploid cell, the genome consists of two complete sets of chromosomes. If stretched out end to end, the DNA would span nearly 2 meters in length. This remarkable scale presents a fundamental biological challenge: how to store such abundance of DNA within the confines of a nucleus only 6 micrometers in diameter and, what is even more challenging, keep this material organised and accessible, a problem roughly equivalent to fitting 50 kilometres of headphone cord into a pocket without it becoming hopelessly tangled.

To solve this problem, the genome is organised into a structure known as chromatin, consisting of the DNA, **histones** (relatively small and positively charged proteins), and non-histone proteins. The most basic unit of chromatin is the **nucleosome**, formed by wrapping DNA around a histone octamer, a protein cylinder consisting of two copies each of histones H2A, H2B, H3, and H4. This octamer interacts with the negatively charged phosphate backbone of DNA, facilitating tight DNA bending and wrapping. Adjacent nucleosomes are connected by stretches of linker DNA, typically 20–80 base pairs long, resulting in a regular spacing of approximately one nucleosome every 200 base pairs. The addition of linker histone H1 further stabilises this structure and promotes higher-order chromatin folding. Moreover, each of the core histones has a somewhat unstructured N-terminal amino acid tail, which extends out from the core and serves as key site for various modifications. The four core histones undergo several key types of covalent or posttranslational modifications (PTMs), including acetylation, methylation, phosphorylation, ADP-ribosylation, monoubiquitylation, and SUMOylation (addition of small ubiquitin-like modifiers). These chemical alterations are not merely decorative, they play crucial roles in regulating chromatin architecture and function. By influencing how tightly or loosely DNA is wound around histone proteins, these modifications help control gene expression, DNA repair, replication, and other essential genomic processes. Each type of modification can act as a signal that recruits specific proteins or protein complexes, effectively shaping the cellular response to various physiological and environmental signals.[3] Histones' vital role in DNA function explains why they are among the most conserved proteins in eukaryotes. For instance, the histone H4 protein in pea and human is remarkably similar, with only two differences in its 102 amino acids.[1]

The nucleosome is a constantly changing structure where the DNA repeatedly unwraps from the histone octamer, briefly remains exposed (10-50 milliseconds), and then rewraps around the histones. Cellular processes further loosen DNA-histone complexes through the action of various **chromatin-remodelling complexes**. These complexes hydrolyse ATP to move DNA relative to the histone octamer. This temporary change in nucleosome structure, through repeated "sliding" cycles that pull the DNA helix along the nucleosome core, ensures that all DNA sequences within chromatin are potentially accessible for binding by other proteins inside the cell. In addition, some of the remodelling complexes have the ability to either exchange histones or completely removing the histone core form the nucleosome.[1]

So far, by wrapping DNA onto histones the genome and creating the so called "beads on a string" structure an estimated 3 to 10 fold reduction in length was achieved.[1], [3] Because the disk-like nucleosome structure has a 10 nm diameter the structure was named 10-nm fibril. In order to fit into the nucleus, it is further condensed into 30-nm fibre by supercoiling every 6 to 7 nucleosomes forming a zigzag ribbon structure and thus achieving additional 4 to 6 fold reduction. Likely, histone modifications play a crucial role in transitioning from the 10 nm fibril to the 30 nm fibre state and *vice versa* thanks to the N terminal tails facilitating interactions between nucleosomes.

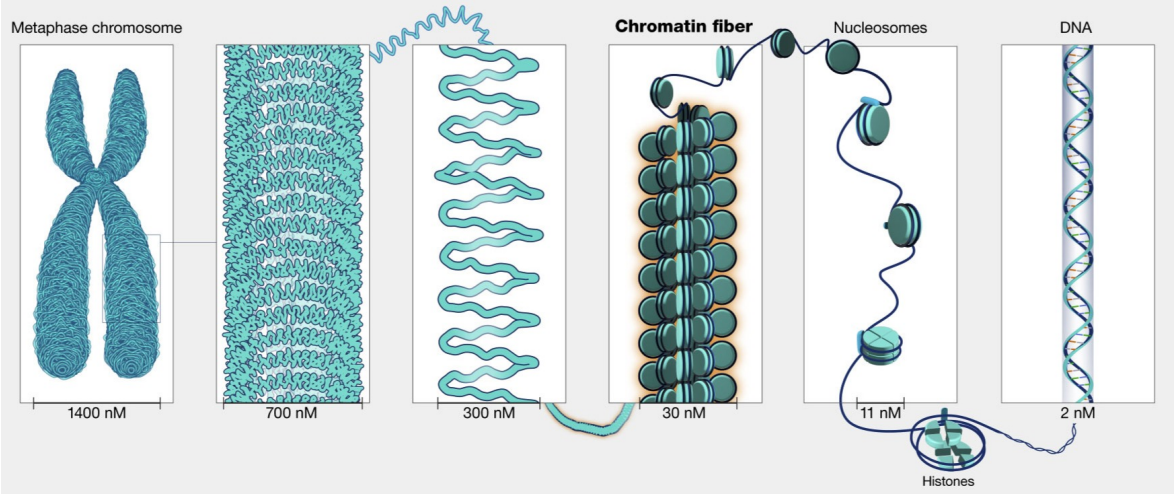

Figure 2.1: Schematic overview of different levels of DNA organisation.[9]

The fibres appear not to exist as a linear structure but rather appear to be organised into loops or domains from 50,000 to 200,000 bp in size each, further packing the chromatin into a shorter but thicker structure.[4] It is noteworthy, that despite being widely recognised and described in well-established textbooks, the 30 nm structure fails to be observed *in vivo* in experimental studies and researchers propose alternative models such as larger globular structures stacking the 10 nm fibres.[5]

A number of these loops have been shown to be attached to the nuclear matrix, a network of proteins and RNA molecules spanning the entire nucleus. Attachment to the nuclear matrix (or scaffold) is possible through special DNA segments, AT-rich regions called **matrix attachment regions** (MARs) or scaffold attachment regions (SARs) which are located at the base of the DNA loops. Each of these loops form a structural domain of the chromatin and some but not all of them appear to correspond to functional domains in which a given loop of DNA transitions from the 30 nm fibre to the 10 nm fibril, before the onset of transcription.[4] Chromatin loops are organised into higher-order structures called **topologically associating domains** (TADs), which are spatial compartments of the genome that interact more frequently within a given TAD than with neighbouring regions, or an enhancer located in one TAD is more likely to interact with local promoters than with those located in neighbouring TADs. TADs are often bounded by the architectural protein CTCF and stabilised by the cohesin complex, creating a three-dimensional genome organization that influences gene regulation.[6] Cohesin complexes extrude chromatin loops until halted by convergently oriented CTCF binding sites, thereby shaping TADs.[7] Research has shown that when CTCF or cohesin is depleted, the structure of TADs in vertebrates is greatly reduced. Interestingly, this disruption has only a small impact on overall gene expression. Although hundreds of genes are affected, fewer than half show increased expression, indicating that losing TAD boundaries can cause random enhancer-promoter interactions. This suggests that TADs only limit the influence of certain enhancers. However, it is still possible that many genes might have less accurate expression in terms of location or timing without TADs, a change that might only be detected using more precise techniques.[8]

On the other hand, experiments in transgenic mice have demonstrated that introducing a gene into a mice genome often results in very low expression. Also, the transcription levels were vastly different depending on the specific site that those genes were placed suggesting important impact of position effect that was not mitigated by incorporating genes together

with their adjacent regulatory elements, a phenomenon known as **position effect variegation (PEV)**. This suggests a major influence of chromatin context that is not fully mitigated by incorporating nearby regulatory elements.[10] A practical implication was observed in humans. Changes in TADs and chromatin loop were established to inducing oncogenesis by activating oncogenes or silencing tumour suppressor genes.[11]

From the functional standpoint, chromatin exists in two major states: **euchromatin** and **heterochromatin**. Euchromatin is transcriptionally active and is typically less condensed. It is associated with nucleosome remodelling, histone covalent modifications, reduced DNA methylation, and the presence of histone variants that facilitate accessibility. Regions of euchromatin are often sensitive to enzymatic digestion by nucleases such as DNase I, a feature used in DNase assays to mark regulatory elements such as enhancers and promoters. These DNase-sensitive sites represent disrupted nucleosome structure, where DNA is transiently unwrapped or replaced by regulatory proteins like transcriptional activators. Heterochromatin, on the other hand, is densely packed and transcriptionally silent. It is enriched in repressive histone modifications and high levels of DNA methylation.[3]

Insulators are genomic loci that separate genes located in one chromatin region from regulation by transcription factors binding to enhancers of chromatin regions in close proximity. They are often bound by the CTCF protein. Together with other proteins, such as cohesin, CTCF regulates the formation of chromatin loops, including regulatory loops. In addition, insulators can act like genomic fences preventing the spread of heterochromatin from silenced genomic regions to transcriptionally active regions, hence their name. Disrupting insulator regions dysregulates transcription and can lead to developmental disorders.[12]

During mitosis, chromatin becomes even more compacted, reaching a compaction level of approximately 8000 fold compared to naked DNA. At this stage, chromosomes adopt a highly ordered looped architecture, where chromatin loops are tightly packed in a structured and by no means random arrangement. This extreme level of compaction renders individual chromosomes microscopically visible, a feature often associated with the classic textbook representation of chromosome structure. This marks the final stage in the hierarchical organization of DNA—from a linear strand to the metaphase chromosome familiar from cytogenetic images.[3], [4]

2.1.2. The process of transcription

The process of DNA-dependent synthesis of RNA, or transcription, is carried out by three enzymes: RNA polymerases I, II and III. Moving around 50 nucleotides per second and working in concert, these enzymes can generate approximately 1,000 transcripts of a single gene within an hour, though elongation rate can vary between 20 and 70 nt/s depending on chromatin context.[1] RNA polymerases I and III primarily transcribe so-called "housekeeping genes" which are involved in essential cellular functions. These genes are generally expressed constitutively so their expression is regulated relatively loosely. In contrast, polymerase II (Pol II) is responsible for the expression of the vast majority of genes and is subject to tight control by intracellular and extracellular signalling.[13]

Transcription initiation is a critical control point in gene expression, as it largely determines which proteins a cell produces and in what quantities. The **transcription start site** (TSS) marks the first nucleotide transcribed into RNA, defining the $5'$ end of a gene, while the $3'$ end corresponds to the site where RNA polymerase disengages from the DNA template. Any sequence in front of the TSS is referred to as “upstream” and “downstream” means after the TSS. In eukaryotes, genes are composed of exons and introns, with the initial transcript (pre-mRNA) undergoing splicing to remove introns and ligate exons into mature mRNA.

The **core promoter** is defined as the genomic region within approximately 50 base pairs surrounding a TSS. This compact DNA stretch is vital as it provides the binding sites necessary for the formation of the pre-initiation complex. While bacterial RNA polymerase requires only a single transcription-initiation factor to begin transcription, eukaryotic RNA polymerases require many such factors, collectively called **general transcription factors** (GTFs).[4]. In humans, more than 50 different proteins bind to the core promoter and form a massive 4 MDa assembly of basal transcriptional machinery. By facilitating this complex's formation, the core promoter precisely positions the catalytic site of Pol II, thereby determining the exact location where transcription will begin. Notably, Pol II cannot directly recognise specific DNA sequences. Therefore, the exact location of the TSS is not determined by Pol II itself, but rather by the precise physical positioning of Pol II in relation to the protein complex. [13]

For many years, it was believed that a single motif upstream of TSS served as a universal promoter, known as the **TATA box**. Now we know that promoter recognition in eukaryotes is much more varied and complex, yet TATA box remains the most studied element of the core promoter and stands as a classical example. It serves as the binding site for the **TATA-binding protein** (TBP). TBP recognises the TATA box only when the promoter region is accessible, i.e., relatively free of nucleosomes. Upon binding, TBP recruits a set of associated proteins known as **TBP-associated factors** (TAFs), which together with TBP form the multi-subunit TFIID complex. This complex not only anchors the transcription machinery but also facilitates local chromatin remodeling via the histone acetyltransferase activity of TAF1. Once TFIID is bound to DNA, it serves as a landmark for the ordered assembly of other general transcription factors, including TFIIA, TFIIB, TFIIE, TFIIF, and TFIIH, as well as RNA polymerase II, completing the formation of the pre-initiation complex.[14]

However, only around 10 to 20 percent of genes have a TATA or TATA-like box. In such instances, additional regulatory elements, such as the initiator (Inr) and the downstream promoter element (DPE), directing the assembly of the transcriptional complex, facilitating initiation of basal transcription.[15] Notably, approximately 60-70 percent of human promoters are associated with CpG islands.[16] Core promoter sequence elements are often conserved across orthologous genes, however, the diversity among mammalian promoters make reliable prediction of TSS regions without experimental data a separate problem.

Another essential component of the transcriptional machinery is the **Mediator complex**, a large multiprotein assembly that functions as a bridge between DNA-bound regulatory factors and the polymerase II apparatus. The complex comprise more than 30 subunits and it is not binding DNA itself but rather is recruited by transcriptional activators bound to enhancers or promoter-proximal elements. Once engaged, it facilitates the assembly and stabilization of the pre-initiation complex by physically interacting with both RNA polymerase II and general transcription factors. The head and middle modules connect directly to RNA polymerase II, while the tail module interacts with upstream activator proteins, enabling Mediator to transmit regulatory signals to the core promoter. The CDK8 module can reversibly associate with the Mediator complex and inhibit its interaction with Pol II, thereby exerting context-dependent repressive or tuning functions.[4], [13]

In addition to its role in transcription initiation, Mediator is also involved in facilitating transcription elongation. It contributes to the recruitment of elongation factors such as P-TEFb, which phosphorylates the C-terminal domain (CTD) of Pol II, allowing the polymerase to transition into productive elongation. Because of its strategic position and modular structure, the Mediator complex functions as a hub for signal integrating, coordinating diverse upstream inputs and determining transcriptional outcomes in a context-specific manner.[4], [13]

Studies have also revealed additional rate-limiting steps in transcription beside recruting Pol II and GTFs, including the release of promoter-proximal paused Pol II. Promoter-proximal pausing by RNA polymerase II is a well-established mechanism to control the timing, rate, and possibly the magnitude of transcriptional responses. After Pol II is recruited and begins transcribing DNA into RNA, it frequently stops synthesising the new RNA strand when it's only about 20 to 60 nucleotides long. Pol II then stays in this paused state until it receives further signals that encourage it to continue the elongation process. This process is particularly important in stress and inflammation response.[17]

Together, the coordination of general transcription factors, chromatin state, and regulatory cofactors like Mediator ensure that transcription is responsive to both internal signals and external stimuli, thereby making it a central control point for gene expression.

2.1.3. Mechanisms of regulation of gene expression

The cell can regulate gene expression at every step along the path from DNA to functional protein. This regulation includes: precise control over when and how much a gene is transcribed (transcriptional control), RNA processing and splicing, exporting complete RNA from the nucleus to the cytosol, selective mechanisms for RNA translation, degradation of RNA in the cytoplasm, and ultimately, the controlled degradation or inactivation of the proteins themselves. Among these, transcriptional control is often considered the predominant step, as it is the only point at which the cell can prevent the production of unnecessary or redundant intermediates. Schematic overview of the regulatory code is shown in fig. 2.2.

Sequence-specific transcription factors

The transcriptional activity of a gene is largely determined by the binding of specific proteins to DNA sequences that are often called ***cis*-regulatory sequences**, because they must be on the same chromosome (or *in cis*) to the genes they control. These proteins, known as **transcription factors** or **regulators**, are central players in the control of gene expression. Depending on the effect they have, we distinguish between positive transcription regulators (**activators**) that facilitate transcription, and negative regulators (**repressors**) that stop this process.

In contrast to the few general transcription factors that assemble around the promoter of every gene, there are approximately 1600 human genes encoding different transcription factors, amounting to around 8% of all protein-coding genes. Each TF uses a highly conserved DNA-binding domain (DBD) to recognise specific closely related motifs, typically 6 to 12 nucleotide pairs in length, often described by position weight matrices or logos. However, in vivo binding depends only partially on the presence of the motif. Most TFs bind only a small fraction of their motif matches across the genome. Factors that modulate actual occupancy include chromatin accessibility, nucleosome positioning, local DNA shape, flanking sequences, and cooperative interactions with other TFs.[18]

However, a typical 6-8 bp motif may occur thousands of times by chance across the genome, making it difficult to distinguish functional binding sites from background sequences. To overcome this limitation, many transcription factors bind DNA as **dimers**, a strategy that significantly enhances both the affinity and specificity of recognition. Dimerisation effectively doubles the length of the recognised sequence, allowing proteins to discriminate more selectively between potential binding sites. Homodimers often bind palindromic or tandem motifs, while heterodimers, composed of two different TFs, can recognise asymmetric or composite motifs. This combinatorial flexibility enables the same transcription factor to participate in

distinct regulatory programs depending on its dimerisation partner, greatly expanding regulatory diversity without requiring a proportional increase in the number of TF genes.[18]

In a typical scenario, the dimerisation units form multiple, strong noncovalent bonds and dimers or heterodimers form stably in solution and bind as preassembled units. However, when subunit interactions are weak, the proteins exist mainly as monomers in solution and assemble into dimers only upon encountering the correct DNA motif, a phenomenon known as **cooperative binding**, where the binding of one subunit enhances the affinity of the other for DNA. In this case, binding becomes cooperative, and the fraction of DNA bound as a function of protein concentration follows a sigmoidal curve rather than a hyperbolic one. Therefore, cooperative binding introduces a threshold-like response, such that *cis*-regulatory elements are typically either largely unbound or fully occupied, with few intermediate states.[1]

Transcription regulators typically bind to DNA within nucleosomes with less affinity than to naked DNA, primarily for two reasons. First, the specific regulatory DNA sequence might be oriented inward, toward the histone core, making it inaccessible to the regulator. Second, even when the regulatory sequence faces outward, many transcription regulators induce subtle conformational changes in the DNA, such as bending or kinking, that are resisted by the DNA tightly wrapped around the histone core. Nucleosome remodelling can change the nucleosome structure to improve access for transcription regulators, but even without remodelling, the DNA at nucleosome edges *breathes*, temporarily exposing sequences and allowing regulators to bind. This breathing is much more frequent near the nucleosome edges, making them the easiest site to access by TFs.[19] This dynamic promotes cooperative binding among transcription regulators. When one regulator binds during a breathing event, it loosens the DNA-histone interaction, making it easier for a second regulator to bind nearby. When regulators have sufficient affinity and concentration, they can exploit nucleosome breathing to invade nucleosomes destabilizing them.

There is a class of transcription factors that are especially efficient in binding to histone-bound DNA, termed **pioneer factors**. They can directly bind nucleosomal DNA and are often the first to engage with DNA at previously inactive gene sites. Although they usually destabilise nucleosomes upon binding, their main role is likely to recruit other proteins, such as chromatin-remodelling complexes. When such a complex is attracted, further changes in the conformation of the DNA are made allowing other TFs to efficiently bind nearby.[20] Their ability to change the structure of the chromatin is showcased in factor combinations such as OSKM (Oct4, Sox2, Klf4, and c-Myc) that can reprogram somatic cells to induce pluripotency.[21]

The interaction between transcription regulators and DNA is not a static event but rather a highly dynamic equilibrium characterised by rapid binding and dissociation. Single-molecule tracking experiments in living cells reveal that these regulators exist in distinct states: a high-mobility, unbound state while diffusing through the nucleoplasm, and a transient, low-mobility state when bound to DNA. The duration of this bound state is directly proportional to the regulator's binding affinity for a specific DNA sequence. High-affinity interactions with *cis*-regulatory sites result in longer residence times, while weak, non-specific interactions are short-lived.[22] Any protein that binds to tightly to a given DNA sequence will also bind, much more weakly, to any sequence. This weak binding facilitates a scanning mechanism that allows regulators to search the genome for their targets. Consequently, the sustained regulation of a target gene is not achieved by a single, permanently bound molecule, but is instead the statistical outcome of a continuous cycle of many regulator molecules associating with and dissociating from their specific DNA sequences over time.[1]

Cis-regulatory elements

In some genomic regions, a number of *cis*-regulatory sequences occur in close proximity, forming a cluster at which many regulators converge. Those clusters are known as ***cis*-regulatory elements** (CREs). In general, CREs recognised by the regulators can be divided on the basis of the role they play in the transcription. However, the same CRE can be bound by regulators that either increase or decrease the transcription rate and therefore serve different roles depending on the context.[23]

Enhancers are clusters of *cis*-regulatory motifs that attract multiple transcription factors and upregulate the formation of the transcription complex and facilitate binding to the core promoter. They can act over great distances, typically up to 100,000 bp away from the TSS in human genome, with some acting over distances in the megabase range.[24] Their activity likely relies on chromatin looping, which brings enhancers into spatial proximity with target promoters. Recent evidence suggests that enhancer–promoter communication is not always a simple matter of physical contact: loops may pre-exist prior to activation or dynamically form in response to developmental or environmental cues.[25]

In contrast, **silencers** are regulatory elements that inhibit linked promoter activity and reduce transcription - they are the repressive counterparts of enhancers. Although discovered decades ago, and despite evidence of their importance in development and disease, silencers have been much less studied than enhancers. Silencers are often bifunctional regulatory elements that can also act as enhancers. They downregulate gene expression either by recruiting repressor proteins that interfere with preinitiation complex formation or by passively preventing transcription factor binding.[26] Mutations in the *cis*-regulatory regions have been shown to cause phenotypic changes in human, leading to so-called enhanceropathies. Examples include thalassaemia, a blood disease in which imbalanced quantities of α- and β-globin are produced, caused by misregulation of the human β-globin gene transcription.[27]

Given that both enhancers and silencers can act over long genomic distances, their activity must be restricted to appropriate target genes, a role fulfilled by **insulators**. In addition to their role in blocking heterochromatin spread, they prevent communication between an enhancer and a promoter when located between them.[28] Beyond linear models of regulation, transcriptional control now includes higher-order genome organization, where genes, enhancers, and regulatory complexes integrate within dynamic nuclear condensates or "transcription factories," allowing spatially coordinated gene activation.[29] Together, these findings underscore that *cis*-regulation is governed not only by DNA sequence motifs but also by chromatin context, 3D genome architecture, and dynamic nuclear organization.

Cooperative interactions between transcription factors and cofactors

Regulation of eukaryotic gene expression rarely depends on the action of a single transcription factor acting in isolation. Instead, transcriptional control is achieved through the combined and coordinated action of multiple sequence binding proteins, which together determine when, where, and to what extent a gene is transcribed. This principle, referred to as **combinatorial control**, underlies the complexity and specificity of gene regulation in higher eukaryotes.[1] Because the numerous *cis*-regulatory sequences governing a typical gene's expression are often distributed across large genomic regions, the term gene control region is used to refer to the entire stretch of DNA responsible for regulating and initiating that gene's transcription. This region encompasses both the promoter, where general transcription factors and RNA polymerase assemble, and all the additional CREs that serve as binding sites for transcription factors, which modulate the efficiency and rate of transcription initiation at the promoter.

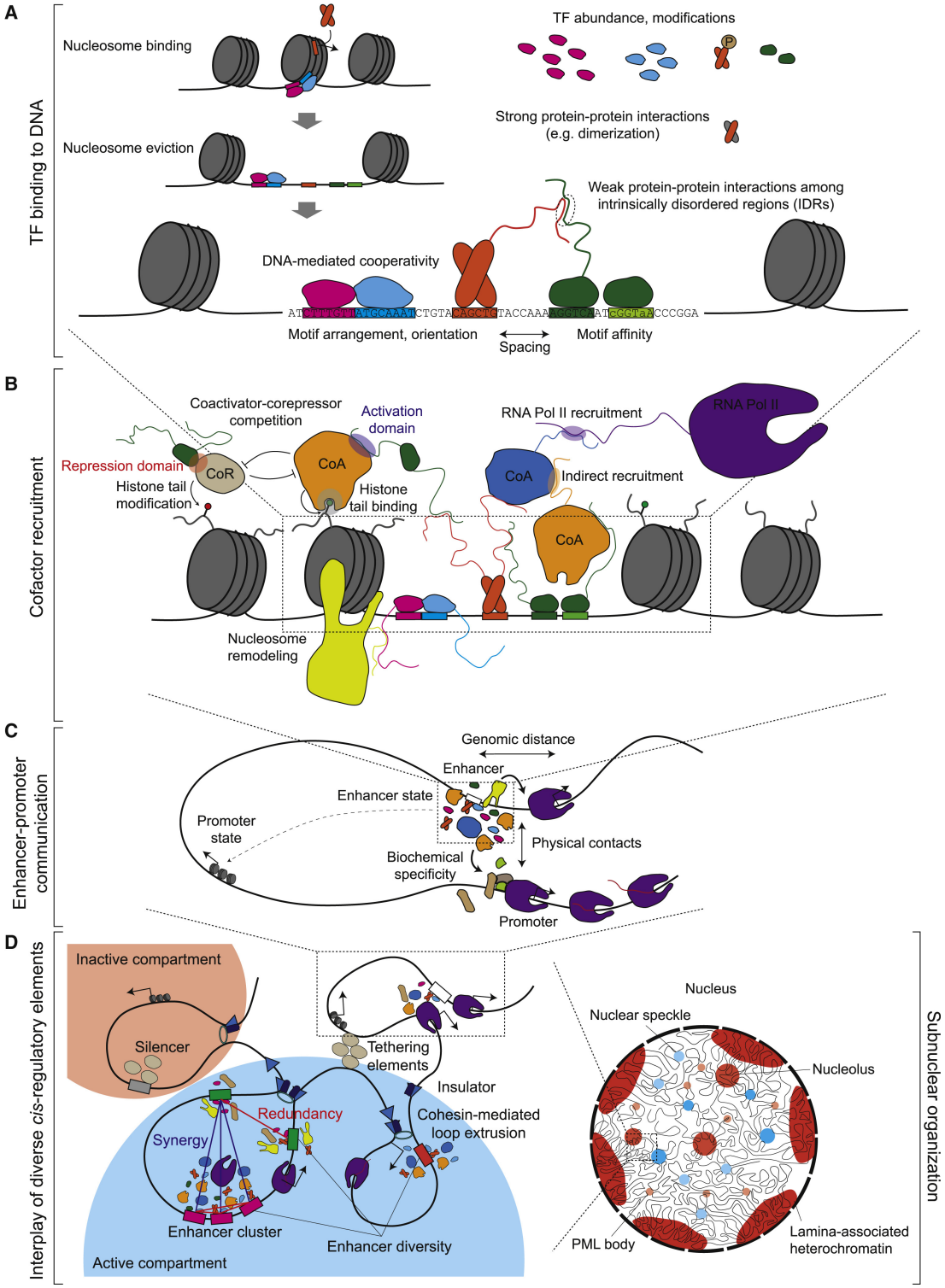

Figure 2.2: **Hierarchy of the cis-regulatory code.** (A) Transcription factor (TF) binding is determined by DNA sequence specificity, nucleosome positioning, and cooperative interactions between TFs. (B) TFs recruit cofactors (activators or repressors) to regulatory elements, often facilitated by protein-protein interactions. (C) The specificity of interactions between enhancers and promoters is regulated by biochemical compatibility and genomic distance. (D) The 3D organization of the nucleus and local chromosomal neighbourhoods facilitates interactions between various regulatory elements. [18]. Licensed under CC BY-NC-ND 4.0.

Transcription factors often function cooperatively to overcome chromatin barriers and establish active regulatory regions, rather than acting independently or sequentially.

Beyond the pioneer factor model, many TFs can jointly bind within the span of a single nucleosome (∼150 bp), stabilizing each other's association with DNA. Such cooperation can arise through different mechanisms, including protein–protein interactions, DNA-mediated cooperativity, and nucleosome-mediated cooperativity, each contributing to enhancer activation in context-dependent ways. [18]

Transcription factors rarely influence transcription directly. Instead, they function primarily as molecular adapters that recruit large multi-subunit protein complexes, collectively known as **cofactors**, to specific *cis*-regulatory elements. These cofactors, comprising coactivators (CoAs) and corepressors (CoRs), mediate the regulatory outcomes of TF binding by modifying chromatin structure, influencing Pol II dynamics, and shaping the local transcriptional environment.[18] Cofactors can be broadly divided into three major functional classes: (i) the **Mediator complex**, which physically links DNA-bound regulators to the general transcription machinery and facilitates Pol II activation, (ii) **chromatin remodellers**, which reposition or remove nucleosomes to maintain accessibility at promoters and enhancers, and (iii) **histone modifiers**, which add or remove specific post-translational modifications on histone tails to establish transcriptionally permissive or repressive chromatin states.[1] Individual TFs often recruit multiple cofactors (in some cases both CoAs and CoRs) through distinct or overlapping protein interaction domains. The same TF can thus produce different transcriptional outcomes depending on which cofactors are locally available and which additional TFs are co-bound at a given enhancer.[30] The specificity of TF–cofactor interactions contributes to the context-dependence of gene regulation: distinct TFs preferentially associate with different cofactor families, and certain cofactors are limiting in abundance, creating competition and hierarchy among enhancers. In genomic contexts, these partnerships are further modulated by chromatin environment and three-dimensional genome organization, allowing cofactors not only to tune transcriptional output locally but also to mediate enhancer–promoter communication over long distances. Together, these findings emphasise that transcriptional regulation emerges from a continuum of interactions among TFs, cofactors, and chromatin rather than a simple one-to-one correspondence between TF binding and gene activation. The combinatorial and context-specific recruitment of cofactors thus represents a crucial mechanistic link between the molecular action of TFs at individual regulatory elements and the coordinated behaviour of gene regulatory networks.[18]

Together, these findings emphasise that transcriptional regulation emerges from a continuum of interactions among TFs, cofactors, and chromatin rather than a simple one-to-one correspondence between TF binding and gene activation. This poses a challenge of how should we accurately discover, describe and model those complex interaction governing the cell fate?

2.2. Gene regulatory networks

Finally, gene expression is dynamic and stochastic: transcription occurs in bursts whose frequency and amplitude depend on promoter architecture, chromatin state, and TF kinetics.[31] These molecular interactions collectively define the regulatory logic encoded in the genome—the foundation of **gene regulatory networks** (GRNs) that coordinate cellular identity and responses.

2.2.1. Definitions

There is no universal definition of a gene regulatory network. In general, GRN models encompass all of the molecular species and regulatory interactions necessary to fully describe observed patterns of gene expression. [32]

In mathematical terms, a network is often described as a graph, which is defined as a pair (V, E), where V is the set of nodes (or vertices) and E is the set of edges (or links) connecting them. Gene regulatory networks are interpretable computational representations of the regulatory relationships governing gene expression, structured in the form of such graphs. [33]

In GRNs, nodes can represent various components involved in gene regulation, such as transcription factors, splicing factors, long non-coding RNAs, microRNAs, metabolites, and genes themselves, while edges represent regulatory interactions between them. Edges in GRNs can encode rich information about regulatory relationships. They may be weighted, to represent the strength or confidence of the interaction; directed, to indicate the flow of regulation from a regulator to its target. Casuality is not easily determined, although can be inferred by appropriate data or experimental design (e.g. perturbation studies, time-series analysis, or causal inference frameworks such as Granger or Pearl causality). Edges can also be signed, distinguishing between activating (positive) and repressing (negative) effects. Such edge attributes allow GRNs to capture not only who is connected to whom, but also how strongly, in which direction, and with what regulatory outcome the interaction occurs. Ideally, an edge corresponds to a direct regulatory relationship rather than an indirect association. [34]

The simplest representation of a GRN is a bipartite graph, where transcription factors form one set of nodes, and the genes they regulate form the other, and egdes are pairs TF-TG (target gene). These two sets are disjoint by definition. Bipartite graphs are the most extensively studied in the literature and remain one of the most widely adopted frameworks due to their simplicity and interpretability. However, the classification of an entity as a TF is not always clear-cut and can depend on the specific context. A common approach is to supply a curated list of genes with experimentally validated regulatory activity. [34], [35]

2.2.2. Approaches and data used to reconstruct GRNs

Understanding how GRNs function has been a central challenge in biology since the early discoveries of gene regulation, such as the seminal work on characterization of the lac operon in the 1960s.[36] With the rise of systems biology, reconstructing large-scale GRNs has become a major objective, supported by advances in high-throughput experimental techniques and computational approaches. Traditionally, GRNs were either compiled from experimentally validated regulatory interactions stored in curated databases or inferred *de novo* from gene co-expression patterns in bulk transcriptomic data.[35]

Bulk transcriptomic data

Methods for inferring GRNs from transcriptomic data, historically from microarrays and, more recently, from bulk or single-cell RNA sequencing, are among the earliest, most widely used, and most extensively developed approaches in the field of network biology.[37] Owing to the widespread availability of gene expression data, a large number of GRN inference tools rely solely on transcriptomic measurements. These methods typically assume that genes with similar expression profiles are likely to participate in shared regulatory programs, a principle often referred to as *guilt by association*. Consequently, most approaches are grounded in pairwise measures of statistical dependence, such as correlation or mutual information, to identify

potential regulatory interactions. When sufficient transcriptomic data are available, the inferred GRNs can be tailored to a specific biological context, providing more relevant insights than the general-purpose networks obtained from databases. While effective for capturing co-expression patterns at scale, these methods are intrinsically limited: they do not distinguish between direct and indirect interactions, struggle to infer causality or directionality, and fail to capture many regulatory layers, such as transcription factor protein abundance and binding to DNA, cooperation between TFs and cofactors, alternative splicing, post-translational modifications, and the chromatin structure and accessibility of the genome, which are not reflected at the mRNA level. Incorporating these additional regulatory modalities has the potential to produce GRNs that more accurately reflect gene regulation in vivo. The advent of single-cell and multi-omic technologies has provided new types of data to capture these regulatory layers. For instance, integrating chromatin accessibility data allows the refinement of TF-gene interactions by accounting for whether a gene locus is open and by including *cis*-regulatory elements in the inference process. [38]

Chromatin accessibility and TF binding

Beyond transcriptomic data, other regulatory modalities can provide additional layers of information to improve GRN inference. Transcription factor binding can be directly measured across the genome using assays such as chromatin immunoprecipitation followed by sequencing (ChIP-seq) or cleavage under targets and tagmentation (CUT&Tag). These data enable the construction of GRNs by linking TF binding sites to putative target genes. However, profiling TF binding remains costly, is limited to TFs for which high-quality antibodies are available, and often relies on assigning binding events to the nearest gene, which can overlook distal regulatory interactions. As an alternative, chromatin accessibility data, most commonly measured via ATAC-seq, can be used to infer putative regulatory elements that are potentially targeted by TFs. Methods leveraging chromatin accessibility generally follow a two-step approach: first, TFs are assigned to accessible regulatory elements based on motif matching; second, these elements are linked to genes, typically based on genomic proximity or known enhancer-promoter interactions. By combining transcriptomic measurements with TF binding or chromatin accessibility data, these approaches can refine TF-TG assignments and capture regulatory relationships that are not evident from gene expression alone, producing GRNs that potentially reflect the underlying biology in vivo more accurately. [35]

Despite their utility, GRN inference methods based on bulk transcriptomic or multi-omic data share inherent limitations. Analyses using bulk measurements average signals across many cells, making it difficult to capture cell type-specific regulatory information and masking underlying cellular heterogeneity [38].

Single-cell resolution

Many of the limitations associated with GRN inference from bulk omics data have been addressed by the advent of single-cell omics technologies. While these approaches share core principles with their bulk predecessors, single-cell techniques provide measurements at the level of individual cells, and multi-omic single-cell methods can capture multiple molecular modalities from the same cell. These approaches offer a high-resolution view of the cellular and molecular landscape within tissues, revealing heterogeneity that bulk sequencing methods cannot capture. GRN reconstruction methods applied to single-cell data enable the inference of cell type-specific TF–gene interactions, as well as the dynamic regulatory changes that occur across developmental processes or different biological conditions. [38]

2.2.3. Characterisation of methods for GRN inference

A variety of computational methods have been developed to infer gene regulatory networks from omics data, each with different assumptions, input requirements, and modelling strategies. Given the inherent noise, dimensionality, and correlational nature of biological data, no single inference algorithm is universally superior. A landmark study even demonstrated that consensus predictions from an ensemble of different methods, "wisdom of crowds", yield more robust and accurate network models than any individual algorithm. [37] Broadly, inference methods can be grouped into several categories based on the type of statistical or computational framework they employ. Here, we follow the classification proposed by [38], which organises methods based on their general inference strategy.

Correlation-based methods are among the earliest approaches and rely on the principle of guilt by association: genes with correlated expression profiles are likely to participate in shared regulatory programs. For example, we assume that co-expressed genes are functionally related or co-regulated or that correlation between CRE accessibility and putative target gene's expression suggest a regulatory relationship. Common statistical measures for these associations include Pearson's correlation, which captures linear relationships, and Spearman's correlation, a non-parametric measure capable of detecting monotonic or non-linear associations. Linear correlation excels in identifying situations where changes in TF expression or CRE accessibility lead to proportional changes in gene expression, whereas Spearman correlation can capture more complex regulatory patterns among TFs, CREs, and genes. Mutual information, an information-theoretic, non-parametric metric, is another common approach for quantifying dependence between variables. While effective for detecting co-expression patterns, these methods cannot distinguish between direct and indirect interactions and do not provide causal directionality.[38]

Regression-based methods model the expression of a target gene as a function of multiple potential regulators, such as TF expression or CRE accessibility. By estimating the effect of each regulator separately, the resulting coefficients can be interpreted in terms of both strength and direction of association. Non-parametric variants, such as tree-based regression, do not rely on assumptions about the underlying data distribution and can capture complex, non-linear relationships. However, regression-based methods have notable limitations. The large number of predictors can easily lead to model overfitting and poor generalization. Additionally, predictors are often correlated with one another, reflecting co-regulation or cooperative activity, which can complicate the inference of independent regulatory effects and reduce interpretability. Furthermore, the accuracy of regression-based inference strongly depends on the sample size relative to the number of genes and regulators.[38]

Probabilistic methods to GRN inference explicitly model the data using statistical distributions, in contrast to correlation- or regression-based methods that rely primarily on measures of dependence. One common class is Gaussian Graphical Models (GGMs), which assume that gene expression follows a multivariate normal distribution. In this framework, the precision matrix (inverse of the covariance matrix) encodes partial correlations between genes, representing direct associations while controlling for indirect effects. Regularization techniques, such as the graphical lasso, are typically employed to estimate sparse networks in high-dimensional settings, enhancing interpretability. While GGMs have proven effective, they rely on assumptions of normality and linear relationships, and estimating large precision matrices from limited data can be challenging. **Bayesian networks** offer a complementary approach by modelling the joint probability distribution of gene expression as a product of local conditional probabilities. Each gene is associated with a set of parent regulators, creating a directed acyclic graph (DAG) that captures directional dependencies. This framework allows

the incorporation of prior knowledge and provides a probabilistic characterization of network uncertainty, often yielding an ensemble of plausible network structures weighted by their posterior probability. However, learning Bayesian networks is computationally demanding due to the combinatorial complexity of the network space, the requirement to enforce the DAG constraint, and the fact that the same distribution can be represented as different networks. Despite these challenges, Bayesian approaches prove effective, particularly when prior information is available or uncertainty quantification is important.[39]

Dynamical models put forward a key aspect of gene regulation, its temporal nature: cells dynamically respond to internal and external signals, and gene expression changes over time. Dynamical GRN inference methods explicitly leverage time-series data to capture these processes. One widely used class is **Dynamic Bayesian Networks (DBNs)**, which extend standard Bayesian networks to handle temporal data. By unrolling the network (only allowing edges to connect vertices at different time points) across time points, DBNs ensure the directed acyclic graph (DAG) constraint is satisfied while allowing feedback loops to be represented across successive time points. DBNs facilitate inference of causal, time-directed regulatory interactions, but remain computationally demanding, and assumptions of linearity and equal time-scales can limit biological realism. **Differential equation models** provide an alternative, mechanistically motivated framework, describing the rate of change of gene expression as a function of the current state of the system. Linear DE models, such as $dx/dt = Ax$, are commonly used, where the interaction matrix A encodes regulatory effects between genes. DE methods approximate derivatives from time-series data and recover network structure via regularised regression, while other approaches employ Gaussian processes or basis function expansions to infer parameters in a Bayesian framework.[39]

Deep learning methods have recently emerged as a powerful class of approaches for GRN inference, leveraging their capability to learn complex, high-dimensional patterns from large-scale omics data. Unlike traditional correlation, regression, or probabilistic methods, deep learning techniques treat GRN inference typically as classification or regression tasks that model intricate relationships between genes with minimal prior assumptions. These methods employ architectures such as convolutional neural networks (CNNs), graph neural networks (GNNs), and autoencoders among others, often incorporating additional biological information or temporal dynamics to improve accuracy. Deep learning approaches are particularly effective in handling noisy single-cell RNA sequencing data and integrating heterogeneous data types including imaging and spatial transcriptomics. Despite significant promise, challenges remain related to model interpretability, overfitting, and computational demand. A variety of deep learning models have been developed that advance beyond simple pairwise gene interactions, enabling the inference of complex regulatory patterns at scale and thereby enhancing the reconstruction of regulatory networks from both static and temporal omics datasets.[40]

2.2.4. Details of selected methods

GENIE3 GENIE3 ("GEne Network Inference with Ensemble of trees") is a GRN inference method introduced in 2010 [41]. Four implementations of GENIE3 are currently available: in `Python`, `MATLAB`, `R` using the `randomForest` package, and in `R/C`. GENIE3 achieved state-of-the-art performance and was ranked first in the DREAM4 In Silico Multifactorial Network Inference Challenge [42]. The goal of GENIE3 is to infer a directed gene regulatory network, represented as a directed graph with p nodes, where each node corresponds to a gene. A directed edge from gene i to gene j indicates that gene i regulates the expression of gene j. GENIE3 infers only unsigned edges; that is, it determines whether a regulatory link exists but

does not classify it as activation or repression. The input to the algorithm is a gene expression matrix measured across multiple experimental conditions or samples.

Let the gene expression data be represented as a matrix:

$$\mathbf{X} \in \mathbb{R}^{p \times n},$$

where p is the number of genes, n is the number of samples (conditions, experiments, or cells), X_{ij} is the expression level of gene i in sample j.

GENIE3 decomposes the GRN inference problem into p separate regression problems, one for each target gene $j \in \{1, \ldots, p\}$. Let the set of predictor genes be all other genes, $\{1, \ldots, p\} \setminus \{j\}$, and denote the corresponding predictor matrix by

$$\mathbf{X}_{\cdot,-j} \in \mathbb{R}^{(p-1) \times n}.$$

The regression problem is formulated as

$$\mathbf{x}_{\cdot j} = f_j\big(\mathbf{X}_{\cdot,-j}\big) + \boldsymbol{\varepsilon}_j,$$

where $\mathbf{x}_{\cdot j} \in \mathbb{R}^n$ is the expression vector of the target gene j, f_j is an unknown predictive function, and $\boldsymbol{\varepsilon}_j$ is a noise vector. The function f_j is assumed to depend only on the direct regulators of gene j, making this a feature selection problem.

Each function f_j is then estimated using an ensemble of regression trees, either Random Forests or Extra-Trees. Formally, for a single problem, the objective is to find a function $\hat{f}_j^{(t)}$ that minimises:

$$\sum_{k=1}^{n} \Big(x_{kj} - \hat{f}_j^{(t)}(\mathbf{X}_{\cdot,-j}^{(k)}) \Big)^2,$$

where $\mathbf{X}_{\cdot,-j}^{(k)} \in \mathbb{R}^{p-1}$ is the vector of predictor gene expression values for sample k. In a tree model, the learning sample is recursively split with respect to one variable in order to minimise varinac ein resulting subsamples. A Random Forest consists of an ensemble of T such regression trees $\{\hat{f}_j^{(t)}\}_{t=1}^{T}$ and each tree is built on a bootstrap sample from the original learning sample and, at each test node, K attributes are selected at random among all candidate attributes before determining the best split. A measure of variance reduction is used of the form

$$I(\mathcal{N}) = \#S \operatorname{Var}(S) - \#S_t \operatorname{Var}(S_t) - \#S_f \operatorname{Var}(S_f),$$

where S denotes the set of samples that reach node $\mathcal{N}$, S_t (S_f) denotes its subset for which the test is true (false) and $\operatorname{Var}(\cdot)$ is the variance of the output variable in a subset. For a given variable, the value is obtained by summing I over nodes. The predictions of the ensemble are averaged to produce the final estimate, which reduces variance and improves robustness compared to a single tree.

The contribution of each predictor gene $i \neq j$ to the prediction of gene j is quantified using the ensemble's variable importance measure, resulting in an importance score w_{ij}. To prevent bias towards genes with high expression variance, all gene expression values are standardised to have unit variance in the training set before fitting the ensembles. After computing the importance scores for all target genes, they are aggregated into a weighted adjacency matrix $\mathbf{W} \in \mathbb{R}^{p \times p}$:

$$\mathbf{W}_{ij} = w_{ij}, \quad \mathbf{W}_{jj} = 0.$$

The output of A higher value of w_{ij} indicates stronger evidence that gene i regulates gene j. The final GRN can then be extracted by selecting the top-ranked interactions or applying a threshold, resulting in a sparse, directed network.

GRNBoost2 GRNBoost2 is a GRN inference algorithm introduced in 2019 [43]. GRNBoost2 is implemented within the `Arboreto` framework, an open-source Python package that offers user-friendly interfaces and supports scalable network inference workflows. It extends the GENIE3 framework by employing gradient boosting machines instead of random forests, improving scalability and performance for large datasets, including single-cell RNA-seq. The goal of GRNBoost2 is to infer a directed network between p genes, where nodes represent genes and a directed edge from gene i to gene j indicates a putative regulatory relationship. Also, it infers only unsigned edges.

Similar to GENIE3, GRNBoost2 decomposes GRN inference into a series of regression problems, training one GBM regression model per target gene to predict its expression based on candidate regulators. Unlike GENIE3's random forests, GRNBoost2 employs *gradient boosting* [44], an additive model that combines weak learners (shallow decision trees) sequentially to improve predictive accuracy. Unlike random forests, gradient boosting builds trees sequentially, where each tree is fit to the residuals of the previous ensemble to iteratively reduce prediction error:

$$\hat{f}_j^{(t)} = \hat{f}_j^{(t-1)} + \eta \cdot \text{Tree}_t,$$

where η is the learning rate and Tree_t is the t-th regression tree. Each tree partitions the predictor space recursively, effectively minimizing the residual sum of squares.

A key feature of GRNBoost2 is its *stochastic gradient boosting* implementation with an early-stopping regularization heuristic. Specifically, each new tree in the ensemble is trained on a random subset (e.g., 90%) of the samples, while the remaining out-of-bag samples (10%) are used to estimate loss function improvement. If the average improvement of recent iterations falls below zero, training halts, preventing overfitting and reducing computational cost.

After training the model for target j, the algorithm calculates the *feature importance* of each predictor gene $i \in \{1, \ldots, p\} \setminus \{j\}$. This importance score, denoted w_{ij}, is typically the reduction in loss function brought by all the splits on gene i across all trees in the boosted ensemble. The final output is the complete, weighted adjecency matrix $\mathbf{W}$. All $p(p-1)$ potential regulatory links are ranked by their weights to produce the final network prediction.

PIDC PIDC (Partial Information Decomposition and Context) is a GRN inference algorithm presented in 2017, specifically designed to leverage single-cell transcriptomic data. [45] It is rooted in *partial information decomposition* (PID), a multivariate information theory framework that decomposes the mutual information shared among triplets of genes into unique, redundant, and synergistic components. By capturing these higher-order interactions, PIDC provides a more nuanced understanding of gene-gene dependencies compared to pairwise mutual information methods.

For explaining PIDC we need to introduce some information theory measures. The entropy, $H(X)$, quantifies the uncertainty in the probability distribution, $p(x)$, of a random variable X. For a discrete random variable,

$$H(X) = -\sum_{x \in \mathcal{X}} p(x) \log p(x)$$

which is maximal for a uniform distribution. For example, a gene that is expressed differently across cells will have a higher entropy.

In case of two variables, the information that one variable gives abotut the other is quantified with the Mutual Information (MI):

$$I(X;Y) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \log \left(\frac{p(x,y)}{p(x)p(y)} \right) = H(X) + H(Y) - H(X,Y),$$

which is the difference between the joint entropy of X, Y and joint entropy given their independence. It is non-negative and symmetric. For a pair of genes with dependent pattern of expression, their observed joint entropy will be lower, and so they will have greater mutual information. Given a third variable, Z, the conditional mutual information (CMI),

$$I(X;Y|Z) = H(X,Z) + H(Y,Z) - H(X,Y,Z) - H(Z)$$

quantifies the information between X and Y given knowledge of Z. A number of measures were derived to quantify information between multiple variables. One of the most popular is intersection information (II):

$$II(X;Y;Z) = I(X;Y|Z) - I(X;Y) = I(X;Z|Y) - I(X;Z) = I(Y;Z|X) - I(Y;Z),$$

which quantifies the information gain with a known third variable compared to when it is not known.

Analysing only pairs of genes is often insufficient to capture complex regulatory dependencies. Therefore, PIDC employs the *Partial Information Decomposition* (PID) framework, which extends mutual information to triplets of variables. Given two source variables X and Y and a target variable Z, PID decomposes the total information that X and Y provide about Z into four components: Redundant information, which is shared by both X and Y, Unique information, which is provided only by one of the sources (either X or Y), Synergistic information, which arises only from the joint observation of X and Y and cannot be obtained from either alone. Together, these components sum to the total mutual information:

$$I(X,Y;Z) = \text{Syngergy}(Z;X,Y) + \text{Unique}_Y(Z;Y) + \text{Unique}_X(Z;X) + \text{Redundancy}(Z;X,Y).$$

To compute the PID terms, the redundant information is first estimated using the *specific information*, I_{spec}, which quantifies the information that one variable provides about a specific state of another variable. In the context of gene expression, the "state" of a gene in a given cell refers to the discrete bin in which its expression level falls after discretization, done by a Bayesian block method.

The specific information that variable X provides about a particular state z of variable Z is defined as:

$$I_{\text{spec}}(z;X) = \sum_{x \in X} p(x \mid z) \left(\log\left(\frac{1}{p(z)}\right) - \log\left(\frac{1}{p(z \mid x)}\right) \right).$$

For a set of source variables $S = \{X, Y\}$ and a target variable Z, the redundant information is calculated by aggregating the minimum information provided by any variable in S about each state z of Z:

$$\text{Redundancy}(Z;X,Y) = \sum_{z \in Z} p(z) \min_{S \in \{X,Y\}} I_{\text{spec}}(z;S).$$

The unique information terms can then be derived from the redundant information and the pairwise mutual information (MI). In particular:

$$I(X;Z) = \text{Unique}_Y(Z;X) + \text{Redundancy}(Z;X,Y),$$

which means that we partitioned the MI into unique and redundant terms given a third variable. Finally, the synergistic information is obtained using the interaction information as:

$$II(X;Y;Z) = \text{Synergy}(Z;X,Y) - \text{Redundancy}(Z;X,Y).$$

Key idea used by PIDC is that for gene pairs that are truly connected (regulatory relationship), the unique information between them is consistently higher relative to redundant components across many triplets. PIDC focuses on the *unique information* shared between pairs of genes, which better reflects direct regulatory relationships than raw mutual information that can be inflated by indirect or redundant associations. For each pair of genes (X, Y), PIDC computes the *Proportional Unique Contribution* (PUC) defined as the average ratio of unique to total mutual information over all possible third genes Z:

$$u_{X,Y} = \sum_{Z \neq X,Y} \left(\frac{\text{Unique}_Z(X;Y)}{I(X;Y)} + \frac{\text{Unique}_Z(Y;X)}{I(X;Y)} \right),$$

which quantifies the extent to which the pairwise mutual information is uniquely attributable to a direct relationship. In the inference algorithm, the redundant and qunique information is estimated for every triplet of genes and then the PUC is calculated for every pair of genes. To select the most relevant edges, PIDC incorporates *network context* by estimating empirical distributions of PUC scores for each gene. The confidence score of an edge between genes X and Y is then given by the product of their cumulative distribution functions evaluated at the observed PUC:

$$c = F_X(u_{X,Y}) \times F_Y(u_{X,Y}),$$

where F_X and F_Y are the cumulative distribution functions of PUC scores involving X and Y, respectively (Gamma distribution is assumed). This step ranks edges relative to each gene's context instead of using a global threshold. This approach mirrors the concept used in the Context Likelihood of Relatedness [46] algorithm and compensates for gene-specific variability, improving specificity and reducing false positives.

Chapter 3

Materials and methods

3.1. Single-cell datasets

This study makes use of several publicly available single-cell datasets that provide complementary perspectives on transcriptional regulation, chromatin accessibility, and the relationship between sequencing depth and downstream inference quality. These datasets were selected to provide complementary perspectives on transcriptional regulation: differentiation trajectories (dynamic) versus distinct cell types (static), and single-modality (RNA) versus multi-modality (RNA+ATAC). Together, these datasets allow us to systematically explore how sequencing budget allocation influences the recovery of biological structure and the inference of GRNs. Below is a short description of the datasets we used.

3.1.1. hHep (human hepatocyte differentiation)

One of the most popular datasets used for benchmarking comes from [47]. It captures the transcriptomic landscape of human induced pluripotent stem cells (iPSCs) differentiating into hepatocyte-like cells. A total of 425 single-cell transcriptomes were sequenced using the *Fluidigm C1* platform, across multiple time points during differentiation. It was included in the original BEELINE framework.

3.1.2. Kim23 (human liver organoids)

To compare how advances in sequencing technologies and the increasing volume of generated data impact GRN inference, we selected data from a tissue type similar to hHep. For this purpose, we used the dataset provided by [48]. The study also used expandable human liver organoids generated from induced pluripotent stem cells. Organoids were cultured in two main conditions: Hepatic Medium (HM), expandable condition used for maintenance, and Differentiation Medium (DM), more differentiated condition applied to induce further maturation. The dataset consists of single-cell RNA-seq and single-cell ATAC-seq profiles generated from the same biological sample through independent experiments. Because the two modalities were not measured on the same cells, it is not a multiome dataset in the strict sense. Instead, it represents two parallel assays that provide matched population-level transcriptomic and chromatin accessibility information. Four organoid samples were analysed for both sequencing types - two biological replicates for the HM condition and two biological replicates for the DM condition. A total of 39,300 cells were analysed by scRNA-seq and $\tilde{3}$7,000 nuclei by scATAC-seq. For our benchmarking, we focused on a single biological replicate of the HM condition (*kim23-hm1*), consisting of 9,988 sequenced nuclei.

3.1.3. Buenrostro18 (haematopoietic differentiation)

To assess performance in a well-studied developmental system distinct from the liver, we utilised the dataset from [49]. This dataset describes the differentiation of human haematopoietic stem cells (HSCs) into various blood lineages. It represents a classic "branching" trajectory where multipotent progenitors decide between myeloid, erythroid, and lymphoid fates. Like hHep, this dataset provides full-length transcript coverage and high sensitivity. Cells were sequenced using the *Smart-seq2* protocol on FACS-sorted cells. The total numer of the sequenced cells was 14,432 for scRNA-Seq and a total of 2,034 scATAC-Seq samples were sequenced.

3.1.4. PBMC10k (peripheral blood mononuclear cells)

To evaluate multi-omic inference strategies in a fully integrated setting, we utilised the "PBMC from a Healthy Donor (Granulocytes Removed)" dataset provided by 10x Genomics [50]. Unlike the Kim23 dataset, where RNA and ATAC data came from different cells, this dataset was generated using the *Chromium Single Cell Multiome ATAC + Gene Expression* technology. This platform allows for simultaneous profiling of chromatin accessibility and gene expression within the *same* biological nucleus, establishing a direct cell-specific link between the epigenetic landscape and transcriptional output.

Biologically, this dataset represents a highly heterogeneous mixture of fully differentiated immune cell types—including T-cells, B-cells, monocytes, and NK cells, rather than a continuous developmental trajectory. This allows us to test the ability of motif-informed algorithms to resolve cell-type-specific regulatory networks in a system defined by discrete, stable clusters rather than pseudotemporal gradients. The dataset contains 11,898 joint profiles, providing a robust test bed for high-throughput multi-omic integration.

3.2. Evaluation framework

The goal of this study is to evaluate how sequencing budget allocation, defined primarily by the number of profiled cells and the sequencing depth per cell, influences the quality of gene regulatory network (GRN) inference from single-cell data. To ensure methodological consistency and fair comparison across experimental conditions, we built our workflow around the BEELINE framework [51], which provides a standardised protocol for constructing reference networks, defining evaluation criteria, and assessing GRN inference methods. Furthermore, we evaluated regulatory network inference methods that leverage chromatin accessibility data to determine whether multiomic experiments represent a reasonable strategy for allocating sequencing budget.

While BEELINE was originally developed for benchmarking inference algorithms across fixed datasets, we extend its principles to a budget-centric simulation setting. Our analysis integrates cell subsampling, read downsampling, and metacell aggregation scenarios, allowing us to mimic a range of realistic experimental designs under limited sequencing resources.

3.2.1. The BEELINE Benchmarking Framework

The BEELINE framework consists of three core components: (1) curated single-cell datasets, (2) corresponding reference networks, or "gold standards," and (3) a standardised evaluation pipeline. While we utilised our own datasets, we adopted its procedures for reference network generation and performance evaluation.

Inference Methods for Benchmarking

The field of GRN inference is populated by dozens of computational methods, each built on different assumptions and algorithmic principles (e.g., correlation-based, regression-based, information-theoretic). The BEELINE study itself provides a benchmark of 12 distinct methods. For our study, we selected a subset of these methods based on their strong performance in the BEELINE benchmark and their widespread adoption, namely GENIE3, GRNBoost2 and PIDC. As documented in our initial results (Figure 4.1), we also performed a preliminary evaluation of other methods implemented in BEELINE, including SCODE, SINCERITIES, and Ppcor. These methods showed substantially lower performance on our datasets and were therefore excluded from the further analyses to focus on the behaviour of high-performance algorithms.

Reference networks (gold standards)

A fundamental challenge in GRN inference is the lack of a complete, universally true network for validation. BEELINE addresses this by constructing reference networks from high-confidence, experimentally validated or literature-curated interaction databases. Originally, RegNetwork, TRRUST and DoRothEA were considered as ground truth for non-specific networks. ENCODE, ChIP-Atlas and ESCAPE were scanned to generate cell-type specific networks, and STRING database was used for functional networks. Other evaluation strategies exist, perturbation analysis being the most prominent among them, yet data from perturbation studies is not as readily available. Choice of database derived ground truth can be further justified by its popularity in GRN benchmarking studies.[52]–[54]

After exploratory analysis, we opted to adapt the functional network as our reference network due to its superior performance. Following this methodology, our reference networks were derived from the **STRING database (v12.0)** [55]. STRING aggregates known and predicted protein-protein interactions, including both direct (physical) and indirect (functional) associations. We filtered these interactions to include only those with a high confidence score (combined score > 700), representing a trade-off between network density and evidential quality. It is crucial to acknowledge that these reference networks are inherently incomplete and contain a mixture of regulatory types (e.g., protein-protein, metabolic), not just the TF-gene interactions we aim to infer. Nevertheless, they represent the best available proxy for a biological "ground truth."

While the original BEELINE framework relied on earlier iterations of the database (v10.5 / v11), we opted for the most recent release to ensure our ground truth reflects the current state of human regulatory knowledge. STRING v12.0 integrates millions of new data points from recent high-throughput human studies and improved text-mining of the biomedical literature, totalling around 400,000 interacting pairs.

The Evaluation Universe

A key innovation of the BEELINE framework, which we adopted, is the concept of the "evaluation universe." GRN inference is typically performed not on the entire genome, but on a subset of genes, such as the most highly variable genes (HVGs). It would be unfair to penalise an algorithm for failing to find an edge if one of the genes in that edge was not provided as input, which was illustrated in this work in Figure 4.4.

Therefore, for each inference task, the evaluation is restricted to a specific **evaluation universe**. This universe is defined as the intersection of genes present in both the inference input (e.g., the 1000 selected HVGs) and the reference network. The gold standard is subsetted

to this universe, and all metrics are calculated only on the potential edges within this common set of genes. This ensures a fair comparison, as seen in our analysis in Figure 4.4.

TFs curation

To restrict the search space of the inference algorithms to biologically plausible regulators, BEELINE authors utilised a curated list of 1,563 human transcription factors that we adopted. This reference list was compiled from multiple high-confidence databases of regulatory interactions, ensuring that candidate regulators have prior functional evidence. The list integrates transcription factors from: RegNetwork, aggregating transcriptional and post-transcriptional regulatory relationships, TRRUST v2, a manually curated database of literature-mined regulatory interactions, DoRothEA, specifically including TFs with high-confidence regulons (confidence levels A, B, and C). This filtering step focuses the inference task on regulators with established biological relevance.

3.2.2. Experimental pipeline

Our computational analysis was structured around a series of experiments designed to simulate different resource allocation strategies. The general pipeline for downsampling experiment was as follows:

1. **Data subsetting (simulation):** The full, quality-controlled datasets were subsetted to create a simulated input matrix. This involved (a) **cell subsampling**, where a specified number of cells was drawn uniformly at random, and (b) **read downsampling**. To simulate shallower sequencing, the UMI counts for the selected cells were downsampled by modeling the new count for each gene as a draw from a **binomial distribution** $B(n, p)$, where n is the original UMI count and p is the desired proportion of reads to retain (e.g., $p = 0.7$ for 70% depth).

2. **Preprocessing:** For each simulated matrix, we performed standard single-cell preprocessing using `scanpy` [56]. This included library size normalization (to 10,000 counts per cell), log-transformation, and selection of the top N highly variable genes (HVGs) plus all known transcription factors that were considered highly variable.

3. **GRN Inference:** The resulting expression matrix was used as input for the selected GRN inference methods.

4. **Evaluation:** The ranked edge list produced by each method was evaluated against the corresponding STRING-derived gold standard (subsetted to the correct evaluation universe).

Fixed-budget allocation While the previous experiment assessed variables independently, sequencing is often bound by a finite financial cap. To model this, we designed a "Fixed-budget" simulation where the total number of UMIs (B_{total}) is held constant, necessitating a direct trade-off between the number of sequenced cells and the depth per cell.

We defined four budget tiers based on the total available UMIs in the quality-controlled Kim23 dataset: **Nano** (10% of total, $\approx$ 7M UMIs), **Low** (25%, $\approx$ 18M UMIs), **Medium** (50%, $\approx$ 36M UMIs), and **High** (75%, $\approx$ 54M UMIs). Within each tier, we defined an experimental strategy parameter $\alpha \in [0, 1]$ to represent the allocation preference. The lower bound, Depth-First ($\alpha = 0$), minimises the cell count to maximise read depth (up to the original sequencing levels), while the upper bound, Cells-First ($\alpha = 1$), maximises the cell count at the expense

of read depth. Intermediate strategies were simulated by sampling random subsets of cells and applying binomial thinning to their counts to precisely match the target B_{total}.

Metacells Another analysis evaluated whether computational aggregation could mitigate the sparsity issues associated with low sequencing depth. Unlike previous experiments that altered raw data quantity, this approach modifies data topology by grouping transcriptionally similar cells into metacells. We evaluated two distinct aggregation algorithms: **K-Means**, a geometric approach grouping cells based on Euclidean distance in PCA space, and **SEACells** [57], a manifold-aware method utilizing archetype analysis to strictly preserve biological phenotypes. To determine the optimal level of aggregation, we varied the granularity parameter k (the target number of cells per metacells). This allowed us to assess the trade-off between the denoising benefits of high aggregation (high k) and the retention of biological resolution (low k).

Further details of the specific simulations, are described in their respective sections in Chapter 4.

Implementation and reproducibility

To address dependency requirements of each tool, we employed an environment isolation strategy using **Conda**.

- **Execution environment:** The core expression-based methods (GENIE3, GRNBoost2, PIDC) were executed within the standardised BEELINE (v1.0) framework to ensure comparability with established benchmarks.
- **Other methods integration:** Newer GRN inference methods and other tools were deployed in dedicated, isolated Conda environments to satisfy their specific library dependencies. Specific verisons used were: CellOracle (v0.20.0), SCENIC+ (v1.0a1), SEACells (v0.3.3)
- **Custom wrappers:** To integrate new methods into a unified pipeline, we developed custom Python wrappers. These wrappers managed data input/output standardisation, allowing methods with different requirements to be evaluated against the common BEELINE scoring metrics.

Code for this work, including wrapper code and preprocessing steps with scanpy and figures generation are available in the project repository[1].

3.2.3. Evaluation Metrics

GRN inference methods do not output a binary network. Instead, they either produce a ranked list of all possible regulatory edges, ordered from most confident to least confident, or return a subset of interactions that passed some threshold to be considered an edge with a corresponding numerical values representing the method's confidence. The task is therefore to evaluate the quality of this ranking. This is a classic binary classification problem, for which BEELINE employs two primary metrics: the Area Under the Receiver Operating Characteristic (AUROC) and the Area Under the Precision-Recall (AUPR) curve.

Given a ranked list, a threshold is moved from the top-ranked edge to the bottom. At each threshold, all edges above the threshold are considered "predicted positives," and all below are "predicted negatives." These are compared against the gold standard to compute:

[1] `https://github.com/szysztof17/thesis`

- **True Positives (TP):** Edges present in both the prediction list (above threshold) and the gold standard.
- **False Positives (FP):** Edges in the prediction list but not in the gold standard.
- **True Negatives (TN):** Edges absent from both the prediction list and the gold standard.
- **False Negatives (FN):** Edges absent from the prediction list but present in the gold standard.

From these counts, we can calculate the key rates:

- **True Positive Rate (TPR)**, or **Recall**: The fraction of all true edges that are correctly identified.
$$TPR = \text{Recall} = \frac{TP}{TP + FN}$$
- **False Positive Rate (FPR):** The fraction of all false edges that are incorrectly identified as positive.
$$FPR = \frac{FP}{FP + TN}$$
- **Precision:** The fraction of predicted positive edges that are truly positive.
$$\text{Precision} = \frac{TP}{TP + FP}$$

Area Under the ROC Curve (AUROC) The ROC curve plots the TPR (y-axis) against the FPR (x-axis) at all possible thresholds. The **AUROC** is the area under this curve. An AUROC of 1.0 represents a perfect classifier, while an AUROC of 0.5 represents a classifier that is no better than random guessing. This metric is robust and widely used, and it is the primary metric used in our analyses (e.g., Figure 4.1).

Area Under the Precision-Recall Curve (AUPR) The Precision-Recall (PR) curve plots Precision (y-axis) against Recall (TPR, x-axis). The **AUPR** is the area under this curve. Gene regulatory networks are extremely sparse, meaning the number of true negative edges (TN) vastly outnumbers the true positive edges (TP + FN). In such highly imbalanced datasets, AUROC can sometimes be overly optimistic. AUPR is considered a more informative metric in these scenarios because it does not depend on the number of True Negatives and is more sensitive to the performance at the top of the ranked list, which is often what biologists are most interested in.

AUPR Ratio One limitation of the raw AUPR score is that, unlike AUROC, its baseline value for a random predictor is not fixed at 0.5. Instead, the random baseline varies depending on the sparsity of the network; specifically, it equals the background density of positives in the evaluation universe ($P/(P+N)$). This makes it difficult to compare AUPR values across different datasets or biological contexts where the network density might differ.

To address this, we utilise the **AUPR Ratio**, which normalises the AUPR against the random baseline:

$$\text{AUPR Ratio} = \frac{\text{AUPR}_{\text{model}}}{\text{AUPR}_{\text{random}}}$$

where $\text{AUPR}_{\text{random}}$ represents the proportion of true edges in the total set of possible edges within the evaluation universe. An AUPR Ratio of 1.0 indicates performance equivalent to random guessing, while a value greater than 1.0 indicates that the model is successfully enriching for true regulatory interactions.

Early Precision Ratio (EPR) While AUROC and AUPR measure global performance across the entire ranked list, in practical applications, researchers are often only interested in the most confident predictions for experimental validation. To capture this, we calculate **Early Precision**, defined as the fraction of true positives among the top k edges, where k is typically set to the number of true edges in the ground truth network (this choice of k was also chosen by BEELINE authors).

To ensure comparability across datasets with varying sparsity, we report the **Early Precision Ratio (EPR)**. This metric normalises the Early Precision against the background network density, or percantage of edges from all possible vertices pairs:

$$\text{EPR} = \frac{\text{Precision@}k}{\text{Background Density}}$$

An EPR value greater than 1.0 signifies that the algorithm's top-ranked predictions are enriched for true regulatory edges compared to a random selection. This metric is particularly critical for assessing a method's utility in generating high-confidence candidates for downstream analysis.

Chapter 4

Results

4.1. Comparison of selected methods

While a comprehensive evaluation of GRN inference methods was not the primary focus of this work, a limited comparison of selected algorithms was included as part of the analysis to assess their suitability for downstream applications. To guide this selection, we referred to the BEELINE benchmark study, which offers a standardised comparison of GRN inference methods across datasets and metrics. We selected six representative methods from this benchmark for an initial comparison: **GENIE3**, **GRNBoost2**, **PIDC**, **SCODE**, **SINCERITIES**, and **Ppcor**. This group includes methods with diverse algorithmic strategies. Although our project's primary focus is on inferring static networks (a task suited for methods like GENIE3, GRNBoost2, PIDC, and Ppcor), we also included methods that are designed to leverage pseudotime or time-series data (SCODE and SINCERITIES). We included this latter group to provide a comprehensive baseline, as the *hhep* dataset consists of multiple time points that can be used to construct a simple trajectory.

We first benchmarked these six methods on the *hhep* dataset using the AUROC curve, as shown in Figure 4.1.Methods such as GRNBoost2, GENIE3, and PIDC demonstrated consistently higher performance across datasets, achieving relatively strong AUROC values regardless of network size or reference standard. In contrast, methods such as SCODE, SINCERITIES, and Ppcor performed substantially worse in this benchmark, exhibiting low predictive accuracy and limited reliability across evaluation scenarios. For this reason, they were excluded from further analysis.

As discussed in our Methodology, AUPR is often considered a highly relevant metric for GRN inference due to the extreme class imbalance of the data. We therefore also evaluated performance using AUPR. As anticipated, the absolute AUPR values are substantially lower than the AUROC scores (e.g., 0.05-0.1 for best performing methods on dataset-specific networks). These seemingly low scores are not a sign of poor performance, but rather an expected and direct consequence of the sparsity of the gold-standard network. In a highly sparse network where the fraction of true positive edges might be very small, the AUPR for a random classifier is not 0.5, but is instead equal to this very small fraction (e.g., baseline AUPR ≈ 0.01). To account for this low baseline, the BEELINE authors and others often use an **AUPR Ratio**, calculated as the observed AUPR divided by the random baseline. Therefore, the scores achieved by our top methods represent a $\approx$ 5-fold improvement over random guessing, confirming their strong predictive power.

Another metric used in benchmarking studies is **Early Precision** a metric that evaluates the precision of the top-k ranked predictions (where k is the number of true edges as per

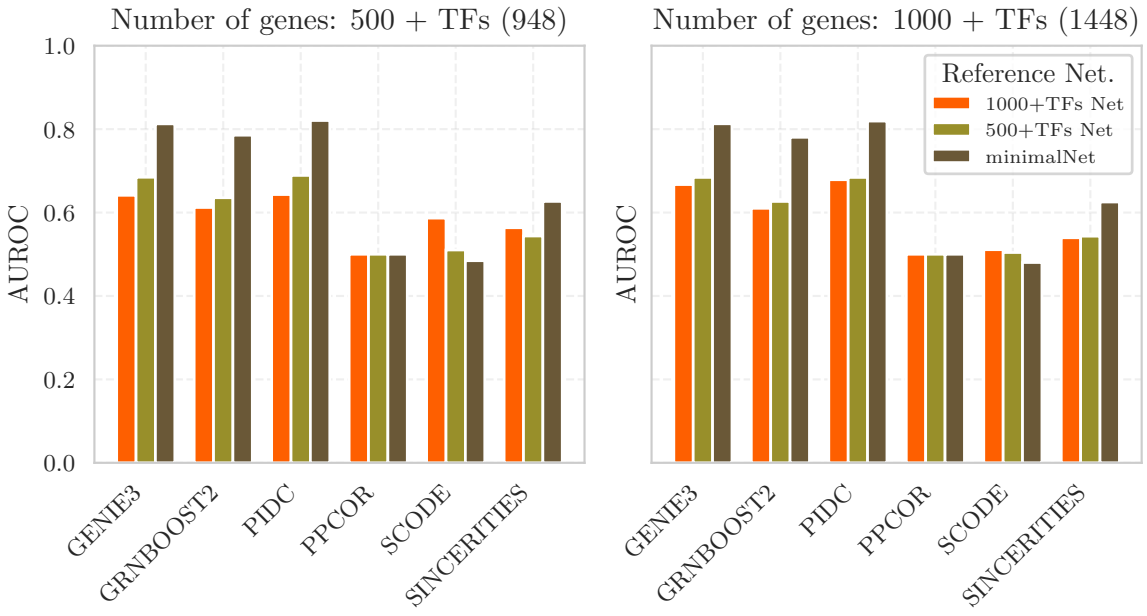

Figure 4.1: Performance comparison of GRN inference methods based on AUROC scores for networks inferred from 500 and 1000 most variable genes (plus variable transcription factors) from the hHep data, evaluated against three reference networks: 1000+TFs Net (STRING12 subset induced by the 1000+TFs gene dataset), 500+TFs Net (induced by 500+TFs genes), and minimalNet (only 100 most highly variable genes intersected with the StringDB). Each bar represents the AUROC score of a method's predictions compared to the given reference network. The x-axis shows the inference method, while the bar colors correspond to the evaluation network used. GRNBoost2, GENIE3, and PIDC demonstrate consistently strong performance, whereas SINCERITIES, SCODE, and PPCOR underperform across both gene sets and evaluation networks.

BEELINE). Similar to AUPR, the raw early precision value is not informative about the performance without the knowledge of gold standard network structure. Therefore, **Early Precision Ratio** is used, where the EP values are compared to those of a random predictor.

What is most important for our purposes is that the relative ranking of the methods remains identical to the AUROC analysis: GENIE3, GRNBoost2, and PIDC are the clear top-tier, while the other three methods perform at or near the random baseline. This consistency across both metrics gives us high confidence in our method selection. Also, based on the preliminary analysis, we decided that AUROC is the most suitable primary metric. It is more stable and interpretative, its 0.5 baseline is intuitive and constant, and it provides a clearer signal for comparing experimental strategies. We therefore deliberately chose AUROC as a primary metric for subsequent analyses.

A final consideration in method selection is the computational budget, or the time and resources required to run the analysis. Table 4.2 details the user time used by each method. The results reveal a striking trade-off between performance and cost.

The two top-performing tree-based methods, GENIE3 and GRNBoost2, are also the most computationally expensive, with GENIE3 taking over 6 hours of user time for the 1000-gene set. SCODE also falls into this high-cost category. In stark contrast, PIDC, which performed as well or better than the tree-based methods, is remarkably fast, using less than 30 minutes

Table 4.1: Precision-based metrics for the six GRN inference methods on the hHep dataset. The top-tier methods (GENIE3, GRNBoost2, PIDC) consistently outperform the others across all metrics.

	AUPRC		AUPRC Ratio		EPR	
Dataset Algorithm	500+TFs	1000+TFs	500+TFs	1000+TFs	500+TFs	1000+TFs
GENIE3	0.08	0.06	4.85	5.07	9.45	11.44
GRNBOOST2	0.07	0.04	3.79	3.78	7.59	8.78
PIDC	0.10	0.07	5.54	5.67	11.30	13.54
SCODE	0.02	0.01	1.00	1.02	0.66	0.78
SINCERITIES	0.02	0.01	1.14	1.13	0.87	0.82
PPCOR	0.02	0.01	1.00	1.00	0.00	0.00

Table 4.2: Computation times (in minutes) by algorithm and dataset.

	500+TFs		1000+TFs	
Time [min] Algorithm	User	Wall	User	Wall
GENIE3	201.57	62.42	389.22	95.47
GRNBOOST2	52.40	6.00	125.19	10.23
PIDC	6.95	6.97	29.93	29.96
SCODE	124.78	124.78	198.98	199.00
SINCERITIES	3.70	3.70	7.50	7.52
PPCOR	0.15	0.16	0.47	0.47

of user time for the large dataset. The other low-performing methods, SINCERITIES and PPCOR, were also very fast, however useless. This makes PIDC a clear standout, offering both state-of-the-art accuracy and a highly efficient runtime. It is important to note that computational cost can vary substantially depending on the implementation. For instance, both GENIE3 and GRNBoost2 are parallelised, which significantly reduces their elapsed wall time. In contrast, PIDC could potentially be optimised to better leverage high-performance computing resources.

Summing up the benchmark, we chose GENIE3, GRNBoost2 and PIDC as our expression based GRN inference algorithms, due to their better performance and having demonstrated the nuances of precision-based metrics, we will use AUROC as the primary metric for the rest of this thesis. It is the most stable, its 0.5 baseline is intuitive and constant, and it provides the clearest signal for comparing the relative performance of different experimental strategies.

4.2. Evaluation of pre-processing parameters and dataset specificity

4.2.1. Impact of quality control strategies on inference accuracy

Quality control is an essential first step in single-cell RNA sequencing data analysis, playing a crucial role in ensuring the accuracy of downstream analyses. We aimed to assess how quality control can influence GRN inference, as evaluated by the AUROC metric. Quality control

thresholds are not universal - they vary depending on the specific experiment, tissue type, and cell population analysed. One commonly used metric is the percentage of mitochondrial gene expression, often assigned a fixed cut-off of 5%. Although this threshold is widely adopted in many studies and is set as the default in analysis software, it has been shown to introduce biases and unnecessarily exclude certain cell types.[58], [59]

A widely used strategy for quality control is data-driven, where threshold values are determined by examining the characteristics of the actual dataset. For instance, the percentage of mitochondrial gene expression is often used as an indicator of dying or otherwise unfit cells. However, this metric varies substantially across tissues. Healthy human liver cells can have mitochondrial gene counts up to four times higher than those of healthy lung epithelial cells.[60] Some researchers opt for fixed thresholds, while others apply statistical criteria such as standard deviations from the mean to define cutoffs. It is also important to consider the inherent heterogeneity of cell populations, as different cell types exhibit distinct gene expression patterns based on their biological roles, activity levels, and metabolic demands. Consequently, identifying optimal quality control parameters remains a complex and nuanced task.[61]

As our dataset, we used kim23. A summary of its key QC metrics is shown in fig. 4.2. The primary parameter of interest was n_genes_by_counts, which reflects the total transcript count per cell. This metric displayed a bimodal distribution, a feature of particular significance for downstream analysis. Additionally, the mitochondrial gene content in the cells was markedly higher than the commonly used thresholds of 5% or even 10%. To avoid discarding potentially valuable cells, we adopted a more lenient threshold of 25%, which we found to be a reasonable compromise.

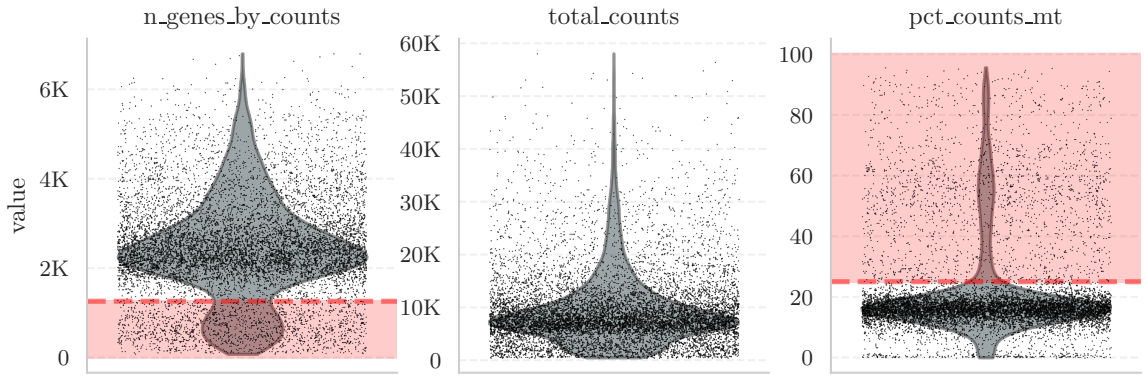

Figure 4.2: Quality control metrics for the kim23 dataset. The vertical line marks the local minimum at 1259 in the distribution of detected genes per cell, used as a gene count threshold. The horizontal line indicates the fixed mitochondrial content threshold of 25%.

To assess the influence of QC on the accuracy of GRN inference, we used two approaches: applying a fixed threshold and discarding cells based on percentiles. The results are shown in fig. 4.3. The first and rather obvious observation is that the standard 5% mitochondrial cut-off is entirely unsuitable in our case, leading to the loss of nearly all information. The results also show that discarding cells with high mitochondrial content is a crucial step, significantly improving performance. Our chosen threshold of 25% appears to be an effective choice. Our initial observation of the bimodal distribution in gene counts led us to select the local minimum between the two modes as a cut-off value. However, as seen in the results, this choice leads to a decrease in performance. This may be attributed to the heterogeneity of the analysed cells and their transcriptional activity at the time of sequencing. The bi-modality might reflect

distinct biological states rather than technical artifacts or unfit cells. The last plot compares tow filtering strategies, discarding both extremes versus discarding only low gene count cells. Both strategies show a general improvement in AUROC as the lower percentile threshold increases up to 10%, beyond 10%, performance slightly drops or plateaus. The low-high filtering (red) slightly outperforms low-only (green) in most, but not all, cases. However, the difference between the two strategies is minor, suggesting that the upper percentile cut-off has a limited effect on GRN inference accuracy. The lower tail of the gene count distribution (cells with very few genes detected) clearly contains low-quality or uninformative cells, and removing them improves GRN inference. Removing cells from the upper tail (high gene counts) brings only marginal benefit. These may represent biologically distinct or highly active cells more often than artifacts. Therefore, a lower-tail filtering strategy alone (low-100) appears to strike a good balance between retaining data and improving inference quality.

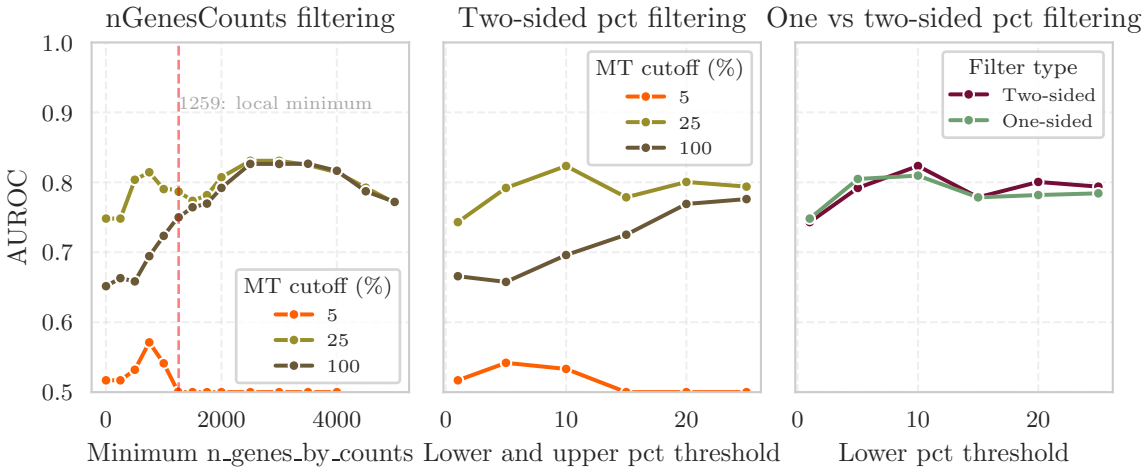

Figure 4.3: Effect of different QC filtering strategies on GRN inference accuracy. (1) Fixed minimum gene count threshold filtering: cells with fewer than a specified number of expressed genes are removed. (2) Two-sided percentile-based filtering: both low and high outliers (based on gene count distribution) are excluded. (3) One-sided vs two-sided percentile filtering: comparison of discarding only low-expression cells vs. discarding both low and high outliers. AUROC values are shown for each strategy.

4.2.2. Influence of HVG selection on network recovery

To evaluate how the number of input genes affects the performance of gene regulatory network inference, we tested multiple thresholds for selecting highly variable genes (HVGs). For each HVG cutoff, a subset of genes was selected and used to infer a network, which was then evaluated against two networks: a reference network adapted to the gene set - that is, only edges between selected genes were considered in the reference, in line with the BEELINE benchmarking framework, and also a common network, where only genes present in all gene sets were considered. In general, the common network yields higher AUROC values across all HVG cutoffs, consistent with previous findings.

As shown in fig. 4.4, increasing the number of genes results in a slight but consistent decline in AUROC, regardless of the reference network used. This trend likely reflects the dilution of the regulatory signal by the inclusion of less informative genes with lower variability. When fewer highly variable genes are used, the signal-to-noise ratio seems higher, leading

to better predictive performance. While AUROC remains relatively stable, the differences across settings suggest that performance evaluation is sensitive to the relationship between the learning and evaluation universes. This is less pronounced when reference network is fixed (and does not increase in size with increasing HVGs number) and becomes clearer with networks adapted to the prediction space, as is the case in BEELINE-like evaluation.

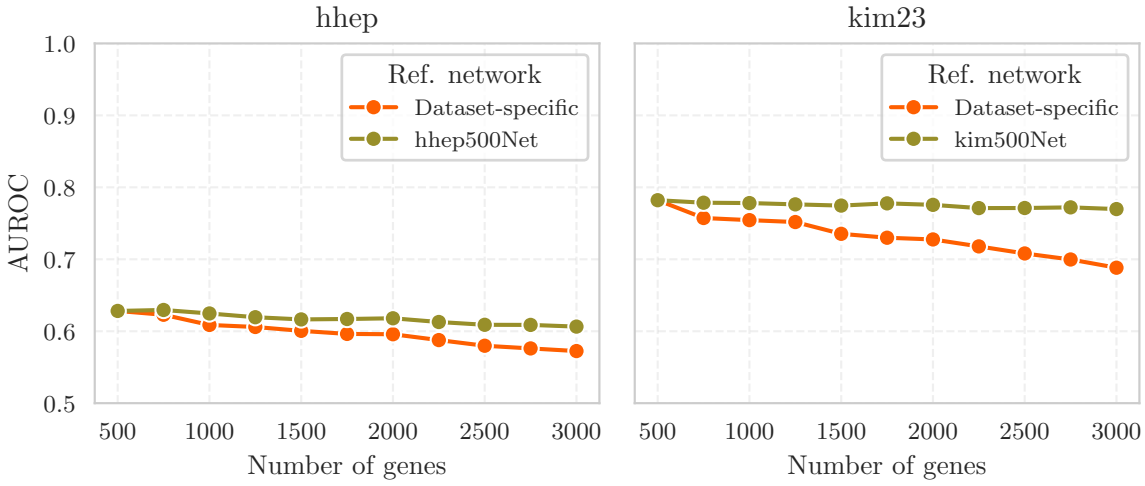

Figure 4.4: Comparison of AUROC values for GRN inference across different cutoffs for highly variable genes, evaluated against two types of reference networks: dataset-specific (based on STRING12) and intersection of networks across all datasets (hhep500Net and kim500Net, respectively).

4.2.3. Assessment of cross-dataset generalisability

To evaluate how well GRN inference generalises across datasets, we performed a comparison between *kim23* and *hhep* — two scRNA-seq datasets derived from human liver organoid tissue. Although the biological origin is the same, the datasets differ significantly in size, sequencing depth, and experimental protocols. The *kim23* dataset is more recent and contains substantially more cells, potentially enabling the inference of more comprehensive and accurate regulatory relationships.

For both datasets, we constructed dataset-specific gold standard networks using STRING-derived interactions. GRN inference models were evaluated both within-dataset and cross-dataset settings, that is, using predictions from one dataset and evaluating them on the reference network of the same or the other dataset. As shown in fig. 4.5, models performed substantially worse in the cross-dataset setting for kim23, despite the shared tissue origin. When using *kim23-hm1* predictions evaluated on its own network, AUROC scores exceeded 0.75, but dropped to $\tilde{0}.55$ when scored against the *hhep*-derived network. Notably, this is not true for the reverse comparison - AUROC values are similar regardless of the evaluation network.

This is consistent with the network structure statistics. The *hhep*-based network contains 14,524 regulator–target interactions, while *kim23-hm1* contains only 4,636. They share only 1,726 edges, and only 152 genes overlap as regulators or targets. Such limited overlap confirms that the inferred networks are dataset-specific, highly dependent on the choice of variable genes used for inference.

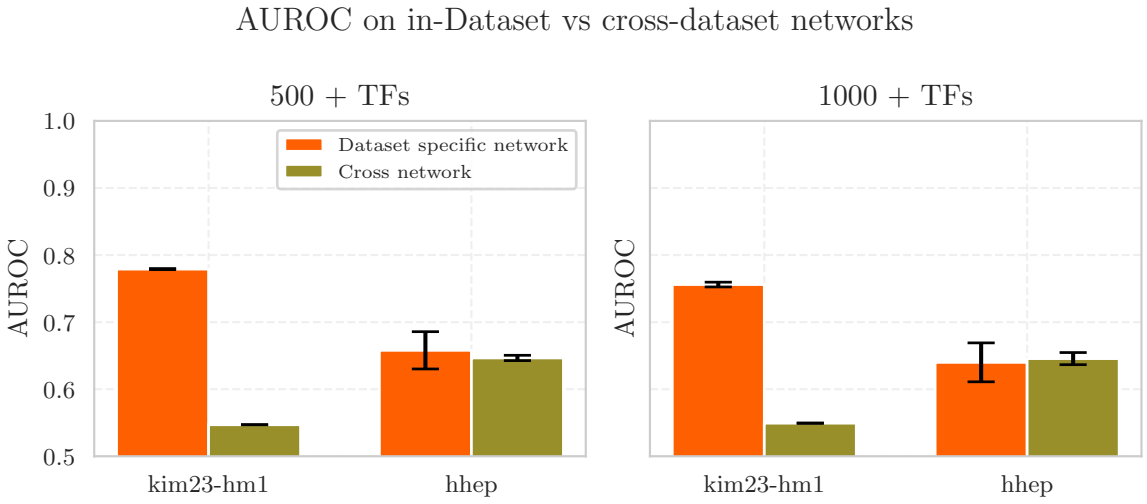

Figure 4.5: AUROC scores for GRN inference models evaluated on networks derived from the same dataset and from the other dataset, shown separately for 500 and 1000 genes settings. A substantial performance drop is observed in the cross-dataset setting.

To further isolate the effect of dataset quality, we performed an additional control experiment in which the gene sets and reference networks were fixed to those from *hhep*, while the input expression data came either from *hhep* or from *kim23-hm1*. Specifically, we used the same 500 and 1000 highly variable genes (HVGs) selected in *hhep* and applied them to *kim23-hm1* expression profiles (referred to as *kim23-hm1-extgenes*). As expected, the networks used for evaluation were identical in both cases. This isolates the effect of the dataset alone, keeping both the evaluation network and the gene set constant. The hhep gene set was selected because of the 1448 HVGs, only 16 were not present in the kim23 dataset, whereas only half of the HVGs from kim23 were present in the hhep dataset.

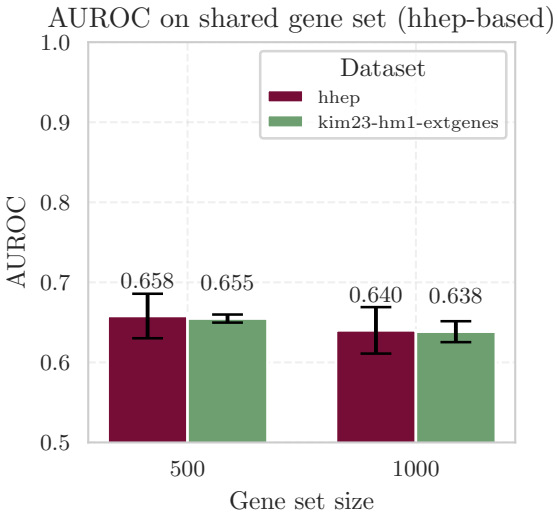

Figure 4.6: AUROC scores when using the same gene set and hhep-based network, but different expression datasets. No consistent performance improvement is observed with the newer *kim23-hm1-extgenes* data.

The results are shown in fig. 4.6. AUROC was slightly lower when using *kim23-hm1-extgenes* compared to *hhep*, both for 500 HVGs (0.638 vs. 0.658) and 1000 HVGs (0.640 vs. 0.655). These results indicate that switching to a newer or larger dataset does not lead to

consistent performance improvement when the gene set and the evaluation network remain constant. Notably, the variance in performance across methods appears to be reduced when using the larger *kim23-hm1* dataset, suggesting that increased cell numbers may help stabilise inference results.

4.3. Metacell aggregation

Transitioning from bulk analysis to single-cell resolution inherently introduces a significant increase in data sparsity which can be addressed with pseudobulking. This approach reduces technical noise by grouping cells, often by cell type, and pooling signals from a given group of cells, which improves statistical power for differential gene expression or chromatin accessibility analyses. However, pseudobulk loses the single-cell resolution, making it difficult to capture rare cell types or intermediate states and dynamic relations between them.[62]

Another common strategy to address the challenges of data sparsity and noise in scRNA-seq data is to aggregate cells into *metacells*. This approach creates average expression profiles for small, homogeneous groups of cells, which can potentially denoise the data, reduce computational burden, and improve the accuracy of downstream analyses like GRN inference. To investigate this, we compared the performance of GRNs inferred from metacell-aggregated data against a baseline of networks inferred from the full single-cell dataset.

We evaluated two distinct metacell generation techniques: a standard **K-Means** clustering algorithm and **SEACells**, a more sophisticated method based on a neural network model [57]. The performance of three GRN inference methods was assessed across a range of metacell size (k), where a higher k signifies a higher degree of data aggregation.

The results, shown in Figure 4.7, reveal that the utility of metacell aggregation is highly dependent on the chosen GRN inference method and to a lesser degree the aggregation technique itself.

Perhaps the most striking finding is the catastrophic impact of aggregation on **PIDC** results. While PIDC achieved the highest baseline AUROC of all tested methods (0.785), its performance worsened with metacell generating and plummeted to near-random levels when applied to highly aggregated data, a phenomenon not observed in other methods, regardless of the aggregation technique. This suggests that the statistical information PIDC relies upon is fundamentally lost with the metacells aggregation process. Perhaps its unique statistical mechanism makes PIDC particularly vulnerable to the loss of cell-to-cell variability introduced by aggregation. Since PIDC relies on estimating mutual information across continuous expression distributions, the metacell generation process effectively loses the natural fluctuation in gene expression that drive these estimates. As a result, the subtle nonlinear relationships that PIDC is designed to detect become indistinguishable from background variation, causing the method to collapse toward random performance. This sensitivity suggests that PIDC may depend more heavily than other GRN inference tools on fine-grained single-cell heterogeneity, making it incompatible with strong aggregation strategies.

Similarly, **GRNBoost2** did not benefit from data aggregation. For all values of k, its performance on metacell data was consistently lower than its baseline AUROC of 0.752. This indicates that, like PIDC, GRNBoost2 relies on information present at the single-cell level that is erased by aggregation.

In contrast, **GENIE3** is the only algorithm that robustly benefits from the aggregation approach. For nearly all levels of aggregation, its performance surpassed its single-cell baseline AUROC of 0.760. This suggests that GENIE3 algorithm is less sensitive to the loss of single-cell resolution and can effectively leverage the denoised signal present in metacell profiles to

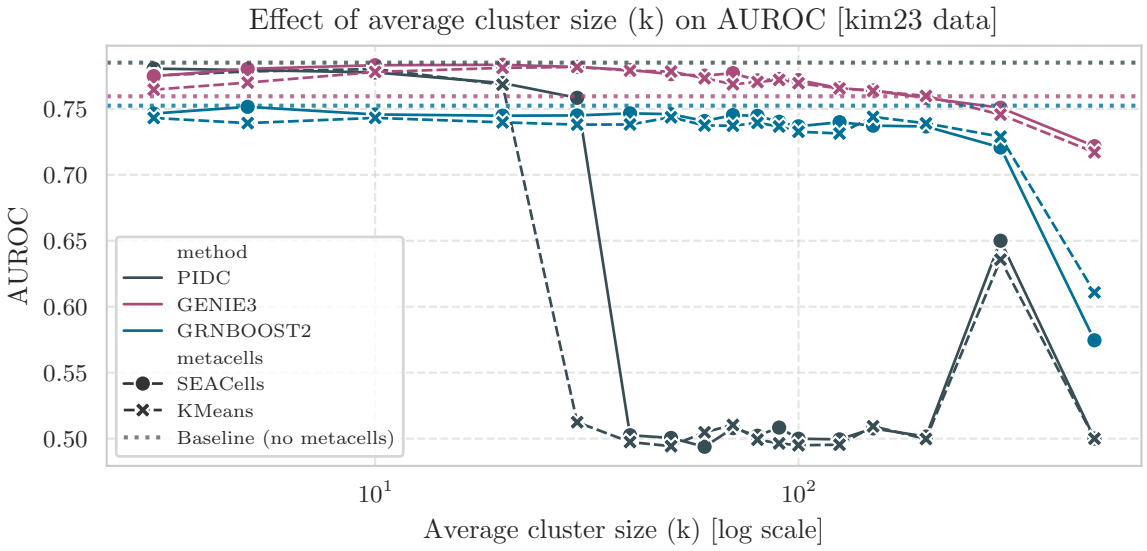

Figure 4.7: The effect of metacell aggregation on GRN inference accuracy. AUROC scores are plotted against the number of metacells (k) for three inference methods: PIDC (gray), GRNBoost2 (blue), and GENIE3 (rouge). Performance is shown for two aggregation techniques: SEACells (circles) and KMeans (crosses). Horizontal lines indicate the baseline performance for each method on the non-aggregated single-cell data. The results show that the impact of metacells is highly method-specific.

improve its predictions.

While the primary determinant of success was the inference algorithm, the choice of aggregation method also had a clear impact. Comparing the two techniques, the more sophisticated neural network-based **SEACells** method consistently yielded better outcomes than the standard **K-Means** clustering approach. This suggests that SEACells does a better job of preserving the relevant biological structure within the data during aggregation, although this comes at a higher computational cost - running time reached days for clustering into small metacells (a few cells in size) while it remained seconds for K-Means.

An additional insight from the analysis is the role of the number of metacells, k, particularly for GENIE3, the only method that benefited from the aggregation. The relationship between performance and degree of aggregation is not monotonic. Rather, the data suggest an optimal range. As shown in Figure 4.7, the performance of both aggregation techniques increases with the degree of aggregation and peaks around an intermediate value of $k = 50$. As aggregation increases further, performance begins to plateau or decrease slightly. This pattern suggests the existence of an optimal trade-off: sufficient aggregation is needed to denoise the data, but excessive aggregation may begin to obscure important biological heterogeneity, leading to a decline in accuracy. However, for PIDC and GRNBoost2, this trade-off is not observed, as any amount of aggregation proved to be detrimental.

While metacell aggregation is often motivated by the potential for improved accuracy, it also has significant implications for computational cost. By reducing the number of input data points from thousands of cells to a few hundred metacells, the runtime of downstream GRN inference can be drastically reduced. However, the metacell generation process itself introduces a computational cost. The time-accuracy trade-off reveals different outcomes for each method. For GENIE3, metacell aggregation presents a clear "win-win" scenario, simul-

taneously improving both performance and computational speed. In contrast, GRNBoost2 presents a practical compromise: while accuracy slightly decreases, the runtime is substantially reduced, offering a worthwhile trade-off for large-scale analyses, as shown in Figure 4.8. For PIDC, however, the severe drop in accuracy makes metacell aggregation an unfavorable strategy, regardless of any speed improvements.

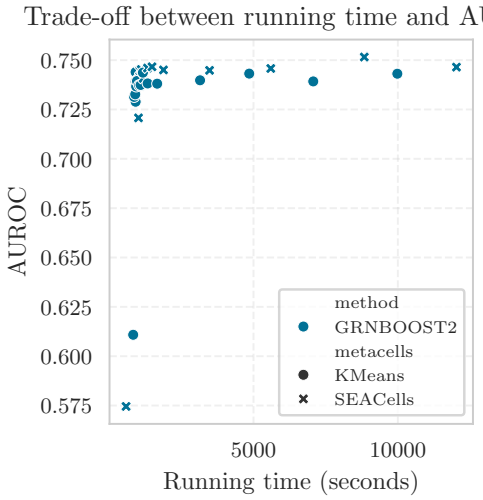

Figure 4.8: The trade-off between running time and AUROC for GRNBoost2. The plot exemplifies that it may be worth trading accuracy for substantial complexity cost reduction.

In summary, our findings show that metacell aggregation is not a universally beneficial preprocessing step. Its utility is highly dependent on the specific GRN inference algorithm used. For methods like PIDC and GRNBoost2, aggregation consistently degraded performance. In contrast, GENIE3's accuracy was robustly improved by the approach. Therefore, whether to use metacells should be carefully considered in the context of the chosen inference pipeline.

4.3.1. Context dependency and generalisability

To verify our initial findings and better understand the conditions under which metacell aggregation acts beneficially, we extended our evaluation to two contrasting datasets: **Buenrostro18**, which characterises continuous hematopoietic differentiation, and **PBMC10k**, composed of different types of mature immune cells.

This comparison demonstrates that the utility of metacells is not universal, but is strictly dictated by the underlying topology of the data. In the **Buenrostro18** dataset, we observed the strongest gains from aggregation. As shown in Figure 4.9, applying SEACells aggregation led to a significant 14.5% increase in AUROC for GENIE3 (rising from 0.673 to 0.771) and a notable 5.2% increase for GRNBoost2.

We hypothesise that this success is due to the continuous nature of the differentiation manifold. In such dynamic systems, single-cell data is often "gappy," meaning intermediate states are underrepresented. Aggregation via SEACells effectively performs *manifold smoothing*, filling in the dropouts along the trajectory from progenitor to mature cells.

On the other hand, the **PBMC10k** dataset showed no such benefit. As detailed in Figure 4.9, performance for all inference methods declined or stagnated as aggregation (k) increased. This suggests that for datasets with discrete, well-separated cell types, the "denoising" benefit of metacells is negligible. Unlike continuous trajectories, discrete clusters do not possess intermediate transition states to be smoothed. Consequently, aggregation effectively reduces sample size without adding structural information, leading to a loss of statistical

power for inferrence methods. The finding that is observed universally is the catastrophic effect of metacell generation on PIDC performance. We attribute this to the statistical properties of PIDC's mutual information estimator, which relies on distributional variance that aggregation eliminates.

This discrepancy highlights that the success of metacell aggregation is highly dependent on specific dataset characteristics, particularly the underlying biological heterogeneity. While pseudobulking strategies can be powerful, they are not a "one-size-fits-all" solution and must be applied with the specific topological properties of the input data in mind.

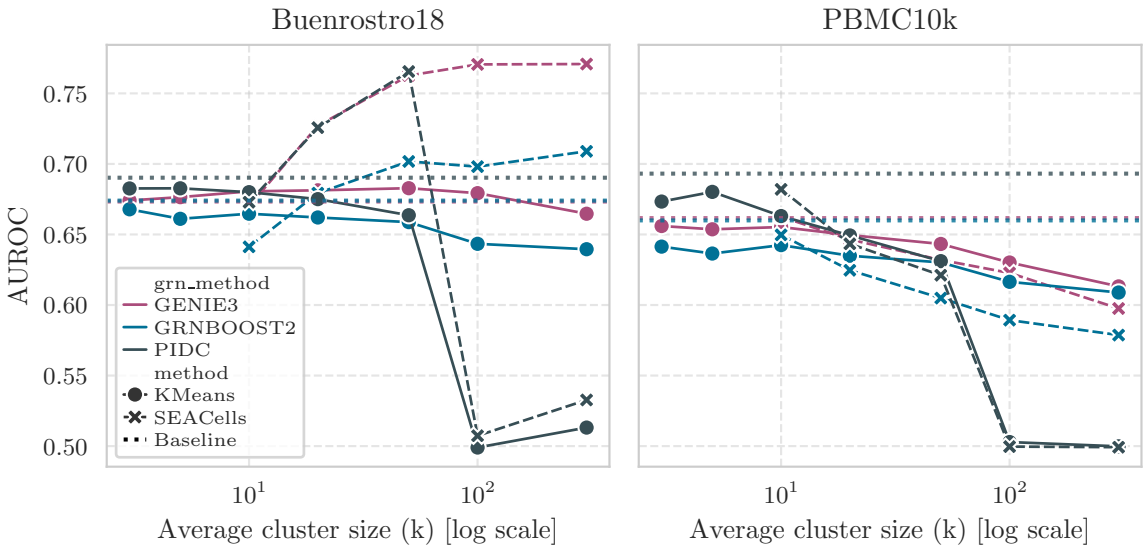

Figure 4.9: The effect of metacell aggregation on GRN inference accuracy. AUROC scores are plotted against the number of metacells (k). Two additional datasets were included: buenrostro18 (left), a dataset characterizing continuous hematopoietic differentiation, and PBMC10k (right), a dataset composed of distinct, mature immune cell types. Substantial improvement in AUROC values is seen for buenrostro18 data with GENIE3 and GRNBoost2. For pbmc10k dataset, consistent drop in performance is seen.

4.4. Evaluation of sequencing depth and sample size trade-offs

4.4.1. Separating the effects of cell number and sequencing depth

To build a foundational understanding of how resource limitations affect GRN inference, we first designed an experiment to independently assess the impact of reducing cell numbers and sequencing depth. The central idea was to strategically worsen the data to simulate real-world experimental constraints. By computationally reducing either the number of cells in the dataset or the number of reads per cell, we can mimic how a biological experiment would perform if fewer cells were sequenced or if each cell was sequenced more shallowly. This downsampling approach allows us to address a fundamental question: what is the relative importance of sample size versus sequencing quality?

The experiment was structured as a grid search. We systematically downsampled the full quality-controlled dataset by retaining a certain percentage of the original cells (from 10% to 100%) and, for each of those subsets, retaining a certain percentage of the original

reads (from 10% to 100%). This created a grid of 100 distinct experimental scenarios, each representing a unique combination of cell count and sequencing depth. For each condition, standard preprocessing steps were performed, a GRN was inferred, and its accuracy was evaluated against the reference network.

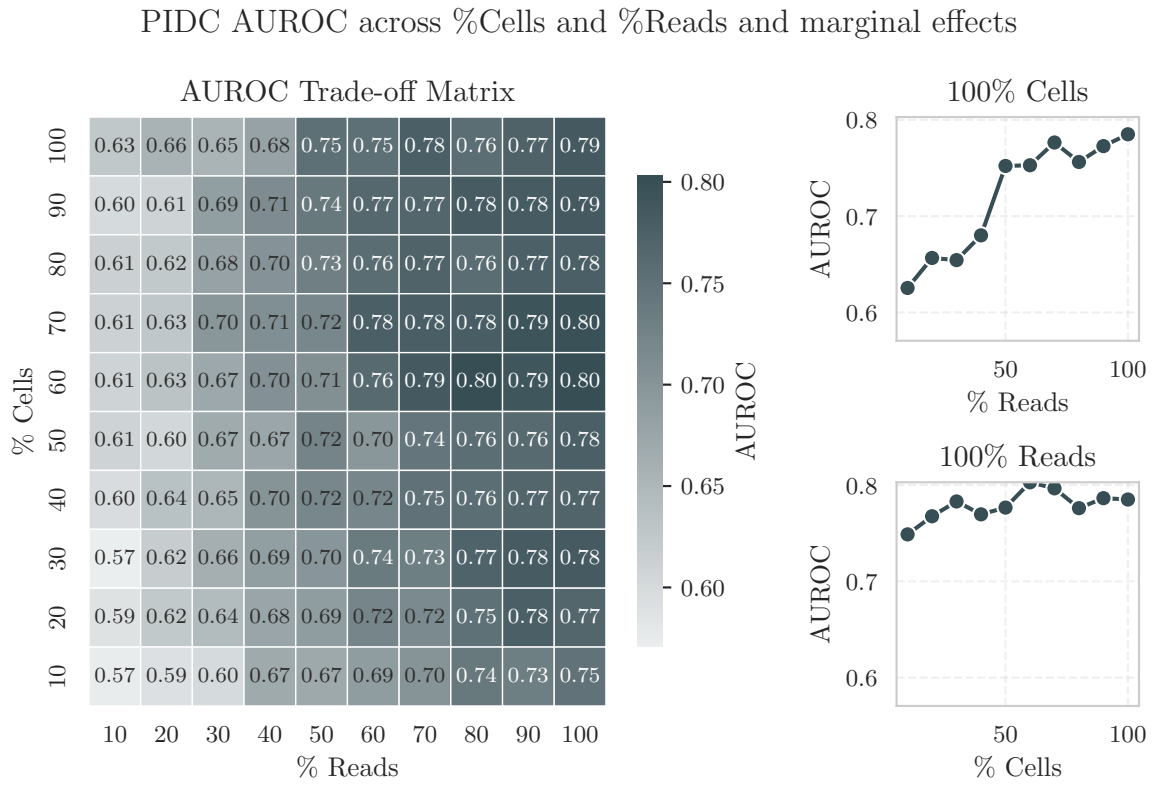

Figure 4.10: The effect of downsampling on PIDC performance. The main heatmap shows the AUROC score for each combination of cell and read percentages. AUROC values over 0.75 are marked in white. The marginal plots on the right and bottom isolate the effect of reducing reads (at 100% cells) and reducing cells (at 100% reads), respectively. The data clearly shows that performance is more resilient to a loss of cells than to a loss of sequencing depth.

The results of this experiment, visualised for the PIDC method in fig. 4.10, provide a clear and compelling answer. The heatmap of AUROC scores reveals that performance is far more sensitive to a reduction in sequencing depth (% Reads) than to a reduction in the number of cells (% Cells).

This is most evident when examining the marginal effects. The "100% Cells" plot shows a steep, almost linear decline in performance as sequencing depth is reduced while keeping 100% of the cells. The AUROC score drops from a high of 0.79 with 100% of reads to approximately 0.63 with only 10% of reads. In stark contrast, the "100% Reads" plot shows a much more resilient performance profile. When reducing the number of cells while maintaining 100% of the reads, the AUROC score remains high, only dropping from 0.79 to about 0.75, even with just 10% of the original cells.

This stark difference suggest that a smaller sample of high-quality, deeply sequenced cells is substantially more valuable for GRN inference than a large sample of low-quality, shallowly sequenced cells. For instance, an experiment with only 30% of the cells but 100% of the

reads (AUROC $\approx$ 0.78) is far superior to one with 100% of the cells but only 30% of the reads (AUROC $\approx$ 0.65). The heatmap shows that to achieve a high AUROC score (>0.75), a sequencing depth of at least 60-70% is required, whereas a high score can still be achieved with as few as 20-30% of the total cells.

Universality across datasets and algorithms

To determine whether the preference of sequencing depth is a specific trait of the PIDC algorithm or a global property of GRN inference, we extended our evaluation to include a second dataset (Buenrostro18, which characterises continuous hematopoietic differentiation) and across chosen inference methods (PIDC, GRNBoost2 and GENIE3).

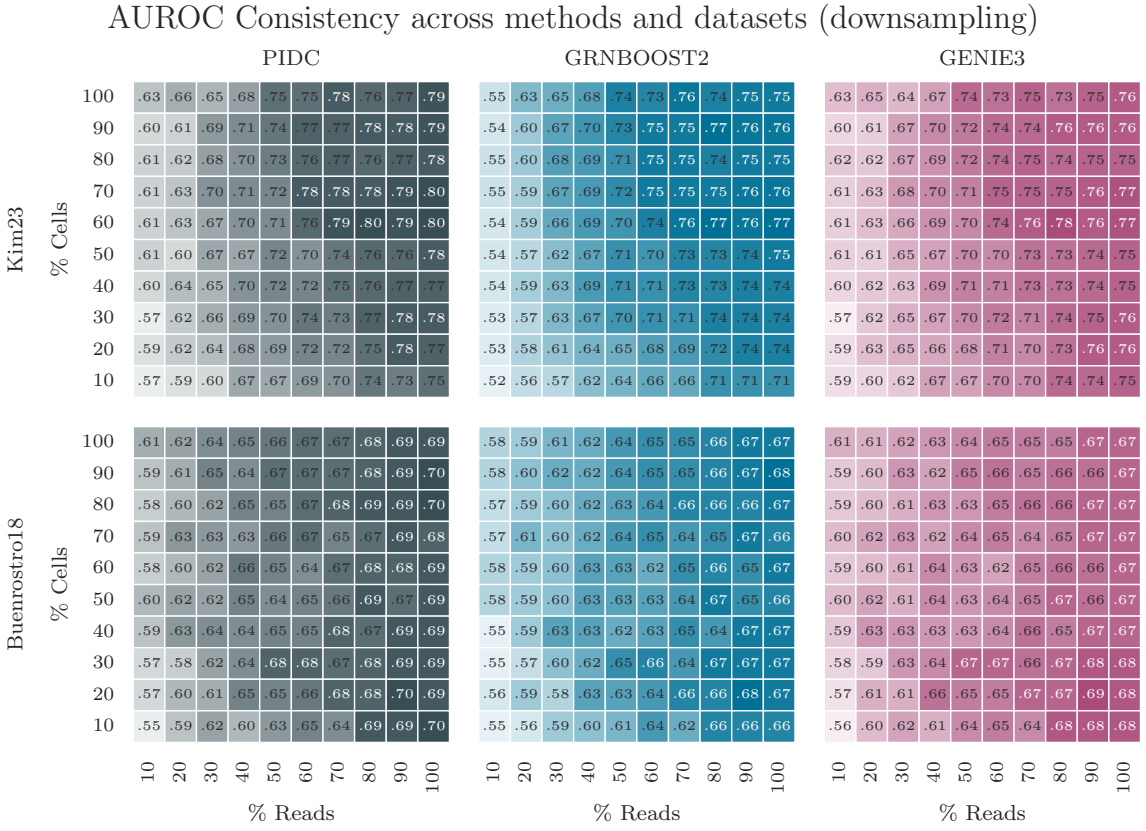

Figure 4.11: The effect of downsampling across datasets and algorithms. Columns represent different inference algorithms. The top row corresponds to Kim23 data and the bottom row corresponds to Buenrostro18 data. AUROC values within 0.02 from maximal values in each panel are marked in white. While some minor variability is visible between adjacent cells, a natural consequence of the stochastic downsampling process, a dominant and consistent trend emerges across all six panels. This trend confirms that inference performance is more robust to reductions in cell count (y-axis) and more sensitive to the loss of sequencing depth (x-axis).

The results, summarised in Figure 4.11, demonstrate that the trends observed in the initial experiment are universal. Across all six scenarios (2 datasets $\times$ 3 methods), the performance landscapes are generally similar. We consistently observe a "vertical resilience," where accuracy is maintained despite severe reductions in cell count (moving top-to-bottom), contrasted with a "horizontal collapse," where accuracy degrades rapidly with the loss of read depth (mov-

ing right-to-left). Notably, this pattern holds regardless of the underlying framework of the inference method. While PIDC (information-theoretic) and GENIE3/GRNBoost2 (random forest-based) utilise fundamentally different statistical approaches to detect gene associations, they are all critically dependent on the quality of the expression profile rather than the quantity of profiles observed. This backs the hypothesis that the "depth over cells" principle is not an artifact of a specific tool or biological context, but rather a general constraint of scRNA-seq-based regulatory inference.

Deeper profiles reduce stochastic transcript dropout and better capture lowly expressed regulators, stabilizing statistical dependencies. Adding many shallow profiles instead amplifies sparsity (more zeros per gene), increases noise, and obscures the regulatory signal. It should be noted that computational downsampling is a simplification as it does not introduce new sequencing artifacts, but only reduces depth or sample size. Therefore, the observed trends may differ under real experimental conditions. However, the practical conclusion seems to be clear and consistent: sequencing depth should be prioritised over maximizing the number of cells. However, this experimental design does not account for a fixed budget. An experiment with 100% cells and 100% reads is far more "expensive" than one with 30% cells and 30% reads. This leads to the next logical inquiry: what if the total number of UMIs is fixed? Are deeply sequenced cells still better than shallow profiles spread across many cells when the total sequencing effort is constant? This question forms the basis of the following section.

4.4.2. Fixed-budget sampling strategies

An additional question that emerged concerns how the trade-off between sequencing depth and the number of cells manifests under a fixed-budget scenario, quantified as the total number of reads in the dataset. To systematically investigate this trade-off, we designed an in silico experiment based on a "fixed-budget" simulation. The core of this design is the concept of an experimental strategy, which defines how a given budget is spent. We define this strategy on a continuous axis from 0% to 100% as follows: Strategy 0% (Depth-First): this represents a strategy that prioritises sequencing depth above all else. For a given budget, we use the absolute minimum number of cells theoretically required to achieve that budget, thereby maximizing the number of reads sequenced per cell and Strategy 100% (Cells-First): this represents the opposite extreme, prioritizing sample size. Here, we use the maximum number of cells available in our dataset, spending the budget to sequence each one as shallowly as possible while still meeting the budget constraints. By sampling points along this strategy axis, we can simulate a range of balanced approaches between these two extremes.

To understand how resource availability influences the optimal strategy, we performed these simulations across four distinct budget tiers, defined as a percentage of the total UMIs available in the full, quality-controlled dataset. We chose four thresholds: Nano - 10% of total UMIs, Low - 25%, Medium - 50% and High - 75%. For each defined budget and strategy point, the corresponding number of cells was calculated. A random sample of cells of that size was drawn from the full dataset, and the reads within this sample were then proportionally down-sampled to precisely match the target total UMI count for that budget. This process ensures that every simulation within a budget tier represents an alternative way of spending the exact same total sequencing resources, allowing for a fair and direct comparison of the resulting GRN inference accuracy. It needs to be noted, that due to random cell sampling and because reads could not be "up-sampled", the 0% Strategy tends to have the total UMI count 1-2% lower than other strategies. As before, each individual sample was then subject to standard preprocessing steps.

This experimental design allows us to directly test two primary hypotheses: first, that,

similarly to the previous experiment, down-sampling reads results in faster deterioration of the results than cell sampling and second, that the optimal balance is dependent on the total sequencing budget. The results are summarised in fig. 4.12.

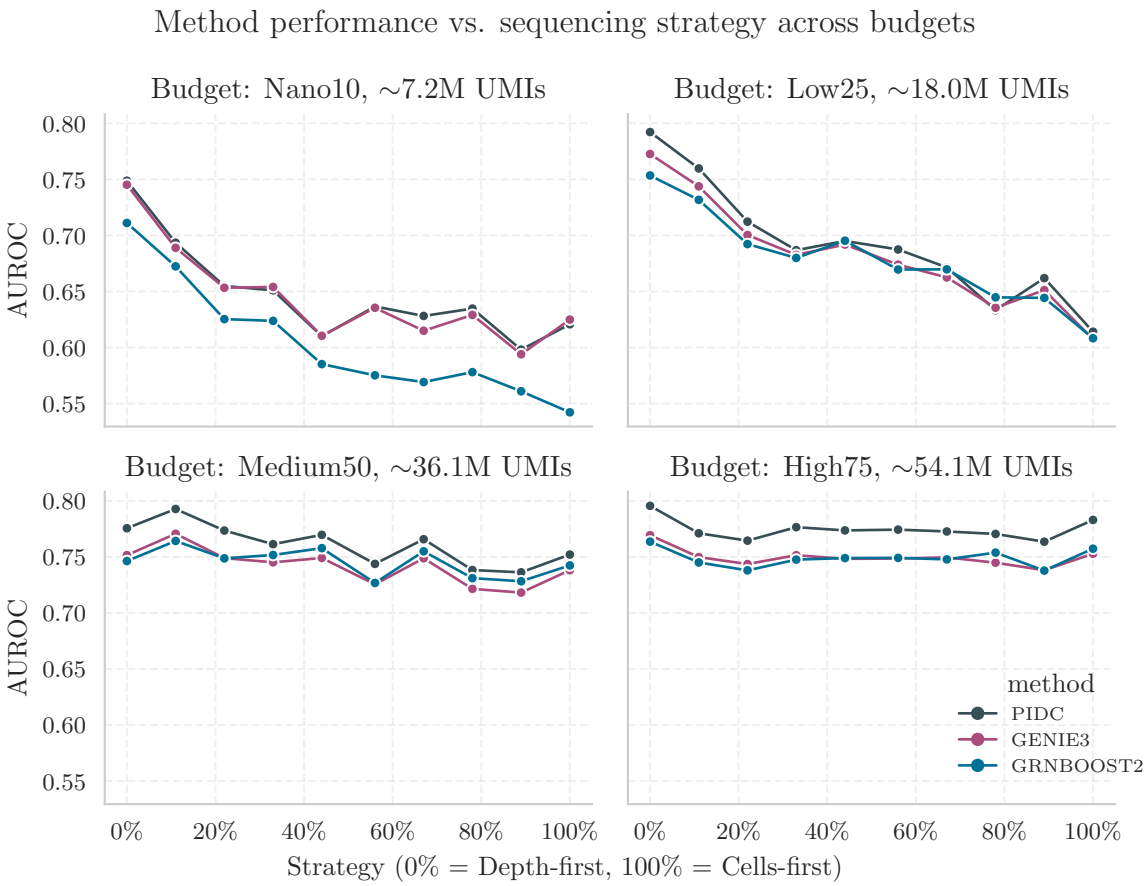

Figure 4.12: The depth - number of cells trade-off's effect on GRN inference accuracy. AUROC scores are plotted against different Strategies for three inference methods: PIDC (gray), GRNBoost2 (blue), and GENIE3 (rouge) and four budgets. The results show that the optimal allocation of resoucrces is highly dependent on the budget.

For every budget tier, the relationship between strategy and performance forms a distinct curve demonstrating that the optimal approach to budget allocation is highly conditional on the available resources. Under the "High75" budget, representing a well-resourced experiment with approximately 54 million UMIs, all methods deliver strong and relatively stable performance. The curve is fairly horizontal, peaking at the two extremes (0 and 100% Strategy), but the drop in performance is not spectacular. Also, PIDC consistently achieves the highest AUROC scores, establishing itself as the top-performing method when the budget is not a major constraint. The relative flatness of all three curves indicates that with a high budget, there is considerable flexibility in experimental design without a significant loss of accuracy. In the "Medium50" budget scenario (approximately 36 million UMIs), the trade-offs become more apparent. PIDC again shows the best performance, but by a smaller margain. However, as the strategy shifts towards prioritizing more cells (100% strategy), the performance of all methods begins to decline more noticeably than in the high-budget scenario. ALso, we see that the reults are less stable and possibly subject to random variation. This suggests that at

a medium budget, while a balanced approach is effective, prioritization of sequencing depth is the most advantageous strategy. The "Low25" budget (approximately 18 million UMIs) marks a critical turning point. Here, the performance of all methods becomes highly sensitive to the experimental strategy. The optimal approach is clearly a "Depth-First" (0%) strategy, where all three methods achieve their peak performance. Difference between methods is clear only with full-depth cells and diminishes with read subsampling. As the strategy shifts towards a "Cells-First" approach, there is a dramatic drop in performance for all methods, with AUROC scores plummeting from 0.75-0.80 to around 0.60 at the 100% strategy mark. This is a key finding: when the budget is limited, the loss of sequencing depth is far more detrimental to GRN inference accuracy than a reduction in the number of cells. The data strongly indicates that for low-budget experiments, one should prioritise sequencing a smaller number of cells as deeply as possible. In the most budget-constrained scenario, the "Nano10" (approximately 7 million UMIs), the overall performance is significantly lower. The trends observed in the low-budget setting are even more exaggerated here. The only viable strategy is a strong "Depth-First" approach. Any move towards a "Cells-First" strategy results in a catastrophic loss of performance. Notably, GRNBoost2 performance collapses in this scenario, with its AUROC dropping to nearly 0.54 at the 100% strategy mark, suggesting it is particularly ill-suited for low-depth, low-budget data. This confirms that with a minimal budget, the only effective strategy is to maximise sequencing depth.

In conclusion, the optimal strategy for allocating a sequencing budget is not fixed but is instead highly dependent on the total resources available. When the budget is high, there is a large "sweet spot" of effective experimental designs. However, as the budget becomes more constrained, the importance of sequencing depth increases dramatically. For low-resource experiments, a "Depth-First" strategy is not just optimal, it is the only viable approach to achieving meaningful results.

4.4.3. Targeting high-content cells

The grid-search experiments demonstrated that reducing sequencing depth is more detrimental than reducing cell numbers. However, real-world datasets exhibit significant heterogeneity in sequencing depth across cells and some cells are naturally sequenced more deeply than others due to technical variation or biological content. To validate whether high-depth cells are indeed the primary drivers of GRN inference accuracy, we performed an experiment comparing two selection strategies under a fixed total UMI constraint.

We defined a series of UMI budget thresholds (percentages of the total dataset). For each threshold, we constructed two subsets of data:

1. **Deepest selection:** We selected cells with the highest individual UMI counts until the cumulative UMI budget was exhausted. This results in a smaller number of high-quality cells.

2. **Random selection:** We selected cells uniformly at random until the same UMI budget was reached. This results in a larger number of cells with average sequencing depth.

The results, summarised in fig. 4.13, confirm the hypothesis that cells with higher transcript counts contribute disproportionately to inference accuracy. As shown in the comparison, for any given budget, the “Deepest selection” strategy yielded consistently higher AUROC scores compared to the “Random Selection” strategy. Notably, the Deepest strategy achieved these high scores using significantly fewer cells. For example, a subset of only 550 high-depth cells (accounting for 20% of total UMIs) yielded an AUROC of 0.753, which approaches the

baseline performance of the full dataset (0.785). In contrast, the random sampling strategy required approximately 4,500 cells (accounting for 60% of total UMIs) to achieve comparable accuracy. Furthermore, analysis of the performance trajectories reveals distinct saturation patterns: the Deepest strategy reaches a performance plateau at just 20% of the UMI budget, whereas the Random strategy exhibits a gradual incline, requiring the full dataset to maximise accuracy.

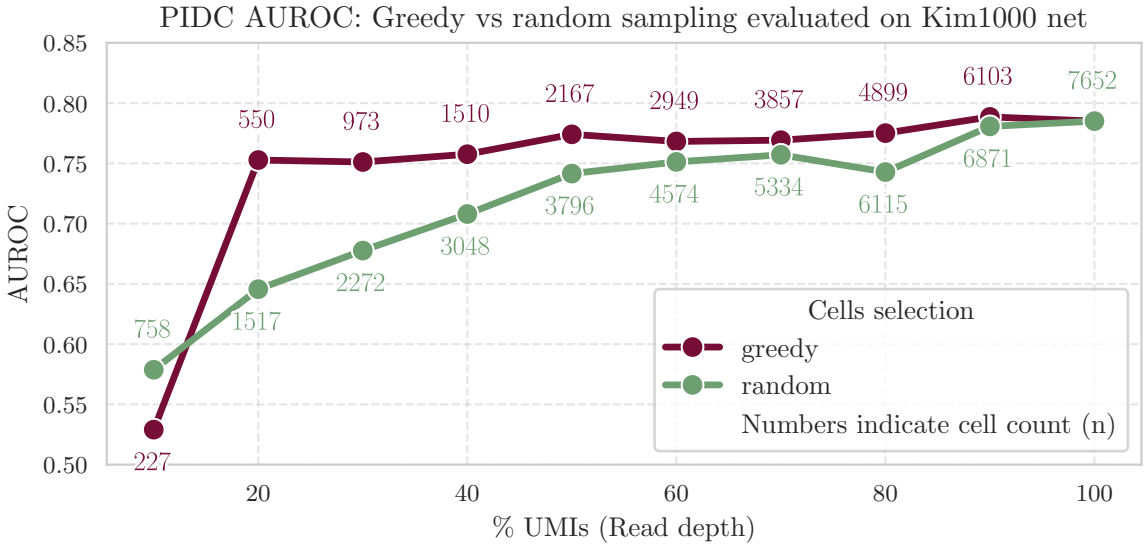

Figure 4.13: Comparison of PIDC performance (AUROC) on Kim23 using two selection strategies under fixed UMI budgets. The x-axis represents the percentage of UMIs retained, and the numbers annotating each point indicate the number of cells (n) included in that subsample. The "greedy" strategy (maroon) selects cells with the highest individual UMI counts (Deepest selection), while the "random" strategy (green) selects cells uniformly at random. The results illustrate that the greedy strategy achieves maximum performance saturation with significantly fewer cells and a lower total UMI budget than random sampling.

This empirical finding bridges the gap between our theoretical downsampling simulations and practical data properties. It suggests that the "long tail" of deeply sequenced cells in a standard scRNA-seq dataset may contribute most signal to GRN inference algorithms, with shallow cells effectively diluting the high-fidelity information present in the top-tier cells. From an experimental design perspective, these findings suggest that for GRN reconstruction, protocols that yield fewer but more deeply sequenced cells offer a superior return on investment compared to high-throughput, low-coverage strategies.

4.5. Evaluation of multi-omic strategies

While statistical methods like PIDC and GENIE3 infer networks solely from transcriptomic covariation, a separate class of algorithms attempts to anchor these predictions in biological reality using transcription factor binding motifs. We sought to evaluate whether incorporating this prior biological knowledge, ranging from static motif databases to cell-specific chromatin accessibility, improves inference accuracy against the STRING12 ground truth.

The initial phase of our evaluation focused on establishing the performance hierarchy using standard transcriptomic inputs and assessing the baseline efficacy of motif-informed methods

in non-multi-omic settings. The **Kim23** (liver organoid) and **Buenrostro18** (hematopoietic differentiation) datasets, representing distinct biological contexts, were used for this comparison. CellOracle was selected as a primary representative due to its widespread adoption and its accessible implementation distinguishing it from other multi modal inference algorithms.

As established in section 4.1, the expression-only methods (PIDC, GENIE3, and GRNBoost2) consistently define the upper performance tier, demonstrating robust predictive power across varied biological contexts (e.g AUROC values ≈ 0.785 on Kim23). The objective of the current benchmark was to determine whether incorporating genomic sequence information and chromatin accessibility data could lead to increase in evaluation metrics.

We first assessed this approach using CellOracle in its base configuration. In this mode, the algorithm constructs a "base GRN" by scanning the genome for TF binding motifs at promoter regions and then uses the scRNA-seq data to refine these connections. Next, we constructed a dataset specific base GRN using ATAC-Seq data coming from the same biological replicate as expression data. In contrast to the expression-based baselines, CellOracle demonstrated a consistent and significant performance limitations across both datasets. Results of this evaluation are summarised in fig. 4.14 and section 4.5.

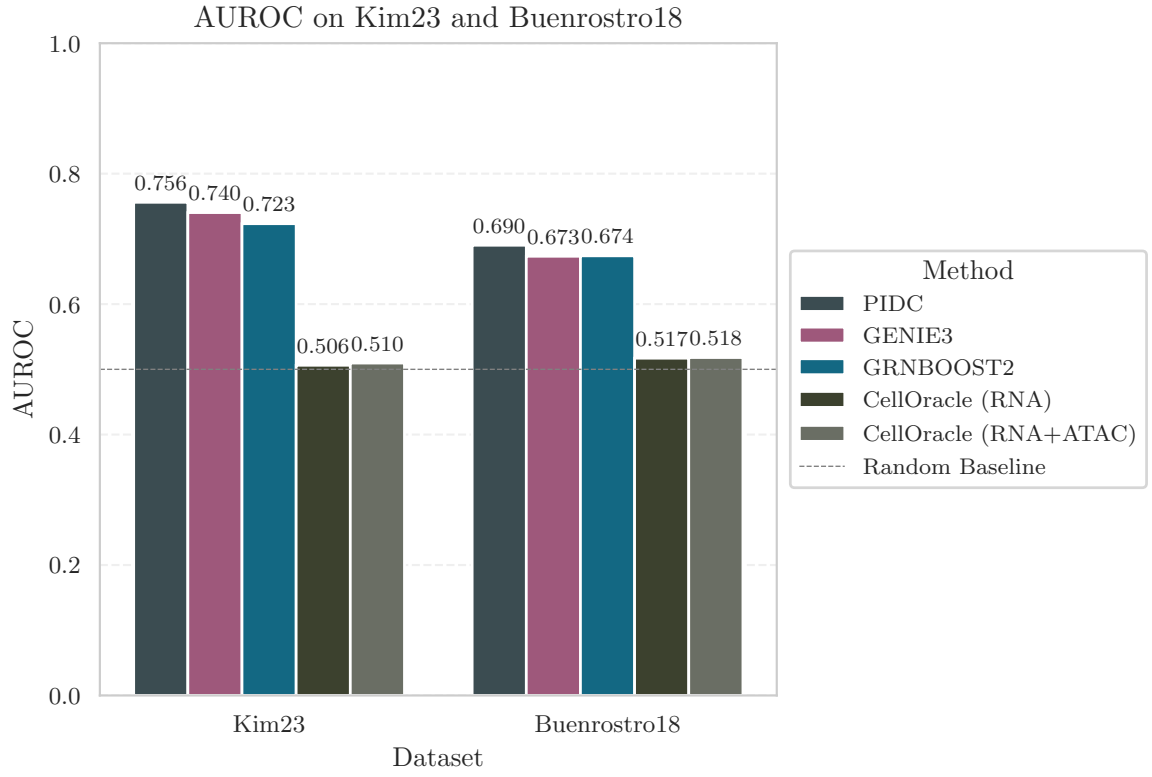

Figure 4.14: AUROC performance for GRN inference methods on the Kim23 and Buenrostro18 datasets. The expression-only baselines are maintained, while the motif-informed method (CellOracle) consistently performs near the random baseline.

When assessing performance using AUROC, both variants of CellOracle show poor results across datasets. On Kim23, AUROC values are 0.506 for the non-paired version and 0.510 for the paired-data version, while on Buenrostro18, they are 0.517 and 0.518, respectively. These values are only slightly above random, and considerably lower than those obtained by established methods such as GENIE3, GRNBoost2, or PIDC (0.67 to 0.75).

	Buenrostro18		Kim23	
	AUPRR	EPR	AUPRR	EPR
CellOracle (RNA)	1.18	2.09	1.09	2.27
CellOracle (RNA+ATAC)	1.18	2.40	1.12	2.49
GENIE3	5.95	11.74	10.57	20.62
GRNBOOST2	5.57	10.60	8.23	16.57
PIDC	6.50	18.85	20.92	47.32

Table 4.3: AUPR Ratio and EPR for all methods across datasets.

This limited performance could possibly be explained by the characteristics of networks returned by CellOracle. The method is designed to produce small, conservative networks capturing only the strongest regulatory interactions. However, precision-oriented metrics do not substantially improve this picture. AUPR ratios and Early Precision Ratios (EPR) remain low across datasets, for instance, EPR on Kim23 is 2.27 (RNA) and 2.49 (RNA+ATAC), and on Buenrostro18, 2.09 and 2.40. These results indicate that even among the top-ranked edges, CellOracle recovers few true positives. Although the inclusion of ATAC information leads to a slight increase in performance, the improvement is marginal.

We hypothesise that while this design leads to sparse predictions that poorly overlap with the full reference networks, resulting in low AUROC and precision metrics, it might be useful for highlighting the most confident edges. In this sense, CellOracle's outputs may still potentially be valuable for focused downstream analysis of high-confidence interactions. Nonetheless, they do not perform well when evaluated against comprehensive benchmarking metrics.

Extending the evaluation to paired multiomic data

To further assess whether motif information or chromatin accessibility can enhance GRN inference accuracy, we expanded our benchmark to the **PBMC10k** dataset, which is a of paired multiome data (simultaneous profiling of scRNA-seq and scATAC-seq from the same nuclei), and represents a mixture of fully differentiated immune cell types. This dataset provides an additional biological context and allows us to evaluate newer approaches, including **SCENIC+** (and its earlier version tailored to transcriptomic input, SCENIC), and **LINGER**, methods that rely on multiomic data, alongside the previously analysed CellOracle configurations. We hypothesised that the cell-level chromatin landscapes of these cell populations might provide a more favorable setting for inference methods, representing the theoretical "golden scenario" for regulatory inference.

Consistent with our observations in the Kim23 and Buenrostro18 datasets, the AUROC values for those newer methods remain close to random. For CellOracle, AUROC on PBMC10k reaches only 0.513 (RNA-only) and 0.516 (RNA+ATAC), indicating a marginal and practically negligible improvement when incorporating chromatin accessibility. SCENIC+ and SCENIC show similarly modest AUROC values around 0.51. These results mirror the earlier trend: integrating motif priors or ATAC information does not translate into improved global ranking performance, especially when compared to expression-based baselines such as GENIE3 (AUROC = 0.63).

A similar pattern emerges in the AUPR Ratio results. CellOracle achieves only slight enrichment over random expectations (AUPRR $\approx$ 1.22–1.30), with the RNA+ATAC variant improving by approximately 6%. Other multiomic methods, such as SCENIC+, remain within

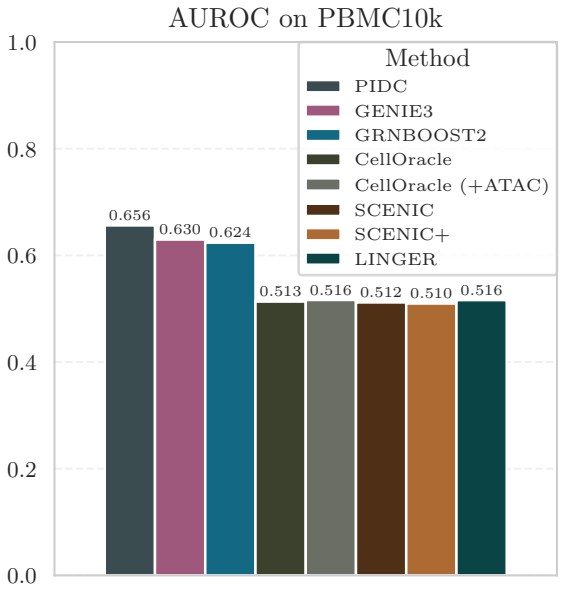

Figure 4.15: AUROC performance for GRN inference methods on the PBMC10k dataset. Near random AUROC values are observed for the multiomic methods.

dataset	PBMC10k	
	AUPRR	EPr
CellOracle	1.22	3.24
CellOracle (+ATAC)	1.30	3.62
SCENIC	1.37	5.95
SCENIC+	1.27	7.48
LINGER	1.20	3.34
GENIE3	5.61	12.11
GRNBOOST2	4.84	11.28
PIDC	5.71	13.60

Table 4.4: AUPRR and EPR on multiome PBMC10k

a similar range. These results confirm that precision-weighted metrics, which emphasise early recovery of true interactions, also fail to show meaningful gains from incorporating motif or accessibility information.

The only metric in which multiomic strategies show a clearer benefit is the Early Precision Ratio (EPR). While CellOracle again underperforms (EPR 3.24 - 3.62), SCENIC+ achieves substantially higher EPR values (5.95-7.48). Although still considerably lower than expression-based baselines (e.g. GENIE3: 12.11), this behaviour indicates that SCENIC+ is better able to prioritise a subset of high-confidence edges. This suggests that multiomic integration may offer value primarily in the top-ranked portion of the edge list, despite weak global performance.

Overall, the PBMC10k results reinforce the patterns observed in Kim23 and Buenrostro18. Motif-informed and multiomic GRN inference methods consistently exhibit: (i) near-random AUROC performance, (ii) only modest improvements in performance metrics when ATAC information is included, and (iii) more noticeable, but still limited, gains in early precision for SCENIC+.

We hypothesise that these limitations arise from both methodological and benchmarking considerations. First, methods such as CellOracle and SCENIC+ aim to identify strong motif-supported regulatory interactions and therefore return sparse, conservative networks with limited overlap to dense reference sets such as STRING. Second, motif-inferred regulatory potential does not necessarily correspond to observed protein interactions, co-expression, or genetic evidence, all of which contribute to STRING's interaction scores. This mismatch likely suppresses the global accuracy of motif-based GRNs when evaluated against a non-regulatory benchmark.

In summary, while multiomic and motif-informed strategies may still hold promise for identifying a small subset of biologically plausible high-confidence edges, they do not achieve

competitive performance in comprehensive benchmarks. Their greatest utility may lie in downstream, mechanism-oriented analyses rather than in recovering global GRN structure.

Ground-truth bias Despite the sophistication of modern multi-omic architectures and the integration of chromatin accessibility data, performance metrics remain stagnantly low. This counter-intuitive result may not strictly reflect inference quality. Instead, we hypothesise that it points to a fundamental misalignment between regulatory potential and the available benchmarks, a ground truth bias. Multimodal and multi-omic methods (CellOracle, SCENIC+) are designed as aggressive filters. They enhance specificity by enforcing the constraint that an interaction must be physically plausible (i.e., TF motif must be in an accessible chromatin region). However, this hard constraint leads to a drastic loss of sensitivity when evaluated against the STRING v12 ground truth.

The STRING database aggregates functional associations, including indirect regulation, cofactor complexes, and downstream signalling effects that do not require a direct TF-to-Promoter binding event. By enforcing strict physical constraints, multi-omic methods likely remove valid functional connections. In effect, the models are penalised by the benchmark for "correctly" removing physically implausible edges that are nonetheless recorded as functional in the gold standard. In contrast, methods like PIDC and GENIE3, which rely on non-linear statistical dependencies, are agnostic to the physical mechanism. This flexibility allows them to capture the broader "functional logic" of the gene regulatory network, resulting in significantly higher overlap with the ground truth. However, these benchmarks (that may also be based on different databases like DoRothEA or TRRUST), represent the standard adopted in current GRN inference research. We therefore adhered to this established framework despite its flaws, because a comprehensive benchmark requires a universal baseline and the STRING database is widely used for this purpose. Nonetheless, we have to keep limitations of such approach in mind.

In conclusion, our benchmark demonstrates that, in the context of recovering functionally defined regulatory networks, the increased complexity and constraint provided by multi-omic integration do not yield a corresponding improvement in predictive power. For this problem domain, expression-based statistical methods remain the most effective approach.

Chapter 5

Discussion and conclusions

This study systematically evaluated how preprocessing decisions, sequencing budget allocation, metacell aggregation, and algorithmic choices influence the accuracy of GRN inference from single-cell RNA sequencing data. Across multiple datasets and methodological settings, several clear patterns emerged. Here, we aim to interpret these findings in the context of existing literature, outline their implications for experimental design, and discuss limitations and opportunities for future work.

Only a subset of GRN inference methods provide reliable performance. Across all datasets, GENIE3, GRNBoost2, and PIDC were the only algorithms that consistently produced accurate GRNs. This aligns with prior benchmarks showing that tree-based and information-theoretic approaches are more robust to noise and sparsity compared to other models in real-world datasets. In particular, PIDC's strong performance likely reflects its suitability for capturing non-linear co-variation in high-dimensional datasets. Conversely, methods such as SCODE or SINCERITIES rely on assumptions that are fragile under dropout-heavy scRNA-seq data, explaining their poor performance. Notably, we only included real scRNA data. On simulated datasets, that are used in benchmarking studies, those methods that failed to prove effective in our work are better at recovering the underlying networks [51]. Those networks are known beforehand and serve as ground truth, approach significantly different than comparing to known interaction databases.

Quality control decisions strongly affect downstream GRN inference. QC filtering has long been recognised as essential for clustering and differential expression, but its importance for GRN inference is less well studied. Our results show that removing low-quality cells improves network accuracy, while removing high-gene-count cells has inconsistent effects. The failure of the standard 5% mitochondrial threshold on the kim23 dataset demonstrates that QC practices cannot be universally applied: mitochondrial proportions depend on tissue type, sequencing protocol, and cellular physiology. These findings emphasise that GRN-focused preprocessing requires dataset-specific tuning.

Larger HVG sets degrade global GRN accuracy. We observed that increasing the number of Highly Variable Genes (HVGs) consistently decreased AUROC values. This likely occurs because large HVG sets introduce genes with lower variance or weak regulatory coupling, which dilute the regulatory signal. GRN inference relies on capturing coordinated variation; genes with minimal variance contribute little signal while adding noise to statistical associations. Our findings support the use of a compact, high-variance gene subset when the goal is global GRN reconstruction.

However, we noted that the drop in accuracy is less pronounced when evaluating smaller subnetworks. This observation points to a fundamental limitation not only of this work but

of the GRN inference field at large: **the "Ground Truth Bias."** No universally accepted, experimentally validated human GRN exists. As acknowledged by the authors of the BEELINE framework, which remains the most comprehensive benchmarking study to date, among others [51], [53], relying on interaction databases like STRING imposes constraints that may not reflect cell-type-specific regulation. This issue was also highlighted in a recent study published during the preparation of this thesis [52]. Another evaluation that may seem to overcome some limitations of our approach may be perturbation analysis. While perturbation studies could offer an alternative ground truth, e.g. form knock-out experiments, such data remains unavailable for many biological contexts. We utilised the BEELINE framework with STRING12 fully aware of these limitations, identifying global network prediction as just one aspect of GRN inference. A key question for future research is whether these trends would hold if validated against causal perturbation data rather than static databases.

GRNs are highly dataset-specific. We observed minimal overlap between networks inferred from biologically similar datasets (e.g., two hepatocyte datasets). This reflects both biological and technical factors: GRNs differ across experimental systems, sequencing platforms, differentiation states, and donor backgrounds. Technical variability such as library preparation or depth can also distort co-expression structure. This result confirms that GRN inference lacks strong cross-dataset generalisability and must be interpreted within the context of individual experiments.

Metacells aggregation is beneficial only under specific conditions. Pseudobulking is a strategy widely adopted in GRN inference methods. It can reduce noise and grossly decrease computational cost of downstream analyses. Nonetheless, clear benefit in terms of accuracy is yet to be determined. In our work, metacells improved GENIE3 performance for some datasets with continuous trajectories (Buenrostro18, Kim23) but reduced performance for discrete populations such as PBMCs. Possibly, trajectory datasets benefit from averaging because it reduces stochastic noise without collapsing distinct states. In contrast, for discrete datasets, metacells can obscure meaningful differences between cell types. Moreover, the impact on accuracy is highly method-specific, PIDC's collapse under metacells being a prominent example that likely arises because aggregation alters the marginal distributions that PIDC depends on. Our results underline that metacells are not universally advantageous and should be applied cautiously.

Multiomic complexity fails to translate into improved accuracy. A striking finding of our study is that multiomic methods (CellOracle, SCENIC+), despite leveraging chromatin accessibility data and motif priors, failed to outperform expression-only baselines when evaluated against the STRING reference network. Moreover, the addition of ATAC-seq data yielded negligible improvement in AUROC scores, with performance often stagnating near the random baseline. While it can be argued that functional databases like STRING penalise motif-based methods by including indirect associations, we suggest that this result reflects a genuine limitation of current multiomic inference strategies rather than a bias in the benchmark.

Gene regulatory networks are fundamentally functional entities: they describe the flow of information that determines cellular identity. The STRING database, by aggregating diverse lines of evidence, provides a robust approximation of this functional logic. The superior performance of expression-based methods, suggests that statistical learning on the transcriptome is currently the most effective way to reconstruct these functional dependencies. These algorithms capture the *outcomes* of regulation, the coordinated changes in gene expression, without being constrained by the static and often over-conservative definitions of binding motifs. By strictly filtering for physical evidence in open chromatin, multiomic methods likely discard valid functional interactions that occur via indirect mechanisms or distal enhancers

not linked by simple proximity rules.

This strict physical filtering is further compromised by a fundamental temporal discordance between the two modalities. Chromatin accessibility is a highly dynamic process; mammalian genes are typically expressed in short, minute-scale bursts [63] during which chromatin is transiently accessible. An ATAC-seq experiment, which represents a single snapshot of the cellular state, is statistically prone to missing these transient regulatory events. In contrast, mRNA has a median half-life of approximately 10 hours [64], effectively acting as a temporal integrator that records the history of transcriptional activity long after the chromatin has closed. Consequently, the reliance on ATAC-seq snapshots to validate regulatory potential may inherently limit sensitivity, as the "physical" door does not need to be open at the exact moment of sequencing for the functional message to persist in the cell.

This interpretation is strongly supported by external evidence from the very recent geneRNIB study [52]. Unlike our work, which relied on database interactions, geneRNIB benchmarked methods using perturbation data, widely considered the gold standard for *causal* regulatory inference. Three methods form our work, namely GRNBoost2, SCENIC+ and CellORacle were included in the geneRNIB framework and in the final classification GRNBoost2 proved to be the most accurate method in terms of performance metrics and also the most effective computionaly. The fact that GRNBoost2 wins across two distinct benchmarking paradigms: our functional evaluation using STRING and geneRNIB's causal evaluation, validates our findings. It confirms that transcriptomic signal alone can provide a more accurate reconstruction of global gene regulatory networks than complex multiomic integration. Summing up, a core result of our work is that when allocating budget, including other sequencing modalities in the experimental design probably would not be beneficial in terms of GRN accuracy.

Sequencing depth drives inference accuracy more than cell count. Another core result of this work is the asymmetric impact of sequencing resources on GRN inference. As observed in the grid-search experiments (fig. 4.10 and fig. 4.11), AUROC scores exhibited "vertical resilience" but "horizontal collapse." Specifically, inference accuracy remained relatively stable even when the number of cells was reduced by 70-80%, provided that sequencing depth was maintained. In contrast, reducing sequencing depth while maintaining the full complement of cells resulted in a rapid and monotonic degradation of performance across all tested algorithms.

This trend was further confirmed in the fixed-budget simulations (fig. 4.12). In scenarios where sequencing resources were scarce, the "Depth-First" allocation strategy consistently yielded significantly higher AUROC scores than the "Cells-First" strategy. In a well-resourced experiment the trade-off was not so obvious, suggesting we may reach a plateu where we get all the information and further increasing the numer of cells or reads may be redundant.

These results suggest that for the specific task of gene regulatory network inference, the *quality* of the expression profile (reduced sparsity and accurate transcript counts) is more critical than the *quantity* of observations (number of cells). Statistical inference methods, whether based on mutual information or variance reduction, rely on detecting covariation between regulators and targets. Shallow sequencing exacerbates technical dropout, which effectively breaks these statistical dependencies. Our findings indicate that adding more shallowly sequenced cells fails to compensate for this loss of signal, as it essentially increases the sample size of noise rather than recovering the underlying regulatory correlations. This was empirically validated by our targeted selection analysis, which showed that a small subset of the deepest-sequenced cells in the dataset consistently outperformed larger, randomly selected populations of equivalent total sequencing cost. These findings reinforce that deeper sequencing delivers more regulatory information per unit cost than sampling additional shallow cells.

Practically, this means that experiments focusing on GRN reconstruction should allocate as much sequencing depth per cell as feasible.

In summary, this study contributes to the ongoing effort to optimise gene regulatory network inference from single-cell data. Our findings indicate that experimental design choices, particularly the prioritization of sequencing depth, may have a more pronounced impact on inference accuracy than simply increasing sample size. Additionally, the limited performance gains observed with current multi-omic integration strategies highlight the complexity of bridging physical chromatin data with functional regulatory logic, as well as the challenges inherent in standard benchmarking frameworks. We hope these observations offer practical reference points for allocating sequencing budgets and stimulate further discussion on how best to evaluate computational predictions in systems biology.

Bibliography

[1] B. Alberts, *Molecular biology of the cell*, 7th ed. New York: W. W. Norton & Company, 2022, ISBN: 978-0-393-88482-1.

[2] F. Crick, "On protein synthesis," eng, *Symposia of the Society for Experimental Biology*, vol. 12, pp. 138–163, 1958, ISSN: 0081-1386.

[3] P. J. Kennelly, K. M. Botham, O. P. McGuinness, V. W. Rodwell, and P. A. Weil, Eds., *Harper's illustrated biochemistry*, eng, 32nd ed. New York: Mcgraw Hill, 2023, ISBN: 978-1-260-46995-0.

[4] D. S. Latchman, *Gene control*, eng, 2nd ed. New York: Garland science, 2015, ISBN: 978-0-8153-4503-9.

[5] K. Maeshima, S. Ide, and M. Babokhov, "Dynamic chromatin organization without the 30-nm fiber," en, *Current Opinion in Cell Biology*, vol. 58, pp. 95–104, Jun. 2019, ISSN: 09550674. DOI: 10.1016/j.ceb.2019.02.003.

[6] E. E. M. Furlong and M. Levine, "Developmental enhancers and chromosome topology," en, *Science*, vol. 361, no. 6409, pp. 1341–1345, Sep. 2018, ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.aau0320.

[7] G. Fudenberg, M. Imakaev, C. Lu, A. Goloborodko, N. Abdennur, and L. A. Mirny, "Formation of Chromosomal Domains by Loop Extrusion," en, *Cell Reports*, vol. 15, no. 9, pp. 2038–2049, May 2016, ISSN: 22111247. DOI: 10.1016/j.celrep.2016.04.085.

[8] S. S. Rao, S.-C. Huang, B. Glenn St Hilaire, *et al.*, "Cohesin Loss Eliminates All Loop Domains," en, *Cell*, vol. 171, no. 2, 305–320.e24, Oct. 2017, ISSN: 00928674. DOI: 10.1016/j.cell.2017.09.026.

[9] National Human Genome Research Institute, *Chromatin [Illustration]*, Available at: https://www.genome.gov/genetics-glossary/Chromatin.

[10] A. Williams, N. Harker, E. Ktistaki, *et al.*, "Position effect variegation and imprinting of transgenes in lymphocytes," en, *Nucleic Acids Research*, vol. 36, no. 7, pp. 2320–2329, Feb. 2008, ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gkn085.

[11] X. Wang and F. Yue, "Hijacked enhancer–promoter and silencer–promoter loops in cancer," en, *Current Opinion in Genetics & Development*, vol. 86, p. 102 199, Jun. 2024, ISSN: 0959437X. DOI: 10.1016/j.gde.2024.102199.

[12] D. G. Lupiáñez, K. Kraft, V. Heinrich, *et al.*, "Disruptions of Topological Chromatin Domains Cause Pathogenic Rewiring of Gene-Enhancer Interactions," en, *Cell*, vol. 161, no. 5, pp. 1012–1025, May 2015, ISSN: 00928674. DOI: 10.1016/j.cell.2015.04.004.

[13] C. Carlberg, *Gene Regulation and Epigenetics: How Science Works*, eng, 1st ed. Cham: Springer Nature Switzerland, 2024, ISBN: 978-3-031-68729-7. DOI: 10.1007/978-3-031-68730-3.

[14] J. Zlatanova and K. E. Van Holde, *Molecular biology: structure and dynamics of genomes and proteomes*, eng, 2nd ed. Boca Raton, FL Abingdon, Oxon: CRC Press, Taylor and Francis Group, 2023, ISBN: 978-0-367-67408-3.

[15] V. X. Jin, G. A. Singer, F. J. Agosto-Pérez, S. Liyanarachchi, and R. V. Davuluri, "Genome-wide analysis of core promoter elements from conserved human and mouse orthologous pairs," en, *BMC Bioinformatics*, vol. 7, no. 1, p. 114, Dec. 2006, ISSN: 1471-2105. DOI: 10.1186/1471-2105-7-114.

[16] J. Langerman, D. Lopez, M. Pellegrini, and S. T. Smale, "Species-Specific Relationships between DNA and Chromatin Properties of CpG Islands in Embryonic Stem Cells and Differentiated Cells," en, *Stem Cell Reports*, vol. 16, no. 4, pp. 899–912, Apr. 2021, ISSN: 22136711. DOI: 10.1016/j.stemcr.2021.02.016.

[17] X. Liu, W. L. Kraus, and X. Bai, "Ready, pause, go: Regulation of RNA polymerase II pausing and release by cellular signaling pathways," en, *Trends in Biochemical Sciences*, vol. 40, no. 9, pp. 516–525, Sep. 2015, ISSN: 09680004. DOI: 10.1016/j.tibs.2015.07.003.

[18] S. Kim and J. Wysocka, "Deciphering the multi-scale, quantitative cis-regulatory code," en, *Molecular Cell*, vol. 83, no. 3, pp. 373–392, Feb. 2023, ISSN: 10972765. DOI: 10.1016/j.molcel.2022.12.032.

[19] D. Winogradoff and A. Aksimentiev, "Molecular Mechanism of Spontaneous Nucleosome Unraveling," en, *Journal of Molecular Biology*, vol. 431, no. 2, pp. 323–335, Jan. 2019, ISSN: 00222836. DOI: 10.1016/j.jmb.2018.11.013.

[20] A. Barral and K. S. Zaret, "Pioneer factors: Roles and their regulation in development," en, *Trends in Genetics*, vol. 40, no. 2, pp. 134–148, Feb. 2024, ISSN: 01689525. DOI: 10.1016/j.tig.2023.10.007.

[21] K. Takahashi and S. Yamanaka, "Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors," en, *Cell*, vol. 126, no. 4, pp. 663–676, Aug. 2006, ISSN: 00928674. DOI: 10.1016/j.cell.2006.07.024.

[22] C. Carlberg and F. Molnár, *Mechanisms of Gene Regulation: How Science Works*, en. Cham: Springer International Publishing, 2020, ISBN: 978-3-030-52320-6. DOI: 10.1007/978-3-030-52321-3.

[23] J. Erceg, T. Pakozdi, R. Marco-Ferreres, *et al.*, "Dual functionality of *cis* -regulatory elements as developmental enhancers and Polycomb response elements," en, *Genes & Development*, vol. 31, no. 6, pp. 590–602, Mar. 2017, ISSN: 0890-9369, 1549-5477. DOI: 10.1101/gad.292870.116.

[24] G. Bower, E. W. Hollingsworth, S. H. Jacinto, *et al.*, "Range extender mediates long-distance enhancer activity," en, *Nature*, vol. 643, no. 8072, pp. 830–838, Jul. 2025, ISSN: 0028-0836, 1476-4687. DOI: 10.1038/s41586-025-09221-6.

[25] T. Pollex, A. Rabinowitz, M. C. Gambetta, *et al.*, "Enhancer–promoter interactions become more instructive in the transition from cell-fate specification to tissue differentiation," en, *Nature Genetics*, vol. 56, no. 4, pp. 686–696, Apr. 2024, ISSN: 1061-4036, 1546-1718. DOI: 10.1038/s41588-024-01678-x.

[26] J. A. Segert, S. S. Gisselbrecht, and M. L. Bulyk, "Transcriptional Silencers: Driving Gene Expression with the Brakes On," en, *Trends in Genetics*, vol. 37, no. 6, pp. 514–527, Jun. 2021, ISSN: 01689525. DOI: 10.1016/j.tig.2021.02.002.

[27] A. Claringbould and J. B. Zaugg, "Enhancers in disease: Molecular basis and emerging treatment strategies," en, *Trends in Molecular Medicine*, vol. 27, no. 11, pp. 1060–1073, Nov. 2021, ISSN: 14714914. DOI: 10.1016/j.molmed.2021.07.012.

[28] J.-J. M. Riethoven, "Regulatory Regions in DNA: Promoters, Enhancers, Silencers, and Insulators," in *Computational Biology of Transcription Factor Binding*, I. Ladunga, Ed., vol. 674, Series Title: Methods in Molecular Biology, Totowa, NJ: Humana Press, 2010, pp. 33–42, ISBN: 978-1-60761-853-9. DOI: 10.1007/978-1-60761-854-6_3.

[29] M. L. Negri, S. D'Annunzio, G. Vitali, and A. Zippo, "May the force be with you: Nuclear condensates function beyond transcription control: Potential nongenetic functions of nuclear condensates in physiological and pathological conditions," en, *BioEssays*, vol. 45, no. 10, p. 2 300 075, Oct. 2023, ISSN: 0265-9247, 1521-1878. DOI: 10.1002/bies.202300075.

[30] G. Stampfel, T. Kazmar, O. Frank, S. Wienerroither, F. Reiter, and A. Stark, "Transcriptional regulators form diverse groups with context-dependent regulatory functions," en, *Nature*, vol. 528, no. 7580, pp. 147–151, Dec. 2015, ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature15545.

[31] A. Raj and A. Van Oudenaarden, "Nature, Nurture, or Chance: Stochastic Gene Expression and Its Consequences," en, *Cell*, vol. 135, no. 2, pp. 216–226, Oct. 2008, ISSN: 00928674. DOI: 10.1016/j.cell.2008.09.050.

[32] H. Bolouri, *Computational modeling of gene regulatory networks : a primer*. London: Imperial College Press, 2008, ISBN: 978-1-84816-220-4.

[33] F. Conte, G. Fiscon, V. Licursi, *et al.*, "A paradigm shift in medicine: A comprehensive review of network-based approaches," en, *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, vol. 1863, no. 6, p. 194 416, Jun. 2020, ISSN: 18749399. DOI: 10.1016/j.bbagrm.2019.194416.

[34] D. Mercatelli, L. Scalambra, L. Triboli, F. Ray, and F. M. Giorgi, "Gene regulatory network inference resources: A practical overview," en, *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, vol. 1863, no. 6, p. 194 430, Jun. 2020, ISSN: 18749399. DOI: 10.1016/j.bbagrm.2019.194430.

[35] P. Badia-i-Mompel, L. Wessels, S. Müller-Dott, *et al.*, "Gene regulatory network inference in the era of single-cell multi-omics," en, *Nature Reviews Genetics*, vol. 24, no. 11, pp. 739–754, Nov. 2023, ISSN: 1471-0056, 1471-0064. DOI: 10.1038/s41576-023-00618-5.

[36] F. Jacob and J. Monod, "Genetic regulatory mechanisms in the synthesis of proteins," en, *Journal of Molecular Biology*, vol. 3, no. 3, pp. 318–356, Jun. 1961, ISSN: 00222836. DOI: 10.1016/S0022-2836(61)80072-7.

[37] D. Marbach, J. C. Costello, R. Küffner, *et al.*, "Wisdom of crowds for robust gene network inference," *Nature Methods*, vol. 9, no. 8, pp. 796–804, Aug. 2012, ISSN: 1548-7105. DOI: 10.1038/nmeth.2016.

[38] D. Kim, A. Tran, H. J. Kim, Y. Lin, J. Y. H. Yang, and P. Yang, "Gene regulatory network reconstruction: Harnessing the power of single-cell multi-omic data," en, *npj Systems Biology and Applications*, vol. 9, no. 1, pp. 1–13, Oct. 2023, Publisher: Nature Publishing Group, ISSN: 2056-7189. DOI: 10.1038/s41540-023-00312-6.

[39] V. A. Huynh-Thu and G. Sanguinetti, "Gene Regulatory Network Inference: An Introductory Survey," in *Gene Regulatory Networks*, G. Sanguinetti and V. A. Huynh-Thu, Eds., vol. 1883, Series Title: Methods in Molecular Biology, New York, NY: Springer New York, 2019, pp. 1–23, ISBN: 978-1-4939-8881-5. DOI: 10.1007/978-1-4939-8882-2_1.

[40] J. Dong, J. Li, and F. Wang, "Deep Learning in Gene Regulatory Network Inference: A Survey," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 21, no. 6, pp. 2089–2101, Nov. 2024, ISSN: 1545-5963, 1557-9964, 2374-0043. DOI: 10.1109/TCBB.2024.3442536.

[41] V. A. Huynh-Thu, A. Irrthum, L. Wehenkel, and P. Geurts, "Inferring Regulatory Networks from Expression Data Using Tree-Based Methods," en, *PLoS ONE*, vol. 5, no. 9, M. Isalan, Ed., e12776, Sep. 2010, ISSN: 1932-6203. DOI: 10.1371/journal.pone.0012776.

[42] D. Marbach, R. J. Prill, T. Schaffter, C. Mattiussi, D. Floreano, and G. Stolovitzky, "Revealing strengths and weaknesses of methods for gene network inference," en, *Proceedings of the National Academy of Sciences*, vol. 107, no. 14, pp. 6286–6291, Apr. 2010, ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.0913357107.

[43] T. Moerman, S. Aibar Santos, C. Bravo González-Blas, *et al.*, "GRNBoost2 and Arboreto: Efficient and scalable inference of gene regulatory networks," en, *Bioinformatics*, vol. 35, no. 12, J. Kelso, Ed., pp. 2159–2161, Jun. 2019, ISSN: 1367-4803, 1367-4811. DOI: 10.1093/bioinformatics/bty916.

[44] J. H. Friedman, "Greedy function approximation: A gradient boosting machine.," *The Annals of Statistics*, vol. 29, no. 5, Oct. 2001, ISSN: 0090-5364. DOI: 10.1214/aos/1013203451.

[45] T. E. Chan, M. P. Stumpf, and A. C. Babtie, "Gene Regulatory Network Inference from Single-Cell Data Using Multivariate Information Measures," en, *Cell Systems*, vol. 5, no. 3, 251–267.e3, Sep. 2017, ISSN: 24054712. DOI: 10.1016/j.cels.2017.08.014.

[46] J. J. Faith, B. Hayete, J. T. Thaden, *et al.*, "Large-Scale Mapping and Validation of Escherichia coli Transcriptional Regulation from a Compendium of Expression Profiles," en, *PLoS Biology*, vol. 5, no. 1, A. Levchenko, Ed., e8, Jan. 2007, ISSN: 1545-7885. DOI: 10.1371/journal.pbio.0050008.

[47] J. G. Camp, K. Sekine, T. Gerber, *et al.*, "Multilineage communication regulates human liver bud development from pluripotency," en, *Nature*, vol. 546, no. 7659, pp. 533–538, Jun. 2017, ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature22796.

[48] J.-H. Kim, S. J. Mun, J.-H. Kim, M. J. Son, and S.-Y. Kim, "Integrative analysis of single-cell RNA-seq and ATAC-seq reveals heterogeneity of induced pluripotent stem cell-derived hepatic organoids," en, *iScience*, vol. 26, no. 9, p. 107 675, Sep. 2023, ISSN: 25890042. DOI: 10.1016/j.isci.2023.107675.

[49] J. D. Buenrostro, M. R. Corces, C. A. Lareau, *et al.*, "Integrated Single-Cell Analysis Maps the Continuous Regulatory Landscape of Human Hematopoietic Differentiation," en, *Cell*, vol. 173, no. 6, 1535–1548.e16, May 2018, ISSN: 00928674. DOI: 10.1016/j.cell.2018.03.074.

[50] 10x Genomics, *PBMC from a Healthy Donor (Granulocytes Removed) - Single Cell Multiome ATAC + Gene Expression Dataset, Cell Ranger ARC v2.0.0*, Available at https://www.10xgenomics.com/resources/datasets, 2019.

[51] A. Pratapa, A. P. Jalihal, J. N. Law, A. Bharadwaj, and T. M. Murali, "Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data," en, *Nature Methods*, vol. 17, no. 2, pp. 147–154, Feb. 2020, ISSN: 1548-7091, 1548-7105. DOI: 10.1038/s41592-019-0690-6.

[52] J. Nourisa, A. Passemiers, M. Stock, *et al.*, *geneRNIB: A living benchmark for gene regulatory network inference*, en, Mar. 2025. DOI: 10.1101/2025.02.25.640181.

[53] M. Saint-Antoine and A. Singh, *Benchmarking Gene Regulatory Network Inference Methods on Simulated and Experimental Data*, en, May 2023. DOI: 10.1101/2023.05.12.540581.

[54] A. R. Sonawane, D. L. DeMeo, J. Quackenbush, and K. Glass, "Constructing gene regulatory networks using epigenetic data," en, *npj Systems Biology and Applications*, vol. 7, no. 1, p. 45, Dec. 2021, ISSN: 2056-7189. DOI: 10.1038/s41540-021-00208-3.

[55] D. Szklarczyk, R. Kirsch, M. Koutrouli, *et al.*, "The STRING database in 2023: Protein–protein association networks and functional enrichment analyses for any sequenced genome of interest," en, *Nucleic Acids Research*, vol. 51, no. D1, pp. D638–D646, Jan. 2023, ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gkac1000.

[56] F. A. Wolf, P. Angerer, and F. J. Theis, "SCANPY: Large-scale single-cell gene expression data analysis," en, *Genome Biology*, vol. 19, no. 1, p. 15, Dec. 2018, ISSN: 1474-760X. DOI: 10.1186/s13059-017-1382-0.

[57] S. Persad, Z.-N. Choo, C. Dien, *et al.*, "SEACells infers transcriptional and epigenomic cellular states from single-cell genomics data," en, *Nature Biotechnology*, vol. 41, no. 12, pp. 1746–1757, Dec. 2023, ISSN: 1087-0156, 1546-1696. DOI: 10.1038/s41587-023-01716-9.

[58] L. Heumos, A. C. Schaar, C. Lance, *et al.*, "Best practices for single-cell analysis across modalities," en, *Nature Reviews Genetics*, vol. 24, no. 8, pp. 550–572, Aug. 2023, ISSN: 1471-0056, 1471-0064. DOI: 10.1038/s41576-023-00586-w.

[59] A.-M. Galow, S. Kussauer, M. Wolfien, *et al.*, "Quality control in scRNA-Seq can discriminate pacemaker cells: The mtRNA bias," en, *Cellular and Molecular Life Sciences*, vol. 78, no. 19-20, pp. 6585–6592, Oct. 2021, ISSN: 1420-682X, 1420-9071. DOI: 10.1007/s00018-021-03916-5.

[60] D. Osorio and J. J. Cai, "Systematic determination of the mitochondrial proportion in human and mice tissues for single-cell RNA-sequencing data quality control," en, *Bioinformatics*, vol. 37, no. 7, A. Mathelier, Ed., pp. 963–967, May 2021, ISSN: 1367-4803, 1367-4811. DOI: 10.1093/bioinformatics/btaa751.

[61] A. Subramanian, M. Alperovich, Y. Yang, and B. Li, "Biology-inspired data-driven quality control for scientific discovery in single-cell transcriptomics," en, *Genome Biology*, vol. 23, no. 1, p. 267, Dec. 2022, ISSN: 1474-760X. DOI: 10.1186/s13059-022-02820-w.

[62] J. U. Loers and V. Vermeirssen, "A single-cell multimodal view on gene regulatory network inference from transcriptomics and chromatin accessibility data," en, *Briefings in Bioinformatics*, vol. 25, no. 5, bbae382, Jul. 2024, ISSN: 1467-5463, 1477-4054. DOI: 10.1093/bib/bbae382.

[63] D. M. Suter, N. Molina, D. Gatfield, K. Schneider, U. Schibler, and F. Naef, "Mammalian Genes Are Transcribed with Widely Different Bursting Kinetics," en, *Science*, vol. 332, no. 6028, pp. 472–474, Apr. 2011, ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1198817.

[64] E. Yang, E. Van Nimwegen, M. Zavolan, *et al.*, "Decay Rates of Human mRNAs: Correlation With Functional Characteristics and Sequence Attributes," en, *Genome Research*, vol. 13, no. 8, pp. 1863–1872, Aug. 2003, ISSN: 1088-9051. DOI: 10.1101/gr.1272403.

List of Figures

List of Tables