Nonequilibrium Physics of Single-Cell Genomics

Dissertation
zur Erlangung des akademischen Grades
Doctor rerum naturalium (Dr. rer. nat.)

vorgelegt
der Fakultät Physik
Bereich Mathemathik und Naturwissenschaften
der Technischen Universität Dresden
von

Fabrizio Olmeda
geboren am 07.06.1992 in Rom, Italien

Max-Planck-Institut für Physik komplexer Systeme

Dresden, 2021

Wissenschaftlicher Betreuer:
Dr. Steffen Rulands und Prof. Dr. Frank Jülicher

Eingereicht am 29.11.2021
Verteidigt am 11.05.2022

Gutachter:
1. Prof. Dr. Frank Jülicher
2. Prof. Dr. Marc Timme
3. Prof. Dr. Benjamin David Simons

List of Abbreviations

DNAme	DNA methylation
CpG	The base pair cytosine-guanine
mESC	Mouse embryonic stem cell
scNMT-Seq	Single-cell nucleosome, methylation and transcription sequencing
BS-Seq	Bisulphite sequencing
RNA-Seq	RNA sequencing
RG	Renormalization group
Sequence space	One dimensional lattice with CpGs as sites
Physical space	The three dimensional space of the cell nucleus
GRN	Gene regulatory network
MSRJD	Martin-Siggia-Rose-Janssen-De Dominicis
DAT	De Almeida Thouless

Abstract

The self-organisation of cells into complex tissues relies on the tight regulation of molecular processes governing their behaviour. Understanding these processes is a central questions in cell biology. In recent years, technological breakthroughs in single-cell sequencing experiments have enabled us to probe these processes with unprecedented molecular detail. However, biological function relies on collective processes on the mesoscopic and macroscopic scale, which do not necessarily obey the rules that govern it on the microscopic scale. Insights from these experiments on how collective processes determine cellular behaviour consequently remain severely limited. Methods from nonequilibrium statistical physics provide a rigorous framework to connect microscopic measurements to their mesoscopic or macroscopic consequences.

In this thesis, by combining for the first time the possibilities of single-cell technologies and tools from nonequilbrium statistical physics, we develop theoretical frameworks that overcome these conceptual limitations. In particular, we derive a theory that maps measurements along the linear sequence of the DNA to mesoscopic processes in space and time in the cell nucleus. We demonstrate this approach in the context of the establishment of chemical modifications of the DNA (DNA methylation) during early embryonic development. Drawing on sequencing experiments both *in vitro* and *in vivo*, we find that the embryonic DNA methylome is established through the interplay between DNA methylation and 30-40 nm dynamic chromatin condensates. This interplay gives rise to hallmark scaling behaviour with an exponent of 5/2 in the time evolution of embryonic DNA methylation and time dependent, scale-free connected correlation functions, both of which are predicted by our theory. Using this theory, we successfully identify regions of the DNA that carry DNA methylation patterns anticipating cellular symmetry breaking *in vivo*.

The primary layer determining cell identity is gene expression. However, read-outs of gene-expression profiling experiments are dominated by systematic technical noise and they do not provide "stochiometric" measurements that allow experimental data to be predicted by theories. Here, by developing effective spin glass methods, we show that the macroscopic propagation of fluctuations in the concentration of mRNA molecules gives direct information on the physical mechanisms governing cell states, independent of technical bias. We find that gene expression fluctuations may exhibit glassy behaviour such that they are long-lived and carry biological information. We demonstrate the biological relevance of glassy fluctuations by analysing single-cell RNA sequencing experiments of mouse neurogenesis.

Taken together, we overcome important conceptual limitations of emerging technologies in biology and pioneer the application of methods from stochastic processes, spin glasses, field and renormalization group theories to single-cell genomics.

Zusammenfassung

Die Selbstorganisation von Zellen zu komplexen Geweben beruht auf der strengen Regulierung molekularer Prozesse, welche ihr Verhalten bestimmen. Diese Prozesse zu verstehen ist eine zentrale Frage der Zellbiologie. In den letzten Jahren haben technologische Durchbrüche bei Einzelzell-Sequenzierungsexperimenten uns ermöglicht, diese Prozesse mit noch nie dagewesenen molekularen Details zu untersuchen. Biologische Funktionen beruhen jedoch auf kollektiven Prozessen auf der mesoskopischen und makroskopischen Ebene, die nicht unbedingt auf den selben Prinzipien basieren, denen sie auf der mikroskopischen Skala unterliegen. Die Erkenntnisse aus diesen Experimenten über die Bestimmung des zellulären Verhaltens durch kollektive Prozesse bleiben daher stark begrenzt. Methoden der statistischen Nichtgleichgewichtsphysik bieten einen präzisen Rahmen, um mikroskopische Messungen mit ihren mesoskopischen oder makroskopischen Konsequenzen zu verbinden. In dieser Arbeit kombinieren wir zum ersten Mal die Möglichkeiten der Einzelzelltechnologie mit den Werkzeugen der statistischen Physik in Nichtgleichgewichtssystemen und entwickeln einen theoretischen Rahmen, der diese konzeptionellen Einschränkungen überwindet. Insbesondere leiten wir eine Theorie ab, die Messungen entlang der linearen Sequenz der DNA auf mesoskopische Prozesse in Raum und Zeit im Zellkern abbildet. Wir demonstrieren diesen Ansatz im Zusammenhang mit der Etablierung chemischer Modifikationen der DNA (DNA-Methylierung) während früher Embryonalentwicklung. Anhand von Sequenzierungsexperimenten sowohl in vitro als auch in vivo, stellen wir fest, dass das embryonale DNA-Methylom durch das Zusammenspiel von DNA-Methylierung und 30-40 nm großen dynamischen Chromatinkondensaten gebildet wird. Dieses Zusammenspiel führt zu einem charakteristischen Skalierungsverhalten mit einem Exponenten von 5/2 in der zeitlichen Entwicklung der embryonalen DNA-Methylierung und zeitabhängigen, skalenfreien Korrelationsfunktionen, die beide von unserer Theorie vorhergesagt werden. Mit Hilfe dieser Theorie gelingt es uns DNA-Regionen zu identifizieren, die DNA-Methylierungsmuster tragen, welche zelluläre Symmetriebrechungen in vivo vorhersagen. Genexpression ist die primäre Ebene, die die Zellidentität bestimmt. Die Ergebnisse von Experimenten zur Erstellung von Genexpressionsprofilen werden jedoch durch systematisches technisches Rauschen dominiert und liefern keine „stochiometrischen“Messungen, welche eine Vorhersage der experimentellen Daten durch Theorien ermöglichen würden. Durch die Entwicklung effektiver Spinglass-Methoden können wir

unabhängig von technischen Verzerrungen zeigen, dass die makroskopische Ausbreitung von Fluktuationen in der Konzentration von mRNA-Molekülen direkte Informationen über die physikalischen Mechanismen liefert, die den Zelltyp festlegen. Wir stellen fest, dass Fluktuationen in der Genexpression ein glasartiges Verhalten aufweisen können, sodass sie langlebig sind und biologische Informationen enthalten. Wir demonstrieren die biologische Relevanz glasartiger Fluktuationen durch die Analyse von Einzelzell-RNA-Sequenzierungsexperimenten während der Neurogenese in Mäusen. Insgesamt überwinden wir somit wichtige konzeptionelle Beschränkungen aufkommender Technologien in der Biologie und leisten Pionierarbeit bei der Anwendung von Methoden aus den Bereichen Spin-Gläser, stochastische Prozesse, Renormierungsgruppen- und Feldtheorien auf die Einzelzellgenomik.

Acknowledgments

I have been told many times that I am not particularly good in doing compliments. This is probably a great opportunity to redeem myself and thank all the people that made this PhD journey possible.

First of all, I would like to thank my PhD supervisor, Steffen. You allowed me to work in a great environment and you always supervised my research trusting on my capacities. You gave me the freedom to explore many different fields and you left the doors of your office always open for long scientific discussion. Without you, I would have probably missed many fascinating biological and physical questions. To this end, I would like to thank all the present and past members of the Rulands lab. Thank Adolfo, Aida, Alvaro, Bahareh, Fabian, Felix, Ivan, Matteo, Misha, Yiteng and Ruslan, as you have always stimulated my curiosity with your unique expertise, that I will carry with me throughout my future career. We shared great memories and I hope we will drink together a Gluhwein as soon as possible. Thanks for teaching me how to better prepare a scientific talk and discussion. In particular, I want to to thank Yiteng, Fabian, Misha and Bashwar for their work, which has become part of the thesis. Misha you were an extremely talented and funny student and I hope to work with you again in the future. Don't worry, I will never like maccaroni massala! I thank Wolf Reik and all the members of his lab for an exciting and fruitful collaboration and for introducing me to the beautiful field of epigenetics. My PhD journey would not have begun without the help of Lorenzo, who suggested me to apply for this PhD position: you have been been a reliable friend and one of the best physicist I have ever met. I hope we can work together in the futures, so, you are booked for a paper! I would like to thank Ulrike and Anna, I hope I haven't stressed you out about my last-minute bureaucracy.

I started the PhD at the same time as Adolfo and I really thank you for all the inputs you gave me, not only from a scientific point of view but also for sharing great times. You have a deep understanding of science from a multiple perspective and always asking yourself extremely fascinating questions. You made me discover chess, one of my greatest passions and you have were an amazing teacher, but I will beat you one day, don't worry! I would like to thank Fabian, the senior in the lab that everyone would love to have. I am sure I will miss our lunch discussion at the university.

After one year in Dresden I had the luck to have the best office mate I could have asked for, Giacomo. First of all, I bothered you quite enough, so I'm sorry for that.

Thanks to you now I can stand a jazz song for more than one minute. Without irony, you were always close to me in tough personal and work situations. I will miss our vegan kebab dinners, discussing about any topic and thank for always bringing positive thinking in the office. I would like to thank Andy: you are truly an amazing person and you have been a nice shoulder in tough situations. I would like to thank Laura, Giovanni, Elisa and Francesca: our Italian pizza dinners have always been a refreshing moment in the week and thank for listening to a rather egocentric person. I would like to thank Alberto, Anthony, Eleonora, Filippo and Monica as you opened my mind to many different social topics and made me, in general, a more conscious person. I always recall with joy our out-of-tune singing dinner! Thank you Alberto as you have always brought a lot of refreshing energy. Ah, I will never define sup a sport! Anthony, I am looking forward for our next football training as they have really helped me in this last period. I know that I am a bad goalkeeper, but do not try to humiliate me, shooting from 30 meters. Thank Matteo for having read almost all the thesis. You definitely know your scientific value and I am sure you will focus on your talent. To this end, I would like to thank Steffen, Giacomo, Adolfo, Ivan, Alison, Matteo, Frank for their great comments on the thesis. Thank Charlie, as you were always there when I had silly questions on RG and field theories. I would like to thank all the aforementioned people, along with Ana, Roberto, Dora, Stefano for making my time in Dresden much nicer.

Nothing I have ever done or achieved would have been possible without my family and my long-time friends. Thank you as you saw me changing in many different ways and you have always be there to guide me during this process.

I would like to thank my parents. You have always pushed me to do everything that made me feel happy, upon relying, probably wrong, to the fact that I would be able to find the right directions in my life. You have always been supportive, but never intrusive, a quality that I really love. *So bene quanto e' stato difficile crescermi in una situazione difficile. Mi avete sempre dimostrato cosa significa essere forti nella vita e ad affrontare di petto ogni situazione. Grazie per aver sempre creduto, stimolato le mia capacita' e spronato a trovare le mie passioni, che mi hanno salvato in piu' circostanze nella mia vita. Mi avete sempre fatto incuriosire del mondo che mi circondava e senza di voi non avrei mai iniziato questa carriera.* I know that both of you, would have not be happy if you had not met great partners. They have been supportive with me, accepting some weird behaviour of a difficult child and teenager as I was. I really thank you for that.

I was an only child since I was 14 when my sister Sofi was born: I've always loved you and I'm aware that your life looks like a mess right now, but your intelligence and great sensitivity are hard to find. I know that whatever you choose to do it will make you happy, or better to say, you will turn it into somethings that makes you

happy. Please, never hesitate to call me as I am a happier person if I can help you. Few years later my first brother, Adriano, was born, then Fabio, and lastly Marcello. In a blink of an eye, I had three brothers and one sister. Adriano, Fabio and Marcello: I'm not sure if you know how much I love you. Unfortunately, we haven't lived in the same city for as much as I want. You have to remember that you are always in my mind and speeches. Each of you has his own unique talents, passions and personalities. Only with time, you will discover how great you are. My only hope is that I will have the opportunity to accompany you through the important steps. I would like to thank Marika: you have been a spark in my life and your enthusiasm has breathed new life into me. Everything is moving fast for you at the moment, but I know that you will make the right decisions. I want to thank Daniele, Federico (Pata), Ivano, Federico (Risone) e Stanislao: You have always made me feel at home whenever I come back to Rome as if the years had never passed. We have known each other since 15 years and I don't know how you can still stand me. I know, I should call you more often, but as said, you know me well enough. Thank you for being always there and I have never had any doubt whom I should call for sharing joy and sadness. You have all made such an impact in my life that as a physicist, I will never be able to quantify. Thank you Alison for having supported me during the last months of the PhD, having me was one of the best thing that could have happened. I really enjoy the time we spend together and I have no idea how you are able to control my stormy personality, but you manage. Thanks for helping me in changing some bad habits. You put a smile back on my face and I hope I will be able to do the same. I'm quite sure I should have thanked other persons, but I always find it difficult to deal with emotions. I have a long wondered how to write this last sentence, but there is nothing I can do to make it better. You have inspired me since the day we met and made me believe in love again. Although you are not longer with us, you are still able to shape my life in better, to guide and teach me. To Sidney, with love.

Contents

1. Introduction

1.1. Bridging physics and single-cell genomics

What is life? This is not only a rhetorical question that you will hear by someone trying to flirt in a club, but it is also the title of a fascinating book by Erwin Schrödinger [1], who asked himself the same questions, for possibly deeper reasons. If we just try to pause for a moment and think about what life is, we would have a hard time even to figure out where to start. Possibly, depending on your interests and experience, you would have a different starting point: the origin of life, the social life or the development of life. Personally, what fascinated me the most and let me start the PhD journey is the latter one. In particular: how do cells self organise into tissues and organs? How can they make precise decisions in space and time, despite the noisy environment they live in? Understanding the mechanisms underlying the regulation of cell fate is pivotal, not only for a comprehension of the processes responsible for development, regeneration and ageing, but also for diseases that occur upon disregulation of these processes, such as cancer. In order to characterize many biological processes that occurs during the life of an individual - cell decisions are not an exception - it is fundamental to consider their energetic costs. Biological processes require a flux of energy from the environment, which is essential for life. But how do different organisms process this flux of energy? Schrödinger and many others after him would have replied that energy is essential to form structures, so eventually to make order. There is a constant battle throughout our life between energy, which builds order, and entropy, which destroys it. We already know which one is the winner, as eventually, the death is the inevitable triumph of entropy. A net flux of energy, in a physical language, is associated with systems out of thermal equilibrium. Nonequilibrium physics then arises as a natural framework to understand development and cell behaviour [2]. In order to study the development of life we cannot withdraw from a theoretical understanding of systems out of equilibrium. As mentioned, we still need a starting point as the factors involved in cells behaviour are quite too many to be covered in a thesis, or even 100 of them. We thus look at the minimal scale that current technologies allow us, which is the molecular one. On the molecular level, different processes are involved in regulating cellular fate [3]. One factor, which historically has been extensively covered is the expression of genes [4]. In the last decades, it has become clear that there are additional, epigenetic layers of

regulation: dynamic changes in the way the DNA is folded, modifications to the protein complexes around which the DNA is wrapped and chemical modifications of the DNA itself play a role in determining cell fate.

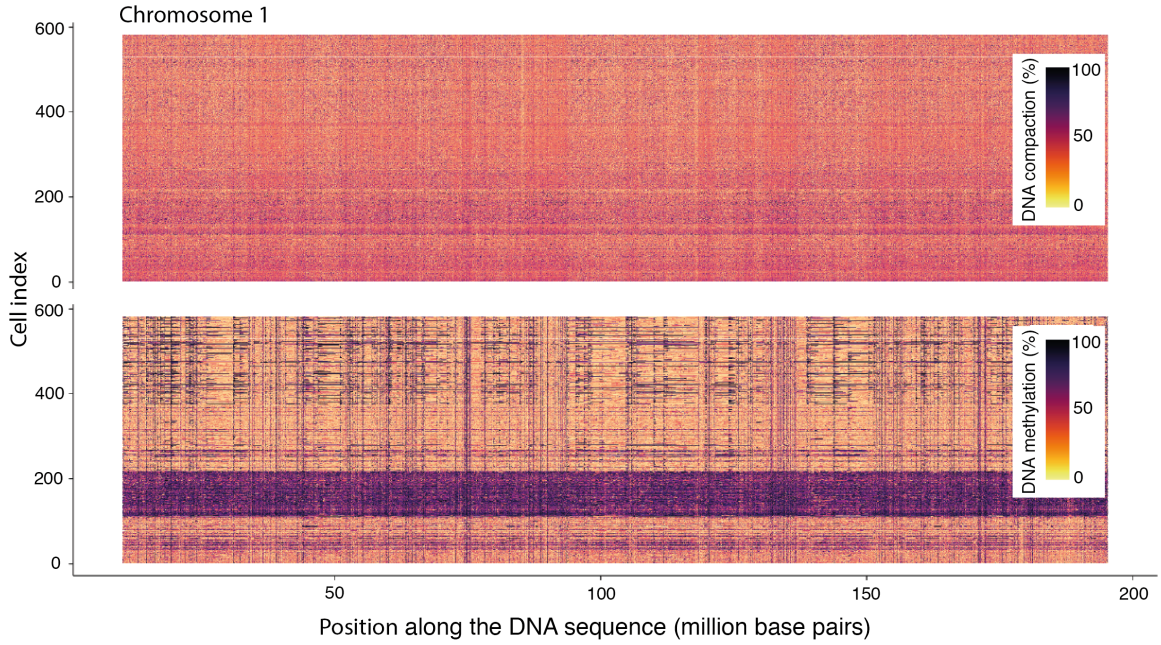


Figure 1.1.: DNA methylation and DNA compaction of chromosome 1 of the mouse embryo at a base pair resolution for hundreds of cells from single-cell sequencing experiments.

This is our starting point, but we always have to keep in mind that no process is completely decoupled from the other, such that the more we explore the neighbours of our starting point the more connections we will find. In this thesis, I will offer a unified theoretical and conceptual framework to deal with some of them. Even though all these layers of cell regulation have been known for decades, only in recent years technological breakthroughs in single-cell biology allow probing them with unprecedented microscopic detail in living organisms (*in vivo*) [5]. In sequencing experiments we obtain information on the expression of thousands of genes, a detailed molecular profile of epigenetic modifications and even the spatial organisation at single-cell resolution. The breakthrough is that we can profile all these layers of regulation, at the same time, for a single cell (Fig. 1.1), which we refer to as multi-omics. Single-cell multi-omics technologies have led to detailed descriptions of the molecular processes underlying cellular behaviour [6]. Such a breakthrough posed some questions about the way we were thinking about cell types, changing it from a discrete representation to a continuous one. Going even deeper, we still struggle to properly define what a cell type is, or whether this concept might be obsolete. As nothing is completely decoupled, even the knowledge of all these layers is not enough to get a full comprehensive idea. Biological functions, such as cell differentiation or proliferation are determined by emergent (collective) states on

the cellular and tissue scale that arise from interactions between processes occurring at the molecular scale, but the collective properties of interacting many-particle systems do not necessarily obey the rules that govern the microscopic scales. Schrödinger himself would have had a hard time to understand how ferromagnetism works by solving the associated Schrödinger equation describing detailed atomic interactions. In biology, the problem we are facing is similar, as the collective dynamics underlying biological function cannot be straightforwardly inferred from detailed molecular measurements. Hence, despite the excitement brought by novel developments in single-cell genomics ("2018 breakthrough of the year" by Science), insights from these technologies remain descriptive until matched with methods to identify collective degrees of freedom. We thus arrived to the central question of this thesis: How can we unveil from detailed quantitative information of the microscopic scale the emergent processes that determine biological function at the cellular and tissue scale? For an untrained eye, this looks like the solution to a puzzle for which we have all the constituent building blocks accessible via multi-omics experiments. Why can we not solve the puzzle? Rephrasing the question in more daily life terms: would we be able to understand how an engine works by looking at its individual components? The challenge is not only to rebuild the engine, but to understand the consequences of taking away just one particular screw, which in the analogy is as if we were provoking a disease in a body. The knowledge of the building blocks is not enough, but we need to understand how they work together, a phenomena know as emergence of collective behaviour. Collective behaviour is key to understand processes such as symmetry breaking: how can two identical cells make different decisions? What are the factors involved in these decisions? Non equilbrium physics of complex systems deals with the solution of these puzzles, by studying how emergence of macroscopic collective behaviour arises from microscopic interactions. In particular, field theory and renormalization group theories [7, 8], which give a theoretical framework to infer different contributions of microscopic degrees of freedom, come as a natural and powerful framework to begin to understand the collective processes underlying cellular behaviour and symmetry breaking *in vivo*. Recent technological advances, such as deep learning, may suggest that the analysis of big single-cell data set, as in Fig. 1.1, has a better performance when done by an artificial intelligence [9]. If an artificial intelligence can humiliate the best Go players in the world, it can do the same with physicists [10]. Artificial intelligence can only be used as a powerful predictive tool, but, due to the lack of general conceptual framework, it remains descriptive and not suited to bridge the scales ranging from microscopic to macroscopic. In particular, it will not provide insight into the underlying rules governing the mechanisms of cell behaviour for the following reasons: First, it requires solving a difficult "inverse" problem which involves mapping given sequencing profiles to one out of an infinite number of processes in space and time. Solving this problem computationally involves probing

a large number of such processes for consistency with the sequencing data. Secondly, emergent properties of interacting complex systems do not usually obey the rules that act on its constituents (emergence). A mechanistic understanding of biological processes does not only allow us to still give accurate predictions, but when something changes with respect to our predictions, we can pinpoint the causes of the change and eventually open up the world to something unexpected. As an example, the breaking of a certain mechanism can give insight into the particular causes of a disease and the understanding of it will greatly simplify the work to find a cure for it. In this thesis, we will pursue an interdisciplinary approach and combine novel technologies in single-cell genomics and nonequilibrium statistical physics to understand collective processes underlying cellular behaviour. Our work will aid to overcome important conceptual limitations in genomics inferring the emergence of macroscopic and collective behaviour from molecular measurements, and provides a framework for understanding the function of key biological processes underlying cellular regulation. At the same time, we will take an interdisciplinary approach to tackle fundamental questions in nonequilibrium physics. Specifically, we will develop and solve original theories for out of equilibrium field theories with non-local and non-linear interactions, scale invariance of processes that are not close to a critical point, disordered systems without symmetries and multi scale interacting complex systems. In the following part of the introduction, we will describe the biology of gene expression and DNA methylation as fundamental mechanisms underlying cell behaviour and their dynamic changes throughout the life span of an individual. Later, we will give a more detailed description of current technologies in genomics and their limitations. Finally, we will introduce the main theoretical concepts and tools of out of equilbrium systems used in this thesis.

1.2. Biological background

1.2.1. Epigenetics

Although there are different definitions of the term epigenetics, here we will adopt the definition: “ Epigenetics is the change in the state of expression of a gene that does not involve a mutation, but that is nevertheless inherited in the absence of the signal or event that initiated the change ” [11]. Example of epigenetic processes are: changes in the chromatin structure, histone acethylation, enhancers or DNA methylation. In this thesis, we will mostly study DNA methylation, which is one of the primary layers of epigenetic modifications. DNA methylation is a chemical modification that affects the nucleic acids of the DNA. In particular, in mammals it mostly affects cytosine (C) when it is next to a guanine (G) adding a methyl group to the cytosine. This base-pair is referred to as CpG where p stands for the phosphorous between cytosine and

guanine, and we refer to it as 5mC whenever it is methylated. Methylation patterns play a crucial role for the development of an individual as well as for its adulthood [12]. Changes in methylation patterns have been associated for example to cancer cells [13] and furthermore the readout of methylated DNA can be used to infer the biological ages of individuals [14]. Their tight regulation is thus essential for the life of an individual.

There are three main processes responsible for the changes of DNA methylation: *de novo*, *maintenance* and *demethylation* DNA methylation. DNA methylation is established by enzymes from the DNMT3 family. In particular, DNMT3a/b actively modify the epigenetic status of CpGs into methylated cytosines Fig. 1.2 A, whilst DNMT3L are used to recruit the DNMT3a/b. Maintenance of DNA methylation is the re-establishment of DNA methylation after cell division. During DNA replication, the new strand is not methylated and DNMT1 reads out the methylated sites on the template strand and copies them into the new one. Demethylation, as we will later specify, is caused by lack of DNMT1 or by active removal of methylation marks by enzymes in the TET family. We will first focus on the establishment of DNA methylation, also referred to as *de novo* DNA methylation. We will later explain the effect of maintenance and demethylation of DNA methylation and its role in ageing and cancer.

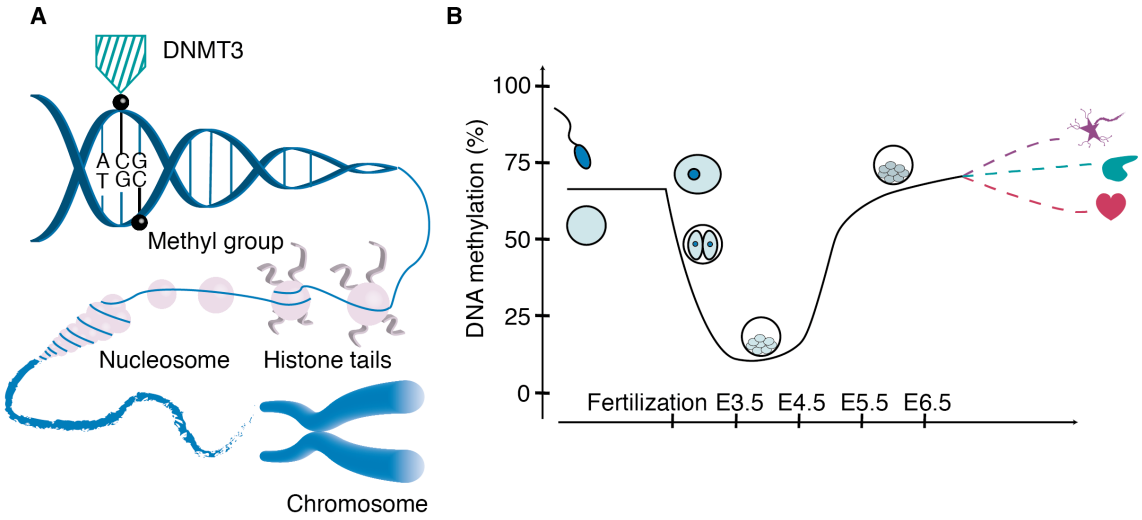


Figure 1.2.: (**A**) Establishment of DNA methylation by DNMT3 enzymes. Methylated CpG are highlighted in black. (**B**) Schematic sketch of DNA methylation dynamics during early mouse development (EN stands for N days after fertilization).

Early embryonic development

During development there are drastic changes at the global and local level of DNA methylation. In particular, after fertilization, embryonic cells inherit paternal and maternal DNA methylation marks. In the first days of the mouse embryo, paternal and

maternal DNAme is erased and it is followed by a wave of *de novo* DNA methylation. In particular, the global level of DNA methylation increases from $\sim 20\%$ to $\sim 80\%$ (Fig. 1.2 B). This wave of *de novo* DNAme precedes the first decisions made by embryonic cells and is one of the factors that determines their fate [15]. A deep understanding of how methylation patterns are established is thus key to understand cell fate decision. The mechanisms of *de novo* DNAme are largely unknown *in vivo* and even in *in vitro*. Only recently, due to the breakthrough of new technologies [16–18] we got access to a detailed profiling of DNAme at the base pair resolution for single cell. In [19] DNAme patterns reveled the specific binding locations of DNMT3A/B at different genomic loci.

DNMT3 enzymes are responsible for *de novo* DNAme. We can identify three different DNMT3 enzymes: DNMT3a, DNMT3b and DNMT3L [20]. These enzymes, during development, have partially non-overlapping biological functions [21]. In particular, DNMT3a and DNMT3L enzymes are required for methylation of most imprinted loci in germ cells [22]. DNMT3b is responsible for methylation of the centromeric region [23] and its mutations are a cause of facial anomalies syndrome, rare autosomal disease [24]. DNM3L is catalytically inactive and cooperates with the other DNMT3 enzymes (Fig. 1.3 A) to methylate the DNA and it is a positive regulator of methylation at the gene bodies of housekeeping genes [25, 26]. DNMT3L is also involved in the release of DNMT3a from dense heterochromatin to make it available to act at imprinted differentially methylated regions [27]. The enzymes in the DNMT3 family act together to establish new methylation marks upon binding to the DNA.

The complex of DNMT3, as shown in Fig. 1.3 B occupies 8-10 bps along the DNA, whereas CpG have a typical distance of $\sim$ 100bps. The differences and similarities between enzymes in the DNMT3 family is evident from the distinct functional domains. The N-terminal part in DNMT3a and DNMT3b contains two defined domains, ADD and PWWP, the latter one being absent in DNMT3Ls. The PWWP part is essential for the targeting of the pericentromeric chromatin and the ADD part part constitutes a platform for protein-protein interactions. The ADD part also interacts with the N-terminal part of histones H3 tails, thereby stimulating DNAme. active transcription. DNMT3L interacts with unmethylated tails of the histone, recruiting DNMT3a at specific loci [28, 29]. The C-terminal domains of all three DNMT3 have the AdoMet-dependent MTase fold, but DNMT3L contains several amino acid exchanges and deletions within the conserved DNA-(cytosine C5)-MTase motifs, stressing the impossibility of catalytic activity by these enzymes [30].

A detailed knowledge of the functions of DNMT3 enzymes, is not sufficient to understand how they interact between each other in order to bind and methylate the DNA during development and how it affects the first cell fate decisions made in the embryo. DNAme provides an epigenetic barrier that reduces developmental potential by promoting different cellular identity [31]. Changes in methylation allow the zygote

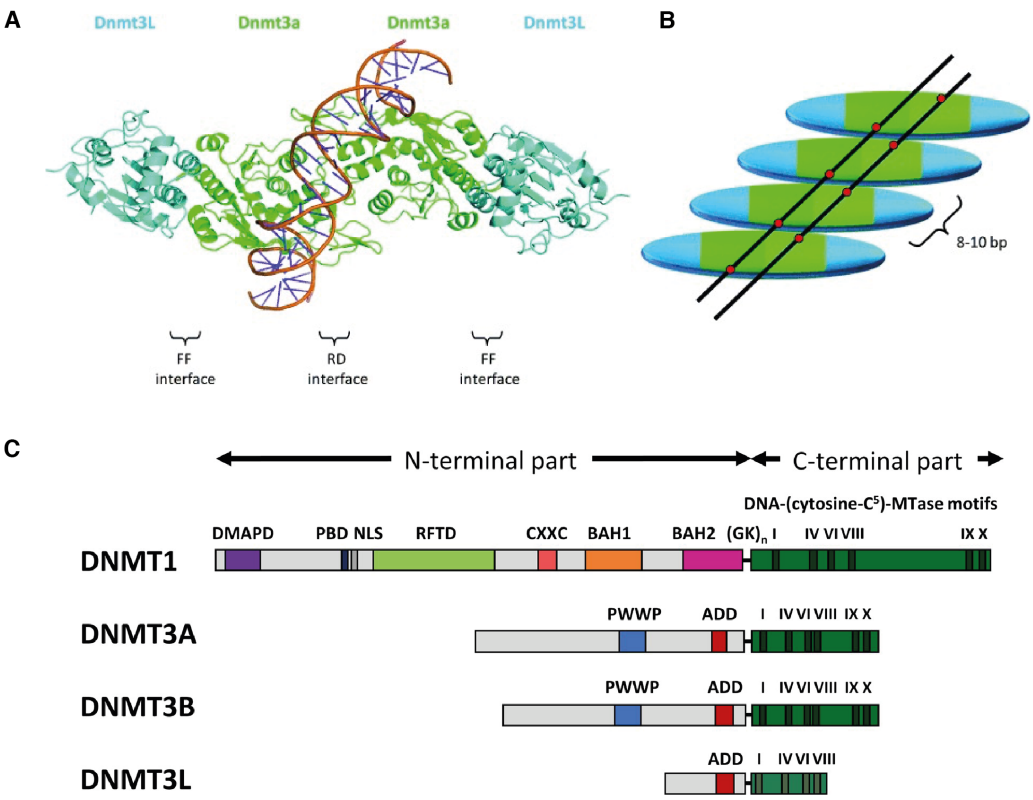


Figure 1.3.: (**A**) DNMT3s and their cooperative action along the DNA. (**B**) Occupation length in bps of the enzyme group. (**C**) Functional domains of DNMT3 enzymes. Figures are adapted from [25, 30].

to erase the epigenetic signature inherited and to regain developmental totipotency. An in-depth understanding of the mechanisms responsible for changes in DNAme is thus crucial to tame potency trajectories and diseases.

Ageing and cancer

DNA methylation patterns are not fixed throughout the life of an individual after *de novo* DNAme. In particular, there are processes which actively or passively remove methylation marks. Passive demethylation is caused by *maintence* of DNAme upon cell division by DNMT1 enzymes [33]. DNMT1 reads out the methylated strands and copies the methylation status to the new strand. Passive demethylation is then caused either by a loss of DNMT1 [34] or by a failure of the "copy-paste" machinery [3]. In Fig. 1.4 A, we show how DNMT1 enzymes recognize DNA methylated cytosines and methylate the cytosines on the other strand. After the first DNA replication, the DNA is said to be hemimethylated before the pattern is restored. It is found [35] that cells with depletion of DNMT1 progressively lose DNAme at the promoter of the tumor-suppressor gene CDKN2A. In human cells, DNAme at the the promoter of CDKN2A is associated with its silencing [36]. In general, DNA methylation can be an early sign to detect cancer

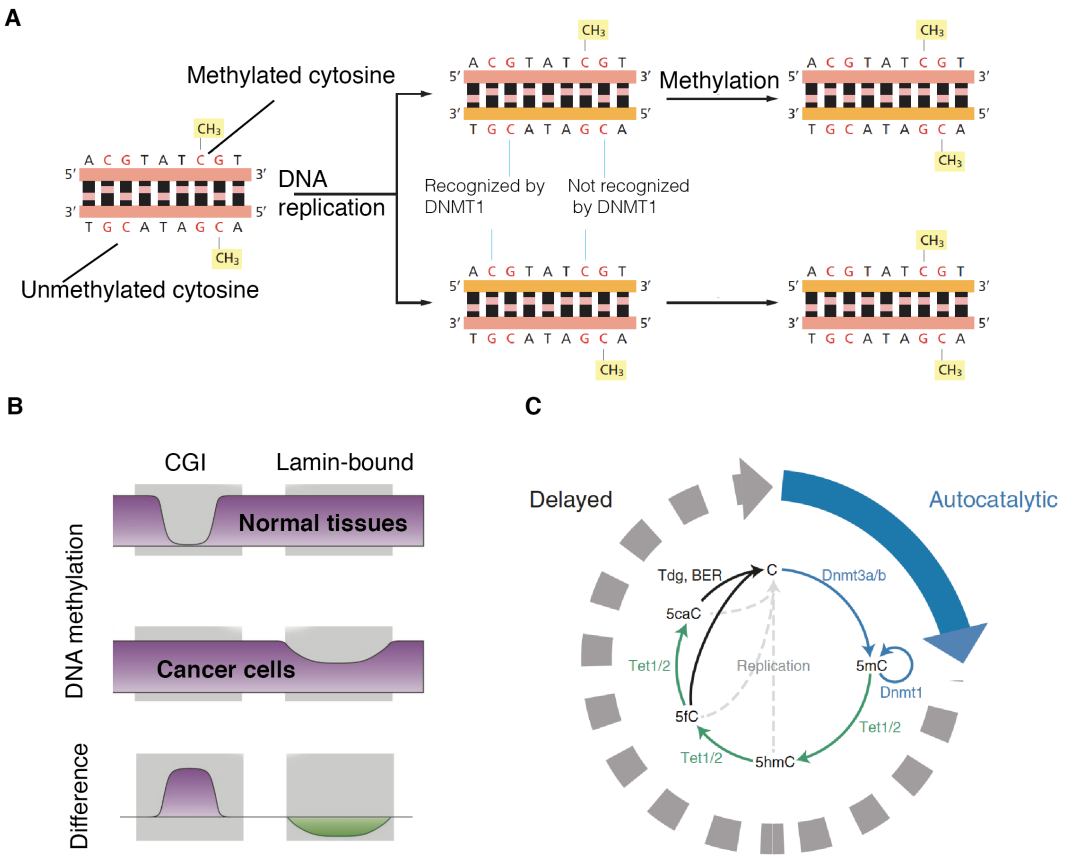


Figure 1.4.: (**A**) Maintenance of DNA methylation by DNMT1 enzymes after cell replication. (**B**) Differential changes of DNA methylation in cancer cells at the level of CpG islands (CGI). (**C**) DNA methylation oscillations driven by *de novo* methylation, active and passive demethylation. Figures are adapted from [3, 13, 32].

cells [13]. Without entering into details, even though the specific DNA methylation patterns are not yet fully understood, regions of the genome with high CpG density (CpG islands) are usually less methylated in comparison to less dense CpG regions [3]. In cancer cells, the situation is almost reversed (Fig. 1.4 B) and CpG islands ($\sim$ 13000 in the human genome) may undergo *de novo* methylation whilst demethylation occurs at the nuclear lamina domains (lamin-bound), which characterize the structure of the chromatin. This demethylation is thought to be mostly passive and it is caused by a failure of DNMT1 machinery due to the fast replication in cancer cells [37] . Active demethylation may happen due to several factors. Here we focus mostly on the role of the TET enzymes family (TET1/TET2/TET3), which add a hydroxyl group onto the methyl group of 5mC (DNAme) such that the new methylated CpG will be referred to as hydroximethylated (5hmC) [38]. CpGs then goes through successive changes of the chemical modification ($C \rightarrow 5mC \rightarrow 5hmC \rightarrow 5fC \rightarrow \ldots$) [39] leading to an active turnover [40]. We can then think of DNA methylation as a clock, Fig. 1.4 C. Methylation turnover leads to oscillations of DNAme which becomes evident during exit from pluripotency [32], such that the understanding of such machinery becomes a

key to unveil symmetry breaking in the embryo. DNAme is not only associated with development or cancer, but changes of DNAme patterns are observed during ageing [41]. We can then use DNAme to quantify chronological and biological age of individuals.

In the last decade there have been improvements in machine learning techniques and multiple epigenetic clocks were built. According to [14] a "DNA methylation clock" - example of an epigenetic clock - is an estimator built from DNAme marks that are strongly correlated ($r \geq 0.8$) with chronological age or time, which can accurately quantify an age-related phenotype. One of the first clocks [42], tested on blood-derived DNA, takes only 71 CpGs in order to infer chronological and biological age. An epigenetic clock constructed in [43] does not need a specific cell line, but it can accurately predict the age with cells from different tissues. These epigenetic clocks have limitations both from sample sizes and from a lack of mechanistic understanding, but they are anyway able to measure the biological age of an individual, which may differ between individuals with the same chronological age.

DNAme then plays an essential role during life and in this thesis we will develop the analytical and statistical tools to investigate and predict its dynamic changes as well as to understand the underlying biophysical mechanisms.

1.2.2. Gene expression

Gene expression is the process that allows cells to translate genetic information into functional proteins. There are many factors involved in the regulation of gene expression and in the following we summarized some key steps that lead to the production of a functional protein [3].

- RNA polymerase (RNAP) binds to the promoter of a gene, thus initializing *transcription*.
- A bound RNAP opens the DNA, exposing the chromatin for roughly 10bps (*transcriptional bubble*), it then slides along the the gene body one base pair at a time until it arrives at the terminator part.
- The outputs of this process are either noncoding RNAs or messenger RNAs (mRNAs).
- mRNA is translated into functional protein by ribosomes.

As all the the previous steps encode noise, even within this simplified picture, the expression of a single gene is subjected to great variability making the distribution of gene expressions possibly long-tailed [44, 45]. On top of that, genes interact with each other, introducing another source of noise, which may facilitate the regulation of all the genes in the nucleus, namely gene regulatory networks (GRNs). Interactions are

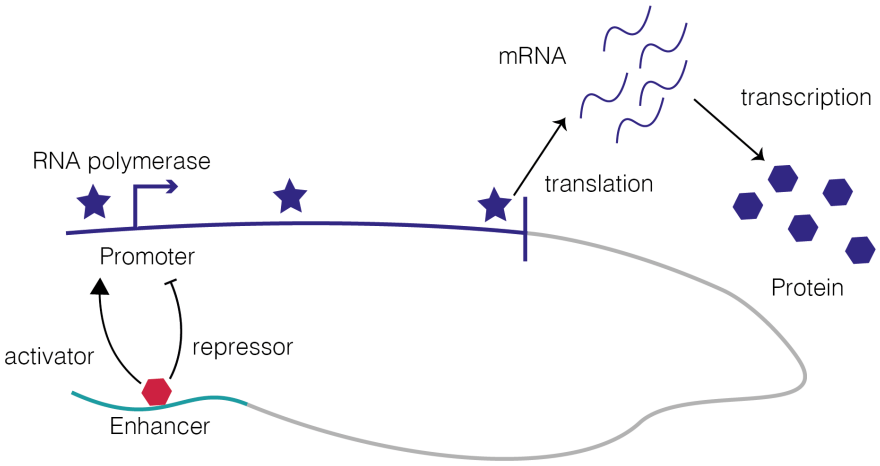


Figure 1.5.: Simplified representation of different steps involved in gene expression. RNA polymerase (star) translate the gene into mRNA which is then transcribed into protein. Proteins of different genes or the gene itself (self-regulating) can act as activators or repressors upon binding to enhancer or promoter regions.

essential in the following sense: if there is no interaction between genes, every gene is transcribed independently from the others. The only way to control transcription would be to change the rates at which every transcription, translation degradation process is happening. Even though this is not an absurd scenario, as we will see when studying the effect of DNAme in gene bodies, it is quite reductive and likely cannot cope with the time scales of a living cell. The simplest form of interaction we know are via *activator* and *repressor* transcription factor proteins (TFs). Activators bind to either operator or enhancers and, as the name suggest, they increase the rate of transcription. The detailed role of enhancers is beyond the scope of this thesis, but it is fascinating to know that these genetic regions can be even millions of base pairs far from the promoter of the targeted gene and recruit specific TFs and loop in the three dimensional space of the nucleus to interact with the the promoter [46]. Repressor proteins, on the other hand, prevent transcription by binding to operators or enhancers, Fig. 1.5. Activators and repressors proteins thus give rise to interactions between genes, changing the distributions of protein and mRNA in a living cell. When interactions are taken into account, genes form a network where each node interacts with other by activating or repressing each other. Not every gene plays a role in the transcription of the other, as two genes may not interact at all. As an example, in E. Coli, the expected number of genes activating or repressing a target one is $2-3$ [47], making this GRN a sparse network. How does the knowledge of GRNs help in understanding cell fate decision?

Before giving an answer to the last question, we need to know what a cell state is. It is hard to point out at a single definition of cell state or cell type [48]. We can generally

say that a given cell type is characterized by the expression of certain genes and the downregulation of others. Understanding how different GRNs work helps to infer cell state transitions and, even more generally, cell types. In particular, during development, the potency of progenitor cells is increasingly restricted as they undergo numerous fate decisions Fig. 1.6 A. After fertilization, an organism is comprised of a single cell, which keeps on dividing. A single cell, upon the first division, becomes two identical cells, then four, eight, etc... Individuals are clearly not ball of identical cells and something must happen, in a way that cells take different decisions and form different structures. This phenomenon is referred to as *symmetry breaking*. It looks like a complicated concept, but is not and we all experience that. Imagine that we put a ball carefully on the top of an hill between two valleys. Where will the ball eventually fall (if it does)? We would say with conviction that it will fall with 50 % probability in one of the valleys. This is partially true. If the landscape and the ball itself are perfectly symmetric (rotationally symmetric), then 50 % should be the outcome. We said partially true because when all the details are taken into account we could realise that the hill is actually not as smooth as we thought and there might be a little bit of wind blowing in a particular direction. If we had known all these conditions, we would have been able to tell in which valley the ball will fall. In physics, for example, the ball may pictorially represent a ferromagnetic material and the valleys possible orientation of the magnetization. If a temperature is quenched from a temperature higher to one lower than the Curie temperature, the ferromagnetic material will have a net magnetization in a given direction, breaking the rotational symmetry. Similarly, cells can be pictorially represented as balls rolling down a hill with many valleys, where a valley is a cell type (Fig. 1.6 B). This representation is referred to as epigenetic or Waddington landscape [49]. There have been many attempts to quantify the Waddington landscape [50–52] and so to relate the pictorial representation to actual quantitative data. As an example, in (Fig. 1.6 B) there are no axes as it is not clear, as far as our knowledge is concerned, if they can be uniquely identified in terms of biological parameters. Even though the Waddington landscape gives a simple picture to hierarchically order cell types, it does not capture how they are defined cell state transitions with respect to the microscopic parameters.

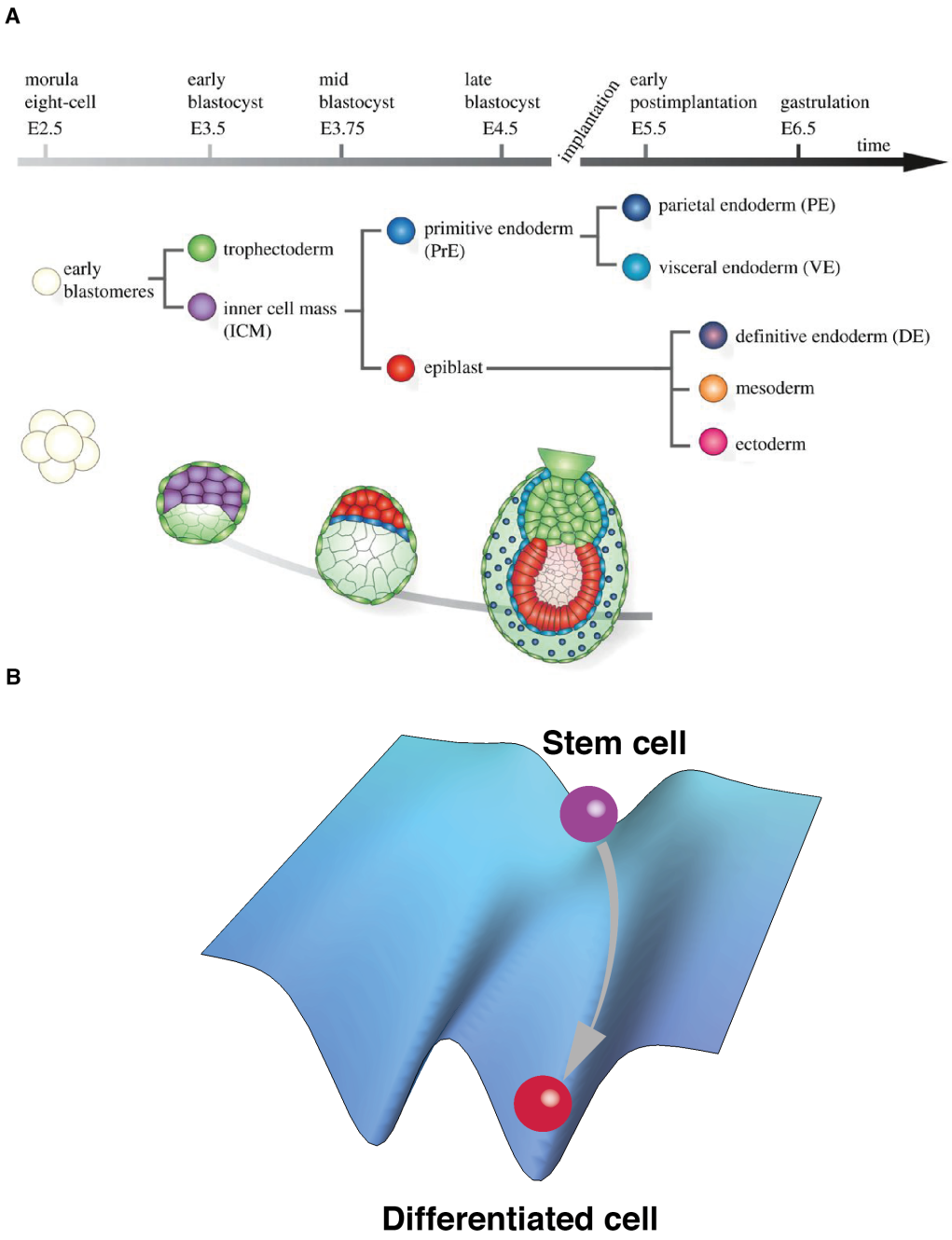


Figure 1.6.: (**A**) Cell fate decisions during development. Epiblast cells (red) are embryonic progentior cells and have the remarkable capacity to generate any somatic cell type. Figure adapted from [53]. (**B**) The Waddington landscape as a pictorial representation of cell fate decisions.

1.2.3. Sequencing of DNA and RNA in single cells

In the previous section we introduced two main layers of regulation of cellular symmetry breaking: gene expression and DNA methylation. Nowadays, sequencing experiments allow us to quantify both layers at the same time for single cell [6]. We will discuss mainly two types of sequencing experiment: RNA sequencing (RNA-Seq) and bisulphite

sequencing (BS-Seq), as the main sequencing concepts are shared between different experiments. In RNA-Seq and BS-Seq experiment the molecular content of a cell is first extracted from the nucleus, respectively RNA and DNA, or with current multi-omics technologies, both at the same time. As a first step, cells need to be sorted. One possible method is flow-activated cell sorting (FACS), that tags cells with a fluorescent monoclonal anti-body, which recognizes specific surface markers [5]. The content is initially fragmented and then amplified, as it would be extremely hard to get any signal, especially for single cells. This steps introduces technical biases (batch effects) as the cells are not equally amplified and they have to be corrected in the analysis of the resulting data set. Batch effects are then a source of technical variability that makes it extremely hard to compare a data set of the same experiment, done for example, in different laboratories. As the content was fragmented, there is a further step of alignment to map back each fragment to the sequence along the DNA or RNA [54, 55]. The information is then stored and further quality control steps must be done in order to have an informative data set. In particular for DNA sequencing, due to the amplification steps, whenever the fragments are mapped back to the genome of reference, there might be multiple fragments for the same part of the genome. The number of fragments multiplied by the average fragment length and divided by the length of the genome gives the coverage, which is used for quality control as it is a measure of the information contents. All the previously introduced steps of sequencing come along with technical noises and biological variability, thus requiring machine learning and sophisticated pipelines to obtain quantitative predictions from sequencing experiments. In the next sections we are going to give more details of the two technologies, but we have to keep in mind that we will analyse single-cell multi-omics experiment as well, where these technologies are used together for individual cells.

Bisulphite sequencing

Bisulphite sequencing allows us to obtain information on the methylation status of CpGs along the DNA sequence [56]. We do not enter into details of the technical steps of bisulphite sequencing, but we outlined them briefly in Fig. 1.7. In particular, a DNA strand passes through a bisulphite treatment where all the non-methylated cytosines are transformed into uracil. Each strand is amplified via polymerase chain reaction (PCR) and uracil is transformed into thymine, such that, the only cytosines that are left are the one that were originally methylated. As it is the case for most of the sequencing experiments, there are mainly two types of bisulphite sequencing: bulk and single cell. Bulk sequencing pushes together fragments of different samples such that the individual methylation status of each DNA strand is lost. Single-cell sequencing maps methylated cytosines of single-cell DNA strands. We will begin by

showing methods for the analysis of bulk sequencing, as it is the most challenging, and we will later deal with single-cell data. In Fig. 1.7 we show the outcome of a typical bulk bisulphite sequencing data. For each sample (sample.id) we have the length of the fragment as well as its start and end position on the chromosome (seqnames). We then have the number of CpG per each fragment as well as the number of informative CpG (number of methylated + number of non-methylated). This is a very crucial part for sequencing analysis, as due to technical errors, there might be misreads such that we lose information of some cytosine or even fragments in the process. The number of informative CpGs in bulk sequencing can be larger than the total number of cytosines as we might have multiple reads per position from different samples. We then need to be very careful and compute every statistical observable accounting for different information contents. One way to do so is to tile the data, as in Fig. 1.7, such that there are at least 10 informative CpGs per tile. Then, a different weight to each tile is defined as $\frac{p+n}{n.cpg}$, with p, n the number of methylated and unmethylated cytosines respectively and $p + n$ is the information per fragment. Individual methylation for each fragment is computed as $\frac{p+1}{p+n+2}$ [16]. Every global quantity can be computed upon weighted averages over individual tiles and eventually filtering out regions which are on the tails of the weights distributions. In single-cell sequencing we do have information for each individual CpG per cell, but the same procedures of weightings can be applied, particularly to identify not informative regions of the DNA.

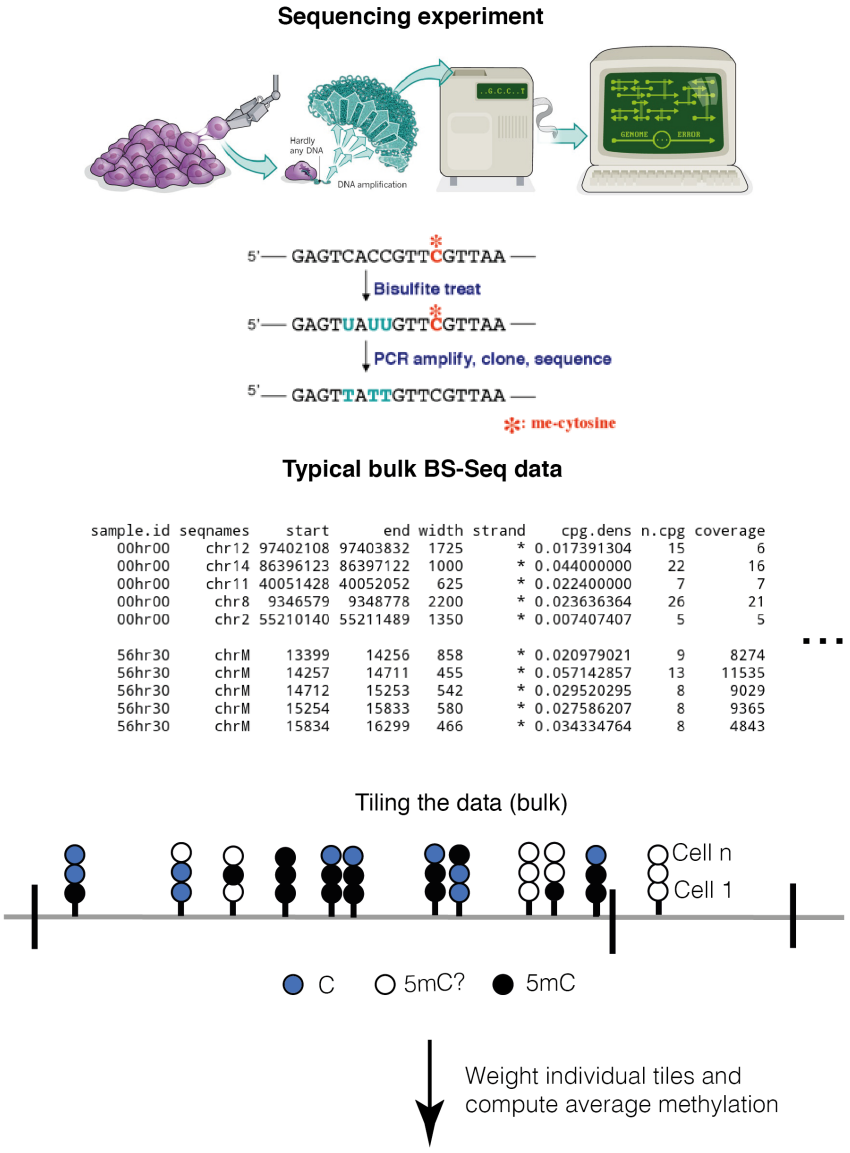


Figure 1.7.: Sketch of the bisulphite sequencing treatment and data processing (adapted from [57]). After the bisulphite treatment, the data is collected and regions which do not have enough coverage are filtered out. In the table we show bulk BS-Seq, where there is no information on DNA methylation at single CpG resolution but rather for different fragments. The data is then tiled upon merging different fragments and finally analysed.

RNA sequencing

As mentioned in the previous section, we have not yet defined a way of how to quantify gene expression and what the limits of recent technologies are. Specifically, during the last years, RNA sequencing [58] has come out as a powerful tool to investigate gene expression and in particular RNA abundances at single-cell resolution [59]. The main steps of single-cell sequencing are the extraction of RNA followed by mRNA enrichment, which gives a total number of roughly 10-30 million fragments per sample. These

fragments, encoding information on single gene transcripts, must be first aligned and then analysed. Via all these steps, we might have already lost a lot of information due to technical imperfections and, as every fragment of the sample is amplified, there is no clear relationship between the readout after sequencing with the actual stochiometric value of the transcripts. We construct a $n \times m$ matrix with n is the number of samples (cells in case of single-cell sequencing) and m the number of genes. The entries are RNA abundances of single genes. As a first step we need to filter samples as done for bisulphite sequencing experiments. Specifically, we remove: cells with low reads (tail of the distribution), cells in which only few genes are expressed and cells with high number of reads of mitochondrial genes as they are signatures of senescent (non proliferating) cells (Fig. 1.8). Due to different batch effects, a $log10$ normalization is applied to the entire matrix. Typically, a gene that does not vary across samples is a sign of technical errors in the sequencing procedure. In order to avoid keeping genes with these feature, a certain number of highly variable genes (HVGs)is selected (typically 1000-2000). The new matrix then is $n \times h$, where h is the number of HVGs. The content of every cell lies in an h-dimensional space and often some visualization techniques are implemented in order to have a clearer representation of the data set. In particular, not all the dimensions may contribute equally to the identification of cell states, developmental trajectories, etc. In order to better visual the high-dimensional data, a dimensional reduction is in generally performed. In Fig. 1.8 we show two historically used dimensional reduction techniques: a principal component analysis [60] (PCA) and uniform manifold approximation and projection [61] (UMAP). Both techniques have different advantages and disadvantages, but they achieve the same outcome of visualising cells in a lower dimensional space (typically two-dimensional). We do not enter into details of the need of different dimensional reduction techniques as it is beyond the scope of this thesis. Briefly, PCA is an orthogonal linear transformation where the first principal components are defined in terms of covariances in the original space. On the other hand, UMAP is non-linear and topology-preserving algorithm which requires that the original data is uniformly distributed on a Riemann manifold. Interestingly, these tools may capture structures in the data that are not obvious in higher dimensions. In order to understand a possible practical use of dimensional reduction, in Fig. 1.8 we use PCA and UMAP to show different biologically relevant quantities. As an example, in the PCA plot we show the total number of gene counts and in the UMAP the expression of the gene Dio3 for all the cells. As we will see later, we can follow trajectories along dimensional reduction plots to understand how a particular gene changes along them and, going back to the cells lying on that trajectory, we might find information on different underlying biological processes. In order to give structure to the RNA-Seq outcomes, algorithms like k-means clustering [62] are often used. They represent samples (cells in scRNA-Seq) in a k-nearest neighbor graph, and identify clusters via

module optimization. Without entering into the specific algorithm details, in general, a matrix with distances between samples is constructed (samples are expressed as vectors of gene expression), where the metric can be chosen. After the distances are computed the closest k samples to one sample are found and then a weighed edge between two samples based on distances and shared neighbours is produced [63, 64]. After the graph is built, Leiden and Louvain algorithm [65, 66] use module optimization, by minimizing a functional on the graph, to define different communities or clusters. Both in k-means and Leiden or Louvain algorithm an arbitrary parameter is introduced by the choice of the value k for the first one and on the number of communities to detect for the latter ones. It becomes clear that, apart from the technical variability of sequencing all these further steps of analysis introduce biases which must be overcome or smoothed for a deeper understanding of biological process analysed with RNA-Seq. To conclude, in this section, we outlined some steps for a bioinformatic and statistical analysis of RNA-Seq and BS-Seq data set. We showed how both sequencing techniques have their limitations and in the next section we are going to present some of the theoretical methods of nonequilibrium systems used in this thesis which will allow us to overcome these limitations and shed light on underlying biological and physical mechanisms that are hidden in data sets constructed with sequencing experiments.

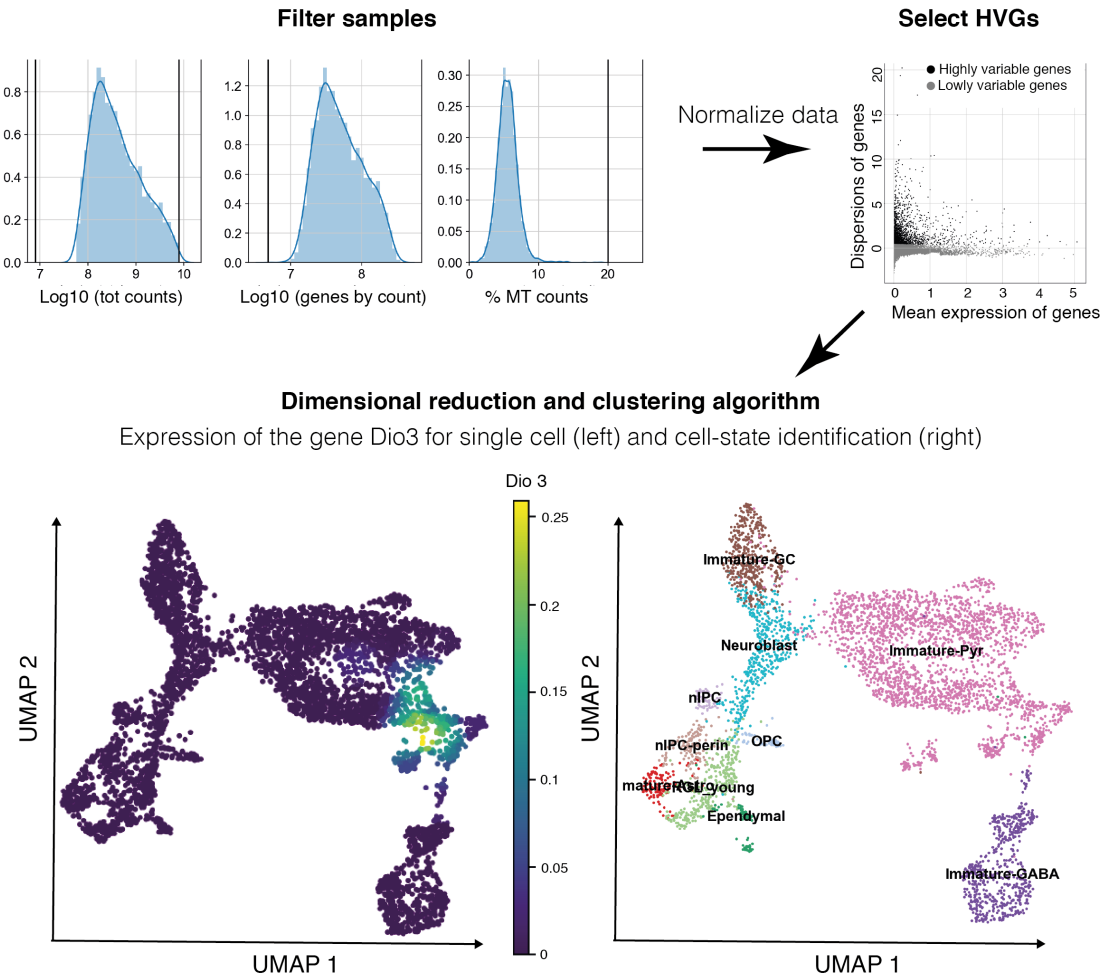


Figure 1.8.: Simplified workflow of standard bionformatics analysis of an RNA sequencing experiment. Cells are first filtered based on technical quality: cells with low number of total RNA counts, low number of total expressed genes (>0) and with high percentage of mitotic RNA are filtered out. After normalization, informative genes are selected (highly variable genes) and then dimensional reduction (here UMAP) and clustering methods algorithms (here Leiden) are used to visualize the data and identify communities. The data are adapted from [67].

1.3. Theoretical background

1.3.1. Field theoretical methods in nonequilibrium physics

In this section we outline the main field theoretical methods for nonequilibrium statistical physics used throughout the thesis. Field theories, path integrals and renormalization group are powerful tools that allow us to predict many statistical and physical observable [68]. Field theories provide a framework to study biological and physical problems with arbitrary non-linearities and different sources of noises for spatially extended stochastic systems. In nonequilibrium physics we mostly deal either with Langevin

equations, describing the stochastic dynamics of a field, which is a coarse-grained description of microscopic or macroscopic variables, or master equations, describing the spatio-temporal changes of a probability distribution itself. Langevin equations may be described via Fokker Planck equations, describing the probability distribution of the field [69], in a similar manner as the master equation. The main difference between the two formulations of a stochastic process is that master equations allow discrete changes in the variables, whilst Fokker Planck equations are in a continuous form, so we have to choose the right framework based on the problem we are dealing with. Path integrals are a closed representation of field theories and their use greatly simplifies, as we will see, the computation of physical observables. We will first construct a path integral starting from the Langevin equation, known as Martin-Siggia-Rose-Janssen-De Dominicis [70–72] and later we will construct it starting from a master equation, known as Doi-Peliti [73, 74].

Path integral representation of the Langevin equation

We start with a general stochastic partial differential equation describing the dynamics in space ($\vec{x}$) and time t of a field $\phi(\vec{x},t)$,

$$\partial_t\phi(\vec{x},t) = L(\phi(\vec{x},t)) + G(\phi(\vec{x},t))\eta(\vec{x},t)\,. \tag{1.1}$$

L and G are general functions of the fields. $\eta(\vec{x},t)$ is the noise, which in a simple form is Gaussian and uncorrelated with zero mean and unitary variance, such that $\langle\eta(\vec{x},t)\eta(\vec{y},t')\rangle = \delta(t-t')\delta(\vec{x}-\vec{y})$, where $\langle\ldots\rangle$ is the average over multiple realisations of the noise. This is an arbitrary complicated equation, without a general solution and we need to find a starting point to construct a path integral. As a wise man told me once "when you don't know what to do, just insert an identity!". There are many ways to add an identity, but we only know that the field $\phi(\vec{x},t)$ must satisfy Eq. (1.1). We can then write

$$1 = \int \mathcal{D}\left[\phi(\vec{x},t)\right] \prod_{\vec{x},t} \delta(\partial_t\phi(\vec{x},t) - L(\phi(\vec{x},t)) - G(\phi(\vec{x},t))\eta(\vec{x},t))\,, \tag{1.2}$$

where $\mathcal{D}\left[\phi(\vec{x},t)\right]$ stands for all the possible values the field can take. We thus found the identity, at the price of adding a delta function. We can then use the functional integral representation of the delta function

$$1 = \int \mathcal{D}\left[\phi(\vec{x},t)\right] \mathcal{D}\left[i\tilde{\phi}(\vec{x},t)\right] e^{\int d\vec{x} \int dt\, \tilde{\phi}(\vec{x},t)[-\partial_t\phi(\vec{x},t)+L(\phi(\vec{x},t))+G(\phi(\vec{x},t))\eta(\vec{x},t)]}\,, \tag{1.3}$$

where $\tilde{\phi}(\vec{x},t)$ is, at this stage of description, a dummy variable introduced as the

functional representation of the delta function and it has no clear physical meaning. In the last step we rewrote Eq. (1.1) as a path integral, which is still hard to handle. We are only interested in a general statistical observable of a physical quantity, that we defined as $O\left[\phi(\vec{x},t)\right]$. Such a quantity may be the average of the field $\langle\phi(\vec{x},t)\rangle$, time autocorrelations of a field or spatial correlations between the field at different values of $\vec{x}$. As the process we are studying is stochastic, we are interested in computing averages over the noise distributions. In particular, $\langle O\left[\phi(\vec{x},t)\right]\rangle$ can be expressed as,

$$\langle O\left[\phi(\vec{x},t)\right]\rangle = \frac{\int \mathcal{D}\left[\eta(\vec{x},t)\right] P(\eta(\vec{x},t)) O\left[\phi(\vec{x},t)\right]\cdot 1}{\int \mathcal{D}\left[\eta(\vec{x},t)\right] P(\eta(\vec{x},t))\cdot 1}\,, \tag{1.4}$$

where P is the probability distribution of the noise. We multiplied numerator and denominator by 1, which will be substituted with the definition of the identity, Eq. (1.3). Upon writing the probability distribution of the Gaussian noise,

$$\begin{aligned}\langle O\left[\phi(\vec{x},t)\right]\rangle \propto \int \mathcal{D}\left[\eta(\vec{x},t)\right] \int \mathcal{D}\left[\phi(\vec{x},t)\right] \mathcal{D}\left[i\tilde{\phi}(\vec{x},t)\right] e^{-\frac{1}{4}\int d\vec{x}\int dt\,\eta(\vec{x},t)\eta(\vec{x},t)} \\ e^{\int d\vec{x}\int dt\,\tilde{\phi}(\vec{x},t)[-\partial_t\phi(\vec{x},t)+L(\phi(\vec{x},t))+G(\phi(\vec{x},t))\eta(\vec{x},t)]} O\left[\phi(\vec{x},t)\right]\,,\end{aligned} \tag{1.5}$$

and integrating over the quadratic term in $\eta(\vec{x},t)$,

$$\langle O\left[\phi(\vec{x},t)\right]\rangle = \int \mathcal{D}\left[\phi(\vec{x},t)\right] \mathcal{D}\left[i\tilde{\phi}(\vec{x},t)\right] O\left[\phi(\vec{x},t)\right] e^{-S[\phi,\tilde{\phi}]}\,, \tag{1.6}$$

with

$$S[\phi,\tilde{\phi}] = \int d\vec{x}\int dt\,\tilde{\phi}(\vec{x},t)\left[\partial_t\phi(\vec{x},t) - L(\phi(\vec{x},t)) - G(\phi(\vec{x},t))\tilde{\phi}(\vec{x},t)\right]\,, \tag{1.7}$$

we arrive to the MSRJD path integral formulation of the Langevin equation. In this formalism, any observable is represented as a path integral over two different fields, $\tilde{\phi},\phi$. In the next section we will arrive to a very similar path integral starting from a general master equation. We will then show a practical example and how to compute desired observables.

Path integral representation of the master equation

In this section we present the formalism developed by Doi [73] and Peliti [74] to represent master equation in a path-integral formulation. We start with the description of particles evolving on a one dimensional lattice with N sites. This choice is useful for didactic reasons and close to the specific theories we will develop in the rest of the thesis. The generalization to higher dimension is straightforward, and not useful for the purposes of this thesis. We define $\mathbf{D} = (D_1,\ldots,D_N)$ the discrete values of the quantity D across the lattice. At each time step, there are mainly three different process that

can happen.

- **D** can increase its value by discrete amounts in any lattice sites, or multiple at once, for example $D_i \to D_i + 1$.
- **D** can decrease in a similar manner or
- **D** can be "reshuffled", for example, $D_i \to D_i - 1\,, D_j \to D_j + 1$.

The third case can be recast in the first two, as it is a decrease and increase at the same time at two different lattice sites. The probability distribution describing **D** thus changes in time according to,

$$\partial_t P(\mathbf{D}, t) = L^+[P(\mathbf{D}, t)] + L^-[P(\mathbf{D}, t)]\,, \tag{1.8}$$

where $L^\pm$ are operators describing the creation or annihilation of a particle. To make a concrete simple example, if there is a spontaneous creation of a particle at any site with rate k^+ and annihilation of two particles with rate k^-, the master equation is,

$$\begin{aligned}\partial_t P(\mathbf{D}, t) = &\sum_i k^+ \left[P(\mathbf{D}, D_i - 1, t) - P(\mathbf{D}, t)\right] + \\ &\sum_i k^- \left[\frac{(D_i+2)(D_i+1)}{2} P(\mathbf{D}, D_i + 2, t) - \frac{D_i(D_i-1)}{2} P(\mathbf{D}, t)\right]\,,\end{aligned} \tag{1.9}$$

where $P(\mathbf{D}, D_i + n, t)$ stands for $P(D_1, D_2, \ldots, D_i + n, \ldots, D_N, t)$. Before proceeding with the path integral representation we introduce a Fock space, in which the probability distribution is formally written as,

$$|P(t)\rangle = \sum_{\mathbf{D}} P(\mathbf{D}, t) a_1^{\dagger D_1} \ldots a_N^{\dagger D_N} |0\rangle\,. \tag{1.10}$$

The operator $a_i^{\dagger D_i}$ is the creation operator and formally represents a creation event at a given site. D_i denotes the number of bound enzymes at site i. The creation and annihilation operators $a_i, a_i^\dagger$ act on the basis $|D\rangle$ as follows,

$$\begin{aligned} a_i^\dagger |D_i\rangle &= |D_i + 1\rangle\,, \\ a_i |D_i\rangle &= D_i |D_i - 1\rangle\,, \end{aligned} \tag{1.11}$$

and they follow standard commutation rules, $[a_i, a_i^\dagger] = 1$. Using this notation we can formally rewrite the master equation (1.8) in terms of the operators, $a_i^\dagger, a_i$,

$$\partial_t |P(t)\rangle = -H |P(t)\rangle\,, \tag{1.12}$$

where H is determined by the specific processes taken into consideration. In particular, H is written only in terms of the creation and annihilation operators. In the specific

example of Eq. (1.9), $H = \sum_i \left[k^+(1-a_i^\dagger) + k^-(1-a_i^{\dagger\,2})a_i^2/2\right]$. The expectation value of any observable $O(\mathbf{D},t)$ by definition is

$$\langle O(\mathbf{D},t)\rangle = \sum_{\mathbf{D}} O(\mathbf{D})P(\mathbf{D},t)\,. \tag{1.13}$$

After some algebra (Appendix A) we arrive at a closed expression in terms of H as a path integral,

$$\langle O(\mathbf{D})\rangle = \int \mathcal{D}[\phi]\mathcal{D}[\hat{\phi}]O(\boldsymbol{\phi},\hat{\boldsymbol{\phi}}=1)e^{-S[\hat{\phi},\phi]}\,, \tag{1.14}$$

with

$$S[\hat{\phi},\phi] = -\sum_i \phi_i(t_f) + \int_0^{t_f} \mathrm{d}t \sum_i \left(\hat{\phi}_i(t)\partial_t\phi_i(t) + H_i[\hat{\phi},\phi]\right)\,, \tag{1.15}$$

In H the following replacements are made due to the coherent state formulation: $a_i^\dagger \to \hat{\phi}_i\,, a_i \to \phi_i$, $\boldsymbol{\phi} = (\phi_1 \dots \phi_N)$ and similarly for $\hat{\boldsymbol{\phi}}$. Every observable $O(\mathbf{D})$ has first to be written in terms of creation and annihilation operators $a^\dagger, a$, which are then replaced with the fields $\phi, \hat{\phi}$ and finally the conjugate field $\hat{\phi}$ must be set to one [8]. If we compare equations (1.14),(1.6) we realise that they have a very similar form, which is extremely convenient as it overcomes several problems. First of all, starting with two completely different approaches from different physical scenarios, we arrive to a common expression that we can treat with the same theoretical tools, such as renormalization group. The Langevin description is a an approximation for a continuous field of its dynamics, whilst the master equation is an exact representation of all the possible microscopic processes. Either way, we realise that physical and biological observables may still, under certain condition, be described with the same theoretical tools... Is that not fascinating? Before getting too excited, we have to be careful in the interpretations. The fields ϕ, $\hat{\phi}$ in both representations still have an unclear physical meaning. The field ϕ in the MSRJD path integral is directly related to the real field, whilst in the Doi-Peliti it is only related via noise averages. By looking at a path integral, we would need to know the original problem (master or Langevin equation) before getting into any computation. In this thesis we will not further comment on different interpretations, as they are fully discussed in several textbooks [7, 8, 68], we just mention that $\tilde{\phi}$ is related in both formalism to response functions. Our main goal is to use these theoretical tools to predict genomic and physical observables, such that we will be mostly interested in computing noises averages described in terms of the fields ϕ. In particular one observable that we will face throughout the thesis is connected correlation functions, which we will discuss in the next section. Connected correlation functions describe how fields are correlated in space or in time, such that they encode information of the spatio-temporal dynamics.

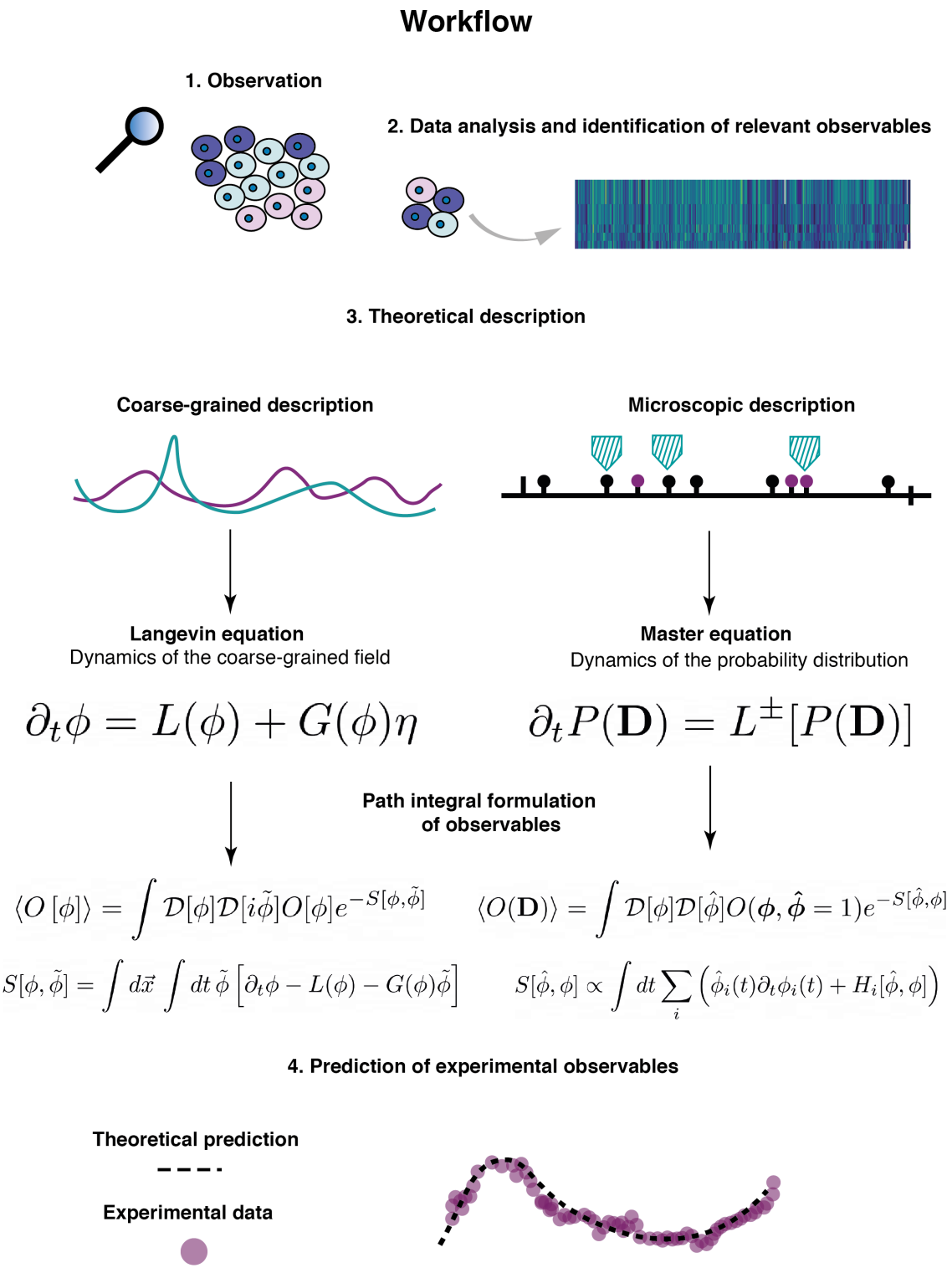


Figure 1.9.: Highlights of the main theoretical methods applied to genomics in this thesis. After observation, the data is analysed and the main relevant observables are identified. The theoretical description is derived in terms of a coarse-grained field or from microscopic processes. Different descriptions lead to different path integral formulations, which are then used to infer and predict experimental data.

Cumulant generating functional for correlation functions

In the previous sections we were able to write every observable of a stochastic process in a path integral formalism. Specifically, correlation functions $C(\vec{x},\vec{y},t,t')$ of a field $\phi(\vec{x},t)$ are defined as,

$$\begin{aligned} C(\vec{x},\vec{y},t,t') &= \langle\phi(\vec{x},t)\phi(\vec{y},t)\rangle \\ C_c(\vec{x},\vec{y},t,t') &= \langle\phi(\vec{x},t)\phi(\vec{y},t)\rangle - \langle\phi(\vec{x},t)\rangle\langle\phi(\vec{y},t')\rangle \doteq \langle\phi(\vec{x},t)\phi(\vec{y},t)\rangle_c\,, \end{aligned} \tag{1.16}$$

C_c are connected correlations and they describe correlations of the fields independently of changes in their averages. The latter ones are the correlations we are interested as they are essential, as we will see, to describe processes where the mean does not have a clear physical meaning. As these are observables of a generic process, we can write them in the form Eq. (1.14),(1.6). We use the same notation as for the MSRJD ($\tilde{\phi}$), but the same concepts apply to the Doi-Peliti path integral. Specifically, we define a generating functional of correlations, $Z[\mathbf{h},\phi,\tilde{\phi}]$, as

$$Z[\mathbf{h},\phi,\tilde{\phi}] = \frac{\int \mathcal{D}[\phi,\tilde{\phi}] e^{-S[\phi,\tilde{\phi}]+\int d\vec{x}\int_0^{tf} dt\left[h(\vec{x},t)\phi(\vec{x},t)+\tilde{h}(\vec{x},t)\tilde{\phi}(\vec{x},t)\right]}}{\int \mathcal{D}[\phi,\tilde{\phi}] e^{-S[\phi,\tilde{\phi}]}}\,. \tag{1.17}$$

We introduced auxiliary fields $\mathbf{h}=(h,\tilde{h})$, which will eventually be set to zero as they are not originally present in the path integrals. Path integrals are though essential to simplify the computation of correlation functions. Connected correlation functions can then be derived by taking functional derivatives of the logarithm of generating function (cumulant generating functional) as,

$$\langle\phi(\vec{x},t)\phi(\vec{y},t')\rangle_c = \frac{\delta^2}{\delta h(\vec{x},t)\delta h(\vec{y},t)}\ln(Z[\mathbf{h},\phi,\tilde{\phi}])|_{\mathbf{h}=0}\,. \tag{1.18}$$

It can be easily checked that Eq. (1.18) describes physical correlation functions even for the Doi-Peliti path integral [75]. The actions $S[\phi,\tilde{\phi}]$ or $S[\phi,\hat{\phi}]$ are arbitrary complicated functional of the fields. There are often cases, where the actions contain terms that are quadratic in the fields. We will refer to the quadratic terms of the action as the bare action S_0. Let us consider the case in which the action can be split into a quadratic part, which we will refer to as Gaussian or bare and a second part, which contains all the terms above the second order in the fields, $S = S_0 + S_I$. If we neglect for one moment the S_I we can evaluate the generating functional. Upon first transforming the field theory into Fourier space $\vec{x}\to\vec{q}$ and $t\to\omega$ as it is easier to compute correlation functions in this space, the generating functional is compactly written as,

$$Z[\mathbf{h},\phi,\tilde{\phi}] \propto \int \mathcal{D}[\phi,\tilde{\phi},\mathbf{h}] \exp\left[\int_{\vec{q},\omega} -\frac{1}{2}\begin{pmatrix}\tilde{\phi} & \phi\end{pmatrix}\hat{S}_0\begin{pmatrix}\tilde{\phi}\\ \phi\end{pmatrix} + \begin{pmatrix}\tilde{h} & h\end{pmatrix}\begin{pmatrix}\tilde{\phi}\\ \phi\end{pmatrix}\right], \tag{1.19}$$

where $\hat{S}_0$ is a matrix containing the operators in Fourier space associated to the dynamics and to the second orders terms in the fields. We are left with a Gaussian functional which leads to,

$$Z[\mathbf{h}] = \exp\left[\int_{\vec{q},\omega} \left(\tilde{h}\, h\right) \hat{S}_0^{-1} \begin{pmatrix} \tilde{h} \\ h \end{pmatrix}\right] . \tag{1.20}$$

It is straightforward to take functional derivatives and find connected correlation functions. Many interesting physical processes are not often formalised in a way such that the action is purely Gaussian. In these cases we need to find a way to deal with higher order terms. We start from a simple didactic case for in order to present how non-linearities can be treated in stochastic systems expressed as path integrals. Specifically, we take Eq. (1.1) with $L(\phi(\vec{x}), t) = a\phi(\vec{x}) + b\phi(\vec{x})^3 + D\partial^2_{\vec{x}}\phi(\vec{x}, t)$ and $G(\phi(\vec{x}), t) = \sqrt{D}$. The only term that will not be quadratic is the cubic term, whilst the quadratic terms in a, b (quadratic after multiplying by $\tilde{\phi}(\vec{x}, t)$) will be part of the bare action, such that bare connected correlations $C_0(\vec{x}, \vec{y}, t, t')$ are in Fourier space,

$$C_0(\vec{q}, \omega) = \frac{2D}{\omega^2 + D^2(a + \vec{q}^2)^2} . \tag{1.21}$$

We omit from now on the lower index c to indicate connected correlations. The only non-linear term in ϕ^3 must be evaluated. If the parameter b is small enough, we may expect that this term is subleading with respect to the quadratic terms, such that we can expand the action as,

$$e^{-S[\phi,\tilde{\phi}]} \simeq e^{-S_0}\left[1 - S_I + \frac{1}{2}S_I^2 + O(S_I^3)\right] , \tag{1.22}$$

with $S_I = b \int d\vec{x} \int dt\, \phi(\vec{x}, t)^3$. All the terms in the previous equation are then higher order moments over the Gaussian distribution e^{-S_0} and they can be computed by means of Feynman diagrams as we will see. We then developed a procedure to consistently compute correlation functions from any field theory upon writing the path integral formulation. We thus have a powerful consequence: as long as we are able to write the biological process either as a master equation or as a Langevin equation, we have a workflow to compute every observable no matter how complicated the problem is. In Fig. 1.9 we outline this workflow schematically. It is clear that the computation of all the terms in Eq. (1.22) is extremely cumbersome and in the next section we present how renormalization group methods greatly simplify the task.

1.3.2. Renormalization group theory

Whenever an action is expanded as in Eq. (1.22), there is always a question that we need to address: to which order should we stop the expansion? Someone would say that

as physicist the second order is more than enough, but the first one is sometimes satisfactory as well. Unfortunately, these expansions are justified when they are controlled, meaning that we know which terms we are omitting and why we are omitting them. Let us consider for example that the non-quadratic terms in the action are the cubic one and a spatially dependent one, e.g. $S_I = b \int d\vec{x} \int dt\, \phi(\vec{x},t)^3 + c \int d\vec{x} \int dt\, \phi(\vec{x},t)^2 \partial_x^2 \phi(\vec{x},t) + \ldots$. We would first need to evaluate all this terms at linear order, then at the second order and so on. Without entering into details, it is clear that this would be a vast program and we will luckily have to find something different than the path integral formulation. Moreover, this expansion is not often well behaved or well defined [7]. Renormalization group theory does the job for us, by selecting which terms in the path integrals are relevant and quantifying how these terms affect connected correlation functions or more generally, critical exponents associated with a field theory. Renormalization has one requirement: the field theory has to be close to a critical point or has to be scale invariant. We have not yet defined what a critical point of a field theory is. For our purposes, a critical point is a point in the parameter space of the field theory where observables, such as correlation functions, become scale invariant. When a quantity is scale invariant, it is not anymore possible to define a length scale, or differently, the length scale is infinite. Even though, scale invariance or self-similarity may seem hard concepts, we have often faced them during our life: fractals are typical scale invariant objects, an example are the roman broccolis. By looking at their shape, we realise that whenever we zoom in a particular region of the broccoli, it repeats itself almost indefinitely, such that the zoomed part is a representation of the whole at a smaller scale. By looking at a smaller part of the object, we can in principle reconstruct the entire object by zooming out. Renormalization group does exactly the same procedure with field theories, it hence provides a way to connect microscopic theories with macroscopic description, taking into account fluctuations from all the length scales. Here we outline the main step of Renormalization group theory. Let's imagine that we observe a certain field theory at a given resolution a. We first decrease this resolution, it is the opposite of a microscope, we could call it a *macroscope*, arriving to a length scale $l \cdot a$, with $l > 1$. By doing so, we formally integrate out fluctuations that are smaller than the new resolution. The field is then renormalized as $\phi(x) \to \hat{\phi}(x)$. As we zoomed out, we need to rescale the length scales such that $x' = x/l$. As we changed the length scales, we need to recover the size of fluctuations such that $\phi'(x') = \hat{\phi}(x) l^{\xi}$, with ξ a yet unknown exponent. The action after an RG transformation is changed to: $S \underset{dl}{\to} S' = R[S]$, with R the renormalization procedure. If we make l small enough the action will perform a trajectory $S \to S' \to S'' \to \ldots$, known as *RG flow*. An RG fixed point is an attractor (S^*) of the RG flow, Fig. 1.10 A such that the action remain structurally invariant under further RG transformations.

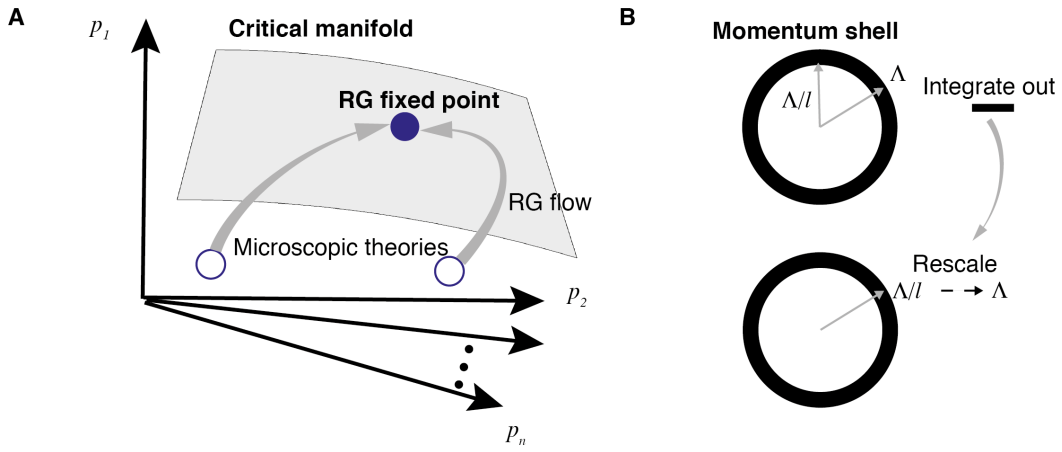


Figure 1.10.: (**A**) The RG fixed point is the attractor of different microscopic theories defined in terms of their parameters. (**B**) The momentum shell is divided into two parts: high and low momenta. The high momenta are integrated out and this step is repeated till eventually the microscopic theory reaches the RG fixed point.

In particular, different microscopic theories may lead to the same attractor under RG and that's where the usefulness of Renormalization group theories lies. Microscopic theories with different parameters and even different shapes of interactions, close to the critical point, have the same critical exponents. Field theories which share same critical exponents are grouped into *universality classes*.

Even though the outlined procedures might look like an hard task, Wilson developed a method, known as *momentum shell renormalization*, which will greatly simply the tasks we outlined in words [76]. In particular the idea of Wilson was that whenever we coarse grain the fields, we effectively integrate out degrees of freedom with wavelength shorter than the wavelength associated with the resolution we coarse grain to. We define a momentum shell with the momentum $k \in [0, \Lambda]$, where Λ is the maximum momentum naturally given by the lattice spacing of the theory. Momentum shell renormalization follows the same steps we outlined before but in Fourier space (Fig. 1.10 B). Specifically, we first

- integrate out momenta modes in $\Lambda/l < k < \Lambda$. These are the modes, smaller than our resolution (high momenta corresponds to short wavelengths),
- rescale distances $x' = x/l$ and so momenta $k' = k\,l$ and
- rescale fields $\phi'(x') = \phi(x)l^{\xi}$.

In summary, Wilson momentum shell renormalization group allows to relate different microscopic theories to the same macroscopic behaviour for different field theories close to a critical point. It is then a powerful technique that we will implement whenever we need to go infer macroscopic behaviour from microscopic measurements. In this

first part of the theoretical introductions we show a way to describe physical observable of field theories in path integral formulations and tools that allow to deal with non-linearities. Complex systems cannot often been described by the spatio-temporal dynamics of a field, but they often involved field theories interacting on multiple scales. In the remaining part of the introduction we introduce the concepts of disorder via spin glass theories, which will serve as a theoretical foundation to deal with complex networks of interactions, such as the gene regulatory networks.

1.3.3. Theories for disordered systems

Biological systems form complex networks where interactions between several components are governed by several and often unknown factors. In this section, we introduce concepts in spin glass theories that will be widely used in Chapters 4, 5 which deal with complex networks of interactions with inherited disorder. Disorder arises, for example, when interactions between different components of the system are uneven. Spin glass theories have the advantage of being minimal representation of disordered systems that we can study and can give us insights into emerging properties of such systems. In this thesis, we are mostly interested in the effect of disorder in complex systems, such as gene regulatory networks or multi-scale interacting systems. A minimal model for the role of disorder is the Sherrington-Kirkpatrick model [77], which exact solution was given by Parisi in [78]. Let us consider that the system is described by a set of possible states or variables $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_N)$, with N the system size. ξ_i may represent for example the expression of the gene i or the methylation status of a CpG site i. We consider for now that these variables are binary ($\xi_i = \pm 1$). We associate to every realization of ξ an energy of the form: $H = f(\boldsymbol{\xi})$, where H is the Hamiltonian and it is a generic function of the state of the system. A form of the Hamiltonian, that encodes minimal interactions, such that every sequence $\boldsymbol{\xi}$ is represented as

$$H = \sum_{i,j} J_{i,j} \xi_i \xi_j \,. \tag{1.23}$$

$J_{i,j}$ are for now arbitrary value, which represents interaction between state variables ($\boldsymbol{\xi}$). If $J_{i,j} = -J$ we recover the Curie-Weiss model or fully connected Ising model. The specific form of $J_{i,j}$ encodes the minimal energy states of the system. In particular, if $J_{i,j}$ are all negative, the energy is minimized whenever ξ_i and ξ_j have the same value. Whenever $J_{i,j}$ are all negative, the outcome is that the minimum of the energy is achieved by the sequence $\boldsymbol{\xi} = (1, \ldots, 1)$ or $\boldsymbol{\xi} = (-1, \ldots, -1)$. In technical terms, the ground state has a two-fold degeneracy as the minimal energy can be achieved in two ways. Even though we have not described yet what happens to systems described by the Hamiltonian (1.23) in contact to an heat bath, at this level of description they

seem trivial. To go one little step further, we consider a group of three isolated spins ($i = 1, 2, 3$) with symmetrical interactions ($J_{i,j} = J_{j,i}$) and $J_{1,2} = J_{1,3} = J_{2,3} = 1$. It can be checked that the minimum of the energy has an 8-fold degeneracy as eighth different sequences have the same energy, $\boldsymbol{\xi} = (1, 1, -1)$ or $\boldsymbol{\xi} = (-1, -1, 1)$ and so on. This property is referred to as frustration [79] and generally refers to constraints such that it is not possible to minimize all the bonds $J_{i,j}\xi_i\xi_j$ at the same time. We consider for now only a very small system $N = 3$, if we take $N \gg 1$ then the situation may get really complicated. In particular, imagine that system is in a given configuration ξ^*, which does not minimize the energy. Let us consider that we can only change a variable (ξ_i) at a time and accept this change only if the energy is lowered by the change. There might situations in which we cannot change any variable as the energy will increase, but the sequence ξ^* is not the the absolute minima of the energy. This is known as a metastable state, as roughly speaking, even though is not the minimum of the energy is minimal enough such that it is hard to change and get out of that. As mentioned, the thermodynamical behaviour might be very different, but we will discuss it in Chapter 4. Disorder is encoded into the random distribution of interactions and from now on, we take $J_{i,j}$ to Gaussian distributed with mean $\mu = 0$ and variance σ^2. The probability of observing a particular configuration, if the system is in a thermal bath, is given by the Boltzmann distribution,

$$P(\boldsymbol{\xi}) = e^{-\beta \sum_{i,j} J_{i,j}\xi_i\xi_j}/Z\,, \tag{1.24}$$

where Z is the normalization and $\beta = k_bT$ (k_b is the Boltzmann constant and T is the temperature). It is easy to check that if $J_{i,j}$ are distributed with $N(0, \sigma^2)$ the Hamiltonian is not extensive. We thus need to fix this issue and take $J_{i,j}$ distributed with variance $\sigma^2/\sqrt{N}$. Spin glasses are not different from standard statistical physics problems, as every thermodynamical quantity can be computed from the free energy F via the partition function Z,

$$F = -\lim_{N\to\infty} \frac{1}{\beta N} \log Z\ , Z = \sum_{\boldsymbol{\xi}} e^{-\beta \sum_{i,j} J_{i,j}\xi_i\xi_j}\,, \tag{1.25}$$

We immediately realise that there is a major problem, namely, we are computing the free energy just for a specific realisation of the couplings (disorder) $J_{i,j}$, whilst it will be way better to find it for the couplings distributed according to $N(0, \sigma^2/\sqrt{N})$. In this way, we can study all the systems with the couplings having the same variance at once, way more convenient! To do so, we take the average of the partition function $\overline{Z}$ over the distribution of couplings,

$$\overline{Z} = \int d\mathbf{J} P(\mathbf{J}) \sum_{\boldsymbol{\xi}} e^{-\beta \sum_{i,j} J_{i,j}\xi_i\xi_j}\,, \tag{1.26}$$

with $P(\mathbf{J})$ the distribution of $J_{i,j}$, Gaussian here, and $\sum_{\boldsymbol{\xi}}$ stands for the sum over $\xi_i = \pm 1, \forall i$. This is called *annealed* average and unfortunately will not give in general reasonable results. We will not enter more into details, but in order to have an intuitive understanding of the reason of such failure, in Eq.(1.26) we are averaging over the couplings and sum over the spins at the same time. We are then considering the couplings to be equally "important" as the variables, in other words, we let the couplings change with the variables, whilst we would like to keep them fixed and perform the average later. In order to solve this issue, we can average the free energy over the couplings rather than the partition function,

$$\overline{F} = -\lim_{N\to\infty} \frac{1}{\beta N} \overline{\log Z} = -\lim_{N\to\infty} \frac{1}{\beta N} \lim_{n\to 0} \frac{1}{n} \log \overline{Z^n}\,, \tag{1.27}$$

where

$$\overline{Z^n} = \int d\mathbf{J} P(\mathbf{J}) \sum_{\boldsymbol{\xi^1}\dots\boldsymbol{\xi^n}} e^{-\beta\left[H(\boldsymbol{\xi^1})+\dots+H(\boldsymbol{\xi^n})\right]}\,. \tag{1.28}$$

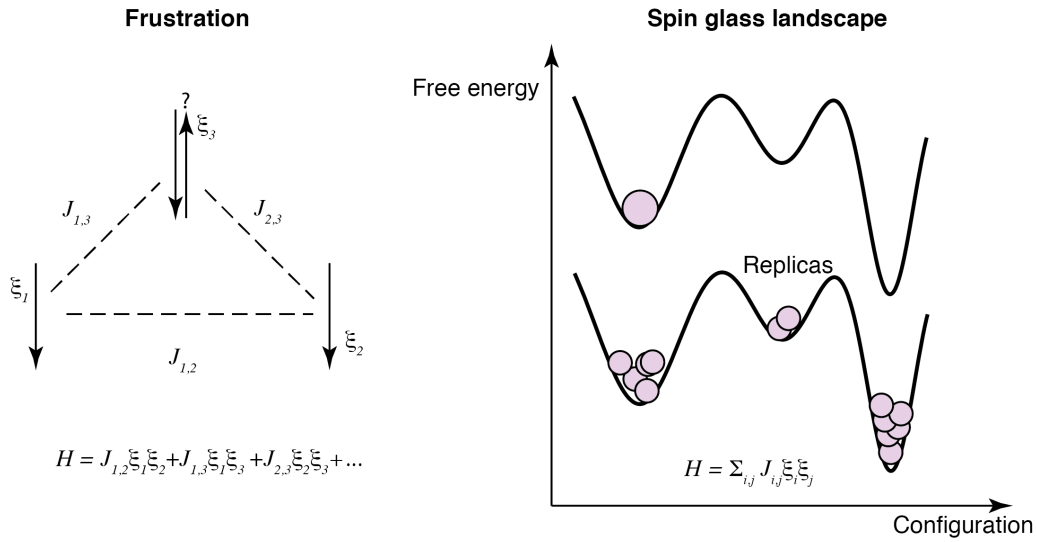


Figure 1.11.: (left) Simple example of frustration for a group of three spins in a spin glass model ($J_{1,2} = J_{1,3} = J_{2,3} = 1$). Once the two spins are oriented in order to minimize the energy, the bonds between them and the third spin (ξ_3) are frustrated as they do not minimize the energy irrespective of its sign. (right) Schematic view of the spin glass free energy landscape in terms of replicas (pink circles) where the couplings $J_{i,j}$ of the Hamiltonian H are quenched Gaussian random variables.

In the last step we made use of the replica trick [80], rather than computing an average over the log of the partition function we just compute the average over n times the partition function (referred to as replicated). This technique greatly simplify the computation of the free energy, which will be cumbersome otherwise. Moreover, as we will see in later chapters, there is a strong connection between replicas and actual

physical or biological systems. In particular, the free energy (*quenched*) is given by

$$\overline{F} = -\lim_{N\to\infty}\lim_{n\to 0}\frac{1}{\beta N n}\log\left[\int d\mathbf{J}P(\mathbf{J})\sum_{\boldsymbol{\xi^1}\dots\boldsymbol{\xi^n}} e^{-\beta\left[H(\boldsymbol{\xi^1})+\dots+H(\boldsymbol{\xi^n})\right]}\right]. \tag{1.29}$$

We started with the study of the thermodynamic equilibrium of one system and we arrived to the computation of the thermodynamic of n uncoupled identical systems. This is pretty weird, as we might expect that computing n times the partition function shouldn't give more insights rather than computing it once. Actually this is not the case. As we outlined in this section, the minima of the free energy are multiple in glassy systems, and depending on the initial conditions or other external factors, the system may remain trapped in one of them. Intuitively if we replicate the system n times, some copies may end up in one minima, some in other, such that replicated the partition functions, actually may be used as a way to study the free energy landscape in its full glory. Yes... we need then to take the limit $n \to 0$, but this is another story and we will not discuss it here [79]. There are more concepts in spin glass theories that we will analyse in the remaining part of the thesis. From now on, whenever we refer to free energy in spin glass or disordered systems, we will adopt the definition given by Eq.(1.29).

1.4. Overview of the thesis

In this thesis we overcome conceptual limitations of genomics and sequencing using methods from nonequilibrium statistical physics. By developing novel frameworks in non equilbrium systems, we are able to predict scaling behaviour for long range interacting particle systems and asymmetric spin glasses, which are in agreement with numerical and experimental results. Specifically, in Chapter 2 we infer chromatin structures in the three dimensional space of the nucleus upon developing a theoretical framework based on one dimensional multi-omics genomic data profiling DNA methylation,chromatin accessibility and gene expression. Our theoretical framework describes DNAme kinetics in terms of a master equation that is mapped to a quantum problem with hard-bosons. Upon using renormalization group and path integral methods, we show that the experimentally observed scaling behaviour of connected correlation functions and average DNAme in the mouse embryo, both *in vitro* and *in vivo*, is correctly predicted by our theory. At the end of the chapter and in Chapter 3 we challenge the theory to infer the relationship between DNA methylation and gene expression for cells exiting from pluripotency. We then extend the results for systems with long range interactions to set a theoretical background for synchronization phenomena found during the first cell fate decisions. As gene expression is also regulated via gene regulatory

networks, in Chapter 4 we construct a master equation describing the effect of different molecular processes in those networks. Specifically, by starting from a master equation we map the dynamics of fluctuations to an asymmetric bipartite spin glass. We show that these systems exhibit both a static and a dynamical phase transition between a phase where fluctuations are uncorrelated to a glassy phase with non trivial correlations. Relying on single-cell sequencing data we show that cells may lie in a glassy phase and transitions between cell states are regulated via correlations of fluctuations which emerge in this phase. At the end of this thesis, Chapter 5, we extend the theoretical results found throughout the thesis to derive a minimal, yet general theory of multi scale interacting complex systems. The theoretical results are general, such that they can be applied to complex systems ranging from embryonic development to ecosystems and social systems. Finally, in Conclusions and future perspectives we summarize in detail all the findings of this thesis and we give possible future research directions inspired by this thesis.

2. From Sequence to Space and Time in Single-Cell Genomics

2.1. Introduction

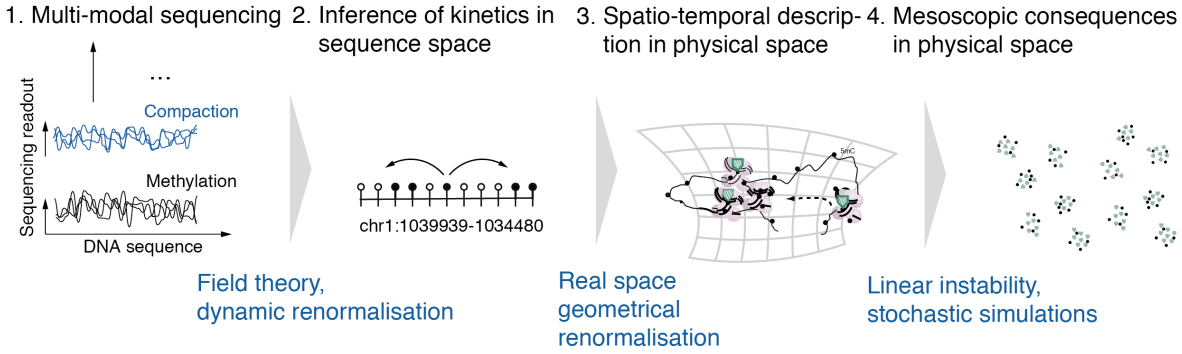


Figure 2.1.: Highlights of the main theoretical and computational steps used to infer *de novo* DNA methylation dynamics and mesoscopic structures from single-cell sequencing experiments.

Recent breakthroughs in single-cell sequencing allows us to probe molecular mechanisms with unprecedented details. As we discussed in Sec. 1.2.3, BS-Seq can be used to measure CpG methylation. This sequencing technique gives as an outcome the methylation status of almost every CpG of the DNA (~ 70 %). This results in an huge and comprehensive amount of data and information. Every time such immense data set are built, there is always a question that puzzles us: What can we learn from them? It is clear that if we want to find our personal name written on the DNA with a methylation "alphabet", we will probably succeed. What we measure is a one dimensional (sequence space) string of bits of information (methylated or not), but we know that processes in the cell happens on the three dimensional space of the nucleus (physical space). Importantly, these processes cannot currently be inferred by conventional computational methods for the following reasons. First, it requires solving a difficult "inverse" problem which involves mapping given sequencing profiles to one of an infinite number of processes in space and time. Solving this problem computationally involves probing a large number of such processes for consistency with the sequencing data. As the simulation of these processes involves unspecified, non-local interactions and therefore is

computationally very costly, solving this inverse problem using conventional tools is not feasible. Secondly, emergent properties of interacting complex systems do usually not obey the rules that act on its constituents (emergence) [81] the spatio-temporal processes underlying biological function cannot straightforwardly be inferred from detailed molecular measurements as provided by single-cell genomics. In this chapter, by applying methods from nonequilbrium physics to single-cell genomics, we develop a unique and general theoretical framework to connect detailed molecular measurements in single-cell genomics to emergent phenomena in space and time. Specifically, we use tools such as renormalization group and stochastic field theories to show that emergent phenomena in physical space, such as phase separation, can be unveiled from single-cell epigenome sequencing along the one-dimensional DNA sequence. We demonstrate this approach by revealing the interplay between the establishment of epigenetic marks (*de novo* DNA methylation) during early mouse development and nanoscale topological changes in chromatin structure. In particular, in Sec. 2.2 we show how statistical relevant quantities can be inferred from single-cell and BS-Seq. Based on these findings in Sec. 2.3,2.4 we will develop a theoretical framework to infer collective epigenetic mechanisms in early development aimed at understanding and predicting experimental data. In Sec. 2.5 we develop a tool to infer dynamics in the cell nucleus from sequencing data. Finally, in Sec. 2.6 we challenge the theoretical results by predicting experimental results and in Sec. 2.7 we apply these methods to *in vivo* mouse embryo data showing the effect of methylation patterns on symmetry breaking of cells exiting from pluripotency [82]. All these steps are outlined in Fig. 2.1. Finally, in Sec. 2.8 we extend the theoretical framework to account for methylation changes during adulthood and ageing.

2.2. Analysis of sequencing data of DNA methylation

As we outlined in the introduction (Sec. 1.2.1), upon exit from pluripotency which occurs around implantation of the embryo, upregulation of the *de novo* methyltransferase genes Dnmt3a/b leads to massive and rapid *de novo* DNA methylation to a genome average of 80% per CpG Fig. 2.2 A [25].

In order to study how DNAme is established during early development, our collaborators in the Wolf Reik laboratory cultured mouse embryonic stem cells (mESCs) in 2i culture conditions, where cells assume a naïve pluripotent state and DNA methylation is globally low. Cells were then released into serum conditions Fig. 2.2 B, where Dnmt3a/b genes are upregulated [32], meaning that DNMT3 enzymes are in high abundance in the cell nucleus. The transition from 2i to serum conditions has been shown to recapitulate the epigenetic and transcriptional changes occurring during transition

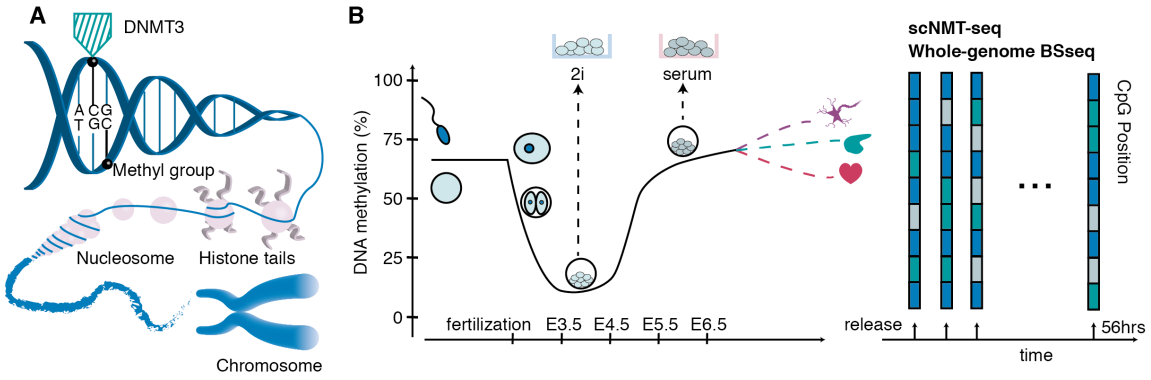


Figure 2.2.: (**A**) Establishment of DNA methylation by DNMT3 enzymes. Methylated CpG are highlighted in black. (**B**) Schematic sketch of DNA methylation dynamics during early mouse development and illustration of 2i-release experiments and sequencing data.

to formative and primed pluripotency in the embryo [83].

After release into serum conditions we performed two complementary sets of experiments (Fig. 2.2 B). Firsty, we performed a *bulk* whole-genome bisulfite-sequencing (BS-seq) time course of 31 time points over a period of 56 hours giving access to high-coverage information with high temporal resolution and secondly we performed a single-cell NMT-sequencing (scNMT-Seq) experiment of 288 mESCs with lower temporal resolution (0h, 24h and 48h). scNMT-Seq is a multi-omics sequencing experiment which enables joint profiling of the genomic distribution of DNA methylation marks, DNA accessibility (GpC methylation) and the transcriptome (gene expression) at a single-cell resolution [16].

We initially focus on the first experiment and we use the second experiment to further verify of our findings. As *de novo* DNA methylation is associated with a gain in average methylation, from roughly 15 % to 80 % (the percentage is always in number of methylated CpG divided by total number of CpGs), we calculated the increase of average methylation (methylation density) across the whole genome, Sec. 1.2.3 (technical details in Appendix B.1). In Fig. 2.3 we show a subset of the data showing the global increase of DNAme over a wide genomic region.

As the establishment of methylation marks is known to be regulated locally by a number of different factors, including CpG density [84], transcription and histone modifications [85], the global increase of methylation density may be not informative as there is a risk to average over totally distinct functional genomic domains. Basically, we should avoid to mix pizza with pineapple. We then computed changes in DNA methylation densities along many of such distinct genomic features (promoters, gene bodies, etc...). We found that functionally distinct genomic regions acquired average DNA methylation levels at different rates Fig. 2.4 A

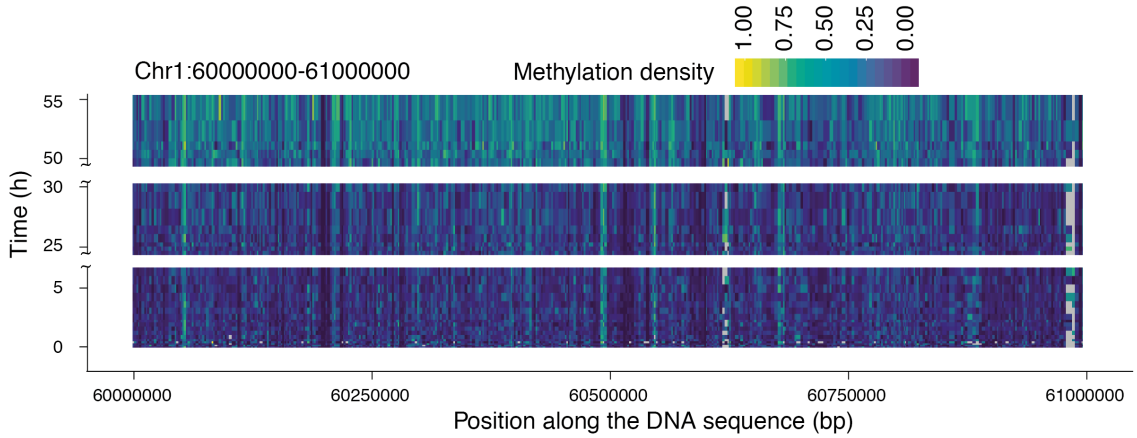


Figure 2.3.: Methylation density changes along a portion of chromosome 1 of mESCs for 56 hours after release into 2i from serum conditions.

We then asked whether all the curves Fig. 2.4 A have the same functional form and we made the following scaling ansatz,

$$\langle m \rangle_{g,\rho_{CpG}} = a_{g,\rho_{CpG}} + f(\mathbf{b}_{\mathbf{g},\rho_{\mathbf{CpG}}}, t)\,, \tag{2.1}$$

with $\langle m \rangle_{g,\rho_{CpG}}$ the methylation density for the genomic feature g with CpG density ρ_{CpG}. $a_{g,\rho_{CpG}}$ is the initial base value after erasure of paternal and maternal DNA methylation and $f(\mathbf{b}_{\mathbf{g},\rho_{\mathbf{CpG}}}, t)$ is a generic function which may depend on a set of parameters $\mathbf{b}_{\mathbf{g},\rho_{\mathbf{CpG}}}$. We then looked for the simplest possible model which is in accordance with experimental data. We were able to rescale time (Appendix B.1) for each time series in such a way that all curves collapsed onto a single curve Fig. 2.4 B, a phenomenon referred to as scaling behaviour.

The surprising emergence of scaling suggests that there is one generic mechanism of how DNA methylation is established genome-wide. This implies that, apart from a rate and an initial base value, the mechanisms is the same in every genomic region and it is independent of the CpG density. Notably, the simplest model gives a power law with an exponent of 5/2, $f(t) = b_{\rho_{CpG},d}\tau^{5/2}$, with τ being the rescaled time. We are in a position to write down a very compact and generic form of how the methylation density $\langle \tilde{m} \rangle$, upon subtracting the base value a and rescaling by b, changes with respect to the rescaled time,

$$\langle \tilde{m} \rangle = \tau^{5/2}\,. \tag{2.2}$$

The time evolution of average DNA methylation levels therefore is scale-invariant, i.e. its mathematical form does not change on time intervals of different lengths (self-similarity). Temporal scale-invariance and scaling behaviour with a specific exponent

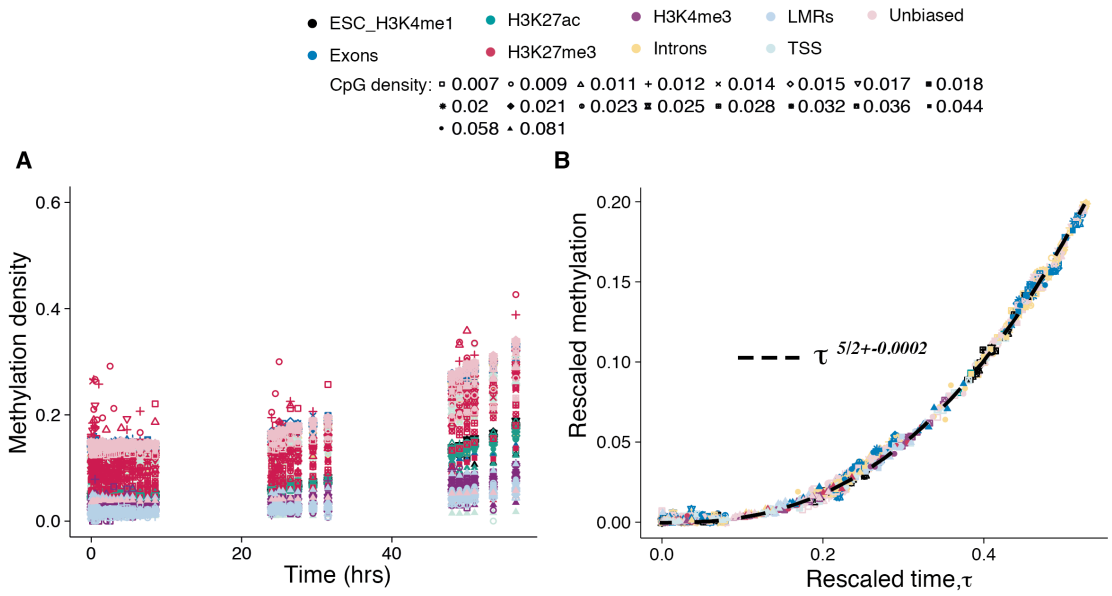


Figure 2.4.: (**A**) Methylation density changes during 56hrs time-course of different genomic regions (colors), further subdivided by their CpG density (shapes). CpG densities bins are defined using quantiles of the distribution of densities over all the fragments. Error bars are smaller than the size of the points. (**B**) Scaling analysis of the curves in (**A**) allow to collapse them into a single master curve which follows a power law with exponent 5/2.

of 5/2 are a signature of collective, self-organisation processes, suggesting that DNA methylation marks are established via a collective mechanism involving interacting DNMT3 enzymes [25, 86]. Whenever this interactions connect different genomic loci, this results in spatial coordination of *de novo* DNAme along the genome [87, 88].

Dynamical quantities, such as the increase of the average methylation with respect to time are often not sufficient to rule out independence between methylation binding events. In order to strengthen the hypothesis that there is a collective process underneath *de novo* DNAme, we computed equal time connected two point correlation functions (Appendix B.1), as they provide a way to infer the potential spatial coordination between DNA methylation marks. Connected correlations functions are defined as,

$$\langle m_i m_j \rangle_c = \langle m_i m_j \rangle - \langle m_i \rangle \langle m_j \rangle \,, \tag{2.3}$$

with i the genomic position of a CpG site and m_i its methylation density. The average in Eq. (2.3) is performed over different cells at the same time point and over pairs of CpGs at a distance $|j-i|$. We focus now on genome wide correlations and we will later derive correlations for specific genomic regions. The main reason for this choice relies on the specific data we are looking at. Bulk BS-Seq experiment have high temporal resolution (56 hrs, sampled almost hourly, with some sleeping), but not high spatial resolution as we do not have information for single CpGs. This definition of corre-

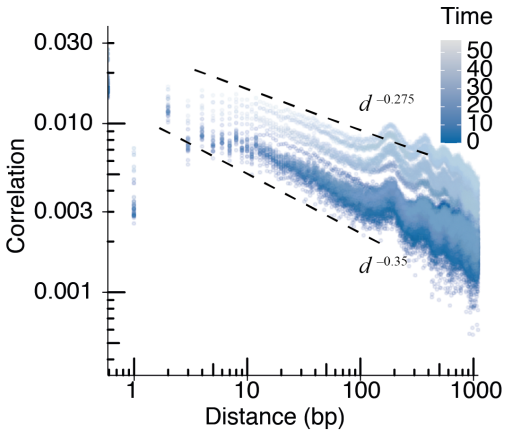


Figure 2.5.: Connected two-point correlation functions between DNA methylation marks at different times during the 2i-release experiment (double logarithmic plot). Oscillations are caused by nucleosomes positioning at $\approx 150 - 200$ bps. Dashed black lines are power laws with exponents fitted with a power law model. The exponents, which determine how correlations between DNA methylation marks decreases with respect to their distance in sequence space, range from -0.35 for cells sequenced directly after release to -0.275 for cells sequenced 56 hours later.

lations allows to infer correlations between methylation marks, independently of the average. As has been observed previously [89, 90], we found that in addition to weak oscillations reflecting nucleosome positions, these correlations follow a power law over several orders of magnitude in sequence space Fig. 2.5. A power law decay of connected correlation functions is a signature of strong correlations of DNA methylation marks over distances of thousands of base pairs and which come about presumably by yet unknown interactions between DNMT3 binding events over extended genomic domains. The exponents changes with respect to time and so average methylation, ranging from -0.35 to -0.275, which for a reader trained in critical phenomena may be quite puzzling. In Sec. 2.6 we theoretically predict how different exponents are time dependent and how they are connected to non criticality and to the effective dimensionality of the system. Taken together, the emergence of scaling behaviour suggests that *de novo* DNA methylation is a collective phenomenon that is spatially coordinated along the genome.

We are now in a position to develop a theoretical framework to study *de novo* DNA methylation based on sequencing experiment. In the next section, using field theoretical methods we will infer the stochastic kinetics of *de novo* DNA methylation in the space defined by the one dimensional sequence of CpGs (sequence space) and we will predict both the increase of average methylation with respect to time as well as the shape of connected correlation functions.

2.3. Nonequilibrium theory of *de novo* DNA methylation

To reveal the biological mechanisms underlying the interactions between DNMT3 binding events we developed a theoretical approach that allows inferring collective processes in the physical space of the nucleus from measurements along the one-dimensional sequence of the DNA. In contrast to hypothesis-driven approaches typically used to model biological systems our framework deduces the kinetics from sequencing data. Briefly, we start by describing the stochastic kinetics of *de novo* DNA methylation in the sequence space via a master equation formulation. Upon mapping the master equation to an hard-boson like path integral, we will be able to predict scaling behaviour of connected correlation functions and the dynamic changes of the average DNA methylation. Later, we will employ a dynamic geometric mapping between distances in the sequence space and distances in a projected physical space to derive the kinetics in the three dimensional space of the nucleus. At the end, we will use of theoretical results to accurately predict kinetics, experimental correlation and cross-correlation functions *in vitro* and *in vivo*.

Specifically, we begin by a defining a physical and mathematical framework for general out-of-equilibrium stochastic enzymes kinetics incorporating:

- binding and unbinding of enzymes to the DNA,
- chemical modifications of the DNA and
- general and unknown interactions of enzymes along the DNA sequence.

In the context of *de novo* DNA methylation, the chemical modification is DNAme at CpG sites and the enzymes are the DNMT3a/b. In the following we will work in the context of *de novo* DNA methylation, but the theory can be applied to other epigenetic processes. We begin with the last of the previous terms of enzymes kinetics as it will turn out to be the only important to infer *de novo* DNA methylation dynamics and it is the most theoretically challenging. To this end define an interaction kernel with the less possible assumptions. In particular, we assume that DNMT3 enzymes, bind to a particular CpG site with a rate that depends on the positions of the other enzymes that are already bound in the vicinity of the site. Second, we make no distinctions between DNMT3a/b and for the rest of the thesis, unless specified otherwise, we will refer to them as DNMT3. These assumptions greatly simplify the theoretical analysis and will be justified at posteriori as our model will correctly predict emergent spatio-temporal statistics. We then consider a positive feedback and cooperation between DNMT3 enzymes. We initially restrict the model to interaction kernels that depends only to the distance between the possible binding site and closest already bound enzymes. This

choice will turn out to be sufficient to predict experimental observables such that we do not need to add further interactions. The binding rate at a given CpG site i is then given by the closest sites to the right and to the left with a bound enzymes on top Fig. 2.6.

.

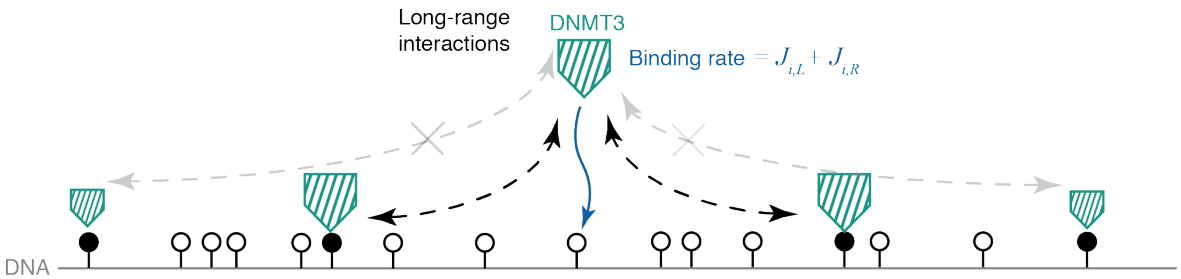


Figure 2.6.: Illustration of the first term of the master equation (2.4). CpG sites are depicted as circles, where an empty and a filled circle indicate a non methylated and a methylated CpG respectively. The black dashed lines indicate the interactions encoded in the kernel in Eq. (2.4), whilst the gray line are the neglected contributions to the kernel.

Specifically, if the nearest bound sites are at positions L and R we write the binding rate as $J_{i,L} + J_{i,R}$, where the two terms correspond to the contribution from the left and right bound neighbour, respectively. $J_{i,j}$ are in general functions of the distance between the binding sites, such that we write $J_{i,j} = J_{i,j}(|j - i|)$. Interaction with the closest occupied sites on the left and right effectively restrict these interactions in range [91, 92]. Omitting unbinding and demethylation terms for brevity, the time evolution of the probability of finding a given DNMT3 binding and DNA methylation profile, $P(\mathbf{D}, \mathbf{m}, t)$ follows a master equation of the form

$$\begin{aligned}\frac{\partial P(\mathbf{D},\mathbf{m},t)}{\partial t} =& \sum_{i=1}^{N}\sum_{l=i+1}^{N} J_{i,i+l}(l)\left(\prod_{j=1}^{l-1}\bar{D}_{i+j}\right) D_l[D_i P(\bar{\mathbf{D}}_i,\mathbf{m},t) - \bar{D}_i P(\mathbf{D},\mathbf{m},t)] \\ &+ \text{l.n.n.} \\ &+ \text{methylation and unbinding processes}\,,\end{aligned}$$

where $\mathbf{D}$ and $\mathbf{m}$ are a binary vectors describing DNMT3 occupancy and DNA methylation, respectively, such that, for example, $D_i = 1$ if site i is occupied and $D_i = 0$ otherwise. $\bar{\mathbf{D}}_i$ is the same vector where, at position i, D_i is replaced by $1 - D_i$ and $\bar{D}_i = 1 - D_i$. N is the number of lattice sites of the sequence space, such that $i = 1$ indicate the first CpG, $i = 2$ the second and so on, irrespective of the distance neighbouring CpG sites in base pairs (Fig. 2.6). Written in this form, the distance between CpGs is topological and not metric [93], but the distance in base pairs (metric) between two CpG sites can be easily included in the theory upon rescaling the interaction kernel

$J_{i,i\pm l}$. The interaction with the left nearest neighbors has the same form as the term shown and is abbreviated by "l.n.n". In this ansatz, we implicitly assumed that there are no further non-linearities or spatial correlations relevant on large enough length scales, for example from the catalytic activity (the deposition of methyl groups) of bound enzymes on the DNA or from de-methylation processes. We will show below that such additional processes are irrelevant under rescaling and we omit them in the following. We are then left with only binding kinetics and methylation rates by DNMT3 enzymes. In the context of *de novo* DNA methylation the non-linear nature of DNMT3 binding is supported by the literature [88].

In the following, as linear and uncorrelated processes do not contribute to the results obtained below, we will consider the marginal distribution

$$P(\mathbf{D},t) = \sum_{\mathbf{m}} P(\mathbf{D}|\mathbf{m},t)P(\mathbf{m},t)\,. \tag{2.4}$$

Following the definition of the model ansatz we have $P(\mathbf{D}|\mathbf{m},t) = P(\mathbf{D},t)$ such that the master equation for the marginal distribution has the same form as above,

$$\frac{\partial P(\mathbf{D},t)}{\partial t} = \sum_{i=1}^{N}\sum_{l=i+1}^{N} J_{i,i+l}(l)\left(\prod_{j=1}^{l-1}\bar{D}_{i+j}\right) D_l[D_i P(\bar{\mathbf{D}}_i,t) - \bar{D}_i P(\mathbf{D},t)] + \text{l.n.n.}\,. \tag{2.5}$$

The further complexity of this master equation (a part from long range interactions) is on the intrinsic sites exclusions. Indeed, a DNMT3 enzymes can bind only to free, i.e. non bound, CpG sites, such that standard Doi-Peilti path integral methods based on bosonic commutation rules cannot be applied. In particular, we need to take into account that a CpG site can be either occupied by a DNMT3 enzyme or not, such that multiple DNMT3s can not bind at the same site. We refer to the last constraint as site restriction and historically there are two ways to deal with it. The first one is based on writing the master equation in terms of fermionic operators [94], which by definition encode this constraint, but the resulting path-integrals are often very cumbersome and it is extremely hard, even for master equations simpler than the one we presented, to derive explicit analytical results. The second approach is based on adding the site restrictions with Dirac delta functions, allowing to work with bosonic operators [95]. In the following paragraph we will explain in details how to construct a path integral and how to derive a field theory for master equations with site restrictions

2.3.1. Path integral representation

In order to rigorously write down the path integral formulation of the master equation Eq. (2.4), we introduce a Fock space, Sec. 1.3.1, in which the probability distribution

is formally written as

$$|P(t)\rangle = \sum_{\mathbf{D}} P(\mathbf{D}, t) a_1^{\dagger D_1} ... a_N^{\dagger D_N} |0\rangle \,. \tag{2.6}$$

The operator $a_i^{\dagger D_i}$ is the creation operator and formally represents a binding event at a given site. D_i denotes the number of bound enzymes at site i. The creation and annihilation operators $a_i, a_i^\dagger$ act on the basis $|D\rangle$ as

$$\begin{aligned} a_i^\dagger |D_i\rangle &= |D_i + 1\rangle \,, \\ a_i |D_i\rangle &= D_i |D_i - 1\rangle \,, \end{aligned} \tag{2.7}$$

and they follow standard commutation rules $[a_i, a_i^\dagger] = 1$. Using this notation we can formally rewrite the master equation in terms of the creation operators, $a_i^\dagger$

$$\partial_t |P(t)\rangle = -H |P(t)\rangle \,, \tag{2.8}$$

where H is from Eq. (2.5)

$$H = -\sum_{i=1}^{N} \sum_{l=1}^{N-i} \prod_{j=1}^{l-1} J_{i,i+l}(l) \hat{\delta}_{D_{i+j},0} \hat{\delta}_{D_{i+l},1} [a_i^\dagger \hat{\delta}_{D_i,0} - \hat{\delta}_{D_i,0}] + l.n.n.. \tag{2.9}$$

The terms $\hat{\delta}_{D_i,0}$ are equal to 1 if an enzyme is not present at the CpG site i and 0 otherwise and restrict binding to a single enzyme per site. Before proceeding further, we need to understand how the $\hat{\delta}$ operator acts in the Fock space. To do so, we introduce a coherent state basis (Section 1.3.1) for which the identity is given by

$$1 = \int d\phi d\hat{\phi} e^{-\hat{\phi}\phi} e^{\phi a^\dagger} |0\rangle \langle 0| e^{\hat{\phi} a} \,. \tag{2.10}$$

In this new basis, following the rules of [95], the $\hat{\delta}$ operator acts on the coherent state basis as

$$\begin{aligned} \langle \phi | a^\dagger \hat{\delta}_{\hat{n},m} |\phi\rangle &= \frac{1}{m!} \hat{\phi} \left(\hat{\phi}\phi\right)^m e^{-\phi\hat{\phi}} \,, \\ \langle \phi | a \hat{\delta}_{\hat{n},m} |\phi\rangle &= \frac{1}{(m-1)!} \phi \left(\hat{\phi}\phi\right)^{m-1} e^{-\phi\hat{\phi}} \,, \\ \langle \phi | \hat{\delta}_{\hat{n},m} |\phi\rangle &= \frac{1}{m!} \left(\hat{\phi}\phi\right)^m e^{-\phi\hat{\phi}} \,. \end{aligned} \tag{2.11}$$

We now express every observable $A(\mathbf{D})$ as a path integral of the form (Section 1.3.1),

$$A(\mathbf{D}) = \int \mathcal{D}[\phi] \mathcal{D}[\hat{\phi}] A(\boldsymbol{\phi}, \hat{\boldsymbol{\phi}} = 1) e^{-S[\hat{\phi},\phi]} \,, \tag{2.12}$$

with

$$S[\hat{\phi},\phi] = -\sum_i \phi_i(t_f) + \int_0^{t_f} \mathrm{d}t \sum_i \left(\hat{\phi}_i(t)\partial_t\phi_i(t) + H_i[\hat{\phi},\phi]\right) , \tag{2.13}$$

and where we performed a partial integration in time. The Hamiltonian in this action reads

$$H_i[\hat{\phi},\phi] = (1-\hat{\phi}_i)e^{-\hat{\phi}_i\phi_i}\left[\sum_{l=1}^{N-i} J_{i,i+l}(l)\hat{\phi}_{i+l}\phi_{i+l}e^{-\sum_j \hat{\phi}_{i+j}\phi_{i+j}} + \sum_{l=1}^{i} J_{i,i-l}(l)\hat{\phi}_{i-l}\phi_{i-l}e^{-\sum_j \hat{\phi}_{i-j}\phi_{i-j}}\right] . \tag{2.14}$$

Defining the generating functional of correlations $Z[\mathbf{h},\phi,\bar{\phi}]$ as

$$Z[\mathbf{h},\phi,\hat{\phi}] = \int \mathcal{D}[\mathbf{h},\phi,\hat{\phi}]e^{-S[\phi,\hat{\phi}]+\int_0^y ds\int_0^{t_f} dt\left[h(s,t)\phi(s,t)+\bar{h}(s,t)\hat{\phi}(s,t)\right]} , \tag{2.15}$$

expectation values of products of observables, such as correlation functions, can be expressed as functional derivatives with respect to the auxiliary external field $\mathbf{h} = (h(s,t),\bar{h}(s,t))$,

$$\begin{aligned}\langle\phi(s,t)\rangle &= \frac{\delta}{\delta h(s,t)}Z[\mathbf{h},\phi,\hat{\phi}]|_{\mathbf{h}=0} ,\\ \langle\phi(s,t)\phi(y,t')\rangle &= \frac{\delta^2}{\delta h(s,t)\delta h(y,t)}Z[\mathbf{h},\phi,\hat{\phi}]|_{\mathbf{h}=0} .\end{aligned} \tag{2.16}$$

We have to be careful as $\phi(s,t)$ is not $D(s,t)$, i.e. the density of enzymes at position s. $\phi(s,t)$ is for now just the field of the coherent state basis. Luckily, it is easy to show that $\langle D(s,t)\rangle = \langle\phi(s,t)\rangle$, and similarly for correlations [75], which makes the use of field theoretical methods very usefull in the context of master equations for nonequlibrium systems. We are now in a position to derive the time evolution of the average density of bound enzymes and later of methylation density.

2.3.2. Semiclassical solution of the path integral

In this section we infer the functional form of the interaction kernel of enzyme binding events from a semiclassical solution of the field theory in order to derive moments equations for $\langle D(s,t)\rangle, \langle m(s,t)\rangle$. In a first step, we rewrite the Hamiltonian Eq. (2.14) in continuous space, i.e. Riemann integration. Upon introducing a spatial discretisation $\sum_i \Delta s \to \int ds$ the Hamiltonian in the action (2.14) is given by

$$\begin{aligned}H[\hat{\phi}(s),\phi(s)] = J(1-\hat{\phi}(s))e^{-\phi\hat{\phi}}\Big[&\int_0^{N-s}\mathrm{d}y\frac{\hat{\phi}(s+y)\phi(s+y)}{y^\lambda}e^{-\int_{z=0}^{y}\mathrm{d}z\hat{\phi}(s+z)\phi(s+z)}\\ &+\int_0^{s}\mathrm{d}y\frac{\hat{\phi}(s-y)\phi(s-y)}{y^\lambda}e^{-\int_{z=0}^{y}\mathrm{d}z\hat{\phi}(s-z)\phi(s-z)}\Big] ,\end{aligned} \tag{2.17}$$

where we specified the binding kernel for a long-range process as $J_{i,i\pm}(l) = \frac{J}{l^\lambda}$ and took the binding rates equal throughout the genome. The choice of the kernel is a minimal choice for a long-range kernel of interactions. We will later show that this choice is sufficient for a theoretical prediction of experimental observables. The integrals in Eq. (2.17) are not solvable in general and we need to develop a method to deal with them. We can then look for a clever Taylor expansion of the functions of the fields. These results will turn out to be correct for the first moments of the master equations (methylation density), but will fail to predict higher moments (correlation functions). In particular, we expand to first order in the exponential and to the second order in the terms $\frac{\hat{\phi}(s\pm y)\phi(s\pm y)}{y^\lambda}$. As the last step we extend the limit of integration to infinity and the Hamiltonian is simplified to

$$H[\hat{\phi},\phi] = J\Gamma(1-\lambda)\left(1-\hat{\phi}\right)e^{-\phi\hat{\phi}}\left[2\left(\hat{\phi}\phi\right)^\lambda + \left(\phi\hat{\phi}\right)^{\lambda-3}\left(2-3\lambda+\lambda^2\right)\frac{\partial^2\left(\phi\hat{\phi}\right)}{\partial s^2}\right], \tag{2.18}$$

where $\Gamma(x)$ is the Euler gamma function. In order to derive the last equation we made use of the known integral: $\int_0^\infty dx e^{ax}/x^\lambda = a^{\lambda-1}\Gamma(1-\lambda)$ for $0 < \lambda < 1$. It can be shown [75] that computing the first moment $\langle D(s,t)\rangle = \langle\phi(s,t)\rangle$ from Eq. (2.16) is equivalent to extremising the action $S[\phi,\hat{\phi}]$ in Eq. (2.13), making possible to derive physical observables from averages of the coherent states. The extremisation of the action is known as semiclassical solution or, in other context, mean field solution. We will refer to semiclassical solution when we derive it from a path integral and mean field when it is directly derived from the master equation. Upon extremising the action, $\frac{\delta S}{\delta\hat{\phi}(x)}|_{\hat{\phi}(x)=1} = 0$ while setting $\hat{\phi}(x) = 1$ for probability conservation [8], we obtain a partial differential equation describing the time evolution $\phi(s,t)$ as

$$\frac{\partial\phi(s,t)}{\partial\tilde{t}} = \phi(s,t)^\lambda + D\phi(s,t)^{\lambda-3}\partial_s^2\phi(s,t)\,. \tag{2.19}$$

Here, $\tilde{t} = 2Jt\Gamma(1-\lambda)$ and $D = (2-3\lambda+\lambda^2)/2 > 0$ is the diffusion constant. It can be shown that this solution is a minima of the action such that $\frac{\delta^2 S}{\delta\hat{\phi}(x)^2}|_{\hat{\phi}(x)=1} > 0$. We thus found the semiclassical dynamics for the coherent state $\phi(s,t)$. In can be further shown [8] that in Eq. (2.19) we can substitute $\phi(s,t)$ with $\langle D(s,t)\rangle$ such that Eq. (2.19) describes the dynamics of the average DNMT3 occupancy and so it is a field theory for the average of an observable of the master equation. In the hard-boson path-integral representation, Eq. (2.17), the term $\exp\left[-\int_{z=0}^{y}\mathrm{d}z\hat{\phi}(s+z)\phi(s+z)\right]$ in the Hamiltonian, for a slowly varying field can be approximated by $e^{-y\langle\hat{\phi}(s)\phi(s)\rangle}$, where $\langle\hat{\phi}(x)\phi(x)\rangle$ is the average product of the fields. This defines an effective exponential cutoff to the interactions at a characteristic length. This interpretation of the exponential term justifies the expansion within the limit of integration. We replaced a spatial average with

an ensemble average, which for this spatial translation invariant systems is justified, as long as some conditions are fulfilled, which we are gonna talk about later. Moreover, the analysis done in the previous section was combining ensemble average (different cells) with spatial averages (averaging over genomic features).

2.3.3. Inference of the interaction kernel

To infer the interaction kernel $J_{i,i+l}$ we compute first and higher order moments of the local methylation density $m(s,t)$ which, in the semiclassical limit, follows from $\partial_t m(s,t) = k\phi(s,t)$, where k is the methylation rate (note that we omit demethylation). Having computed these moments for a general class of interaction kernels we can then match theoretical predictions with experimental data to infer the functional form of interactions between enzyme binding events. Here, we focus on the time evolution of the first moment of the global DNA methylation level, $m(t) = \sum_{s=1}^{N} m(s,t)/N$. To begin, we sum Eq. (2.19) to obtain a differential equation for $\phi(t) = \sum_{s=1}^{N} \phi(s,t)/N$, which is solved by

$$\phi(t) = t^{1/(1-\lambda)}\,, \tag{2.20}$$

where time is made adimensional upon diving Eq. (2.19) by J and we refer, with a slight abuse of notation, to the adimensional time as t. Average level of DNAme can be straightforwardly derived as

$$\partial_t m(s,t) = k\phi(s,t)\,, \tag{2.21}$$

arriving to the final expression:

$$m(t) = m(t=0) + k\frac{1-\lambda}{2-\lambda}t^{1+1/(1-\lambda)}\,. \tag{2.22}$$

Therefore, in order to match the experimentally obtained exponent of $5/2$ we find that $\lambda = 1/3$, such that the interaction kernel is

$$J_{i,i\pm l} = \frac{1}{l^{1/3}}\,, \tag{2.23}$$

Our parameter free model is then able to predict first moments observed experimentally Fig.2.4.

2.3.4. Failure of the perturbative expansion of the action

Sometimes, negative results are more useful than positive results. In this section, we will show how typical field theoretical approaches to compute connected correlation

functions fails with an Hamiltonian like Eq. (2.17). We remind (Section 1.3.1), that in order to computed two-point spatial correlation functions the following quantity is evaluated from the field theory:

$$\langle \phi(s,t)\phi(y,t')\rangle = \int \mathcal{D}[\phi]\mathcal{D}[\bar{\phi}]\phi(x,t)\phi(y,t')e^{-S[\phi,\bar{\phi}]}\,. \tag{2.24}$$

When computing correlation functions we typically need a bare action [8], with terms up to quadratic order and higher order terms which are treated perturbatively. The action Eq. (2.13) lacks quadratic terms in the Hamiltonian. We may then look for perturbation around the semiclassical solution, which we indicates as ρ, by means of a Gauge transformation: $\phi = (\rho+\psi)e^{-\bar{\psi}}, \hat{\phi} = e^{\bar{\psi}}$. The Hamiltonian in the new fields is

$$H[\bar{\psi},\psi] = 2\Gamma(1-\lambda)(1-e^{\bar{\psi}})e^{-(\rho+\psi)}\Big[\rho^{\lambda}\left(1+\frac{\psi}{\rho}\right)^{\lambda} + D\rho^{\lambda-3}\left(1+\frac{\psi}{\rho}\right)^{\lambda-3}\frac{\partial^2\psi}{\partial x^2}\Big]\,. \tag{2.25}$$

Starting from Eq. (2.25) we add another term in the Hamiltonian accounting for DNMT3 unbinding. As this term will be shown to be irrelevant it serves only as the purpose of having a well defined stationary state. Such term is of the form $H_u = u\sum_{i=1}^{N}(1-a_i)\hat{\delta}_{D_i,1}$, which results in a term in the action: $u\phi(\hat{\phi}-1)e^{-\hat{\phi}\phi}$. Expanding around the small perturbations $(\psi,\bar{\psi})$ and retaining only leading terms,

$$H = f(\rho)\left(\bar{\psi}+\frac{1}{2}\bar{\psi}^2\right)(-\frac{1}{2}\psi^2+\psi-1)\Big[1+\lambda\frac{\psi}{\rho}+\frac{\lambda^2-\lambda}{2}\frac{\psi^2}{\rho^2}+D\rho^{-3}\left(1+(\lambda-3)\frac{\psi}{\rho}\right)\frac{\partial^2\psi}{\partial x^2}\Big]\,, \tag{2.26}$$

with $f(\rho) = 2\Gamma(1-\lambda)\rho^{\lambda}e^{-\rho}$. The correlation functions for the new fields are computed through the generating functional (Section 1.3.1)

$$Z[\mathbf{h},\psi,\bar{\psi}] = \int \mathcal{D}[\mathbf{h},\psi,\bar{\psi}]e^{-S[\psi,\bar{\psi}]+\int_0^y dx\int_0^{tf} dt[h(x,t)\psi+\bar{h}(x,t)\bar{\psi}(x,t)]}\,. \tag{2.27}$$

Correlation and response functions are respectively derived as previously

$$\begin{aligned}\langle \psi(s,t)\psi(y,t')\rangle &= \frac{\delta^2}{\delta h(x,t)\delta h(y,t)}Z[\mathbf{h},\psi,\bar{\psi}]|_{\mathbf{h}=0}\\ \langle \psi(s,t)\bar{\psi}(y,t')\rangle &= \frac{\delta^2}{\delta h(x,t)\delta \bar{h}(y,t)}Z[\mathbf{h},\psi,\bar{\psi}]|_{\mathbf{h}=0}\,.\end{aligned} \tag{2.28}$$

We then split the action into a Gaussian and an "interactive" part. The linear terms $(-f(\rho)\bar{\psi}+ue^{-\rho}\rho\bar{\psi})$ act as a shift of the external field $\bar{h}$. In Fourier space the Gaussian action, with $\psi_{\pm} = \psi(\pm q,\pm\omega)$ and equivalently for $\bar{\psi}_{\pm}$ can be recast in a matrix form as

$$S_0 = \frac{1}{2}\begin{pmatrix}\bar{\psi}_- & \psi_-\end{pmatrix}\hat{S}_0\begin{pmatrix}\bar{\psi}_+\\ \psi_+\end{pmatrix}\,, \tag{2.29}$$

with

$$\hat{S}_0 = \begin{pmatrix} -f(\rho)\left(1+u\frac{\rho^{-\lambda+1}}{2\Gamma(\lambda-1)}\right) & -i\omega + f(\rho)\left(D\rho^{-3}q^2+\frac{\lambda}{\rho}-1-u\frac{\rho^{-\lambda+1}}{2\Gamma(1-\lambda)}\right) \\ i\omega + f(\rho)\left(D\rho^{-3}q^2+\frac{\lambda}{\rho}-1-u\frac{\rho^{-\lambda+1}}{2\Gamma(1-\lambda)}\right) & 0\,. \end{pmatrix} \tag{2.30}$$

Taken all terms together the moments generating functional is

$$Z[\mathbf{h},\psi,\bar{\psi}] = \int \mathcal{D}[\psi,\bar{\psi},\mathbf{h}] \underbrace{\exp\left[\int_{q,\omega} -\frac{1}{2}\left(\bar{\psi}_-\psi_-\right)\hat{S}_0\begin{pmatrix}\bar{\psi}_+\\ \psi_+\end{pmatrix} + \begin{pmatrix}\bar{h}+f(\rho) & h\end{pmatrix}\begin{pmatrix}\bar{\psi}_+\\ \psi_+\end{pmatrix}\right]}_{Z_0} e^{-S_1}\,, \tag{2.31}$$

where S_1 contains the remaining non-quadratic term of the action. The bare (from the Gaussian part) connected correlations are from Eq. (2.31)

$$C_0(q,\omega) = \frac{f(\rho)\left(1+u\frac{\rho^{-\lambda+1}}{2\Gamma(1-\lambda)}\right)}{\omega^2+W(q)^2} \tag{2.32}$$

With $W(q) = f(\rho)\left[\left(\frac{\lambda}{\rho}-1\right)+D\rho^{-3}q^2-u\frac{\rho^{-\lambda+1}}{2\Gamma(\lambda-1)}\right]$

Once we found the Gaussian part for correlation functions, we can start to treat S_1 perturbatively (Section 1.3.1) by means of Feynman diagrams. Feynman diagrams are a powerful theoretical tools when computing higher order corrections to observables after a perturbation expansion of the action. In particular, the perturbed partition function is in general written as a power series

$$Z[\mathbf{h},\psi,\bar{\psi}] = \int \mathcal{D}[\mathbf{h},\psi,\bar{\psi}]Z_0(1-S_1+\frac{1}{2}S_1^2+\ldots)\,. \tag{2.33}$$

As an example, the first correction to connected correlation functions is formally

$$\langle\psi(s,t)\bar{\psi}(y,t')\rangle = \langle\psi(s,t)\bar{\psi}(y,t')\rangle_{Z_0} - \int \mathcal{D}[\mathbf{h},\psi,\bar{\psi}]Z_0\psi(s,t)\bar{\psi}(y,t')S_1\,, \tag{2.34}$$

where $\langle\psi(s,t)\bar{\psi}(y,t')\rangle_{Z_0}$ stands for the bare correlation functions, Eq. (2.32). We just need to compute the second term on the r.h.s of Eq. (2.34). As Z_0 is Gaussian, this first higher order term is an higher order moment of the form,

$$\int \mathcal{D}[\mathbf{h},\psi,\bar{\psi}]Z_0\psi(s,t)\bar{\psi}(y,t')\int_{s'',t''} F(\psi^l(s'',t)\bar{\psi}^m(s'',t''))\,, \tag{2.35}$$

where we respedented the perturbative contribution of the action S by writing an integral of a general function F of the fields. In general, not every diagram contributes to the connected correlations functions for reasons we don't discuss here [8]. For the purpose of our analysis, the first relevant terms that contribute to connected correla-

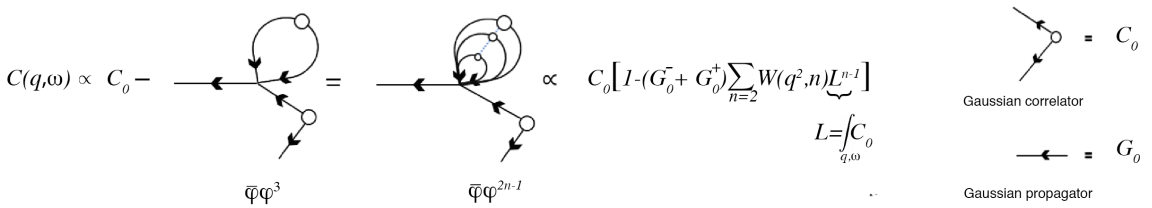


Figure 2.7.: Perturbative expansion of the connected correlation functions, Eq. (2.38). A closed loop indicates integration over internal variables.

tions are the one in the 4th power of the fields: $\psi^l\bar{\psi}^m, l+m=4$, the other terms are not present due to the space translation and left/right symmetry. All of the surviving terms are in the form (in Fourier space),

$$-\left(A_{lm}q^2+B_{lm}\right)\psi^l\bar{\psi}^m\,. \tag{2.36}$$

The first correction (*one loop*) is,

$$\int\mathcal{D}[\mathbf{h},\psi,\bar{\psi}]Z_0\psi(k)\bar{\psi}(k')\int_{k'',k''',k''''}\left(A_{13}q''^2+B_{13}\right)\,\bar{\psi}(k'')\psi(k''')\psi(k'''')\psi(-k''-k'''-k''')\,, \tag{2.37}$$

where we denoted with k the Fourier components as $k=(q,\omega)$. The only non vanishing contribution to the correlation functions is, to one loop (Fig. 2.7),

$$C(k)=C_0(k)\left[1-6(G_0(k)+G_0(-k))\left(A_{3,1}q^2+B_{3,1}\right)\int_{k'}C(k')\right] \tag{2.38}$$

$G_0(k)$ is the null-model for the propagator. We haven't discussed the propagator yet, for now it is simply: $G_0=\langle\bar{\psi}\psi\rangle$.

We then expand the action to higher order terms and the only relevant contributions are given by $\{l=2n-1, m=1\}$ with $n\in\mathrm{N}$. This expansion gives rise to one-loop 2n-point vertex only, Fig. 2.7. If we keep all the term in the expansion the one-loop contribution to correlation functions can be computed has:

$$C(k)=C_0(k)\left[1-(G_0(k)+G_0(-k))\sum_{n=2}^{\infty}c_{2n}\left(A_{2n-1,1}q^2+B_{2n-1,1}\right)\left(\int_{k'}C(k')\right)^{n-1}\right] \tag{2.39}$$

Where $c_{2n}=(2n)!$ is the combinatorial factor coming from the possible contractions for the 2n-point vertexes and,

$$A_{l,1}=f(\rho)D\rho^{-3}\sum_{k=0}^{l-1}\frac{(-1)^k}{\rho^{l-k-1}}\frac{(\lambda)_{l-k+1}}{k!(l-k-1)!(\lambda)_2}\qquad B_{l,1}=f(\rho)\sum_{k=0}^{l}\frac{-1^k}{\rho^{l-k}}\frac{(\lambda)_{l-k}}{k!(l-k)!(\lambda-l+k)}\,, \tag{2.40}$$

with $(\lambda)_l = \lambda(\lambda - 1)...(\lambda - l)$

We obtained a closed expression for the connected correlation functions, with the Gaussian part identical to the Ginzburg-Landau field theory for the dynamics of a non-conservative field [96]. It is known [8] that in such models, power law correlation functions with non trivial exponents arise close to a critical point and our field theory does not have any critical point for $\lambda = 1/3$ [92, 97]. It can be shown that the field theory defined by the Gaussian action Eq. (2.29) can never be scale invariant as the mass term $W(q = 0)$ is never zero [98]. On top of that, we argue that the perturbative approach gave contributions as the one of models A/B [96] for which critical exponents are known [8] and they are not in accordance with experimental data or numerical simulations as we will show. We want to stress that we have power law behaviour of connected correlation functions in a system that does not have a critical point as this is often assumed and sometimes hard to justify for systems with such non trivial behaviour [99, 100]. Intuitively there is a deeper reason why this approach fails to correctly predict correlation functions. After we expand the fields around a base state integrals of the form Eq. (2.17) inside the integrals there is the length scale, $1/\langle\phi\rangle$, which is associated with the effective cutoff of interactions. In other terms, we neglect how the integral behave above and below the natural cutoff introduced by the average occupancy, as expected for long-range interactions restricted to the closest neighbours. In order to overcome this problem in the next sections we will adopt a different approach which will turn out to predict experimental connected correlation functions Fig. 2.5.

2.4. Spatial correlation functions of DNA methylation marks

The key insight to calculate the correlation function is that Eq. (2.17) gives rise to two spatial regimes in sequence space. For short distances interactions are long-range following a power law decay with exponent 1/3, while for distances larger than $1/\langle\phi\rangle$ interactions decay exponentially and are effectively local. In the following, we will therefore derive the correlation function separately for these two regimes using renormalization group methods and we will confirm these results with numerical simulations.

2.4.1. Short tail scaling

In order to develop a method that is capable to describe the short-distance regime we consider the action Eq. (2.17) and, after taking the semiclassical approximation, we

expand it to first order in ϕ, and ∂_s. After this we obtain

$$\partial_t\phi(s,t) = \int_0^s dy\,\phi(y)|s-y|^{-\lambda}e^{-\int_{z=0}^{s-y}dz\,\phi(z)} + \int_0^s dy\,\partial_y\phi(y)|s-y|^{1-\lambda}e^{-\int_{z=0}^{s-y}dz\,\phi(z)}\,, \quad (2.41)$$

where s is the position in sequence space. For the sake of brevity we omitted the noise terms and integrals of the same form describing interactions with the right nearest bound site. The interaction kernel Eq. (2.17) has the form $|s-y|^{-\lambda}e^{-\int_{z=0}^{s-y}dz\,\phi(z)}$, with $\lambda = 1/3$ in the case of *de novo* DNA methylation. Upon considering a perturbation $h(s,t)$ around a mean field dynamical solution, $\phi_0(t)$, i.e. $\phi(s,t) = \phi_0(t) + h(s,t)$ taking terms up to second order in $h(s,t)$ and after some algebra, we find that the equation for $h(s,t)$ is described by, Appendix C.1.1,

$$\partial_t h(s,t) = e^{-\phi_0(t)}\int_0^s dy\,h(y)|s-y|^{-\lambda}e^{-\phi_0(t)}\int_0^s dy\int_y^s dw\,h(y)|s-y|^{-\lambda}h(w-y) + \xi(s,t)\,. \quad (2.42)$$

Here we introduced the noise $\xi(s,t)$, such that this equation is in the form of a Langevin equation. Before going further we need to make a couple of remarks. First, we did not add noise, the noise is always present as the field described by the master equation is stochastic. Second, there are different ways in order to derive noise starting from the path integral Eq. (2.17). The noise can be derived by identifying terms that are proportional to $\hat{\phi}^2$. Why is it so? If we would start from the master equation and write the path integral formulation than the noise term in the Langevin equation acquires a term in $\hat{\phi}^2$ in the path integral (Sec. 1.3.1). Here, we are doing the inverse workflow. Even though, this might look trivial and completely justified, it is not in many situations, due to the different interpretations of the fields in the Doi-Peliti and MRSJD path integral formulation [101]. We carefully look in the expansion of the action Eq. (2.17) terms that are both non-conservative, i.e. proportional to $\hat{\phi}^2$ and conservative, i.e. proportional to $\hat{\phi}^2\partial_s^2\phi$ (terms in $\partial_s\phi$ are not present due to symmetries). By collecting these terms, the noise in Eq. (2.42) has correlations $\langle\xi(s,t)\xi(s',t')\rangle = \delta(t-t')(2\Gamma_{NC} - 2\Gamma_C\partial_s^2)\delta(s-s')$ and zero mean. Γ_C and Γ_{NC} are the noise strengths for conservative and non conservative noise, respectively. They are just functions of J and their specific value is not important in the following. After this little preamble we are in a position to derive connected correlations functions at short distances. After some algebra, Appendix C.1.1, they have the shape

$$\langle h(s,t)h(s',t)\rangle = \left[2\left(|s-s'|^2\Gamma_C + \Gamma_{NC}\lambda(1+\lambda)\right)|s-s'|^{-2-\lambda}\cos(\pi\lambda)/2\Gamma(\lambda)\right]\,, \quad (2.43)$$

where $\Gamma(\lambda)$ is the gamma function. We therefore obtain for the scaling of the correlation function $\langle h(s,t)h(s',t)\rangle \sim |s-s'|^{-\lambda}$ with $\lambda = 1/3$ in the case of de novo DNA methylation.

We expect that higher order corrections to this result will depend on the parameters of the model, in particular on average enzyme occupancy, ϕ_0. In order to understand this point, we may reason that the previous derivation is exact at low values of the average occupancy, ϕ_0, whilst for larger values of ϕ_0 we expect higher order corrections to become relevant. In the context of spatial correlations we use the terms enzyme occupancy and DNAme synonymously from now on. We can already notice from the bare propagator and correlator in that in the short wavelength regime diffusion takes over and in case of conservative noise we expect correlation functions to be described by other exponents. In particular, if we express $\langle h(s,t)h(s',t)\rangle \sim |s'-s|^{2\chi}$ it is possible to realize by dimensional argument [102] that there is another value for the critical exponent, $\chi = \frac{-(1+d+\lambda)}{3}$, which follows from taking into account higher order non linearities and it will be the correct exponent for the long tail of the correlation functions. In the following we are going to prove this simple scaling argument with renormalization group methods.

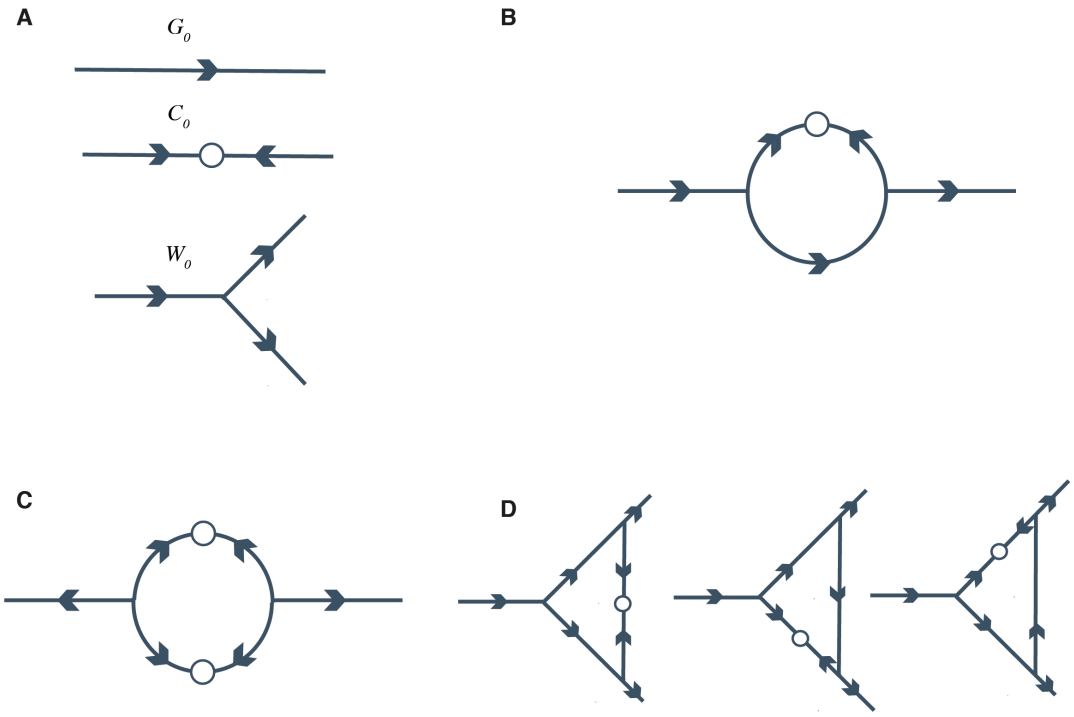


Figure 2.8.: (**A**) Bare propagator (G_0), correlator (C_0) and vertex (W_0). (**B,C,D**) one loop corrections to the propagator, correlator and vertex respectively. Closed loops indicate integration over the internal variables.

2.4.2. Long tail scaling

In order to calculate the exponents for the long distance regime we begin with Eq. (2.41) which, as above, is regularised by including the lowest order spatial derivative that is

in agreement with the symmetries of the model. After linearisation we obtain

$$\begin{aligned}\partial_t h(s) =&\partial_s^2 h(s) + \int_0^s dy h(y)|s-y|^{-\lambda} + \\ &- \int_0^s dy h(y)|s-y|^{1-\lambda}(h(s) - (s-y)\frac{1}{2}\partial_s h(s)) + \xi(s,t)\,.\end{aligned} \tag{2.44}$$

where we used $\int_a^b dx\, f(x) \approx (b-a)(f(a)+f(b))/2$. After some algebra (Appendix C.1.2) we can write the previous equation in Fourier space as,

$$G_0(\mathbf{q},\omega)^{-1} h(\mathbf{q},\omega) = \xi(\mathbf{q},\omega) - \nu \int_{\mathbf{k},\omega'} W(\mathbf{q},\mathbf{k}) h(\mathbf{k},\omega) h(\mathbf{q}-\mathbf{k},\omega'-\omega)\,, \tag{2.45}$$

where $h(\mathbf{q},\omega) = \int d\mathbf{s} \int dt h\,(\mathbf{s},t)\, e^{i\mathbf{q}s} e^{i\omega t}$, $G_0^{-1} = \left(i\omega + D_0\mathbf{q}^2 + J|\mathbf{q}|^{-\lambda}\right)$. We reintroduced the dimensional parameters from the adimensional Eq. (2.44) as we are interested in how different terms in the field theory scale under renormalization. In Eq. (2.45) we defined the vertex, which accounts for the non linear part, as

$$W(\mathbf{q},\mathbf{k}) = \frac{1}{2}\left[\frac{\mathbf{k}(\mathbf{q}-\mathbf{k})}{|\mathbf{k}-\mathbf{q}|^{3-\lambda}} + \frac{(\mathbf{q}-\mathbf{k})\mathbf{k}}{|\mathbf{q}|^{3-\lambda}}\right]. \tag{2.46}$$

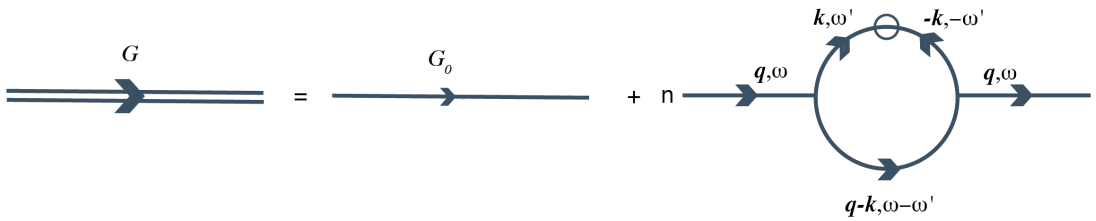
In the limit $\mathbf{k} \to 0$ (hydrodynamic limit) the vertex scales as $\mathbf{k}$, which implies non renormalization of the vertex function. We now proceed with standard renormalization group technique as outlined in Sec. 1.3.2. RG describes the flow of parameters under renormalization. Due to the simple form of Eq. (2.45), which does look like a KPZ equation with non-local kernel [103, 104], we don't even have to go to the path integral formulation as it is quite easy to derive perturbative expansions of the vertex, propagator and correlator [105]. In particular, we can notice that Eq. (2.45) is a recursive equation for the field h, as it enters on the r.h.s in the convolution integral such that by recursion,

$$G_0(\mathbf{q},\omega)^{-1} h(\mathbf{q},\omega) = \xi(\mathbf{q},\omega) - \nu \int_{\mathbf{k}} W(\mathbf{q},\mathbf{k}) G_0(\mathbf{q},\omega)\xi(\mathbf{q},\omega)\ldots\,. \tag{2.47}$$

We haven't specified what the dots in the previous equations are. It will take a page, or actually an infinite page to write them all. Moreover, this is just the recursion relation for h and we would need to write other recursions for G, C, W, really a tough program! Luckily, Feynman diagrams are again powerful tools to write this recursion relationship in a compact and clear form. In particular, to *one loop* (cutting the recursion after one iteration), the perturbative expansion of the field theory (2.45) is represented by means of Feynman diagrams in Fig. 2.8. These diagrams then serves as a way to find corrections to the parameters under RG. As an example, the perturbative expansion of the propagator by taking into account the first order correction (Fig. 2.8B) is,

$$G(\mathbf{q},\omega) = G_0(\mathbf{q},\omega) + 4\nu^2 G_0(\mathbf{q},\omega)^2 \int_{\mathbf{k},\omega'} W(\mathbf{q},\mathbf{k}) h(\mathbf{k},\omega) \\ h(\mathbf{q}-\mathbf{k},\omega'-\omega) W(\mathbf{q},-\mathbf{k}) C(\mathbf{k},\omega') G_0(\mathbf{q}-\mathbf{k},\omega-\omega') \tag{2.48}$$

Without any technical computations, we notice that the integral is diverging, that's were RG comes into play. As outlined in Sec. 1.3.2, the integral is computed in the momentum shell $[\Lambda e^{-l}, \Lambda]$, and the divergence is "cured". After the evaluation of the integral, we need to compare the l.h.s and r.h.s of Eq. (2.48). Specifically, all the terms of the order q^2 will renormalize the diffusion coefficient D_0 and so the terms in q^λ will renormalize J. We have to do the same procedure for the renormalization of the correlator and the vertex, which first order corrections are respectively given in (Fig. 2.8 C,D). We skip further technical details as they are quite lengthy and are similar to KPZ like equations found in many textbooks [8, 68] and papers [106, 107]. We outline the standard workflow procedure of RG in Fig. 2.9.



Write the corresponding equation in momentum shell renormalization (integrate out long wavelengths)

$$G(\mathbf{q},\omega)^{<} = G_0(\mathbf{q},\omega) + n\nu^2 G_0(\mathbf{q},\omega)^2 \int^{>}_{\mathbf{k},\omega'} W(\mathbf{q},\mathbf{k}) h(\mathbf{k},\omega) h(\mathbf{q}-\mathbf{k},\omega'-\omega) W(\mathbf{q},-\mathbf{k}) C(\mathbf{k},\omega') G_0(\mathbf{q}-\mathbf{k},\omega-\omega')$$

Compare parameters from both sides of the equations (set Λ to 1 eventually)

$$i\omega + D_0 q^2 + J|\mathbf{q}|^\lambda = i\omega(\Lambda,l) + D_0(\Lambda,l) q^2 + J(\Lambda,l)|\mathbf{q}|^\lambda$$

Determine RG flow equations for all the Feynman diagrams

Figure 2.9.: Steps to derive the RG flow equations. In this example we evaluate the correction to the propagator. n are the numbers of diagrams (in this case $n = 4$). For simplicity $\int^{>}$ is the integral were the variables are computed for high momenta $\mathbf{k} \in [\Lambda e^{-l}, \Lambda]$, and the opposite for $G^{<}$.

Finally, the renormalization of the parameters leads to the following RG flow equa-

tions,

$$\begin{aligned}
\partial_l D_0 &= \left[z - 2 - \frac{K_d \nu^2}{d D_0^3} \left[(d-2)\,\Gamma_{NC} + (d-3)\,\Gamma_C \right] \right] D_0\,, \\
\partial_l \nu &= \left[z + \chi - 2 + (3-\lambda) \right] \nu\,, \\
\partial_l \Gamma_C &= \left[z - 2\chi - d - 2 - \frac{K_d \nu^2}{2 d D_0^3 \Gamma_C} (1+d)(\Gamma_{NC} + \Gamma_C)^2 \right] \Gamma_C\,, \\
\partial_l \Gamma_{NC} &= \left[z - 2\chi - d \right] \Gamma_{NC}\,,
\end{aligned} \tag{2.49}$$

where $K_d = S_d/(2\pi)^d$ and S_d is the area of a d dimensional sphere. From the non renormalization of the non conserved noise and of the couplings we get the exact exponent identities $\chi = \frac{(-1-d+\lambda)}{3}$ and $z = \frac{(-2+d+2\lambda)}{3}$. χ is the exponents describing decay of spatial equal time connected correlation functions of the fields and z is the dynamical critical exponents, describing how temporal and spatial correlation length scale with respect to each other at criticality [102]. In $d = 1$ and for long distances correlations then decay with an exponent $2\chi = -\frac{10}{9}$. Taken together, we find that the correlation function in sequence space decays in two algebraic regimes,

$$C(s - s') = \begin{cases} |s - s'|^{-\left(\frac{1}{3}\right)}, & \text{for } |s - s| \ll 1/\langle m \rangle\,, \\ |s - s'|^{-\left(\frac{10}{9}\right)}, & \text{for } |s - s| \gg 1/\langle m \rangle\,. \end{cases} \tag{2.50}$$

The cross-over between these regimes is related to the cutoff of the long-range interactions. The position of the cross-over scales with the only length scale in the system, the typical distance between neighbouring methylated CpGs, $1/m$. Intuitively, this length scale separates a regime dominated by active feedback between DNA methylation and topology and a regime characterised by passive, conservative fluctuations. This last point will be clearer in the next section. The numerical prefactor of the proportionality between the position of the crossover and $1/\langle m \rangle$ depends on the statistics of distances between neighbouring CpGs in base pairs. We confirm theoretical predictions and determine this prefactor using stochastic simulations of master equation Eq. (2.4) with disordered represented by the actual CpG positions in the mouse genome and found that the position of the crossover is approximately equal to $350/\langle m \rangle$ (see below). The algebraic decay of connected correlation functions arises naturally in our system due to the shape of the interaction kernel, which results in a system far from equilibrium. In equilibrium statistical mechanics, power law correlation functions are expected to be observed when a system is closed to a critical threshold. On the other hand, in a non-equilibrium setting, power law behaviour of connected correlation may be observed in a non critical regime [98]. Even though our theory is in agreement with experimental data, predicting scaling behaviour of average DNA methylation and correlation

functions, the interaction kernel that we inferred, Eq. (2.23), has not yet a physical or biological meaning. In the following, we are going to derive the consequences in the space of the nucleus of the interaction kernel and provide a framework to infer chromatin structures from lower dimensional data and models.

2.5. Inference of mesoscopic processes in physical space

In the previous sections we inferred the dynamics in the sequence space of the DNA, but we have clear in mind that biological processes, and DNAme is no exception, happens along the 3D structure of the chromatin. How can we say something regarding topological structures on the DNA if we just study a master equation for binding and methylation kinetics in 1D sequence space? This is the gap that we want to bridge in this section. Specifically, we give a theoretical tool to answer the more general genomic question: How can we learn from low dimensional data about high dimensional structures?

On top of this very general problem, when we computed experimental connected correlation functions in Sec. 2.2, we found a dependence of the power law of correlation functions to time and so, average methylation. Even though, via theoretical analysis of the master equation (2.4) we were able to predict different spatial regimes and capture exponents of connected correlation (Sec. 2.4.1,2.4.2), we are still lacking an understanding of how they depend on the average methylation. In this section we develop a field theory to understand the dynamics in the physical space of the nucleus and the exponents will be characterised within this framework.

2.5.1. Geometric consequences of the interaction kernel

In Sec. 2.3 we infer the kernel from sequencing data and in particular from its first moment (average DNAme) and we obtain Eq. (2.23). Mathematically this kernel is well defined and allowed us to compute spatial correlation functions. However, in this form, the kernel is a pure mathematical object without any clear biological interpretation. Mathematically the kernel is proportional to the rate of binding at a given CpG i for a DNMT3 enzymes and we found that it is proportional to the distance to the closest bound enzymes to the power of $-1/3$, i.e. $J_i = 1/L^{1/3} + 1/R^{1/3}$, with L, R the distance to the left and right nearest bound DNMT3 respectively. Instead of looking at the binding rate at a given CpG site, we ask what is the binding rate in a region surrounding that enzyme, i.e., between the other already bound enzymes. The total binding rate in a region of size l around bound sites is, $\sum_{i=1}^{|R-L|} J_i$, therefore scales

as $l^{2/3}$. $l^{2/3}$ is the surface to volume ratio of an object with volume l, such that if a genomic region of l base pairs were compacted $l^{2/3}$ base pairs would be accessible on the surface, Fig. 2.10 A. Therefore, notably, the inferred interaction kernel describes the compaction of the DNA around methylated sites and the preferential binding of DNMT3 to compacted regions, resulting in positive feedback, Fig. 2.10 B , which is fully consistent with biochemical studies showing that DNAme leads to attractive forces between tetra-nucleosomes *in vitro* [108]. Starting from a master equation in the one dimensional sequence space we arrive to a geometrical interpretation on the three dimensional scale of the nucleus, which we will refer from now on as physical space. In this section we are going to rigorously map the dynamics on these two different spaces.

2.5.2. Field theory in physical space

To systematically derive the dynamics in the space of the nucleus (physical space) from the inferred kinetics in sequence space we start with the partial differential equation describing the spatio-temporal dynamics in sequence space in the semiclassical limit, Eq. (2.19). To begin, we will ask how small length elements in physical space evolve in time for a given position in sequence space. Based on this, we will calculate effective local fluxes in DNA methylation density in physical space and employ a real space renormalization scheme to absorb directed, non local fluxes. Specifically, we seek to define a function g_i that describes the evolution of length elements in a properly defined physical space with respect to changes in DNA methylation density,

$$\delta\Delta x_i = \Delta x_i - g_i(\Delta x_i), \tag{2.51}$$

with initial conditions $\Delta x^0 = \Delta s$. Δs and Δx are, respectively, length elements in sequence and physical space. $g_i(\Delta x_i)$ is for now an unknown function that depends on processes that changes the topological structure of the DNA. Our approach is different with respect to standard differential geometrical approaches. The main reason for introducing a different approach is that we want to derive a general representation of topological dynamics of the DNA which is robust with respect to local forces acting on the DNA. These local forces are often unknown, especially *in vivo* and we want to derive a minimal model, such that we don't need to claim which particular force is relevant. The procedure goes as follow: starting from a description in sequence space given by the semiclassical time evolution of the field $\phi(s,t)$, we aim to derive a master equation in physical space. Let $\Delta\xi_i$ be the length of a discrete element in physical space. A DNA methylation event (understood in a coarse-grained fashion) causes a local compaction of the DNA, such that this length element is contracted by

$$\Delta\xi_i' = \Delta\xi_i^{1/3}\,. \tag{2.52}$$

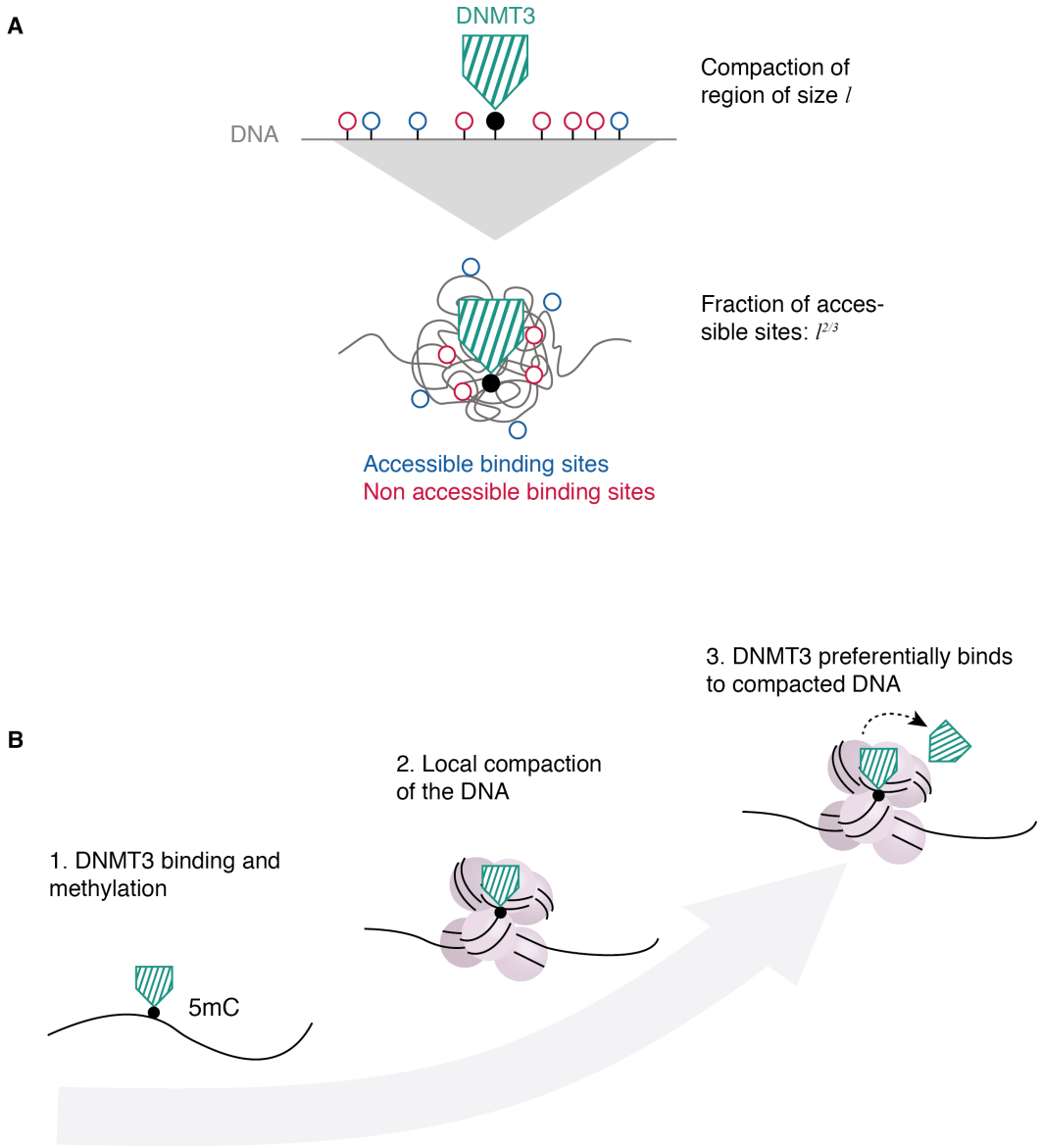


Figure 2.10.: (**A**) Geometric and physical interpretation of the local kernel in Eq. (2.23). The binding rate scales as the distances l between two neighbouring enzymes to the power of 2/3 which is the fraction of accessible sites due to local compaction of the chromatin. $l^{2/3}$ is proportional to the surface to volume ration of a sphere of radius l. (**B**) Local feedback between DNA methylation and topology which gives rise to *de novo* DNA methylation.

Physical space is then understood as the projection of a 3D space onto a 1D space. We argue that this projection is a very good representation of the whole 3D space as long as the predicted structure spans over short length scales, shorter than DNA loops, Fig. 2.11.

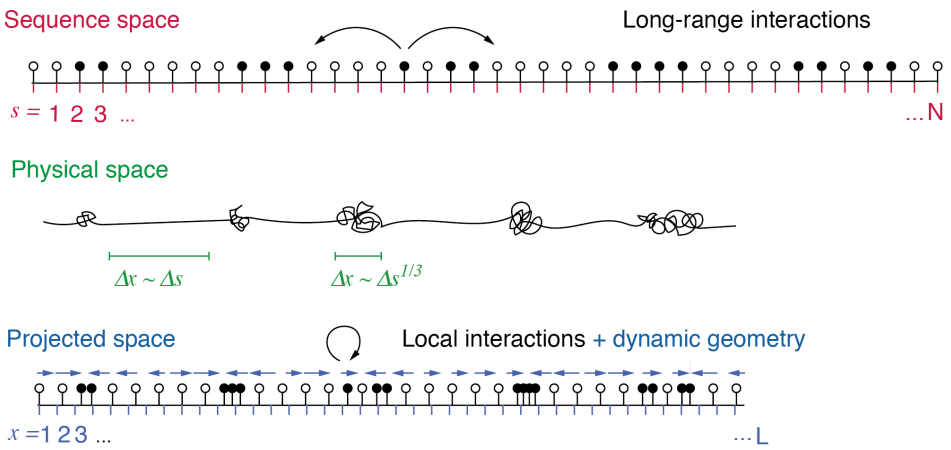


Figure 2.11.: Physical space as projection of the 3D space of the nucleus. The master equation in sequence space (2.5) is dominated by long-range interactions between DMNT3 enzymes. Upon coarse graining in a projected 1D space, the interactions become local (in this space) and give rise to condensation in the physical space of the nucleus, Eq. (2.58).

In the absence of demethylation continous *de novo* methylation will therefore continously compact the DNA locally such that the total length of the DNA in physical space, $\Lambda = \sum_i \Delta\xi_i$, decreases over time. This gives rise to a flux with a velocity that locally depends on the entire concentration field left and right of a given position. To avoid such difficulties, we define a dynamic real space renormalization scheme with renormalized length elements Δx_i such that the total length of the DNA, $L = \sum_i \Delta x_i$, remains constant over time. To achieve this, for a given *de novo* methylation event, Δx_i is first contracted according to $\Delta x_i' = \Delta x_i^{1/3}$ and then rescaled by $\Delta x_i'' = b\Delta x_i'$, with the rescaling factor $b > 1$ given by

$$b = \frac{\Delta x_i + \Delta x_{i+1} + \Delta x_{i-1}}{\Delta x_i' + \Delta x_{i+1}' + \Delta x_{i-1}'} \,. \tag{2.53}$$

After renormalizing back such that the total length is unchanged, we get an effective flux of methylated sites Fig. 2.12, in the vicinity of the contracted domain. With this, we obtain an updating scheme for the concentration of methylated sites at position i, ρ_i, at each DNA methylation event. Specifically, the updated concentration at a given position, ρ_i', is given by contributions from the original concentrations and symmetric

fluxes from the adjacent left and right length elements,

$$\rho_i' = \rho_i + (\rho_{i+1} + \rho_{i-1}) \frac{\Delta x - b\Delta x^{1/3}}{2\Delta x} \,. \tag{2.54}$$

We now consider the joint probability $P(\boldsymbol{\rho}, t)$ to find a given concentration profile $\boldsymbol{\rho}$ at a time t. On time scales much larger than the time scales associated with microscopic DNA methylation and compaction events, we can define the rate of *de novo* methylation in a given length element i, $W(\rho_i) = \rho_i^\lambda$ with $\lambda = 1/3$. In this limit, the time evolution of $P(\boldsymbol{\rho}, t)$ is then given by a master equation for the redistribution of DNA methylation marks in physical space. It takes the form

$$\begin{aligned} \partial_t P(\boldsymbol{\rho}, t) = &\sum_i \left[W(\rho_i - r\rho_{i-1}) P(\rho_i - r\rho_{i-1}, \rho_{i-1} + r\rho_{i-1}, \boldsymbol{\rho})\right] \\ &+ \sum_i \left[W(\rho_i - r\rho_{i+1}) P(\rho_i - r\rho_{i+1}, \rho_{i+1} + r\rho_{i+1}, \boldsymbol{\rho}, t)\right] \\ &- 2\sum_i W(\rho_i) P(\boldsymbol{\rho}, t) \,, \end{aligned} \tag{2.55}$$

where $r = \frac{(\Delta x - b\Delta x^{1/3})}{(2\Delta x)}$ is a dimensionless parameter describing the effective flux of DNA methylation in physical space as a result of a DNAme event. Here, $P(\boldsymbol{\rho}, t)$ is the probability of a given density profile at time t, and we introduce a notation where $P(\rho_{i+1} - 1, \boldsymbol{\rho}, t)$ signifies the probability of a density profile $\boldsymbol{\rho}$, under the condition that at position $i+1$ the density is equal to $\rho_{i+1} - 1$. The term $W(\rho_i)$ accounts for the rate of the DNA methylation, $W(\rho_i) = \rho_i^\lambda$.

The first moment of the master equation (density of enzymes in the projected space) is (Appendix C.2),

$$\partial_t \phi(x, t) = -r a_0^2 W(\phi(x, t)) \partial_x^2 \phi(x, t) \,. \tag{2.56}$$

This partial differential equation describes the flux of DNA methylation density in physical space stemming from the local, methylation-dependent compaction of the DNA. How do the other terms in Eq. (2.19) evolve in physical space? In general, the procedure outlined above leads to a change in the functional form of long-range interactions in physical space. On a mean-field level, however, such interactions again give rise to a local and a diffusion term with potentially different non-linear dependencies on $\phi(x, t)$. Following the calculations we performed in sequence space in Section 2.3 the exponents describing these non linearities in the Langevin equation are not independent of each other. As the first moment, which is a global average on the genome scale, must be identical in sequence space and in physical space, the local term and therefore also the diffusive term, must have the same form in sequence and in physical space. Therefore, in physical space, the time evolution of $\phi(x, t)$ is described by a combination

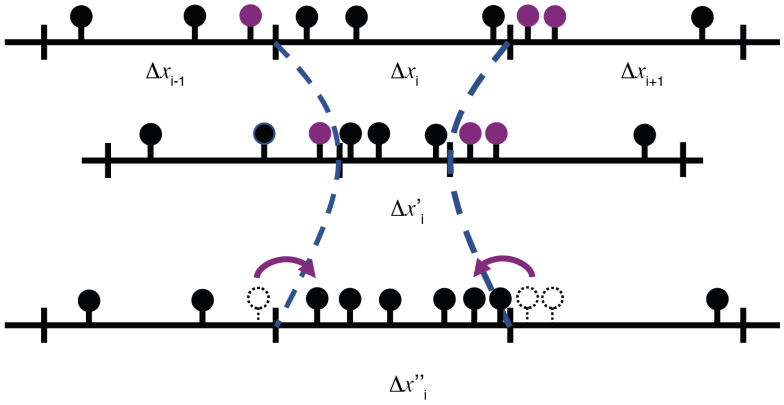


Figure 2.12.: Effect of changes in DNA compaction on the redistribution of methylated CpGs in physical space (black and purple circles). The method consists of two steps: first, after a methylation event the chromatin gets contracted (blue dashed line). Secondly, we renormalise space such that the total length of the system remain invariant (blue dashed line). With this procedure methylated sites of neighboring domains (purple circles) effectively lead to a flux into the contracted domain in physical space.

of processes identical in sequence space and an additional term, Eq. (2.56), describing the flux of DNA methylation in physical space due to changes in DNA topology. Taken together, after substituting the definition of $W(\phi)$, $W(\phi(x)) = \phi(x)^\lambda$, we arrive at a partial differential equation describing the time evolution of $\phi(x,t)$ in renormalized physical space,

$$\partial_t \phi(x,t) = \phi(x,t)^\lambda + \phi(x,t)^{\lambda-3}\partial_x^2\phi(x,t) - r\phi(x,t)^\lambda\partial_x^2\phi(x,t)\,. \tag{2.57}$$

By taking into account the next highest order in the Van Kampen expansion we derive a term for the noise which, due to the conservation of methylation in the renormalization procedure, is conservative,

$$\partial_t \phi(x,t) = \phi(x,t)^\lambda + \phi(x,t)^{\lambda-3}\partial_x^2\phi(x,t) - r\phi(x,t)^\lambda\partial_x^2\phi + \eta(x,t) + \partial_x\left[g(\phi(x,t))\xi(x,t)\right]\,. \tag{2.58}$$

The noise terms have correlations $\langle \xi(x,t)\xi(x',t')\rangle = 2\Gamma_C\delta(t-t')\delta(x-x')$ and $\langle \eta(x,t)\cdot\eta(x',t')\rangle = 2\Gamma_{NC}f(\phi(x,t))\delta(t-t')\delta(x-x')$. We include the non-conservative noise term as it comes from binding kinetics described in the previous sections. As in the following we will mostly interested in perturbations around a dynamical homogeneous solutions the particular dependencies in $g(\phi)$ and $f(\phi)$ is not relevant to lowest order for the remainder of our analysis.

The partial differential equation describing the time evolution of the field $\phi(x,t)$ is structurally similar to Eq. (2.19), but contains a new term $-r\phi(x,t)^\lambda \partial_x^2 \phi$ which is an anti-diffusive term counteracting the diffusion term. We expect that with increasing values of ϕ this term will dominate the diffusion term and potentially lead to the formation of highly methylated regions in physical space. Although we did not explicitly state higher order terms in the spatial derivatives these terms must exist. In the next section we will investigate how such terms affect the formation of methylation condensates.

2.5.3. Formation of condensates in physical space

Eq. (2.58) highlights the emergence of new physical phenomena with respect to the dynamics in sequence space. In this section, we will systematically investigate whether spatial structures emerge in physical space. To this end, we will investigate whether a homogeneous field in physical space is linearly unstable, which would imply the emergence of a characteristic length scale resembling DNA methylation condensates. To this end we take into account the next highest order term in ϕ, $\epsilon\partial_x^4\phi$, which describes restoring forces counteracting DNA compaction at a finite length scale. Condensation happens if a spatial perturbation, $\delta\phi$, of a homogeneous solution, ϕ_0, is unstable. Following standard procedures [109] we linearised Eq. (2.57) and made a general ansatz for the time evolution of the field $\phi(x,t)$ upon perturbation with wave vector k,

$$\phi(x,t) = \phi_0 + e^{\omega t} e^{ikx} \delta\phi \,. \tag{2.59}$$

The homogeneous state is unstable if $\omega > 0$. We obtain a dispersion relation relating the rate of growth of the instability to the wavelength of the perturbation of the form

$$\omega(k) = \lambda\phi_0^{\lambda-1} - \left(\phi_0^{\lambda-3} - r\phi_0^\lambda\right) k^2 - \epsilon\phi_0^\lambda k^4 \,, \tag{2.60}$$

which is depicted in Fig. 2.13 A for a fixed value of r and varying values of ϕ_0. Clusters of methylated DNA can form if the maximum of this function is greater than zero for a finite values of k. This latter condition is ensured by the second derivative being negative. We find that the homogeneous state becomes unstable if $\phi_0 > r^{-\frac{1}{3}}$. The strongest growing mode at the point of the instability is $k = \frac{\sqrt{r\phi_0^3 - 1}}{\left(\sqrt{6}\epsilon\phi_0^{3/2}\right)}$, which gives an indication of the expected typical length scale of the resulting pattern. In summary, we expect an instability in the form of finite size methylation condensates to arise if the average DNA methylation concentration, ϕ_0, exceeds a threshold above which anti-diffusion takes over diffusion given by $r^{-\frac{1}{3}}$. We performed numerical simulations of the Langevin equation (2.58) and found that methyl condensates do indeed form in

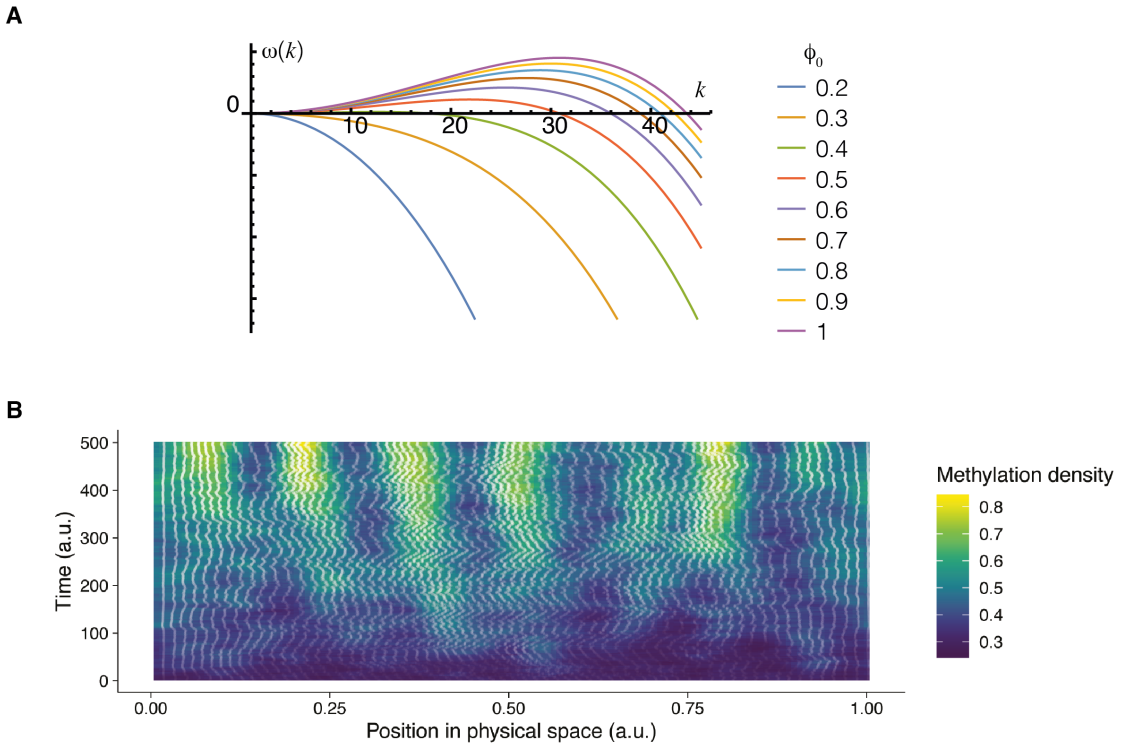


Figure 2.13.: (**A**) Dispersion relation of Eq.(2.58) with the addition of a fourth order term in the derivative ($\epsilon\phi(x,t)^{\lambda}\partial_x^4\phi(x,t)$). The parameter values are $r = 20, \epsilon = 0.01, \lambda = 1/3$. (**B**) Numerical simulations of Eq.(2.58) are performed with a pseudo-spectral method [110] with the same parameter as in **A**.

the expected parameter range (Fig. 2.13 B).

2.5.4. Order of magnitude estimate of condensate sizes

As the precise values of r and ϵ are unknown the linear stability analysis cannot be used straight forwardly to identify the length scale of the predicted methylation condensates. To get an order of magnitude estimate of these condensates we therefore resort to dimensional analysis. There are three length scales involved in the formation of methylation condensates corresponding to the parameters determining the dispersion relation in the previous section:

1. The typical distance between methylated CpGs, l_{5mC}. With an average CpG density of roughly 1% and average DNAme level of 50% we estimate that $l_{5mC} \approx 500\,\text{bp}$.

2. The typical length scale over which DNAme locally affects DNA compaction, l_c. From the cross-correlation function between DNAme and accessibility we find

that $l_c \approx 1000\,\text{bp}$, Fig. 2.18D. We expect that this length scale affects the size of condensates positively.

3. A length scale describing the restoring force counteracting DNA compaction, l_r. We expect this to be of the same order of magnitude as the DNA persistence length, $l_r \approx 100\,\text{bp}$, and to affect the size of condensates negatively.

Taken together, the length scale reflecting the typical size of condensates from these three length scales is given by $l \approx l_{5mC} l_c / l_r$, which is approximately equal to 5000bp.

To test if dynamics of local feedback between DNAme and compaction also leads to the formation of higher-order chromatin structures (methyl-condensates) on larger spatial scales with increasing levels of DNA methylation we reasoned that such condensates should be identifiable as an excess of mid-range physical contacts between pairs of genomic loci in highly methylated regions as measured in chromatin conformation capture experiments. We therefore analysed single-nucleus methyl-3C sequencing data of mouse serum grown ESCs [111]. We tiled the mouse genome into windows of 100kb and, for each window, calculated average DNAme levels and the probability distribution of contact distances. Notably, we found an abrupt increase in mid-range contacts between 3000bp and 5000bp (translating to roughly 30-40nm in diameter) for regions exceeding an average DNAme level of 40 % (Fig. 2.14), in agreement with our prediction of a spatial instability and the emergence of DNAme associated chromatin structures. The sizes of these structure are again consistent with our theoretical estimate and with those estimated from super-resolution imaging studies [112–114].

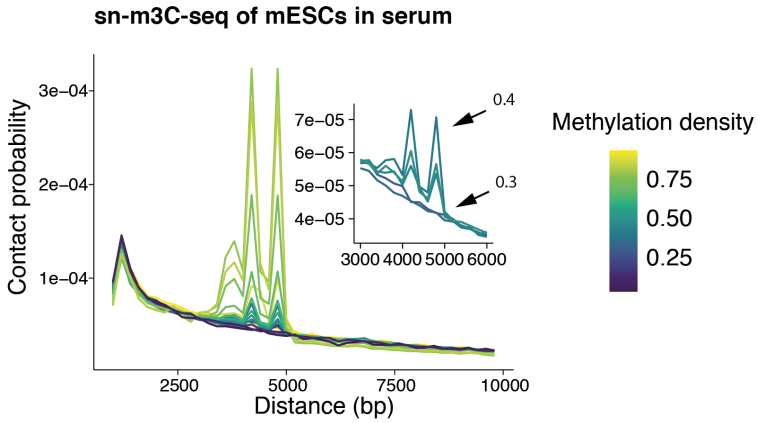


Figure 2.14.: Contact probability of genomic regions for distance (0-10000 bps). We subdivided genomic fragments by their average DNA methylation. There is a notable increase of the contact probability at distances between 3000 and 5000 bps, which is in agreement with our dimensional argument Sec. 2.5.4. The bump of contact probabilities is present only for regions with average DNAme greater than 40% (inset), in agreement with the stability analysis derived in Eq. (2.5.4).

2.6. Prediction of experimental correlation functions

Having inferred how *de novo* DNAme dynamics affects the chromatin structure in the three dimensional space of the nucleus we are in the position to answer the remaining open question about methylation dependent exponents of connected correlations. We focus on the short tail corrections as the corrections to the long tail are identical and the same reasoning will apply.

To understand these corrections to the exponents it is convenient to temporarily consider the correlation function in non-renormalised physical space, and then transform back to sequence space. We begin by noticing that in sequence space the correlation function decays as $|s-s'|^{-1/3}$ for vanishing values of average local DNAme. In non-renormalised physical space for vanishing average DNA methylation the correlation function must scale in the same way as in sequence space, i.e. $\sim |\xi - \xi'|^{-1/3}$. We now consider the effect of n DNA methylation events. According to Eq. (2.52), this leads to a contraction $\Delta\xi' = \Delta\xi^{-(1/3)^n}$. Going back to sequence space, where length elements scale as $\Delta s \sim \Delta\xi^3$ we obtain that the correlation function decays as $|s-s'|^{-(1/3)^{n+1}}$. n is a monotonically increasing function of the local average DNA methylation level which vanishes for $\phi_0 \to 0$. Expanding to first order we obtain approximately $n \approx \alpha\phi_0 + \ldots$ where α is a parameter that we determined to be approximately equal to 1 numerically. Correlation functions then scale as $\langle h(s,t)h(s',t)\rangle \sim |s-s'|^{-(1/3)^{1+\phi_0}}$. If we now replace the average occupancy with the average methylation, as they are straightforwardly directly proportional Eq. (2.21), we obtain that two point connected correlation functions are described by two power laws with methylation density exponents

$$C(s-s') = \begin{cases} |s-s'|^{-\left(\frac{1}{3}\right)^{1+\langle m\rangle}}, & \text{for } |s-s| \ll 1/\langle m\rangle\,, \\ |s-s'|^{-\left(\frac{10}{9}\right)^{1+\langle m\rangle}}, & \text{for } |s-s| \gg 1/\langle m\rangle\,. \end{cases} \tag{2.61}$$

An exponent, that depends on the average might be definitely something nor common or trivial. Intuitively, global level of average DNAme play the role of dimension in field theory, and exponents do depend on the dimensionality of the system for standard field theories. We then can argue that we started from a one dimensional theory and realised that the dimensions are effectively changing along with average methylation, such that absence of DNA methylation gives the one dimensional results and a completely methylated DNA would be a singularity with zero dimension. This is of course quite absurd and not biologically relevant, but yet a smoother transition between dimensions is expected. In Fig. 2.15 A we summarise different regimes of the correlations function and their biophysical meaning. In Fig. 2.15 B we compare theoretical prediction with numerical simulation of the master equation (2.4). We validate our theoretical and

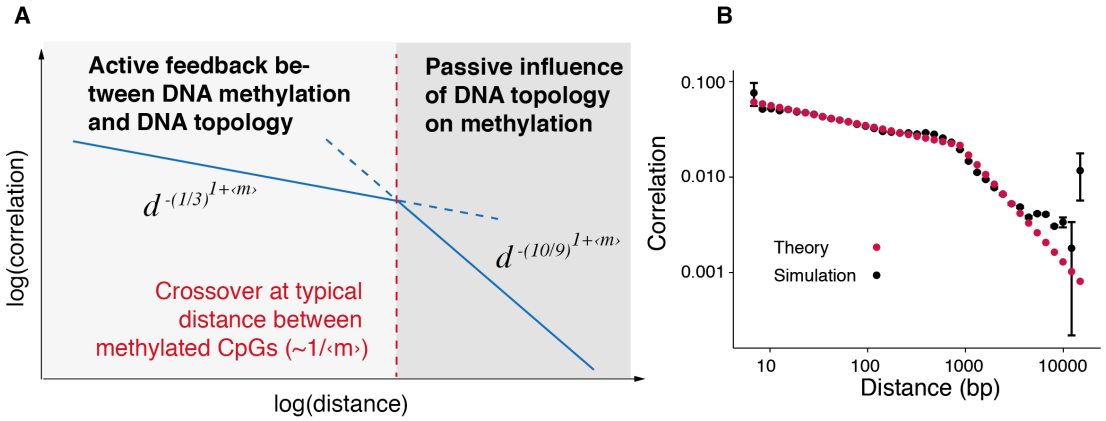


Figure 2.15.: (**A**) Theoretical prediction of connected correlation functions from renormalization group and field theoretical methods. Connected correlation functions decay as power laws for two different spatial regimes dominated by active or passive influence of topology on DNA methylation respectively. The crossover between these two regimes is set by the inverse of the average DNA methylation. (**B**) Gillespie simulations [115] of the master equation (2.4) with binding rate $J = 1$ and methylation rate $k = 1$. We omit unbinding and demethylation. Numerical simulations are performed over a lattice with the biological distribution of distances of CpGs from chromosome 1 of the mouse. Numerical correlations functions are computed as in experimental bulk BS-Seq (Sec. 2.2). Here we plot correlations functions for a global average methylation of 0.5. Theoretical correlation functions, Eq. (2.50) correctly captures the exponents and the cross-over of the numerical simulations.

numerical predictions with two different experiments. The first experiment, mentioned in the beginning is a scNMT-Seq 2i-release experiment of mESCs in which cells are sampled up to two days after release (D0,D1,D2). The second experiment is a scBS-Seq of mESCs cultured *in vitro* in serum conditions. Experimental data for the latter experiment was processed identically to [16]. The results are shown in Fig. 2.16, where we computed two point connected correlation functions in different genomic regions (features) and for different methylation density. We were able to achieve the first data split as we have high spatial resolution in single-cell sequencing and we further divide by local methylation density as the predicted exponents do depend on average methylation. Our parameter free theory (dashed black line) is in excellent agreement with experimental data (Fig. 2.16) further strengthening the hypothesis of local feedback between DNA methylation and topology via long-range interactions.

It is quite mesmerising that a parameter free theory predicts methylation patterns in different genomic regions and at different stages in development. Puzzled by this finding we sought to challenge our theory further. In particular scNMT-Seq allows to obtain molecular information of chromatin structures (GpC methylation) and gene expression (mRNA-Seq) at the same time as CpG methylation (BS-Seq). With standard bioinfor-

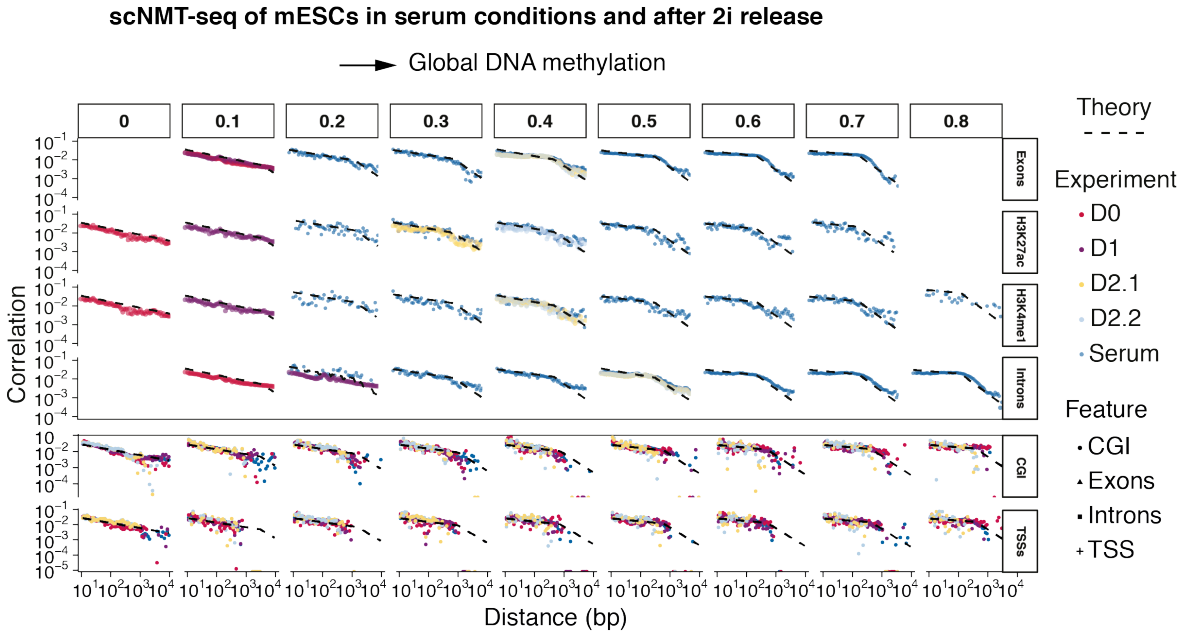


Figure 2.16.: Connected correlation functions of mESCs in serum conditions and after 2i release. Colors stands both for different days after release in the scNMT-Seq 2i-release experiment as well as to distinguish serum and 2i-release. Correlations functions are computed dividing different features by their DNAme $(0-0.8)$. Theoretical predictions (dashed black line) are in perfect agreements with the data as they captures both the length scales that divides the different regimes as well as the exponents across all the features and different DNAme. Promoters (TSS) and CpG islands (CGI) are shorter compared to other features such that long tails could not be resolved.

matics techniques Appendix B.2.2 we analyse at first the transcriptome Fig. 2.17 A,B and found that Dnmt3 genes are upregulated during development as well as other genes associated to pluripotency, which is consistent with their active role in *de novo* DNA methylation. In the next section we will analyse and predict chromatin structure from GpC methylation, which is a measure of DNA accessibility as GpC methyltransferase enzymes bind to regions where nucleosomes are depleted [116].

2.6.1. Cross-correlation functions

scNMT-Seq gives detailed molecular information for three layers of regulation, CpG DNAme, RNA expression and chromatin structure. In particular, it maps the 1D DNA sequence to a binary sequence where each element of the sequence is either accessible or not. This information is stored in GpC methylation, making this layer of information analysable as CpG methylation. Predicting accessibility is thus the last challenge to our theory, which allowed us to infer chromatin structure from one dimensional sequencing data. To infer GpC methylation, we then proceed in the same fashion as done for CpG methylation. We first compute global changes in GpC methylation, which we will refer

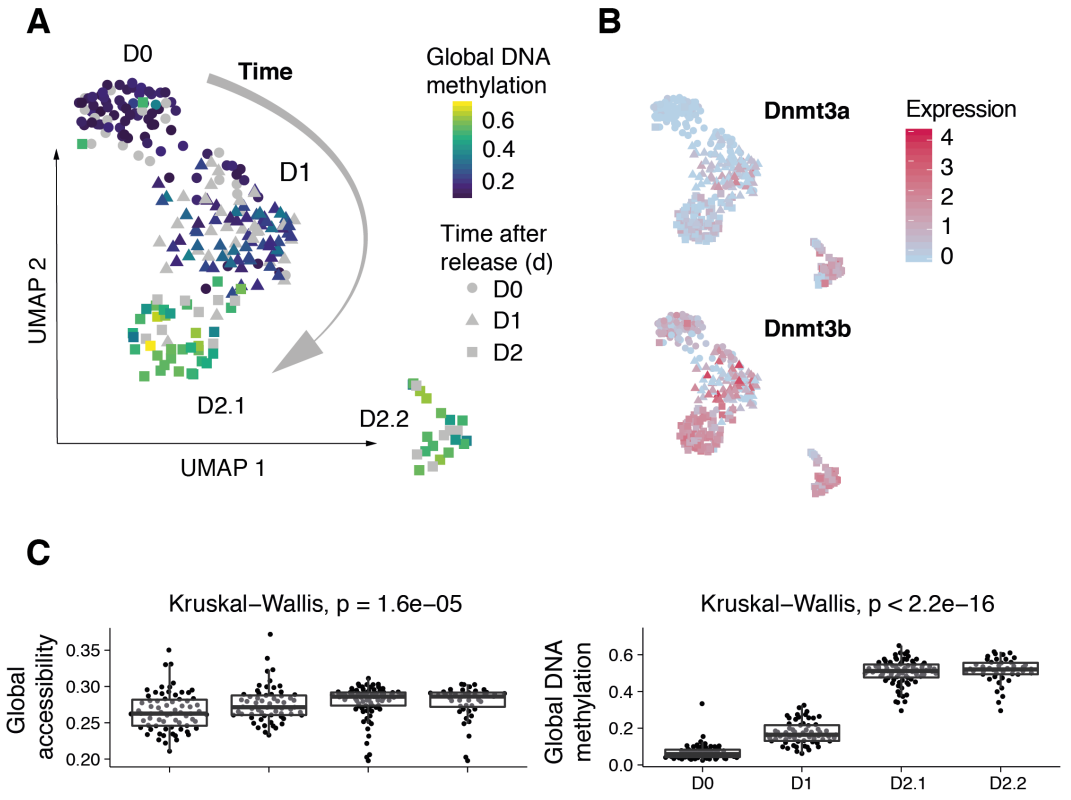


Figure 2.17.: (**A**) Low dimensional representation (UMAP) of the gene expression space for mESCs during *de novo* DNA methylation. Gradient colors are global averages of DNAme. (**B**) Normalized expression of Dnmt3a and Dnmt3b genes during development. (**C**) Changes over time of accessibility compared to DNA methylation.

to as accessibility. In Fig. 2.17 C we compare the first moment of accessibility (average accessibility) for different stages in development to the average DNA methylation (CpG methylation). DNA methylation increases in the same fashion as found in bulk BS-Seq, Fig. 2.4. GpC methylation increases as well, even though much slower. This result seems in contradiction with what we found previously. In order to understand this apparent contradiction, we have to bear in mind that we predicted that DNAme locally compact the DNA on very small structures, whilst this may not be the case if one considered structures on larger length scales. We can formalise this qualitative argument by simply considering the binding kernel. We indeed found that the fraction of accessible sites scales as $l^{2/3}$, with l the size of a compacted region. If we consider another methylation event inside this region, we expect the region to be split in two parts,(left an right) with respect to the newly methylated site. The two regions have sizes l_1, l_2 and of course $l_1 + l_2 \approx l$. These two regions are compacted as well, such that the total number of accessible sites is $l_1^{2/3} + l_2^{2/3}$. This must be compared with $l^{2/3}$. It is easy to check that $l_1^{2/3} + l_2^{2/3} > l^(2/3)\ \forall(l_1, l_2)$, when $l_1 + l_2 = l$.

Having theoretically explained this seemingly unexpected behaviour of global accessibility, we ought to find the behaviour of connected cross-correlation functions, which will be the last challenge to our model. To derive the cross-correlation we begin with the master equation (2.5) and introduce a complementary binary vector $\mathbf{a}$, $a_i \in \{0, 1\}$, which describes whether a site i is accessible ($a_i = 1$) or not ($a_i = 0$). We then couple this vector to the DNA methylation dynamics in the simplest form compatible with our model interpretation in physical space. We begin by considering the expectation value of the product $m_i a_j$ (cross-correlation) and for the sake of simplicity in the notation we take $i < j$. Cross-correlation are by definition

$$\langle m_i a_j \rangle = P(m_i = 1, a_j = 1) = P(a_j = 1 | m_i = 1) P(m_i = 1)\,. \tag{2.62}$$

$P(a_j = 1|m_i = 1)$ is the conditional probability that a site at position j is accessible whenever a site at position i is methylated. $P(a_j = 1|m_i = 1)$ cannot be computed directly as it implicitly depends on other values of $\mathbf{a}$ and $\mathbf{m}$.

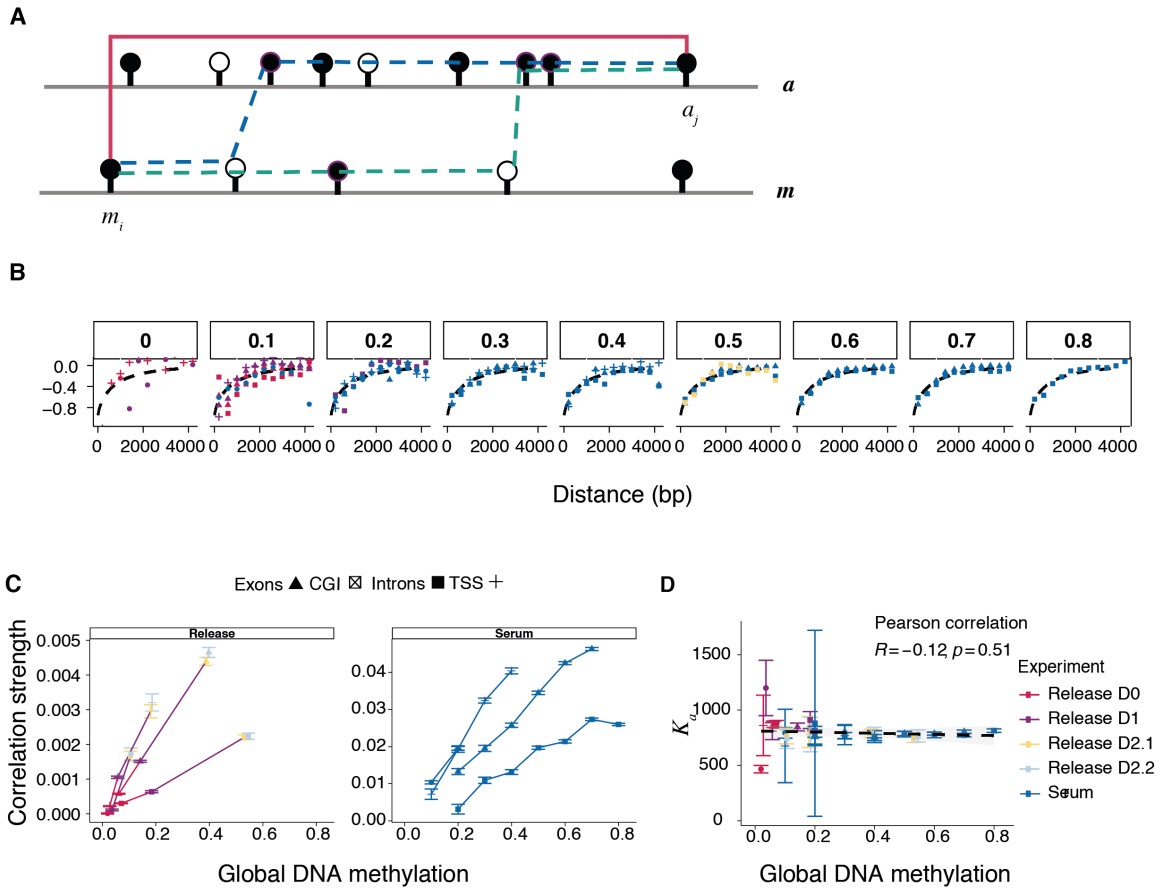


Figure 2.18.: (**A**) Methylation-accessibility model. The red line is the only relevant contribution for connected cross-correlations functions and all the other contributions (dashed lines) are irrelevant. (**B**) Theoretical prediction of connected cross-correlation functions, Eq. (2.68), are in excellent agreement with different experimental data (serum and 2i-release NMT-Seq). (**C**) The strength of cross-correlation functions is linear in the average methylation as predicted in Eq. (2.68) ($\alpha\langle m\rangle$). (**D**) The typical length scale K_a at which cross-correlation decays is independent of DNA methylation in agreement with Eq. (2.68).

To proceed, we therefore in a first step "integrate in" the random variable describing accessibility at position a_{j-1} ,

$$\langle m_i a_j\rangle = \sum_{a_{j-1}} P(a_j|a_{j-1},m_i)P(a_{j-1}|m_i)P(m_i=1)\,. \tag{2.63}$$

Reiterating this procedure for the second factor we find

$$\langle m_i a_j\rangle = \sum_{a_{j-1},a_{j-2}} P(a_j=1|a_{j-1},m_i=1)P(a_{j-1}|a_{j-2},m_i)P(a_{j-2}|m_i=1)P(m_i=1)\,. \tag{2.64}$$

We reiterate again these steps $|j-i|$ times and we obtain

$$\langle m_i a_j \rangle = \sum_{a_k, i \leq k < j} P(a_j|a_{j-1}, m_i) P(a_{j-1}|a_{j-2}, m_i) \dots P(a_{i+1}|a_i, m_i) P(a_i|m_i) P(m_i = 1) \,. \tag{2.65}$$

We are not yet in a position to give physical expression for the conditional probabilities as they have an unknown dependence on m_i. For simplicity, we assume for now that the mechanical coupling between base pairs is much stronger than the coupling of DNA methylation marks:

$$\langle m_i m_{i+1} \rangle \ll \langle a_i a_{i+1} \rangle \,. \tag{2.66}$$

Therefore, $P(a_j|a_{j-1}, m_i) \approx P(a_j|a_{j-1})$ for $i \neq j$. We will elaborate on this assumption in more details below.

In the DNAme model, the probability of binding decays according to Eq. (2.23). The physical interpretation of this kernel was that it reflects the binding probability of enzymes at compacted sites. In our model, the probability that a site is compacted then decays with the distance to the nearest methylated site to the power of $\lambda = -1/3$. Therefore, in order to reflect this kernel, the conditional probability $P(a_j|a_{j-1})$ should decay as $P(a_j|a_{j-1}) \propto a_{j-1}(1 - K_a/k^\lambda)$, where a_{j-1} plays the role of a delta function and k is the distance between a CpG and GpC site. Taken together, we obtain

$$\langle m_i a_j \rangle = P(m_i) P(a_i|m_i) \prod_{k=1}^{|j-i|} \left(1 - \frac{K_a}{k^\lambda}\right) \,. \tag{2.67}$$

In our derivation we implicitly assumed that the conditional probabilities $P(a_j|a_{j-1})$ do not explicitly depend on DNA methylation values at positions other than i. In principle, in order to derive simple expressions for the conditional probabilities we would have had to not only sum over intermediary positions in the accessibility vector, $\mathbf{a}$, but also over all positions in the DNA methylation vector $\mathbf{m}$, ultimately giving a sum over exponentially weighted paths between m_i and a_j Fig. 2.18 A. In the limit, where the mechanical coupling K_a is much stronger than the coupling between DNA methylation events, this sum over exponentials is dominated by the path with the highest contribution of K_a (orange line in Fig. 2.18 A). With $P(m_i) = \langle m_i \rangle$ and defining the local coupling between DNA methylation and accessibility $\alpha = P(a_i|m_i)$, an expression for cross-correlation functions is given as

$$\langle m_i a_j \rangle = \alpha \langle m \rangle \exp\left(-K_a |i-j|^{2/3}\right) \,. \tag{2.68}$$

In summary, we find that the strength of cross-correlation is linearly proportional to the average DNA methylation level. By contrast, the length scale of the decay, $K_a^{3/2}$, is

independent of average DNA methylation. Both results, as well as the functional form of the cross-correlation function, are in excellent agreement with the experimental data Fig. 2.18 B,C,D. We are now in a luxurious position where our theory predicts both DNAme correlation functions and cross-correlation between DNAme and chromatin accessibility for several sequencing experiments *in vitro*. Moreover we are able to predict the shape of this functions for different functional genomic regions, with a parameter free theory. This is suprising, maybe too surprising. How can a biological simple theory predicts higher order statistic in the whole genome? If that's true, it would mean that *de novo* DNAme is a simple process and no special information can be stored in a genomic region. Indeed, such a mechanism alone cannot encode biological information in DNAme patterns beyond the binding affinity of DNMT3 enzymes to the DNA and chromatin. This mechanisms alone, as we are going to show, does not allow to store information locally on the genome via DNAme. We then want to find if our theory cannot predict DNAme patterns in some genomic regions and then analyse them in more details. The regions individuated by the theory, will then be the one that are locally regulated and something more or different than our simple biological model is happening. In the next section, we will analyse *in vivo* data from the mouse embryo and identify which are those processes.

2.7. Anticipating symmetry breaking during exit from pluripotency via DNA methylation marks

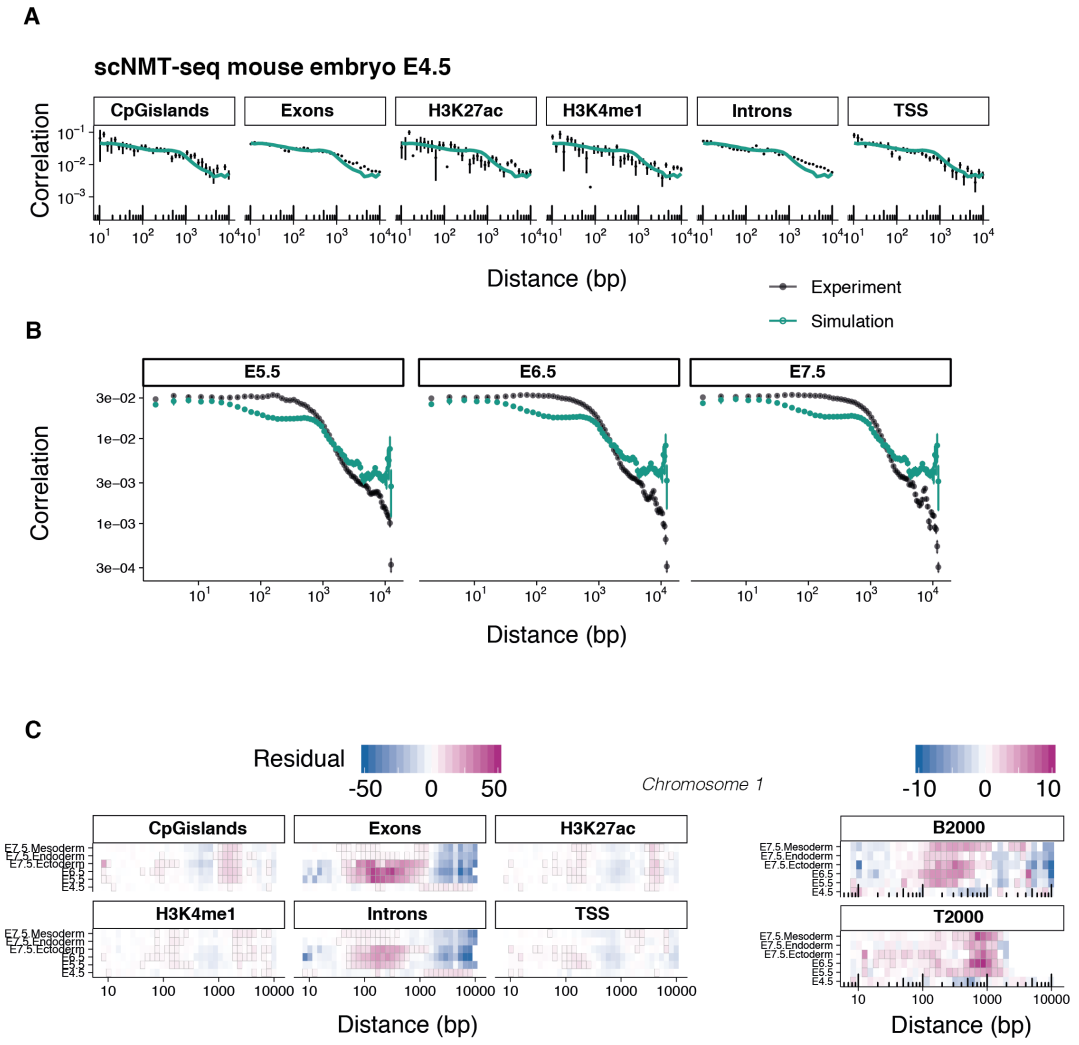


Figure 2.19.: (**A**) Comparison between theoretical and numerical connected correlation functions for mESCs sequenced 4.5 days after fertilization (E4.5). (**B**) Connected correlations functions for later stages of development (E5.5-E7.5). The two spatial regimes of the shape of connected correlation functions is preserved but there is an enrichment of correlations between 100-1000bps. (**C**) Enrichment of connected two points correlations with respect to the "null" model (E4.5). Each row represents cells in a certain stage during development and E7.5 cells are divided into mesoderm, ectoderm and endoderm. Residuals are based on the difference between numerical simulation and experimental data.

Having inferred and tested the kinetic rules governing *de novo* methylation in mESCs *in vitro* we then asked whether we could predict the establishment of 5mC marks during exit from pluripotency and early gastrulation in vivo. The model derived above

describes a single mechanism establishing DNAme genome-wide. Therefore, we expect that when cells become primed for differentiation from E5.5 and carry lineage-dependent DNAme patterns [6] additional mechanisms targeting DNA methylation must be in place. We reasoned that, by quantifying statistical patterns of deviations from our model describing generic, genome-wide DNAme dynamics ("null model"), we could identify genomic regions being specifically regulated by additional processes. To address this question, we analysed scNMT-Seq data from mouse exit from pluripotency and initial cell fate decisions during gastrulation [6] (Appendix B.4). As expected, the model predicts the distribution of DNAme marks in pluripotent cells at E4.5, Fig. 2.19 A. During later stages of development (E5.5-E7.5), when cells undergo cell fate transition, we observed systematic deviations between theory and experiments: even though correlation functions still roughly follow the pattern of two distinct short and longer-distance spatial regimes, we found an enrichment correlations in DNA methylation on a scale between 100 and 1000 bps, Figures 2.19 B. To examine whether this pattern occurs genome-wide or is restricted to specific genomic regions we systematically quantified the difference between theory and experiment (residuals), normalised by the experimental standard error, for any distance between CpGs and different genomic annotations. We found that the enrichment in correlated DNAme marks was specific to gene bodies and in particular to genes silenced between E5.5 and E7.5, but not active genes, Fig. 2.19 C.

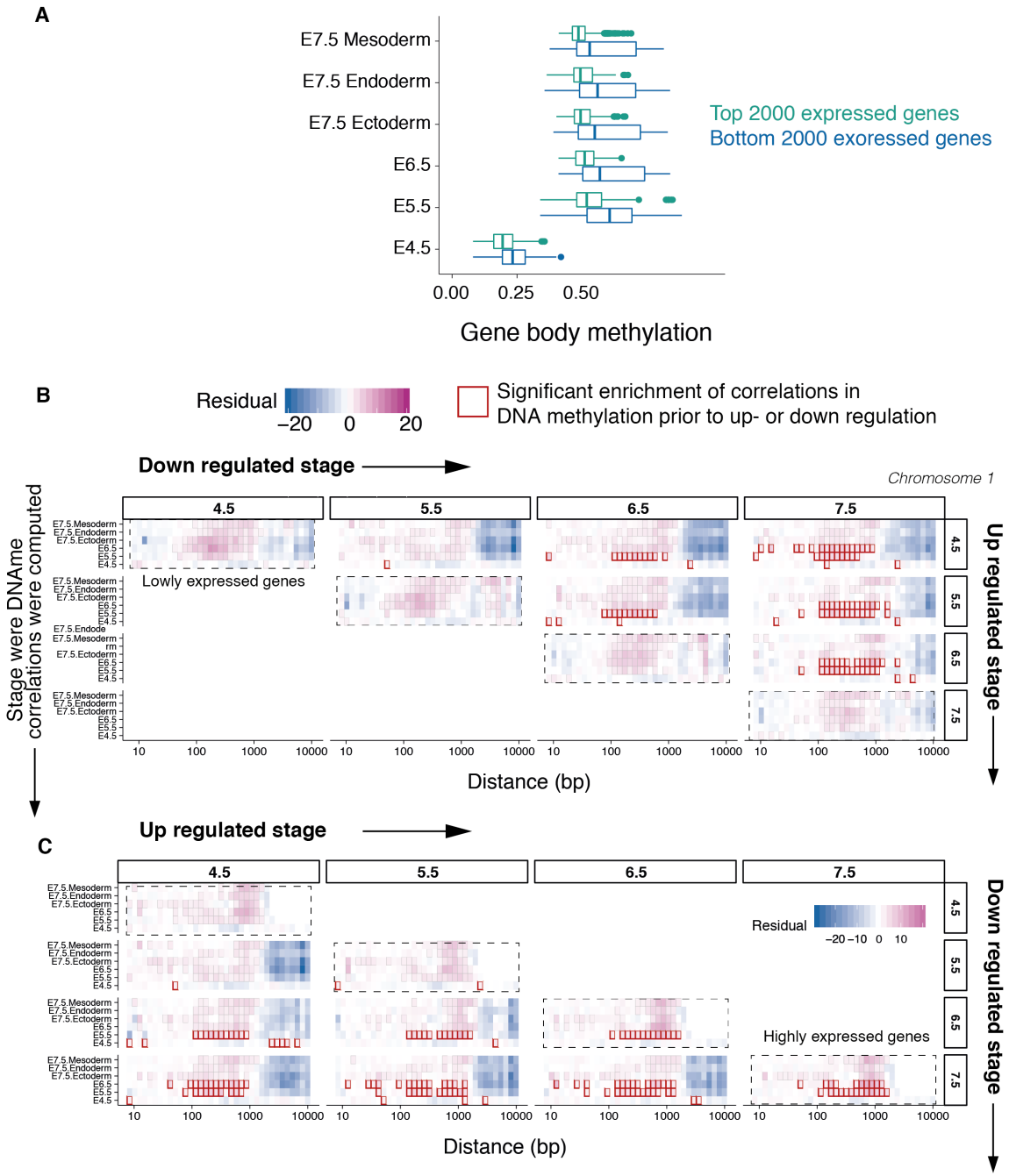


Figure 2.20.: (**A**) Boxplot of average gene body DNA methylation between active and inactive genes during development for all the cells in the embryo. Median, first and fourth quantile are plotted as a box plot, where the median is the horizontal thick line and quantiles are the two boxes heights. The p-value shows significant difference. (**B**) Residuals for groups of genes that are differentially downregulated between pairs of embryonic stages. (**C**) Residuals for groups of genes that are differentially upregulated between pairs of embryonic stages. Significant deviations in (**B**,C) are marked by red squares ($p<0.05$, t-test).

It is well known that differences in global levels of DNAme in gene bodies affect

transcriptional activity [117]. This effect may depend on CpG density and it is in general associated to promoter DNA methylation. Here, we found that promoters and CpG islands follow the biophysical theory defined the previous sections, whilst gene bodies are found to statistically deviate. In order to capture whether global gene body DNA methylation affects transcriptional activity or the main contribution comes from the distribution of the methylation marks independently on the average, we computed absolute levels of DNAme in active and silenced genes. We found that they differed only slightly Fig. 2.20 A and therefore cannot fully explain systematic deviation of gene bodies DNA methylation marks to theoretical predictions.

We then asked whether such deviation is a consequence of gene silencing between E5.5 and E7.5, or whether it temporally precedes the silencing of genes during differentiation. To this end, we determined differentially expressed genes between each pair of embryonic stages and calculated for each set of genes the enrichment or depletion in spatial correlations between DNAme marks in all stages and lineages. We found that for silenced genes which are downregulated between a pair of embryonic stages these changes in DNAme patterns emerge up to two days before changes in the transcriptome appear, suggesting that these marks could play an instructive role by priming the genes for silencing during differentiation Fig. 2.20 B. By contrast, we identified the DNAme pattern characteristic for active genes only after genes had been activated, but not before Fig. 2.20 C. We found that these patterns apply in particular to pluripotency genes Fig. 2.21 A, but also to a set of silenced genes which are not annotated as pluripotency genes Fig. 2.21 B. While polycomb (H3K27me3) [118–120] or H3K9me3 [121, 122] pathways might be candidates for premarking silenced genes, further mechanistic studies will be necessary to elucidate the detailed molecular pathways behind this process. Taken together, our framework to infer collective epigenetic processes involved in *de novo* methylation allows identifying epigenetic patterns preceding transcriptional silencing during differentiation. DNA methylation (5mC) is not the only epigenetic modification of CpG sites during exit from pluripotency, but as we outlined in the introduction (Sec. 1.2.1), enzymes other than DNMT3s, such as DNMT1s and TETs can modify the status of CpG sites. In the remaining part of the chapter, we extend the nonequilibrium enzyme kinetic models introduced in Sec. 2.3 and Sec. 2.4 to take into account the activity of different enzymes of the active methylation turnover.

Figure 2.21.: Residuals for gene bodies of pluripotency genes (**A**) and of of the top and bottom 2000 expressed genes excluding pluripotency genes (**B**).

2.8. Active turnover of DNA methylation

In the previous sections we showed how topological DNA changes lead to long-range interactions between CpG sites and hence to global changes in DNA methylation during the loss of pluripotency. We showed how DNMT3s enzymes cooperate and interact to deposit DNA methylation marks. On the other hand, methylation marks can be actively removed by TETs or passively by DNMT1s enzymes (Sec. 1.2.1) [32, 40]. Specifically, for epiblast cells that are beginning to exit from pluripotency, DNMT3s and TETs are coexpressed [31]. TETs enzymes modify the methylated cytosines oxidating 5mC to hydroxymethyl-cytosine (5hmC) then to formyl-cytosine (5fC) and finally to carboxyl-cytosine (5caC), such that CpGs sites perform a biochemical cycle passing through different states (Fig. 2.22 A), the first one being DNA methylation. In order to get further insights into changes of DNA methylation in a developing embryo, we thus need to study how this biochemical cycle (DNA methylation turnover) is regulated. In this section, we map the DNA methylation turnover to coupled stochastic non linear oscillators. Specifically, we develop a theory of oscillators with power law long range interactions and non local interaction kernel, which has the shape inferred in the previous sections. As these systems may show global and local synchronization, we analyse the conditions under which synchronisation arises.

2.8.1. Phase oscillators with restricted long-range interactions

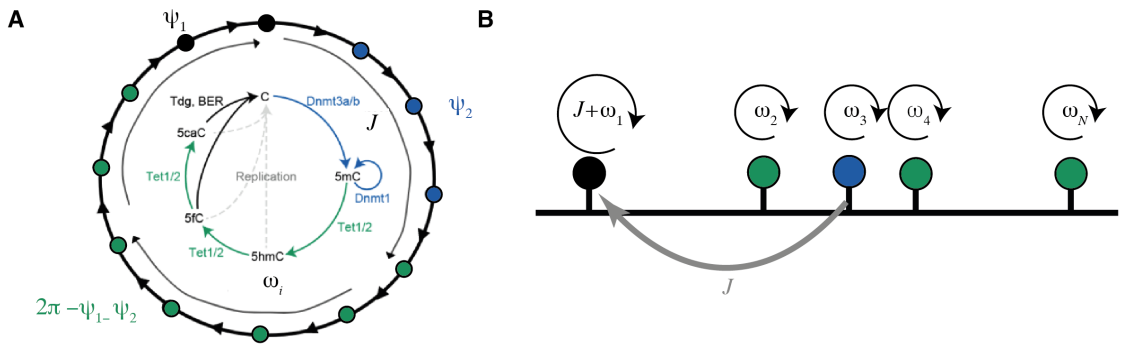


Figure 2.22.: The methylation cycle is divided into three phases: $\psi_1, \psi_2, 2\pi - \psi_1 - \psi_2$. CpGs in the first phase interact with restricted long-range interactions with CpGs that are in the second phase. CpGs also oscillate with an intrinsic frequency ω_i, where i indicates the lattice position.

CpGs sites perform a biochemical cycle passing through different states (Fig. 2.22 A) and as there are just a discrete number of these states, the DNA methylation turnover is effectively described by a discrete phase oscillator [123]. In the previous sections

we found that long range interactions arise between CpG, where a DNMT3 enzyme is bound and free CpG (C). This is the only form of interaction that we know and that is in accordance with experimental evidences, and we do not add other interactions between different states of the cycles. We can thus think of the biochemical cycle as a discrete phase oscillators interacting via long range interactions only in particular phases of the clock, ψ_1, ψ_2. These two phases, sketched in Fig. 2.22 B, are not necessarily defined by one discrete state, but they may incorporate multiple states. If we define a discrete oscillators with n discrete states, the "macrophase" ψ_1 is a shortcut for the first ψ_1 discrete states, ψ_2 for the next ψ_2 and the last phase is comprised of $n-\psi_1-\psi_2$ states. As an example, if $n = 20$, ψ_1 may stand for the first two states and ψ_2 for the next three. We anticipate that in the continuous limit, ψ_1 and ψ_2 take finite ranges and the last phase occupies a total range of the clock, $2\pi-\psi_1-\psi_2$. The master equation describing the DNA methylation turnover, which takes into account long range interactions as in Eq. (2.4), is (Fig. 2.22)

$$\begin{aligned} \partial_t P(\boldsymbol{\phi}) &= \sum_{i=1}^{N} \left[\left(\omega_i + k_i\left(\boldsymbol{\phi}, \phi_i - 1\right)\right) P(\boldsymbol{\phi}, \phi_i - 1) - \left(\omega_i + k_i\left(\boldsymbol{\phi}\right)\right) P(\boldsymbol{\phi}) \right] \\ k_i\left(\boldsymbol{\phi}\right) &= J \sum_{k=1/k\neq i}^{N} \delta_{\phi_i,\psi_1}\delta_{\phi_k,\psi_2} \frac{e^{-\rho_2|k-i|}}{|k-i|^{\lambda}} \\ \rho_2 &= \frac{1}{N}\sum_{j=1}^{N} \delta_{\phi_j,\psi_2}\,, \end{aligned} \tag{2.69}$$

where we omitted the implicit time dependency on $P(\boldsymbol{\phi})$ and we used a mean-field version of the interaction kernel. The delta function δ_{ϕ_i,ψ_2} and similarly for δ_{ϕ_i,ψ_1} in Eq. (2.69) is one whenever $\phi_i \in \psi_2$, meaning that the state ϕ_i belongs to one of the states identified by ψ_2, and it is zero otherwise. ρ_2 and ρ_1 are then counting functions of sites with discrete states in ψ_2 and ψ_1 respectively. We used the notation $f(\boldsymbol{\phi}, \phi_i \pm 1)$ to compactly indicate the state $[\phi_1, \ldots, \phi_i \pm 1, \ldots \phi_N]$. The frequency ω_i denotes the intrinsic oscillations rate of a given CpG i and it is proportional to the typical time that a CpG site i takes to complete the biochemical cycle. The lattice sites indicate, as in the previous section, the topological position of each CpG along the DNA, such that $i = 1$ is the first CpG , $i = 2$ the second, etc... Eq. (2.69) fully defines the dynamics of the oscillators, but it has no exact solution and so we proceed with a system size expansion. The key idea is to divide ϕ into a stochastic and a deterministic part and study their coupled dynamics. By doing so, we formally make an analytical continuation of the discrete clock [124] given by

$$\phi_i = \Omega\Phi_i(t) + \Omega^{1/2}\xi_i(t)\,, \tag{2.70}$$

with Ω, the number of states of the clock. After equating equal order in Ω in the master equation, Appendix D.1, we arrive to a Langevin equation describing the time evolution of the phase of each clock:

$$\frac{d\phi_i}{dt} = w_i + f_1(\phi_i) \sum_{k=1,k\neq i}^{N} \frac{Je^{-\rho_2|k-i|}}{|k-i|^{\lambda}} f_2(\phi_k) + \sqrt{2\omega_i}\xi_i(t), \tag{2.71}$$

$\xi_i(t)$ is a Gaussian white noise with zero mean and unitary variance, $\langle \xi_i(t)\xi_j(t')\rangle = \delta(t-t')\delta_{i,j}$ and we promoted δ_{ϕ,ψ_i} to be general functions of the phases $(f_1(\phi), f_2(\phi))$ and it will be discussed later. Eq. (2.71) has the form of a stochastic Kuramoto model [125, 126] with partial interactions (only between two phases of the clock) and screened long range interactions. One hallmark question on such process is whether and how this system achieves synchronisation. In Fig. 2.23 we show numerical simulations of Eq. (2.71) without noise and we can already notice the particular biochemical patterns arising from such interactions. The methylation clocks seem to be synchronised in a band like structure, such that groups of CpG oscillate together and others seem to not be synchronised at all. This phenomena may remind chimera states observed in spatially interacting Kuramoto oscillators in one dimension [127].

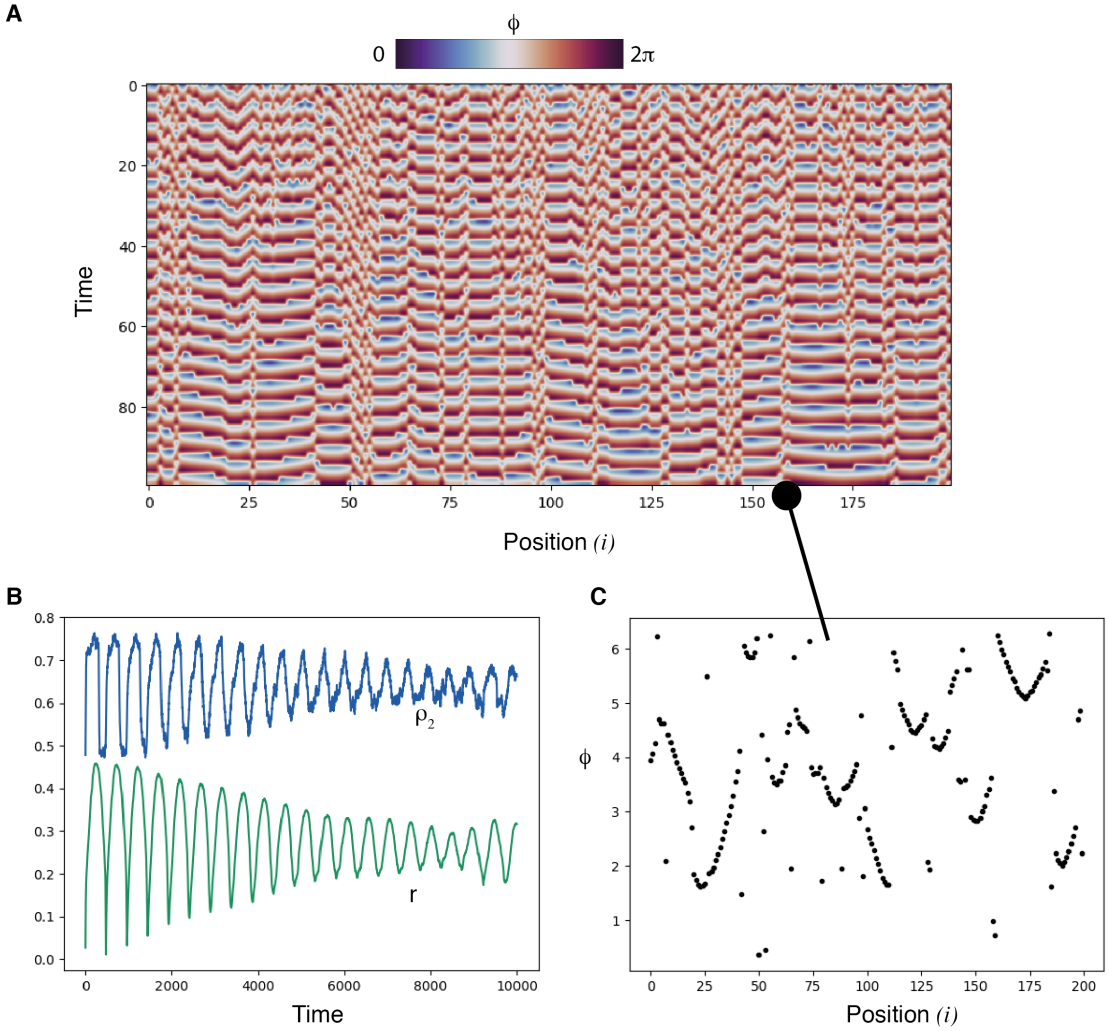


Figure 2.23.: (**A**) Synchronization of the deterministic part of Eq. (2.71) for $g(\omega) = \delta(\omega - 1)$ and $J = 10$ starting from an asynchronous state.(**B**) Dynamics of the Kuramoto order parameter r (green) and the average DNA methylation ρ_2 (blue). These functions exhibit a sustained oscillations in time. (**C**) Individual oscillators phases ϕ_i after 10^5 time steps with $dt = 0.01$. The lattice is divided into separated regions of synchronised oscillators.

Analytically, in order to see at which values of the parameters this model can exhibit a transition from an asynchronous state to a synchronized one, we need to study the behaviour of the Kuramoto order parameter

$$r(t)e^{i\psi(t)} = \frac{1}{N}\sum_i e^{i\phi_i(t)}\,. \tag{2.72}$$

To do so we first define the generator of moments [128–130],

$$H^m_{k,q} := \frac{1}{N}\sum_{j=1}^{N} \overline{\langle e^{ik\phi_j} e^{iqj}\rangle w_j^m} \tag{2.73}$$

where $\overline{(.)}$ is the average over the distribution of intrinsic frequency ω_i and $\langle(.)\rangle$ is an average over the possible realisation of the noise. In the previous equation we employed a Fourier transform of the fields. Hence, k is dual to ϕ and q is dual to the DNA sequence. Upon defining a function $\chi(\theta, y, z, t)$ as

$$\chi(\theta, y, z, t) = \sum_{k=-\infty}^{\infty} \sum_{m=\infty}^{\infty} \sum_{q=-\infty}^{\infty} e^{-ik\theta} e^{-iqz} \frac{y^m}{2\pi m!} H_{k,q}^m, \tag{2.74}$$

its time evolution is given (Appendix D.2) by

$$\partial_t \chi(\theta, y, z, t) = -\frac{\partial}{\partial \theta}[\nu(\theta, y, z, t)\chi] + \frac{\partial}{\partial y}\frac{\partial}{\partial \theta} D(y) \frac{\partial}{\partial \theta}\chi(\theta, y, z, t) - \frac{\partial \chi}{\partial \theta \partial y}\,, \tag{2.75}$$

where $\nu(\theta, y, z, t) = y + J\left[f_1(\theta) \int d\hat{\theta} \int d\hat{z} \frac{e^{-\rho_2 \hat{z}}}{|\hat{z}|^\lambda} f_2(\hat{\theta}) \chi(\hat{x}, y = 0, z - \hat{z}, t)\right]$ is a drift-term and $D(y) = 2y$. Defining $\rho(\theta, \omega, z, t)$ as the density of oscillators with phase θ, position z and frequency ω at time t, and taking $y = 0$ (deterministic limit of Eq. (2.71)), its continuity equation is given by exploiting the relationship $\chi(\theta, y, z, t) = \int d\omega g(\omega) e^{y\omega} \rho(\theta, \omega, z, t)$ as

$$\begin{aligned}
&\frac{\partial \rho(\theta, \omega, z)}{\partial t} = -\frac{\partial}{\partial \theta}\left[\tilde{\nu}\rho(\theta, \omega, z, t)\right]\right] \\
&\tilde{\nu} = \omega + J f_1(\theta) \left[\int\int\int d\omega d\hat{z} d\hat{\theta}\, g(\omega) \frac{e^{-\rho_2 \hat{z}}}{|\hat{z}|^\lambda} f_2(\hat{\theta}) \rho(\hat{\theta}, \omega, z - \hat{z}, t)\right] \\
&\rho_2 = \frac{\int\int\int d\theta' d\omega' dz' g(\omega') f_2(\theta') \rho(\theta', \omega', z')}{\int\int\int d\theta' d\omega' dz' \rho(\theta', \omega', z') g(\omega')} \\
&re^{i\psi} = \int\int\int d\theta d\omega dz \rho(\theta, \omega, z) g(\omega) e^{i\theta}\; z \in (0, 1),
\end{aligned} \tag{2.76}$$

where $g(\omega)$ is continuous form of the distribution of intrinsic frequencies ω_i.
As $\int d\theta \rho(\theta', \omega', z') = 1$ and $\int d\omega g(\omega) = 1$, the denominator of ρ_2 is always one. Eq. (2.76) is a self-consistency equation for the density as ρ_2 depends on ρ. In order to close the equation we need to find the functional form of f_1, f_2. As in the original discrete oscillators model, f_1 and f_2 were delta functions over a set of consecutive states ($f_{1,2} = \delta_{\phi, \psi_{1,2}}$), their continuous form is $f_1(\theta) = H(\Theta(\psi_1) - \mathrm{mod}(\theta, 2\pi))$ and $f_2(\theta) = H(\mathrm{mod}(\theta, 2\pi) - \Theta(\psi_1))H(\Theta(\psi_2) - \mathrm{mod}(\theta, 2\pi))$, where $H(x)$ is the Heaviside step function and $\mathrm{mod}(a, b)$ is the rest of the division a/b and $\Theta(\psi_{1,2})$ is the value of the last discrete state of the phase $\psi_{1,2}$. As an example, if the clock is composed of 12 discrete sites and ψ_1 cover the first three sites, $\psi_1 = \frac{\pi}{2}$. In the following, whenever we refer to biological control function we take the last definition of f_1, f_2 with $\psi_1 = [0, \frac{\pi}{2})$ and $\psi_2 = [\frac{\pi}{2}, \frac{3\pi}{2}]$.

2.8.2. Stationary solutions of synchronised states

Eq. (2.76) does not admit asynchronous stationary solutions. The main reason is that the control functions break the rotational symmetry such that the equations are not invariant under a shift of the fields (oscillators). Indeed, the first integral in Eq. (2.76) will never vanish for an asynchronous solution as long as $\int \mathrm{d}\theta f_2(\theta) \neq 0$. Moreover, we have to consider that for biological consistency, the distribution of frequencies for each oscillators has to be positive. Generally this problem is solved by moving to a rotating frame with a certain angular velocity, typically the median of $g(\omega)$ [130]. In our specific problem, due to the lack of rotational symmetry, this transformation does not simplifies the analytical calculations. For general functions $f_1(\theta_i)f_2(\theta_j)$ is sometimes possible to change to a coordinate system which allows the asynchronous solution to be the stable solution [131], as long as the product $f_1(\theta_i)f_2(\theta_j)$ can be written as $g(\theta_i - \theta_j)$, with g a function that is 2π periodic and well defined. As $f_{1,2}(\theta)$ in the problem we are considering are step functions, it is not always possible to write in the rotationally symmetric form g. Moreover, as the asynchronous solution is never a stable solution, a stability analysis as in [132, 133] or a power series expansion [134] around this state does not give any useful insight for this process, as the stationary states are non-trivial even for vanishing strength of interactions J. We thus need to look for an order parameter different from the Kuramoto one which is even experimentally more accessible. A natural order parameter is ρ_2, which biologically is the average DNA methylation. For the biological control functions, $\rho_2 = 1/2$ in a phase where the interactions do not play any role on synchronization and should be greater than $1/2$ whenever $J > J_c$. We thus first seek for a stationary solution of Eq. (2.76) and we found out that there are two possible solutions. One solutions where oscillators are phase locked at frequencies such that $\tilde{\nu} = 0$ and a second solution where oscillators rotate in a non collective manner around these frequencies and they will satisfy the stationary condition $\tilde{\nu}\rho = C(\omega)$, where $C(\omega)$ is a constant that is determined upon normalization [130]. All together we find that

$$\begin{aligned} \rho &= \begin{cases} \delta\left[\omega + J\rho_2^{\lambda}\Gamma(1-\lambda,0,\rho_2)H(\pi/2-\theta)\right] & \omega \in [-J\rho_2^{\lambda}\Gamma(1-\lambda,0,\rho_2),0] \\ \frac{C(\omega)}{|\omega+H(\pi/2-\theta)A(\lambda)J|} & \text{elsewhere} \end{cases} \\ A(\lambda) &= \left(\pi \int d\omega g(\omega)C(\omega)/\omega\right)^{\lambda} \int_0^{\pi \int d\omega g(\omega)C(\omega)/\omega} \mathrm{d}y e^{-|y|}/|y|^{\lambda}\,, \end{aligned} \tag{2.77}$$

where $\Gamma(1-\lambda,0,\rho_2) = \int_0^{\rho_2} \mathrm{d}y e^{-|y|}/|y|^{\lambda}$ and $H(x)$ is the Heaviside step function. $C(\omega)$ is set by the normalization such that, $\int_0^{2\pi} d\theta\rho = 1$ and we obtain, $C(\omega) = \frac{2|\omega||A(\lambda)J+\omega|}{\pi(|\omega|+3|A(\lambda)J|)}$. To get some useful analytical insights regarding changes in the synchronization of the system, we compute ρ_2 from its definition, Eq. (2.75) and inserting the stationary distribution ρ from Eq. (2.77). The first branch of solutions gives a null contributions

as $g(\omega)$ has a domain $\omega \in [0, \infty]$ such that

$$\rho_2 = 2 \int_0^\infty d\omega \frac{A(\lambda)J + \omega}{3A(\lambda)J + 4\omega} g(\omega) \,. \tag{2.78}$$

As there are no clear symmetries that can be exploited, this integral does not have a simple solution for a general distribution of frequencies. We initially work in the small J limit. As $A(\lambda)$ depends on J, we can not simply expand the integrand in series of J, but we need to find first the solution of the self consistency equation for A and later expand. In particular, we find that $A(\lambda) = \rho_2^\lambda \Gamma(1-\lambda, 0, \rho_2)$ and we can plug this result back into the self consistency equation for ρ_2. The expansion of the integrand in Eq. (2.78) in powers of J has a divergency $1/\omega$ such that the integral might be not well defined for a general distribution of frequencies. To avoid such difficulties, we fix the distribution of frequencies to be an exponential with mean μ for which the exact self consistency equation for ρ_2 is

$$\rho_2 = \frac{1}{8}\left[4 - \rho_2^\lambda \Gamma[1-\lambda, 0, \rho_2] J\mu e^{\frac{3\rho_2^\lambda \Gamma(1-\lambda,0,\rho_2)J\mu}{4}} \mathrm{Ei}\left(-\frac{3}{4}\rho_2^\lambda \Gamma(1-\lambda, 0, \rho_2) J\mu\right)\right] , \tag{2.79}$$

where $\mathrm{Ei}(x) = -\int_{-x}^{\infty} \mathrm{d}y e^{-y}/y$ is the exponential integral function. In the low interaction limit $J \to 0$ we find $\rho_2 = \frac{1}{2}$ and for high interaction $J \to \infty$, $\rho_2 = \frac{2}{3}$. The first of this limit is clear as for no interactions ρ_2 simplifies to $\rho_2 = \int \mathrm{d}\theta f_2(\theta)/2\pi$, which for the biological control function gives the value 1/2. On the other hand when J goes to infinity, oscillators that are in the phase ψ_1 - the only one affected by interactions - will pass an infinitesimal time in that phase such that the density of oscillators is effectively restricted in the phase domain $[\pi/2, 2\pi]$. The oscillators in this last domain experience no interactions such that on average $\rho_2 = \int \mathrm{d}\theta f_2(\theta)/(3\pi/2)$, which for the biological control function gives the value 2/3. Eq. (2.79) for finite and non-zero values of J cannot be further simplifies, but it can be solved numerically and in Fig. 2.24 A we show the analytical prediction compared to numerical simulations for different system sizes (number of oscillators).There is still a little discrepancy between numerical simulations and theoretical results (Fig. 2.24 A inset) and we argue that it is caused either by finite size effects as the numerical curves approach the theoretical one for increasing values of the system size or by the formation of spatial structures such that the homogeneous solution is never stable. Only a systematic study of non-homogeneous solutions will tell us which hypothesis is correct. The Kuramoto order parameter r, once the value of $A(\lambda)$ is known, is found numerically from equations (2.77) and (2.76). Specifically, as ρ_2 saturates at a constant value, r can never been 1 such that full synchronization is never achieved. In Fig. 2.24 B we show numerical simulations of Eq. (2.76) for an exponential distribution of frequency with $\mu = 1$ and in Fig. 2.24

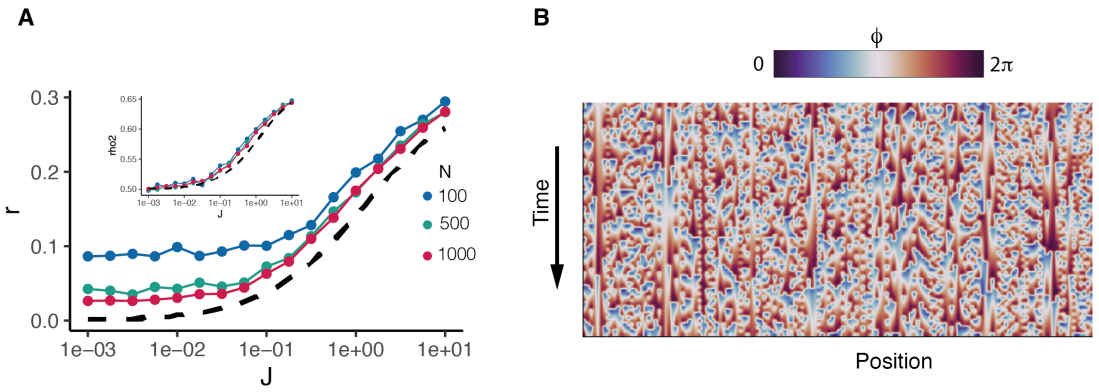


Figure 2.24.: (**A**) Kuramoto order parameter $\langle r \rangle$ and $\langle \rho_2 \rangle$ (inset) as a function of the interactions strength J. The average ($\langle \ldots \rangle$) is taken over 100 realisations of numerical simulations of the model in Eq. (2.71). Different colours represent different system sizes, i.e. the number of lattice sites (N). Theoretical predictions (2.79) and (2.76) are shown as a black dashed line. Numerical simulations of Eq. (2.71) are performed using a 4th order Runge-Kutta algorithm with $\lambda = 1/3$ and the frequencies ω_i are random numbers drawn from an exponential distribution with unitary mean. In (**B**) we show one realisation of the simulations with parameters $J = 1, N = 1000$ (150 sites are shown) and a 4th order Runge-Kutta algorithm with time step $dt = 0.01$.

A we compare theoretical prediction to numerical values of ρ_2 and r. Here, we found that methylation turnover does not exhibit any transition between an asynchronous to synchronous state at finite values of the interactions strength (J) between CpG sites (oscillators). In particular, we found that synchronization and methylation increase non-linearly with respect to the interaction strength and there is no evidence general scaling with respect to the frequency distribution of oscillators. Moreover, we found that full synchronization, in case of partially interacting oscillators, can never be achieved as the average DNA methylation ρ_2 saturates at a constant value.

2.8.3. Partial synchronization in the genome

From numerical simulations we see that long range interactions leads to particular phase-locked spatial structures along the genome Fig. 2.24 A, which are stable in the long time limit. Indeed, we found out that $r(t)$, which measure the degree of synchronization saturates at a constant level lower than one, even for high value of the interactions J Fig. 2.24 B. This behaviour, which can only be explained through an analytical study of Eq. (2.71) beyond linear stability has an important biological consequences: CpG sites will never synchronised genome-wide, making it possible to easily change local structure if needed, without having to change them genome-wide, such that this model make DNA methylation plastic against perturbation and stable at the same time. In order to exploit if the genomic distribution of CpGs plays an important role for synchronisation, we did numerical simulations of Eq. (2.71) with distances be-

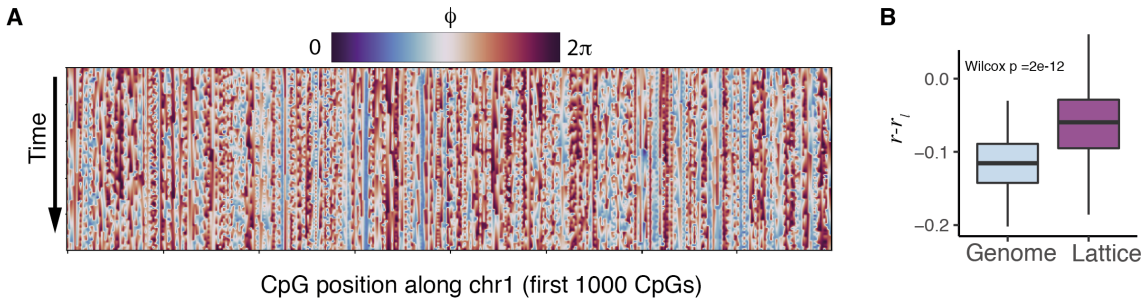


Figure 2.25.: (**A**) Numerical simulations of Eq. (2.71) with a 4th order Runge-Kutta algorithm ($dt = 0.01$) where the lattice distances are taken from the CpG distribution of chromosome 1. We rescaled each distances such that the average distance is 1, to make it comparable with the lattice model. (**B**) Comparison of the difference between global and local (Eq. (2.80)) Kuramoto order parameter for lattice simulation and simulation on the genome with $n_l = 10$. Median, first and fourth quantile are plotted as a box, where the median is the horizontal thick black line and quantiles are the two boxes heights. The p-value shows significant difference.

tween sites taken from chromosome 1 of the mouse genome. A part from a qualitative difference between genome simulations and lattice based ones (Fig. 2.25 A), we ought to find an observable which better describes local synchronization. To this end we define a local order parameter r_l as

$$r_l = \frac{1}{L/n_l} \sum_{k=0}^{(L-n_l)/n_l} \frac{1}{n_l} \sum_{j=kn_l}^{(k+1)n_l} e^{i\phi_j} , \tag{2.80}$$

where the first CpG position is at $j = 0$. We argue that these defined parameter catches local changes in synchronisation with respect to global ones. We found that the local order parameter is significantly higher for genome simulations with respect to lattice one (Fig. 2.25 B), suggesting the possibility that the distribution of CpGs along the genome favours local synchronisation, further strengthening the difference already introduced by long range interactions between local and global synchronisation.

2.9. Summary and discussion

In this chapter we applied methods from non equilbrium statistical physics to infer emergent spatio-temporal processes from multi-omics genomic experiments. In particular, we apply our theory to early dynamics of DNAme during development before one of the first symmetry breaking events in the mouse embryo. This transition, happening between E5.5 and E6.5 after implantation, sets the exit of epiblast cells out of the naive pluripotent state. In Sec. 1.2.1 we analysed 2i-release experimental data of mouse embryonic stem cells and upon studying the increase of the average DNA methylation

and the spatial arrangement of methylated sites, we show how DNA methylation is established via a collective mechanisms involving long range interactions. The data suggest that these interactions are non trivial as collective degrees of freedom emerge during the entire process of *de novo* methylation. Puzzled by these findings, in Sec. 2.3 and Sec. 2.4, we derived a theory of out of equilbrium theory of enzyme kinetics with general and unknown interaction kernel between enzymes, which we applied to unveil the mechanisms of *de novo* DNAme. Upon writing the resulting master equation, deriving the path integral formulation and finally applying renormalization group methods, were able to infer the kernel of interactions between DNMT3 enzymes (responsible of *de novo* DNA methylation). Moreover, our theory correctly predicts the increase of the average methylation in time before cellular symmetry breaking as well as the shape of connected correlation functions for several functional genomic regions (Sec. 2.6). Our theory shows that *de novo* DNA methylation is established via an interplay between enzymes binding kinetic and local chromatin structure. In Sec. 2.5 we apply a newly developed geometrical renormalization scheme to infer the dynamics in the three dimensional space of the nucleus (physical space) from one dimensional sequencing data (sequence space). We were able to show that there is a positive between between DNAme and chromatin compaction, predict the size of local chromatin structures (condensates) and how they emerge from interactions. We tested our results on several experimental data, which confirmed our theory. In 2.7, we challenge our theory with *in vivo* experiments. In particular we wanted to know whether the predicted mechanisms are broken locally, such that specific information can be encoded in the genome via DNAme. We found out, that these mechanisms is indeed broken along gene bodies upon exiting from pluripotency. Specifically, we show that gene bodies of genes that are going to be downregulated have a different arrangement of methylation marks (irrespective of the mean) with respect to other functional genomic regions or to genes that are going to be upregulated, which instead follow the general mechanism predicted by out theory. Surprisingly, we show that the specific arrangement of methylation marks appear two days before the actual shut down of the gene and the consequent cell fate transition. All together we were able to derive a systematic theory to unveil mechanisms in the three dimensional space of the nucleus from one dimensional sequencing data. Our theory is able to predict several genomic observables in the sequence space as well as in the physical space, from *in vitro* and *in vivo* data of mouse embryo. The theory, which we apply in the context of *de novo* DNA methylation, can be generalized to any other epigenetic process, hence providing a general framework to infer enzyme kinetics and emergence of collective degrees of freedom from sequencing data. Finally, in Sec. 2.8 we study the combined role of different enzymes, DNMT3 included, that modify the status of CpG base pairs (Sec. 1.2.1). In particular, by taking the long range interaction nature of the process as derived in the previous sections, we were able to map the

methylation cycle of CpG sites as a Kuramoto model with non-linear, non-local and partial phase interacting kernel. We analysed when synchronisation can be achieved in such systems, and found out that the long range interactions facilitate synchronization of local structures, yet preventing full synchronization of the genome.

3. Scaling and Memory during Transcriptional Activity

In Chapter 2, we showed how DNA methylation marks are established through the interplay between DNA methylation and chromatin conformation. In particular, by combining novel methods from single-cell genomics and nonequilibrium field theories we were able to unveil processes underlying the formation of the embryonic methylome. This distribution was encoded in the shape of connected correlation functions, which exhibit scale free behaviour. In Sec. 2.7 we found that there are regions along the genome that break this general mechanisms such that their correlation functions are not scale free. We found that regions were these mechanisms is broken are gene bodies prior to downregulation during early development. Puzzled by these findings, our aim is to have a better biological and theoretical understanding of the interplay between gene body DNA methylation and gene expression. In particular, in Sec. 3.1 we demonstrate scaling relations between gene expression, DNA methylation and gene length. To understand the biophysical origin of these relations, we constructed a biophysical model based on collective polymerase movements on the gene bodies coupled to dynamic binding-unbinding of DNMT3s typically acting as obstacles to the polymerase movement [117]. The model can successfully address the observed scaling relation generally without any free parameters. We found, in accordance with the results of the previous section, that the correct scaling exponents are observed when taking into account how the three dimensional distance between base pairs along the DNA scales with respect to their linear distance. The observed scaling is reminiscent of a self avoiding random walk of the polymer in three dimensions and it is responsible for the observed scaling in the methylation levels revealed by analytical calculations and simulations. Even though the model correctly predict scaling exponents it does not capture changes in transcriptional output with respect to changes in DNAme patterns, with same global methylation levels. In Sec. 3.2 we extend the model by allowing memory effects of RNA polymerases during transcription and found that they can possibly explain the difference in methylation patterns between downregulated and upregulated genes.

3.1. Derivation of scaling laws

In order to explore the interplay between transcription and DNA methylation dynamics at the genome scale, we analysed recently published scNMT-Seq datasets in mouse embryonic stem cells which provide the mRNA and DNA methylation level at genome scale with single-cell resolution [16]. Obtaining a biophysical understanding of the interplay between DNA methylation and gene expression is complicated by the fact that scRNA-Seq experiments produce non-stochiometric read outs (Sec. 1.2.3). To infer quantities predictable by biophysical theories we used a simple stochastic model for gene expression with a poison-beta approximation to fit the distribution to obtain the parameter values of the model for every individual gene [135].

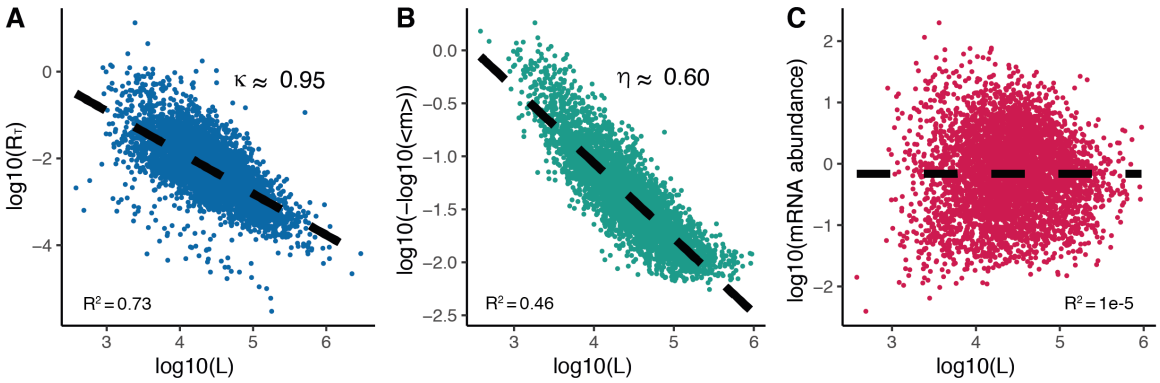


Figure 3.1.: Different scaling relationships are plotted for individual genes (dots) that are left after filtering procedures averaged across cells. The adjusted R^2 values of linear regressions are indicated at the bottom of each graph. (**A**) Scaling between the elongation rate (R_T) and genes length (L) and between average methylation and genes length (**B**), where the slopes of the linear fit are $-\kappa \approx 0.95, -\eta \approx 0.6$ respectively. (**C**) mRNA abundance does not show any clear scaling with respect to the gene length ($R^2 \approx 1e^{-5}$). All the exponents are computed by fitting a linear regression model (lm function in R).

mRNA molecules are produced every time a polymerase ends its activity by reaching the end of the genes that it is transcribing. Intuitively, the longer the genes, the smaller would be the transcription elongation rates (RT) in case there is a roughly constant and equal number of polymerases for each genes. We first preprocess and filter data as done in [16] (Sec. 1.2.3). We then compute the transcription elongation rates (RT) for the set of genes that are left in our analysis and show that elongation rates are negatively correlated with the length of the genes (L) such that $\log(R_T) \sim -\kappa \log(L)\,, \kappa \approx 0.95 \pm 0.02$ (Fig. 3.1 A) .

However, the average mRNA level does not display any such correlation with the gene length (Fig. 3.1 C) and this can be explained by the fact that the average mRNA level depends on the promoter ON/OFF rates which are not influenced by the length of the

gene. The dependency of the elongation rate with respect to the length suggests that the number of polymerases is roughly conserved for every gene and that the timescale of movement of the polymerases on the gene is much slower than the transcription initiation rates on the promoter. We then proceed to compute the average methylation for each gene. The average DNA methylation, computed as number of methylated CpGs divided by the total number of CpGs, should not have any dependence on the gene length. Surprisingly, this is not the case (Fig. 3.1 B) as the average methylation $\langle m \rangle$ has a stretched exponential dependency on L given by $\log(-\log(\langle m \rangle)) \sim -\eta \log L$, $\eta \approx 0.60 \pm 0.01$.

As of the output of single-cell sequencing experiments, we can correlate scaling exponents of each individual cell with the mRNA levels of individual genes belonging to the same cell. This procedure is similar to what we did in Sec. 2.7. In particular, we test whether the scaling relationships hold in every individual gene of every individual cell. We found that there are few genes which show high correlation with the scaling exponents (Fig. 3.2 A). In order to find out which genes are then responsible for the observed scaling, we performed a gene ontology enrichment analysis [136] taking the top 100 genes according to their correlation values. We observed that genes associated with chromatin silencing and epigenetic factors appear predominantly at the top of the enrichment analysis Fig. 3.2 B. These results are again in line with the previous chapter, pointing at an interplay between chromatin structure and scaling of transcriptional elongation rate.

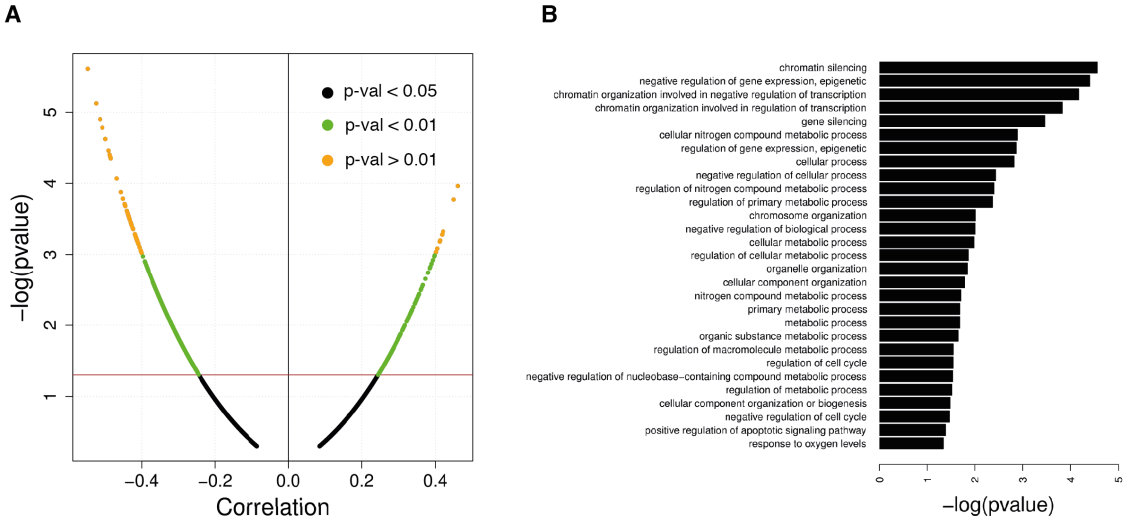


Figure 3.2.: (**A**) Spearman correlation between individual gene exponents (dots) and predicted scaling exponent. Few genes are significant (above the red line) for which the p-value is gretear than 0.05. Colors indicates p-values thresholds. (**B**) Identification of the significant genes which are found to be mostly associated with chromatin silencing and epigenetic factors.

In order to investigate the mechanism for the observed scaling, we constructed a

biophysical model based on the totally asymmetric exclusion process (TASEP) [117, 137] for the movement of the polymerases coupled with dynamic binding-unbinding of methyl binding enzymes. The methyltransfer enzymes can not bind to the CpG site unless it is devoid of any polymerase and, similarly ,the polymerase can move to a CpG site only if it is not occupied by an enzyme thus, posing as an obstacle to the movement of polymerases. These features of the model constitute a two way competition between the polymerase and enzymes kinetics (Fig. 3.3 A). We consider that there are N_T total RNA polymerases, assumed constants for the reasons explained before. Each RNAP can be in two states, bound or not bound. We define N_f the number of free or unbound RNAP and $N_b = N_T - N_f$ the number of bound RNAP. In particular, RNAPs bind at the promoter with a rate α, such that the total binding rate is αN_f. This binding rate sets a time scale τ between two successive RNAP binding events, $\tau = \frac{1}{\alpha N_f}$. If the RNAP slides along the gene body with a constant speed r calculated in bps/s, then the typical distances between two RNAP is $d = r/\tau$. For a gene of length L, the average number of bound RNAP is, $N_b = LN_f\alpha/r$ and as $N_b + N_f = N_T$, the number of free RNAP as a function of N_T is

$$N_f = \frac{N_T}{1 + L/L_0}\,, \tag{3.1}$$

with $L_0 = r/\alpha$, which is a characteristic length scale. The elongation rate (R_T) is by definition the inverse of the time between two consecutive mRNA production events and it is given by

$$R_T = \frac{1}{\tau} = \frac{1}{d/r} = \frac{\alpha N_T}{1 + L/L_0}\,. \tag{3.2}$$

As L_0 is proportional to the speed of RNAP r, if $r \ll \alpha$, meaning that the speed of RNAP polymerase is slower compared to its typical binding rate then we get the scaling relation

$$\log(R_T) = -\log(L) + \log(rN_T)\,. \tag{3.3}$$

Eq. (3.3) is useful as we don't need to know - and it is not possible to obtain it from RNA-Seq experiments - the speed or the total number of RNAP to understand how the transcription rate scales with respect to the gene length. In particular, the term $\log(rN_T)$, which is the unknown, can be found as the intercept of the linear fit between $\log(R_T)$ and $\log(L)$.

The scaling relationship (3.3) seems to be independent of DNA methylation, but nothing ensures that r or N_T are independent of DNA methylation. Moreover, there could be an hidden dependency and in the following we are going to highlight how they are connected. In particular, Eq. (3.3) was derived under the assumption of constant rates of RNAP transcriptional activity, but there might be factors that can change the

rates. For example, if there is another enzyme, such as DNMT3, bound to the gene body, then both the speed r of the RNAP can decrease as well as the typical number of bound RNAP N_b. In the following, we consider the effect of methyl binding enzymes on transcriptional activity since we have experimental information of DNAme, which is established by these enzymes. The following model can be straightforwardly extent to other enzymatic activities.

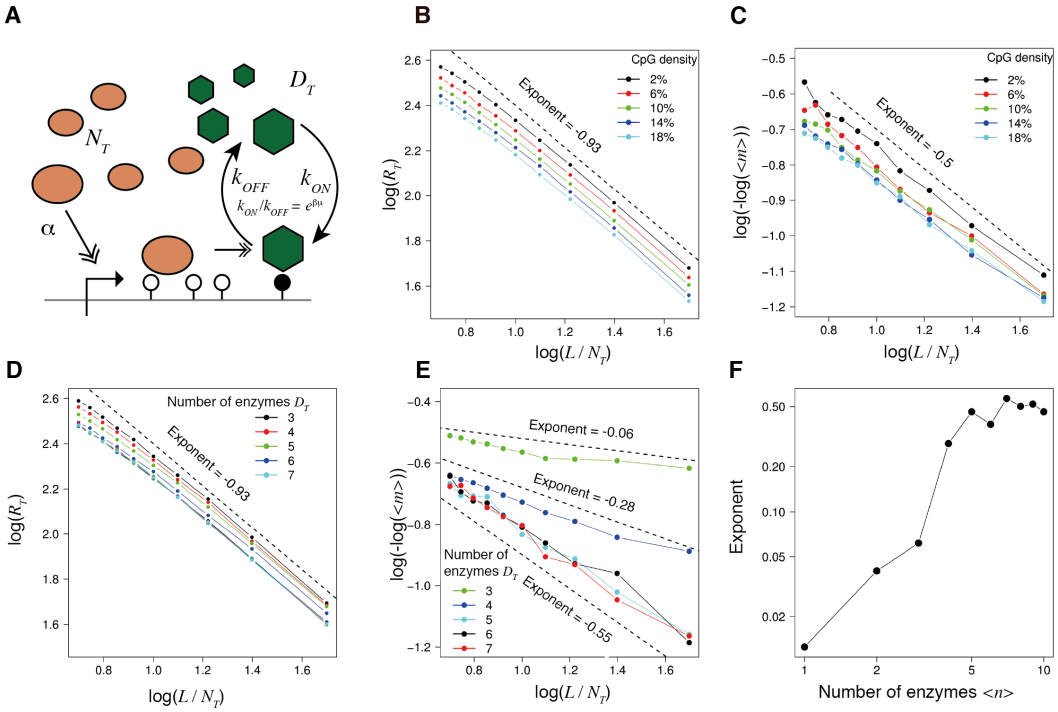


Figure 3.3.: (**A**) Sketch of the model and interplay between RNA polymerases (orange) and methyl binding enzymes (green). The rate of binding k_{ON} and unbinding k_{OFF} are given by the equilbrium condition $k_{ON}/k_{OFF} = \exp(\beta\mu)$. In (**B**) we show the numerical results for the scaling between elongation rate and RNAP density and in (**C**) between DNAme and RNAP density obtained from numerical simulations for different CpG densities. The dashed lines are the theoretical exponents. (**D**) The scaling between elongation rate and RNAP density does not change with increasing numbers of enzymes (D_T). On the other hand, the scaling exponent between methylation and RNAP density, Eq. (3.8), increases with respect to the total number of enzymes (**E**), saturating at ≈ 0.55 (**F**). The CpG density is 10% (Number of CpG/L) in all the simulations unless specified otherwise and the gene length is $L = 50$.

In order to reveal how methyl binding enzymes bound on the DNA change transcriptional activity, we consider D_T total enzymes, of which n are bound enzymes on the gene body (Fig. 3.3 A). As an enzyme does not have necessarily to bind at the promoter, but can bind anywhere along the gene, then the number of ways n enzymes can bind to L_f free sites is

$$L_f(L_f - 1)\dots(L_f - n)\,. \tag{3.4}$$

Moreover there are $\binom{D_T}{n}$ ways to choose n enzymes out of a total of D_T. If we associate to every bound enzyme an energy μ, the partition function is

$$Z = \sum_{n=0}^{D_T} \binom{D_T}{n} L_f^n e^{-\beta\mu n} = (1 + L_f e^{-\beta\mu n})^{D_T} \,. \tag{3.5}$$

In the previous equation we made two assumptions. First, we assume that the speed of RNAP is slow (as before) such that the number of binding sites is $L_f \approx L$ and if $D_T \ll L_f$, then $L_f(L_f - 1)\dots(L_f - n) \approx L_f^n$. Secondly, we consider that the system is in equilbrium with a thermal bath at a temperature T and $\beta = \frac{1}{k_b T}$ and with an energy proportional to the number of bound enzymes, namely μn. μ is the energy required to bind an enzyme at a give site, which is given by the catalytic domain via ATP (Sec. 1.2.1). From the partition function we can then obtain the average enzyme occupancy as

$$\langle n \rangle = \frac{\langle E \rangle}{\mu} = D_T \frac{L_f}{e^{\beta\mu + L_f}} \,. \tag{3.6}$$

where the average energy $\langle E \rangle = -\frac{\partial log Z}{\partial \beta}$. As DNA methylation is proportional to the average occupancy of methyl binding enzymes, the average methylation is given by

$$\langle m \rangle = \frac{\langle D \rangle}{L_f} = D_T \frac{1}{e^{\beta\mu} + \gamma L} \,, \tag{3.7}$$

where γ is the combination of the number of CpG sites and the sites not occupied by RNAP. In particular, the number of free CpG is given by $L_f = \gamma' L - \frac{L}{d}\gamma' L = \gamma' L(1 - \frac{1}{d})$. γ' is the average CpG density and it is typically, $\gamma' \approx 0.01$ and we set $\gamma = \gamma'(1 - \frac{1}{d})$. We need to find how does the free energy cost for a biding events μ depends on the parameters of the model. In particular, it can only depend on the length of the gene as the other parameters are related to RNAP such that $\mu = KL^\eta$, where η is unknown for now and K is a constant. We finally arrive to a compact scaling relation between average DNA methylation and gene length

$$\log(-\log\langle m \rangle) \sim -\eta \log(L) \,, \tag{3.8}$$

where the terms after the dots incorporate unknowns parameters as before. As one of the simplest choice for η we take typical scaling factor of a polymer carrying out a self avoiding random walk in three spatial dimensions ($\eta \approx 0.6$) [138]. In Fig. 3.1 B we plotted this scaling relation and indeed find that it matches very well the experimental data, where the inferred slope is $\eta_{\text{exp}} = 0.6$.

In order to reproduce the analytical result numerically, we performed a set of simulations by keeping the length fixed but varying the polymerase number which is an

equivalent way of changing the polymerase density. Specifically we fix $L = 100$ and set $K = 1$ for all the simulations. The simulation results show that the transcription elongation rate scales inversely with the polymerase density (Fig. 3.3 B) whereas average DNA methylation level scales as Eq. (3.8) with $\eta \approx 0.56$ with respect to the polymerase density (Fig. 3.3 C). In fact, we observed that the scaling exponents are independent of the CpG density of the gene body (Fig. 3.3 B,C). Here, methyl binding enzymes create obstacles to the movement of polymerases. Next, we performed the simulation for a different number of obstacles and observed that the scaling in transcriptional elongation is retained (Fig. 3.3 D) but the scaling exponent between DNA methylation level and length of the gene body increases with the obstacle number (Fig. 3.3 E,F) saturating at around 0.55 for high obstacles number, illustrating the rational for the observed correlation shown in Fig. 3.2 B.

As discussed in the previous section, the explanation of the observed scaling relations requires two fundamental assumptions in the model, namely, the constant polymerase number per gene as well as the scaling between three dimensional distance and the linear distance on the DNA based on the self avoiding random walk of the polymer. In order to investigate the scaling, we analysed a recently published single-cell Hi-C data set at 100 Kb resolution [139]. The experimental data provides us with a contact map of all the mouse embryonic stem cells chromosomes at single-cell level. From the data, the scaling between the three dimensional distance and linear distance is calculated over a distance of 2Mb (Fig. 3.4 A) for several regions of the 16 chromosomes for 8 different mESC cells (Fig. 3.4 B). In fact, the average scaling exponent is found to be very close to around 0.56 corroborating our assumption that the DNMT3 binding is hindered by the presence of polymerases close by and the binding rate is proportional to the distance in physical space. Next, we focused on the other assumption of a constant number of RNAPs per gene which can be explored using a genome wide pol2 ChIP-seq data. Here, we analysed the polymerase densities over the genes of different lengths and observed that indeed the polymerase density scales with an exponent -0.93 in accordance with the simulation result (Fig. 3.4 C).

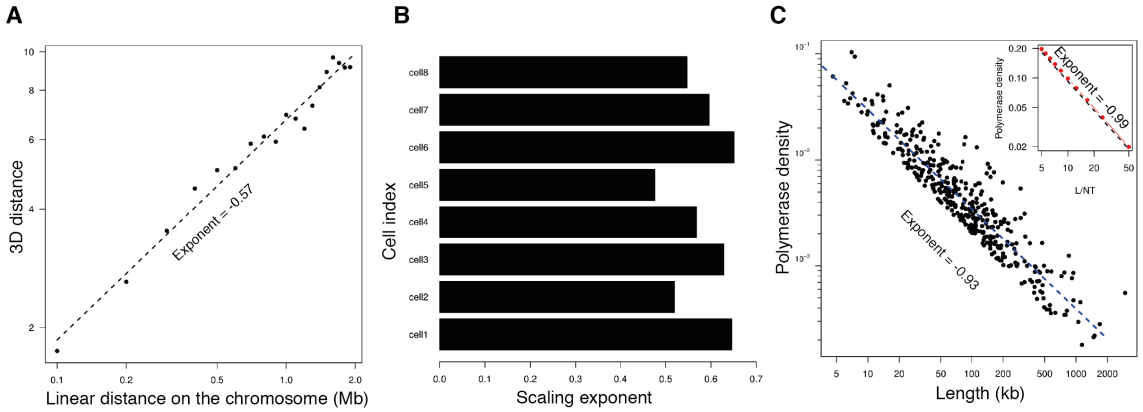


Figure 3.4.: (**A**) Scaling of the linear distance between base pairs along the 1D sequencing of the DNA and their 3D distance from HiC data. (**B**) The measured exponents of the scaling for different cells are in agreement between each other and suggest that the binding energy of methyl binding enzymes is indeed proportional to the three dimensional spatial distance as explained in the previous chapter. (**C**) The experimental scaling of RNAP density with respect to the length of the genes is in accordance with our theoretical prediction.

3.2. Effects of memory of DNA methylation marks during transcriptional output

Until now we have focused on the effects of the average level of DNAme in gene bodies on the transcriptional output. We now investigate whether for a given DNAme level the genomic arrangement of DNAme marks can influence transcription. As the previous scaling arguments suggest, equations (3.8) and (3.3), the minimal interaction between methylation marks and transcription is encoded in the transcriptional speed r. We take then r_i (transcriptional speed at site i) to be a function of the gene body methylation pattern $\boldsymbol{m}$, $r_i = r_i(\boldsymbol{m})$. We then expand this functions in terms of the first moments of the distribution of methylation resulting in

$$r_i(\boldsymbol{m}) \approx r_0 + f(\langle m_i \rangle) + g(\langle m_i m_j \rangle)\,. \tag{3.9}$$

Where f, g are for now two functions and $\langle m_i \rangle$ and $\langle m_i m_j \rangle$ are respectively average methylation and spatial correlation functions. If the particular structure of the pattern does not influence transcription then $g = 0$ and we expand $f(\langle m \rangle)$, around 0 such that, $r(\boldsymbol{m}) = \sum_i r_i/L \sim r_0(1 - \lambda\langle m \rangle)$, where L is the gene length. The negative sign indicates that methylation slows down transcription [117]. If $\lambda \ll 1$ then

$$r(\boldsymbol{m}) \sim r_0 e^{-\lambda\langle m \rangle} \tag{3.10}$$

In order to understand the effect of spatial correlations in DNA methylation patterns on transcriptional rate, we proceed with a microscopic description. In particular, at a first level of description, DNA methylation patterns are defined by their typical spatial correlation length. As a simple form of non-local interaction between methylation and transcription which encodes the correlation length of DNA methylation pattern we consider the kernel

$$K_i = \sum_{j=0}^{i} |m_j - m_{j+1}|/L\,. \tag{3.11}$$

The kernel K_i counts all the variations in the methylation pattern from the initial site of transition ($i = 0$) to the site i (Fig. 3.5 A) and as we will see later it encodes spatial correlation functions of DNA methylation. We expect that variations in DNAme patterns further slow down transcriptional activation such that it has the same sign as f, which will be justified at posteriori. The average transcriptional speed is

$$r(\boldsymbol{m}) = \langle \sum_i (1 - \lambda(m_i + \nu K_i)) \rangle / L\,, \tag{3.12}$$

where ν sets the relative contribution of the kernel K_i with respect to the term that scales with the average DNA methylation. We proceed with a mean field approximation, $P(\mathbf{m}) = P(m_1) \times P(m_2) \ldots \times P(m_2)P(m_L)$ such that r is approximated as

$$r(\boldsymbol{m}) \approx \sum_i (1 - \lambda \left(\langle m_i \rangle + \nu \langle K_i \rangle \right)) \rangle / L. \tag{3.13}$$

We consider the system to be translational invariant ($\langle m_i \rangle = \langle m \rangle$) and the kernel is evaluated as follows: if methylation patterns have a typical length scale of ξ then the number of signs changes up to sites i scales as i/ξ such that

$$r(\boldsymbol{m}) \sim r_0 e^{-\lambda(\langle m \rangle + \frac{\nu}{\xi})}\,. \tag{3.14}$$

We then find that the correlation length and translational speed are positively related. Putting equations (3.14),(3.8),(3.3) together we obtain a relationship between the transcriptional elongation rate and the correlation length of DNA methylation patterns,

$$\log R_T = \frac{1}{\eta} \log(-\log\langle m \rangle) - \lambda(\langle m \rangle + \frac{\nu}{\xi}) + \ldots\,. \tag{3.15}$$

For low values of average DNA methylation $\langle m \rangle$, the first term on the r.h.s is dominant, whilst it becomes irrelevant at high methylation levels. The third term on the r.h.s, as it is not dependent on average DNA methylation, cannot be detected from our previous analysis. In order to experimentally verify these predictions we analysed correlation length in each individual gene and later divided the genes between the 1000 most expressed (Top) and less expressed (Bottom). As the typical number of CpGs in each

gene body is quite limited ($\sim 500 - 600$ CpG per gene body on average), we use a spectral density approach to quantify the correlation length of a given gene i as $\tilde{\xi}_i = I_0/\gamma' L_i$, with I_0 the spectral density for zero frequency. This method requires a spatial series where observation are equally spaced, such that we later rescaled correlation length by the average distance between CpGs as, $\tilde{\xi}_i = \xi_i \, L_i/\gamma_i' L_i = \xi_i/\gamma_i$. The results are shown in Fig. 3.5 B and they do strengthen the hypothesis that transcriptional output is correlated to the correlation length of the DNA methylation patterns. In particular, we find that an increase in the correlation length between DNA methylation marks on the gene bodies is associated with an increase of transcriptional output.

Here, we showed how the interplay between DNA methylation and gene body length affects transcriptional output. We found that the logarithm of elongation rate is linearly proportional to logarithm of the gene body length, such that genes with longer gene bodies are on average less transcribed. Upon introducing a biophysical model of RNA polymerases activity affected by methyl binding enzymes kinetics, we showed that the average enzymes occupancy is a stretched exponential function of the gene body length. Specifically, the average DNA methylation follows the same scaling relationship with a stretching exponent that depends only on the spatial arrangement of base pairs in the physical space. Later, we extend the biophysical model to account for changes of DNA methylation marks arrangement with the same average DNA methylation. We found that memory in transcriptional output determined by the correlation length scale of DNA methylation marks affects the transcriptional output such that higher correlation lengths are associated with highly expressed genes. All the results are supported and in accordance with experimental evidences.

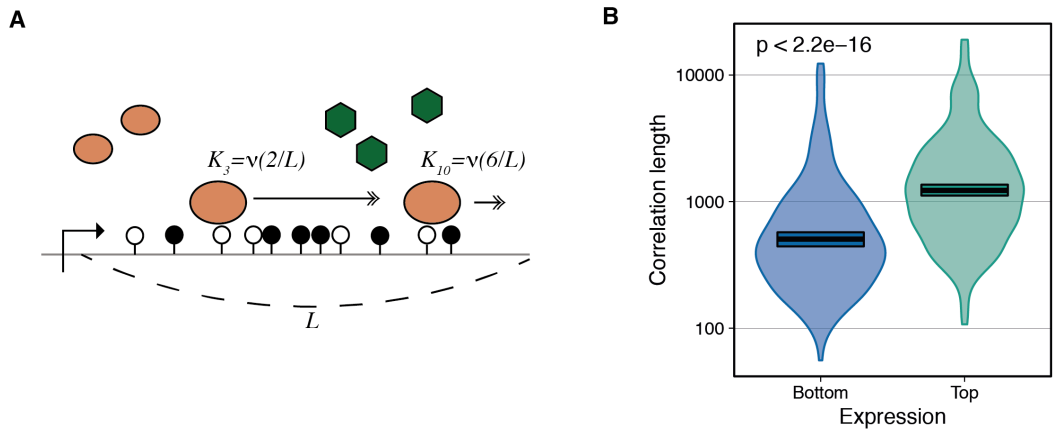


Figure 3.5.: (**A**) Memory kernel K_i for RNAPs at two different binding sites ($i = 3, 10$). (**B**) Estimated correlation length between methylated CpGs for highly and lowly expressed genes. In the violin plots the median is encoded in the horizontal thick black and first and fourth quantiles are encoded in the boxes heights. The p-value, obtained with a t-test between the correlation lengths of the two set of genes, shows significance.

3.3. Summary and discussion

In Sec. 2.7 we concluded that DNAme plays a crucial role in determining cell state transitions, in particular at the level of gene body. In this chapter we provided a theoretical framework, always based on experimental results, to tame the complex interplay between DNAme and gene expression. In particular, in Sec. 3.1 we exploit these intimate relationship by studying scaling laws in gene expression. We initially analyse sequencing experiments of mESCs in serum conditions and derived relationships between three fundamental quantities in gene expression" transcription rate, gene body length and average DNA methylation. Upon developing a minimal and simple thermodynamical model for transcription factor kinetics and DNMT3 binding we were able to explain all the experimental scaling relationships. Specifically, transcription factor kinetics is obstructed by the presence of DNMT3, which gives rise to a competitive interaction, such that non trivial scaling relationship are observed. Surprisingly, theoretical and experimental scaling relationship matches if we consider the 3D structure of the chromatin and its relationship with DNAme as derived in the previous chapter. We then argued whether the only relevant part for transcription was the global average methylation or the structure of methylation patterns. In Sec. 3.2 we include in our model a term which accounts for transcriptional memory along the gene body, concluding that the correlation length of methylation pattern plays a role in transcriptional output. Upon computing correlation lengths for different genes, we found that our prediction that translational speed is inversely proportional to the correlation length, making sense of the results derived in the previous chapter. Even though the theoretical modeling was sufficient to explain the correlations of transcriptional output with DNAme, it was lacking a dynamical perspective, in particular how to regulated transcription via changes in DNAme.

4. Glassy Fluctuations in Gene Regulatory Networks

Throughout this thesis, drawing on the analysis of genomic single-cell sequencing data we developed theories of non equilbrium systems that are in agreement with experimental findings and that are able to predict mechanisms underlying cellular behaviour. In particular, we were able to infer emergent mesoscopic structures of the DNA and their interplay with chemical DNA modifications by analysing one dimensional sequencing experiments (BS-Seq). Apart from DNA methylation, another layer of regulation of cell fate are gene regulatory networks. Single-cell RNA sequencing experiments allow to get quantitative profile of gene expressions, but they are extremely challenging for a direct inference of statistical quantities (Sec. 1.2.3). Systematic technical errors, lack of stochiometric observables and biological variability do not allow accurate theoretical prediction from sequencing experiments. We thus reason that the theoretical approach to infer emergent processes hided in RNA sequencing measurements has to follow the reverse path we took so far.

Here, we start from a theoretical understanding of gene regulatory networks and the coarse-grained theory will suggest the informative statistical observable to infer gene expression fluctuations dynamics in gene regulatory networks. Specifically, in Sec. 4.1 we describe general gene expression dynamics in terms of a master equation, which incorporates interactions between two of the experimentally accessible layers of regulation, mRNA and protein. We will show that the propagation of fluctuations in gene regulatory networks can be universally mapped into bipartite spin glass theories. We thus can use methods developed in the context of glass systems to infer gene expression fluctuations. In particular, we will find the analytical phase diagram which divides fluctuations of gene expression in two main categories: paramagnetic and glassy. The statistical observable which encodes this information is the overlap distribution, which measures heterogeneity between cells which belongs to the same statistical ensemble. Having identified the relevant observable, which is parameter free, in Sec. 4.2 we will draw on three different publicly available scRNA-Seq experiments to quantify heterogeneity in cell states and by comparison to the previously derived phase diagram we will be able to identity glassiness of cell states. We will highlight how neural induced progenitor cells of mouse show potential glassy behaviour such that it is possible to

encode information in gene expression fluctuations. In Sec. 4.3, we find another phase in which fluctuations of gene expression may reside by studying their out of equilibrium dynamics. Specifically, in paramagnetic-like cell states, fluctuations may still be long-lived, such that they are autocorrelated in time for time scales longer than the one described by individual molecular processes. We then ask what are the potential role of correlated fluctuations in cell state transitions and in Sec. 4.3.1 we analyse a self-regulating gene as a paradigmatic example of a genetic switch. We show that one potential role of correlated fluctuations is to regulate the switching between different states of the gene favouring and fixating one of the two states. Finally, in Sec. 4.4, drawing again on single-cell sequencing experiments we show how transition between different cell states is captured by a sharp increase of gene-gene correlations, which is qualitative predicted by minimal theories of cellular symmetry breaking. All together, we derived a theory of gene expression fluctuations in gene regulatory networks that identify statistical observables, namely overlaps, in single-cell RNA sequencing experiments, which are stable against technical and biological variability. Moreover, overlaps are both a measure of the heterogeneity of different cell states and of correlated fluctuations in gene networks.

4.1. Phase diagram of gene expression fluctuations

We consider a set of genes which are, following the central dogma of molecular biology, transcribed to mRNA molecules and then translated to proteins with respective molecular abundances m_i and n_i and degradation rates γ_i and d_i (Fig. 4.1). mRNA molecules are translated to proteins at a rate g_i. The transcription of a gene depends on the protein abundance of other genes via a nonlinear function $f_{ji}(n_j)$, which, as a result of cooperative promoter binding, typically takes the form of a Hill function, $f_{j,i} = \left(n_j^{\alpha_{ji}}\delta_{j,a(i)} + \delta_{j,r(i)}\right) / \left(\nu_{ji}^{\alpha_{ji}} + n_j^{\alpha_{ji}}\right)$ with threshold value ν_{ji} and degree of cooperativity α_{ji}. $\delta_{j,a(i)} = 1$ if j is activated by the transcription of i and 0 otherwise. The same applies to $\delta_{j,r(i)}$, meaning that a gene j inhibits the transcription of i. The time evolution of the probability $P(\mathbf{n}, \mathbf{m}, t)$ of finding protein and mRNA abundances $\mathbf{n} = (n_1, \ldots, m_N)$ and $\mathbf{m} = (m_1, \ldots m_N)$, respectively, is given by a master equation of the form,

$$\begin{aligned}\frac{\partial P(\mathbf{n},\mathbf{m})}{\partial t} &= \sum_{i=1}^{N} d_i \left(E_i^1 - 1\right) n_i P(\mathbf{n},\mathbf{m}) \\ &+ \Omega \left[p_i + \sum_{j\in e(i)} f_{ji}^e\left(\frac{n_j}{\Omega}\right)\right]\left(F_i^{-1} - 1\right) P(\mathbf{n},\mathbf{m}) \\ &+ \left(F_i^1 - 1\right)\gamma_i m_i P(\mathbf{n},\mathbf{m}) \\ &+ \Omega\left(E_i^{-1} - 1\right) g_i m_i P(\mathbf{n},\mathbf{m}) \ .\end{aligned} \tag{4.1}$$

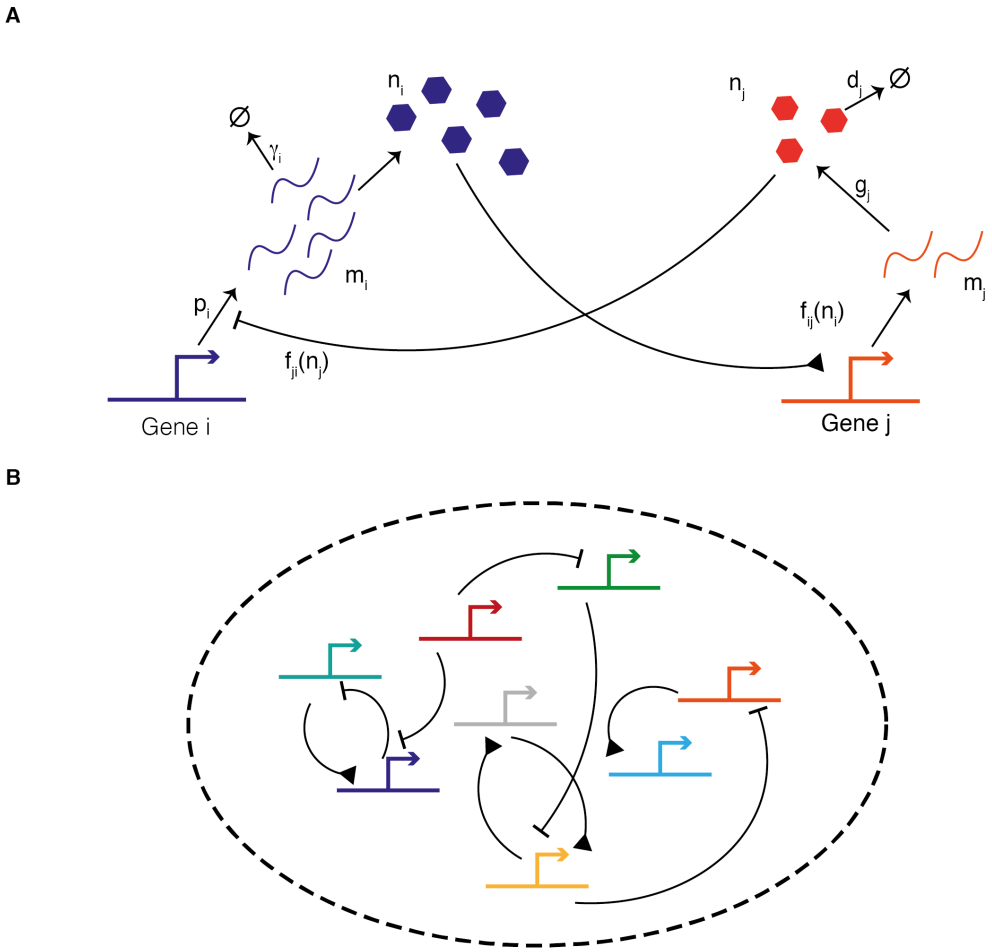


Figure 4.1.: (**A**) Schematic of the master equation (4.1). The mRNAs of a gene i (m_i) are produced with a rate than depends on the proteins of other genes ($f_{ji}(n_j)$) which act as repressors (⊣) or activators (◀) of the target gene and with a basal production rate p_i, independent of the other proteins. mRNAs are translated into protein n_i at a rate γ_i. mRNAs and proteins degrade with rates γ_i and d_i respectively. (**B**) Representation of interactions between different genes inside the cell nucleus (dashed black ellipse).

$j \in e(i)$ in the second terms means that the sum goes over all the genes j that are either activators or inhibitors of i. $E_i^{\pm}$ is a step operator that acts on any function to the right as $E_i^{\pm} g(\mathbf{n}, \mathbf{m}) = g(\{\mathbf{n}, n_i \pm 1\}, \mathbf{m})$ and similarly for $F_i^{\pm}$ which acts on the mRNA dependent part. As far as our knowledge concerns, this is the first study of gene regulatory networks dynamics with arbitrary interactions between genes and which takes into account mRNAs and proteins.

Master equations like Eq. (4.1) are not solvable exactly, such that we need to seek for approximations to estimate protein and mRNA levels. Different techniques can be applied to approximate master equations of the form (4.1), such as field theoretical descriptions [140, 141], WKB approximations [142], spectral methods [143] and system size expansion [124]. We perform a multivariate system size expansion, as it is suited to study the dynamics of fluctuations around steady state protein and mRNA concentrations. To this end, we express both mRNA and protein concentrations in terms of a deterministic and a stochastic component as $n_i(t) = \Omega\phi_i(t) + \Omega^{1/2}\xi_i(t)$ and $m_i(t) = \Omega\psi_i(t) + \Omega^{1/2}\eta_i(t)$ where Ω is the systems size (volume of the cell nucleus). $\phi_i(t)$ and $\psi_i(t)$ will follow a determinate dynamic as we are going to show in the following, whilst the dynamics of $\xi_i(t)$ and $\eta_i(t)$ are stochastic. The system size expansion proceeds as follows: upon substituting the decomposition of mRNA and protein into the master equation we equate terms with the same order in Ω on both sides of the master equations. To highest order in Ω we obtain the following differential equations,

$$\begin{aligned} \frac{\partial \phi_i}{\partial t} &= g_i\psi_i - d_i\phi_i \\ \frac{\partial \psi_i}{\partial t} &= p_i - \gamma_i\psi_i + \sum_{j\in a(i)} f_{ji}^{a}(\phi_j) + \sum_{j\in r(i)} f_{ji}^{r}(\phi_j) \ . \end{aligned} \tag{4.2}$$

Equations (4.2) describe the mean field behaviour of the protein and mRNA levels, i.e. their mean value. The first equation describes the production of protein i upon translation of mRNA i with rate g_i and degradation of protein with rate d_i. The second of those equations describe mRNA production with basal rate p_i, which is independent on the protein and mRNA level. The second term accounts for mRNA degradation, whilst the other two terms are the only non-linear term and they describe mRNA production due to the activation or inhibition from the protein of the genes which acts as transcription factors for the targeted gene. These last terms, as we are going to show, makes GRN very rich in dynamical and even static behaviour. Before proceeding with the study of the full interacting system, we study its behaviour if the genes were not

interacting. Equations (4.2) reduce to

$$\begin{aligned}\frac{\partial \phi_i}{\partial t} &= g_i\psi_i - d_i\phi_i\,,\\ \frac{\partial \psi_i}{\partial t} &= p_i - \gamma_i\psi_i\,.\end{aligned} \tag{4.3}$$

The exact mean field dynamical solution for mRNA concentration is (dropping the index i),

$$\psi(t) = \frac{e^{-\gamma t}\left[\gamma\psi_0 + p(e^{\gamma t}-1)\right]}{\gamma}\,, \tag{4.4}$$

where ψ_0 is the initial mRNA concentration. The solution for $\phi(t)$ is lengthy and we omit it for the sake of simplicity. The initial concentration for this very simple case does not affect long-term behaviour, we then set them to $\psi_0 = \phi_0 = 0$ such that,

$$\begin{aligned}\psi(t) &= \frac{e^{-\gamma t}\left[p(e^{\gamma t}-1)\right]}{\gamma}\,,\\ \phi(t) &= \frac{gp\left[\gamma\left(e^{-dt}-1\right) + d\left(1-e^{-\gamma t}\right)\right]}{\gamma d(d-\gamma)}\,.\end{aligned} \tag{4.5}$$

mRNA concentration first increases exponentially till eventually saturate to the stationary value $\psi^* = p/\gamma$, whilst protein levels saturate exponentially to $\phi^* = pg/\gamma_i d$. As there is only one unique steady state for each gene, it means that there exists only one unique steady state of the gene networks for constant values of the parameters, which is of course absurd. Interactions are then essential to span the whole complexity of cellular states. We proceed now the the study of the effect of those terms on GRN. Equations (4.2) do not have any exact solutions for general functions f_{ji} even for the steady state as f_{ji} can be arbitrary non linear functions. The system size expansion of the master equation gave to highest order that the deterministic part components of protein and mRNA concentration must satisfy the systems of equations (4.2), which in general admits multiple steady states. In order to study the dynamics of GRNs fluctuations we then proceed with an expansion of the master equation to the next highest order term (Ω^0) around one of the possible steady states, denoted by $\boldsymbol{\phi}^*$ and $\boldsymbol{\psi}^*$. From the expansion, we obtain a Fokker-Planck equation describing the time evolution of the joint probability of mRNA and protein fluctuations in the vicinity of a stationary state (Appendix E.1),

$$\frac{\partial P\left(\boldsymbol{\xi},\boldsymbol{\eta}\right)}{\partial t} = \sum_{i=1}^{N}\left[D_i^n\frac{\partial^2}{\partial\xi_i^2} + D_i^m\frac{\partial^2}{\partial\eta_i^2} + \frac{\partial}{\partial\xi_i}v_i^n + \frac{\partial}{\partial\eta_i}v_i^m\right]P\left(\boldsymbol{\xi},\boldsymbol{\eta}\right)\,, \tag{4.6}$$

where $D_i^n = d_i\phi_i^* + g_i\psi_i^*$, $D_i^m = \gamma_i\psi_i^* + p_i + \sum_{j\in e(i)} f_{ji}(\phi_j^*)$, $v_i^n = d_i\xi_i - g_i\eta_i$, $v_i^m =$

$-\sum_{j\in e(i)} f'(\phi_j^*)\xi_j+\gamma_i\eta_i$ are effective diffusion and drift coefficients, respectively ($f'(\phi) = \mathrm{d}f(\phi)/\mathrm{d}\phi$). Fluctuations are described by a Fokker Planck equation, whilst mean values were not, as fluctuations encodes all the stochastic contributions in the GRN around a stationary state. Stationary solutions of the Fokker-Planck equation are seek in the form $P = \exp(-H)/Z$. As typical mRNA abundances are much smaller than protein abundances, fluctuations in mRNA abundances are much stronger than fluctuations in protein abundances, $D_i^m \gg D_i^n$. In this limit, we obtain in Appendix E.1,

$$H = \sum_{i=1}^{N} \frac{d_i \xi_i^2}{2D_i^n} + \sum_{j=1}^{N} \frac{\gamma_i \eta_j^2}{2D_i^m} - \sum_{i=1}^{N} \frac{2g_i}{D_i^n}\xi_i\eta_i + \sum_{ij} \frac{f'_{j,i}(\phi_j^*)}{D_i^m}\xi_j\eta_i\,. \tag{4.7}$$

We define the couplings: $\frac{f'_{j,i}(\phi_j^*)}{D_i^m} = J_{ij}$. The couplings depend only on the specific interactions between genes, e.g. whether transcription factors bind as monomers, trimers, etc... and on the steady state values of the attractors of the dynamics. The specific interactions are fixed once a steady state is chosen and may change, as we studied in the previous sections, due to epigenetic modifications, which we will neglect for not. This implies that the main sources of variations for the couplings are the steady state values from which they are defined. As the specific form of the couplings is unknown, as a minimal form, we take them to be Gaussian distributed with vanishing mean and variance $\sigma/\sqrt{N}$. With this notation (4.7) is formally equivalent to the Hamiltonian of a bipartite spin-glass [144, 145], such that we can transfer methods from spin glass theory (Sec. 1.3.3) to understand the propagation of fluctuations in gene networks. We neglect here the sparse nature of the network, but rather consider a fully connected network. In general, either we consider couplings $J_{i,j} = \pm 1$ and only a finite fraction of edges [146] or as we did, takes a fully connected network with vanishing variance with respect to the system size. We choose the second approach as the couplings can be more easily related to experimental data. The couplings of the Hamiltonian can be indeed inferred from gene-gene correlations [147–149] as continuous valued.

The formal similarities between GRN and spin glasses are not new [140, 150], but as it will be clearer later, we develop a different framework which will allow to bridge the gap between RNA-Seq experiments and theoretical predictions which other theories are lacking. Specifically, we found that we can not map the full dynamics of gene expressions to a spin glass, but only the dynamics of gene expression fluctuations around a stationary state. These findings point out which statistics one has to study to better understand GRNs. In particular, cells would have to be as close as possible to a stationary state and we would need to study fluctuations around those states. Moreover, as we will see, dealing with fluctuations overcomes many difficulties in the analysis of RNA-Seq data. We then develop a theory which does not assume Boolean networks

[151, 152], specific interactions between genes or small gene networks and can deal with the full complexity of GRNs. We now proceed with the study of the probability distribution of fluctuations. We define the average correlation of fluctuations between realisations (replicas) a and b, $Q^{\xi}_{ab} = N^{-1}\sum_i \langle \xi^a_i \xi^b_i \rangle$ and $Q^{\eta}_{ab} = N^{-1}\sum_j \langle \eta^a_i \eta^b_i \rangle$, namely overlaps. With this definition we can formally define a free energy [79] (Section 1.3.3) by

$$F = -\lim_{n\to 0^+, N\to\infty} (\beta n N)^{-1} \int \mathrm{d}\xi\, \mathrm{d}\eta\, \mathrm{d}Q^{\eta}_{ab} \mathrm{d}Q^{\xi}_{ab} e^{-\beta n N H}\,. \tag{4.8}$$

In formal analogy to statistical physics, by taking derivatives of F we can compute macroscopic properties of GRN. The Hamiltonian is (Appendix E.2)

$$\begin{aligned} H = &\frac{\sigma^2}{2}\sum_{a\neq b}\left(Q^{\xi}_{ab}Q^{\eta}_{ab}\right) + \frac{\sigma^2}{2}\sum_a \left(Q^{\xi}_{aa}Q^{\eta}_{aa}\right) \\ &- \log \int \mathrm{d}\xi\, \mathrm{d}\eta \Psi_{\eta,\xi}\Psi_{\eta}\Psi_{\xi}\,, \end{aligned} \tag{4.9}$$

with

$$\begin{aligned} \Psi_{\xi} &= \exp\left[\sum_{a\neq b}\frac{\sigma^2}{2}Q^{\xi}_{ab}\xi^a\xi^b + \sum_a \frac{\sigma^2}{2}Q^{\xi}_{aa}\xi^a\xi^a + \sum_a V^{\xi}(\xi^a)\right] \\ \Psi_{\eta} &= \exp\left[\sum_{a\neq b}\frac{\sigma^2}{2}Q^{\eta}_{ab}\eta^a\eta^b + \sum_a \frac{\sigma^2}{2}Q^{\eta}_{aa}\eta^a\eta^a + \sum_a V^{\eta}(\eta^a)\right] \\ \Psi_{\eta,\xi} &= \exp\left[\sum_a \tilde{K}_{\xi,\eta}\xi^a\eta^a\right]\,. \end{aligned} \tag{4.10}$$

The rich behaviour of fluctuations of GRNs is then entirely described by the structure of the overlaps Q^{ξ}_{ab} and Q^{η}_{ab}. At this stage, we have no idea of what the values of each entries $Q^{\xi/\eta}_{ab}$ are. This problem was solved, for the case of a unipartite SK model, by Parisi [78] and a mathematically rigorous proof was given in [153, 154]. In order to obtain insight into the statistical ensembles of gene expression fluctuations we now deal with the structure of Q_{ab} for asymmetric bipartite spin glasses described by the Hamiltonian (1.23) and show that it is the key to understand GRNs. To this end, we substitute an ansatz of the form $Q^{\xi}_{ab} = (1-q^{\xi}_0)\delta_{ab} + q^{\xi}_0$ and analogously for Q^{η}_{ab} (replica symmetric ansatz). If there is a unique solution of the minima of F with $q_0 = 0$ the system is in a paramagnetic like phase where replicas are uncorrelated between each other. On the other hand, when the replica symmetric ansatz for Q_{ab} is a solution of F, but F is concave, then F admits local minima such that fluctuations may remain trapped in metastable states. Fluctuations may be of two kinds, constrained and not constrained. We reason that protein and mRNA fluctuations can never be unconstrained, meaning that individual fluctuations cannot have domain $[-\infty, \infty]$. This is clear as the basin of attraction of stable steady state are finite.

As fluctuations scales as $\Omega^{1/2}$, in the continuous limit we can impose a spherical constraint such that, $\sum_i \xi_i^2 = N, \xi_i \leftrightarrow \eta_i$, making this model a bipartite version of the p-spin spherical spin glass [155]. This constraint can be interpreted as the tendency of a cell to control fluctuations in order to fix a stationary state, or can be extrinsically imposed. Instead of constraining global fluctuations, we can as well constrain them locally such that, $\xi_i, \eta_i = \pm 1$. In Appendix E.3 we analyze both cases and find that for GRN there is an analytical expression for the boundary between a phase where the only solution is $q_0 = 0$ and a phase where q_0 is positive. The boundary of this region is given by,

$$\begin{aligned} \alpha^2 &= 1 + \tanh \tilde{K}_{\xi,\eta} \quad \text{(binary)} \\ 1 &= \frac{\alpha}{2\sqrt{2}} \frac{3 + \sqrt{1 + 8\tilde{K}_{\xi,\eta}^2 \alpha^2}}{\sqrt{1 + 2\tilde{K}_{\xi,\eta}^2 \alpha^2 + \sqrt{1 + 8\tilde{K}_{\xi,\eta}^2 \alpha^2}}} \quad \text{(spherical)} \end{aligned} \tag{4.11}$$

where and $\tilde{K}_{\xi,\eta} = \frac{2g_i}{D_i^n}$ and $\alpha = 1/\sigma$. We neglect gene to gene variability in the chemical rates, which as long as the rates are Gaussian distributed is merely a shift of σ [156]. As previously outlined, both phase boundaries divide the phase space into two regions. Above the grey or red line in Fig. 4.2 there is a unique solution given by $Q_{ab} = 0$ and below the boundaries the replica symmetric solution is not a unique minima of the free energy. The existence of a replica symmetric solution doesn't imply that the free energy F is always stable. Specifically, the replica symmetric solution might correspond to a maximum of the free energy. We then need to study the stability of the free energy with respect to the replica symmetric (RS) solution [157] and we find analytically the boundary where the replica solution is unstable for the spherical spin glass. In particular, the condition for stability is given by [157],

$$1 - \sigma^2 \overline{(1 - 2\langle \xi^a \xi^b \rangle + \langle \xi^a \xi^b \xi^c \xi^d \rangle)} > 0 \,, \tag{4.12}$$

with

$$\overline{\langle \xi^a \xi^b \xi^c \xi^d \rangle_{RS}^n} = \int Dz Dw \left(\frac{\int d\xi^a d\eta^a \xi^a e^{-H_{RS}(\xi^a, \eta^a)}}{\int d\xi^a d\eta^a e^{-H_{RS}(\xi^a, \eta^a)}} \right)^n \,, \tag{4.13}$$

and $\langle \ldots \rangle_{RS}$ indicating the average over the overlap matrix where we substituted the entries in accordance with the replica symmetric ansatz. These equations can be solved numerically and the phase boundary is shown as a black solid line in Fig. 4.2. Therefore, gene expression fluctuations in GRNs exhibit a phase transition from a phase where fluctuations are uncorrelated to a glassy phase characterised by strong correlations of fluctuations. Interestingly, mRNA translation leads to the existence of this phase boundary that is absent in the p-spin spherical glass with $p = 2$, whilst it is present in case of binary fluctuations (SK).

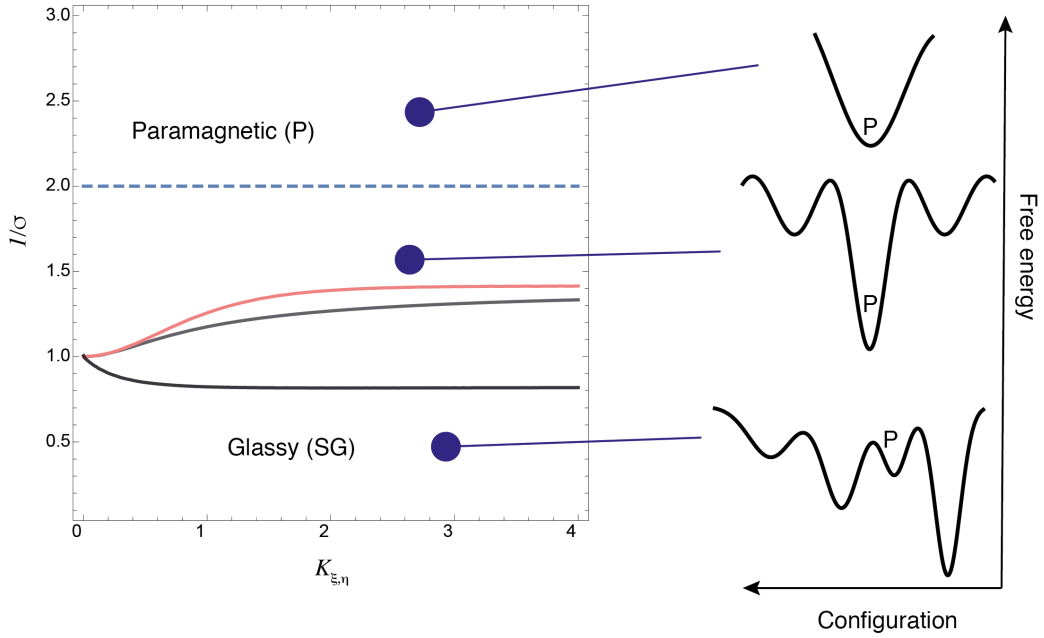


Figure 4.2.: Phase boundary between a paramagnetic like phase (P) and a glassy like phase (SG). Solutions $q_0 = 0$ for the spherical spin glass (grey line) and bipartite SK (red line). The black line denotes the De Almeida Thouless (DAT) line of the instability of the replica symmetric solution for the spherical spin glass. In the region between the black line and the grey line the replica symmetric solution is stable and unstable otherwise. The dashed blue line signals the dynamical transition below which time autocorrelation functions of protein and mRNA fluctuations have a non-vanishing plateau, Eq. (4.23). On the right hand side we show a low dimensional representation of the free energy in terms of the possible configurations. The paramagnetic macrostate is the only minima of the free energy above the blue line, whilst below the blue line other metastable states exists. Below the DAT line the paramagnetic macrostate is not a minima of the free energy.

A hallmark of glassy behaviour is non ergodicity, so that even over very long time scales such systems do not explore the entirety of phase space. Therefore, in the glassy phase, gene expression fluctuations are localised in the gene expression space. As we will discuss below, the localisation of fluctuation has potentially significant implications for the stability of gene expression states. All together, we found that mRNA and protein fluctuations around stationary cell states may exhibit glassy behaviour depending on the parameter of the gene networks, which are as well determined by the particular stationary state. In the next section we aim at identifying different steady states via RNA-Seq experiments and study where cell lie in the phase diagram, Fig. 4.2. In particular, we want to identify which cells show glassy behaviour, if any, and which do not. The identification of these states, it will be shown to be essential to understand cell fate transitions and cell regulation.

4.2. Evidence of glassy fluctuations from RNA sequencing experiments

In the previous sections we studied possible distributions of protein and mRNA fluctuations patterns for cells in a stationary state. We found out that there might emerge two phases, which in the statistical physics language we characterized as paramagnetic and glassy. As outlined at the beginning we proceed with a bottom up and approach and we analyse single-cell gene expression sequencing data to understand in which context these two different phases emerge. In particular, the overlap distributions will tell us in which phase a given steady state is. In the paramagnetic phase, the overlap distribution is unimodal centered around $q = 0$ and in the glassy phase it is broader. As RNA sequencing data only allows us to obtain information on mRNA and not protein abundance, we will from now on focus solely on the mRNA levels. We have already shown in the previous sections, that the transition between paramagnetic and glassy happens for both mRNA and protein at the same boundary, such that inferring the glassy phase for mRNA abundance is enough to infer in which phase the whole gene regulatory network is. We recall the definition of overlaps between two replicas for mRNA fluctuations ($Q^{\eta}_{a,b}$),

$$Q^{\eta}_{a,b} = \frac{1}{N}\langle \sum_i \eta_i^a \eta_i^b \rangle \,, \tag{4.14}$$

where the sum runs over all genes and the angle brackets are the thermal average over the Boltzmann distribution. a, b are the replica indices. At this point, it is still not that clear what a replica in single-cell sequencing experiments is. Before proceeding we need to find a proper biological definition for replicas, bearing in mind that we are computing statistical quantities for fluctuations.

We omit the average over the disorder, meaning that every thermodynamical quantities must be computed averaging over all the possible realisation of the interaction couplings J_{ij} in Eq. (4.7). As thermodynamic quantities are self-averaging, we can compute statistical observable in the steady state for a specific realisation of the couplings, which for now is unknown, and later generalise to different couplings realisations, as long as the mean and variance are the same. We have then found constraints of where observables should be computed, which means that we need to find attractors of the gene networks dynamics, which are identified as cell types. We finally have to compute overlaps for pairs of replicas, which are now identifiable as different individual cells within the same cell types, and finally get the overlaps distributions.

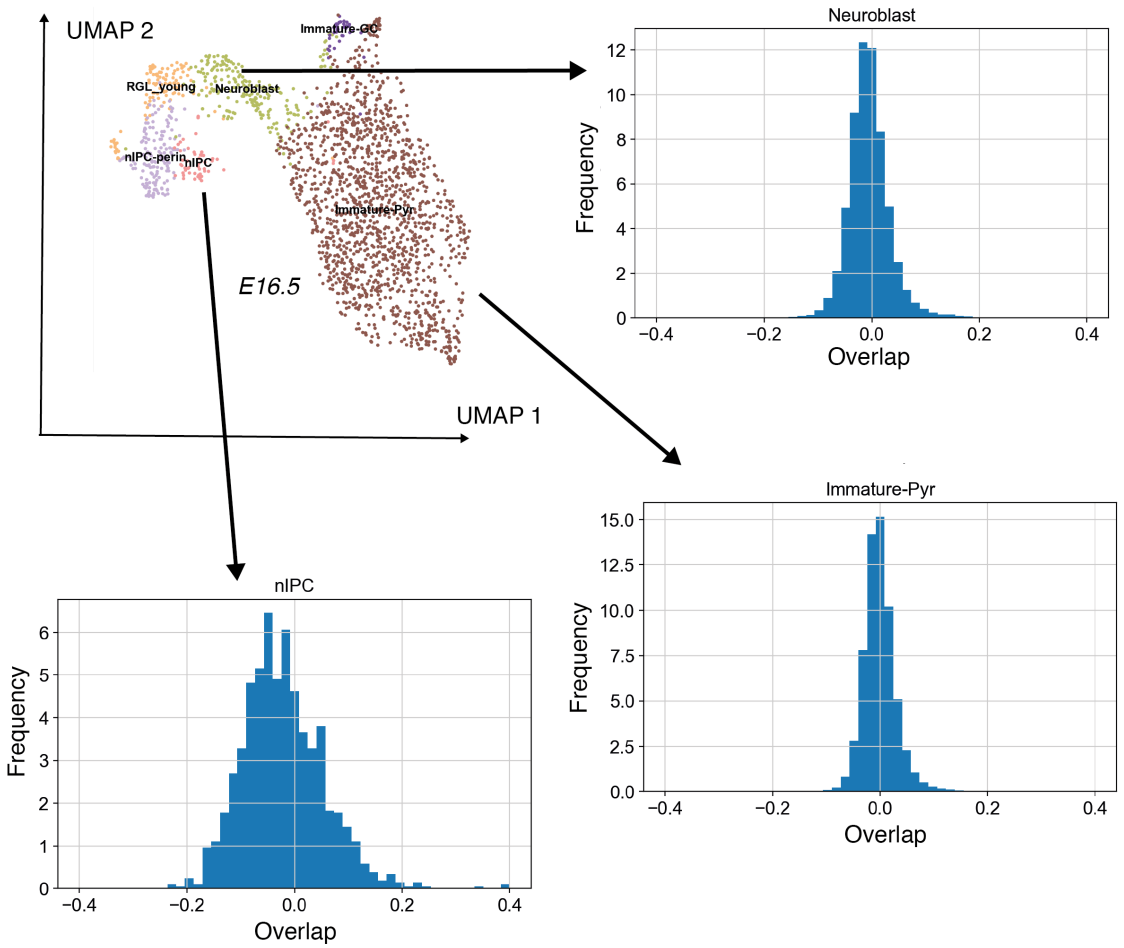


Figure 4.3.: UMAP dimensional reduction for mouse brain cells sequenced at E16.5. For each cluster of cells (colours) identified with the Leiden community detection algorithm we compute the overlap distribution for every pair of cells. nIPCs show a broader distribution of overlaps than neuroblast cells.

Specifically we analyse three different single-cell RNA-Seq data set from the mouse brain [67]. We initially process all data set, as explained in Sec. 1.2.3. Briefly, we filter cells, log-normalize the data and then select 1000 highly variable genes. For all the data sets only cells with reads $\in [10^3, 2\,10^4]$, less than 20% of mitochondrial RNA and with at least 800 genes expressed are kept (this choices are the one of [67]). Later we proceed with the UMAP dimensional reduction Fig. 4.3 for data visualization. We perform a k-means clustering [62] and Leiden community detection [65] to find out which cells belongs to the same cluster and we label them. Each cluster is then read as a cell type. We interpret clusters as stationary states and we can proceed with the study of the overlaps and then compare with analytical results.

In order to obtain mRNA fluctuations, we first center mRNA values by subtracting the mean of each gene for cells in a given cluster and divide by the variance and then

compute overlaps between centered mRNA. Overlaps encode the degree of similarity between cells in the same cluster, as well as the position of a cluster in the phase diagram (Fig. 4.2). In Fig. 4.3 we plot the overlap distribution for three different clusters: neural intermediate progenitor cells (nIPCs), immature pyramidal cells (Pyr) and neuroblasts of mouse brain cells sequenced at E16.5. The overlaps and UMAPs for the other two data set are presented in Fig. 4.4. The overlaps distribution quantifies in a direct and physical way the similarities between cells in the same cluster. In all the data set analysed, the distribution of overlaps for nIPCs appears to be broader than for immature and differentiated cells. Before discussing the biological and physical implication of this finding, we need need to quantify the broadness of the distribution and rule out possible finite size effects. In order to quantify the broadness of the distribution we compute the overlap width as the standard deviation of the overlap distributions for all the cells in a given cluster. In Fig. 4.5, we show the scaling of the overlap width with respect to the number of genes the overlap is computed over (from 10 to 10^3 highly variable genes). In particular, if the fluctuations of RNA in two different cells are uncorrelated, we expect the scaling to follow the central limit theorem (blue dashed line), meaning that the standard deviation should scale as $1/\sqrt{\text{number of genes}}$. The scaling of nIPCs is the only one that show consistent deviation from the central limit theorem in all the data set analysed, giving a further confirmation that the broadness of their overlap distributions is not a finite size effect. On the other hand, clusters of nIPCs are comprised of very few cells (~ 50), compared for example with neuroblast ($\sim 200 - 450$). In order to rule out the possibility that this deviation from the central limit theorem is given by a finite size effect due to limit number of cells, we compute the scaling for subsamples of the biggest clusters (in number of cells). Specifically, we select 50 cells from the bigger clusters in all three data set and compute the scaling of the overlap width with respect to the number of genes for different subsamples. The subsamples are generated by choosing a cell in the bigger clusters randomly and selecting the k nearest neighbours in the graph constructed with k-means clustering, where k is equal to the number of cells in the smaller cluster (nIPCs). In Fig. 4.6 A we show that for the different subsamples there is no consistent deviation from the central limit as it was observed for nIPCs in Fig. 4.5. In Fig. 4.6 B we show the resulting histograms of overlap widths for different subsamples for 10^3 genes and compare them to nIPCs (dashed line). We see again that nIPCs have always almost a higher overlap widths compared to immature Pyr and granule mature cells at different sampling times. This finding rules out the possibility that the scaling observed in nIPCs is a finite size effect. Taken together, we strengthened the finding that nIPCs have a broader distribution of overlaps compared to the other cell types in these data set of mouse brain.

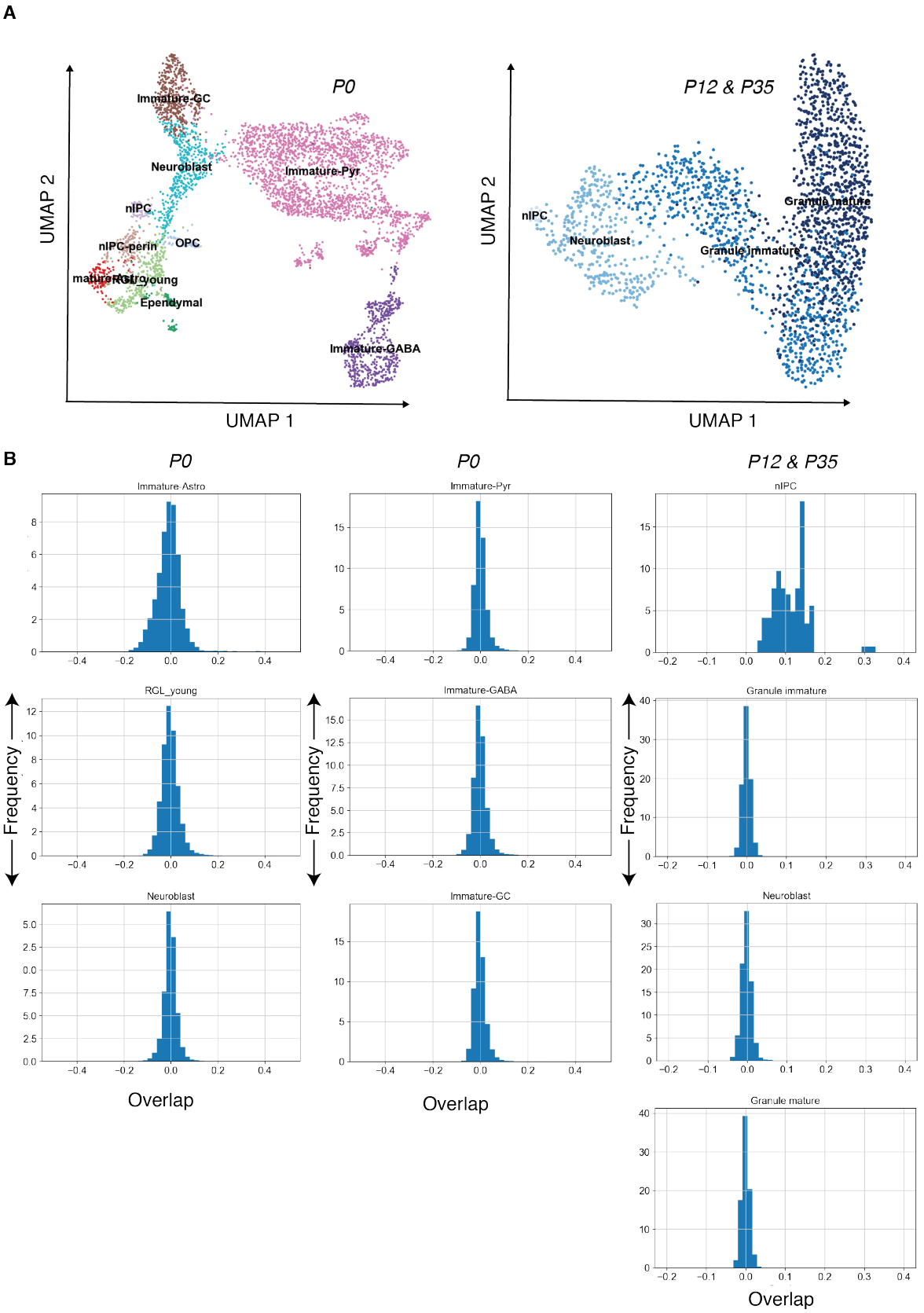


Figure 4.4.: (**A**) UMAP dimensional reduction for mouse brain cells sequenced at P0 (left) and P12 and P35 (right). (**B**) Overlaps distribution of genes in cells belonging to the clusters indicated in (**A**).

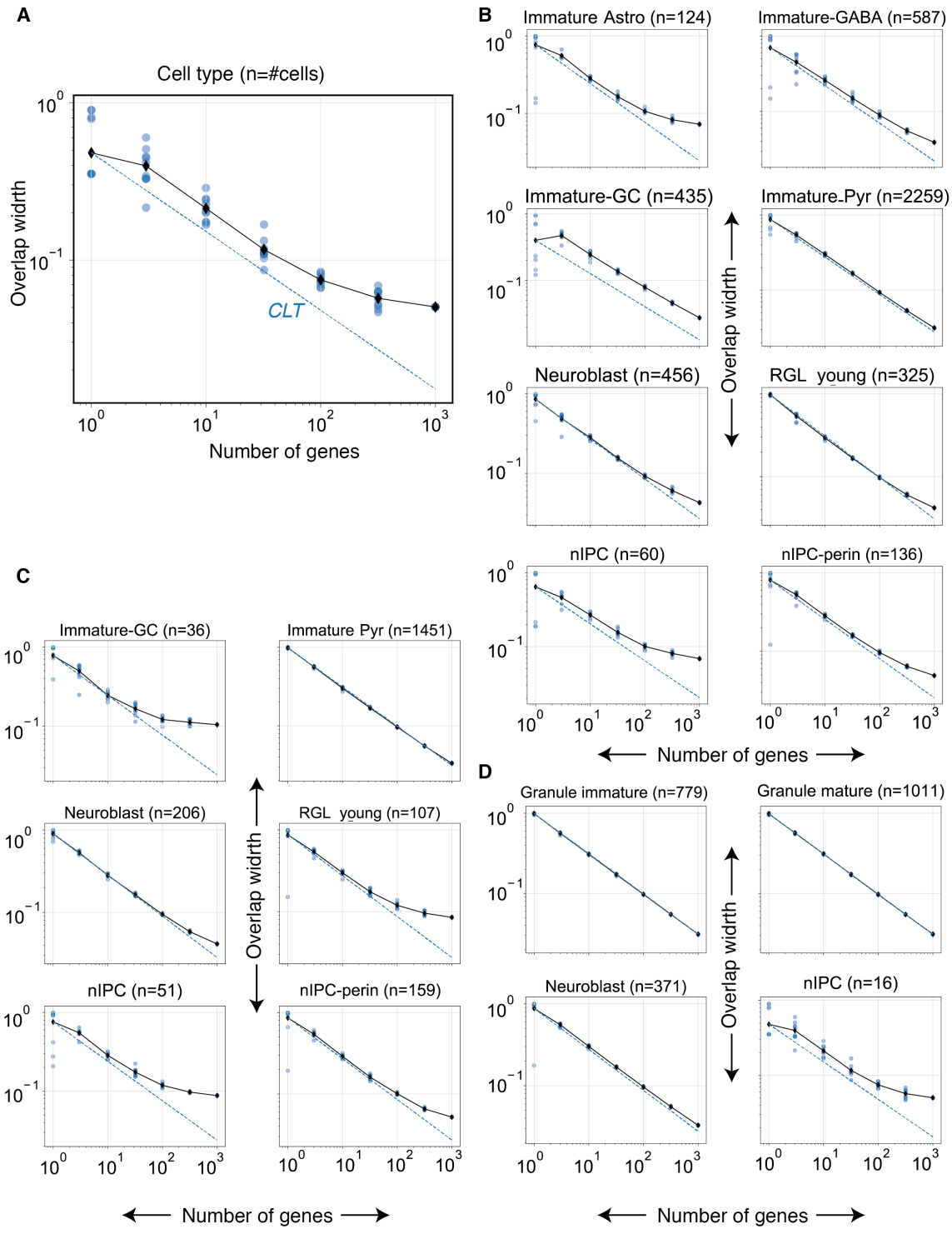


Figure 4.5.: (**A**) Example of the scaling between overlap width (standard deviations of the distribution) and number of genes over witch the overlap is computed. The black solid line is the average over different choices of the genes (dots). The dashed blue line is the central limit theorem (CLT) prediction, $\sim (\text{Number of genes})^{-1/2}$. In all the plots we show the overlap widths scaling for cells in the same cluster for the all the data set: E16.5 (**B**), P0 (**C**), P12 and P35 (**D**).

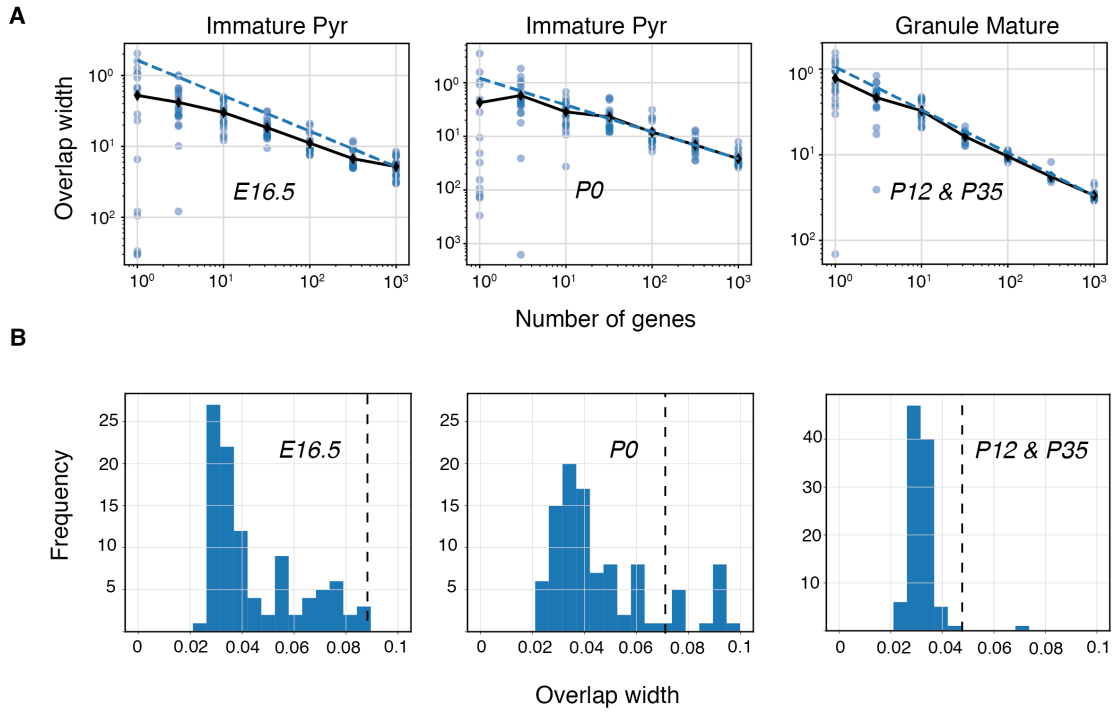


Figure 4.6.: (**A**) Scaling of the overlap width with respect to the number of genes for different subsamples (dots) of the larger clusters (indicate on top). The cells in the subsamples are of the same number of IPCs (n_{IPC}) in all the respective data set and they are chosen by randomly select a cell and selecting the $n_{IPC}-1$ nearest neighbour in the k-means graph. The dashed blue line is the CLT prediction, $\sim (\text{Number of genes})^{-1/2}$. (**B**) The frequencies distribution of the overlap widths for different subsamples, determined as in **A**, for all the data set are shown and compared with the overlap widths of the IPCs (dashed black line).

A broader distribution of overlaps indicates that there is an underlying spin glass landscape, in particular, that these cells resides in a glassy phase, where the dynamics between different realisations of mRNA fluctuations is slower compared to a paramagnetic landscape. On the other hand, for cells that have a narrow distribution of overlaps, which is paramagnetic like, the replicas are uncorrelated on average. Cells that are in a glassy phase are less heterogeneous between each other than the one in a paramagnetic phase, but as we will see in the next sections, the dynamics of gene expression fluctuations for "glassy cells" will have potential implications on plasticity and cell state transitions. We recall that the overlap distribution does not require any dimensional reduction or other clustering algorithms, but only data normalization. Overlap distributions are then statistical observables, that we can compute without the need to tune or choose any free parameters as for dimensional reductions or clustering algorithm, and they are then an essential tools to quantify actual similarities between cells with respect to the structure of the landscape and in particular where certain cell types resides in the phase diagram Fig. 4.2. It will be of great interest in the future to study which cell states in which context have glassy behaviour and why is it so. Here,

we found a statistical tool to identify cell states in terms of glassiness and find the parameters that guides such transitions. Having identified cells with respect to their overlap distributions in a particular cell state, changes the way we have to interpret transitions between cell states in terms of epigenetic (Waddington) landscapes. The loss of stemness cannot be seen even pictorially as the transition from a paramagnetic like phase to a ferromagnetic one Fig. 1.6, but rather as transitions from a rough landscape to another one. Stem cells occupy one of the multiple valleys and differentiated cells occupies valleys of a possibly smoother landscape, Fig. 4.7. We remind that our theory is for gene expression fluctuations of cells which are close to a steady state, so they belong to the same statistical ensemble. We can then study how the different landscapes change in terms of the parameters of the gene network by quantifying glassiness of cell states, but we cannot say much about the way these landscapes change during cell state transitions. Before digging more into the high complexity of cell state transition we have to clearly identify what is the role of glassy fluctuations in gene regulatory networks and how they change stability of gene networks.

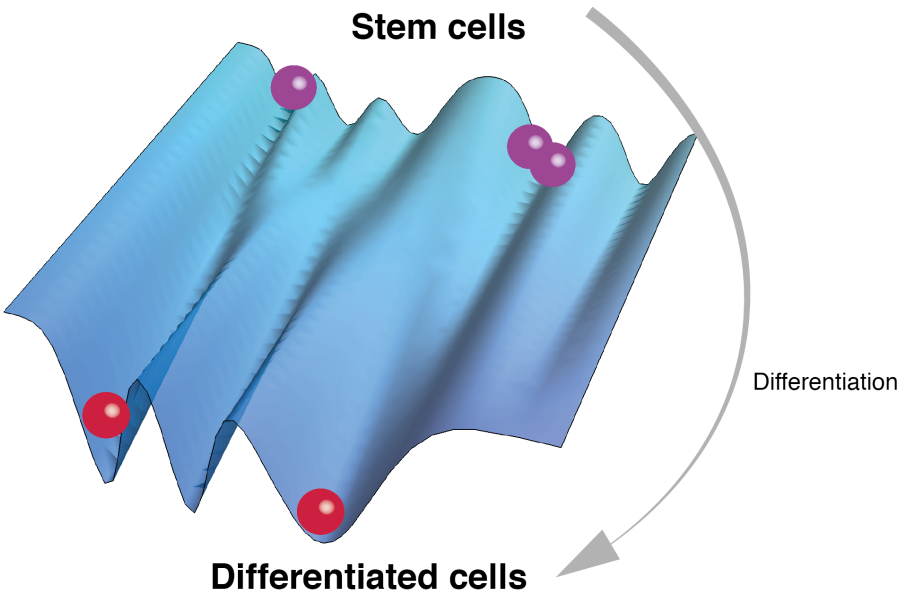


Figure 4.7.: Different interpretation of the Waddington landscape: stem cells occupy one of the multiple minima of the free energy landscape and a quench or loss of stemness leads to a change of the landscape.

4.3. Out of equilibrium dynamics of gene expression fluctuations

To understand the kinetic consequences of glassy fluctuations we study the dynamics of mRNA and protein fluctuations described by the Fokker Planck equation (4.6)

$$\begin{aligned}\partial_t \xi_i &= g_i \eta_i - d_i \xi + \sqrt{D_i^n} W_i^{\xi}\,,\\ \partial_t \eta_i &= -\gamma_i \eta_i + \sum_{j\in e(i)} f'_{j,i}(\phi_j^*)\xi_j + \sqrt{D_i^m} W_i^{\eta}\,.\end{aligned} \tag{4.15}$$

Upon redefining the couplings as $f'_{k,j}(\phi_k^*) = J_{kj}$, the time evolution of ξ_i and η_i is described by the Langevin equations (Appendix E.4)

$$\begin{aligned}\partial_t \xi_i &= g_i \eta_i - d_i \xi_i + \sqrt{D_i^n} W_i^{\xi}\,,\\ \partial_t \eta_i &= -\gamma_i \eta_i + \sigma^2 \lambda \int \mathrm{d}t' \chi_{\eta_i}^{\xi_i}(t,t')\xi_i(t') + \sigma W_c^{\eta_i} + \sqrt{D_i^m} W_i^{\eta}\,.\end{aligned} \tag{4.16}$$

W_i^{η} and W_i^{ξ} are Gaussian uncorrelated white noises with zero mean and unit variance and
$\langle W_c^{\eta_i}(t) W_c^{\eta_i}(t')\rangle = C_i^{\xi}(t,t')$, $\overline{J_{ij}} = 0$, $\overline{J_{ij}^2} = \sigma^2/N$, and $\overline{J_{ij}J_{ji}} = \lambda\sigma^2/N$. $C_i^{\xi}(t,t')$ and $\chi_{\eta_i}^{\xi_i}(t,t')$ are respectively the auto-correlation and response functions of the protein noise. In Eq. (4.16) there is no quadratic terms that accounts for gene gene interactions. The powerful MSRJD method, Sec. 1.3.1, allows us to pass from $2N$ coupled equations (4.15) to N groups of two coupled equations (4.16), at the price of adding a colored noise and a temporal and non-local kernel, which depend on the correlators and propagators of the system. Eq. (4.16) are the noise dynamics without any constraints, spherical or binary. If the couplings are asymmetric such that J_{ij} and J_{ji} are uncorrelated for all pairs of genes then $\lambda = 0$ and in Fourier space

$$\begin{aligned}\xi_j(\omega) &= \frac{g_j \eta_i(\omega) + \sqrt{D_i^n} W^{\xi}(\omega)}{i\omega + d_j}\,,\\ \eta_j(\omega) &= \frac{W_c^{\eta_i}(\omega)}{i\omega + \gamma_j}\,.\end{aligned} \tag{4.17}$$

Correlation in Fourier space, defined as $C_i^{\xi}(\omega,\omega') = \langle \xi_i(\omega)\xi_i(\omega')\rangle$ with, $\xi \leftrightarrow \eta$ are,

$$\begin{aligned}C_i^{\xi}(\omega,\omega') &= \frac{g_j^2 C_i^{\eta}(\omega,\omega') + \delta(\omega+\omega')D_i^n}{(i\omega + d_j)(i\omega' + d_j)}\\ C_i^{\eta}(\omega,\omega') &= \frac{\sigma^2 C_i^{\xi}(\omega,\omega') + D_i^m \delta(\omega+\omega')}{(i\omega + \gamma_j)(i\omega' + \gamma_j)}\,.\end{aligned} \tag{4.18}$$

which, if time translational invariance holds simplifies to,

$$C_i^{\xi}(\omega) = \frac{D_i^m g_i^2 + D_i^n \gamma_i^2 + D_i^n \omega^2}{d_i^2\gamma_i^2 - g_i^2\sigma^2 + d^2\omega^2 + \gamma_i^2\omega^2 + \omega^4}\,. \tag{4.19}$$

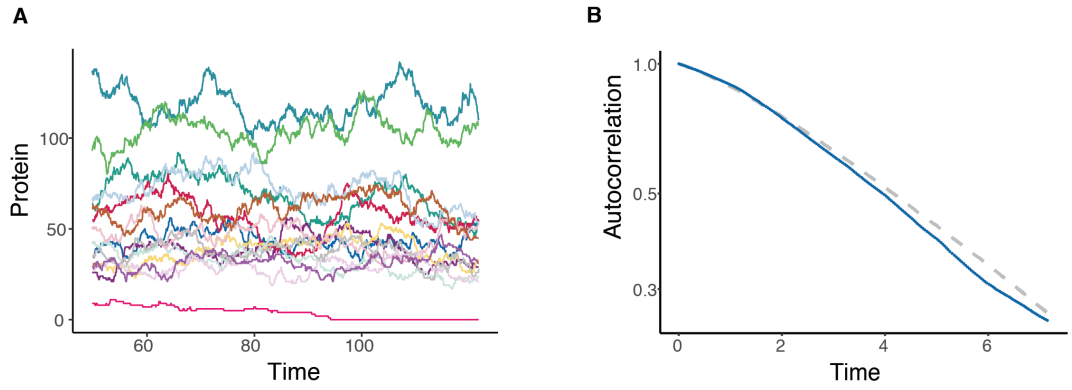


Figure 4.8.: (**A**) Gillespie simulations of the master equation (4.1) with parameters $g_i = d_i = \gamma_i = 1, p_i = 0.1$, $N = 100$ and uniformly distributed exponents $\alpha_{j,i} \in [1, 6]$. Colored lines indicate protein abundances of different genes. (**B**) Autocorrelation functions of protein fluctuations, averaged across all genes, computed after the system reached a steady state (blue line) are in agreement with theoretical predictions (dashed grey line) from Eq. (4.19).

This analytical result is in good agreement with Gillespie simulations of the full stochastic system, Fig. 4.8 A,B. Therefore, for short times, $\omega \to \infty$, the autocorrelation of protein fluctuations decays exponentially as $C_i^{\xi}(|t - t'|) \sim \tau_i^{(1)} e^{-\frac{t}{\tau_i^{(1)}}}$ on a characteristic time scale $\tau_i^{(1)} = [4/(d_i^2\gamma_i^2 - g_i^2\sigma^2)]^{1/4}$. Similarly, in long-time limit, $\omega \to 0$, is also exponential with a characteristic time $\tau_i^{(2)} = \sqrt{(d_i^2 + \gamma_i^2)/(d_i^2\gamma_i^2 - g_i^2\sigma^2)}$ [158]. In contrast to physical glassy systems typical time scale at which temporal correlations in gene expression fluctuations decay is independent of the noise strength. Therefore, protein fluctuations decay rapidly if $d_i\gamma_i \gg g_i\sigma$ and exhibit long-term memory if $d_i\gamma_i \approx g_i\sigma$. $\tau_i^{(1)}$ and $\tau_i^{(2)}$ are purely imaginary if $d_i\gamma_i < g_i\sigma$ meaning that in this case the attractor of the deterministic dynamic is unstable.

We have to bare in mind that we found the shape of connected time autocorrelation in the approximation of unconstrained fluctuations. Even though this is a reasonable approximation in the paramagnetic like state, it is expected to fail in the glassy phase where constraints on fluctuations may change the relaxation dynamics. We will now discuss analytical results of spherically constrained fluctuations. Analytical solutions of the dynamics of binary fluctuations are very hard to get [159] and not of particular interest for the biological system we are studying as they are a crude approximation, so we will not refer anymore to them. In order to study the effect of a glassy

phase dynamically we study Eq. (4.16), with a spherical constraint for the fluctuations, $\sum_i \xi_i(t)^2 = N\,, \sum_i \eta_i(t)^2 = N$, [160–162]. This is formally introduced via Lagrange multipliers $\mu^{\xi}(t), \mu^{\eta}(t)$ as

$$\begin{aligned} \partial_t \xi_i(t) &= g_i \eta_i(t) - d_i \xi(t) - \mu_i^{\xi}(t)\xi(t) + \sqrt{D_i^n} W_i^{\xi}(t) \\ \partial_t \eta_i(t) &= -\gamma_i \eta_i(t) - \mu_i^{\eta}(t)\eta(t) + W_c^{\eta_i}(t)\,, \end{aligned} \tag{4.20}$$

with $\langle W_c^{\eta_i}(t) W_c^{\eta_i}(t')\rangle = D_i^{\eta} + \sigma^2 C_i^{\xi}(t,t')$. From Eq. (4.20) we obtain the equations for the dynamics of autocorrelation functions:

$$\begin{aligned} \partial_t C^{\xi_i}(t,t') &= g_i C^{\xi,\eta}(t,t') - d_i(t) C_i^{\xi}(t,t') - \mu^{\xi}(t) C_i^{\xi}(t,t') + D_i^n \langle W_i^{\xi}(t)\xi_i(t')\rangle \\ \partial_t C_i^{\eta}(t,t') &= -\gamma_i C_i^{\eta}(t,t') - \mu^{\eta}(t) C_i^{\eta}(t,t') + \langle W_c^{\eta_i}(t)\eta_i(t')\rangle\,, \end{aligned} \tag{4.21}$$

with $C_i^{\xi,\eta}(t,t') = \langle \xi_i(t)\eta_i(t')\rangle$. After some algebra (Appendix E.5), and dropping the dependence on i by taking a delta distributions of the parameters, we show that for long enough time, the equations (4.21) that satisfy the fluctuation dissipation theorem (FDT) and time translational invariance ($t' > t, t' - t = \tau$) are simplified to

$$\begin{aligned} \partial_\tau C^{\xi}(\tau) &= g\left(C^{\xi,\eta}(\tau) - C^{\xi,\eta}(0) C^{\xi}(\tau)\right) - \frac{D^n}{2} C^{\xi}(\tau) \\ \partial_\tau C^{\eta}(\tau) &= -\frac{D^m}{2} C^{\eta}(\tau) + \frac{2\sigma^2}{D^m}(1 - C^{\eta}(\tau))^2 C^{\xi}(\tau)\,. \end{aligned} \tag{4.22}$$

As autocorrelation functions should be decreasing functions with respect to τ, FDT does not hold whenever $\partial_\tau C^{\xi}(\tau) > 0, \xi \leftrightarrow \eta$. This last condition is not satisfied when

$$\sigma > \sigma_c = \frac{D^m}{2}\,. \tag{4.23}$$

The autocorrelation functions of bipartite asymmetric spin glasses with spherical constrained fluctuations thus exhibit a plateau in this regime making them long-lived. In Fig. 4.9 A we perform numerical simulation of Eq. (E.66) and show the scaling of protein fluctuations autocorrelation functions for different values of σ and how the plateau emerges for σ close to the critical value. It is interesting to notice that such behaviour is not present for the spherical 2-spin spin glass with unipartite lattice, i.e. only ξ or η. Thus, the two-step process involving protein production may be essential to regulate fluctuations dynamically as a dynamical transition exist even with two-body interactions.

Here, upon mapping the out of equilibrium of gene expression fluctuations to a bipartite asymmetric spherical spin glass, we found that, even for cells in a paramagnetic phase there is another phase defined by a transition line in the phase diagram, where autocorrelation functions of protein and mRNA fluctuations are not exponentially de-

caying at typical length-scales given by the molecular processes in the gene network, but long-lived. Thus the interaction between genes is a key to store information in fluctuations via the emergent collective dynamics of gene expression fluctuations. The existence of a plateau for autocorrelation functions, even outside the glassy phase, signals that memory can be stored in cell states that do not exhibit glassy properties. All together, fluctuations may tell us the specific function of a biological attractor (cell state) as well as its plasticity. As the transition between cell states goes beyond our theory, in the next section, we will take simple biological models of cell state transitions and highlight the potential effect of autocorrelated fluctuations in the regulation of such transitions.

4.3.1. Biological function of correlated fluctuations

To investigate one potential biological function of strongly correlated protein fluctuations we consider a paradigmatic example of a bistable genetic switch. In the simplest case such a switch is given by a self activating gene [163, 164],

$$\partial_t \phi = \alpha \frac{\phi^n}{1+\phi^n} - d\phi + \gamma + \xi(t)\,, \tag{4.24}$$

where $\xi(t)$ represents correlated fluctuations with correlator $C(t,t') = \langle \xi(t)\xi(t')\rangle$. As we found previously, $\langle \xi(t)\xi(t')\rangle = \frac{D_\xi}{\zeta}\exp\{(-|t-t'|/\zeta)\}$ in the paramagnetic state, where ζ is a characteristic time after which fluctuations become uncorrelated ($d \to 1/\zeta$, $D^n \to D^\xi/\zeta^2$). As the exact correlation of glassy fluctuations are extremely hard to get analytically, we define a glassy limit when ζ is greater than the characteristic times set by molecular processes ($\zeta \gg 1/d$). This is far from being an exact form of glassy fluctuations, but it's the simplest form for a process that shows non trivial correlations. We take into account as well multiplicative noise coming from changes in degradation rate by replacing $d \to d + \sqrt{2\lambda}\eta(t)$, with $\langle \eta(t)\eta(t')\rangle = \delta(t-t')$. The system described by Eq. (4.24) has two stable fixed point, corresponding to, respectively, low and high concentrations of proteins. We then ask what is the effect of correlated noise in the transition between these two states. We fixed the parameters as: $\alpha = 5$, $\gamma = 0.1$, $d = 2$, $\lambda = 0.5$ and $n = 2$. We study the statistic of the mean first passage time (MFPT) between the two stable steady states, which is given by writing the approximate Fokker Planck equation (AFPE) associated with Eq. (4.24) [165, 166],

$$\partial_t P(\phi,t) = -\frac{\partial}{\partial\phi}\left[f(\phi)P(\phi,t)\right] + \frac{\partial^2}{\partial^2\phi}\left[D(\phi)P(\phi,t)\right] \tag{4.25}$$

with $D(\phi) = \frac{D_\xi}{2(1-\zeta f'(\phi_s))} + \lambda\phi^2$ and $f(\phi) = \alpha\frac{\phi^n}{1+\phi^n} - dx + \gamma - \lambda\phi$, where ϕ_s is the deterministic steady state. The stationary solution of the AFPE are in the form $P(x) =$

$e^{-\Phi(\phi)}/Z$ and after integrating the equation for the stationary solution over dx and substituting the ansatz exponential we obtain,

$$\Phi(x) = \int dx \, \frac{\partial_x D(x) - f(x)}{D(x)} = ln(D(x)) - \int dx \frac{f(x)}{D(x)} \, . \tag{4.26}$$

Upon inserting the definition of $f(x)$ and $D(x)$, we get the exact expression for the stationary probability,

$$P(x) = \exp\left(-\frac{(d+\lambda)\log\left(Q+\lambda x^2\right)}{2\lambda} + \frac{(Q(\alpha+\gamma)-\gamma\lambda)\tan^{-1}\left(\frac{\sqrt{\lambda}x}{\sqrt{Q}}\right)}{\sqrt{\lambda}\sqrt{Q}(Q-\lambda)} - \frac{\alpha\tan^{-1}(x)}{Q-\lambda} \right)/Z \tag{4.27}$$

where $Q = \frac{D^{\xi}}{2(1-\zeta f'(x_s))}$. From Eq. (4.27) we obtain an expression for the approximated MFPT as an Arrenhius law,

$$\tau_{1\to 2} = \frac{1}{2\pi\Phi''(\phi_1)\Phi''(\phi_2)} e^{-[\Phi(\phi_1)-\Phi(\phi_2)]} \, 1 \leftrightarrow 2 \, , \tag{4.28}$$

where ϕ_1 and ϕ_2 are the two attractors of the deterministic dynamics and $\Phi''(\phi_i) = \partial^2_{\phi_i}\Phi(\phi)|_{\phi=\phi_i}, i = 1, 2$. The escape rates are plotted in Fig. 4.9. In particular the escape rate $\tau_{1\to 2}$ is non monotonic in λ, meaning that a cell can regulate the state of the gene (in terms of expression) by varying the autocorrelation time. In particular, the mean first passage time between two state of the genes initially decreases with increasing correlation time of protein fluctuations, meaning that the transition between gene expression states can happen at time scales faster with respect to uncorrelated fluctuations. On the other hand, in the limit of glassy fluctuations and so long-lived fluctuations, both mean first passage times have higher values compared to non correlated fluctuations, thus fluctuations fix the self-regulating gene in one of the two states.

In this section, we thus characterized the possible implications of glassy fluctuations on the dynamics of genetic switches by studying the stability of a self regulating genes. We found that a potential role of correlated fluctuations is to decrease, in comparison to uncorrelated fluctuations, the mean first passage times from one steady state of the self regulating gene to the other one. In contrast the mean first passage time for the opposite transition is increased. This process thus select one steady state and can play a role in fixing a particular cell state by limiting its plasticity. As we shown in this chapters transition between cell states are as well regulated by interactions between genes, which may lead to drastic changes of the expression of multiple genes, that a self-regulating gene model cannot explain. In the last section, we quantify changes in gene expression during cell state transitions and develop a minimal theory which is in qualitative agreement with the data.

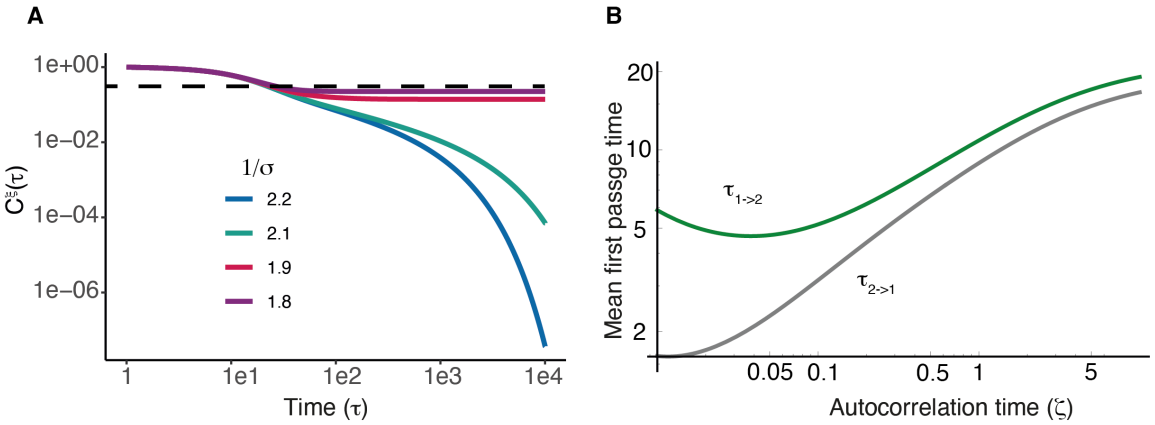


Figure 4.9.: (**A**) Protein autocorrelation functions for different values of σ. A plateau (dashed black line) is observed when the interaction strength σ approaches the critical value $\sigma_c = \frac{D^m}{2}$ from below, Eq. (E.69). The parameters are $D^m = 3, g = 1, D_n = 0.1$ and Eq. (E.69) are solved with a 2nd order Runge Kutta algorithm with $dt = 0.01$. (**B**) Escape times $\tau_{1\to2}, \tau_{2\to1}$ following Arrenhius law as described by Eq. (4.28). $\tau_{1\to2}$ is a non monotonic function of the protein fluctuations autocorrelation time ζ.

4.4. Cell state transitions

In order to study transitions between cell states we need to first characterize them in terms of RNA-Seq experiments. A limitations of single-cell sequencing is that once the cell is sequenced, it dies, such that we do not have any information about the dynamics changes in gene expression for a single cell. To overcome this problem, we use algorithms that defines a pseudo-time [167], which map an effective time (no unit of measure) to each cell. By doing so, we can interpret gene expression changes between cells as the dynamic of expression for a single cell. We can then visualize on the UMAP the pseudotime (Fig. 4.10 A) and define potential trajectories. Once a pseudotime is defined we can follow cells during the trajectory, in particular can see how they evolve in the phase diagram Fig. 4.2. We first define an overlap for individual cells, by computing the overlap width of the distribution given by the overlaps between the 50 nearest neighbour in the UMAP of the given cell. The results are given in the top right panel of Fig. 4.10 B for all the data sets. Even though the overlap is heterogeneous we don't see any significant change along the transition. This result confirms that cell state transitions cannot be seen as adiabatic changes in the spin glass landscape, rather as transitions between different landscapes, where each landscape identifies a particular cell state and its heterogeneity. This does not come as a surprise, as the developed theory was expected to hold only close to a steady state of the gene regulatory network.

To quantify the changes in the spin glass landscapes we compute the Spearman correlation coefficient between all pairs of genes between cells in a given neighbour of

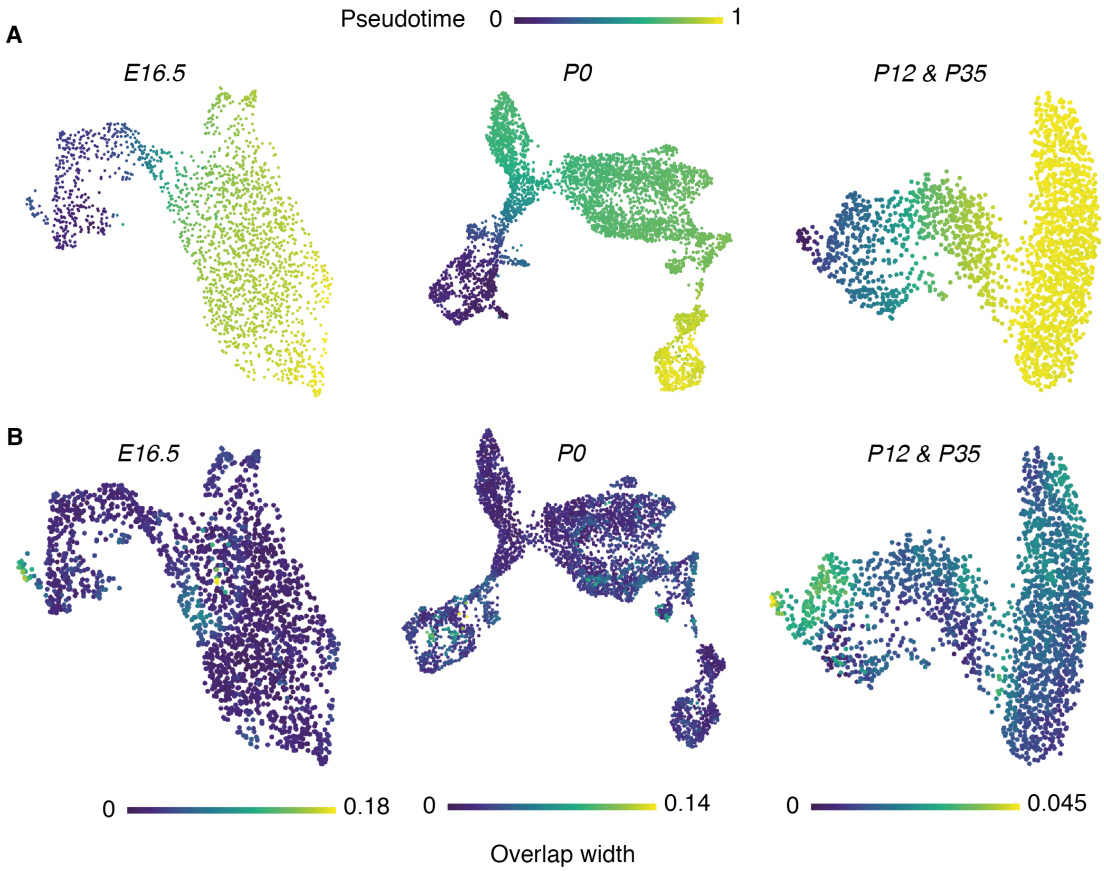


Figure 4.10.: (**A**) Pseudotime for single-cell sequencing shown in the UMAP representation. (**B**) Overlap widths for individual cells are computed as the standard deviation of the overlaps distribution of all the pairs of cells from the first fifty nearest neighbours in the k-means graph of a given cell and the cell itself.

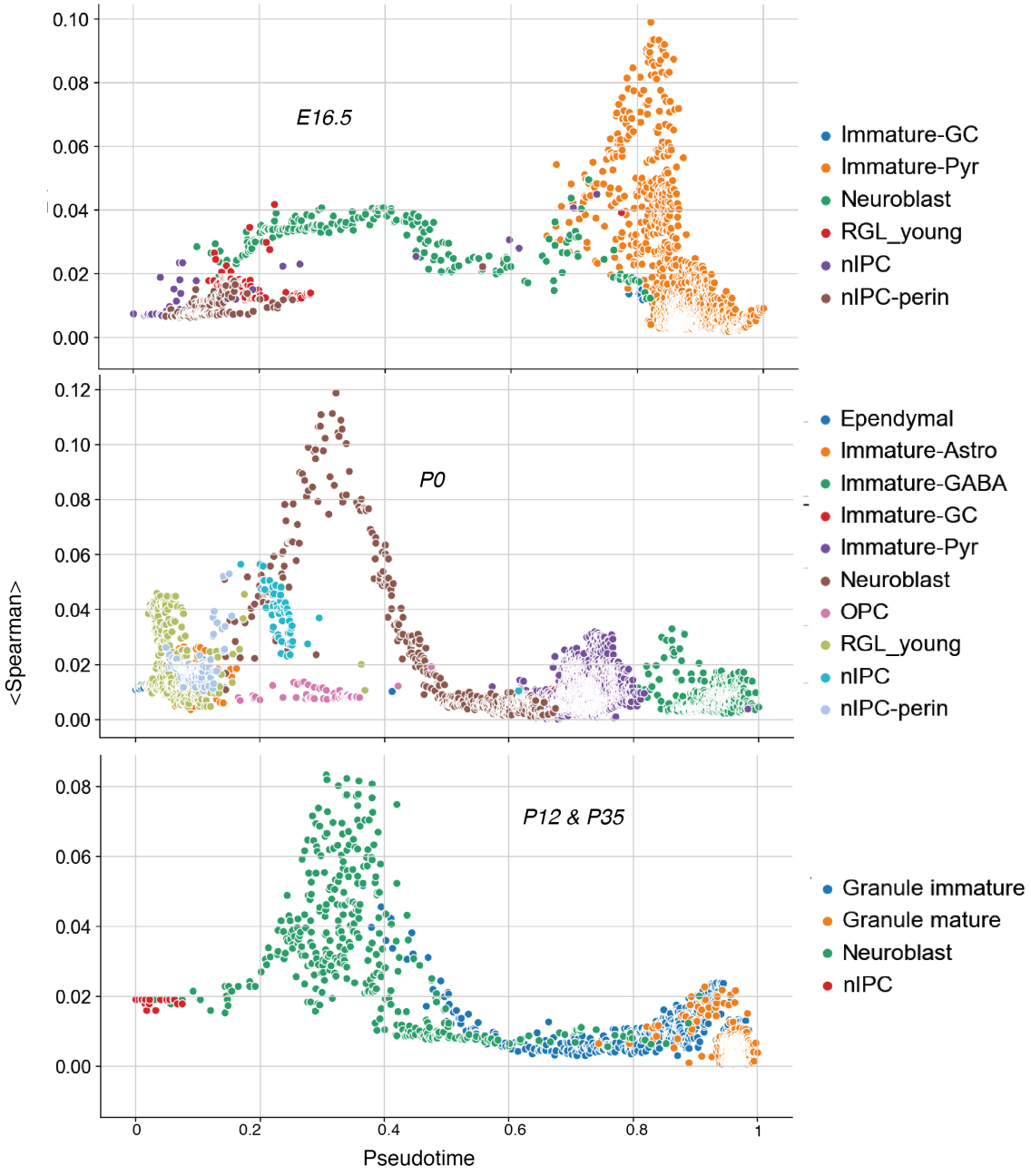


Figure 4.11.: The average Spearman correlation coefficient is calculated along the pseudotime trajectories for all the cells which are the 50 neighbours of a given cell in the k-means graph. The increase in the correlation values is associated to a transition in the high dimensional gene expression space.

a specific cell, this defines a Spearman coefficient for each cell. We plot its absolute value (the sign just reflects activation/inhibition) along pseudotime and we find that the Spearman coefficient has a bump at a given pseudotime which in the UMAPs corresponds to linkages between clusters (Fig. 4.11 bottom) for all the data set analysed. In order to understand how this bump emerges during cell state transitions, we take as a model for cell fate decision the toggle switch (Fig. 4.12 top left), which is a well studied minimal network which exhibit non trivial behaviour, such as bifurcation [168]. A toggle switch is defined by a couple differential equation of the form,

$$\begin{aligned} \partial_t n_1 &= r_{21} \frac{1}{1+n_2^{\alpha_{21}}} + p_1 - d_1 n_1 \\ \partial_t n_2 &= r_{21} \frac{1}{1+n_1^{\alpha_{12}}} + p_2 - d_2 n_2 \,. \end{aligned} \tag{4.29}$$

n_1, n_2 in Eq. (4.29) are the protein concentrations of two genes which inhibit each other with rate r_{12}, r_{21} and via Hill functions. p_1, d_1 are respectively protein production and degradation rate for the first gene and similarly for the second. We will for now on, for the sake of simplicity that equal degradation, interaction rate and Hill coefficient α_{ij} for both genes. The toggle switch has the interesting property that the steady state concentration of a protein bifurcate by increasing the interaction rate (quench), such that it will take one of the two branches in Fig. 4.12 B (bottom left, blue line), if the production rate is fixed. On the other hand the other protein will take the opposite branch such that the symmetry is broken. The solution where both genes have same steady state values of the protein is unstable after a value r_c (dashed blue line) [169, 170]. This is not the only way to get a switch as we can in principle fix the interaction rate and change the ratio between the production rates, such that when whenever the ratio is 1 there is a transition between the most expressed genes. We thus model cell fate transition by either quenching the interaction rate keeping fixed all the other rates or by quenching the production rate of one of the two genes. We then compute the Spearman correlation coefficient between the two genes at fixed rate along the quench. The results are shown in the right panel of Fig. 4.12. We can see a bump in the Spearman correlation, similar to the experimental results, whenever there is a quench of the interaction rate. On the other hand, there is no sign of transition whenever the quench is on the production rate. A possible speculative interpretation of these results in agreement with our theories is the following: a quench in interaction rate is due to the change of intrinsic factors, such as DNA methylation, chromatin structure etc... whilst a quench in production, or upregulation of a given gene is an extrinsic factors, as for example signalling molecules. We thus might have a way to quantify the different roles of intrinsic and extrinsic factors during cell state transitions. This result would

not clearly rule out the importance of extrinsic factor during cell state transition, but it would rather point at an essential role of intrinsic factors. As these are only preliminary results, we cannot make any strong point yet, but they are promising enough to further explore them in the future.

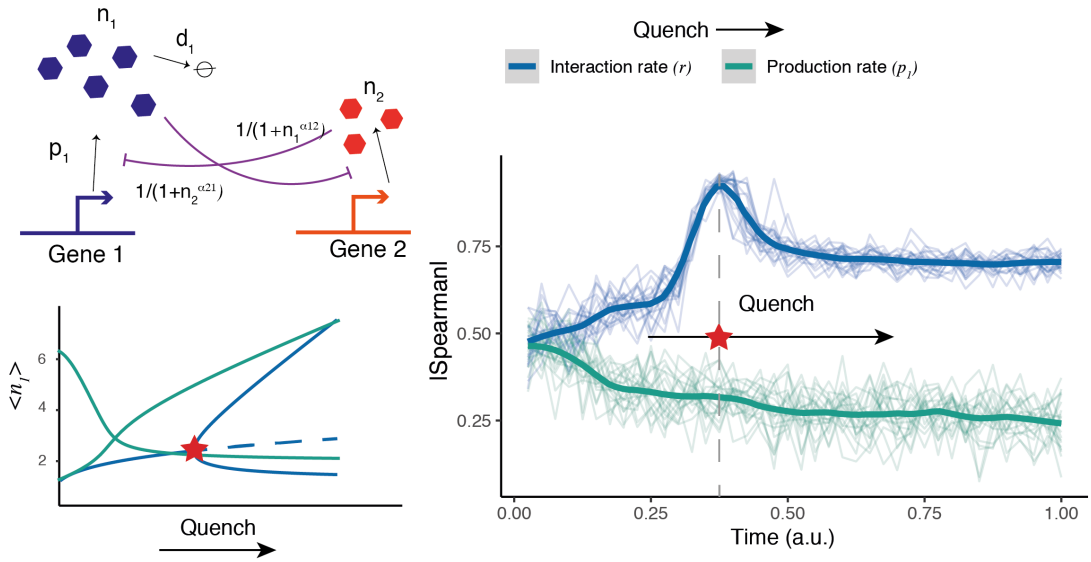


Figure 4.12.: Toggle switch (top left) and stable fixed point of Eq. (4.29) for different quenches of the parameters (bottom). The blue lines are steady state for a single gene (gene 1), the other genes has the same steady state value before the bifurcation and take the opposite branch after the transition. (Right) Spearman correlation coefficient (solid lines) averaged over ten different realizations (light curves) of Eq. (4.29). In both quenches the rates goes from zero to the. For the interaction quench the rate are $p_1 = p_2 = 1, r_{12} = r_{21} = 1$ and for the production quench they are $p_2 = 1, r_{12} = r_{21} = 1$. The Hill coefficients are in all the simulations $\alpha_{12} = \alpha_{21} = 2$ and degradation rates are one. The quenching protocol is an increases of either r or p_1 every 20000 time steps up to the transition point by 0.05 and then it is increase by 0.1. All simulations are run for 1e6 time steps with Euler Mayorana method with time step $dt = 0.01$. Different realisation are obtained by adding Gaussian uncorrelated white noise in both equations (4.29) with variance 0.1.

4.5. Summary and discussion

In this chapter, we developed a theory to study fluctuations of gene expression in gene regulatory networks based on RNA sequencing data, which, as outlined in Sec. 1.2.2, encodes information on the expression of individual genes. As these experiments come along with technical uncertainties, we proceeded with a reverse approach with respect to the previous chapters. In Sec. 4.1 we developed a theoretical framework for general gene regulatory networks, where mRNA are transcribed, translated into proteins and proteins of a given gene may interact with other genes as transcription

factors changing the transcriptional output of the targeted gene. We were able to map the dynamics of genes fluctuations close to a steady state of the network to the dynamics of an Ising bipartite spin glass model. Upon using methods originally developed in glassy systems (Sec. 1.3.3), we study the emergent behaviour of such system and found out that gene regulatory networks may exhibit a transition from a phase with uncorrelated fluctuations to one glassy phase, were fluctuations are non trivially correlated. In Sec. 4.3 we study the specific form of these correlations and found out that depending on the structure of the network as well as the particular steady state they may be exponentially autocorrelated in time or exhibit a plateau, such that autocorrelations are long-lived in time. In Sec. 4.2, we used all the theoretical results to seek which experimental observables are relevant for the study of GRNs. Specifically, we first identify, with machine learning clustering methods, the steady states of the gene network as broadly defined cell states. Later, we study overlaps distributions, for three different mouse brain RNA-Seq experiments, which measures the similarities between replicas of a system (individual cells) in the same state and found that stem cells exhibit a broader distributions with respect to differentiated one. Surprisingly, these results point out to a new way of analysing RNA-Seq experiments and the need to rethink epigenetic landscapes and cell state transitions. Specifically, cell states can be thought as rough landscapes, where valley are possible macrostates of fluctuations with the same statistical ensemble. In some regimes, the dynamics between different fluctuations macrostates is glassy, such that fluctuations are long-lived and it is possible to store information in gene expression fluctuations. In order to investigate the potential role of glassy dynamics, we studied a paradigmatic model of a self-regulating gene and we find that correlated fluctuations play a role in the regulating the changes between different expression states of a gene, by controlling the mean first passage times between different steady states value. Finally, we quantify changes in gene expression between different cells state by measuring gene-gene correlations along cell state transitions. We found out that transition between cell states are associated with a peak in gene-gene correlations which is in qualitative agreement with theoretical gene-gene correlations of prototypical models of cellular symmetry breaking All together, the picture we have in mind of cell states and cell state transitions is very rough and intricate, as we found how different layers of regulations act together in a non-trivial way to determine possible cell states. It would be great to have a theory which encodes all the layers and interactions between layers at different scales. Unfortunately, we do believe it is quite unreasonable and maybe useless to develop such complicated model where the theoretical understanding is minimal and the space of parameters is too huge to explore. We would like to cheat a bit and skip all these complexities underlying cell fate and try to develop in the next section a rather easy, but much more understandable theory of multiscale interacting complex systems.

5. Collective Dynamics of Multiscale Interacting Complex Systems

Complex systems often interact on multiple scales and the emergence of a collective behaviour is often understood when all these different scales are taken into account. As a paradigmatic example we can think of development. In the previous chapters we zoomed inside a cell, arriving to the nucleus and we studied the role of epigenetic factors during cell state transitions. These are, of course, not the only relevant factors that play a role to determine which is the state of the cell. As mentioned in the introduction, signalling factors, induce cells to transit to a particular state. It is then outstanding and mesmerising, that despite all these complicated multi scale interactions, cells do take precise decision in space and time and an organism is formed. In this last chapter, we derive a general theory for complex systems interacting on multiple scales. In particular, we derive bounds that divide regions of the phase space where the system is in a particular macroscopic state. In the first section we review previous and outstanding works done in complex systems and point out at the limitation that brought us to develop a new theory.

5.1. The May bound

In his work, ahead of time [171], May described how complex systems reache stability when governed by a huge number of interactions. In particular, let's consider a complex systems with continuous components $\boldsymbol{\rho} = (\rho_1, \dots, \rho_L)$. Such a complex system is described by a network of interactions, such that we can write the deterministic time evolution of the components as a set of N coupled differential equations

$$\partial_t \rho_i = f_i(\boldsymbol{\rho})\,. \tag{5.1}$$

Eq. (5.1) describes such general interacting system with arbitrary functions f_i, which encode how the components ρ_i are changed by the other components. Such system may reach a stationary state such that $\partial_t \rho_i = 0, \forall i$. Without entering deeply into dynamical systems theory [109], it is enough to know that there might be multiple of such stationary states and that they are divided into two big classes: stable and unstable. When a system reaches a stable stationary state, that we indicate as $\boldsymbol{\rho}^*$, if

every component is minimally perturbed, $\rho_i = \rho_i^* + \delta\rho_i$, with $\delta\rho_i \ll 1$, the system will return, after a certain time, to the stationary values. Instead, a complex system in an unstable stationary state, when minimally perturbed, will go far away from it till it reaches a stationary state or attractor. Technically we can express the stability of a complex systems via

$$\partial_t \boldsymbol{\delta\rho} = \hat{A}\boldsymbol{\delta\rho}\,, \tag{5.2}$$

with the components of the matrix $\hat{A}$, $A_{i,j} = \frac{\partial f_i(\rho_j^*)}{\partial \rho_j}$. May realised that whenever this matrix has a pretty simple form the stability of the system is fully characterised. In particular, if the entries of the matrix are symmetric and distributed according to a distribution with finite mean μ and variance σ^2, the stability of the system is compactly described with a simple relation: the *May bound*. The system is almost certainly stable if for large L (number of components) [171, 172],

$$\sigma < L^{-1/2}\,. \tag{5.3}$$

We arrived to a very general condition for stability for a complex system described by a continuous field $\boldsymbol{\rho}$ satisfying Eq.(5.1). Even though this is a very general and powerful result, there are still some details that must be explored more deeply. In particular, in this thesis we have often faced field theoris describing a continuous field $\rho_i(\vec{z})$, where the variable $\vec{z}$ may play the role of spatial coordinate or any continuous variable. The argument by May still applies to this field as long as there are no processes which shape the field along the $\vec{z}$ directions, such that $\partial_{z_n}\rho_i(\vec{z}) = 0\,\forall i \in [1, L], n \in [1, d]$, with L the number of components of the field $\boldsymbol{\rho}(\vec{z})$ and d the spatial dimension. As we have shown, many biological processes do not satisfy this constraint and in the next section we will deal exactly with the effect of the breaking of this constraint in complex systems. Moreover, complex systems may interact in the $\vec{z}$ space on multiple length scales, such that new phenomena emerge, such as pattern formation.

5.2. Field theory of multiscale processes

We consider a stochastic particle system of N particles, where each of them, indexed by n, is characterised by a categorical variable $i \in [1, L]$ and a position $\vec{z}_n$ in a space endowed with a Euclidean metric (metric space). We then define the density of particles (field) at position $\vec{z}$ as $\rho_i(\vec{z}) = \sum_n \delta(\vec{z}_n^i - \vec{z})$, where $\vec{z}_n^i$ are the positions of the particles of type i.

The time evolution of $\rho(t)$ is determined by conservative processes, which maintain the global density, and non-conservative processes. In the metric space these processes

are characterised by different typical length scales, ζ, which give a typical distance over which particles interact. We denote their rate by a vectorial functional $\boldsymbol{f}_\zeta[\boldsymbol{\rho}]$. In the mean-field or deterministic limit, the time evolution of $\boldsymbol{\rho}$ obeys the following equation

$$\partial_t \boldsymbol{\rho}(\vec{z},t) = \int d\vec{z}' \int_0^\infty \mathrm{d}\zeta\, e^{-|\vec{z}-\vec{z}'|/\zeta} \boldsymbol{f}_\zeta[\boldsymbol{\rho}(\vec{z}',t),\vec{z}']\,. \tag{5.4}$$

In Eq.(5.4), we required that there is a length scale associated to the interactions along the metric space $\vec{z}$. As a minimal choice for this interactions we take an exponential with a cutoff ζ. As $\boldsymbol{\rho}(\vec{z},t) = (\rho_1(\vec{z},t),\dots,\rho_L(\vec{z},t))$, the functional $\mathbf{f}$ must have the same number of vectorial components. Moreover, in general $\mathbf{f}$ has a dependency on the cutoff scale ζ, as different length scales may not contribute equally to the dynamics.

In order to define the transition rates we write $\mathbf{f}_\zeta[\boldsymbol{\rho}] = \mathbf{f}_\zeta^C[\boldsymbol{\rho}] + \mathbf{f}_\zeta^{NC}[\boldsymbol{\rho}]$. We formally divided the time scales at which the only physical processes (conservative and not conservative) happen. As we introduced categorical and finite range components of the density $\boldsymbol{\rho}$, in the following we define contributions from conservative and non-conservative processes for these two sources of interactions.

To begin, as a minimal model for a non-conservative processes, we take the non-local birt-death process [173] such that $\mathbf{f}_\zeta^{NC}[\boldsymbol{\rho}] = \lambda(2\mathbf{h}\,[\boldsymbol{\rho},\vec{z}] - 1) \circ \boldsymbol{\rho}(\vec{z})$, which we will refer to as $f_i[\rho_i]$. We argue that this is a minimal non-linear process which may contain different length scales due to non-locality of the kernel $\mathbf{h}$.

On the other hand, we express conservative processes as a stochastic dynamics for single particle in a given potential, which may depend on the full state of the system. In general, such processes can be described as the dynamics of individual particles in the continuous variable $\vec{z}$ as

$$\partial_t \vec{z}_n^{\,i} = -\sum_{j=1}^{L}\sum_{m}^{N} \boldsymbol{\nabla} V(\{\vec{z}_m^{\,j}\}) + \sqrt{2D_{\rho_i}}\vec{\xi}_i^{\,n}(t)\,, \tag{5.5}$$

with $\langle \vec{\xi}_i^{\,n}(t)\vec{\xi}_j^{\,m}(t')\rangle = \delta(t-t')\delta_{n,m}\delta_{i,j}$. The only assumptions we made, which greatly simplifies the dynamics while keeping it general, are the overdamped limits of the dynamics and the Gaussian white noise. The last introduced process (5.5) thus conserves the number of particles. We will represent it as $f_\zeta^{C,i}[\boldsymbol{\rho}]$ and it is derived from the density representation of Eq. (5.5) [174, 175].

Hence, we are left to describe global scales encoded in the categorial dependence of the functional $\mathbf{f}_\zeta[\boldsymbol{\rho}]$. As the simplest non trivial form, we consider a quadratic term in $K_{ij}\rho_i\rho_j$.

Before studying the analysis of the stochastic dynamics we need to find a form for the kernel $\mathbf{h}$ that can encode different length scales. We then expand the kernel $\mathbf{h}$ to first order as

$$\mathbf{h}[\boldsymbol{\rho},\vec{z}] \approx \mathbf{h}_0 + \mathbf{h}_1 \int \mathrm{d}\vec{y}\,\nu e^{-|\vec{z}-\vec{y}|/\zeta_1}\boldsymbol{\rho}(\vec{y}) + \mathcal{O}(\rho^2)\,, \tag{5.6}$$

where ν is a parameter stemming from the expansion. We represent the second term in the expansion Eq. (5.6) as a solution $\phi(\vec{z})$ of a partial differential equation of the form [176]

$$\alpha_i \phi_i - D_{\phi_i} \nabla^2 \phi_i = \nu_i - \gamma_i \rho_i \,, \tag{5.7}$$

which simplifies Eq. (5.4) as it reduces the convolution integral to the solution of a production-degradation-diffusion equation. The constants are arbitrary, so we set them conveniently to: $h_0^i = \nu_i/\alpha_i$, $h_1^i = -\gamma_i \sqrt{D_{\phi_i} \alpha_i}$ and $\zeta_1^i = \sqrt{D_{\phi_i}/\alpha_i}$.

Combining all this processes together (Fig. 5.1) and taking a two-body potential which depends only on the distance between particles in the same global space i, i.e. $V = V(|\vec{z}_n^i - \vec{z}_m^j|)\delta_{i,j}$, we obtain a stochastic partial differential equation for $\rho_i(\vec{z})$ that is

$$\partial_t \rho_i = f_i\left[\rho_i\right] + \sum_{j \neq i} K_{ji} \rho_j \rho_i + D_{\rho_i} \nabla^2 \rho_i + \boldsymbol{\nabla} \cdot \left[\rho_i \int d\vec{y}\, \boldsymbol{\nabla} V \rho_i(\vec{y}) \right] + \eta_i + \boldsymbol{\nabla} \cdot \vec{\xi}_i \,, \tag{5.8}$$

where ξ and η are Gaussian white noise with

$$\begin{aligned} \langle \eta_i(\vec{z}, t) \eta_j(\vec{z'}, t') \rangle &= \lambda_i \rho_i(\vec{z}, t) \delta(t - t') \delta(\vec{z} - \vec{z}') \delta_{i,j} \,, \\ \langle \xi_i(\vec{z}, t) \xi_j(\vec{z'}, t') \rangle &= 2 D_{\rho_i} \rho_i(\vec{z}, t) \delta(t - t') \delta(\vec{z} - \vec{z}') \delta_{i,j} \,. \end{aligned} \tag{5.9}$$

$\eta(\vec{z}, t)$ is a multiplicative noise coming from the system size expansion of the birth-death process [124, 173] while $\vec{\xi}_i(\vec{z}, t)$ comes from the density representation of Eq. (5.5).

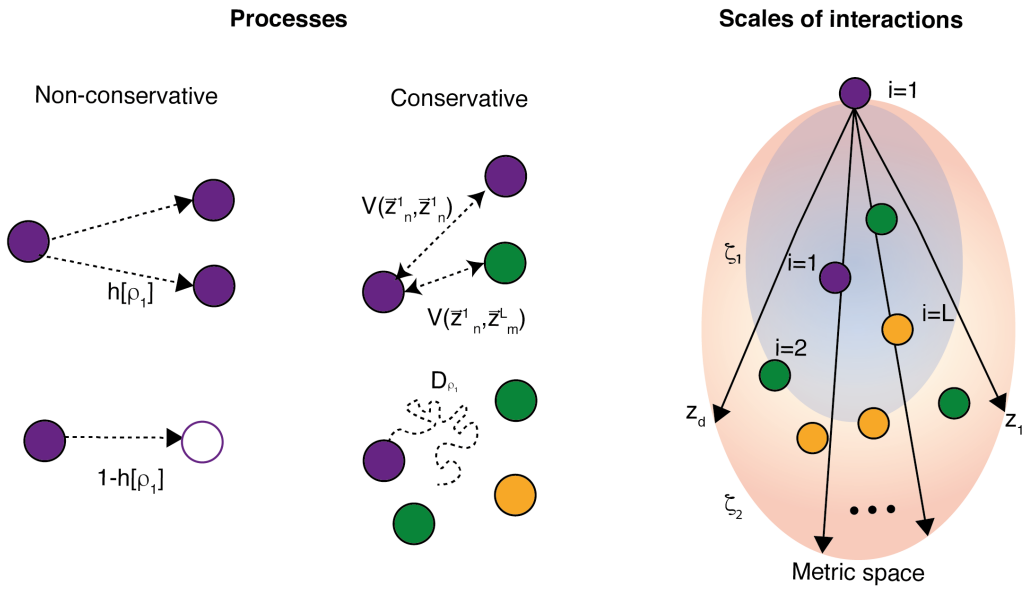


Figure 5.1.: Schematic representation of the main different multi scale processes. The processes are: birth-death with a non-local kernel and diffusion in the metric space $\vec{z}$. Particles are then labelled with different categorical variables $i = 1, \dots L$ and they interact on the metric and categorical space, with dynamics described by Eq. (5.8).

The two-body potential has a similar form as the intermediate-range potential and so, for sake of simplicity we neglect it in favour of the non-conservative process. We do the same for the infinite range interaction term, as it would have been possible to add a term as $\nabla^2 \left[\sum_{j\neq i} L_{ij}\rho_i\rho_j\right]$. An equally feasible approach would have been to neglect the intermediate and global range interactions in the non conservative processes in favour of the conservative ones. We argue that for scalar fields there is not a zero-range scale associated to conservative processes whilst in case of vectorial fields such processes can be, for example, rotational diffusion [177]. A more detailed study of the effect of these terms on such systems will be done in further studies. Here, we set $L_{ij} = 0$ and $V = 0$ everywhere.

We begin by considering the most simple, but yet instructive case, in which we neglect global space components such that $\boldsymbol{\rho} \to \rho$. As explained in the introduction of this chapter we first study the stationary solution of the deterministic part of Eq. (5.8). We find only two stationary solutions: $(\rho^*, \phi^*) = \left(0, \frac{\nu}{\alpha}\right)$ and $\left(\frac{2\nu-\alpha}{2\gamma}, \frac{1}{2}\right)$, the latter being meaningful only for $\nu > \alpha/2$. We then ask whether these solutions are stable against a local perturbation. In the small noise limit, we perform a linear stability analysis of Eq. (5.8). By dropping the index i (as there is no global component), we set $\rho(\vec{z}, t) = \rho^* + \delta\rho e^{i\vec{k}\vec{z}} e^{\omega t} + c.c.$ and similarly for $\phi(\vec{z}, t)$. The $(0, \nu/\alpha)$ steady state solution is stable whenever $\nu < \alpha/2$ and the stability of the non-trivial fixed point leads to the condition

$$\omega = \lambda(2\phi^* - 1) - D_\rho k^2 - 2\lambda\rho^* \frac{\gamma}{D_\phi k^2 + \alpha} > 0\,. \tag{5.10}$$

As the sign of the terms in k are always negative, there might be a Type III instability in the Cross-Hoenberg classification [109] only when $\nu < \frac{\alpha}{2}$. However, the latter inequality would violate the condition of existence of the fixed point. As a consequence, when this solution exists it is always stable.

In the following, we give an intuitive explanation of the possible consequences on the stability due to global components. Later, we will give a more detailed analysis of the dynamics of the system governed by Eq. (5.8). In the presence of the global components, there might be multiple stationary solutions such that $\phi^* \neq \frac{1}{2}$. If one such fixed point exists and they contribute to the stability for some range of parameters, then the instability condition is satisfied for some k_c if $\phi^* > 1/2\,, \rho^* > 0$, Fig. 5.2.

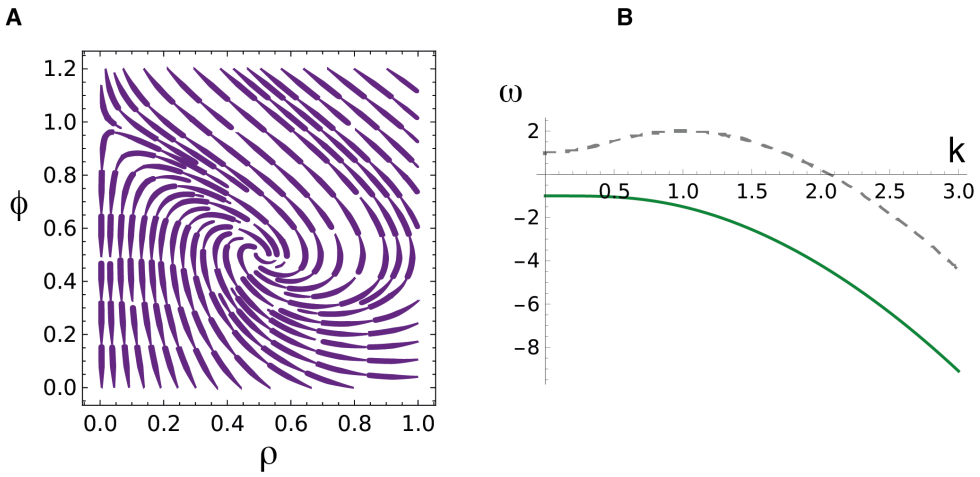


Figure 5.2.: (**A**) Phase portrait of the one category simplified model derived from Eq. (5.8). Streamlines around the stable and unstable fixed point. (**B**) Dispersion relation from the stability of the simplified model (green) which shows no instability for any wavevector k. Dispersion relation in case another non-trivial fixed point exists (dashed gray). In the latter case, a non-zero wavevector associated with the fastest growing mode exists suggesting that a pattern is possible when infinite range interactions are included in the model.

5.2.1. Stationary distributions

In order to obtain further analytical insights into the dynamics of (5.8), we consider the effect of global components $K_{ij}\rho_i(\vec{z})\rho_j(\vec{z})$ on the dynamics of Eq. (5.8). We take disorder into account, embodying it into K_{ij}, which is in turn taken to be Gaussian distributed with mean μ/L, variance σ^2/L and covariance equal to $\epsilon\sigma^2/L$. With this choice, we can interpret the dynamics of Eq. (5.8) as the one of a soft spin glass where the site is given by i. Upon writing the Martin-Siggia-Rose-Jannsen-De Dominicis field theoretical representation of the Langevin equation (5.8) [178, 179] (Sec. 1.3.1), we arrive at the following representation of the dynamics of individual densities

$$\partial_t \rho_i = L\left[\rho_i\right] + \epsilon\sigma^2\rho_i \int_{t',\vec{z}'} \chi_i(t,t',\vec{z},\vec{z}')\rho_i(\vec{z}',t') + W_i\,. \tag{5.11}$$

$L\left[\rho_i\right]$ encodes all the terms associated with the local space $\vec{z}$. W_i is a Gaussian colored noise with zero mean and $\langle W_i(\vec{z},t)W_j(\vec{z}',t')\rangle = \sigma^2 C_i(\vec{z},\vec{z}',t,t')$. The response and correlation functions of the individual densities are respectively, $\chi_i(\vec{z},\vec{z}',t,t') = \langle\frac{\partial\rho_i(\vec{z},t)}{\partial W_i(\vec{z}',t')}\rangle, C_i(\vec{z},\vec{z}',t,t') = \langle\rho(\vec{z},t)\rho(\vec{z}',t')\rangle$. We further define $M(\vec{z},t)$ as the average density of individuals, which will come in hand later. MRSJD path integral formulation has the advantages of passing from $2L$ coupled equations to only L groups of two coupled equations at the price of adding a coloured noise W_i. We look for a stationary homogeneous solution of Eq.(5.11) such that we can set $\int \mathrm{d}t'\,\chi_i(t,t') = \chi_i^*$ and $C_i(t,t') = \langle\rho^{*2}\rangle = c_i$, consequentially $W^* = \sqrt{\sigma^2 c_i}w$ [180, 181], where w is a Gaussian random variable with unitary variance and zero mean. We initially focus on the deterministic part of Eq.(5.11), where multiplicative noise contributions coming from density fluctuations and birth-death processes are neglected. On the other hand, the noise stemming from the infinite range interactions is retained as it is dominant and it comes from the reduction of the deterministic part of systems of equations (5.8). In this regime, there are two pairs of homogeneous solutions of Eq. (5.11), namely $(\rho^*,\phi^*) = (0,\nu/\alpha)$ and

$$\rho^* = \alpha\frac{w\sigma\sqrt{c}+r}{\Delta(\chi^*)}\Theta\left(\frac{w\sigma\sqrt{c}+r}{\Delta(\chi^*)}\right)\,, \tag{5.12}$$

with $\Delta(\chi^*) = 2\lambda\nu - \alpha\epsilon\sigma^2\chi^*$, $r = \lambda(2\nu - \alpha^{-1})$ and $\Theta(x)$ the Heaviside theta. We dropped the index i by assuming delta distribution of the parameters. As w is a Gaussian random variable ρ^* are distributed according to a truncated Gaussian. Eq. (5.12) is supported by the self-consistency equation of stationary response, correlation functions and average density. Assuming that $\Delta(\chi^*) > 0$, which will be justified at posteriori, they are written as:

$$\begin{aligned}
\rho_s &= \int_{-\kappa}^{\infty} Dw \\
\chi^* &= \frac{\alpha}{\Delta(\chi^*)}\int_{-\kappa}^{\infty} Dw \\
M^* &= \frac{\alpha\sigma\sqrt{c}}{\Delta(\chi^*)}\int_{-\kappa}^{\infty} Dw\,(w+\kappa) \\
1 &= \frac{\alpha^2\sigma^2}{\Delta(\chi^*)^2}\int_{-\kappa}^{\infty} Dw\,(w+\kappa)^2\,,
\end{aligned} \tag{5.13}$$

where $Dw = dwe^{-w^2/2}/\sqrt{2\pi}$ and $\kappa = r/\sigma\sqrt{c}$. These are a set of four closed equations which can be self-consistently solved for different values of the parameters.

We simply reason that due to the form of the non-conservative processes, there is no

spontaneous creation of particles such that the stationary state $(\rho^*, \phi^*) = (0, \nu/\alpha)$ is an absorbing state [102]. Heuristically, the probability that k out of L global spaces are not null is given by a binomial with parameter ρ_s. This is possible thanks to representation of Eq. (5.4) in the form of Eq. (5.11) as the equations for single particles are decoupled. We are then left with the original question: how does the stability of the systems changes due to the presence of global and local components? It must not come as a surprise that the analysis of correlation functions will come in hand to address this question.

5.2.2. Spatial correlation functions and density fluctuations

To conclude this chapter, we want to analyse the role of non-locality in correlation functions of individual densities in the local space, $\langle \delta\rho(\vec{z}, t)\delta\rho(\vec{z}', t')\rangle$. A linear equation for a spatio-temporal perturbation satisfies

$$\begin{aligned}\partial_t \delta\rho(\vec{z}, t) =& \rho^* \left[g(\vec{z})\delta\rho(\vec{z}) + \epsilon\sigma^2 \int_{t'} \chi(t, t', \vec{z}, \vec{z}')\delta\rho(\vec{z}', t') \right] + \\ &+ \boldsymbol{\nabla} \cdot \sqrt{2D_\rho \rho^*}\xi(\vec{z}, t) + \delta W(\vec{z}, t) + \sqrt{\lambda\rho^*}\eta(\vec{z}, t)\,,\end{aligned} \tag{5.14}$$

where $g(\vec{k})$ is the Fourier transform of $g(\vec{z})$: $g(\vec{k}) = -\left(\frac{D_\rho}{\rho^*}\vec{k}^2 + \frac{2\lambda\gamma}{\alpha + D_\phi \vec{k}^2}\right)$ and $\delta W(\vec{z}, t) = \sigma^2 \langle \delta\rho(\vec{z}, t)\delta\rho(\vec{0}, 0)\rangle$. Upon Fourier transforming the previous equation, correlation function are written in Fourier components as

$$C(\vec{k}, \vec{k}', \omega, \omega') = \frac{\Lambda(\vec{k})\delta(\vec{k} + \vec{k}')\delta(\omega + \omega')}{\left\langle |\, i\omega/\rho^* - \Omega(k)|^{-2} \right\rangle_+^{-1} - \rho_s\sigma^2}\,, \tag{5.15}$$

where $\Lambda(\vec{k}, \omega) = \rho_s(\lambda + k^2\sqrt{2D_\rho})\rho^*, \Omega(\vec{k}, \omega) = g(\vec{k}) + \epsilon\sigma^2\chi(\vec{k}, \omega)$ and $\langle \dots \rangle_+$ is the average over the noise performed only on the expected fraction of survived particles ρ_s. In order to justify the last statement we need to study how correlations around the absorbing state (no agents) behave. Specifically, the stability of Eq.(5.11) around the absorbing state (ρ_0) are given by

$$\partial_t \rho_0(\vec{z}, t) = \lambda\left(\frac{2\nu}{\alpha} - 1\right)\rho_0(\vec{z}, t) + \frac{2\lambda\gamma}{D_\phi}\rho_0(\vec{z}, t)\boldsymbol{\nabla}^{-2}\rho_0(\vec{z}, t) + D_\rho\nabla^2\rho_0(\vec{z}, t)\,, \tag{5.16}$$

which has the form of s chemotactic equation [107]. When the sign of the linear term in Eq. (5.16) is negative, the extinct states do not contribute, at linear order, to the fluctuations of correlation functions. The stability of the homogenous state is obtained by noticing that the correlation functions Eq. (5.15), in the limit of $\omega, \vec{k} \to 0$, are diverging whenever $\Omega(0, 0)^2 = \rho_s\sigma^2$. The solution of this last equation coupled to

(5.13) gives the critical value $\sigma_c = 2\sqrt{2}\lambda\nu\big/\alpha(1+\epsilon)$. For $\sigma < \sigma_c$ the homogeneous state is stable against linear perturbations and unstable otherwise. Interestingly, this relationship encodes all the scales in a compact form: λ regulates the local potential, ν and α are related to the strength and the length scale of the intermidiate range interaction and ϵ and σ encode the structure of the matrix $\hat{K}$. Following [180, 182], the small ω expansion of correlation functions hence gives

$$C(\vec{k},\omega) = \frac{\lambda + \vec{k}^2\sqrt{2D_\rho}}{\frac{\Omega(\vec{k})^2}{\rho_s} + \pi|\omega|p^+(0)\rho_s\chi^* - \sigma^2}\,. \tag{5.17}$$

The decay of correlated fluctuations are given at the critical boundary by an asymptotic expansion of Eq. (5.17) such that $C(\vec{k},\omega) \sim \frac{1}{\vec{k}^2|\omega|}$. Correlation functions thus slowly decay in time $\sim |t-t'|^{-2}$ and in space $\sim |\vec{z}-\vec{z}'|^{-(d-2)}$ for dimensions $d > 2$. There is no critical transition for $\epsilon = -1$ as it implies $\sigma_c \to \infty$, meaning that in this case the non trivial solution is always stable. Once the value of $\chi^* = \alpha(\epsilon+1)/4\lambda\nu$ is found, we go back to the linear stability and look at how perturbations grow in the local space. In particular, as $\langle\rho^*\rangle_+ = M$, we get

$$\frac{\omega}{M} = -\frac{D_\rho\vec{k}^2}{M} - \frac{2\lambda\gamma}{\alpha + D_\phi\vec{k}^2} + \frac{\epsilon(1+\epsilon)\sigma^2\alpha}{4\lambda\nu}\,. \tag{5.18}$$

This dispersion relationship admits a value of $\vec{k}_c$, such that $\omega(\vec{k})|_{\vec{k}=k_c} = 0$ and $\frac{d^2\omega}{dk^2}|_{\vec{k}=\vec{k}_c} = 0$, signaling a pattern instability for a certain combination of the parameters. Whenever $-1 \leq \epsilon \leq 0$, the homogeneous stochastic state is stable against any perturbation in the local scale $\vec{z}$ as the dispersion relationship is always non-positive for any value of $\vec{k}$. This results shows that a small asymmetry in the infinite range scale of interaction is able to destroy local order that may arise from the other scales. Moreover, Eq. (5.18) has a similar form as the May bound [171] for which only the categorical variable was considered. wiith the addition of a metric components that the May bound is lacking. In Fig. 3, we show the resulting pattern and distribution of individuals for two different values of σ above and beyond the critical line. When $\sigma > \sigma_c$ we observe multi-modality where one mode is centered around zero (extinction) and the higher modes are related instead to pattern instabilities. When $\sigma < \sigma_c$ higher modes disappear, but the effect of giant density fluctuations make the spatial pattern still non trivial and typical of birth-death like processes [183]. Giant density fluctuations arise even in the case of a spatially stable solution due to the effect of multiplicative noise [184]. Indeed, an expansion of Eq. (5.15), even far from the critical point, shows fat tails as correlations decay asympotically as $1/\vec{k}^2$.

5.3. Summary and discussion

The previous chapters raised the questions of how to properly incorporate multiple interactions scales in a stochastic system. In this chapter we addressed this questions starting with a general description which can later be applied to specific cases. In Sec. 5.1 we introduced the work of May on the study of stability of complex systems, composed of many interacting components. The May stability criterion is an powerful tool for systems interacting on a single scale, but doesn't incorporate multiple scales of interaction. In Sec. 5.2, starting from a microscopic description of a process interacting on multiple scales and upon identifying local and global components, we were able to write down a field theory for such processes. In Sec. 5.2.2 we analyse the stability of such systems, identifying the main parameters that drive different regions and the role of the combination of scales. In particular, the phase diagram showed that in addition to the May stability criterion, a new phase arises when global scale interactions are sufficiently high compared to the local scales. In this phase, different local patterns arise for the multiple components, namely a pattern instability in the local space. Interestingly, the phase where these patterns do not emerge is not trivial as it is governed by giant density fluctuations driven by the different sources of noise.

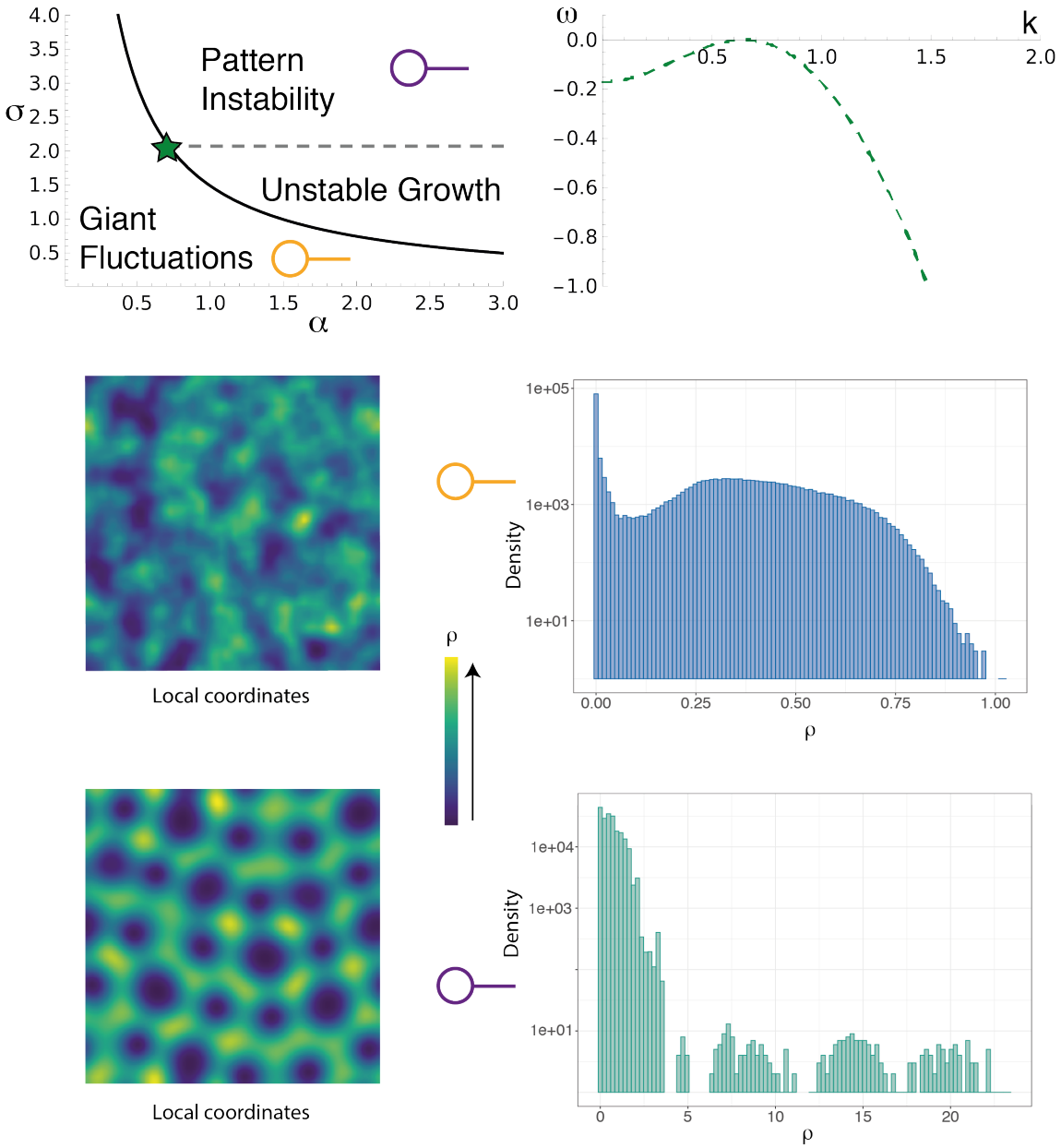


Figure 5.3.: (**A**) The phase diagram of Eq. (5.11) divides the phase space in three different regions depending on the value of the parameter α (strength of local interactions) and σ (strength of non-local interactions). (**B**) Numerical simulation of the deterministic part of Eq. (5.8). (Top) Giant density fluctuations arise even when $\sigma < \sigma_c$. The distribution of particles densities ρ_i is bimodal and exhibit giant fluctuations. (Bottom) When $\sigma > \sigma_c$, there is a pattern instability and multi-modality of the distribution of particle densities. All the simulations are performed in two dimensions using the Euler-Mayorana alogrithm and finite central difference with integration steps $dt = 10^{-3}, dx = 1$. The other parameters are $L = 50$ (number of categorical variables) and 64 lattice sizes per dimension.

6. Conclusions and future perspectives

We said goodbye before we said hello

Richard Wright, Pink Floyd

Biological systems rely on the interactions between several layers spanning over multiple scales. Understanding how different layers are tightly regulated and how interactions lead to collective behaviour is thus key to understand biological functions in living organisms. Nowadays, due to technological breakthroughs in single-cell sequencing, we are able to profile detailed molecular measurements of many of these layers with unprecedented detail. In particular, via single-cell multi-omics technologies, we are able to profile chemical modifications of the DNA, the expression of genes and the structure of chromatin at the same time for individual cells. However, a direct inference of biological functions from direct sequencing measurements remain not well understood.

In this thesis, by applying and developing theories in non equilbrium statistical physics we provided a rigorous framework to infer collective process in biological systems from detailed molecular sequencing measurements, thus overcoming limitations of single-cell technologies. Within our framework, we were able to fully understand key molecular processes leading to gastrulation and cellular symmetry breaking *in vivo* and *in vitro* spanning from nanometer to genomic scales.

In Chapter 2 we formalised our theoretical and conceptual framework drawing on original sequencing data of epigenetic modifications of the DNA (DNA methylation) during early embryonic development. DNA methylation marks are established during gastrulation before one of the first cell fate decision. We initially quantify statistical observables - average DNA methylation and spatial correlation between DNA methylation marks - that capture local and global changes of DNA methylation across single cell from DNA methylation sequencing data (BS-Seq). Specifically, we showed that the increase of average *de novo* DNA methylation with respect to time during early development does not depend on the specific genomic region, but the functional relationship is the same genome wide and follows a power law with an exponent of 5/2. The surprising emergence of scaling suggest that the establishment of DNA methylation marks is caused by a collective mechanism, involving genome-wide interactions and it is further

strengthened by a scale free decay of connected correlation functions, which encode the spatial arrangement of DNA methylation marks. Taken together, these findings point out that nonequilibrium physics is a natural framework to infer changes of epigenetic modifications. We then developed a theory of out of equilibrium kinetics of enzymes, which epigenetically modify the DNA encoding unknown long range and non local interactions kernel. The theory is general and can be applied to different epigenetic processes and we showed how to gain mechanistic understanding of DNA methylation. Upon mapping the theory to path-integrals typical of hard bosons problems in quantum mechanics, we were able to infer the shape of the interaction kernel between enzymes from the average DNA methylation. Upon using methods of perturbation theory and renormalization group we found that the theory predicts scale free behaviour of connected correlation functions in two different spatial regimes dominated by active or passive influence of chromatin structure (topology) on DNA methylation. In order to characterize the interplay between topological and epigenetic modifications, we developed a geometric field theoretical approach to relate theories in one dimension to a projected three dimensional space, such that we can infer structures on higher dimension from lower dimensional data and we found that condensates of few thousands of base pairs are formed via DNA methylation kinetics.

Within this framework all the theoretical predictions are in excellent agreement with several experimental sequencing data for several genomic regions (promoters, CpG islands, H3K4me1, introns and exons) of mouse embryonic stem cells *in vitro* and with stochastic numerical simulations. We then challenged our theory with *in vivo* experiments of mouse embryonic cells during gastrulation. The spatial arrangement of DNA methylation of several genomic regions follows our prediction, but gene bodies (introns and exons) DNA methylation patterns have systematic deviations from the theory, thus breaking the general mechanism that we proposed. We found that deviations from the theory are specific for genes that are going to be downregulated and we can surprisingly detect them via DNA methylation patterns two days prior their downregulation.

Within our theoretical framework, for the first time, we connected epigenetic layers of regulation at different spatial scales during early embryonic development via novel theories of nonequilibrium systems. Our framework is general and can be extended to study different layers of regulation.

In Chapter 3 we applied our theoretical framework to understand how gene body DNA methylation and physical properties influence transcription of genes as their interplay is fundamental to understand cellular behaviour. Drawing on sequencing data of mouse cells cultured in serum *in vitro*, we found that transcription is negatively correlated, in a scale free manner, to the length of the gene body. On the other hand, the average DNA methylation of gene bodies is positively correlated to its length with a stretched exponential dependency. Upon developing a theory incorporating the tran-

scription of genes by RNA polymerases and methyl binding enzymes, we found that the observed scaling relationships are observed only if the binding energy of enzymes is proportional to the three-dimensional structure of the chromatin. We then asked whether the specific spatial arrangement of DNA methylation marks, irrespective of the average, plays a role in the transcription of a gene as observed in the previous chapter. Upon introducing a memory in transcription of RNA polymerases, we predicted the experimentally observed changes of expression with respect to the spatial correlation length of DNA methylation marks.

However, gene expression is determined by other factors: for example, their interactions in gene regulatory networks (GRNs). With scRNA-Seq we can profile the expression of thousands of genes, but these measurements are influenced by batch effects and technical biases, such that even the identification of statistically relevant observables is lacking.

In Chapter 4 we developed a theory of gene expression fluctuations incorporating mRNA and protein, as well as unknown interactions between genes in order to overcome limitations of RNA-Seq experiments. Upon mapping the dynamic of fluctuations in GRNs to the ones of asymmetric bipartite spin glasses, we were able to identify a measure of similarities between cells stable against batch and technical effects, namely overlaps. Secondly, theoretical predictions highlighted the possibilities that different cell types may exhibit different scenarios where fluctuations are either long-lived and strongly correlated (glassy) or uncorrelated (paramagnetic). Drawing on RNA-Seq experiments of the mouse brain we found that progenitor cells are consistently more glassy than immature and neuroblast cells. Taken together we developed a framework which accounts for the propagation of fluctuations in GRNs and their correlation. We found that a possible biological relevance of correlated fluctuations is to make the state of GRNs stable and thus avoiding transition between different steady states. At the end of the chapter, we studied correlation between mRNA fluctuations, which are a measures of gene interactions and found that they show a global increase during cell state transitions. With simple paradigmatic models we identified intrinsic factors as the sources of this variation.

Our work thus lead to a comprehensive study of cell states and fluctuations of gene expression that are robust against technical effects of sequencing experiments. We provided a framework to identify different biological states and roles of fluctuations for cell states and how information in cells can be stored in gene expression fluctuations.

In Chapter 5 we derived a general framework to incorporate interactions for local and non-local spatial scales into a field theoretical description. In particular, we describe fields that are interacting with disorders on a categorical space and with multiple scale interactions on a metric space. Relying on the path integral formulations of the resulting stochastic dynamics of the fields, we were able to find the analytical phase

diagram in the space of the parameters. The complex system is divided into a region where there is a pattern instability in the metric space, a region dominated by giant density fluctuations but without any length scales of the patterns and a last region where interactions lead to unstable growth of the fields. We then found how the interplay between different parameters shape the phase diagram and give possible ways to incorporate different physical processes in the field theoretical framework.

In this thesis we systematically derived theories which allow us to infer emergent macroscopic behaviour from microscopic measurements, that are governed by noise and technical variability. We apply the derived theories in the context of cellular symmetry breaking, in particular studying early embryonic development. Our theories can give mechanistic and predictive understanding and, as more technologies are being developing and more data will be available, we think that extensions of this work may be used to describe other biological processes. As we showed, such processes rely on interactions between many different scales and a comprehensive understanding of the interplay between all these scales is only in its beginnings. As an example: how does the spatial organisation of cells play a role in cellular symmetry breaking and how it is coupled to epigenetic factors and gene expression? Moreover, we have shown that scaling behaviour and scale free correlations in biological systems are not necessarily connected to their criticality, but rather to a systematic understanding of their fluctuations. It would be interesting in the future to develop theories that can correctly capture scaling properties without the need of strong assumptions on the parameters. It is hard to draw a clear boundary between phenomena that we might unveil with our current theoretical, technological and statistical tools to phenomena in which we need to develop new theories, which are possibly very far to what we know. This change was made in theoretical physics at the beginning of the 20th century. We will need a lot of courage to again doubt what we think we have learned, but it is likely the only way to explore complex systems. It is definitely a good time to stop and add infinitesimal modifications or more data to already fully studied theories [185], but especially, it is a good time to think about new theories which could be risky or wrong, but have the potential to bring new directions to explore.

A. Construction of Doi-Peliti path integrals

Starting from Eq. (1.13) we can rewrite it as

$$\langle O(\mathbf{D},t)\rangle = \langle 0| \prod_i e^{a_i} O(\mathbf{D}) \, |P(t)\rangle \ , \tag{A.1}$$

where we introduced a coherent state basis $\langle 0| \, e^a$, which has the property to be the left eigenstate of the creation operator $a^\dagger$,

$$\langle 0| \, e^a a^\dagger = \sum_{n=1}^{\infty} \frac{\langle 0|}{n!} a^n a^\dagger = \langle 0| \, e^a. \tag{A.2}$$

Within this basis, it is possible to write the field theory associated with the stochastic process after introducing the identity

$$1 = \int d\phi d\hat{\phi} e^{-\hat{\phi}\phi} e^{\phi a^\dagger} \, |0\rangle \, \langle 0| \, e^{\hat{\phi} a} \, . \tag{A.3}$$

In order to derive the field theoretical representation of the master equation we need to understand how the delta terms in H act on the coherent state basis. We can first formally write a solution of the master equation Eq. (1.12) as

$$|P(t)\rangle = e^{-Ht} |P(0)\rangle \, , \tag{A.4}$$

where $|P(0)\rangle$ is the initial state (i.e. the probability distribution of enzymes binding profiles at time $t = 0$). The exponential is expanded for small Δt as

$$e^{-Ht} = (1 - \Delta t H)^{\frac{t}{\Delta t}} = (1 - \Delta t H) \cdot (1 - \Delta t H) \cdot \ldots \, . \tag{A.5}$$

Upon inserting the identity (A.3) in the coherent state basis between every factor on the right hand side of Eq. (A.4), the solution at any time t_1 can be written as

$$\begin{aligned} |P(t_1)\rangle = \int \prod_i & d\phi_i(t_1 + \Delta t) d\hat{\phi}_i(t_1 + \Delta t) d\phi(t_1) d\hat{\phi}_i(t_1) e^{-\hat{\phi}_i(t_1)\phi(t_1)_i} \\ & e^{-\hat{\phi}_i(t_1+\Delta t)\phi(t_1+\Delta t)_i} e^{\phi_i(t_1+\Delta t) a^\dagger} \\ & |0\rangle \, \langle 0| \, e^{\hat{\phi}_i(t_1+\Delta t) a} \, (1 - \Delta t H) \, e^{\phi_i(t_1) a^\dagger} \, |0\rangle \, \langle 0| \, e^{\hat{\phi}_i(t_1) a} \, . \end{aligned} \tag{A.6}$$

In this equation we have to evaluate quantities in the coherent state basis between the bra and the ket,

$$\begin{aligned}\langle 0| e^{\hat{\phi}_i(t_1+\Delta t)a} \left(1-\Delta t H\right) e^{\phi_i(t_1)a^\dagger} |0\rangle &= e^{\hat{\phi}_i(t_1+\Delta t)} e^{\phi_i(t_1)} - \Delta t \langle 0| e^{\hat{\phi}_i(t_1+\Delta t)a} \left(H\right) e^{\phi_i(t_1)a^\dagger} |0\rangle \\ &\approx e^{\hat{\phi}_i(t_1+\Delta t)} e^{\phi_i(t_1)} e^{-\Delta t H(\hat{\phi}_i(t_1),\phi_i(t_1))}\,,\end{aligned} \tag{A.7}$$

where $H(\hat{\phi}_i, \phi_i)$ is obtained by replacing all a_i with ϕ_i and $a_i^\dagger$ with $\hat{\phi}_i$. Repeating this procedure $t/\Delta t$ times for each factor $(1-\Delta t H)$ we end up with an integral, P_1, over a product of three terms P_2, P_3, P_4. The integral is given by

$$P_1 = \int \prod_i d\hat{\phi}_i(t) d\phi_i(t) d\hat{\phi}_i(t-\Delta t) d\phi_i(t-\Delta t) \cdot \dots \cdot d\hat{\phi}_i(\Delta t) d\phi_i(\Delta t) d\hat{\phi}_i(0) d\phi_i(0) \dots\,, \tag{A.8}$$

which can be rewritten compactly as a functional integral

$$P_1 = \int \mathcal{D}[\phi]\mathcal{D}[\hat{\phi}] \dots\,. \tag{A.9}$$

P_2 is composed of a product of terms which can be rewritten by means of Riemann integration as

$$P_2 = \prod_{t_1=\Delta t}^{t} e^{\hat{\phi}(t_1+\Delta t)\phi(t_1) - \hat{\phi}(t_1)\phi(t_1)} \approx e^{-\int dt \partial_t \hat{\phi}\phi}. \tag{A.10}$$

Finally there are further $t/\Delta t$ terms coming from the Hamiltonian evaluated at each time step which are simplified as

$$P_3 = \prod_{t_1=\Delta t}^{t} e^{-\Delta t H(\hat{\phi}(t_1),\phi(t_1))} \approx e^{-\int dt H(\hat{\phi}(t),\phi(t))}\,. \tag{A.11}$$

The final factor, P_4, represents initial conditions and we refer to [75] for a discussion of this term. Putting all the terms together we arrive to Eq. (1.14).

B. Analysis of sequencing experiments

B.1. Bulk bisulphite sequencing

Whole genome bisulfite sequencing data was processed identically to [32]. Raw sequence reads were trimmed to remove both poor-quality calls and adapters using Trim Galore (v0.4.1, Cutadapt version 1.8.1, parameters: –paired) [186]. Trimmed reads were first aligned to the mouse genome in paired-end mode to be able to use overlapping parts of the reads only once while writing out unmapped singleton reads; in a second step remaining singleton reads were aligned in single-end mode. Alignments were carried out with Bismark v0.14.4 [187] with the following set of parameters: a) paired-end mode: –pbat; b) single-end mode for Read 1: –pbat; c) single-end mode for Read 2: defaults. Reads were then deduplicated with deduplicate_bismark selecting a random alignment for position that were covered more than once. CpG methylation calls were extracted from the deduplicated mapping output ignoring the first 6 bp of each read (corresponding to the 6N random priming oligos) using the Bismark methylation extractor (v0.14.4) with the following parameters: a) paired-end mode: –ignore 6 –ignore_r2 6; b) single-end mode: –ignore 6. SeqMonk version 0.32 was used to compute methylation rates and coverage in annotation genomic regions. To QC BS-Seq data, pairwise Pearson correlation coefficients were calculated using methylation levels averaged over 10kb tiles. Replicates within the same time point were on average more highly correlated than between time points (r=0.885 versus 0.866). For subsequent analyses, replicates were merged. Further statistical analysis was performed by custom scripts in R. We calculated average DNAme levels for a given set of genomic regions defined by their functional annotation and average CpG density using the "Bisulfite methylation over feature" pipeline in Seqmonk. To be able to identify the functional form of average methylation over time only feature sets that had genome-wide more than 1500 reads at a given time point are shown. Averages over genomic regions were weighted by the average number of reads per CpG. To collapse the time series onto a scaling form, we made a scaling ansatz of the form $m = a + bt^{5/2}$ and determined a and b using nonlinear least squares estimate as implemented in the R function nls. With this, the rescaled time, τ, was defined as $\tau = t\, b^{2/5}$. The exponent was estimated using nonlinear least squares. To verify the robustness of the exponent in the presence of negative data

points with respect to log transformation of both axes we estimated the exponent for different values of an offset parameter, c, such that the rescaled average DNA methylation reads $\langle m \rangle = c + \tau^{5/2}$ and all values of the time course are positive. We found that under these transformations the estimation of the exponent was robust.

B.2. scNMT-Seq 2i release data

B.2.1. BS-Seq

Alignments of the single-cell bisulfite sequencing were performed using Bismark as well as subsequent CpG methylation and GpC accessibility calling. Cells with more than 10^5 reads, less than 15% CHH methylation and a mapping efficiency larger than 10% were kept for downstream analysis. Following (46) average DNA methylation in a given genomic interval was calculated as $m = \frac{p+1}{p+n+2}$, where p and n signify the number of positive or negative reads in a given genomic interval, respectively.

B.2.2. RNA-Seq

scRNA-Seq alignments were performed using Hisat 2 [188]. 226 cells with mitochondrial RNA $< 0.15\%$, > 200000 reads and > 2000 detected genes were kept for downstream analysis. Reads were log normalised using the LogNormalise function of the Seurat package version 3.2.0 with standard parameters. For dimensionality reduction, the top 1000 most highly variable genes were selected and a principal component analysis with default parameters of the Seurat package was performed. Uniform Manifold Approximation was performed on the 15 principal components with the highest variance and with a minimum distance of 0.2.

B.3. sn-m3C-seq data

Following [111] we retained cells with more than 5000 cis contacts at distances longer than 10000bp and more than 100000 covered CpGs. We tiled the genome into windows of 100kbp and, for each tile, calculated average DNAme and cis contact histograms with respect to the genomic distance. We then pooled these histograms for genomic windows of similar DNAme levels and normalized by the total number of cis contacts. While contacts are expected to be technically enriched in GC rich regions, which are typically associated with low DNAme levels, we observe an opposite effect in Fig. 3f. This suggests a biological rather than technical origin of the increasing number of cis-contacts with DNAme level.

B.4. scNMT-Seq embryo data

B.4.1. BS-Seq

Data was processed identically to [6]. Genome-wide correlation and cross-correlation functions were computed by dividing samples with respect to the stage (E4.5, E5.5, E6.5) and lineage (E7.5 Mesoderm, Endoderm, Ectoderm).

B.4.2. RNA-Seq

Cells which had a percentage of mitochondrial RNA <0.15%, nCount_RNA>1e5 and more than 2500 genes with at least one read were kept for downstream analysis. Normalisation was performed using the function LogNormalize from the Seurat package (version 3.2.9). The least and most highly expressed genes were determined based on their log-normalised expression value. Differentially expressed genes between pairs of stages were determined using a t-test. To ensure that the statistical sample size was identical for each comparison the top 2000 genes based on p-value were selected for further analysis. This number was chosen to achieve a balance between the biological significance of selected genes and the sample size necessary to calculate correlation functions. Correlation functions for a given set of genes were computed by first obtaining the coordinates of the corresponding gene bodies using biomart 2.44.1, then computing correlation functions for each gene and finally averaging over all the genes in a given stage or lineage. To compare predictions made by our method to the embryo data we used stochastic simulations of the inferred model taking into account the genomic distribution of CpG sites in the mouse genome. Differences between theory and experiment were rescaled by the experimental standard error of the correlation function at a given genomic distance. Differences were considered significant if $p < 0.05$ using a t-test.

C. Path integral representation of *de novo* DNA methylation

C.1. Connected correlation functions

C.1.1. Short tail

By considering a perturbation $h(s,t)$ around the mean field solution $\phi_0(t)$, $\phi(s,t) = \phi_0(t) + h(s,t)$, Eq. (2.41) can be expressed to first order as

$$\begin{aligned}\partial_t\phi_0(t) + \partial_t h(s,t) = e^{-\phi_0(t)} \int_0^s dy\, \phi_0(t)|s-y|^{-\lambda} \left[1 - \int_{z=0}^{s-y} dz\, h(z)\right] + \\ e^{-\phi_0(t)} \int_0^s dy\, h(y)|s-y|^{-\lambda} \left[1 - \int_{z=0}^{s-y} dz\, h(z)\right] + \text{h.o.}\,.\end{aligned} \tag{C.1}$$

The first two terms on the right hand side cancel with the first one on the left hand side, which is the dynamical mean field solution. Taken together, we find

$$\partial_t h(s,t) = e^{-\phi_0(t)} \int_0^s dy h(y)|s-y|^{-\lambda} \left[1 - \int_{z=0}^{s-y} dz\, h(z)\right] + \text{h.o.}\,. \tag{C.2}$$

After a change of variables, $w = z + y$, we obtain

$$\begin{aligned}\partial_t h(s,t) =& e^{-\phi_0(t)} \int_0^s dy\, h(y)|s-y|^{-\lambda} \\ &- e^{-\phi_0(t)} \int_0^s dy \int_y^s dw\, h(y)|s-y|^{-\lambda} h(w-y) + \xi(s,t)\,.\end{aligned} \tag{C.3}$$

In this expression we recognize a convolution of a fractional integral of a function and the function itself and the noise, $\xi(s,t)$ for the perturbation $h(s,t)$ has both conservative and non-conservative contributions, $\langle \xi(s,t)\xi(s',t')\rangle = \delta(t-t')(2\Gamma_{NC} - 2\Gamma_C\partial_s^2)\delta(s-s')$. Γ_C and Γ_{NC} are the noise strengths for conservative and non conservative noise, respectively. As a side remark, in the case of only conservative noise the non-local and non-linear term becomes relevant under renormalization below a critical dimension $d_c = 2(2-\lambda)$, while in case of only non conservative noise the critical dimension is $d_c = 2(3-\lambda)$. The non linear term is identified as a fractional integral,

$$I^\alpha f = \frac{1}{\Gamma(\alpha)} \int (x-y)^{\alpha-1} f(y) \tag{C.4}$$

where $\Gamma(\alpha)$ is the gamma function. Upon identifying $\alpha - 1 = -\lambda$ in Fourier space the value of this integral scales as $q^{\lambda-1}$.

In order to regularise the theory we introduce an auxiliary process. The lowest order spatial derivative consistent with the symmetries of the theory is $\partial_s^2\phi$. Taken together, taking into account interactions with the right nearest bound site we obtain in Fourier space

$$\partial_t h(q,t) = \left(e^{-\phi_0(t)}q^{\lambda-1} - q^2\right) h(q,t) - q^{\lambda-1}e^{-\phi_0(t)}h(q,t)^2 + \xi(q,t)\,. \tag{C.5}$$

From the dynamical mean field solution of the first moment, Eq. (2.20), we know that $e^{-\phi_0(t)} = e^{\left[-t^{1/(1-\lambda)}\right]}$. In the frequency domain we obtain for small times

$$i\omega h(q,\omega) = (q^{\lambda-1} - q^2)h(q,\omega) - q^{\lambda-1}h(q,\omega)^2 + \xi(q,\omega)\,. \tag{C.6}$$

As a side remark, the inverse free propagator is $G_0^{-1} = i\omega + q^2 - q^{\lambda-1}$, which is defined based on the linear part of Eq. (C.6) as

$$\left(i\omega + q^2 - q^{\lambda-1}\right) h(q,\omega) = \xi(q,\omega)\,, \tag{C.7}$$

and the correlator can be written as

$$C_0 = \left(2\Gamma_{NC} + 2\Gamma_C q^2\right)|G_0|^2\,. \tag{C.8}$$

Taken together, the general solution of Eq. (C.6) is given by

$$h(q,\omega) = \frac{q^{1-\lambda}}{2J}\left(-q^2 + q^{1-\lambda} + \sqrt{4q^{1-\lambda}\xi + (q^2 - q^{1-\lambda}) + i\omega} - i\omega\right)\,. \tag{C.9}$$

C.1.2. Long tail

Before proceeding with renormalization, we have to take into account other possible non-linearities. We started by considering the linear order in ϕ and we obtained Eq. (2.44) involving quadratic terms, ϕ^2, due to the expansion of the integral. We must therefore come back to the field theory Eq. (2.17) and keep quadratic terms as well. The only quadratic term in the field theory is

$$-\phi(s)\hat{\phi}(s)\left(1 - \hat{\phi}(s)\right)\left[\int_0^s \mathrm{d}y \frac{\hat{\phi}(s-y)\phi(s-y)}{y^\lambda} e^{-\int_{z=0}^{y}\mathrm{d}z\hat{\phi}(s-z)\phi(s-z)}\right]\,. \tag{C.10}$$

After functional minimization it with an opposite sign, such that both terms cancel out. This is not surprising, because the symmetry of the system we are studying would

not allow a term that breaks the space reversal symmetry $s \to -s$. Taken together, we find

$$\partial_t h(s) = \partial_s^2 h(s) + \int_0^s dy h(y)|s-y|^{-\lambda} + \frac{1}{2}\int_0^s dy h(y)|s-y|^{2-\lambda}\partial_s h(s) + \xi(s,t)\,. \quad \text{(C.11)}$$

Considering both right and left nearest neighbour interactions, the advective terms cancel out. Including the next highest order term we obtain

$$\partial_t h(s) = \partial_s^2 h(s) + \int_0^x dy h(y)|s-y|^{-\lambda} + \frac{1}{2}\partial_s^2 h(s)\int_0^s dy h(y)|s-y|^{2-\lambda} + \xi(s)\,. \quad \text{(C.12)}$$

We can generalize to any spatial dimension by considering the previous equation with a spatial coordinate in vector form, $\mathbf{s}$. In Fourier space the previous equation can then be written in compact form,

$$G_0(\mathbf{q},\omega)^{-1}h(\mathbf{q},\omega) = \xi(\mathbf{q},\omega) - \nu\int_{\mathbf{k},\omega'} W(\mathbf{q},\mathbf{k})h(\mathbf{k},\omega)h(\mathbf{q}-\mathbf{k},\omega'-\omega)\,, \quad \text{(C.13)}$$

where $h(\mathbf{q},\omega) = \int d\mathbf{s}\int dt h\,(\mathbf{s},t)\,e^{i\mathbf{q}s}e^{i\omega t}$, $G_0^{-1} = \left(i\omega + D_0\mathbf{q}^2 + J|\mathbf{q}|^{-\lambda}\right)$ and,

$$W(\mathbf{q},\mathbf{k}) = \frac{1}{2}\left[\frac{\mathbf{k}(\mathbf{q}-\mathbf{k})}{|\mathbf{k}-\mathbf{q}|^{3-\lambda}} + \frac{(\mathbf{q}-\mathbf{k})\mathbf{k}}{|\mathbf{q}|^{3-\lambda}}\right]. \quad \text{(C.14)}$$

We reintroduced the dimensional parameters from the adimensional Eq. (C.13) as we are interested in how they scale under renormalization.

C.2. Geometrical field theory

The master equation (2.55) can be rewritten in terms of lowering and raising operators such that

$$\partial_t P(\boldsymbol{\rho},t) = \sum_i \left[L_i^{-r\rho_{i-1}}L_{i-1}^{r\rho_{i-1}} + L_i^{-r\rho_{i+1}}L_{i+1}^{r\rho_{i+1}} - 2\right]W(\rho_i)P(\boldsymbol{\rho},t)\,, \quad \text{(C.15)}$$

where the operators $L^{\pm\rho_i}$ act on functions on their right as $L^{\pm\rho_i}f(m_i) = f(m_i \pm \rho_i)$. The operator $L^{-\rho_i}$ can be identically written as $L_i^{-r\rho_{i-1}} = e^{-r\rho_{i-1}\partial\rho_i}$ and expanded as $L_i^{-r\rho_{i-1}} = 1 - r\rho_{i-1}\partial_{\rho_i} + \mathcal{O}(r^2)$. We now proceed with a linear noise approximation of the master equation, where the observable ρ_i is split into two components

$$\rho_i = N\phi_i + \sqrt{N}\eta_i\,, \quad \text{(C.16)}$$

where N is the system size. Written in this form the operators become $L_i^{-r\rho_{i-1}} = 1 - rN^{-1/2}\rho_{i-1}\partial_{\eta_i} + \mathcal{O}(N^{-1})$. The terms on the right hand side of the master equation

are then to lowest order in N, $\mathcal{O}(1/\sqrt{N})$,

$$\sum_i r\left[\phi_{i-1}(-\partial_{\eta_i}+\partial_{\eta_{i-1}})+\phi_{i+1}(-\partial_{\eta_i}+\partial_{\eta_{i+1}})\right]W(\phi)\Pi(\eta)\,. \tag{C.17}$$

As fluctuations in ρ are given by fluctuations in η we have $dP(\rho) = d\Pi(\eta)$, i.e. the probability distribution of the entire process is solely determined by its stochastic part. We now take a continuum approximation such that $\partial_{\eta_{i+1}} \approx \partial_{\eta(x)} \pm a_0\partial_x\partial_{\eta(x)}$ and the same for ϕ_{i+1}, where a_0 is the lattice spacing. After these steps we obtain for the right hand side of the master equation,

$$\int dx\, r a_0^3 \partial_x\phi(x,t)\partial_x\left[W(\phi(x))\frac{\delta\Pi(\eta)}{\delta\eta}\right]\quad. \tag{C.18}$$

By applying the same steps to the left hand side of the master equation to the same order in N we obtain

$$\partial_t P(\rho,t) = \partial_t\Pi - \sqrt{N}a_0\int dx\,\frac{d\phi(x,t)}{dt}\frac{\delta\Pi}{\delta\eta}\,. \tag{C.19}$$

Upon integrating by parts the right hand side and taking equal orders on both side of the expanded master equation, we obtain the partial differential equation (2.56) describing the time evolution of the density field $\phi(x,t)$.

D. Oscillations in DNA methylation

D.1. Discrete phase expansion

In this section we give details for the Van Kampen expansion of the master equation (2.69). After rewriting the phase of the clock as in Eq. (2.70), the l.h.s of the master equation becomes

$$\sum_i \frac{dP(\boldsymbol{\phi},t)}{d\phi_i} = \frac{d\Pi}{dt} - \sum_i \left[\Omega^{1/2}\frac{d\phi_i}{dt}\frac{d\Pi}{d\xi_i}\right]. \tag{D.1}$$

We rewrite the term in the r.h.s of the master equation as

$$\sum_i \left[(\omega_i + k_i(\boldsymbol{\phi},\phi_i-1)\right]P(\boldsymbol{\phi},\phi_i-1) = \sum_i E_i^{-1}\left[(\omega_i + k_i(\boldsymbol{\phi},\phi_i)\right]P(\boldsymbol{\phi},\phi_i), \tag{D.2}$$

where the introduced operator $E_i^{\pm}$ acting on everything to the right as

$$E_i^{\pm}G(\boldsymbol{\phi}) = G(\boldsymbol{\phi},\phi_i \pm 1), \tag{D.3}$$

where $G(\boldsymbol{\phi})$ is a general function of the phase. The operators $E_i^{\pm}$ in the system size expansion are approximated to highest order in Ω as

$$E_i^{\pm} \sim 1 \pm \Omega^{1/2}\frac{\partial}{\partial\xi_i} + \frac{1}{2}\Omega^{-1/2}\frac{\partial^2}{\partial\xi_i^2}. \tag{D.4}$$

Upon using the expansion of the operators and to higher order terms in Ω we get

$$\frac{d\Pi}{dt} - \sum_i \left[\Omega^{1/2}\frac{d\Phi_i}{dt}\frac{\partial\Pi}{\partial\xi_i}\right] = \sum_i \left\{-w_i\left[\Omega^{1/2}\frac{\partial}{\partial\xi_i} - \frac{1}{2}\frac{\partial^2}{\partial\xi_i^2}\right]\Pi - \Omega^{1/2}\frac{\partial}{\partial\xi_i}\left[k(\Phi_i) + \Omega^{-1/2}\frac{\partial k}{\partial\Phi_i}\xi_i\right]\Pi\right\}. \tag{D.5}$$

Collecting terms in power of $\Omega^{1/2}$, and using the chain rule we get

$$\frac{d\Phi_i}{dt} = w_i^0 + f_1(\Phi_i)\sum_{k=1}^{N}\frac{J_1 e^{-\rho_2|k-i|}}{|k-i|^{\lambda}}f_2(\Phi_k), \tag{D.6}$$

which is the mean field equation and where we simply made $k(\Phi)$ explicit. The next order Ω^0 encodes the dynamics of fluctuations

$$\frac{\partial\Pi}{\partial t} = \sum_i \left[(w_i^0 + k(\Phi))\frac{\partial^2\Pi}{\partial\xi_i^2} - w_L^0\frac{\partial k(\Phi)}{\partial\Phi_i}\frac{\partial(\xi_i\Pi_i)}{\partial\xi_i}\right]. \tag{D.7}$$

For small values of $k(\Phi)$ the noise is dominated by ω_i and it is a gaussian white noise arriving to Eq. (2.71).

D.2. Derivation of the Fokker Planck equation

In Eq. (2.73) we introduced the moment generating function for ω_i with respect to both stochastic noise and intrinsic noise given by the possible non-delta distribution of intrinsic frequencies. In Ito convention the dynamics of an arbitrary function $F(.)$ of a certain stochastic process ϕ_i ($i = 1, ..., N$) and with noise amplitude $\sqrt{2D_i}$ is

$$\partial_t F(\boldsymbol{\phi}) = \sum_j \left[\partial_{\phi_j} F(\phi) \partial_t \phi_j + \sum_k \frac{\partial^2 F(\boldsymbol{\phi})}{\partial \phi_j \partial \phi_k} \sqrt{D_j D_k} \right] . \tag{D.8}$$

As the previous equation holds true or every function F we compute now $\partial_t \langle e^{ik\phi_j} e^{iqj} \rangle$. We then substitute in Eq. (D.8) the stochastic process which trajectory is given by Eq. (2.71) hence obtaining (after using standard property of stochastic calculus)

$$\frac{\partial \langle e^{ik\phi_j} e^{iqj} \rangle}{\partial t} = ike^{ik\phi_j} e^{iqj} G(\phi_j) - k^2 e^{ik\phi_j} e^{iqj} D_j \, . \tag{D.9}$$

We then multiply the previous equation by ω_j^m, sum over all $j = 1, ...N$, divide by N and average over the frequency distribution,

$$\frac{1}{N} \sum_j \overline{\frac{\partial \langle e^{ik\phi_j} e^{iqj} \rangle \omega_j^m}{\partial t}} = \frac{1}{N} \sum_j \overline{\omega_j^m \left[ike^{ik\phi_j} e^{iqj} G(\phi_j) - k^2 e^{ik\phi_j} e^{iqj} D_j \right]}, \tag{D.10}$$

With $G(\phi_j) = \omega_j + f_1(\phi_j) \sum_{w=1}^{N} \frac{Je^{-\rho_2 |w-j|}}{|w-j|^\lambda} f_2(\phi_w)$. The l.h.s of the previous equation is simply $\partial_t H_{k,q}^m$ as defined in Eq. (2.73). The r.h.s is slightly more complicated. First we introduce the Fourier representation of f_1, f_2 as

$$\begin{aligned} f_1(\phi_j) &= \sum_{n=-\infty}^{\infty} a_n e^{in\phi_j} \\ f_2(\phi_w) &= \sum_{l=-\infty}^{\infty} b_l e^{il\phi_w} \end{aligned} \tag{D.11}$$

Being $\frac{e^{-\rho_2 |w-j|}}{|w-j|^\lambda}$ just a function of the difference we can define its Fourier transform as

$$\frac{e^{-\rho_2 |w-j|}}{|w-j|^\lambda} = \sum_{s=-\infty}^{\infty} r_s e^{is(w-j)} \, . \tag{D.12}$$

The first term on the r.h.s of Eq. (D.10) is given by

$$\frac{J}{N}\sum_j \sum_{n,l,s}\sum_w a_n b_l r_s ike^{i(k+n)\phi_j} e^{i(q-s)j} e^{il\phi_w} e^{isw}\omega_j^m \,. \tag{D.13}$$

Upon noticing that $\frac{1}{N}\sum_w e^{il\phi_w}e^{isw} = H^0_{l,s}$, the previous equation simplifies to

$$(ik)JN\sum_{n,l,s} a_n b_l r_s H^m_{k+n,q-s} H^0_{l,s} \,. \tag{D.14}$$

The other terms can be computed in a similar way and the resulting dynamical equation for the moments is

$$\partial_t H^m_{k,q} = (ik)JN\sum_{n,l,s} a_n b_l r_s H^m_{k+n,q-s} H^0_{l,s} + (ik)H^{m+1}_{k,q} - k^2 H^{m+1}_{k,q} \,. \tag{D.15}$$

In order to obtain a simpler expression we define the generating function

$$\chi(\theta, y, z, t) = \sum_{k=-\infty}^{\infty}\sum_{m=\infty}^{\infty}\sum_{q=-\infty}^{\infty} e^{-ik\theta}e^{-iqz}\frac{y^m}{2\pi m!}H^m_{k,q} \,, \tag{D.16}$$

and we have to show how all the terms of Eq. (D.15) can be rewritten in term of this function. We will analyse, at first, the second term on the r.h.s of (D.15), which can be rewritten as (upon summing over m, k, q and multiplying by $\frac{y^m}{2\pi m!}$)

$$\sum_{k,m,q}(ik)e^{-ik\theta}e^{-iqz}\frac{y^m}{2\pi m!}H^{m+1}_{k,q} \,, \tag{D.17}$$

and further simplifying as

$$-\frac{\partial}{\partial\theta}\sum_{k,m,q} e^{-ik\theta}e^{-iqz}\frac{\partial}{\partial y}\frac{y^{m+1}}{2\pi m!}H^{m+1}_{k,q} \,. \tag{D.18}$$

In term of the function χ the previous equation is simply

$$-\frac{\partial^2}{\partial\theta\partial y}\chi(\theta, y, z, t) \,. \tag{D.19}$$

Following the same procedure, the third term on the r.h.s of Eq. (D.15) is rewritten as

$$\frac{1}{2}\frac{\partial}{\partial y}\frac{\partial}{\partial\theta}D(y)\frac{\partial}{\partial\theta}\chi(\theta, y, z, t) \,. \tag{D.20}$$

A little more cumbersome is the first term. Initially, we can rewrite it as

$$-JN\frac{\partial}{\partial\theta}\left[\sum_{n,l,s} a_n b_l r_s e^{inx}e^{isz}\chi(\theta, y, z, t)H^0_{l,s}\right] \,, \tag{D.21}$$

where we multiplied and divide by $e^{inx}e^{isz}$ so that $H_{k,q} \to H_{k+n,q-s}$. We now define a term

$$\nu(\theta, y=0, z, t) = JN\left[\sum_{n,l,s} a_n b_l r_s e^{inx} e^{isz} H^0_{l,s}\right] \tag{D.22}$$

and we can read Eq. (D.11) inversely and go back to the space of function defined on ϕ (θ in this notation)

$$\nu(\theta, y=0, z, t) = JN\left[f_1(\theta)\sum_{l,s} b_l r_s e^{isz} H^0_{l,s}\right]. \tag{D.23}$$

We then simply explicit all the other terms in the same way as

$$\nu = J\left[f_1(\theta)\int d\hat{\theta}\int d\hat{z}\frac{e^{-\rho_2\hat{z}}}{|\hat{z}|^\lambda} f_2(\hat{\theta})\chi(\hat{\theta}, y=0, z-\hat{z}, t)\right] \; z \in (0,1)\,. \tag{D.24}$$

All together we arrive to Eq. (2.75).

E. Spin glass theories of GRN

E.1. System size expansion

Upon inserting the system size expansion into the master equation (4.1) and by equating order of Ω in the r.h.s and l.h.s to order Ω^1 we find the mean field dynamics of the fields

$$\begin{aligned}\frac{\partial \phi_i}{\partial t} &= g_i \psi_i - d_i \phi_i \,, \\ \frac{\partial \psi_i}{\partial t} &= p_i - \gamma_i \psi_i + \sum_{j \in a(i)} f^a_{ji}\left(\phi_j\right) + \sum_{j \in r(i)} f^r_{ji}\left(\phi_j\right) \,.\end{aligned} \tag{E.1}$$

Systems of equation like (E.1) may exhibit multiple attractors which are defined by mRNA and protein concentrations such that $\partial_t \phi = \partial_t \psi_i = 0 \forall i = 1, \ldots, N$. In particular, close to one of such attractors, the Fokker Planck equation describing protein and mRNA fluctuations takes the form (equating the next lowest order terms in Ω)

$$\frac{\partial P\left(\xi, \eta\right)}{\partial t} = \sum_{i=1}^{N} \left[D^n_i \frac{\partial^2}{\partial \xi_i^2} + D^m_i \frac{\partial^2}{\partial \eta_i^2} + \frac{\partial}{\partial \xi_i} v^n_i + \frac{\partial}{\partial \eta_i} v^m_i \right] P\left(\xi, \eta\right) \,, \tag{E.2}$$

with $D^n_i = d_i \phi^*_i + g_i \psi^*_i$, $D^m_i = \gamma_i \psi^*_i + p + \sum_{j \in e(i)} f_{ji}(\phi^*_j)$,$v^n_i = d_i \xi_i - \gamma_i \eta_i$, $v^m_i = -\sum_{j \in e(i)} f'(\phi^*_j)\xi_j + \gamma_i \eta_i$. The Fokker Planck equation (E.2) is equivalent to the dynamics of the coupled Langevin equations

$$\begin{aligned}\partial_t \xi_i &= g_i \eta_i - d_i \xi + \sqrt{D^n} W^\xi \,, \\ \partial_t \eta_i &= -\gamma_i \eta_i + \sum_{j \in e(i)} f'_{j,i}(\phi^*_j)\xi_j + \sqrt{D^m} W^\eta \,,\end{aligned} \tag{E.3}$$

where W^ξ and W^η are unitary uncorrelated Gaussian white noises. In the limit of fast degradation $\gamma_i \gg 1$ the two coupled equations reduce to a single equation for the protein fluctuations

$$\partial_t \xi_i = -d_i \xi_i + b_i \sum_{j \in e(i)} f'_{j,i}(\phi^*_j)\xi_j/\gamma_j + \sqrt{D^n} W^\xi \,. \tag{E.4}$$

This fast degradation limit is valid in bacteria [45], but does not hold in eukaryotic systems. As the Fokker Planck equation (E.2) cannot be solved analytically, we look

for approximate solutions of the form: $P = P(\xi)P(\eta)$. The approximate stationary solution follows $P = e^{-H}/Z$, where the "Hamiltonian" H is

$$H = \sum_{i=1}^{N_\xi} \frac{1}{D_i^n}\left(\frac{d_i\xi_i^2}{2} - g_i\xi_i\eta_i\right) + \sum_{j=1}^{N_\eta}\frac{\gamma\eta_j^2}{2D_i^m} + \sum_{ij}\frac{1}{D_i^m}J_{ij}\xi_i\eta_j\,. \tag{E.5}$$

The previous solution is approximate as the term in $\xi_i\eta_i$ would give rise to a contribution in the Langevin dynamics of ξ_i and η_j which are not originally present in (E.3) ,

$$\begin{aligned}\partial_t\xi_i &= g_i\eta_i - d_i\xi + \sum_{j\in e(i)}\frac{f'_{j,i}(\phi_j^*)D_i^n}{D_i^m}\eta_j + \sqrt{D^n}W^\xi \\ \partial_t\eta_i &= -\gamma_i\eta_i + \sum_{j\in e(i)} f'_{j,i}(\phi_j^*)\xi_j + g_i\frac{D_i^m}{D_i^n}\xi + \sqrt{D^m}W^\eta\,.\end{aligned} \tag{E.6}$$

If fluctuations in mRNA are dominant then $D_i^m \gg D_i^n$ - this as to be expected as mRNA has more sources of fluctuations compared to protein - and the first equation is reduced to the exact Langevin equation, whilst the second one has still a term in ξ_i which is not originally present. If $D_i^m \gg D_i^n$ for high mRNA abundance this term scales as, $\frac{D_i^m}{D_i^n} \sim \frac{\gamma_i}{g_i}$, which makes the term in the second equation scale as $\gamma_i\xi_i$. We can then rescale the term in i inside $\sum_{j\in e(i)} f'(\phi_j^*)\xi_j$ (corresponding to self-activation or repression) and the dynamics is fully described by the Hamiltonian (E.5). When J_{ij} are Gaussian distributed, the Hamiltonian is the one of a bipartite spin-glass. We can then use techniques introduce in Sec. 1.3.3 to study how the probability distribution of fluctuations behave with respect to the parameters of the system.

E.2. Derivation of the bipartite spin glass

The Hamiltonian (4.7) can be rewritten as

$$H = \sum_{i=1}^{N} V^\xi(\xi_i) + \sum_{j=1}^{N} V^\eta(\eta_j) + \sum_{ij} J_{ij}\xi_i\eta_j + \sum_{i=1}^{N}\tilde{K}_{\xi,\eta}\xi_i\eta_i\,, \tag{E.7}$$

with $J_{ij} = N(0, W)$ and $V^\xi(\xi_i) = \tilde{K}_\xi\,\xi^2/2$ and similarly for V^η. We define the quantity $W = \sigma\big/N^{1/2}$. This ensures the extensivity of the Hamiltonian in the large N limit. We perform the quenched average

$$[Z]^n = \int D\left[J_{ij}\right] D\left[\xi\right] D\left[\eta\right] P(J_{ij})\exp(-H_r)\,, \tag{E.8}$$

where H_r is the replicated Hamiltonian. Separating diagonal and off-diagonal terms

$$\int \prod_{ij} dJ_{ij} e^{-\frac{J_{ij}^2}{2W^2}} e^{-\sum_a J_{ij}\xi_i^a \eta_j^a}\,, \tag{E.9}$$

the result of the Gaussian integration over the couplings J_{ij}, leads to

$$[Z^n] = Tr_n \exp\left[\frac{W^2}{2}\sum_{a,b,i,j}\xi_i^a\xi_i^b\eta_j^a\eta_j^b + \sum_{i,a}V^\xi(\xi_i^a) + \sum_{j,a}V^\eta(\eta_j^a) + \sum_{i,a}\tilde{K}_{\xi,\eta}\xi_i^a\eta_i^a\right]\,, \tag{E.10}$$

where Tr in case of non-binary variables is $\int\prod_{c,d} d\xi_c d\xi_d$.
We further introduce the integral transformation

$$e^{\frac{BC}{\sqrt{2}\omega}} \sim \int dx d\tilde{x} dy d\tilde{y} e^{-\omega(x^2-\sqrt{2}xy+y^2+\frac{1}{2}\tilde{x}^2+\frac{1}{2}\tilde{y}^2)} e^{B(x+i\tilde{x})+C(y+i\tilde{y})} \tag{E.11}$$

in order to simplify the quartic term in Eq.(E.10). Using this transformation to the Hamiltonian with the following variables

$$\omega = N\frac{\sigma^2}{2\sqrt{2}}, B = \frac{\sigma^2}{2}\sum_i \xi_i^a\xi_i^b, C = \frac{\sigma^2}{2}\sum_j \eta_j^a\eta_j^b\,, \tag{E.12}$$

the quenched average partition function is

$$\begin{aligned}
[Z^n] &= \int\prod_{a,b} dx_{ab}d\tilde{x}_{ab}dy_{ab}d\tilde{y}_{ab}e^{-NnF_n}\,,\\
nF_n &= \frac{\sigma^2}{2}\sum_{a\neq b}\left(\frac{x_{ab}^2}{\sqrt{2}}+\frac{y_{ab}^2}{\sqrt{2}}+\frac{\tilde{x}_{ab}^2}{2\sqrt{2}}+\frac{\tilde{y}_{ab}^2}{2\sqrt{2}}-x_{ab}y_{ab}\right)\\
&\quad+\frac{\sigma^2}{2}\sum_a\left(\frac{x_{aa}^2}{\sqrt{2}}+\frac{y_{aa}^2}{\sqrt{2}}+\frac{\tilde{x}_{aa}^2}{2\sqrt{2}}+\frac{\tilde{y}_{aa}^2}{2\sqrt{2}}-x_{aa}y_{aa}\right)\\
&\quad-\log Tr_{n,\xi,\eta}\Psi_\xi\Psi_\eta\Psi_{\eta,\xi}\,,
\end{aligned} \tag{E.13}$$

where

$$\begin{aligned}
\Psi_\xi &= \exp\left[\sum_{a\neq b}\frac{\sigma^2}{2}\left(x_{ab}+i\tilde{x}_{ab}\right)\xi^a\xi^b + \sum_{a,i}\frac{\sigma^2}{2}\left(x_{aa}+i\tilde{x}_{aa}\right)\xi^a\xi^a + \sum_{a,i}V^\xi(\xi^a)+\right]\\
\Psi_\eta &= \exp\left[\sum_{a\neq b}\frac{\sigma^2}{2}\left(y_{ab}+i\tilde{y}_{ab}\right)\eta^a\eta^b + \sum_{a,i}\frac{\sigma^2}{2}\left(y_{aa}+i\tilde{y}_{aa}\right)\eta^a\eta^a + \sum_{a,i}V^\eta(\eta^a)\right]\\
\Psi_{\eta,\xi} &= \exp\left[\sum_a \tilde{K}_{\xi,\eta}\xi^a\eta^a\right]\,.
\end{aligned} \tag{E.14}$$

As the the exponential term carries a factor N we can evaluate the trace and the saddle point equations in the limit $N\to\infty$. To do so, we perform a change of variables

(we already summed over i neglecting variability in local parameters)

$$\begin{aligned} Q^{\xi}_{ab} &= (x_{ab} + i\tilde{x}_{ab})\ Q^{\eta}_{ab} = (y_{ab} + i\tilde{y}_{ab})\ , \\ \hat{Q}^{\xi}_{ab} &= (x_{ab} - i\tilde{x}_{ab})\ \hat{Q}^{\eta}_{ab} = (y_{ab} - i\tilde{y}_{ab})\ . \end{aligned} \tag{E.15}$$

Rewriting the free energy and performing the saddle point over the hatted variables (which do not enter into the traces) the resulting free energy is

$$nF_n = \frac{\sigma^2}{2}\sum_{a\neq b}\left(Q^{\xi}_{ab}Q^{\eta}_{ab}\right) + \frac{\sigma^2}{2}\sum_{a}\left(Q^{\xi}_{aa}Q^{\eta}_{aa}\right) - \log Tr_{n,\eta,\xi}\Psi_{\eta,\xi}\Psi_{\eta}\Psi_{\xi}\,, \tag{E.16}$$

with

$$\begin{aligned} \Psi_{\xi} &= \exp\left[\sum_{a\neq b}\frac{\sigma^2}{2}Q^{\xi}_{ab}\xi^a\xi^b + \sum_{a}\frac{\sigma^2}{2}Q^{\xi}_{aa}\xi^a\xi^a + \sum_{a}V^{\xi}(\xi^a)\right] \\ \Psi_{\eta} &= \exp\left[\sum_{a\neq b}\frac{\sigma^2}{2}Q^{\eta}_{ab}\eta^a\eta^b + \sum_{a}\frac{\sigma^2}{2}Q^{\eta}_{aa}\eta^a\eta^a + \sum_{a}V^{\eta}(\eta^a)\right]\ . \end{aligned} \tag{E.17}$$

E.3. Replica symmetric solution

In this section we provide detailed calculation leading to the phase diagram Eq. (4.11). Starting from Eq. (4.9), we consider the replica symmetric ansatz

$$Q^{\xi}_{ab} = q^{\xi}_0\, Q^{\xi}_{aa} = q^{\xi}_D\, Q^{\eta}_{ab} = q^{\eta}_0\, Q^{\eta}_{aa} = q^{\eta}_D\,. \tag{E.18}$$

Plugging the previous ansatz into the free energy results in

$$nF_n = \frac{\sigma^2}{2}\sum_{a\neq b}\left(q^{\xi}_0 q^{\eta}_0\right) + \frac{\sigma^2}{2}\sum_{a}\left(q^{\xi}_D q^{\eta}_D\right) - \log Tr_{n,\eta,\xi}\Psi_{\eta}\Psi_{\xi}\Psi_{\eta,\xi}\,, \tag{E.19}$$

where

$$\begin{aligned} \Psi_{\xi} &= \exp\left[\frac{\sigma^2}{2}q^{\xi}_0\left(\sum_a \xi^a\right)^2 + \frac{\sigma^2}{2}\left(q^{\xi}_D - q^{\xi}_0\right)\sum_a(\xi^a)^2 + \sum_a V^{\xi}(\xi^a)\right] \\ \Psi_{\eta} &= \exp\left[\frac{\sigma^2}{2}q^{\eta}_0\left(\sum_a \eta^a\right)^2 + \frac{\sigma^2}{2}\left(q^{\eta}_D - q^{\eta}_0\right)\sum_a(\eta^a)^2 + \sum_a V^{\eta}(\eta_a)\right]\ . \end{aligned} \tag{E.20}$$

The replicated quenched averaged partition function is

$$\begin{aligned} [Z^n] = \int dq^{\xi}_0 dq^{\eta}_0 dq^{\xi}_D dq^{\eta}_D \exp\left\{\left[N\left(\sigma^2 q^{\xi}_0 q^{\eta}_0\frac{n(n-1)}{2} + \frac{\sigma^2}{2}nq^{\xi}_D q^{\eta}_D\right)\right]\right\} \\ \exp\left\{\left[-\left(\log\int\prod_a d\xi^a d\eta^a \Psi_{\eta,\xi}\Psi_{\eta}\,,\Psi_{\xi}\right)\right]\right\} \end{aligned} \tag{E.21}$$

where Tr is made explicit. Ψ_ξ and Ψ_η factorize and we are left to compute

$$\int \prod_a d\xi^a \exp\left\{\left[\frac{\sigma^2}{2} q_0^\xi \left(\sum_a \xi^a\right)^2 + \frac{\sigma^2}{2}\left(q_D^\xi - q_0^\xi\right)\sum_a (\xi^a)^2 + \sum_a V^\xi(\xi^a)\right]\right\}. \tag{E.22}$$

Performing and Hubbard Stratonovich transform to $S = \sum_a \xi_a$,

$$e^{-\frac{b^2}{4a}S^2} = \int_{-\infty}^{\infty} dz e^{-az^2+bSz}\,, \tag{E.23}$$

we obtain

$$\int dz e^{-z^2/2} \int \prod_a d\xi^a \exp\left\{\left[-\sum_a H_{RS}^\xi(\xi_a, z)\right]\right\}, \tag{E.24}$$

with

$$H_{RS}^\xi(\xi^a) = -\sigma\sqrt{q_0^\xi} z\xi^a - \frac{\sigma^2}{2}\left(q_D^\xi - q_0^\xi\right)(\xi^a)^2 + V^a(\xi^a)\,, \tag{E.25}$$

and similarly for H_{RS}^η.
We can write the free energy to minimized as

$$nF_n = \frac{\sigma^2}{2} n(n-1) q_0^\xi q_0^\eta + \frac{\sigma^2}{2} n q_D^\xi q_D^\eta - \left(\log Tr \Psi^\xi \Psi^\eta \Psi^{\eta,\xi}\right). \tag{E.26}$$

As fluctuations are typically of order $\sqrt{N}$, we add a spherical constraint to the previous equation by requiring

$$\frac{1}{N}\sum_i \xi_i^2 = 1\,, \frac{1}{N}\sum_i \eta_i^2 = 1\,. \tag{E.27}$$

The spherical constraints implies $q_D = 1$. We then add $2nN$ Lagrange multipliers $(\lambda_a^\xi, \lambda_a^\eta)$ in (E.21) in terms of the integral representation of the delta functions

$$1 = \int d\xi_i^a \delta(\sum_i (\xi_i^a)^2 - N)\ ,, \tag{E.28}$$

and similarly for η. The free energy is

$$nF_n = \frac{\sigma^2}{2} n(n-1) q_0^\xi q_0^\eta + \frac{\sigma^2}{2} n + \sum_a (\lambda_a^\xi + \lambda_a^\eta) - \left(\log Tr \Psi^\xi \Psi^\eta \Psi^{\eta,\xi}\right), \tag{E.29}$$

where

$$H_{RS}^\xi(\xi^a) = -\sigma\sqrt{q_0^\xi} z\xi^a - \frac{\sigma^2}{2}\left(1 - q_0^\xi\right)(\xi^a)^2 + V^a(\xi^a) - \lambda_a^\xi (\xi^a)^2\,, \tag{E.30}$$

and similarly for H_{RS}^η. Moreover, $V(\xi_a)$ and $V(\eta_a)$ are quadratic in η_a, ξ_a and their sum over replicas is a constant that we can safely neglect for the future computations. Consistently, we consider the replica symmetric ansatz for the multipliers: $\lambda_a^\xi = \lambda_0^\xi\, \lambda_a^\xi =$

λ_0^η

E.3.1. Overlaps of spherically constrained fluctuations

In order to find the value of the overlaps that minimize the free energy we take the saddle point equations

$$\frac{\delta F}{\delta q_0^\xi} = 0\,, \frac{\delta F}{\delta \lambda_0^\xi} = 0\,, \frac{\delta F}{\delta q_0^\eta} = 0\,, \frac{\delta F}{\delta \lambda_0^\eta} = 0\,. \tag{E.31}$$

The last equations reduce to the integral solution of four coupled equations

$$\begin{aligned}
q_0^\xi &= \frac{1}{N}\sum_i \langle \xi_i^a \xi_i^b \rangle = \frac{\int DzDw \prod^c d\xi^c d\eta^c \xi^a \xi^b e^{-\sum_c H_{RS}(\xi^c,\eta^c,z)}}{\int DzDw \prod_c d\xi^c d\eta^c e^{-\sum_c H_{RS}(\xi^c,\eta^c,z)}}\,,\\
1 &= \frac{1}{N}\sum_i \langle (\xi_i^a)^2 \rangle = \frac{\int DzDw \prod_c d\xi^c d\eta^c (\xi^a)^2 e^{-\sum_c H_{RS}(\xi^c,\eta^c,z)}}{\int DzDw \prod_c d\xi^c d\eta^c e^{-\sum_c H_{RS}(\xi^c,\eta^c,z)}}\,,\\
q_0^\eta &= \frac{1}{N}\sum_i \langle \eta_i^a \eta_i^b \rangle = \frac{\int DzDw \prod_c d\xi^c d\eta^c \eta^a \eta^b e^{-\sum_c H_{RS}(\xi^c,\eta^c,z)}}{\int DzDw \prod_c d\xi^c d\eta^c e^{-\sum_c H_{RS}(\xi^c,\eta^c,z)}}\,,\\
1 &= \frac{1}{N}\sum_i \langle (\eta_i^a)^2 \rangle = \frac{\int DzDw \prod_c d\xi^c d\eta^c (\eta^a)^2 e^{-\sum_c H_{RS}(\xi^c,\eta^c,z)}}{\int DzDw \prod_c d\xi^c d\eta^c e^{-\sum_c H_{RS}(\xi^c,\eta^c,z)}}\,,
\end{aligned} \tag{E.32}$$

where $Dz = dze^{-\frac{z^2}{2}}, Dw = dwe^{-\frac{w^2}{2}}$ and

$$\begin{aligned}
H_{RS}(\xi^a,\eta^a,z,w) = &-z\sigma\sqrt{q_0^\xi}\xi^a - w\sigma\sqrt{q_0^\eta}\eta^a + \tilde{K}_{\xi,\eta}\eta^a\xi^a\\
&- \sigma^2\left(1-q_0^\xi\right)\frac{(\xi^a)^2}{2} - \sigma^2\left(1-q_0^\eta\right)\frac{(\eta^a)^2}{2} + \lambda_0^\xi(\xi^a)^2 + \lambda_0^\eta(\eta^a)^2\,.
\end{aligned} \tag{E.33}$$

In the limit of strong interactions ($\sigma \to \infty$) we can solve the internal integral with the saddle point method. We then replace the integral with ξ_a^*, η_a^* that satisfies the equations $\frac{\partial H_{RS}}{\partial \xi_a}|_{\xi_a^*,\eta_a^*} = 0$, $\frac{\partial H_{RS}}{\partial \eta_a}|_{\xi_a^*,\eta_a^*} = 0$. In case of RS solution we simplify to

$$q_0^\xi = \int DzDw\eta_{a^2}^*\,. \tag{E.34}$$

As an example, when $\tilde{K}_{\xi,\eta} = 0$

$$q_0^\xi = \left(\frac{\sigma\sqrt{q_0}}{\sigma^2(1-q_0)}\right)^2 \int dze^{-\frac{z^2}{2}}z^2\,, \tag{E.35}$$

with the known solution [79]

$$q_0^\xi = 1 - \frac{1}{\sigma} = q_0^\eta\,. \tag{E.36}$$

The replica symmetric Hamiltonian has the symmetry under the exchange $\xi \longleftrightarrow \eta$, which simplifies the next computation as it implies that $q_0^\xi = q_0^\eta = q_0$. The exact solu-

tion of the coupled equations for the overlap and Lagrange multiplier can be recasted as [156],

$$\begin{aligned}\frac{1}{N}\sum_i \langle \xi_i^a \xi_i^b \rangle &= \int DzDw \left(\frac{\int d\xi^a d\eta^a \xi^a e^{-\sum_c H_{RS}(\xi^c,\eta^c,z)}}{\int d\xi^a d\eta^a e^{-H_{RS}(\xi^a,\eta^a,z)}} \right)^2 \\ \frac{1}{N}\sum_i \langle \xi_i^{a^2} \rangle &= \int DzDw \frac{\int d\xi^a d\eta^a \xi^{a^2} e^{-\sum_c H_{RS}(\xi^c,\eta^c,z)}}{\int d\xi^a d\eta^a e^{-H_{RS}(\xi^a,\eta^a,z)}} ,\end{aligned} \tag{E.37}$$

and the solution is

$$q_0 = 1 - \frac{\frac{1}{\sigma}\left(\sqrt{\frac{8}{\sigma^2}\tilde{K}_{\xi,\eta}^2 + 1} + 3\right)}{2\sqrt{2}\sqrt{\frac{2}{\sigma^2}\tilde{K}_{\xi,\eta}^2 + \sqrt{\frac{8}{\sigma^2}\tilde{K}_{\xi,\eta}^2 + 1} + 1}} . \tag{E.38}$$

Moreover, we found consistently that $\overline{\langle \xi_i \rangle} = 0$, where $\overline{\dots}$ indicates the average over the disorder J_{ij}. Indeed,

$$\overline{\langle \xi_i \rangle} = \frac{\int DzDw \prod_c d\xi_c d\eta_c \xi_a e^{-\sum_c H_{RS}(\xi_c,\eta_c,z)}}{\int DzDw \prod_c d\xi_c d\eta_c e^{-\sum_c H_{RS}(\xi_c,\eta_c,z)}} \tag{E.39}$$

in the limit $\sigma \to 0$, we substitute the internal integral with the saddle point and giving that ξ^* is an odd function of z, w the result of the external integral is zero. The same results apply to $\overline{\langle \eta_i \rangle}$.

Alternative calculation

For a generical 2-spin hamiltonian, we could have started from (E.13) and following [155] inserting the definition of the overlaps as well as the spherical constraints as delta distributions with Lagrange multipliers λ_{ab}. The resulting free energy is

$$nF_n = \frac{\sigma^2}{2}\sum_{a,b} q_{ab}^{\xi} q_{ab}^{\eta} + \sum_{ab} \lambda_{ab}^{\xi} q_{ab} + \sum_{ab} \lambda_{ab}^{\eta} q_{ab} - \log Tr_{n,\eta,\xi} \Psi_\eta \Psi_\xi \Psi_{\eta,\xi}, \tag{E.40}$$

where

$$Tr_{n,\eta,\xi} \Psi_\eta \Psi_\xi \Psi_{\eta,\xi} = \int d\boldsymbol{\xi} d\boldsymbol{\eta} exp\left[\sum_{a,b} \lambda_{ab}^{\xi} \xi^a \xi^b + \sum_{a,b} \lambda_{ab}^{\eta} \eta^a \eta^b + \sum_a \tilde{K}_{\xi,\eta} \right] . \tag{E.41}$$

We omitted the potential due to the spherical constraint. The argument of the exponential in the last expression can be rewritten as

$$(\boldsymbol{\xi}, \boldsymbol{\eta}) \begin{pmatrix} \boldsymbol{\lambda}^{\eta} & \frac{\tilde{K}_{\xi,\eta}}{2}\boldsymbol{I} \\ \frac{\tilde{K}_{\xi,\eta}}{2}\boldsymbol{I} & \boldsymbol{\lambda}^{\eta} \end{pmatrix} \begin{pmatrix} \boldsymbol{\xi} \\ \boldsymbol{\eta} \end{pmatrix} , \tag{E.42}$$

and the free energy is

$$nF_n = \frac{\sigma^2}{2}\sum_{a,b} q^{\xi}_{ab}q^{\eta}_{ab} + \sum_{ab}\lambda^{\xi}_{ab}q_{ab} + \sum_{ab}\lambda^{\eta}_{ab}q_{ab} - \frac{1}{2}\log det(\boldsymbol{\lambda^{\eta}\lambda^{\xi}} - \frac{\tilde{K^2}_{\xi,\eta}}{4}\boldsymbol{I})\,. \tag{E.43}$$

When $\tilde{K}_{\xi,\eta} \gg 1$ (ignoring constants)

$$\log det(\boldsymbol{\lambda^{\eta}\lambda^{\xi}} - \frac{\tilde{K^2}_{\xi,\eta}}{4}\boldsymbol{I}) = \log\left[\left(-\frac{\tilde{K^2}_{\xi,\eta}}{4}\right)^n det(\boldsymbol{I} - \frac{4}{\tilde{K^2}_{\xi,\eta}}\boldsymbol{\lambda^{\eta}\lambda^{\xi}})\right] \approx Tr(\boldsymbol{\lambda^{\eta}\lambda^{\xi}}\frac{4}{\tilde{K^2}_{\xi,\eta}})\,. \tag{E.44}$$

In order to find the value of the overlaps that minimize the free energy we take the saddle point equations

$$\frac{\delta F}{\delta q^{\xi}_{ab}} = 0\,, \frac{\delta F}{\delta \lambda^{\xi}_{ab}} = 0\,, \frac{\delta F}{\delta q^{\eta}_{ab}} = 0\,, \frac{\delta F}{\delta \lambda^{\eta}_{ab}} = 0\,. \tag{E.45}$$

The extremization with respect to λ leads to

$$q^{\xi}_{ab} = \frac{2}{\tilde{K^2}_{\xi,\eta}}\lambda^{\eta}_{ba}\,, \xi \leftrightarrow \eta\,, \tag{E.46}$$

and

$$\frac{\sigma^2}{2}q^{\xi}_{ab} - q^{\eta}_{ab} = 0\,, \xi \leftrightarrow \eta\,. \tag{E.47}$$

A part from the paramagnetic solution $q_{ab} = 0, a, b = 1 \ldots n$, the last equation is always solved if $\sigma = \sqrt{2}$. The calculation can be carried similarly for $\tilde{K}_{\xi,\eta} \ll 1$ and we found a limiting value $\sigma = 1$, as in the previous derivation.

E.3.2. Overlaps of binary fluctuations

We may remove the spherical constraint and consider binary fluctuations, in this case the free energy reduces to the one of a classical bipartite Sherrington-Kirkpatrick model with an additional term that couples ξ_i and η_i, namely $\tilde{K}_{\xi,\eta}$,

$$\begin{aligned} nF_n = &\frac{\sigma^2}{2}\sum_{a\neq b}\left(Q^{\xi}_{ab}Q^{\eta}_{ab}\right) + \frac{\sigma^2}{2}n \\ &- \log Tr_{n,\eta,\xi}\Psi_{\eta,\xi}\Psi_{\eta}\Psi_{\xi}\,, \end{aligned} \tag{E.48}$$

where

$$\begin{aligned}\Psi_\xi &= \exp\left[\sum_{a\neq b}\frac{\sigma^2}{2}Q^\xi_{ab}\xi_a\xi_b\right],\\ \Psi_\eta &= \exp\left[\sum_{a\neq b}\frac{\sigma^2}{2}Q^\eta_{ab}\eta_a\eta_b\right],\\ \Psi_{\eta,\xi} &= \exp\left[\sum_a \tilde{K}_{\xi,\eta}\xi_a\eta_a\right].\end{aligned} \tag{E.49}$$

We perform again a Hubbard-Stratonovich transformation and taking the RS solution we arrive to ($f = F_n$),

$$\begin{aligned}f = -\frac{\sigma^2}{2}(q_0^\xi - 1)(q_0^\eta - 1) - \langle\Big[&4ch(\sigma\sqrt{q_0^\xi}z)ch(\sigma\sqrt{q_0^\eta}w)ch(\tilde{K}_{\xi,\eta})+\\ &4sh(\sigma\sqrt{q_0^\xi}z)sh(\sigma\sqrt{q_0^\eta}w)sh(\tilde{K}_{\xi,\eta})\Big]\rangle_{z,w}\,.\end{aligned} \tag{E.50}$$

Upon finding the saddle point as done previously and expanding the self-consistency equations, we find that the phase boundary between the paramagnetic and glassy phase is given by

$$\sigma^2(1+\tanh\Big(\tilde{K}_{\xi,\eta}\Big)) = 1\,. \tag{E.51}$$

E.4. MSRJD path integral of spin glass dynamics

A generating function for the coupled Langevin equations (4.15) is (Sec. 1.3.1),

$$\begin{aligned}&Z[\boldsymbol{h}^\xi,\boldsymbol{h}^\eta] = \int \mathrm{D}[\boldsymbol{\xi}]\mathrm{D}[\boldsymbol{\eta}]\mathrm{D}[\hat{\boldsymbol{\xi}}]\mathrm{D}[\hat{\boldsymbol{\eta}}]e^{i\int dt\sum_j\left(h_j^\xi\xi_j+h_j^\eta\eta_j\right)}\\ &e^{i\int dt\sum_j\hat{\xi}_j\left[\partial_t\xi_j-g_j\eta_j+d_j\xi-D_j^n\hat{\xi}_j\right]}e^{i\int dt\sum_j\hat{\eta}_j\left[\partial_t\eta_j+\gamma_j\eta_j-\sum_{k\in e(j)}f'_{k,j}(\phi_k^*)\xi_k-\gamma_j b_j-D_j^m\hat{\eta}_j\right]}.\end{aligned} \tag{E.52}$$

We initially isolate the part which includes disorder by replacing $G_j = \sum_{k\in e(j)} f'(\phi_k^*)\xi_k$. By doing so, we formally introduce a new delta

$$\begin{aligned}&Z[\boldsymbol{h}^\xi,\boldsymbol{h}^\eta] = \int \mathrm{D}[\boldsymbol{\xi}]\mathrm{D}[\boldsymbol{\eta}]\mathrm{D}[\hat{\boldsymbol{\xi}}]\mathrm{D}[\hat{\boldsymbol{\eta}}]\mathrm{D}[\hat{\boldsymbol{G}}]\mathrm{D}[\boldsymbol{G}]e^{i\int dt\sum_j\left(h_j^\xi\xi_j+h_j^\eta\eta_j\right)}\\ &e^{i\int dt\sum_j\hat{G}_j^\xi\left(G_j-\sum_{k\in e(j)}f'_{k,j}(\phi_k^*)\xi_k\right)}e^{i\int dt\sum_j\hat{\xi}_j\left[\partial_t\xi_j-g_j\eta_j+d_j\xi-D_j^n\hat{\xi}_j\right]}e^{i\int dt\sum_j\hat{\eta}_j\left[\partial_t\eta_j+\gamma_j\eta_j-G_j-\gamma_j b_j-D_j^m\hat{\eta}_j\right]}.\end{aligned} \tag{E.53}$$

In the dynamical representation of glassy fluctuations time plays a similar role as the replica index in the equilibrium regime. We now integrate over the couplings $f'_{k,j}(\phi_k^*) = J_{kj}$ with statistics, $\overline{J_{kj}} = 0$, $\overline{J_{kj}^2} = \sigma^2/N$, and $\overline{J_{kj}J_{jk}} = \lambda\sigma^2/N$. We then integrate over

J_{kj} ($\overline{\cdots}$) in the exponential

$$\overline{e^{-i\sum_j \int dt \hat{G}_j(t) \sum_{k\in e(j)} J_{kj}\xi_k(t)}}\,, \tag{E.54}$$

which results in

$$\begin{aligned} & e^{-\frac{\sigma^2}{2}N\int dtdt'\left(L(t,t')C^{\xi}(t,t')+\lambda K(t,t')K(t',t)\right)}\,, \\ & L(t,t') = \frac{1}{N}\sum_j \hat{G}_j(t)\hat{G}_j(t')\,, \\ & C^{\xi}(t,t') = \frac{1}{N}\sum_j \xi_j(t)\xi_j(t')\,, \\ & K(t,t') = \frac{1}{N}\sum_j \xi_j(t)\hat{G}_j(t')\,. \end{aligned} \tag{E.55}$$

From now on we can follow [179] and arrive to (4.16) .

E.5. Out of equilibrium dynamics of p-spin spherical asymmetric bipartite spin glasses

Starting from Eq. (4.21),

$$\begin{aligned} & \partial_t C^{\xi}(t,t') = g_i C_i^{\xi,\eta}(t,t') - d_i C_i^{\xi}(t,t') - \mu^{\xi}(t)C_i^{\xi}(t,t') + D_i^n \langle W^{\xi}(t)\xi(t')\rangle\,, \\ & \partial_t C_i^{\eta}(t,t') = -\gamma_i C_i^{\eta}(t,t') - \mu^{\eta_i}(t)C_i^{\eta}(t,t') + \langle W_c^{\eta_i}(t)\eta(t')\rangle\,, \end{aligned} \tag{E.56}$$

the spherical constraint impose $C_i^{\xi/\eta}(t,t')|_{t'\to t} = 1$. $\mu_i^{\xi/\eta}(t)$ can be evaluated by the relationship $\partial_t C_i^{\xi/\eta}(t,t')|_{t'\to t} + \partial_{t'} C_i^{\xi/\eta}(t,t')|_{t\to t'} = 1$

$$\begin{aligned} & \mu^{\xi}(t) = (g_i C_i^{\xi,\eta}(t,t) - d_i) + D_i^n \langle W_i^{\xi}(t)\xi_i(t)\rangle\,, \\ & \mu^{\eta}(t) = -\gamma_i + \langle W_c^{\eta_i}(t)\eta_i(t)\rangle\,. \end{aligned} \tag{E.57}$$

We are then just left to compute the terms in the brackets. In particular with Novikov's theorem [189] the first one is

$$\langle W_i^{\xi}(t)\xi_i(t')\rangle = \int ds \langle W_i^{\xi}(t)W_i^{\xi}(s)\rangle \chi_i^{\xi}(t',s) = \chi_i^{\xi}(t,t')\,, \tag{E.58}$$

where the propagator $\chi_i^{\xi}(t,t') = \langle \frac{\delta\xi_i(t)}{\delta W_i^{\xi}(t')}\rangle$. We take $t' > t$ and so $\chi(t,t') = 0$ for $t' > t$ due to causality and $\lim_{t'\to t}\chi(t,t') = 1$. We then find $\mu_i^{\xi}(t) = g_i C_i^{\xi,\eta}(t,t) - d_i + \frac{D_i^n}{2}$.

The second average is

$$\langle W_i^\eta(t)\eta(t')_i\rangle = \int ds\langle W_i^\eta(t)W_i^\eta(s)\rangle\chi_i^\eta(t',s) = D_i^m\chi_i^\eta(t,t') + \sigma^2\int_{-\infty}^{t'} dsC_i^\xi(s,t)\chi_i^\eta(t',s)\,. \tag{E.59}$$

Assuming again $t' > t$ we arrive to $\mu_i^\eta(t) = -\gamma_i + \sigma^2\int_{-\infty}^{t'} dsC_i^\xi(s,t)\chi_i^\eta(t,s) + \frac{D_i^m}{2}$. We have only to evaluate equal time cross correlations, which are given by

$$\begin{aligned}\partial_t C_i^{\xi,\eta}(t,t') &= g_i C_i^\eta(t,t') - d_i C_i^{\xi,\eta}(t,t') - \mu_i^\xi(t)C_i^{\xi,\eta}(t,t')\,,\\ \partial_t C_i^{\xi,\eta}(t,t') &= -\gamma_i C_i^{\xi,\eta}(t,t') - \mu_i^\eta(t)C_i^{\xi,\eta}(t,t')\,,\end{aligned} \tag{E.60}$$

and subtracting the two equations we arrive to

$$C_i^{\xi,\eta}(t,t') = \frac{g_i C_i^\eta(t,t')}{d_i - \gamma_i + \mu^\xi(t) - \mu_i^\eta(t)}\,. \tag{E.61}$$

It has to be noticed that $\mu^\xi(t)$ depends on the equal cross-correlation,

$$C_i^{\xi,\eta}(t,t') = \frac{g_i C_i^\eta(t,t')}{g_i C_i^{\xi,\eta}(t,t) + \frac{D_i^n}{2} - \sigma^2\int_{-\infty}^{t'} dsC_i^\xi(s,t)\chi_i^\eta(t,s) - \frac{D_i^m}{2}}\,. \tag{E.62}$$

We close the equation for the spherical constraints upon solving the equation for the response of mRNA fluctuations,

$$\partial_t\chi_i^\eta(t,t') = -(\gamma_i + \mu_i^\eta(t))\chi_i^\eta(t,t') + \delta(t,t')\,. \tag{E.63}$$

Once we found the equation for the spherical constraint the equation for the correlations are compactly given by

$$\begin{aligned}\partial_t C_i^\xi(t,t') &= g_i(C^{\xi,\eta}(t,t') - C_i^{\xi,\eta}(t,t)C^\xi(t,t')) - \frac{D_i^n}{2}C_i^\xi(t,t') + D_i^n\chi^\xi(t,t')\,,\\ \partial_t C_i^\eta(t,t') &= -\frac{D_i^m}{2}C^\eta(t,t') + D_i^m\chi^\eta(t,t')_i + \sigma^2(1 - C_i^\eta(t,t'))\int_{-\infty}^{t'} dsC_i^\xi(s,t)\chi_i^\eta(t,s)\,,\end{aligned} \tag{E.64}$$

In the following we drop the index i and take delta distributions of the parameters. For long enough times, we look for solutions such that the FDT holds, and so the last equation is ($t' > t, t' - t = \tau$),

$$\begin{aligned}\partial_t C^\xi(\tau) &= g\left(C^{\xi,\eta}(\tau) - C^{\xi,\eta}(0)C^\xi(\tau)\right) - \frac{D^n}{2}C^\xi(\tau)\,,\\ \partial_t C^\eta(\tau) &= -\frac{D^m}{2}C^\eta(\tau) - \frac{2\sigma^2}{D^m}(1 - C^\eta(\tau))\int_0^\tau duC^\xi(\tau - u)\partial_u C^\eta(u)\,.\end{aligned} \tag{E.65}$$

The last integral can be approximated for $\tau \gg 1$,

$$\begin{aligned}\partial_\tau C^\xi(\tau) &= g\left(C^{\xi,\eta}(\tau) - C^{\xi,\eta}(0)C^\xi(\tau)\right) - \frac{D^n}{2}C^\xi(\tau)\,,\\ \partial_\tau C^\eta(\tau) &= -\frac{D^m}{2}C^\eta(\tau) + \frac{2\sigma^2}{D^m}(1-C^\eta(\tau))^2 C^\xi(\tau)\,.\end{aligned} \tag{E.66}$$

As we are looking at a relaxation problem we expect $\partial_\tau C^{\xi\backslash\eta}(\tau) \leq 0$ and find $C^{\xi,\eta}(0) \approx 1$. The inequalities in this regime are given by (upon substituting the solution of $C^{\xi,\eta}$)

$$\begin{aligned}&\frac{4\sigma^2}{D^{m^2}}C^\xi(\tau)(1-C^\eta(\tau)) > \frac{2g + D^n - D^m}{D^m} or\\ &4C^\eta(\tau)g^2 + C^\xi(\tau)(D^n+2g)\left(D^m - D^n - 2g + \frac{2\sigma^2}{D^m}C^\xi(\tau)\,(1-C^\eta(\tau))\right) \leq 0\,,\end{aligned} \tag{E.67}$$

Being $D^m \gg D^n, g$ the second inequality is always satisfied. The first inequality for $\partial_\tau C^\eta(\tau)$ is

$$\frac{4\sigma^2}{D^{m^2}}C^\xi(\tau)(1-C^\eta(\tau))^2 \leq C^\eta(\tau)\,. \tag{E.68}$$

In the same condition $D^m \gg D^n, g$ we find a dynamical transition at

$$\sigma_c = D^m/2\,. \tag{E.69}$$

If we want to compare these results with the phase diagram (4.11) (Fig. 4.2) we have to divide σ^2 by D^m and the critical line is defined by $\sigma_c = 1/2$.

References

1. Schrödinger, E. & Penrose, R. *What is Life?: With Mind and Matter and Autobiographical Sketches* (Cambridge University Press, 1992).
2. Fang, X., Kruse, K., Lu, T. & Wang, J. Nonequilibrium physics in biology. *Rev. Mod. Phys.* **91,** 045004 (2019).
3. Alberts, B. *Molecular biology of the cell* (Garland Science, New York, NY, 2015).
4. Zhang, J., Nie, Q. & Zhou, T. Revealing Dynamic Mechanisms of Cell Fate Decisions From Single-Cell Transcriptomic Data. *Frontiers in Genetics* **10,** 1280 (2019).
5. Hwang, B., Lee, J. H. & Bang, D. Single-cell RNA sequencing technologies and bioinformaticspipelines. *Experimental & Molecular Medicine* **50,** 1–14 (2018).
6. Argelaguet, R. *et al.* Multi-omics profiling of mouse gastrulation at single-cell resolution. *Nature* **576,** 487–491 (2019).
7. Zinn-Justin, J. *Quantum Field Theory and Critical Phenomena; 4th ed.* (Clarendon Press, Oxford, 2002).
8. Täuber, U. C. *Critical Dynamics: A Field Theory Approach to Equilibrium and Non-Equilibrium Scaling Behavior* (Cambridge University Press, 2014).
9. Grace, K., Salvatier, J., Dafoe, A., Zhang, B. & Evans, O. When Will AI Exceed Human Performance? Evidence from AI Experts. *Journal of Artificial Intelligence Research* **92,** 729–754 (2018).
10. Silver, D. *et al.* Mastering the Game of Go with Deep Neural Networks and Tree Search. *Nature* **529,** 484–489 (2016).
11. Language: Disputed definitions. *Nature* **455,** 1023–1028 (2008).
12. Lee, H. J., Hore, T. A. & Reik, W. Reprogramming the Methylome: Erasing Memory and Creating Diversity. *Cell Stem Cell* **14,** 710–719 (2014).
13. Bergman, Y. & Cedar, H. DNA methylation dynamics in health and disease. *Nature Structural & Molecular Biology* **20,** 274–281 (2013).
14. Bell, C. G. *et al.* DNA methylation aging clocks: challenges and recommendations. *Genome Biology* **20,** 249 (2019).

15. Lee, H. J., Hore, T. A. & Reik, W. Reprogramming the methylome: erasing memory and creating diversity. *Cell Stem Cell* **14,** 710–9 (2014).

16. Clark, S. J. *et al.* scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells. *Nature Communications* **9,** 781 (2018).

17. Macaulay, I. C., Ponting, C. P. & Voet, T. Single-Cell Multiomics: Multiple Measurements from Single Cells. *Trends Genet* **33,** 155–168 (2017).

18. Kelsey, G., Stegle, O. & Reik, W. Single-cell epigenomics: Recording the past and predicting the future. *Science* **358,** 69–75 (2017).

19. Baubec, T. *et al.* Genomic profiling of DNA methyltransferases reveals a role for DNMT3B in genic methylation. *Nature* **520,** 243–7 (2015).

20. Bestor, T. H. The DNA methyltransferases of mammals. *Human Molecular Genetics* **9,** 2395–2402 (2000).

21. Gao, L. *et al.* Comprehensive structure-function characterization of DNMT3B and DNMT3A reveals distinctive de novo DNA methylation mechanisms. *Nature Communications* **11,** 3355 (2020).

22. Kaneda, M. *et al.* Essential role for de novo DNA methyltransferase Dnmt3a in paternal and maternal imprinting. *Nature* **429,** 900–903 (2004).

23. Gopalakrishnan, S., Sullivan, B. A., Trazzi, S., Della Valle, G. & Robertson, K. D. DNMT3B interacts with constitutive centromere protein CENP-C to modulate DNA methylation and the histone code at centromeric regions. *Human Molecular Genetics* **18,** 3178–3193 (2009).

24. Xu, G.-L. *et al.* Chromosome instability and immunodeficiency syndrome caused by mutations in a DNA methyltransferase gene. *Nature* **402,** 187–191 (1999).

25. Jurkowska, R. Z., Jurkowski, T. P. & Jeltsch, A. Structure and Function of Mammalian DNA Methyltransferases. *ChemBioChem* **12,** 206–222 (2011).

26. Neri, F. *et al.* Dnmt3L Antagonizes DNA Methylation at Bivalent Promoters and Favors DNA Methylation at Gene Bodies in ESCs. *Cell* **155,** 121–134 (2013).

27. Jurkowska, R. Z. *et al.* Oligomerization and Binding of the Dnmt3a DNA Methyltransferase to Parallel DNA Molecules: heterochromatin localization and the role of Dnmt3L. *Journal of Biological Chemistry* **286,** 24200–24207 (2011).

28. Law, J. A. & Jacobsen, S. E. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nature Reviews Genetics* **11,** 204–220 (2010).

29. Jia, D., Jurkowska, R. Z., Zhang, X., Jeltsch, A. & Cheng, X. Structure of Dnmt3a bound to Dnmt3L suggests a model for de novo DNA methylation. *Nature* **449,** 248–251 (2007).

30. Jeltsch, A. & Jurkowska, R. Z. Allosteric control of mammalian DNA methyltransferases – a new regulatory paradigm. *Nucleic Acids Research* **44,** 8556–8575 (2016).

31. Seisenberger, S. *et al.* The Dynamics of Genome-wide DNA Methylation Reprogramming in Mouse Primordial Germ Cells. *Molecular Cell* **48,** 849–862 (2012).

32. Rulands, S. *et al.* Genome-Scale Oscillations in DNA Methylation during Exit from Pluripotency. *Cell Systems* **7,** 63–76.e12 (2018).

33. Moore, L. D., Le, T. & Fan, G. DNA Methylation and Its Basic Function. *Neuropsychopharmacology* **38,** 23–38 (2013).

34. Li, E., Bestor, T. H. & Jaenisch, R. Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. *Cell* **69,** 915–926 (1992).

35. Robert, M.-F. *et al.* DNMT1 is required to maintain CpG methylation and aberrant gene silencing in human cancer cells. *Nature Genetics* **33,** 61–65 (2003).

36. Little, M. & Wainwright, B. Methylation and p16: Suppressing the suppressor. *Nature Medicine* **1,** 633–634 (1995).

37. O'Hagan, H. M. *et al.* Oxidative damage targets complexes containing DNA methyltransferases, SIRT1, and polycomb members to promoter CpG Islands. *Cancer Cell* **20,** 606–19 (2011).

38. Wu, X. & Zhang, Y. TET-mediated active DNA demethylation: mechanism, function and beyond. *Nature Reviews Genetics* **18,** 517–534 (2017).

39. He, Y. F. *et al.* Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA. *Science* **333,** 1303–7 (2011).

40. Parry, A., Rulands, S. & Reik, W. Active turnover of DNA methylation during cell fate decisions. *Nature Reviews Genetics* **22,** 59–66 (2021).

41. Johnson, A. A. *et al.* The role of DNA methylation in aging, rejuvenation, and age-related disease. *Rejuvenation Res* **15,** 483–94 (2012).

42. Hannum, G. *et al.* Genome-wide Methylation Profiles Reveal Quantitative Views of Human Aging Rates. *Molecular Cell* **49,** 359–367 (2013).

43. Horvath, S. DNA methylation age of human tissues and cell types. *Genome Biology* **14,** 3156 (2013).

44. Elowitz, M. B., Levine, A. J., Siggia, E. D. & Swain, P. S. Stochastic gene expression in a single cell. *Science* **297,** 1183–6 (2002).

45. Shahrezaei, V. & Swain, P. S. Analytical distributions for stochastic gene expression. *Proceedings of the National Academy of Sciences* **105,** 17256–17261 (2008).

46. Gasperini, M., Tome, J. M. & Shendure, J. Towards a comprehensive catalogue of validated and target-linked human enhancers. *Nature Reviews Genetics* **21,** 292–310 (2020).

47. Ud-Dean, S. M., Heise, S., Klamt, S. & Gunawan, R. TRaCE+: Ensemble inference of gene regulatory networks from transcriptional expression profiles of gene knock-out experiments. *BMC Bioinformatics* **17,** 252 (2016).

48. What Is Your Conceptual Definition of "Cell Type" in the Context of a Mature Organism? *Cell Syst* **4,** 255–259 (2017).

49. Waddington, C. H. *The strategy of the genes: A discussion of some aspects of theoretical biology* (Allen & Unwin, London, 1957).

50. Wang, J., Zhang, K., Xu, L. & Wang, E. Quantifying the Waddington landscape and biological paths for development and differentiation. *Proceedings of the National Academy of Sciences* **108,** 8257–8262 (2011).

51. Bhattacharya, S., Zhang, Q. & Andersen, M. E. A deterministic map of Waddington's epigenetic landscape for cell fate specification. *BMC Systems Biology* **5,** 85 (2011).

52. Philip Greulich, R. S. & MacArthur, B. D. *The physics of cell fate* 189–206 (Academic Press, 2020).

53. Viotti, M., Foley, A. C. & Hadjantonakis, A.-K. Gutsy moves in mice: cellular and molecular dynamics of endoderm morphogenesis. *Philosophical Transactions of the Royal Society B: Biological Sciences* **369,** 20130547 (2014).

54. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29,** 15–21 (2012).

55. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* **26,** 589–595 (2010).

56. Li, Y. & Tollefsbol, T. O. DNA methylation detection: bisulfite genomic sequencing analysis. *Methods Mol Biol* **791,** 11–21 (2011).

57. Owens, B. Genomics: The single life. *Nature* **491,** 27–29 (2012).

58. Lister, R. *et al.* Highly Integrated Single-Base Resolution Maps of the Epigenome in Arabidopsis. *Cell* **133,** 523–536 (2008).

59. Stark, R., Grzelak, M. & Hadfield, J. RNA sequencing: the teenage years. *Nature Reviews Genetics* **20,** 631–656 (2019).

60. Pearson, K. On Lines and Planes of Closest Fit to Systems of Points in Space. *Philosophical Magazine* **2,** 559–572 (1901).

61. McInnes, L., Healy, J. & Melville, J. *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction* 2018.

62. Jin, X. & Han, J. in *Encyclopedia of Machine Learning and Data Mining* (eds Sammut, C. & Webb, G. I.) 695–697 (Springer US, Boston, MA, 2017).

63. Levine, J. H. *et al.* Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis. *Cell* **162,** 184–197 (2015).

64. Xu, C. & Su, Z. Identification of cell types from single-cell transcriptomes using a novel clustering method. *Bioinformatics* **31,** 1974–1980 (2015).

65. Traag, V. A., Waltman, L. & van Eck, N. J. From Louvain to Leiden: guaranteeing well-connected communities. *Scientific Reports* **9,** 5233 (2019).

66. Traag, V. A., Van Dooren, P. & Nesterov, Y. Narrow scope for resolution-limit-free community detection. *Phys. Rev. E* **84,** 016114 (2011).

67. Hochgerner, H., Zeisel, A., Lönnerberg, P. & Linnarsson, S. Conserved properties of dentate gyrus neurogenesis across postnatal development revealed by single-cell RNA sequencing. *Nature Neuroscience* **21,** 290–299 (2018).

68. Kamenev, A. *Field Theory of Non-Equilibrium Systems* (Cambridge University Press, 2011).

69. Gardiner, C. W. *Handbook of stochastic methods for physics, chemistry and the natural sciences* Third, xviii+415 (Springer-Verlag, Berlin, 2004).

70. Martin, P. C., Siggia, E. D. & Rose, H. A. Statistical Dynamics of Classical Systems. *Phys. Rev. A* **8,** 423–437 (1973).

71. Janssen, H.-K. On a Lagrangean for classical field dynamics and renormalization group calculations of dynamical critical properties. *Zeitschrift für Physik B Condensed Matter* **23,** 377–380 (1976).

72. De Dominicis, C. Techniques de renormalisation de la théorie des champs et dynamique des phénomènes critiques. *J. Phys. Colloques* **37,** C1–247–C1–253 (1976).

73. Doi, M. Stochastic theory of diffusion-controlled reaction. *Journal of Physics A: Mathematical and General* **9,** 1479–1495 (1976).

74. Peliti, L. Path integral approach to birth-death processes on a lattice. *J. Phys. France* **46,** 1469–1483 (1985).

75. Cardy, J., Falkovich, G. & Gawedzki, K. *Non-equilibrium Statistical Mechanics and Turbulence* (Cambridge University Press, Cambridge, 2008).

76. Wilson, K. G. The renormalization group and critical phenomena. *Rev. Mod. Phys.* **55,** 583–600 (1983).

77. Sherrington, D. & Kirkpatrick, S. Solvable Model of a Spin-Glass. *Phys. Rev. Lett.* **35,** 1792–1796 (1975).

78. Parisi, G. Infinite Number of Order Parameters for Spin-Glasses. *Phys. Rev. Lett.* **43,** 1754–1756 (1979).

79. Mezard, M., Parisi, G. & Virasoro, M. *Spin Glass Theory and Beyond* (World Scientific, 1986).

80. Edwards, S. F. & Anderson, P. W. Theory of spin glasses. *Journal of Physics F: Metal Physics* **5,** 965–974 (1975).

81. Anderson, P. W. More Is Different. *Science* **177,** 393–396 (1972).

82. Olmeda, F. *et al.* Inference of emergent spatio-temporal processes from single-cell sequencing reveals feedback between de novo DNA methylation and chromatin condensates. *bioRxiv* (2021).

83. Kalkan, T. *et al.* Tracking the embryonic stem cell transition from ground state pluripotency. *Development* **144,** 1221–1234 (2017).

84. Deaton, A. M. & Bird, A. CpG islands and the regulation of transcription. *Genes Dev* **25,** 1010–22 (2011).

85. Cedar, H. & Bergman, Y. Linking DNA methylation and histone modification: patterns and paradigms. *Nature Reviews Genetics* **10,** 295–304 (2009).

86. Auclair, G., Guibert, S., Bender, A. & Weber, M. Ontogeny of CpG island methylation and specificity of DNMT3 methyltransferases during embryonic development in the mouse. *Genome Biology* **15,** 545 (2014).

87. Rajavelu, A., Jurkowska, R. Z., Fritz, J. & Jeltsch, A. Function and disruption of DNA Methyltransferase 3a cooperative DNA binding and nucleoprotein filament formation. *Nucleic Acids Research* **40,** 569–580 (2011).

88. Sneppen, K. & Dodd, I. B. Nucleosome dynamics and maintenance of epigenetic states of CpG islands. *Phys. Rev. E* **93,** 062417 (2016).

89. Zhang, L. *et al.* DNA Methylation Landscape Reflects the Spatial Organization of Chromatin in Different Cells. *Biophysical Journal* **113,** 1395–1404 (2017).

90. Lövkvist, C., Sneppen, K. & Haerter, J. O. Exploring the Link between Nucleosome Occupancy and DNA Methylation. *Front Genet* **8,** 232 (2017).

91. Ginelli, F., Hinrichsen, H., Livi, R., Mukamel, D. & Politi, A. Directed percolation with long-range interactions: Modeling nonequilibrium wetting. *Phys. Rev. E* **71,** 026121 (2005).

92. Hinrichsen, H. Non-equilibrium phase transitions with long-range interactions. *Journal of Statistical Mechanics: Theory and Experiment* **2007,** P07006–P07006 (2007).

93. Ballerini, M. *et al.* Interaction ruling animal collective behavior depends on topological rather than metric distance: Evidence from a field study. *Proceedings of the National Academy of Sciences* **105,** 1232–1237 (2008).

94. Brunel, V., Oerding, K. & van Wijland, F. Fermionic field theory for directed percolation in $(1 + 1)$-dimensions. *Journal of Physics A: Mathematical and General* **33,** 1085–1097 (2000).

95. Wijland, F. v. Field theory for reaction-diffusion processes with hard-core particles. *Phys. Rev. E* **63,** 022101 (2001).

96. Hohenberg, P. C. & Halperin, B. I. Theory of dynamic critical phenomena. *Rev. Mod. Phys.* **49,** 435–479 (1977).

97. Ginelli, F., Hinrichsen, H., Livi, R., Mukamel, D. & Torcini, A. Contact processes with long range interactions. *Journal of Statistical Mechanics: Theory and Experiment* **2006,** P08008–P08008 (2006).

98. Altland, A. & Simons, B. D. *Condensed Matter Field Theory* 2nd ed. (Cambridge University Press, 2010).

99. Mora, T. & Bialek, W. Are Biological Systems Poised at Criticality? *Journal of Statistical Physics* **144,** 268–302 (2011).

100. Mastromatteo, I. & Marsili, M. On the criticality of inferred models. *Journal of Statistical Mechanics: Theory and Experiment* **2011,** P10012 (2011).

101. Benitez, F. *et al.* Langevin Equations for Reaction-Diffusion Processes. *Phys. Rev. Lett.* **117,** 100601 (2016).

102. Hinrichsen, H. Non-equilibrium critical phenomena and phase transitions into absorbing states. *Advances in Physics* **49,** 815–958 (2000).

103. Kardar, M., Parisi, G. & Zhang, Y.-C. Dynamic Scaling of Growing Interfaces. *Phys. Rev. Lett.* **56,** 889–892 (1986).

104. Kechagia, P., Yortsos, Y. C. & Lichtner, P. Nonlocal Kardar-Parisi-Zhang equation to model interface growth. *Phys. Rev. E* **64,** 016315 (2001).

105. Barabási, A. L. & Stanley, H. E. *Fractal Concepts in Surface Growth* (Cambridge University Press, 1995).

106. Medina, E., Hwa, T., Kardar, M. & Zhang, Y.-C. Burgers equation with correlated noise: Renormalization-group analysis and applications to directed polymers and interface growth. *Phys. Rev. A* **39,** 3053–3075 (1989).

107. Mahdisoltani, S., Zinati, R. B. A., Duclut, C., Gambassi, A. & Golestanian, R. Nonequilibrium polarity-induced chemotaxis: Emergent Galilean symmetry and exact scaling exponents. *Phys. Rev. Research* **3,** 013100 (2021).

108. Jimenez-Useche, I. *et al.* DNA Methylation Effects on Tetra-Nucleosome Compaction and Aggregation. *Biophysical Journal* **107,** 1629–1636 (2014).

109. Cross, M. C. & Hohenberg, P. C. Pattern formation outside of equilibrium. *Rev. Mod. Phys.* **65,** 851–1112 (1993).

110. Dennis, G. R., Hope, J. J. & Johnsson, M. T. XMDS2: Fast, scalable simulation of coupled stochastic partial differential equations. *Computer Physics Communications* **184,** 201–208 (2013).

111. Li, G. *et al.* Joint profiling of DNA methylation and chromatin architecture in single cells. *Nature Methods* **16,** 991–993 (2019).

112. Ricci, M. A., Manzo, C., García-Parajo, M. F., Lakadamyali, M. & Cosma, M. P. Chromatin Fibers Are Formed by Heterogeneous Groups of Nucleosomes In Vivo. *Cell* **160,** 1145–1158 (2015).

113. Ou, H. D. *et al.* ChromEMT: Visualizing 3D chromatin structure and compaction in interphase and mitotic cells. *Science* **357** (2017).

114. Xu, J. *et al.* Super-resolution imaging reveals the evolution of higher-order chromatin folding in early carcinogenesis. *Nature Communications* **11,** 1899 (2020).

115. Gillespie, D. T. Exact stochastic simulation of coupled chemical reactions. *The Journal of Physical Chemistry* **81,** 2340–2361 (1977).

116. Kelly, T. K. *et al.* Genome-wide mapping of nucleosome positioning and DNA methylation within individual DNA molecules. *Genome Res* **22,** 2497–506 (2012).

117. Cholewa-Waclaw, J. *et al.* Quantitative modelling predicts the impact of DNA methylation on RNA polymerase II traffic. *Proceedings of the National Academy of Sciences* **116,** 14995–15000 (2019).

118. Di Croce, L. & Helin, K. Transcriptional regulation by Polycomb group proteins. *Nat Struct Mol Biol* **20,** 1147–55 (2013).

119. Lee, T. I. *et al.* Control of developmental regulators by Polycomb in human embryonic stem cells. *Cell* **125,** 301–13 (2006).

120. McLaughlin, K. *et al.* DNA Methylation Directs Polycomb-Dependent 3D Genome Re-organization in Naive Pluripotency. *Cell Reports* **29,** 1974–1985.e6 (2019).

121. Becker, B. *et al.* H/KDEL receptors mediate host cell intoxication by a viral A/B toxin in yeast. *Sci Rep* **6,** 31105 (2016).

122. Wang, C. *et al.* Reprogramming of H3K9me3-dependent heterochromatin during mammalian embryo development. *Nat Cell Biol* **20,** 620–631 (2018).

123. Jörg, D. J. Stochastic Kuramoto oscillators with discrete phase states. *Phys. Rev. E* **96,** 032201 (2017).

124. Kampen, N. V. *Stochastic processes in physics and chemistry* (North Holland, 2007).

125. Kuramoto, Y. *Chemical Oscillations, Waves, and Turbulence* (Dover Publications, 2003).

126. Gupta, S., Campa, A. & Ruffo, S. Kuramoto model of synchronization: Equilibrium and nonequilibrium aspects. *Journal of Statistical Mechanics: Theory and Experiment* **2014** (2014).

127. Smirnov, L., Osipov, G. & Pikovsky, A. Chimera patterns in the Kuramoto–Battogtokh model. **50,** 08LT01 (2017).

128. Pérez, C. J. & Ritort, F. A moment-based approach to the dynamical solution of the Kuramoto model. *Journal of Physics A* **30,** 8095–8103 (1997).

129. Bonilla, L. L., Pérez Vicente, C. J., Ritort, F. & Soler, J. Exactly Solvable Phase Oscillator Models with Synchronization Dynamics. *Phys. Rev. Lett.* **81,** 3643–3646 (1998).

130. Acebrón, J. A., Bonilla, L. L., Pérez Vicente, C. J., Ritort, F. & Spigler, R. The Kuramoto model: A simple paradigm for synchronization phenomena. *Rev. Mod. Phys.* **77,** 137–185 (2005).

131. Daido, H. Order Function and Macroscopic Mutual Entrainment in Uniformly Coupled Limit-Cycle Oscillators. *Progress of Theoretical Physics* **88,** 1213–1218 (1992).

132. Strogatz, S. H., Mirollo, R. E. & Matthews, P. C. Coupled nonlinear oscillators below the synchronization threshold: Relaxation by generalized Landau damping. *Phys. Rev. Lett.* **68,** 2730–2733 (1992).

133. Strogatz, S. H. From Kuramoto to Crawford: exploring the onset of synchronization in populations of coupled oscillators. *Physica D: Nonlinear Phenomena* **143,** 1–20 (2000).

134. Ott, E. & Antonsen, T. M. Low dimensional behavior of large systems of globally coupled oscillators. *Chaos: An Interdisciplinary Journal of Nonlinear Science* **18,** 037113 (2008).

135. Kim, J. K. & Marioni, J. C. Inferring the kinetics of stochastic gene expression from single-cell RNA-sequencing data. *Genome Biology* **14,** R7 (2013).

136. Gene Ontology: tool for the unification of biology. *Nature Genetics* **25,** 25–29 (2000).

137. Derrida, B., Evans, M. R., Hakim, V. & Pasquier, V. Exact solution of a 1D asymmetric exclusion model using a matrix formulation. *Journal of Physics A: Mathematical and General* **26,** 1493–1517 (1993).

138. De Gennes, P. *Scaling concepts in polymer physics* eng (Cornell Univ. Pr., Ithaca [u.a.], 1979).

139. Stevens, T. J. *et al.* 3D structures of individual mammalian genomes studied by single-cell Hi-C. *Nature* **544,** 59–64 (2017).

140. Sasai, M. & Wolynes, P. G. Stochastic gene expression as a many-body problem. *Proceedings of the National Academy of Sciences* **100,** 2374–2379 (2003).

141. Zhang, B. & Wolynes, P. G. Stem cell differentiation as a many-body problem. *Proceedings of the National Academy of Sciences* **111,** 10185–10190 (2014).

142. Assaf, M., Roberts, E. & Luthey-Schulten, Z. Determining the Stability of Genetic Switches: Explicitly Accounting for mRNA Noise. *Phys. Rev. Lett.* **106,** 248102 (2011).

143. Mugler, A., Walczak, A. M. & Wiggins, C. H. Spectral solutions to stochastic models of gene expression with bursts and regulation. *Phys. Rev. E* **80,** 041921 (2009).

144. Hartnett, G. S., Parker, E. & Geist, E. Replica symmetry breaking in bipartite spin glasses and neural networks. *Phys. Rev. E* **98,** 022116 (2018).

145. Brunetti, R., Parisi, G. & Ritort, F. Asymmetric Little spin-glass model. *Phys. Rev. B* **46,** 5339–5350 (1992).

146. Viana, L. & Bray, A. J. Phase diagrams for dilute spin glasses. *Journal of Physics C: Solid State Physics* **18,** 3037–3051 (1985).

147. Castellana, M. & Bialek, W. Inverse Spin Glass and Related Maximum Entropy Problems. *Phys. Rev. Lett.* **113,** 117204 (2014).

148. Tkacik, G., Schneidman, E., au2, M. J. B. I. & Bialek, W. *Spin glass models for a network of real neurons* 2009.

149. Huynh-Thu, V. A. & Sanguinetti, G. Gene Regulatory Network Inference: An Introductory Survey. *Methods in molecular biology* **1883,** 1–23 (2019).

150. Kauffman, S. The large scale structure and dynamics of gene control circuits: An ensemble approach. *Journal of Theoretical Biology* **44,** 167–190 (1974).

151. Font-Clos, F., Zapperi, S. & La Porta, C. A. M. Topography of epithelial–mesenchymal plasticity. *Proceedings of the National Academy of Sciences* **115,** 5902–5907 (2018).

152. Tripathi, S., Kessler, D. A. & Levine, H. Biological Networks Regulating Cell Fate Choice Are Minimally Frustrated. *Phys. Rev. Lett.* **125,** 088101 (2020).

153. Guerra, F. Broken Replica Symmetry Bounds in the Mean Field Spin Glass Model. *Communications in Mathematical Physics* **233,** 1–12 (2003).

154. Talagrand, M. The Parisi formula. English. *Ann. Math. (2)* **163,** 221–263 (2006).

155. Crisanti, A. & Sommers, H. .-J. The sphericalp-spin interaction spin glass model: the statics. *Zeitschrift für Physik B Condensed Matter* **87,** 341–354 (1992).

156. Biroli, G., Bunin, G. & Cammarota, C. Marginally stable equilibria in critical ecosystems. *New Journal of Physics* **20,** 083051 (2018).

157. De Almeida, J. R. L. & Thouless, D. J. Stability of the Sherrington-Kirkpatrick solution of a spin glass model. *Journal of Physics A: Mathematical and General* **11,** 983–990 (1978).

158. Parisi, G., Ricci-Tersenghi, F. & Rizzo, T. Diluted mean-field spin-glass models at criticality. *Journal of Statistical Mechanics: Theory and Experiment* **2014,** P04013 (2014).

159. Cugliandolo, L. & Kurchan, J. The out of equilibrium dynamics of the Sherrington-Kirkpatrick model. *Journal of Physics A Mathematical and Theoretical* **41** (2007).

160. Cugliandolo, L. F. & Kurchan, J. Analytical solution of the off-equilibrium dynamics of a long-range spin-glass model. *Phys. Rev. Lett.* **71,** 173–176 (1993).

161. Cugliandolo, L. F. & Dean, D. S. Full dynamical solution for a spherical spin-glass model. *Journal of Physics A: Mathematical and General* **28,** 4213–4234 (1995).

162. Crisanti, A., Horner, H. & Sommers, H. .-J. The sphericalp-spin interaction spin-glass model. *Zeitschrift für Physik B Condensed Matter* **92,** 257–271 (1993).

163. Huang, M.-C., Wu, J.-W., Luo, Y.-P. & Petrosyan, K. G. Fluctuations in gene regulatory networks as Gaussian colored noise. *The Journal of Chemical Physics* **132,** 155101 (2010).

164. Hornos, J. E. M. *et al.* Self-regulating gene: An exact solution. *Phys. Rev. E* **72,** 051907 (2005).

165. Luo, X. & Zhu, S. Stochastic resonance driven by two different kinds of colored noise in a bistable system. *Phys. Rev. E* **67,** 021104 (2003).

166. Häunggi, P. & Jung, P. in *Advances in Chemical Physics* 239–326 (John Wiley & Sons, Ltd, 1994).

167. Cannoodt, R., Saelens, W. & Saeys, Y. Computational methods for trajectory inference from single-cell transcriptomics. *European Journal of Immunology* **46,** 2496–2506 (2016).

168. Schultz, D., Walczak, A. M., Onuchic, J. N. & Wolynes, P. G. Extinction and resurrection in gene networks. *Proceedings of the National Academy of Sciences* **105,** 19165–19170 (2008).

169. Gardner, T. S., Cantor, C. R. & Collins, J. J. Construction of a genetic toggle switch in Escherichia coli. *Nature* **403,** 339–342 (2000).

170. Strasser, M., Theis, F. J. & Marr, C. Stability and Multiattractor Dynamics of a Toggle Switch Based on a Two-Stage Model of Stochastic Gene Expression. *Biophysical Journal* **102,** 19–29 (2012).

171. May, R. M. Will a Large Complex System be Stable? *Nature* **238,** 413–414 (1972).

172. Wigner, E. P. On the Distribution of the Roots of Certain Symmetric Matrices. *Annals of Mathematics* **67,** 325–327 (1958).

173. Jörg, D. J., Kitadate, Y., Yoshida, S. & Simons, B. D. Stem Cell Populations as Self-Renewing Many-Particle Systems. *Annual Review of Condensed Matter Physics* **12,** 135–153 (2021).

174. Dean, D. S. Langevin equation for the density of a system of interacting Langevin processes. *Journal of Physics A: Mathematical and General* **29,** L613–L617 (1996).

175. Kawasaki, K. & Koga, T. Relaxation and growth of concentration fluctuations in binary fluids and polymer blends. *Physica A: Statistical Mechanics and its Applications* **201,** 115–128 (1993).

176. Robertson, N. & Skeldon, A. *Patterns in a non-local reaction diffusion equation* in (2007).

177. Cates, M. E. & Tailleur, J. When are active Brownian particles and run-and-tumble particles equivalent? Consequences for motility-induced phase separation. *EPL (Europhysics Letters)* **101,** 20010 (2013).

178. Cugliandolo, L. F. *Course 7: Dynamics of Glassy Systems* in *Slow Relaxations and nonequilibrium dynamics in condensed matter* (Springer Berlin Heidelberg, Berlin, Heidelberg, 2003), 367–521.

179. Galla, T. Dynamically evolved community size and stability of random Lotka-Volterra ecosystems. *EPL (Europhysics Letters)* **123,** 48004 (2018).

180. Opper, M. & Diederich, S. Phase transition and 1/f noise in a game dynamical model. *Phys. Rev. Lett.* **69,** 1616–1619 (1992).

181. Biscari, P. & Parisi, G. Replica symmetry breaking in the random replicant model. *Journal of Physics A: Mathematical and General* **28,** 4697–4708 (1995).

182. Roy, F., Biroli, G., Bunin, G. & Cammarota, C. Numerical implementation of dynamical mean field theory for disordered systems: application to the Lotka–Volterra model of ecosystems. *Journal of Physics A: Mathematical and Theoretical* **52,** 484001 (2019).

183. Houchmandzadeh, B. Theory of neutral clustering for growing populations. *Phys. Rev. E* **80,** 051920 (2009).

184. Houchmandzadeh, B. Clustering of diffusing organisms. *Phys. Rev. E* **66,** 052902 (2002).

185. Nurse, P. Biology must generate ideas as well as data. *Nature* **597,** 305 (2021).

186. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17,** 10–12 (2011).

187. Krueger, F. & Andrews, S. R. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* **27,** 1571–2 (2011).

188. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nature Methods* **12,** 357–360 (2015).

189. Novikov, E. A. Functionals and the random-force method in turbulence theory. *J. Exptl. Theoret. Phys. (U.S.S.R.)* **41** (1964).