SISSA
Scuola
Internazionale
Superiore di
Studi Avanzati

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

Spatial organization of DNA across scales: a Bayesian inference and information-theoretic approach

Candidate:
Alex Chen Yi Zhang

Advisors:
Angelo Rosa
Guido Sanguinetti

Academic Year 2023-24

Abstract

Spatial organization plays a fundamental role in many biological processes, from the folding of chromatin within the nucleus and the patterns of gene expression levels that guide cell fate decisions within an embryo during development. Understanding how the spatial arrangement of genetic material is connected to biochemical processes and cellular function is crucial for deciphering mechanisms of gene expression and development. This thesis presents an investigation across multiple scales of these relationships, focusing first on chromatin organization and nuclear biochemistry, and then extending to the properties of gene expression patterns that guide tissue development.

In Chapter 2, SEMPER, a polymer-based model, is introduced to quantify the role of chromatin's polymeric nature in shaping biochemical landscapes. By integrating one-dimensional biochemical data, such as transcription factor binding and histone modifications, into a three-dimensional linear scaffold, SEMPER highlights the influence of chromatin's physical structure on epigenetic patterning.

Chapter 3 shifts perspective and approaches the problem from a complementary angle by introducing bioSBM, a stochastic block model designed to predict chromatin structure starting from epigenetic features. This model provides an interpretable and generalizable framework for linking biochemical data to large-scale chromatin folding. BioSBM successfully predicts chromatin interactions across different cell types and conditions, suggesting that the relationships between epigenetic marks and chromatin structure are broadly conserved.

In Chapter 4, the thesis expands to the tissue level, examining how spatial gene expression patterns along embryonic axes guide cellular differentiation and tissue organization. The concept of correlational information (CI) is introduced to quantify the degree of spatial correlations in gene expression fluctuations and is shown to help cells improve their positional accuracy during development. A non-local decoding framework is proposed, suggesting that cells can leverage information from neighboring cells to refine their own positional decisions.

Acknowledgement

First of all, I'd like to thank my PhD advisors Guido and Angelo for their support and guidance in my research. Working with two people with very distinct fields of expertise was stimulating and certainly not devoid of challenges, but you gave me the freedom and the trust to explore my own research directions and allowed me to make my own mistakes, which is a process through which I certainly learned a lot. Thank you Guido also for the numerous chats about climbing and mountains and for lending me Ptashne's book, which I have to admit I still have to finish. I commit to reading it in the next months, and then you'll finally have it back.

I also thank Gašper Tkačik for welcoming me into his group and for the stimulating discussions together with David Brückner.

Thanks to the gang of office 323: Romina, Eddy, Gigi, Nour, and Sara. It wasn't easy to start a PhD in the middle of a pandemic, but together we managed to make it fun. A special thank goes of course to Eddy for sharing office 418 with me, for being a good friend and for making our time at work feel less like actual work. I am sure that your future is bright and I am so excited for your next adventure in Torino with Lucia. Thank you office 420 (Vitto, Santi, Sam, Matte Forte, Matte Scarso) for welcoming me into your fun office.

Sono profondamente grato alle persone con cui ho condiviso una casa in questi anni passati a Trieste. Iniziamo con Albi, grazie per tutti i caffè turchi, per il pane e per avermi fatto uscire dalla "*bolla*" della SISSA e per avermi fatto conoscere un'altra faccia della città. Grazie a Gigi: anche se non abbiamo vissuto insieme a Trieste, prima di questa città abbiamo condiviso tante case e tanta vita. Grazie Ross, per condividere con me la tua passione per il cibo e per la tua leggerezza, Enrico per tutte le partite a scacchi, e Pennacchio, beh per essere Pennacchio. Grazie Michi per la tua infinita energia nerd e la tua curiosità, e per aver portato la Rota-board a casa. Grazie per avermi fatto amare i Vulfpeck ancora più di quanto non li amassi prima, e grazie per i tentativi su "*La Ruggine*" insieme. Prima o poi la chiuderemo. Grazie Franci per avermi portato in montagna quando ero solo uno sprovveduto ragazzo di

pianura. La nostra prima scampagnata, quella in cui ho perso la borraccia blu e in cui ci siamo infradiciati i calzini, me la ricorderò sempre con tantissimo affetto. E soprattutto, grazie per ridere alle mie battute. Che pesci sono quelli?

Dopo aver ringraziato le persone con cui ho vissuto in questi anni, devo anche menzionare tutti quelli, che pur non avendo condiviso fisicamente quattro mura e un tetto con me, hanno vissuto con me qualche momento in questa città che ormai posso chiamare casa. Grazie Marco, Cri, Andre, Albi di nuovo, Anna, Max, Dario, Fede, Vero e tutti gli altri scoppiati dei Megadito.

Per restare in tema, ringrazio le persone che per prime hanno condiviso una casa con me: la mia famiglia. Voglio iniziare da mio fratello Calvin. Grazie Calvin per essere una persona che guardo con tanta ammirazione, anche se il fratello maggiore sarei io. Hai certamente ereditato l'etica del lavoro dei nostri genitori e sono fiero di te. Grazie per il tempo passato insieme da bambini; il periodo che ricordo con più affetto è quando, ancora a Nonantola, io venivo a prenderti alla fermata del bus dopo scuola e passavamo le ore seguenti a giocare a Prince of Persia sul computer di casa. Poi, voglio ringraziare anche mamma e papà:

谢谢妈妈，谢谢爸爸。真的非常感谢你们给予的一切。感谢你们这些年来对我的支持，以前我一直觉得，我们是不同的人，对生活的看法也有很大的不同。但是随着我长大，我渐渐明白，我们之间的一些差异源于你们所做的那些牺牲，而这些牺牲是我无法想象的。这一切都归功于你们，你们从未让我和弟弟感到缺少过什么，哪怕是每周工作70个小时的代价。自从你们三十年前来到意大利，为了我们的家庭，你们从未停止努力，从未休息过。我知道在这样半生不停地工作之后，要放松下来并不容易，我非常理解这一点。我希望今天，你们看到你们帮助我取得的成就之后，能够决定从此为未来好好享受应得的休息。我爱你们。

你们的大儿子。

Infine, vorrei ringraziare Eugi. Grazie per avermi invitato a quell'aperitivo a casa tua, per non avermi lasciato andare via, per avermi fatto assaggiare il salame con il mascarpone. Grazie perché ami la pizza almeno quanto la amo io. Grazie per le vacanze avventurose on the road, ma anche per le mattinate passate a guardare i film con la colazione a letto. Grazie per esser stata una presenza costante da quando ci siamo conosciuti. Non vedo l'ora di esser tuo coinquilino e di vivere la nostra prossima avventura insieme a Vienna.

Contents

1 Introduction **1**

1.1 The different scales of biology 1

1.1.1 Molecular and nuclear level organization 1

1.1.2 Tissue level organization 2

1.2 Chromatin organization: an overview 3

1.2.1 Experiments 3

1.2.2 Modeling 5

1.3 Bayesian modeling and computation 9

1.3.1 Information Theoretic Tools 9

1.3.2 Graphical models 10

1.4 Computational Methods in Bayesian Inference 13

1.4.1 Sampling Methods 13

1.4.2 Variational Methods 14

2 SEMPER **17**

2.1 Materials and methods 19

2.1.1 *SEMPER*: an integrative model of chromatin organisation . . 19

2.1.2 Estimating model parameters for *SEMPER*: approximate Monte Carlo method 22

2.1.3 Datasets used as inputs for the model and experimental Hi-C data 25

2.2 Results: Epigenetic state 26

2.2.1 Predictions of epigenetic states 26

2.2.2 Influence of polymeric terms in estimating the parameters . . 28

2.2.3 Bottom-up modeling leads to realistic interaction patterns . . 29

2.3 Discussion and conclusions 34

3 bio-SBM **37**

3.1 Stochastic Block Models 39

3.1.1 Generative process of bioSBM 42
3.1.2 Variational inference 43
3.2 Results 44
3.2.1 bioSBM explains the hierarchical organization of the chromatin in terms of epigenomic marks 46
3.2.2 bioSBM predicts the structure of the chromatin in unseen cell lines 48
3.3 Discussion and conclusions 50

4 Correlational information **53**
4.1 What does it mean for a cell to know its position? 56
4.1.1 Positional Information 56
4.1.2 Decoding maps and positional error 57
4.2 Correlational information and patterning utility 60
4.3 Correlational information and non-local decoding 64
4.3.1 Information content of non-local decoder 69
4.4 Non-local decoding mechanisms 71
4.5 Discussions and Conclusions 74

5 Conclusions **77**

A Supplementary material of Chapter 2 **81**
A.1 Model and Methods 81
A.1.1 The polymer model 81
A.1.2 Monte Carlo method 82
A.1.3 Inference method *via* approximate MC sampling 83
A.2 Inclusion of the experimental data in the model 86
A.3 Validation test from synthetic data 86

B Supplementary material of Chapter 3 **95**
B.1 A VEM algorithm for posterior nference 95
B.1.1 Variational E-step 96
B.1.2 Variational M-step 99
B.2 Daset preprocessing 100
B.3 Simulation of predicted maps 101
B.4 Supplemental Table and Figures 102

C Supplementary material of Chapter 4 **105**
C.1 Derivation of U = PI + CI 105
C.2 Structure of covariance matrix of gene expression level fluctuations . . 106

List of Figures

1.1 Experimental breakthroughs for chromatin organization 4
1.2 Co-polymers, epigenetic patterning and magnetic polymers 7
1.3 Chromatin networks 8
1.4 Graphical models 11
1.5 Hierarchical Bayesian Model for Latent Dirichlet Allocation (LDA) . 12

2.1 Schematic illustration of the *SEMPER* model and the inference pipeline 21
2.2 *SEMPER* recapitulates histone modification patterns. 27
2.3 Polymer and spin configurations from varying parameter values. . . . 29
2.4 Posterior distributions for *SEMPER* 30
2.5 The weights magnitudes inferred by including a polymeric term are smaller than the case logistic regression case 31
2.6 Polymer Importance reduces with higher level of coarse-graining . . . 32
2.7 De novo generated contact maps from SEMPER posterior 33

3.1 Hierarchical model describing the generative process of bioSBM . . . 42
3.2 k-means of membership vectors: biological interpretation of block model latent space 45
3.3 Enrichment in biochemical features for k-mean clusters of bioSBM membership vectors 46
3.4 a) Enrichment of mixed membership communities in the biochemical features used in the model. Each community correlates with distinct epigenetic features. b) Interaction intensities between all pairs of communities. c) Membership proportions of communities 4 and 5 as a function of node indices. 47
3.5 bioSBM testing pipeline and example of predicted map for chromosome 19 48
3.6 Predictive performance of bioSBM on unseen chromosomes and cellular conditions. Comparison with a state-of-the-art polymer model 50

4.1 Examples of two ensembles of morphogen profiles carrying different PI 57
4.2 Decoding maps 58
4.3 Spatial correlations of fluctuations on gene expression profiles 63
4.4 Positional information, correlational information, and total patterning utility as a function of the noise parameters 64
4.5 Schematic representation of optimal local and non-local decoding . . 65
4.6 Non-local decoding: effective morphogen gradients are affected by smaller positional error 68
4.7 ΔPI: information gains through non-local decoding 70
4.8 Linear filters that approximate the optimal non-local decoding 72

A.1 Discretization of ChIP-seq tracks 88
A.2 *Posterior distribution for synthetic data* 89
A.3 Comparison between approximate inference methods 90
A.4 Posterior comparison for uncoupled spins 91
A.5 Posterior comparison for coupled spins 92
A.6 Average contact probabilities 93
A.7 Standard deviations of contact probabilities 94

B.1 Hierarchical model describing the generative process of bioSBM . . . 95
B.2 Model selection for block model 102
B.3 The importance of having mixed-memberships: Shannon entropies of MAP estimates of membership vectors, and better recapitulation of biological annotations 103
B.4 Correlations between Hi-C maps: a comparison between the parametrization of Γ in terms of a linear transformation of a neural network 103

1

Introduction

1.1 The different scales of biology

Biological systems are structured across different scales, from molecular interactions within the cell to the coordinated development of entire organisms. This thesis delves into two key aspects of this intricate organization. We begin by studying chromatin interactions, focusing on how the three-dimensional arrangement of DNA within the nucleus is tied to the complex biochemical environment of the cell. Understanding these spatial relationships is key to deciphering how chromatin organization influences cellular function. In the second part, we shift our attention to developmental biology, examining how cells interpret positional information to guide their differentiation and organize into tissues.

1.1.1 Molecular and nuclear level organization

At the molecular level, one of the central objects is the DNA sequence (the genome) which constitutes the code or blueprint of our biological makeup. The complete sequencing of the human genome has been achieved for more than twenty years now [1] however, this knowledge is not enough to understand how our genetic code is read and interpreted: all of our somatic cells share the same DNA sequence but they can differentiate into diverse cell types that have very different functions and characteristics. Indeed our chromosomes are not naked DNA, but they form a tight complex with histone octamers, which can carry posttranslational modifications that affect the binding affinity of specific proteins such as transcription factors (TFs). A variety of such epigenetic factors has been described and shown to play a pivotal role in gene expression, differentiation, and disease [2, 3, 4, 5, 6]. Genetic and epigenetic factors

(such as histone modifications and binding of TFs) can be probed with biochemical assays such as chromatin immunoprecipitation followed by sequencing (ChIP-Seq)[7] which provides us with genome-wide 1D maps of protein-DNA interactions.

These interactions occur on a scaffold made of linear chromosomes (46 in diploid human cells) of roughly 2 meters in total length, tightly packed into a nucleus of 5 to 10 μm of width. The precise way in which the chromatin fiber folds within the nucleus is thought to affect the expression of some genes, notably thanks to the modulation of the interaction between cis-regulatory elements through CTCF-mediated loops and topologically associating domains (TADs)[8, 9, 10, 11]. On the other hand, there is also a growing body of studies that show that this connection may be more nuanced than previously thought [12, 13, 14, 15], as it was observed that the disruption of the 3D organization of chromatin does not necessarily affect gene expression significantly on a global level. This indicates that the strength and direction of the link between biological function and 3D genome architecture still require further study.

Motivated by these reasons in the first part of my thesis I will study the biological role of the spatial organization of the genetic material in the nucleus. First, in **Chapter 2** I will detail the development and application of a novel Bayesian inference framework to assess the importance of the polymer scaffold which is chromatin in mediating biochemical signals. This will highlight the importance of the polymeric nature of chromatin in the establishment of patterns of epigenetic modifications. Second, in **Chapter 3** I will develop a random graph model to describe long-range chromatin interactions and compartmentalization. This probabilistic model links biochemical features to distinct chromatin structural types, providing insights into how biochemical processes influence large-scale chromatin organization.

In section 1.2 of this chapter, I will briefly review the experimental assays that allow us to probe chromatin spatial arrangement and introduce some paradigmatic ways to model this organization.

1.1.2 Tissue level organization

Zooming out and looking at the formation of multicellular systems we will observe that there is an extra layer of complexity to the execution of our genetic code. Embryonic development crucially depends on fate choices of differentiating stem cells, which must be tightly regulated to ensure the precise and reproducible generation of distinct cell types, in correct proportions at the right time and place. The various differentiation events must be coordinated within the population of cells through cell-cell signaling and tissue mechanics, however, the exact way in which this orchestration happens is

still a matter of active research.

More than five decades ago, Lewis Wolpert introduced the concept of *Positional Information* (PI), which describes the fact that cells acquire distinct cell fates based on their position within a tissue. The formalization, many years later, of PI in the mathematical framework of information theory and the development of precision in-vivo microscopy techniques allowed scientists to make quantitative statements about the fundamental limits of the precision attainable by developing embryos [16, 17, 18, 19].

In **Chapter 4** I first review some concepts of developmental biology and the mathematical framework of positional information. Then, I will introduce and characterize an information theoretical quantity that we dub correlational information (CI) that quantifies the spatial correlations in fluctuations of gene expression levels within a developing embryo. Finally, I will discuss in which way cells in a forming tissue can take advantage of these correlations to enhance their developmental precision.

Chapter 5 then concludes with a summary of findings and an outlook on future research.

1.2 Chromatin organization: an overview

1.2.1 Experiments

The road to our current understanding of chromosomes' 3D organization was paved by significant technological advances [20]. Here, I provide a short overview of the key breakthroughs that have fundamentally reshaped our comprehension of how our genome is arranged within the nucleus.

In the late 1800s, the usage of simple light microscopes combined with DNA-binding dyes allowed the observation of distinct mitotic chromosomes. However, during interphase, the dye filled the entire nucleoplasm, leading to the assumption that chromosomes in interphase were randomly organized. This view changed with the advent of FISH (fluorescent in situ hybridization) in the early 1980s [21], a technique that enabled the visualization of the 3D position of specific chromosomal loci. The development of FISH techniques and the simultaneous use of multiple fluorochromes (chromosome painting) confirmed that interphase chromosomes remain segregated in distinct domains known as *chromosome territories* [22, 23, 24].

A major leap forward occurred roughly twenty years ago when Dekker et al [25] developed Chromosome Conformation Capture (3C), a technique that maps in vivo

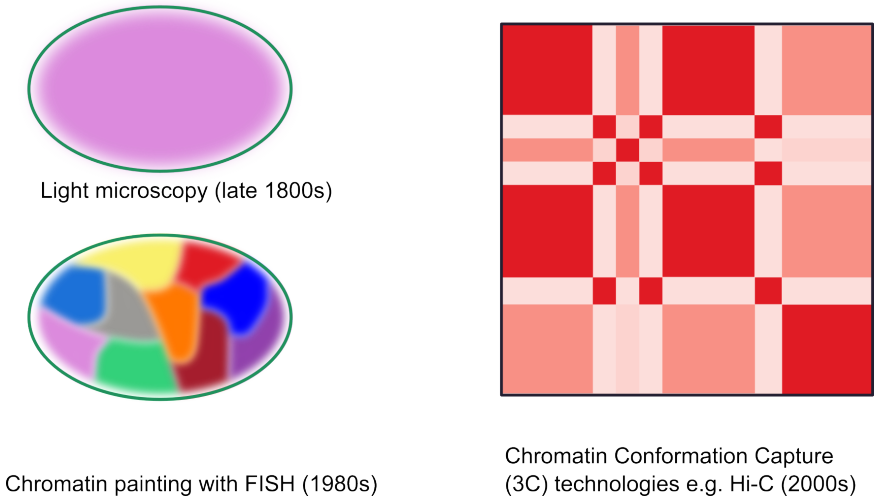


Figure 1.1: A graphical depiction of the evolution of our knowledge about chromatin organization. We went from randomly organized interphase chromosomes in the late 1800s, to a territorial view of chromosomes thanks to FISH experiments. Chromosome conformation capture methods allowed us to have a much more detailed understanding of the folding of our genomes.

interactions between linearly distant chromatin loci that come into 3D spatial proximity within a population of cells. Over the years, numerous variants and derivatives of 3C have been developed, each with technical differences, but all share three fundamental steps: cells are first crosslinked to preserve spatial proximities of DNA loci, then chromatin is fragmented (often with restriction enzymes), and the resulting fragments are isolated and sequenced to create a map of 3D interactions. Among these, Hi-C [26] stands out by combining conformation capture with high throughput sequencing to generate genome-wide contact maps at high resolution. These maps are represented as 2D matrices, where rows and columns correspond to distinct genomic regions, and the entries reflect the contact frequencies between pairs of regions (Figure 1.1).

The introduction of Hi-C provided a different but complementary perspective to microscopy data on chromatin organization. Hi-C confirmed the existence of chromosome territories, as it revealed significantly more intrachromosomal contacts compared to interchromosomal ones. Moreover, Hi-C unveiled that interphase chromosomes are segregated into two major compartments, termed A and B, which are structurally and functionally distinct. On a finer scale, it revealed the presence of sub-megabasepair domains called topologically associating domains (TADs), which are enriched in chromatin contacts. Another remarkable feature identified in Hi-C maps is the presence of off-diagonal high-intensity dots, which are now recognized as signatures of chromatin looping.

Hi-C and its derivatives belong to a class of techniques known as *non-enrichment*

methods, which assay chromatin contacts across the entire genome. To map chromatin folding at a finer scale and to better understand the functional implication of 3D architecture, *enrichment approaches* have been developed. These methods amplify the signal of contacts involving specific proteins (e.g. ChIA-PET [27, 28]) or specific genomic regions (e.g. promoter capture Hi-C, PCHi-C [29, 30, 31]), allowing for detailed mapping of chromatin contacts that would otherwise require up to 20-50-fold more sequencing depth without enrichment.

While traditional FISH protocols inherently determine the positions of specific chromatin loci within nuclei of single cells, their low resolution and limited throughput can restrict their usefulness. On the other hand, sequencing-based techniques like Hi-C offer genome-wide maps with high resolution but sacrifice a big part of the spatial information provided by microscopy. Recently, advances in both super-resolution and high-throughput FISH protocols have emerged to bridge this gap. Techniques like STORM (Stochastic Optical Reconstruction Microscopy) and PALM (Photoactivated Localization Microscopy) [32, 33] surpass the diffraction limit of conventional light microscopy, enabling the observation of chromatin organization with nanometer precision in single cells. Additionally, high-throughput methods such as multiplex FISH [34] are being developed to significantly increase the number of loci that can be analyzed simultaneously. These innovations promise to reconcile the distinct, yet complementary, perspective provided by traditional microscopy and sequencing-based approaches, with the potential of offering a more comprehensive understanding of chromatin architecture.

1.2.2 Modeling

Because of the stochastic and variable nature of chromosome folding *in vivo*, experimental techniques like 3C and FISH, while powerful, provide data that can be challenging to interpret on their own. These methods yield valuable insights but often require theoretical and computational models to fully understand the underlying principles of chromatin organization. Theoretical and computational models provide a valuable framework for making sense of experimental data, allowing researchers to integrate diverse findings, fill in knowledge gaps, and generate predictive insights into the relationship between chromatin structure and function [35]. There exist various ways to model chromatin's spatial organization, many of which can be broadly categorized into two classes: 1) polymer models that explicitly describe chromatin as a filament, and 2) abstract models based on graphs or networks that represent chromatin contacts. Both classes of models are useful as they capture different and complementary aspects of chromatin organization.

Polymer modeling

The study of chromatin organization has benefited immensely from polymer physics [36, 37]. A polymer is in essence a very long one-dimensional chain made of small units called monomers, and depending on whether the single units are identical or not we call it homopolymer or heteropolymer respectively. Models and mechanisms with varying degrees of complexity and levels of detail have been proposed to explain several aspects of chromatin architecture [38, 39]. If we neglect all the details and the heterogeneity associated with DNA sequence and epigenetic modifications, chromosomes form at first approximation a solution of very long and thin homopolymeric chains. Despite the abstraction and the lack of details of these models, the laws of polymer physics are able to some extent to explain the behavior of chromatin organization. For instance Rosa and Everaers [40] explain with extremely minimal polymer simulations that the presence of chromosome territories may be caused by the extremely long time-scales of mitotic chromosomes de-condensation, due to the numerous topological constraints to which polymer chains are subject to. Many other studies show how simple models of homopolymers in solution can recapitulate notable chromosome features [41, 42, 43], however the biological reality contains specificities, such as epigenetic effects, that cannot be accounted for by simple homopolymers. Different heteropolymer models may be better suited depending on the kind of questions one may be interested in. The two essential ingredients of a heteropolymer are 1) the shape of the polymeric scaffold and 2) the types of different monomers. The shape or conformation of the chain describes the spatial arrangement of chromatin within the nucleus, while the type of the single monomers model the different epigenetic states of a certain stretch of DNA. The different states or colors of the monomers can either represent some concrete epigenetic modification such as the presence or absence of a histone modification, or some abstract representation of chromatin epigenetic state [44].

First, several models have been proposed the explain the folding of chromatin with a fixed pattern of epigenetic marks, embodying the widespread idea that the interaction between two stretches of DNA may depend on their epigenetic marks and that the heterogenous patterns of these marks contribute to the non-trivial and non-random motifs in chromosomes spatial organization. Notable examples of such models are block co-polymers [46] and strings-and-binders models [47, 48]. These kinds of models do not take into account that epigenetic marks are dynamically deposited and are transient [49]. To account for this, another class of models studied how patterns of epigenetic modifications self-organize and spread, given the interaction probability between different loci [50, 51]. In order to study the direction and the strength of

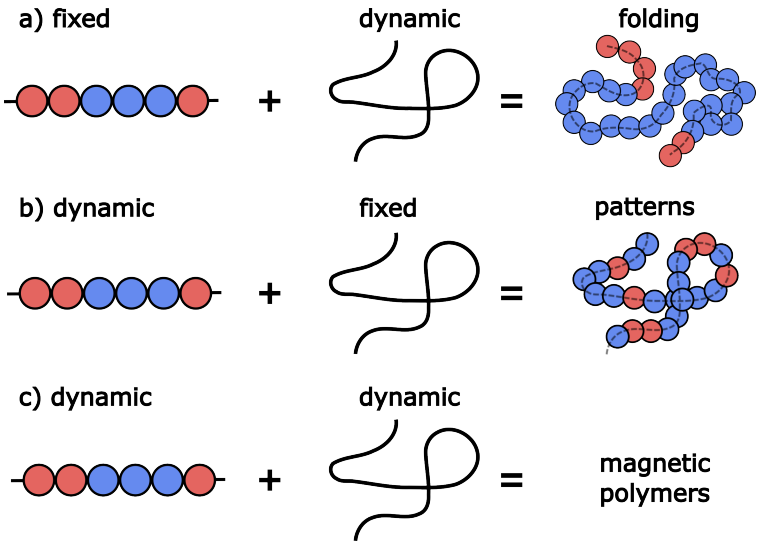


Figure 1.2: Classes of heteropolymer models to describe folding of chromatin and spreading of epigenetic patterns. (a) polymers with a quenched pattern of epigenetic marks model the effect of biochemical modification on the 3D structure of chromosomes (b) Ising-like systems embedded on a quenched substrate model the effect of structure on epigenetics (c) magnetic polymers allow the simultaneous modeling of structure and biochemistry. Figure adapted from Colì et al (2019). [45]

the link between structure and biochemistry, it should be natural to consider both the polymer conformation and the epigenetic marks pattern as not fixed, and the dynamical interplay between these factors. These kinds of models are called magnetic polymers or annealed copolymers (as opposed to quenched copolymers) [52, 45, 53]. The model I study in Chapter 2 belongs to this class and I develop a Bayesian inference framework to quantify the relative importance of the contribution from the structural polymeric nature of chromatin and its biochemical properties.

Network modeling

Alongside polymer models, complementary network/graph theory approaches have gained traction in the last years to describe the genome and the epigenome. Networks are a powerful tool to describe relational data, in which many objects or entities are connected in a non-trivial way [54]. In network theory, the interacting objects are called nodes and the interactions between nodes are represented by edges. When modeling chromatin as a network, the nodes represent fragments of DNA (just like the monomers in polymer models) and edges represent contacts or 3D spatial proximity as probed by chromosome capture data. For instance, we could think of a Hi-C map as the adjacency matrix of a graph representing chromatin contacts (See Fig. 1.3).

While polymer models are more realistic in the fact that they explicitly take into

account (to varying degrees of detail) the physical nature of the linear chromosomes, network approaches offer additional advantages and therefore complement the physics-based approaches.

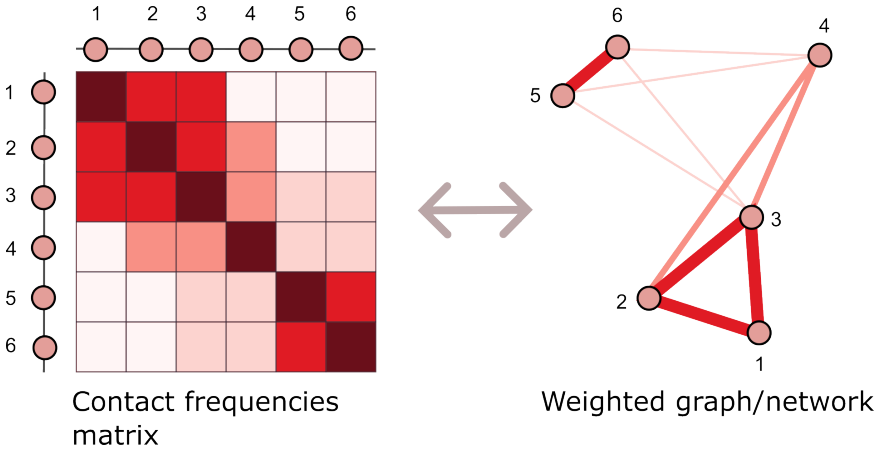


Figure 1.3: Chromsomes, being linear chains, are naturally well-described by polymer physics. An alternative representation of the 3D organization of chromatin is given by networks, where each node corresponds to a genomic region and edges represent contacts. The adjacency matrix of a graph encodes for all pairs of nodes the intensity of their interactions

First of all the vast literature on graph theory provides us with an extensive toolbox to analyze relational data. These tools can be used to extract patterns from conformation capture data that would otherwise be hard to see. One notable example is [55] where the authors used the network theory toolbox to derive an algorithm for precise TAD detection. Furthermore, a description of chromatin in terms of a network may help in identifying precise links between structure and function, both in terms of general principles that regulate our genomic machinery and in the context of clinical applications. This is granted by the simplicity, in the network formalism, in which one can integrate the structural data from conformation capture experiments with 1D data from epigenetic assays, by associating each node in the chromatin network to a vector of biochemical features [56, 57]. In Chapter 3 I develop a generative probabilistic model based on the stochastic block model (SBM) [58, 59] to describe long-range interactions and compartmentalization of the genome. The model allows to perform rigorous statistical inference and reveals a lower dimensional representation of chromatin interactions recapitulating known biology. Furthermore, the model is generative in nature and allows us to predict long-range chromatin patterns starting from biochemical assays.

1.3 An introduction to Bayesian modeling and computation

In this section, I provide a concise overview of methods of Bayesian modeling and computation that will be useful throughout this thesis. Bayesian modeling provides a robust and rigorous framework for statistical inference, parameter estimation and uncertainty quantification. This approach enables the integration of prior knowledge with new data, allowing us to update our beliefs about unknown parameters in a systematic way. Bayesian methods treat parameters as random variables and use probability distributions to quantify uncertainty about them. The cornerstone of Bayesian statistics is Bayes' theorem:

$$P(\theta|y) = \frac{P(y|\theta)P(\theta)}{P(y)} \tag{1.1}$$

which allows us to update our belief about the parameters θ given observed data y. Here, $P(y|\theta)$ is the likelihood (or the model), the probability of observing the data given the parameter θ and $P(\theta)$ is the prior distribution, representing our initial belief about θ before observing any data. The theorem combines prior and likelihood, through the normalization constant $P(y)$, called the marginal likelihood or evidence, to get the posterior distribution $P(\theta|y)$, representing our updated belief about θ after observing data y

1.3.1 Information Theoretic Tools

Because we will work with probability distributions throughout the thesis I introduce here some useful information-theoretic quantities, used to describe the properties of these distributions.

Shannon entropy [60, 61] for a random variable X is defined as

$$S[X] = -\sum_x P(x) \log_2 P(x) \tag{1.2}$$

It was originally introduced in the context of communication theory. Suppose we have a signal x sampled from a random source with probability distribution $P(x)$, then $S[X]$ tells us how many bits are typically needed to encode such a signal. Even outside the context of communication, Shannon entropy can be used to quantify uncertainty and randomness.

The Kullback-Leibler divergence or relative entropy between two probability dis-

tribution $P(x)$ and $Q(x)$ reads

$$D_{KL}[P(x)||Q(x)] = \sum_x P(x) \log_2 \frac{P(x)}{Q(x)} \tag{1.3}$$

It is a measure of how different the two distributions are, which is always greater than or equal to 0. $D_{KL}[P||Q] = 0$ only if $P(x) = Q(x)\ \forall x$.

Finally, another important quantity is the Mutual Information between two random variables X and Y

$$I(X,Y) = \sum_{x,y} P(x,y) \log_2 \frac{P(x,y)}{P(x)P(y)} = D_{KL}[P(x,y)||P(x)P(y)] \tag{1.4}$$

It quantifies the statistical dependence between the two random variables. $I(X,Y) = 0$ if X and Y are independent, and it is positive otherwise.

1.3.2 Graphical models

Graphical models [62] are valuable depictions of independence and dependence relationships within complex probabilistic models involving multiple random variables. In particular, when the underlying graph is a directed acyclic graph (DAG), the model is referred to as a *Bayesian network*. These networks are used to represent the factorization of a joint probability distribution into a product of conditional distributions, which simplifies computations and highlights dependencies. For instance, consider a probability distribution over x_1, x_2 and x_3. This distribution can be factorized as:

$$P(x_1, x_2, x_3) = P(x_1|x_2, x_3)P(x_2|x_3)P(x_3) \tag{1.5}$$

which is represented graphically in Figure 1.4 (a). In this figure, arrows indicate conditional dependencies, with an arrow from one node to another signifying that the probability of the receiving node is conditioned on the sending node. For instance, node 1 has incoming arrows from nodes 2 and 3, which reflects the factor $P(x_1|x_2, x_3)$.

While any valid factorization of a joint probability distribution can be mapped to a graph, not all graphs correspond to valid factorizations. For example, the graph in Figure 1.4(b) represents the factorization $P(x_1, x_2, x_3) = P(x_1|x_2)P(x_2|x_3)P(x_3|x_1)$, which is invalid because it contains cyclic dependencies. Hence, only directed acyclic graphs (DAGs) are permitted in Bayesian networks.

Bayesian networks are particularly useful for visualizing conditional independence relationships. For instance in Figure 1.4 (d) nodes 1 and 2 depend only on node 3. This corresponds to the conditional independence $x_1 \perp\!\!\!\perp x_2|x_3$ (where $\perp\!\!\!\perp$ denotes

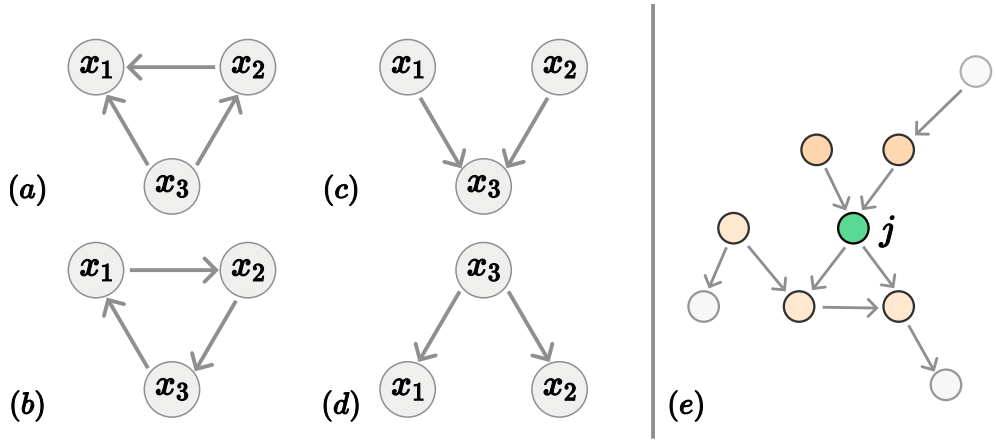


Figure 1.4: (a),(c) and (d) are simple examples of Bayesian networks over three random variables. (b) is an example of a graph representing an invalid factorization, due to the presence of a cycle (loop). Fig.(e) highlights for an example node j (in green), its Markov blanket $MB(x_j)$ in light orange.

statistical independence). In contrast, this relationship does not hold in Figure 1.4 (c). These intuitive examples can be formalized using the concept of *Markov Blanket* (MB). The Markov Blanket of a node x_j is defined as the set of variables that renders x_j conditionally independent from the rest of the variables in the network, given its blanket. Formally:

$$MB(x_j) = \{x_i, i \in \mathcal{J}\} \qquad \text{s.t.} \quad x_j \perp\!\!\!\perp x_k | MB(x_j) \quad \forall k \notin \mathcal{J} \tag{1.6}$$

and it can be shown that the Markov Blanket of x_j consists of:

$$MB(x_j) = Pa(x_j) \cup Pa^{-1}(x_j) \cup coPa(x_j) \tag{1.7}$$

where $Pa(x_j)$ refers to the *parents* of node j, (the nodes with edges directed towards j), $Pa^{-1}(x_j)$ refers to the *children* of j (the nodes with edges directed away from j) and $coPa(x_j)$ are the *co-parents*, (nodes that share a child with j). Identifying the Markov Blanket for each variable in complex models helps simplify computations and often reduces the complexity of inference. This is illustrated in Figure 1.4 where the MB of node j is highlighted in orange, allowing us to quickly see that given the colored nodes, the remaining nodes become irrelevant for determining the probability distribution of x_j.

Hierarchical Bayesian Models

A particularly useful type of Bayesian network is the so-called Hierarchical Bayesian models (HBMs), which offer a robust framework for modeling complex data structures

by introducing multiple *layers* of parameters. These models excel at capturing dependencies and variability across different levels of hierarchy, such as groups, individuals, or repeated measurements. Their strength lies in their ability to accommodate intricate statistical relationships while preserving interpretability through their clear probabilistic semantics. Each parameter in an HBM has a well-defined meaning, and the hierarchical structure facilitates the integration of information at various levels. This balance between flexibility and interpretability makes HBMs particularly valuable for uncovering associations in biological processes [63, 64, 65, 66].

Example: Latent Dirichlet Allocation

I present here a prototypical example of a Hierarchical Bayesian Model, the Latent Dirichlet Allocation (LDA) model [67], which has been adapted for applications in studying transcriptomics and DNA accessibility data [65, 66]. LDA is a generative statistical model originally designed to uncover latent structures in text data. It seeks to identify hidden topics within a collection of documents and associate words with these topics, thereby disambiguating their meanings based on context.

In LDA, each document is assumed to be a mixture of multiple topics, where each topic is characterized by a distribution over words. For instance, in a corpus of scientific papers, topics could represent different research areas, and the words associated with each topic would be terms commonly used in that area. The goal of LDA is to infer these topics from the observed data and determine the proportion of each topic within each document.

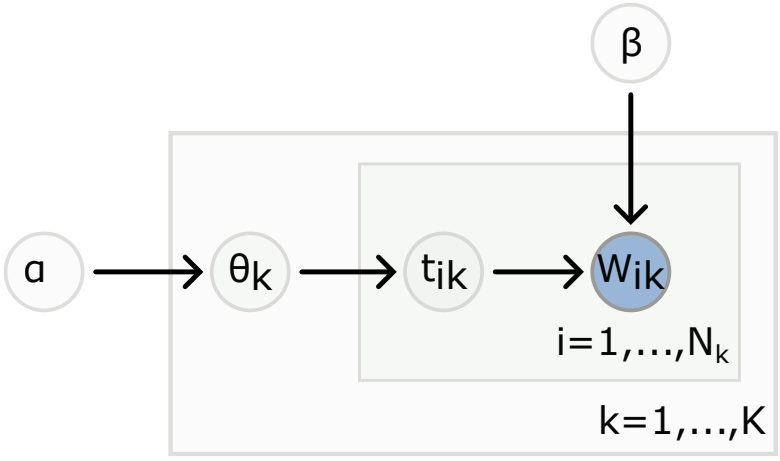


Figure 1.5: Hierarchical Bayesian Model for Latent Dirichlet Allocation (LDA)

The hierarchical structure of LDA involves three main components:

- Document-specific topic distribution θ_k. For each document, a distribution over K latent topics is drawn from a Dirichlet prior with hyperparameter α. This

distribution represents the document's mixture of topics.

- Topic assignments t_{ik}. Each word in each document is assigned a topic based on the document's topics distribution.
- Observed words w_{ik}. Each word in a document is generated according to the topic assignments and the probabilities of words given the topic β.

In Figure 1.5 you can see the depiction of LDA in plate notation, which is a convenient way to graphically represent the hierarchy of variables involved in the model. The plates or rectangles represent repeated measurements: the outer plate surrounds the variables associated with each document, indicating that these variables are repeated for all documents in the corpus. The inner plate represents the repetition of words within a document. The colored circles represent observed variables while the white ones represent the latent variables.

In Chapter 3 I will make use of ideas from LDA to build a random graph model for chromatin long-range interactions. LDA exploits the structured variability in a corpus of documents, in order to infer a short and compact representation of each document in terms of the proportions of a small number of topics. In a similar spirit, the model in Chapter 3 aims at inferring an interpretable low-dimensional representation of chromatin organization starting from interaction patterns as assayed by Hi-C techniques and simultaneously linking this compact representation with nuclear biochemical processes.

1.4 Computational Methods in Bayesian Inference

Bayesian inference often requires computational methods to approximate posterior distributions, especially when the exact calculation is intractable. Two prominent approaches are sampling methods and variational methods.

1.4.1 Sampling Methods

Sampling methods, such as Markov Chain Monte Carlo (MCMC) [68], generate samples from the posterior distribution to approximate it. Among these, the Metropolis-Hastings algorithm is a widely used technique. It proposes new states based on a proposal distribution $q(\theta^{(2)}|\theta^{(1)})$ and accepts or rejects these states based on an acceptance probability:

$$\alpha = \min\left(1, \frac{P(\theta_{\text{new}}|y)q(\theta_{\text{old}}|\theta_{\text{new}})}{P(\theta_{\text{old}}|y)q(\theta_{\text{new}}|\theta_{\text{old}})}\right) \tag{1.8}$$

For complex models the likelihood is often intractable, thereby preventing the direct calculation of the acceptance probability α, however we will see in Chapter 2 that if we can sample from said intractable likelihood, we can come up with approximation strategies to estimate α.

1.4.2 Variational Methods

Variational methods offer an alternative to sampling methods by transforming the inference problem into an optimization problem. A prominent example of this is the Variational Bayes Expectation Maximization (VBEM) [69, 70] also called Variational Message Passing (VMP), derived from the Expectation-Maximization (EM) algorithm [71]. I provide here a general description of VBEM, which will be instrumental for Chapter 3, where I implement a version of this algorithm.

Consider a model with observed variables y, latent or unobserved variables θ, and model parameters ϕ. The logarithmic marginal likelihood, or evidence of the model, is given by:

$$\mathcal{L}(\phi) = \log(P_\phi(y)) = \log \int d\theta P_\phi(\theta, y) \tag{1.9}$$

This can be rewritten as:

$$\begin{aligned} \mathcal{L}(\phi) &= \log \int d\theta \frac{P_\phi(\theta, y) Q(\theta)}{Q(\theta)} \geq \\ &\geq \mathbb{E}_Q \left[\log \frac{P_\phi(\theta, y)}{Q(\theta)} \right] = \hat{\mathcal{L}}\{\phi, Q\} = \text{ELBO} \end{aligned} \tag{1.10}$$

where the Jensen inequality is applied. The term $\hat{\mathcal{L}}\{\phi, Q\}$, known as the Evidence Lower BOund (ELBO), depends on the parameters ϕ and a distribution $Q(\theta)$ over the latent variables θ. Moreover, it can be shown that:

$$\begin{aligned} \hat{\mathcal{L}}\{\phi, Q\} &= \log P_\phi(y) + \int d\theta Q(\theta) \log \frac{P_\phi(\theta|y)}{Q(\theta)} = \\ &= \mathcal{L}(\phi) - D_{KL}[Q(\theta)||P_\phi(\theta|y)] \leq \mathcal{L}(\phi) \end{aligned} \tag{1.11}$$

This equation indicates that the ELBO becomes tighter as the Kullback-Leibler divergence $D_{KL}[Q(\theta)||P_\phi(\theta|y)]$ decreases, or in other words, meaning that $Q(\theta)$ becomes a better approximation of the posterior distribution $P_\phi(\theta|y)$, and the inequality holds as equality only if $Q(\theta) = P_\phi(\theta|y)$. The EM algorithm consists of iteratively maximizing the ELBO with respect to ϕ and $Q(\cdot)$ to compute the MLE (maximum

likelihood estimate) of ϕ. The optimization process starts with an initial value ϕ^0, and at iteration t $Q^t(\theta)$ and ϕ^t are updated through:

1. E-step: set $Q^t(\theta) = P_{\phi^{t-1}}(\theta|y)$
2. M-step: compute $\phi^t = \text{argmax}_\phi \hat{\mathcal{L}}\{\phi, Q^{t-1}\}$

This approach assumes that the posterior distribution is tractable, which is often not the case. When intractable, the E-step becomes:

$$Q^t(\theta) = \text{argmax}_Q \hat{\mathcal{L}}\{\phi^{t-1}, Q\} = \text{argmin}_Q D_{KL}[Q(\theta)||P_\phi(\theta|y)] \tag{1.12}$$

Thus, posterior inference is framed as an optimization problem. Typically, $Q(\theta)$ is chosen from a parametric family of distributions over the latent variables, often with structural assumptions. A common choice is the mean-field approximation, where the variational distribution is factored into independent components: $Q(\theta) = \prod_i Q_i(\theta_i)$. In the VBEM algorithm, the E-step involves setting the functional derivative of $\hat{\mathcal{L}}\{\phi, Q\}$ to zero. Under the mean-field approximation, the following equation holds:

$$\begin{aligned} \frac{\delta}{\delta Q_i(\theta_i)} \hat{\mathcal{L}}\{\phi, Q\} &= \frac{\delta}{\delta Q_i(\theta_i)} \mathbb{E}_Q \left[\log \frac{P_\phi(\theta, y)}{Q(\theta)} \right] \\ &= \int \prod_{j \neq i} dQ_j \log P_\phi(\theta, y) - \log Q_i(\theta_i) - 1 = 0 \end{aligned} \tag{1.13}$$

Leading to the update equation for the variational distribution:

$$Q_i(\theta_i) \propto \exp(\mathbb{E}_{Q_{\neq i}}[\log P_\phi(\theta, y)]) \tag{1.14}$$

Black-Box Varational Inference

In cases where the expectations of Eq. 1.14 are not directly computable, a widely used method that allows direct optimization of the ELBO is called Black-Box Variational Inference (BBVI) [72]. The main idea behind BBVI is to compute the gradient of ELBO to perform gradient ascent, instead of manually deriving update equations that may require intractable expectation. For a variational distribution chosen from a parametric family $Q_\lambda(\theta)$, where λ are the parameters, the gradient of the ELBO with respect to the variational family parameters can be expressed as:

$$\nabla_\lambda \hat{\mathcal{L}}_\lambda\{\phi, Q\} = \mathbb{E}_{Q_\lambda} \left[\nabla_\lambda \log Q_\lambda(\theta) (\log P_\phi(\theta, y) - \log Q_\lambda(\theta)) \right] \tag{1.15}$$

Having this equation, we can compute noisy unbiased estimates of the gradient with Monte Carlo samples from the variational distribution.

$$\nabla_\lambda \hat{\mathcal{L}}_\lambda\{\phi, Q\} \approx \frac{1}{M}\sum_{m=1}^{M} \nabla_\lambda \log Q_\lambda(\theta_m)(\log P_\phi(\theta, y) - \log Q_\lambda(\theta_m))$$
$$\text{where } \theta_m \sim Q_\lambda(\theta) \tag{1.16}$$

This allows BBVI to be widely applied to a variety of models without requiring model-specific derivations.

2

SEMPER: bottom-up data integration in polymer models for chromatin organization

This chapter uses, with small modifications (wording and section organization), the full publication:

Alex Chen Yi Zhang, Angelo Rosa, Guido Sanguinetti. Bottom-up data integration in polymer models for chromatin organization. *Biophysical Journal*, 123(2), 2024

Genomes' tight confinement [73] within the nucleus of the cell during interphase plays a crucial role in many fundamental biological processes *in vivo*, ranging from gene expression and gene regulation to DNA replication [74, 75, 76]. However, how such mechanical constraints on the chromatin fibers interact with the genetic signals which regulate cellular biology at the molecular level, and how the genome $1D$ sequence is integrated within the complex $3D$ nuclear environment to determine the *epigenetic* fate of each cell, remain amongst the most challenging open problems of modern biology [77, 76].

The chromatin filament of each chromosome is organized into distinct motifs that characterize genome folding at different scales in $3D$: topologically-associated domains (TAD's [78]), nuclear (A/B) sub-compartments [26] and chromosome territories [22] all participate in genome regulation. Quite surprisingly, the molecular

origins of these structures appear remarkably "simple", having been explained as the consequence of general physical mechanisms based on polymer physics, such as molecular "bridges" [48, 47], polymer microphase separation [46, 79, 80], DNA-loop extrusion [81, 82] and topological constraints [40, 83]. On the other hand, cells need to execute genetic programs in a robust and coordinated fashion which requires the efficient transcriptional activation of specific, and cell-type dependent, genomic regions, a fact that cannot be accounted for only by general physical principles. So far, the biochemical bases of these processes have been explained with remarkable success in terms of the $1D$ structure of the genome, particularly in terms of binding of sequence-specific transcription factor (TF) proteins [84, 85] and other epigenetic marks, such as covalent modifications of histone tails [2], to the underlying DNA sequence. In this respect, there is now a mature line of research showing how machine learning (ML) algorithms trained on sequence-related features can be used to explain and predict the transcriptional as well as the epigenetic states of the cells [86, 87, 88, 89]. There is also a growing literature showing that biochemical covariates are highly predictive of the $3D$ structure of chromatin [90, 91, 92, 93]. However, these predictions are generally done either disregarding the polymer nature of the chromatin or including $1D$ effects *post hoc* on a polymer scaffold, rather than constructing a bottom-up model which reflects local interactions.

We therefore introduce here *SEMPER (Sequence Enhanced Magnetic PolymER)*, a new statistical model for chromosome conformations that explicitly bridges between the $1D$ sequence-informed structure of the genome and its $3D$ nature as a polymer. It does so by *interpolating* between a spatial magnetic polymer model [45, 53] describing domains formation in chromatin fibers and a statistical predictor [86] of epigenetic states based on TF-binding data. As, however, a direct Bayesian estimation of the parameters introduced in the model is intractable, we develop here a novel approximate inference method to estimate the model parameters and quantify the corresponding uncertainties.

We apply this model to study epigenetic patterns of selected sequences from human chromosome 2 at different levels of resolution. We compare our results in terms of chromatin state predictions with a purely data-driven benchmark [86], leading to both improved predictions and a substantial reappraisal of the role of the various molecular factors in determining the chromatin state.

Additionally, and as a by-product of our methodology, *SEMPER* explicitly provides putative chromatin conformations, hence it can be used to explore the specific patterns that emerge in the chromatin contact maps obtained by Hi-C technologies [26, 78]. Overall, our work demonstrates that a suitable bottom-up polymer

model enhanced with a relatively *small* number of sequence "features" can both reproduce the epigenetic state of the chromosome and its local $3D$ organization, providing a new tool to conceptualize biochemistry and gene regulation at the microscopic level.

The chapter is organized as follows. In Section 2.1 we summarize the details of our computational model, namely: (i) the model Hamiltonian and the physical meaning of the distinct terms and parameters which define it (Sec. 2.1.1), (ii) the (approximate) Monte Carlo procedure used to estimate the parameters and the relative uncertainties (Sec. 2.1.2) and (iii) a brief account (Sec. 2.1.3) of the experimental data sets used in this work, those used as inputs of the model and those used to evaluate its reliability. In Sec. 2.2, we demonstrate the validity and generality of our approach by focusing on some specific sequences of the human chromosome 2: in particular, we discuss the predictions of our model for the epigenetic state of the chromatin fiber (Sec. 2.2.1) and the role that the polymeric nature of chromatin plays in it (Sec. 2.2.2). Then (Sec. 2.2.3), we measure the chromatin contacts within our model conformations and compare them to the corresponding experimental Hi-C matrices. Finally (Sec. 2.3), we conclude by discussing the main advantages of our model with respect to other approaches and point out future possible improvements. In Appendix A, the interested reader will find additional details about the polymer model (Sec. A.1) and the simulation scheme (Secs. A.1.2 and A.1.3), a detailed explanation (Sec. A.2) concerning the treatment of experimental data for their integration in the model, as well as (Sec. A.3) a proof-of-concept for the numerical convergence of our method by making use of synthetic data.

2.1 Materials and methods

2.1.1 *SEMPER*: an integrative model of chromatin organisation

Eukaryotic genomes are organized in single chromosomes, and each chromosome is made of a unique filament of chromatin fiber. Chromatin itself is a sophisticated filamentous DNA-protein complex which serves as a platform for the integration of multiple biochemical and biophysical signals [2].

Starting from these considerations, the working hypothesis of our approach is that the structural organization of the chromatin filament is itself affected by the presence of genetic signals such as sequence composition or transcription factor binding. This has been strongly validated at the correlative level in a number of recent papers [86, 87]. Here instead, we aim at developing a mechanistic polymer model where these

influences are encoded *a priori*, and their relative importance can be gauged by means of statistical inference within the mechanistic model.

The purely data-driven model we build upon is [86], in particular we start from a simple logistic regression (LR) predictor of the presence/absence of the *H3K27me3* histone modification along the DNA sequence from the binding of transcription factors and other $1D$ assays. We then embed this $1D$ model on top of the *SEMPER* polymer model, in order to account for the spatial constraints to which the chromatin filament is subject.

In general, a convenient model to capture the $3D$ organization of the chromatin is the *magnetic polymer model* [52, 45, 53]: a polymer chain in $3D$ is modeled as a self-avoiding random walk (SAW), with nearby (in space) monomers interacting via a Ising-like term modeling the propensity of chromatin regions with the same epigenetic mark (see below) to arrange into spatially close domains. While this model has been shown to recapitulate successfully the statistics of open/compact chromatin domains, it does not account for the dynamical activation of specific chromatin loci as a result of sequence-specific biochemical signals.

To address this issue, we expand the "old" magnetic polymer model into our new *SEMPER* model by introducing monomer-specific *external field* terms, which modulate the probability of each monomer lying in an open or compact region based on its biochemical features, such as the sequence composition or the presence of specific binding sites for transcription factors (TF, see Fig. 2.1).

More precisely, chromosome spatial conformations are modeled as N-steps self-avoiding paths on the $3D$ simple cubic lattice [94] of unit lattice spacing $= d$. For computational convenience, we work with polymer chains with $N = 1000$, while we vary d systematically in order to study the chromatin fiber at different resolution details (see Table 2.1).

Then, single chromosomal conformations at temperature T are sampled from the canonical distribution function $P(\{r\}, \{\sigma\}|\boldsymbol{\theta}) \sim e^{-\mathcal{H}(\{r\},\{\sigma\}|\boldsymbol{\theta})/\kappa_B T}$, where κ_B is the Boltzmann constant and the Hamiltonian $\mathcal{H}$ is given by the expression:

$$\begin{aligned} \mathcal{H}(\{r\}, \{\sigma\}|\boldsymbol{\theta}) &= \mathcal{H}_{\text{SAW}}(\{r\}) - \frac{J}{2}\sum_{i\neq j}\sigma_i\sigma_j\Delta(\boldsymbol{r}_i, \boldsymbol{r}_j) - \sum_i h_i(\boldsymbol{w})\sigma_i \\ &\equiv \mathcal{H}_{\text{SAW}}(\{r\}) + \mathcal{H}_{\text{Ising}}(\{r\}, \{\sigma\}|J) + \mathcal{H}_{\text{ext}}(\{\sigma\}|\boldsymbol{w})\,. \end{aligned} \tag{2.1}$$

In Eq. (2.1), $\{r\} \equiv (\mathbf{r}_1, ..., \mathbf{r}_N)$ and $\{\sigma\} \equiv (\sigma_1, ..., \sigma_N)$ are the compact notations for, respectively, the spatial coordinates of the N monomers composing the polymer chain and the corresponding Ising-like spin variables representing the (epigenetic)

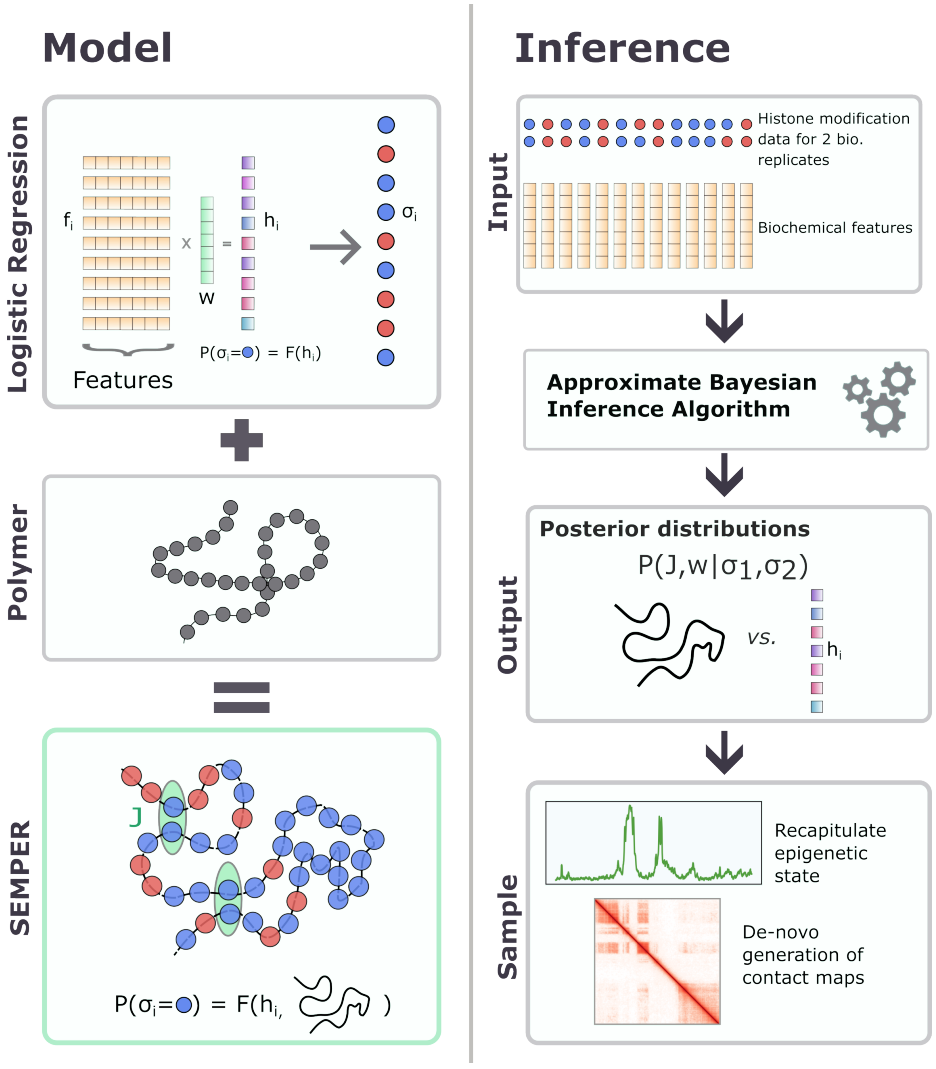


Figure 2.1: Schematic illustration of the *SEMPER* model and the inference pipeline. The model (left column) starts from a simple statistical predictor (logistic regression) of the epigenetic state of chromatin from other biochemical features. In this description each data point corresponds to a genomic locus, characterized by the presence or absence of some epigenetic mark (in our case *H3K27me3*), and each genomic locus has an associated vector of biochemical features such as the binding of specific transcription factors (TFs). SEMPER (Sec. 2.1.1) expands the simple predictor by embedding the system on a linear polymer, which induces correlations between linearly distal chromatin loci. On the right, I schematize the inference procedure. The input data consists of binarized ChIP-seq tracks for *H3K27me3* for a certain genomic region, for which we typically have two biological replicates, and the associated biochemical features. Our novel Bayesian inference algorithm (Sec. 2.1.2) then provides posterior distributions for the model parameters, allowing us to quantify the relative importance of the polymeric constraints. The obtained posterior samples can then be used to predict both, the $1D$ epigenetic state and the $3D$ folding of the chromatin fiber extracted from the Hi-C contact maps (Sec. 2.2). Notably, the model gets no input from $3D$ chromatin folding, so it can be used to get insight on chromatin conformations based only on the information acquired from 1D data (Sec. 2.2).

"states" of the chromatin in terms of the binary values $\sigma_i = 1/0$ (respectively, the presence/absence of certain histone modification, see below and Sec. 2.1.2). The three terms in Eq. (2.1) represent the following: (i) The first term $\mathcal{H}_{\rm SAW}(\{r\})$ enforces strict self-avoidance of the polymer chain on the cubic lattice [94]: it is equal to $+\infty$ whenever any two monomers occupy the same lattice site and $= 0$ otherwise. (ii) The second term $\mathcal{H}_{\rm Ising}(\{r\},\{\sigma\}|J)$, with $J > 0$ and $\Delta(\boldsymbol{r}_i,\boldsymbol{r}_j) = 1$ for $|\boldsymbol{r}_i - \boldsymbol{r}_j| = d$ and $= 0$ otherwise, is the Ising-like interaction generating the magnetic effect. (iii) The *external field* term $\mathcal{H}_{\rm ext}(\{\sigma\}|\boldsymbol{w})$ models the coupling of the local chromatin state σ_i with the underlying biochemistry. In particular the values of the local external fields h_i, $i = 1,\ldots,N$ are expressed as a linear combination of f features,

$$h_i(\boldsymbol{w}) = \sum_{\alpha=1}^{f} F_{i\alpha} w_\alpha \,, \tag{2.2}$$

where the $F_{i\alpha}$ entry of the $N \times f$ matrix $\boldsymbol{F}$ represents the value of the α-th feature at the i-th genomic region while the parameter w_α is a measure of the "weight" assigned to the corresponding feature in determining the chromatin epigenetic state. More precisely, the value of $F_{i\alpha}$ is defined as the *log fold changes over control* of ChIP-seq assays of transcription factor binding and histone modifications. For compactness in the following we will denote by $\boldsymbol{\theta} \equiv (J,\boldsymbol{w})$ the free parameters of the polymer model, namely the spin-spin coupling constant J and the f weights ($\boldsymbol{w} \equiv (w_1, w_2, ..., w_f)$) used in the computation of the external fields.

Finally, it is worth noticing that in the limit when the magnetic polymer structure is ignored (*i.e.*, for $J = 0$), the spin variables decouple and the probability of a particular spin sequence is given by the logistic regression (LR) predictor considered in [86]. In this way, the polymer model described by Eq. (2.1) interpolates between a biophysical model based on magnetic polymers and a data-driven predictor for the epigenetic state of the sequence. For the interested reader, a more detailed account of the polymer model is provided in Sec. A.1.1 in Appendix A.

2.1.2 Estimating model parameters for *SEMPER*: approximate Monte Carlo method

In this work, we choose as our definition of the chromatin state the presence/absence of the histone modification known as *H3K27me3*, notably a proxy [95] for chromatin existing in a highly compact phase (heterochromatin). The presence of *H3K27me3* along the chromatin filament can be detected through state-of-the-art ChIP-seq experiments [7]: here, we adapt the results of certain of these experiments to our model as the following.

Explicitly, we assume to observe a length N spin configuration $\{\tilde{\sigma}\}$ for the state of all monomers (open/compact chromatin). Then, by measuring energy in units of $\kappa_B T = 1$ where κ_B is the Boltzmann constant, the likelihood associated with such a spin configuration at fixed $\boldsymbol{\theta} = (J, \boldsymbol{w})$ reads:

$$P(\{\tilde{\sigma}\}|\boldsymbol{\theta}) = \sum_{\{r\}\in \mathrm{SAW}} P(\{r\},\{\tilde{\sigma}\}|\,\boldsymbol{\theta}) = \frac{e^{-\mathcal{H}_{\mathrm{ext}}(\{\tilde{\sigma}\}|\boldsymbol{w})}}{Z(\boldsymbol{\theta})}\,\Phi(J,\{\tilde{\sigma}\})\,, \tag{2.3}$$

with

$$Z(\boldsymbol{\theta}) = \sum_{\{r\}\in \mathrm{SAW}} \sum_{\{\sigma\}} \exp\left(-\mathcal{H}(\{r\},\{\sigma\}|\,\boldsymbol{\theta})\right)\,, \tag{2.4}$$

$$\Phi(J,\{\tilde{\sigma}\}) = \sum_{\{r\}\in \mathrm{SAW}} e^{-\left(\mathcal{H}_{\mathrm{SAW}}(\{r\})+\mathcal{H}_{\mathrm{Ising}}(\{r\},\{\tilde{\sigma}\}|J)\right)}\,, \tag{2.5}$$

where $P(\{r\},\{\tilde{\sigma}\}|\,\boldsymbol{\theta})$ is the canonical distribution with the Hamiltonian (2.1).

The introduction of the coupling strength J allows for physically motivated correlations between the spin variables σ_i which in the LR model are statistically independent. Unfortunately, the calculation of the likelihood (2.3) entails the marginalization of the polymer configurations which, except for the $J = 0$ case, constitutes an analytically intractable task. Therefore the posterior distribution,

$$P(\boldsymbol{\theta}|\{\tilde{\sigma}\}) \propto P(\{\tilde{\sigma}\}|\,\boldsymbol{\theta})P(\boldsymbol{\theta}) = \frac{e^{-\mathcal{H}_{\mathrm{ext}}(\{\tilde{\sigma}\}|\boldsymbol{w})}}{Z(\boldsymbol{\theta})}\,\Phi(J,\{\tilde{\sigma}\})P(\boldsymbol{\theta})\,, \tag{2.6}$$

also depends on the intractable normalization constants $Z(\boldsymbol{\theta})$ and $\Phi(J,\{\tilde{\sigma}\})$. Under these conditions, exact sampling with standard tools such as the popular Metropolis-Hastings Monte Carlo (MC) algorithm [68] becomes unfeasible: in fact, as known, in the MC algorithm moving from the current state $\boldsymbol{\theta}^t$ of the chain the new state $\boldsymbol{\theta}^{t+1}$ (proposed by adding a random displacement $\boldsymbol{\delta\theta}^t$) is accepted/rejected with a certain probability depending on the ratio "$P(\boldsymbol{\theta}^{t+1}|\{\tilde{\sigma}\})/P(\boldsymbol{\theta}^t|\{\tilde{\sigma}\})$" which also contains said intractable normalization constants.

Various methodologies, collectively known as *likelihood-free* or *simulator-based* inference methods, have been developed over the years to perform approximate Bayesian inference in settings where the likelihood is not analytically tractable. Such methods, recently reviewed in [96, 97], are principally geared towards situations where the stochastic process can only be simulated. The error made in these approximations is typically difficult to control and, generally, these methods incur very high computational costs particularly when the number of parameters to estimate is from medium to high. The main cause of both these drawbacks is that typically these methods treat the simulator as black box sample generators, without taking advantage of the characteristics of the system at hand. Here, instead, we propose to leverage the description

of the system in terms of the Hamiltonian (2.1) to enable a more guided exploration of the parameters space, thus providing a more efficient and scalable algorithm.

For this purpose, we propose an approximate Metropolis-Hastings inference method based on the observation that we may take advantage of the following Taylor expansions:

$$\ln \frac{Z(\boldsymbol{\theta}+\boldsymbol{\delta\theta})}{Z(\boldsymbol{\theta})} \simeq \boldsymbol{\delta\theta}\cdot\nabla_{\boldsymbol{\theta}} \ln Z(\boldsymbol{\theta}) + \frac{\boldsymbol{\delta\theta}^T\cdot \boldsymbol{H}_{\ln Z}(\boldsymbol{\theta})\cdot\boldsymbol{\delta\theta}}{2}\,, \tag{2.7}$$

where $\boldsymbol{H}_{\ln Z}(\boldsymbol{\theta})$ is the Hessian matrix of the function $\ln Z$ evaluated in $\boldsymbol{\theta}$ and

$$\ln \frac{\Phi(J+\delta J,\{\tilde{\sigma}\})}{\Phi(J,\{\tilde{\sigma}\})} \simeq \frac{\partial \ln \Phi(J,\{\tilde{\sigma}\})}{\partial J}\delta J + \frac{1}{2}\frac{\partial^2 \ln \Phi(J,\{\tilde{\sigma}\})}{\partial J^2}\delta J^2\,. \tag{2.8}$$

In particular, we observe that the first and the second derivatives correspond to expectation values with respect to (w.r.t.) probability distributions that we cannot evaluate explicitly but that we can sample from. We have that

$$\begin{aligned}
(\nabla_{\boldsymbol{w}} \ln Z(\boldsymbol{\theta}))_j &= \mathbb{E}_{\{r\},\{\sigma\}|\boldsymbol{w},J}\left[\sum_i F_{ij}\sigma_i\right]\,,\\
\frac{\partial \ln Z(\boldsymbol{\theta})}{\partial J} &= \mathbb{E}_{\{r\},\{\sigma\}|\boldsymbol{w},J}\left[\frac{1}{2}\sum_{i\neq j}\sigma_i\sigma_j\Delta(\boldsymbol{r}_i,\boldsymbol{r}_j)\right]\,,
\end{aligned} \tag{2.9}$$

where $\mathbb{E}_{\{r\},\{\sigma\}|\boldsymbol{w},J}[\cdot]$ stands for the expectation value w.r.t. the distribution of the spins and polymer configurations at fixed $\boldsymbol{w}$ and J, *i.e.* w.r.t. the joint likelihood. The elements of the Hessian matrix in Eq. (2.7) read:

$$[\boldsymbol{H}_{\ln Z}(\boldsymbol{\theta})] = \langle s_i s_j\rangle - \langle s_i\rangle\langle s_j\rangle = \mathrm{cov}(\boldsymbol{s})_{ij}\,, \tag{2.10}$$

where we use the shorthand notation $\langle\cdot\rangle$ for $\mathbb{E}_{\{r\},\{\sigma\}|\boldsymbol{w},J}[\cdot]$ and the array $\boldsymbol{s}$ is defined as:

$$\boldsymbol{s} \equiv \begin{bmatrix} \sum_i F_{i1}\sigma_i \\ \sum_i F_{i2}\sigma_i \\ \vdots \\ \vdots \\ \sum_i F_{if}\sigma_i \\ \frac{1}{2}\sum_{i\neq j}\sigma_i\sigma_j\Delta(\boldsymbol{r}_i,\boldsymbol{r}_j) \end{bmatrix} \in \mathbb{R}^{f+1}\,.$$

We have developed a simulator for *SEMPER* based on [94, 98, 52] that we can employ to obtain samples from the likelihood $P(\{r\},\{\sigma\}|\boldsymbol{\theta})$ and then use these samples to obtain unbiased estimates of the quantities needed to compute the above Taylor expansions. Finally, the ratio of the two partition functions in Eq. (2.7) reads:

$$\frac{Z(\boldsymbol{\theta}+\boldsymbol{\delta\theta})}{Z(\boldsymbol{\theta})} \simeq \exp\left(\boldsymbol{\delta\theta}\cdot\langle\boldsymbol{s}\rangle + \frac{\boldsymbol{\delta\theta}^T\cdot cov(\boldsymbol{s})\cdot\boldsymbol{\delta\theta}}{2}\right) \tag{2.11}$$

with corrections that are of the order of $\sim \mathcal{O}(\|\boldsymbol{\delta\theta}\|^3)$, while for the evaluation of the ratio in Eq. (2.8) we use the expressions:

$$\frac{\partial \ln \Phi(J,\{\tilde{\sigma}\})}{\partial J} = \mathbb{E}_{\{r\}|J,\{\tilde{\sigma}\}}\left[\frac{1}{2}\sum_{i\neq j}\tilde{\sigma}_i\tilde{\sigma}_j\Delta(\boldsymbol{r}_i,\boldsymbol{r}_j)\right], \tag{2.12}$$

and

$$\frac{\partial^2 \ln \Phi(J,\{\tilde{\sigma}\})}{\partial J^2} = \mathbb{V}_{\{r\}|J,\{\tilde{\sigma}\}}\left[\frac{1}{2}\sum_{i\neq j}\tilde{\sigma}_i\tilde{\sigma}_j\Delta(\boldsymbol{r}_i,\boldsymbol{r}_j)\right]. \tag{2.13}$$

The expectation value (2.12) and the variance (2.13) are computed w.r.t. the distribution function:

$$P(\{r\}|J,\{\tilde{\sigma}\}) = \frac{e^{\frac{J}{2}\sum_{i\neq j}\tilde{\sigma}_i\tilde{\sigma}_j\Delta(\boldsymbol{r}_i,\boldsymbol{r}_j)}}{\Phi(J,\{\tilde{\sigma}\})}, \tag{2.14}$$

which describes the probability for the polymer configurations with spin variables $\{\sigma\}$ having values equal to the observed spin configuration $\{\tilde{\sigma}\}$. For additional technical details on the implementation of the simulator (especially regarding the MC moves adopted in the sampling of polymer conformations and the efficient method to bypass long-lived low-energy states), see Secs. A.1.2 and A.1.3 in Appendix A.

2.1.3 Datasets used as inputs for the model and experimental Hi-C data

We employ freely available data for the human chromosome 2 from the H1 embryonic stem (H1-hESC) cell line. As briefly anticipated in Sec. 2.1.2, as *markers* for compact chromatin, we use ChIP-seq tracks of *H3K27me3*. As *predictors* we use a combination of sequence-derived features and measured binding of transcription factor proteins. Namely, as a sequence-derived feature we use the CpG content due to its established biological relevance in the context of epigenetics. Then, motivated by the work [86] where it has been shown that transcription factor (TF) binding data around transcription start sites have good predictive power regarding the presence (or the absence) of histone modifications around the same sites, we employ the same TF binding data studied there to build (for practical details on how to include (coarse-grain) experimental data in the model, see Sec. A.2 in Appendix A) the features matrix $\boldsymbol{F}$ entering (see Eqs. (2.1) and (2.2)) in the definition of the external fields h_i of the *SEMPER* model. This strategy will enable a direct and transparent comparison between the results of the purely data-driven logistic regression model of [86] (which, as noted here, represents a limiting case of *SEMPER*) and the polymer-based approach, in particular in terms of the interpretation of model results.

The histone modification data for *H3K27me3* (in the form of two independent replicates with the same sequence (isogenic)) and the TF binding data were down-

loaded from the *ENCODE project* database [99]. The complete list of data, corresponding accession, and data file numbers are summarized in Table A.1 in Appendix A. Finally, the Hi-C matrices used in comparison to the maps generated with our polymer model (see Sec. 2.2.3) were drawn from the *4DN* data portal [100] under the accession code 4DNFIQYQWPF5.

The computer code implementation of *SEMPER* can be downloaded at the following online repository: `https://github.com/alex-chenyi-zhang/Magpol-Likelihood-Free.git`.

2.2 Results: Epigenetic state and spatial conformations of human chromosome 2

To assess its robustness, we have conducted systematic tests for *SEMPER* by employing *synthetic* data (Sec. A.3 in Appendix A). First, we have considered an artificial chromatin sequence for which the ground truth values of the parameters of the model are known *by construction*. As shown in Fig. A.2 in Appendix A, the method is capable of recovering these values faithfully. Second, we have taken advantage of the fact that for $J = 0$ (namely when the model reduces to a pure LR predictor, see Sec. 2.1.1) an exact evaluation of the likelihood is possible and compare the results to our approximate method and to the recently proposed *Synthetic Likelihood* method by Wood [101]. The results displayed in Figs. A.3, A.4 and A.5 in Appendix A demonstrate the superiority of our method with respect to Wood's and give us confidence in the model.

We are now in the position to apply *SEMPER* on a case study with real data. As anticipated in Sec. 2.1.3, in the rest of the chapter we consider the epigenetic states of selected portions of human chromosome 2. Then, in order to assess the performance of our model, we vary systematically the coarse-grain resolution (details in Table 2.1) of the model chromatin fiber.

2.2.1 Predictions of epigenetic states

Having selected a data set and features, we use the approximate inference method we developed to obtain samples from the intractable posterior distributions for the model parameters. To check if the obtained posterior distributions are meaningful, we compute *predictions* of epigenetic state patterns as average simulated configurations of the *H3K27me3* mark presence. In particular, to generate such predictions in the form of spin configurations, we perform Bayesian model averaging for the posterior

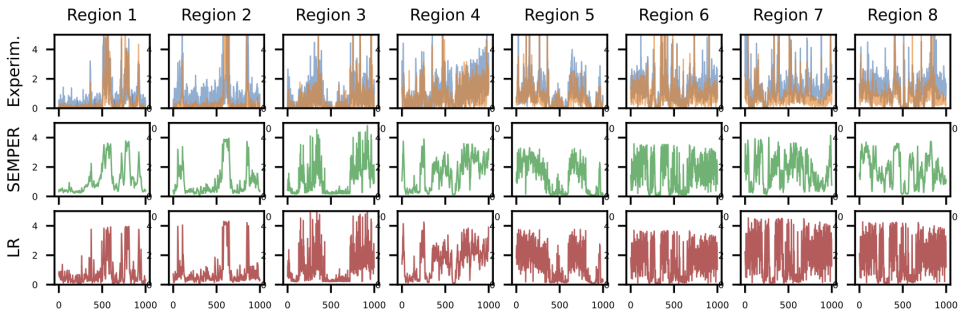


Figure 2.2: *SEMPER* recapitulates histone modification patterns. (Top row) Comparison between the two biological replicates (see Sec. 2.1.3 for details) of the experimentally measured *H3K27me3* ChIP-seq tracks [99]. (Middle and bottom row) Average patterns w.r.t. the predictive posterior distributions of the spin configurations calculated by, respectively, the *SEMPER* and the logistic regression (LR) model.

predictive distribution, *i.e.*

$$P(\{\sigma\},\{r\}|\{\tilde{\sigma}\}) = \int d\boldsymbol{\theta} P(\{\sigma\},\{r\}|\boldsymbol{\theta}) P(\boldsymbol{\theta}|\{\tilde{\sigma}\})\,. \tag{2.15}$$

Noticeably, this estimator provides also a full account of the uncertainties in the parameter estimates and the related uncertainties over the predictions.

The results of our predicted histone modification patterns are shown in Fig. 2.2 (middle row), in comparison to the measured patterns for the two independent replicates (top row) and the predictions from pure LR (*i.e.* $J = 0$, bottom row). Qualitatively, *SEMPER* recapitulates well the *H3K27me3* patterns by performing similarly to LR. Then, in order to obtain a more quantitative measure of the performance of *SEMPER*, we have computed (see Table 2.1) the correlations between the posterior-predicted profiles and the actual ChIP-seq profiles used to train the model, as well as the correlation between the two replicate experimental profiles. By examining the last two columns of Table 2.1, we observe that the adoption of the magnetic polymer model improves noticeably the predictions if compared to the situation (LR method [86]) where the polymeric nature of the chromatin fiber is neglected. The improvement appears particularly marked for shorter length scales of coarse-graining (500bp, *i.e.* of the same order of magnitude as the length of DNA wrapped around a nucleosome [2]), where *SEMPER* achieves a correlation level close to the correlation between experimental replicates. Since $\sigma_i = 1/0$ accounts for the presence/absence of the histone modification, portions of the polymer rich in positive spins have the tendency to condense in agreement with the known fact [102] that *H3K27me3* is associated with transcriptional repression *via* the formation of heterochromatic regions.

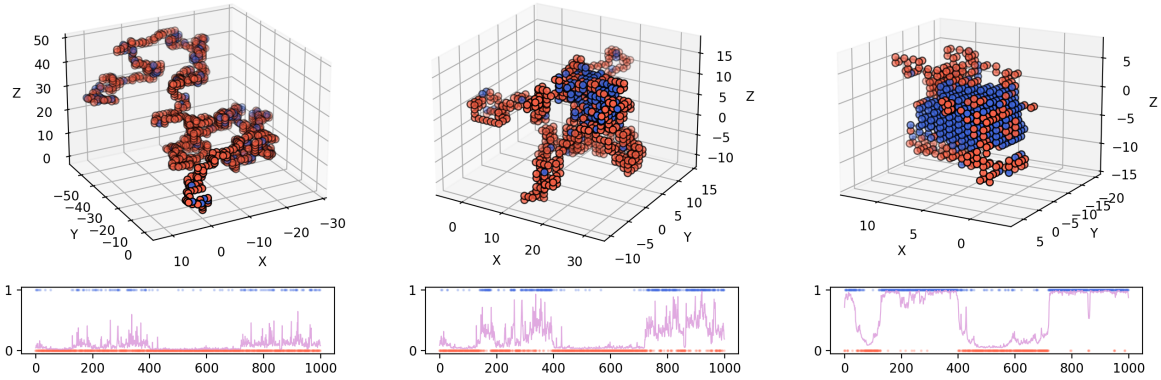
Start (Mbp)	End (Mbp)	Res. (bp)	Replicates	SEMPER	LR
218.6	219.1	0.5k	0.894	0.874	0.785
222.0	222.5	0.5k	0.625	0.786	0.752
130	131	1k	0.818	0.621	0.568
111.2	113.2	2k	0.794	0.656	0.631
85	90	5k	0.853	0.684	0.612
129	134	5k	0.750	0.584	0.531
95	105	10k	0.694	0.562	0.546
105	115	10k	0.761	0.647	0.638

Table 2.1: Mean Pearson correlation coefficients between the predicted epigenetic state patterns and the experimentally measured histone modification ChIP-seq data (two isogenic replicates for each region, see Sec. 2.1.3 for details). For each region of human chromosome 2, we report also the correlation between the two biological replicates. We observe the improvement of the predictions of the *SEMPER* model compared to logistic regression (LR [86]) in which the polymeric nature of chromatin is, by construction, not taken into account. The first two columns contain the extremes of the considered genomic regions reported in Mbp ($= 10^6$ basepairs), while the third column reports the coarse-grain resolution of the polymer model as the number of basepairs (bp) corresponding to each monomer in the model.

2.2.2 Influence of polymeric terms in estimating the parameters

Our results show quantitatively that the incorporation of a polymer structure leads to an improved recapitulation of chromatin state, suggesting an important role of 3D organization in shaping the patterns of histone modifications.

Even more importantly Fig. 2.4 and Fig. 2.5 also suggest a major difference in terms of the possible interpretation of the results obtained. The panels in Fig. 2.4 show samples from joint posteriors over pairs of weights obtained from our approximate Monte Carlo algorithm (blue dots) and from the logistic regression (LR) model previously adopted in [86]. The comparison illustrates a striking difference between the estimates: first of all, the estimate of the interaction parameter J is significantly different from zero, indicating that the data strongly favors an explanation which includes the polymer nature of the chromatin. *SEMPER* estimates of weights tend to have larger uncertainty and to be generally of smaller average magnitude compared to the LR estimates (see Fig. 2.5). Intriguingly, *SEMPER* infers a very small weight for the feature corresponding to CpG density, consistent with the alternative chromatin regulating mechanisms underpinning CpG methylation and polycomb-repressor



(a) $J = 0$ (b) $J \sim P(J|\tilde{\sigma})$ (c) $J = 1.3$

Figure 2.3: Polymer and spin configurations from varying parameter values. For the central column, I sample parameter sets from the posterior distribution and then sample configurations of magnetic polymer at fixed parameter values. In the left column, I still sample the weights for the biochemical features from the posterior but fix the monomer-monomer coupling to $J = 0$, while in the right column, I fix $J = 1.3$. The top row shows examples of such magnetic polymer configurations while the bottom row displays the snapshot of the associated spin configuration and the average magnetization across the sampled ensemble. We can appreciate how the wrong estimation of the polymer importance encoded in the coupling J leads to erroneous epigenetic state recapitulation.

complex-mediated deposition of *H3K27me3* [103]. In turn, discrepancies in weight estimates can lead to very significant differences in terms of interpretation: our results suggest that the importance of TFs in shaping the chromatin landscape might have been overestimated by purely data-driven efforts such as [86], while many conformational features can largely emerge as a result of polymer mechanisms amplifying relatively subtle genetic signals.

2.2.3 Bottom-up modeling leads to realistic interaction patterns

Besides being able to predict epigenetic state patterns from sequence and TF binding data alone, the model provides us with realistic putative chromatin conformations. To demonstrate their plausibility, we compute the corresponding contact maps between chromatin loci and compare those to the ones measured through Hi-C technologies [26, 104]. Hi-C, which is typically based on a suitable combination of cross-linking, ligation and sequencing of nuclear DNA, constitutes a powerful and worldwide established protocol for the systematic detection of chromatin-chromatin interactions. The result of the experiment can be intuitively displayed in the form of *contact maps*, whose entries correspond to the number of detected pairwise contacts between the two generic

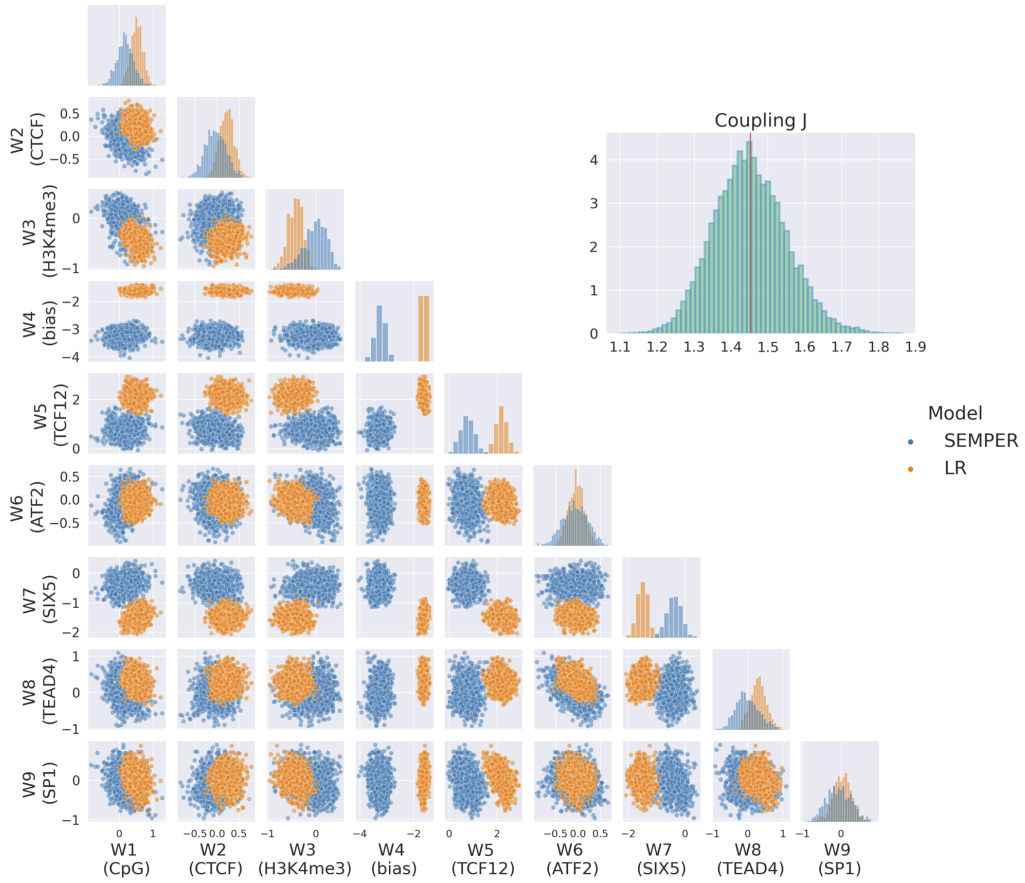


Figure 2.4: *SEMPER* and the approximate Monte Carlo scheme (Sec. 2.1.2) allow us to obtain samples from the otherwise intractable posterior distribution of the model. In this plot, the results for the human chromosomal region *chr2:218,600,000-219,100,000* (see first entry in Table 2.1) are shown. In green, one can see the marginal posterior distribution for the coupling J between neighboring spins. The support of the distribution on non-zero values shows the great importance of the polymer in explaining the 1D *H3K27me3* data. In the smaller squares, we show the one and two-dimensional posterior marginals for the weights w_α (Eq. (2.2)) that appear in the external fields of the Hamiltonian (Eq. (2.1)), each associated with a different biological covariate. Notice the quantitative differences with the logistic regression (LR) posterior distributions.

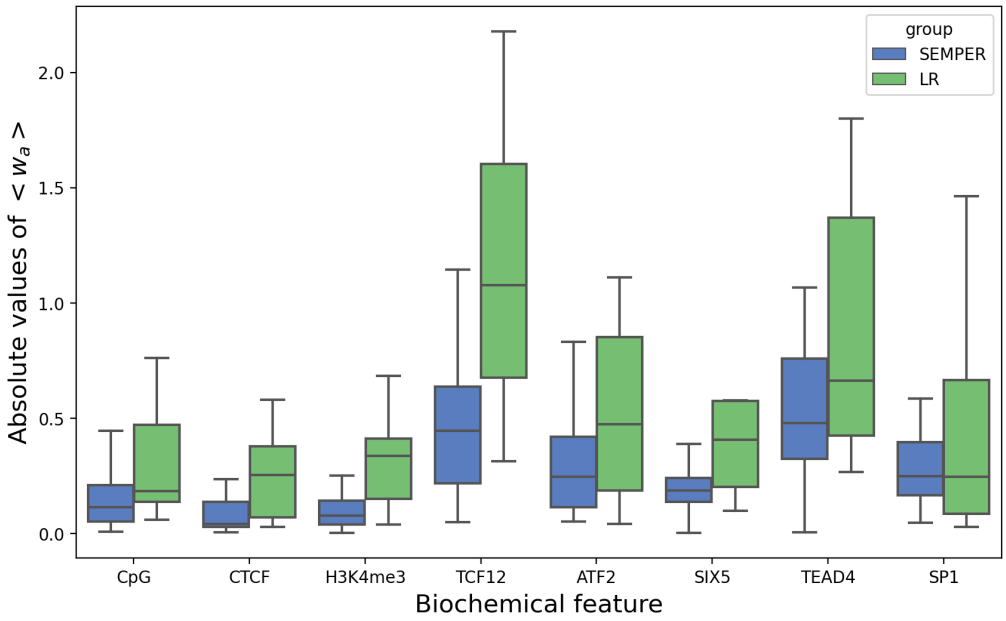


Figure 2.5: For all the analyzed genomic regions, I compute the absolute values of the posterior means for the weights associated with all biochemical features. The comparison between weight magnitudes for the *SEMPER* and LR model shows an overestimation of epigenetic factors' importance by a purely biochemical approach.

fragments i and j along the genomic sequence of interest.

In order to compare with experiments, contact matrices from simulations are computed in the following way. For a given chromatin conformation, a contact between monomers i and j is counted if their mutual spatial distance $||\boldsymbol{r}_j - \boldsymbol{r}_i|| \leq r_c$ where r_c is a suitable cutoff contact distance. Different values of r_c in the range $[\sqrt{5}d - \sqrt{20}d]$, where d is the employed spatial discretization of the chromatin fiber (Sec. 2.1.1), were used without displaying substantial differences in the emerging patterns in the contact maps. The experimental *vs.* the simulated matrices are shown in Fig. 2.7. For comparison, each panel displays also the relative *H3K27me3* track (Fig. 2.2, top row) used to learn the parameters of the polymer model, thus providing an intuitive illustration of how the intrinsically noisy ChIP-seq signal is "transformed" into a contact map. Importantly, like the authors in [102], we never include any structural information about the chromatin organization itself in the model. Our aim is to explore to what extent a bottom-up approach integrating sequence and polymer models can recapitulate the contact maps of the chromatin, rather than obtaining the highest possible predictive accuracy.

From a close inspection of Fig. 2.7, it is clear that not all the details of the experimental maps can be captured from our model, and some spurious contacts are easily discernible at the boundaries of the simulated matrices (likely, the result of

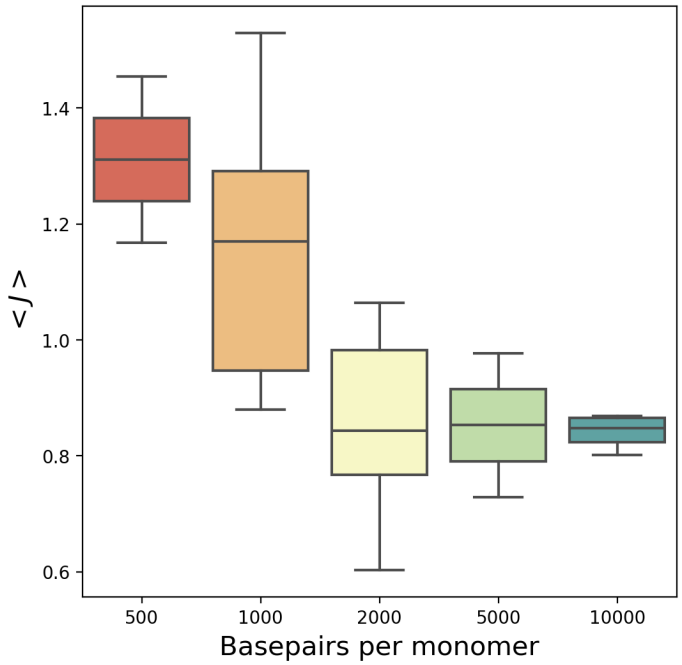


Figure 2.6: For all the analyzed genomic regions I compare the posterior mean of the inferred monomer-monomer coupling for different levels of coarse-graining, i.e., the number of basepairs modeled by a single monomer. This plot shows that the importance of taking into account the polymeric nature of chromatin is particularly marked for finer organization level, while this importance decreases as one looks at a more coarse-grained picture.

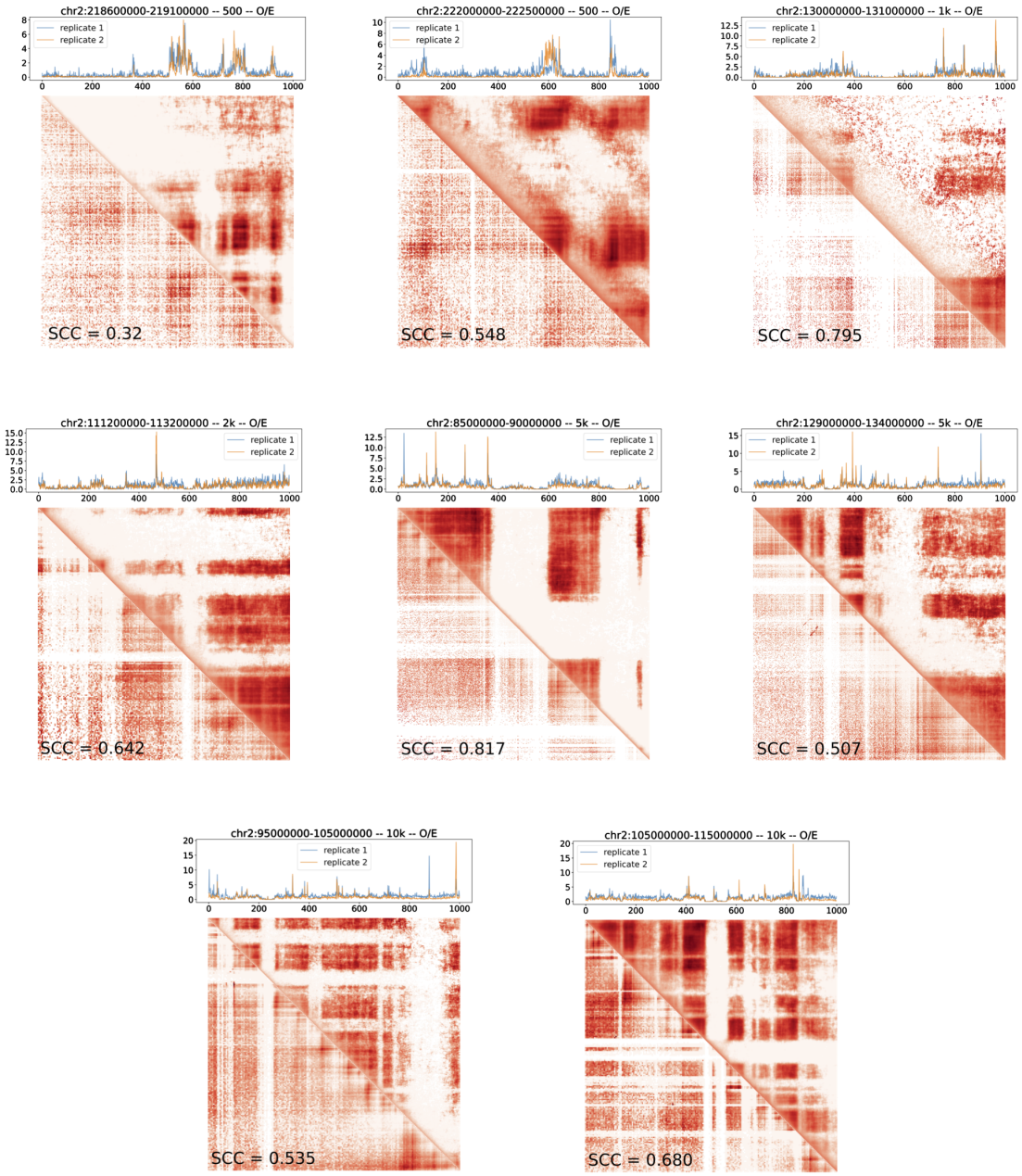


Figure 2.7: Comparison of experimental and calculated ChIP-seq tracks and Hi-C matrices. In each panel: (i) the top plot shows the *H3K27me3* ChIP-seq tracks for the two employed biological replicates; (ii) the bottom plot shows the comparison between the experimentally measured Hi-C maps (lower triangular part) and the calculated ones from our model chromatin conformations (upper triangular part). The intensity of the pixels corresponds to the normalized log-observed-*vs.*-expected number of reads. The reported *Stratum Corrected Correlation* (SCC) coefficient, with range $\in [-1.0 : +1.0]$, quantifies the similarity between experimental and predicted Hi-C matrices (see text for details).

working with a finite polymer chain and not with a full chromosome as in [40]). Despite these mismatches, our simulated conformations display also a number of realistic patterns of interactions fairly consistent with the experimental results. In order to assess then the similarity between experimental and computational Hi-C matrices, we employ the so-called *Stratum Corrected Correlation* (SCC), a metric index with range $\in [-1.0 : +1.0]$ introduced in [105] for the scope: the results displayed in Fig. 2.7 demonstrate, overall, good correlation between the two sets of data. Such correlations are also emerging by comparing the corresponding mean contact probabilities [26] (as well as the Pearson correlation coefficient between the two) between pairs of chromatin loci as a function of their genomic separation (see panels (A) and (B) in Fig. A.6 in Appendix A). Overall, the similarity between experimental and predicted Hi-C matrices is particularly remarkable considering that we have used data for just one (amongst all possible!) histone modification and only a few transcription factors, that, a priori, should not be expected to recapitulate fully the detailed geometry of the chromatin filament. Finally notice that, as a by-product, one can also apply our method to quantify the intrinsic variability of the chromatin fiber (see Fig. A.7 in Appendix A, displaying the posterior-predicted standard deviations of contact probabilities).

2.3 Discussion and conclusions

In this work, we have introduced *SEMPER*, a statistical model for integrating epigenomic data in an intuitive, and truly *bottom-up*, polymer model of the chromatin fiber. Essentially, the model is based on the two following main assumptions (see the Hamiltonian (2.1)): that certain properties of the sequence (like the propensity to host specific transcription factors) determine the epigenetic fate of the sequence itself and that, in turn, chromatin sequences with the same epigenetic mark interact favorably with each other through an Ising-like term. These ingredients lead to, both, a good recapitulation of the epigenetic state of the sequence and the chromatin folding in $3D$ (Sec. 2.2).

Stemming from these assumptions, *SEMPER* aims at establishing a conceptual bridge between the following two lines of research in epigenomic data modeling which have been quite separate up to now: the $1D$ inherent information content of the genome and its $3D$ folding mediated by the polymeric properties of the chromatin fiber, all combined together into a novel Bayesian-based inference method. In this respect, the model can be viewed as a non-trivial extension of the (simpler) logistic regression (LR) predictor [86], and hence it connects to the rich line of research which develops machine learning methods for epigenomic prediction and imputation. We

show that *SEMPER* improves on LR on a number of example regions, likely due to the introduction of physically motivated local correlations between epigenomic marks at different monomers on a self-interacting chromatin filament. On the other hand, *SEMPER* is a *bona fide* statistical polymer model and hence fully mechanistic.

Remarkably, we show that the model is able to predict non-trivial structural features at a very high resolution without using any structural data and without including any *a priori* information or mechanistic knowledge concerning how chromatin folding works. Our work, in particular, confirms and strengthens previous results such as [90, 91, 92] on the importance of the polymeric nature of chromatin, while providing a new framework to integrate seamlessly the (epi)genomic data directly within a polymer model in a transparent and interpretable way.

As already remarked in the Introduction (2), ours is not the first polymer-based set-up where "epigenetically-related" chromatin sequences with like compositions attract each other. Amongst the proposed models, important differences emerge that make the different approaches not equivalent to each other: as an example, in some works [46, 102] the sequence is 'quenched', in others [106, 107, 108] the sequence is 'annealed' by changing in time as the consequence of the dynamic interplay between appropriate epigenetic remodellers and chromatin motion. More in general, all these models rely on some biological hypotheses and their "plausibility" is assessed based on their more or less favorable comparison with the available experimental observations (Hi-C, above all). The number of free parameters employed in the definition of some of these models is non-small and, in general, comparable to the number of parameters (*i.e.*, the number of "features") used in *SEMPER*; assessing parameters' robustness is crucial for the corresponding model, yet in general it is not an easy task. Conversely, the Bayesian-based strategy which *SEMPER* relies on mitigates this problem substantially and allows us to evaluate the parameters of the polymer model in a robust and interpretable fashion. Interestingly (see Sec. 2.2.2), this has led to a novel understanding of the role of the CpG density in histone modification *H3K27me3* w.r.t. the previous study [86] where the polymer was not taken into account.

At the same time, and despite its conceptual simplicity, Bayesian inference in *SEMPER* remains highly intractable, preventing the deployment of standard sampling algorithms. Therefore, a major contribution of our work has been the development of a novel approximate inference method that allows parameter estimation and uncertainty quantification. This algorithm is in principle applicable to a broad class of models where estimation of the partition function relies on computation of moments using model simulations.

While we believe *SEMPER* constitutes an important conceptual step, several improvements and future directions are possible. From the biological point of view, our work has explicitly focussed on a single histone mark, namely *H3K27me3*. It was chosen because of its well-known ubiquitous role in the formation of heterochromatic regions; nonetheless, a possible future development should be to consider multi-component spin degrees of freedom for each monomer ($\mathcal{O}(n)$-model [109]) in order to account for other possible histone marks (such as, for instance, *H3K9me3* or *H3K27ac* [95]) that also contribute shaping the architecture of the chromatin filament. From the modeling point of view, we have chosen to introduce the effect of sequence features via a linear external field. This is natural as a first step because of its connection with the logistic regression framework, but it can certainly be revisited to include more complex nonlinear effects. From the computational point of view, the cost of the simulations based on the pivot algorithm [94] remains high, preventing for the time being accurate fine-grain (and sufficiently large-scale!) treatment of the whole chromosomes. This drawback is maximally due to the need to perform extensive polymer simulations for each step of the posterior sampling algorithm. Future work devoted to finding more efficient sampling procedures by utilizing concepts from, *e.g.*, advanced machine learning [97] will likely further push the boundaries of physics-based models to analyze epigenomic data.

3

bioSBM: a random graph model to integrate epigenomic data in chromatin structure prediction

This chapter is based on the manuscript `arXiv:2409.14425` [110] and is ready to be submitted for peer review.

A characteristic feature of a eukaryotic cell, as opposed to archaea and eubacteria, is the sequestration of the cellular genome in a tight cellular space called the nucleus. In humans, the approximately two-meter-long chromatin filament is tightly packed in a nucleus of 5-10 μm in diameter. It has been clear that this packing is highly organized since early immunofluorescence microscopy experiments, which showed the heterogeneous subnuclear localization patterns of various proteins and histone marks, hinting at the existence of functionally distinct compartments within the nucleus [111, 112, 113, 114].

The advent of Chromatin Conformation Capture (3C) [25], in particular of the 3C derivative Hi-C[26], allowed the mapping of chromatin contacts at a genome-wide level. Thanks to the fixation of nuclei with formaldehyde, which preserves information about the spatial proximity of linearly distal DNA loci, Hi-C experiments generate contact frequency maps that display non-trivial interaction motifs. One notable feature of the organization of chromatin unveiled by Hi-C is the segregation of the genome into two classes of domains dubbed A and B compartments, which are characterized by distinct interaction patterns. On top of these connectivity differences, A and B compartments have been shown to correlate with epigenetic marks associated with

active (A) and silenced (B) transcriptional states. [26]. Despite these correlations, more detailed microscopy studies and information from epigenetics (e.g., histone modifications and preferential binding of transcription factors) [115, 74, 104] suggest that the transcriptional state of the genome is more nuanced and the binary classification into A/B compartments may be an oversimplification. While A/B compartments and the sub-compartments defined by Rao et al. (2014) [104] do provide valuable correlative links between distinct 3D chromatin interaction patterns and different epigenetic marks, a quantitative understanding of the relationship between 3D organization and underlying nuclear biochemical processes remains elusive.

Several attempts to address this issue have utilized polymer-based models, which aim to simulate chromatin structure by incorporating biophysical properties [90, 92, 46, 116]. While such models provide valuable insights by connecting some biochemical processes to mechanisms of genome folding, their reliance on polymer simulations results in high computational costs, particularly when additional parameters need to be estimated (as showcased by the work presented in Chapter 2). In contrast, more recent deep learning models scale efficiently and have achieved impressive success in predicting Hi-C contact maps. Early models make predictions using DNA sequence alone, but these predictions do not account for cell-type-specific variability [93, 117, 118]. More recently, deep learning approaches have started to incorporate 1D epigenetic signals [119], improving predictive accuracy across different cellular conditions. However, these models remain largely uninterpretable, making it difficult to connect their predictive power with underlying biological mechanisms.

In recent years, ideas from the field of network or graph theory [54] have emerged as a promising paradigm to study chromatin organization at the mesoscopic level. These methods avoid the computational overheads of microscopic polymer-based models by abstracting chromatin structure as a network of interactions, where DNA loci are treated as nodes and their contacts as edges. Such graph-based approaches not only were successfully used to reveal structural patterns [120, 55] but also provided interpretable insights into the relationship between chromatin architecture and biological function.[121, 57, 122]

In this chapter, we propose bioSBM, an interpretable network model that directly links chromatin structure with biochemical features. bioSBM is based on the stochastic block model (SBM) [58, 59], a class of generative network models that partition the network into communities based on interaction patterns, making them highly suitable for uncovering latent structures in chromatin interaction maps. bioSBM modulates this community structure by considering biochemical covariates such as histone modifications and transcription factor binding, therefore constructing a quan-

titative framework to understand the relationship between 3D chromatin organization and biochemical processes. Unlike traditional SBMs, which assign each node to a single community, bioSBM allows for mixed memberships, enabling genomic regions to participate in multiple communities simultaneously, thus capturing the context-dependent nature of chromatin interactions.

We apply our model to Hi-C data from the GM12878 lymphoblastoid cell line, where we identify interpretable community structures that include, but go beyond, the conventional A/B compartmentalization and subcompartments [104]. In addition to community detection, we can infer the map from biochemical features to the community composition of entire chromosomes and we learn the interaction patterns that link the various communities. Finally, the results of our inference allow us to show that bioSBM can serve as a generative model capable of predicting Hi-C maps for unseen chromosomes and cellular conditions, further demonstrating its robustness and utility.

The rest of the chapter is organized as follows. In Section 3.1 we provide a detailed overview of the *vanilla* stochastic block model and discuss how it can be adapted to better describe different types of analyzed data. This section also details the data we utilize (distance-corrected or observed over expected Hi-C maps), and our customized version of the SBM. In Section 3.1.2 we explain the inference algorithm developed to compute posterior probabilities for our model. For a more detailed mathematical derivation, refer to Appendix B.1. Finally, we present our results in Section 3.2, demonstrating the biological relevance and predictive power of the bioSBM model.

3.1 Stochastic Block Models

SBMs are a particular class of random graphs. A graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ consists of a set of vertices $\mathcal{V}$, representing entities $1, \ldots, N$ and a set of edges $(i, j) \in \mathcal{E} \subset \mathcal{V} \times \mathcal{V}$, denoting pairwise interactions between these entities. Edges can be binary, indicating the presence or the absence of a link, or they can be weighted, reflecting the *strength* of interactions. Random graphs [123, 124] denote graphs whose structure is itself a random variable, usually as a result of a random attachment process.

SBMs have their roots in the world of social sciences, where they were used to model populations divided into sub-populations or *communities*. The central idea is that interactions between individuals are influenced by their community, creating a non-trivial structure in the interaction graph. This simple yet powerful idea made block models into popular models to study general types of relational data.

Numerous algorithms have been developed to detect community structures in complex networks and to make sense of them [125, 87]. Box. 1 provides a schematic overview of the stochastic block model and of some of its most common extensions.

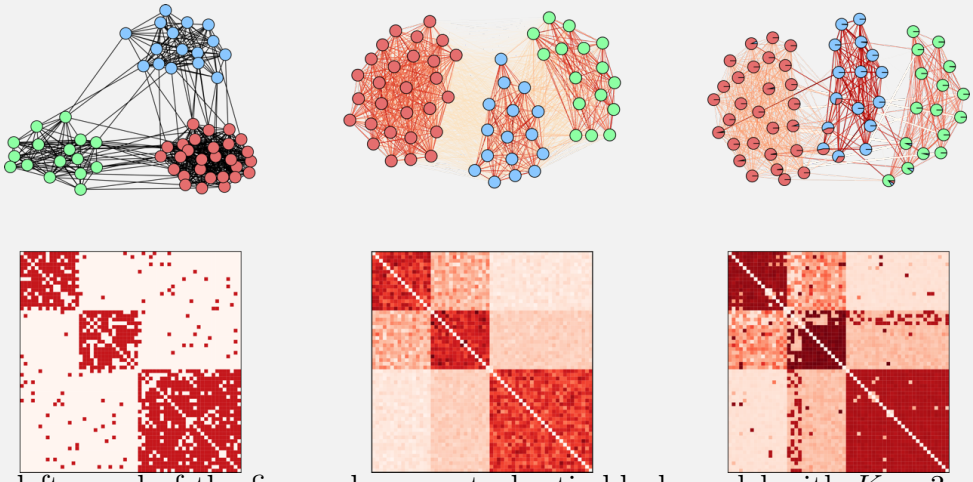
In our case, the relational data is derived from Hi-C contact frequency maps. Early Hi-C experiments demonstrated that contact frequency between genomic loci is strongly influenced by *linear proximity* or genomic distance, with the contact probability exhibiting a power-law decay as a function of this distance [26]. This scaling behavior is understood to arise from fundamental polymer physics mechanisms [41, 126, 83]. Instead of using raw Hi-C data, it is often more insightful to study the so-called *observed over expected* (OE) Hi-C maps. These OE maps are derived as the logarithmic ratio between the actual contact frequencies recorded in Hi-C matrices and the expected contact frequencies based on genomic distances. These maps effectively highlight interaction patterns of chromatin by accounting for and removing the global polymeric effects that contribute to the power-law scaling. Significant interactions can be observed across large genomic scales, sometimes spanning entire chromosomes. To uncover a latent structure in these interaction patterns, we employ a weighted stochastic block model. To capture the complexity of community structures we use a mixed membership version of the SBM, allowing genomic regions to belong to multiple classes.

Box 1: Flavours of Stochastic Block Models (SBM)

Stochastic Block Models (SBMs) are a type of generative model used in network theory to describe the structure of networks by dividing nodes into communities or blocks. Each block represents a group of nodes that have a similar pattern of connections. The SBM assumes that the probability of a connection between any two nodes depends only on the blocks to which the nodes belong. This model helps understand the network's underlying structure and is commonly used for community detection [125, 127, 128]. The generative process defined by an SBM is as follows:

- Determine the number of communities K
- Assign each node to one of the K communities
- For each pair of nodes, generate an edge with a probability that depends on the communities of the nodes. Specifically, generate a Bernoulli random variable with the parameter of the distribution depending on the colors of the two nodes.

Flavours of Stochastic Block Models (SBM)



The left panel of the figure shows a stochastic block model with $K = 3$ communities. In this example, the graph is an *assortative* SBM, meaning that intra-class edge probabilities are higher than inter-class ones.

In its basic version, the SBM is binary i.e., the edges between nodes are either present or absent. However, many real-world networks involve weighted edges, where the connections between nodes have different strengths or capacities. To adapt the binary SBM for valued (weighted) graphs, we can modify the probability distribution of the edges given the colors or communities to which the two involved nodes belong to. Instead of a Bernoulli random variable, we might use Poisson random variables for integer value edges, or Gaussian random variables for real-valued networks, specifying the means and variances of the distributions for each pair of distinct communities [129]. The central panel shows an instance of such weighted SBM.

Another aspect we can tweak is the fact that in the traditional SBM, each node belongs to a single community or block. In many networks, nodes may exhibit characteristics of multiple communities. The Mixed Membership Stochastic block Model (MMSBM) [130] addresses this by allowing nodes to belong to multiple communities by specifying a probability distribution over classes, or membership proportions.

In this chapter, we will work with an SBM that has real-valued edges and mixed membership proportions. The right panel shows an example of weighted MMSBM.

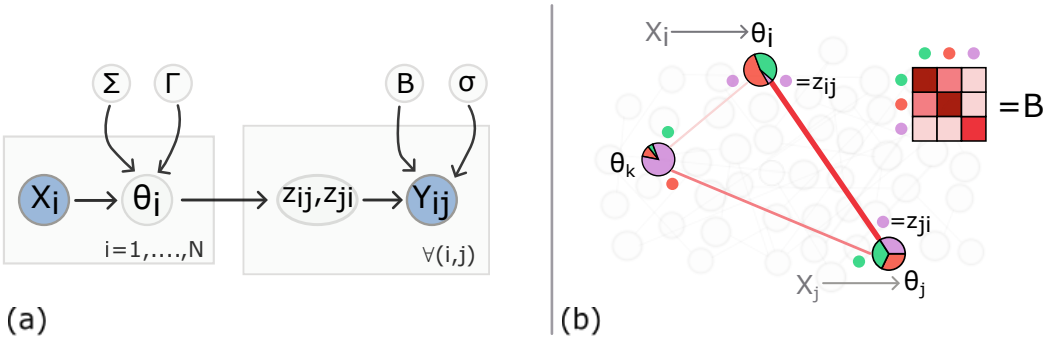


Figure 3.1: Generative process of bioSBM. a) Hierarchical Bayesian model representing the generative process. The colored circles represent observed variables, which are the edge weights Y_{ij} and the node covariates vectors x_i. The connectivity patterns of the graph are determined by per-node latent class membership vectors θ_i and global hyperparameters represented by the circles outside the rectangles. b) An example displays the sampling for the weights of edges between a subset of three nodes.

3.1.1 Generative process of bioSBM

A key difference between standard relational data and the biological setting we consider is the availability of a wealth of additional data in biology. While Hi-C measures contact patterns between chromosomal regions, a variety of biochemical assays, such as ChIP-seq, provide 1D genomic maps of specific epigenomic marks at such regions. In our model, we integrate this ChIP-seq data as a vector of biochemical covariates associated with each node, and which modulates the probability of each node belonging to the different communities.

Formally, bioSBM is a hierarchical Bayesian model [131, 132] over weighted graphs. Graph nodes $i \in \{1, \ldots, N\}$ represent a set of contiguous genomic regions of fixed length. The observed weighted network has adjacency matrix Y, with Y_{ij} representing the logarithmic observed over expected Hi-C contact frequency between region i and j. The observed weighted network is assumed to be generated according to the latent distributions of group memberships for each node/genomic region, as well as the matrices that specify group-group interaction strengths. Each node i has an associated membership proportions' vector θ_i, where $\theta_{i,g}$ denotes the probability of node i belonging to group g, allowing nodes to belong to multiple classes and display interactions that are context-dependent.

The group-group interaction strengths are defined by matrices $\mathbf{B}$ and σ, where $B_{g,k}$ and $\sigma^2_{g,k}$ represent the mean and variance of the strength of interaction between class g and class k. For each pair of nodes (i, j), discrete variables $z_{i,j}$ and $z_{j,i}$ denote the group membership of i when interacting with j, and vice-versa. The edge weight

is then sampled from a Gaussian distribution parameterized by B and σ matrices.

Putting everything together, the generative process for *bio*-SBM proceedes as follows:

- Each node has an associated vector of features P $\boldsymbol{x}_i$, with $\boldsymbol{X} \in \mathbb{R}^{P \times N}$ denoting the covariate matrix. The covariates correspond to biochemical data that can be associated with the different genomic regions, such as data from ChIP-seq assays.

- For every node i we sample a distribution $\boldsymbol{\theta}_i$ from the logistic normal distribution [133, 134] with mean $\boldsymbol{\mu_i} = \Gamma(\boldsymbol{x}_i)$and global covariance Σ

$$\boldsymbol{\eta}_i \sim \mathcal{N}_K(\Gamma(\boldsymbol{x}_i), \boldsymbol{\Sigma}) \tag{3.1}$$

$$\theta_{i,k} = \frac{\exp(\eta_{i,k})}{\sum_{k'=1}^{K} \exp \eta_{i,k'}} \tag{3.2}$$

  $\Gamma : \mathbb{R}^P \to \mathbb{R}^K$ is the parametric function that maps biochemical features to probabilities over group memberships. Notice that Γ and Σ are global parameters shared among all nodes. In our specific implementation, Γ is simply a linear transformation encoded in a $K \times P$ matrix. Notice that Γ and Σ are global parameters shared among all nodes.

- For every pair of nodes (i, j) with $i = 1, \ldots, N$ and $j = 1, \ldots, i-1$

$$z_{i,j} \sim \text{Multinomial}(\boldsymbol{\theta}_i) \text{ membership of } i \text{ interacting with } j \tag{3.3}$$

$$z_{j,i} \sim \text{Multinomial}(\boldsymbol{\theta}_j) \text{ membership of } j \text{ interacting with } i \tag{3.4}$$

- Once the memberships of i and j are sampled the weight of the edge is sampled from a Gaussian of which parameters are encoded in the global B and σ matrices.

$$P(Y_{ij} | z_{i,j} = k, z_{j,i} = g, B, \sigma) = \mathcal{N}(Y_{i,j} | B_{k,g}, \sigma_{k,g}) \tag{3.5}$$

Schematically, the bioSBM model is represented as a graphical model as in Figure 3.1.

3.1.2 Variational inference

To uncover the latent structure of chromatin interactions, we developed a posterior inference algorithm tailored to bioSBM. This algorithm estimates the latent parameters that best explain the observed Hi-C interaction data, integrating both chromatin interaction frequencies and biochemical covariates.

Our approach is based on variational inference, a method well-suited for complex probabilistic models like the bioSBM, where exact inference is intractable. We

optimize a variational lower bound on the model evidence, commonly referred to as the Evidence Lower Bound (ELBO), which enables us to approximate the posterior distribution of the latent variables.

The variational inference procedure optimizes the ELBO, defined as:

$$\mathcal{L}(q, \boldsymbol{\Psi}) = \mathbb{E}_q \left[\log P(\boldsymbol{Y}, \boldsymbol{\eta}_{1:N}, \boldsymbol{Z}|\boldsymbol{\Psi}, \boldsymbol{X}) - \log q(\boldsymbol{\eta}_{1:N}, \boldsymbol{Z})\right] \tag{3.6}$$

where Y represents the observed Hi-C interaction data, X denotes the biochemical covariates, $\eta_{1:N}$ are the latent membership vetors ($\theta_{1:N}$ are the normalized versions), Z represents the community assignments for edges, and Ψ includes the global model parameters. The variational distribution $q(\eta_{1:N}, Z)$ approximates the true posterior distribution $P(\eta_{1:N}, Z|Y, X, \Psi)$.

The algorithm proceeds in two main steps:

1. **Variational E-step**: We update the variational distributions of the latent variables, η_i and z_{ij}, by maximizing the ELBO with respect to the variational parameters. The factorized variational distributions take the form:

$$q(\boldsymbol{\eta}_i) \propto \exp\left\{\log P(\boldsymbol{\eta}_i|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}) + \mathbb{E}_{q(\boldsymbol{Z})}[\log P(\boldsymbol{Z}|\boldsymbol{\eta}_i)]\right\} \tag{3.7}$$

$$q(\boldsymbol{z}_{ij}) \propto \exp\left\{\mathbb{E}_{q(\boldsymbol{z}_{ji})}[\log P(Y_{ij}|\boldsymbol{z}_{ij}, \boldsymbol{z}_{ji}, \boldsymbol{B})] + \mathbb{E}_{q(\boldsymbol{\eta}_i)}[\log P(\boldsymbol{z}_{ij}|\boldsymbol{\eta}_i)]\right\} \tag{3.8}$$

   Here, η_i are the continuous latent membership vectors, and z_{ij} are the discrete community assignments for edges.

2. **Variational M-step**: This step involves optimizing the model parameters $\Psi \equiv (\Sigma, \Gamma, B, \sigma)$ with the current estimates of the variational distributions. The matrix Γ maps the biochemical covariates to the latent space, while Σ is the covariance matrix capturing the variability in the latent memberships. The matrices B and σ describe the mean interaction strengths and variances between communities.

The iterative process of alternating between the E-step and the M-step continues until convergence, at which point the model parameters and variational distributions jointly provide an interpretation of the chromatin interaction patterns. For the detailed mathematical derivations and specific parameter update rules, refer to Appendix B

3.2 Results

To train the model, we perform posterior inference using M pairs of biochemical covariates and Hi-C matrices $(X^\mu, Y^\mu)_{\mu=1}^M$, computing approximate posterior distri-

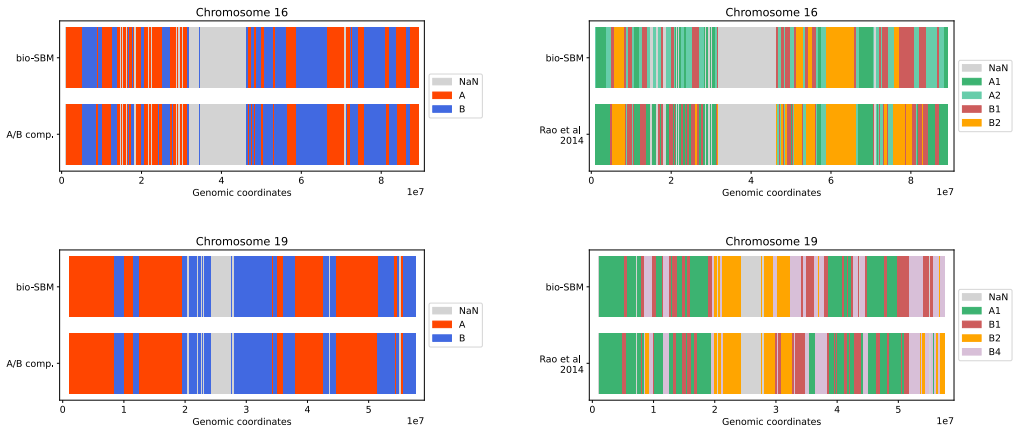


Figure 3.2: The latent representation found through inference of the bioSBM model is biologically meaningful. The top row in each plot represents the clusters obtained by applying k-means on the MAP membership vectors inferred through our algorithm. The bottom rows are biological annotations.

butions over per-node latent membership vectors θ_i and the model parameters. In particular, the two sets of chromosomes we use are the odd-numbered chromosomes from 11 to 21 and the even-numbered chromosomes from 12 to 22. We use the model trained on one set to make predictions on the other and vice-versa. We applied the inference algorithm to data from the GM12878 lymphoblastoid cell line at a resolution of 100 kbps, and through Bayesian model selection, using evidence lower bound (ELBO) as a criterion, we determined that the optimal number of communities for our model is $K = 12$ (Figure B.2). This suggests a more complex organization structure that extends beyond the conventional binary A/B compartmentalization and the more nuanced subdivision in subcompartments (up to six) [104]. At the same time, our model provides a more concise representation of chromatin contacts than the 30 epigenetically distinct "*binding domains*" suggested by Esposito et al [135] using polymer modeling.

The Maximum A Posteriori (MAP) estimates of the vectors θ_i provide the most plausible community membership proportions for each node, based on the observation of experimental Hi-C maps and associated biochemical covariates. Along with the θ_i values, the inference process also estimates the global parameters that characterize the generative model. A key parameter is the matrix Γ, which represents the linear transformation mapping biochemical features to the probabilities of belonging to each community, offering insights into how biochemical factors shape chromatin structure. Additionally, the matrix B encodes the interaction strengths between all pairs of communities.

3.2.1 bioSBM explains the hierarchical organization of the chromatin in terms of epigenomic marks

To evaluate the biological relevance of the communities identified by our model, we performed k-means clustering on the MAP membership vectors θ_i. Clustering into two groups allowed us to compare these clusters to known A/B compartments, while varying the number of clusters, depending on the number of significantly present subcompartments in each chromosome, enabled comparisons with the subcompartments defined by Rao et al. [104].

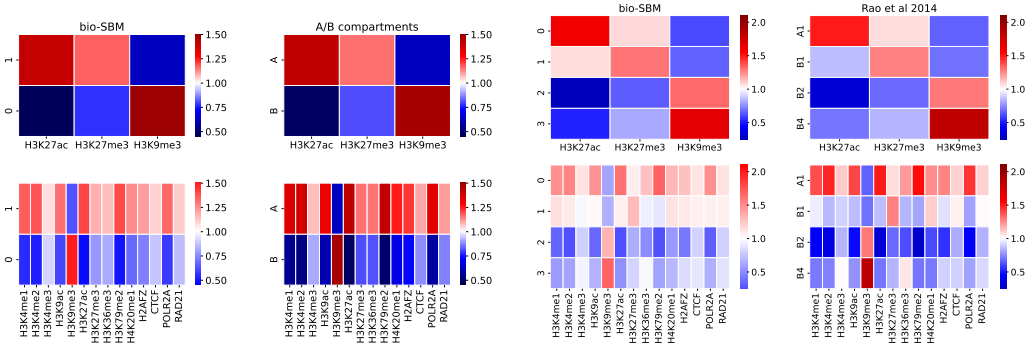


Figure 3.3: Log fold enrichment in biochemical features (such as the presence of histone marks) for chromosome 19. The patterns of enrichment for our clusters are in good agreement with the enrichment found for A/B compartments and subcompartments.

The results of the clustering showed a significant overlap with known biological annotations. Figure 3.2 illustrates this comparison for chromosomes 16 and 19. Both the binary subdivision in A/B compartments and the more granular classification in subcompartments can be captured from the full mixed membership vectors inferred by our model. Further validation of the inferred communities was performed by assessing the enrichment of each k-means-derived cluster in the biochemical features x_i. These enrichments were compared to those observed in A/B compartments and subcompartments annotations (See Figure 3.3), where we can see a near-perfect agreement of bioSBM results with the enrichment of the binary A/B classification and a very good agreement with the enrichment of the subcompartments defined by [104].

We extended this analysis to other chromosomes (see Table B.1 and Section B.2 for details on datasets used) and computed the similarity between the subdivision found by clustering the membership vectors and those based on the biological annotations using the normalized mutual information (NMI) score. We obtained a median NMI score (which ranges from 0 to 1) of $NMI^{AB} \simeq 0.79$ for the binary clustering, and $NMI^{subcomp} \simeq 0.49$ for the subcompartments. These results suggest that while our model's communities largely align with known biological structures, the correspon-

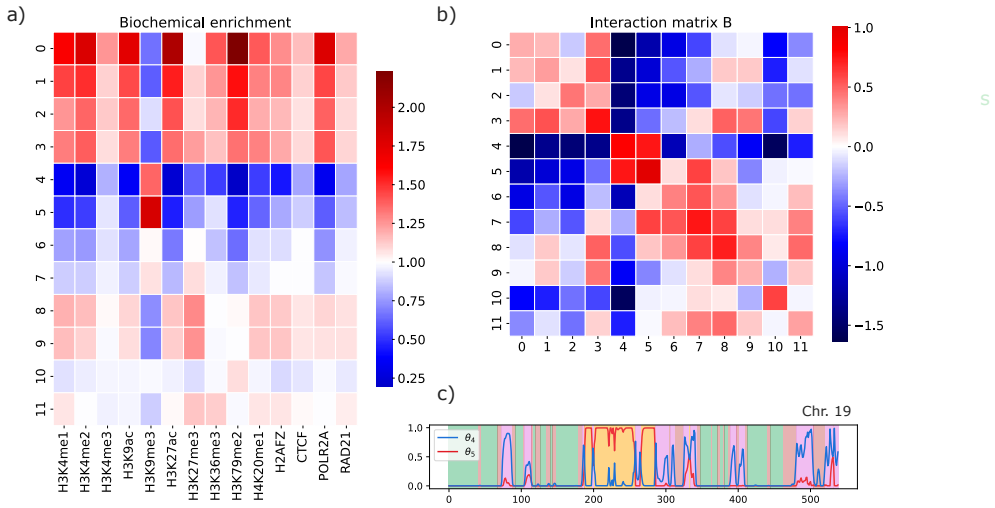


Figure 3.4: a) Enrichment of mixed membership communities in the biochemical features used in the model. Each community correlates with distinct epigenetic features. b) Interaction intensities between all pairs of communities. c) Membership proportions of communities 4 and 5 as a function of node indices.

dence is incomplete. The partial agreement with subcompartments suggests that our model may capture additional layers of chromatin interaction complexity that may be missed by conventional classification methods.

Importantly, the model goes beyond merely segmenting chromatin regions based on their epigenetic features; it also illustrates how these different classes interact, as represented by the matrix B, where each entry B_{kg} encodes the interaction strength between class k and g. Figure 3.4 shows that each of the 12 inferred communities is associated both with distinct epigenomic patterns and with interaction patterns between communities. Interestingly, despite the apparent redundancy for some of the communities at the level of epigenomic profiles, there are clear distinctions between the interaction patterns of different classes. For instance, classes 0 and 2 have similar epigenetic profiles but regions predominantly associated with class 0 have reduced contact frequency with regions primarily linked to class 2, and vice versa. Another notable observation is that classes 4 and 5 share a similar enrichment profile but differ in interaction patterns. Class 4 interacts only with itself and the epigenetically similar class 5, while class 5 interacts positively with classes 6, 7, and 8 as well. Figure 3.4 (c) shows that regions with non-zero probabilities of belonging to classes 4 and 5 overlap with clusters corresponding to the B2 and B4 compartments. Here however nodes need not be categorized into one class or another, they can share properties of multiple communities in different proportions. See Figure B.3 for a comparison between mixed-membership and single-membership SBM.

Altogether these observations show that the bioSBM representation of chromatin

interaction patterns provides a more nuanced description than the one provided by simple segmentation of the genome in different clusters.

3.2.2 bioSBM predicts the structure of the chromatin in unseen cell lines

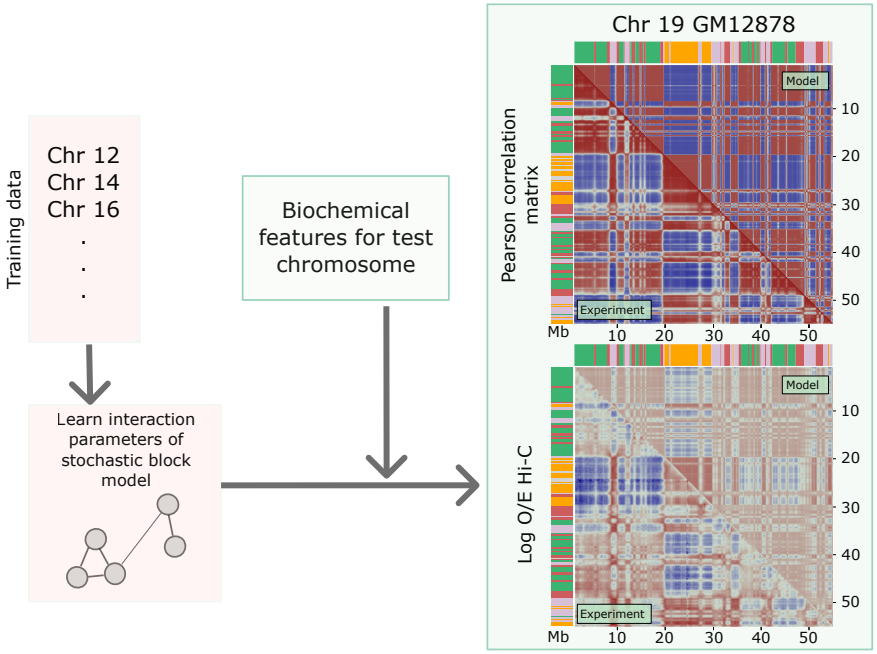


Figure 3.5: The figure shows the Hi-C maps and the Pearson correlation matrix for chromosome 19. These maps are computed with the biochemical covariates associated with the chromosome and the parameters for the stochastic block model learned on a different set of chromosomes. The bottom triangle of each map is the experimentally observed Hi-C or Pearson correlation matrix, the top triangles are the model-generated counterparts. Above each map, the colored bars represent the segmentation of the genome arising from k-means clustering of the model mixed-membership vectors, while the bars on the left of each map are the subcompartment annotations from [104]

The previous section focussed on an analysis of the interpretability of bioSBM, showing that the inferred model parameters recapitulate and extend previous observations on the epigenetic state of the chromatin and its compartments. In this section, we leverage the generative structure of bioSBM to test whether the simple representation of the genome contacts in terms of interactions between a limited number of communities and the fact that these communities can be determined starting from independent measurements of biochemical covariates is enough to reproduce long-range genome-wide chromatin contact patterns.

Specifically, given matrices B and Γ, inferred from some training chromosomes in some conditions, and biochemical measurements for test chromosomes (either differ-

ent chromosomes in the same cell line, or chromosomes from a different cell line), we can compute their community structure as $\theta_i = \Gamma x_i$. With the predicted community structure, we use the matrix B to sample contact maps and compute their expectations, which can then be compared to experimentally obtained Hi-C maps (refer to Section B.3 for technical details).

Figure 3.5 illustrates the testing pipeline, providing an example of the predicted Hi-C map and predicted Pearson correlation matrix for chromosome 19. Here, biochemical covariates associated with chromosome 19 were used, while the block model interaction parameters were inferred from a different set of chromosomes. The strong correlation between experimental and predicted Hi-C maps, as quantified by Pearson correlation coefficients (Figure 3.6), demonstrates the model's predictive accuracy. Interestingly, the difference in accuracy between the training and test sets was marginal, with a one-tailed Mann-Whitney U test yielding a p-value of 0.12. Therefore, bioSBM effectively generalizes across different chromosomes within the same cell line, suggesting that the inferred associations B_{kg} between communities and the biochemistry-to-structure map Γ reflect genuine chromatin interactions (see Figure B.4 for a comparison between the results obtained with a linear Γ and a neural network parametrization of the mapping between covariates and membership proportions).

Further support for this conclusion comes from a comparison with results reported by Esposito et al (2022) [135], where a code linking 1D chromatin marks to the interaction parameters of a polymer model was learned. While the polymer model similarly recapitulates chromatin contact patterns in training data, testing on unseen chromosomes from the GM12878 cell line strongly favors the bioSBM model in terms of predictive performance (see Fig. 3.6).

Finally, we tested the model on data from the human colorectal carcinoma cell line HCT116, where auxin-inducible degron (AID) technology was used to degrade RAD21 [136]. We apply the model as trained on GM12878 to independent covariates from the ENCODE Project [99] obtained from experiments on the HCT116 cell line without AID-induced RAD21 degradation. In this case, we observe a decrease in predictive accuracy, likely due to the mismatch between the learned map Γ, which includes RAD21, and a biological sample where the latter has been degraded. Additionally, RAD21 deletion could have induced slight changes also in the other covariates, further affecting predictive performance. Notably, when the RAD21 covariate was excluded from the model (by setting the corresponding row in the feature matrix X to zero), performance improved significantly (p-value = 0.03), suggesting the flexibility and robustness of bioSBM in capturing chromatin interactions under different cellular conditions.

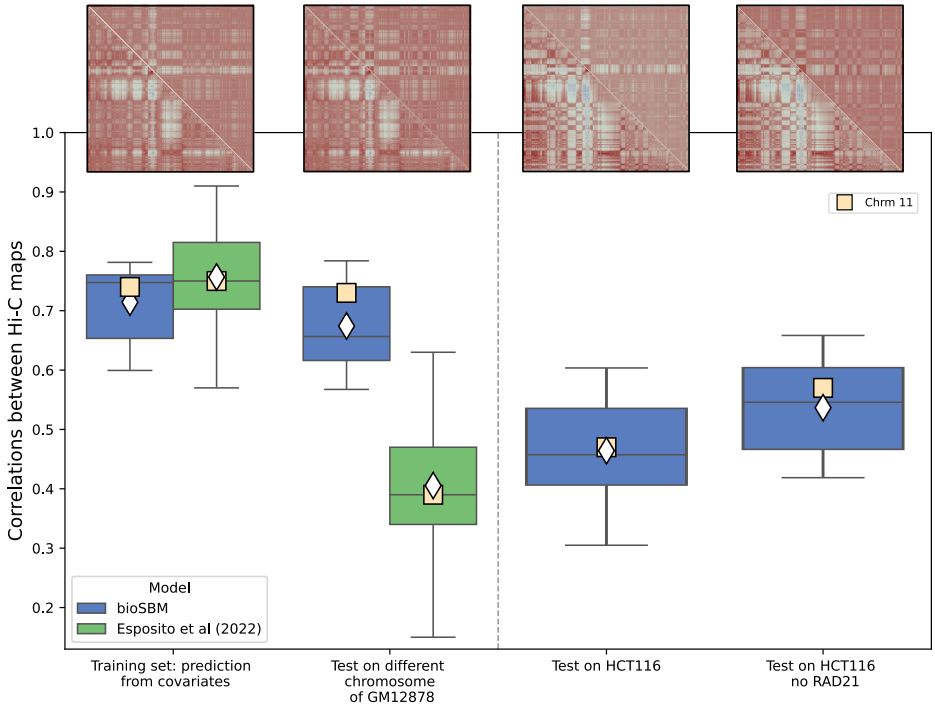


Figure 3.6: Box plots of correlations between experimental maps and model-generated maps, where the individual data points are single chromosomes. White diamonds represent the means of the box plots and the ivory squares are the correlation values for the example chromosome 11. The insets show comparisons of experimental maps (lower triangular part) and model-generated maps (upper triangle). To test whether the training performance was significantly higher or not than the test on different chromosomes on the same cell line we performed a one-tailed Mann-Whitney U test, that yielded a negative answer (p-value = 0.12). For tests on HCT116 we see that the correlations are higher in the RAD21 depleted scenario (tested for significance with a p-value = 0.03). The green box plots refer to correlations reported by Esposito et al (2022) [135] for a polymer-based model on the same dataset.

3.3 Discussion and conclusions

The bioSBM model introduced in this study offers a novel approach for modeling long-range chromatin interactions by integrating Hi-C data with biochemical covariates such as histone modifications and binding of transcription factors. By employing a mixed membership stochastic block model, we capture a more refined and nuanced view of chromatin structure, extending beyond the traditional binary A/B and subcompartments framework. Our results show that the 12 communities identified by bioSBM correlate with known epigenetic features, reinforcing the idea that chromatin interactions are closely tied to the biochemical landscape of the genome. The partial agreement with subcompartments suggests that our model may capture additional layers of chromatin interaction complexity that may be missed by conventional clas-

sification methods.

Although more abstract than polymer models, which explicitly take into account (to varying degrees of detail) the physical nature of the linear chromosomes, the latent representation learned by bioSBM remains biologically interpretable. The inferred associations between biochemistry and structure encoded in the linear map Γ and the communities interactions contained in B provide a starting point for systematically exploring more mechanistic descriptions of how chromatin folding is affected by nuclear biochemical processes.

The predictive power of the model is another important contribution. By leveraging biochemical covariates, bioSBM can accurately predict chromatin contacts across different chromosomes and cell lines, comparing favorably with state-of-the-art polymer approaches [135]. Notably, the model's robustness is highlighted by its performance on the HCT116 cell line, where the removal of the RAD21 covariate improved predictive accuracy, indicating that bioSBM can adapt to different chromatin environments and capture interactions under varying conditions, such as RAD21 depletion. Interstingly, bioSBM also compares favorably with recent deep-learning models. Many deep learning approaches predict Hi-C maps using only DNA sequence data as input features [93, 117], which limits their ability to capture cell-type specific variations. In contrast, bioSB explicitly incorporates epigenetic features, making it more suited for cell-type-specific predictions. Additionally, bioSBM's predictive power is comparable to a recent deep learning model that does include epigenetic signals [119]. However, deep learning models have a distinct advantage in terms of computational power, allowing them to predict Hi-c maps at higher resolutions and model small-scale perturbations, such as the effects of single nucleotide polymorphisms (SNPs) [93]. While our model cannot operate at this fine-grained scale, it remains effective in predicting the effects of larger-scale manipulations such as the AID-induced RAD21 reported above. Crucially, bioSBM stands out for its interpretability: every parameter has a clear probabilistic semantic, allowing their analysis to uncover a clear biological meaning. This makes bioSBM a valuable, flexible tool for studying chromatin organization in diverse cellular contexts.

While this study focuses on Hi-C data, the so-called called enrichment methods [137], such as Promoter Capture Hi-C (PCHi-C) [138] or ChIA-PET [139, 140], may offer more functionally relevant perspectives on chromatin structure. PCHi-C, for example, highlights promoter regions interactions, while ChIA-PET focuses on interactions involving specific proteins. These methods present challenges for polymer models because they capture interactions between non-contiguous regions. However, bioSBM, being graph-based, could accommodate these non-contiguous interactions

with some adaptation. Though some graph-based studies exist that link nuclear biochemistry to chromatin structure as assayed by these enrichment-based data [121, 57], a fully generative predictive model mapping biochemistry to the chromatin interaction patterns, such as that provided by bioSBM for Hi-C, has yet to be developed. With appropriate adjustments, bioSBM could readily be extended to accommodate these data types.

In conclusion, the bioSBM model successfully balances interpretability and scalability, offering a valuable tool for understanding the relationship between chromatin structure and its biochemical underpinnings. By incorporating biochemical features and allowing for mixed memberships, bioSBM provides a more flexible and biologically meaningful representation of chromatin interactions. The model's predictive power and adaptability across different cellular contexts underscore its potential for further applications, such as exploring chromatin dynamics in different developmental stages or disease states.

4

Correlational information and non-local decoding of morphogen gradients

This chapter is based on unpublished work done in collaboration with David Brückner and Gašper Tkačik at the Institute of Science and Technology Austria (ISTA), Klosterneuburg, Austria.

In developing tissues or organisms, cells undergo differentiation, a process in which pluripotent stem cells commit to specific cell fates. To ensure proper biological function, differentiation occurs in spatial patterns that are both precise and reproducible [141, 142]. In the early 1900s, scientists first proposed that cells determine their fates by sensing the presence of chemicals in the cytoplasm, the so-called fate-determining factors [143, 144, 145]. In 1952, Alan Turing introduced the term *'morphogens'* for these factors and proposed mathematical models (reaction-diffusion systems) to describe how spatial patterns of morphogens form [146].

A few decades later, Lewis Wolpert introduced the seminal idea that cells acquire their identities by '*reading out*' the information encoded in graded profiles of these morphogens distributions, a concept he termed *Positional Information* (PI) [147, 148, 149, 150]. In essence, the idea is that the spatial position of each cell within a developing organism or tissue is encoded in the concentration gradients of morphogen molecules. Cells then interpret these gradients to make cell-fate decisions.

Wolpert's intuitive idea remained theoretical until 1988 when Driever and Nüsslein-Volhard provided the first experimental visualization of a morphogen gra-

dient. They observed that the transcription factor *Bicoid* establishes a concentration gradient along the anterior-posterior (AP) axis of the *Drosophila melanogaster* embryo, extending from the anterior end to over 50% of the embryo length [151, 152]. Following this discovery, numerous PI-carrying morphogens were identified in *Drosophila* and other organisms [153, 154, 155]. Perturbations of these morphogens were shown to systematically disrupt morphological features that are otherwise highly reproducible across wild-type embryos [156, 157, 158, 159], establishing a causal but qualitative connection between positional information and the resulting embryonic body plans.

Recent efforts have aimed at developing a quantitative understanding of Positional Information. A critical consideration is that the encoding of cellular positions in the concentrations of patterning molecules is inherently stochastic. Morphogen molecules are often present in low absolute copy numbers, making the measurement of local concentrations by cells noisy. Additionally, variability between embryos, such as differences in morphogen source strength, introduces further randomness. Consequently, the cellular machinery that interprets morphogen gradients must account for this stochasticity. First Gregor et al. [160] quantitatively demonstrated how this stochasticity imposes physical limits on the precision and reproducibility of the information that cells can extract from morphogens. Following this, several works formulated a rigorous mathematical framework for optimal decoding, which aimed at understanding how cells can infer position most accurately given these stochastic constraints [161, 162, 163]. Subsequently, other research extended these ideas by formalizing Positional Information within the framework of information theory. This body of work connected information theory to the concepts of optimal decoding and successfully applied these ideas to experimental data [16, 164, 18, 19].

The theoretical frameworks developed so far have been instrumental in demonstrating that developing organisms can operate near the physical limits of precision when specifying their body plans. These frameworks provide a robust understanding of the constraints imposed by stochasticity on morphogen-based patterning. However, several outstanding questions remain, particularly regarding the origin and refinement of Positional Information (PI).

At the onset of the embryogenesis for a wide range of organisms (such as the mammalian embryo), the developing organism consists of a single fertilized egg that undergoes successive divisions. During these early stages, all cells are nearly identical, and there is no Positional Information. The process by which PI emerges and becomes progressively refined through subsequent developmental steps is not yet fully understood. This refinement likely involves a sequence of processes that gradually increase

the precision of spatial patterning. In many cases, initial pattern-forming processes are thought to be subject to various sources of noise, including both extrinsic factors (such as environmental variability) and intrinsic fluctuations (arising from the inherent randomness of molecular interactions). To counter such noise, downstream operations such as diffusive couplings or cell-cell interactions, may allow cells to collectively reduce these fluctuations. Investigating the dynamics of these pattern-forming systems has led to the development of several mechanistic physical models, including reaction-diffusion, dynamical systems, and mechano-chemical models [165, 166, 167]. However, an approach to quantify the information that may be extracted along the sequence of developmental steps, and to predict the optimal strategy to do so, remains outstanding.

A striking example of how PI can be refined during development is seen in *Drosophila*, where the maternal Bicoid gradient carries considerably lower PI than the PI carried by the spatial profile of the downstream gene Hunchback [16]. One possible explanation is that Hunchback may receive additional PI from inputs other than Bicoid [18]. Another possible reason why Hunchback carries more PI than Bicoid is that the Hunchback readout mechanism performs temporal and spatial (through diffusion) averaging to reduce noise. [160, 168, 169]. Although these mechanisms may play an important role, they are subject to biophysical limits. For instance, spatial averaging is in tradeoff with having sharp spatial gradients, which, as we will see in the next section, contribute positively to a high PI. Spatial averaging can reduce noise if the morphogen gradient is slowly varying, but if the averaging window is too large, information-carrying morphogen gradients would be 'flattened'. Another instance in which spatial averaging is likely to be less useful is when gene expression fluctuations are not independent but are spatially correlated [170].

Given these limitations, we address the question of what alternative strategies that do not solely rely on averaging may be. One intriguing possibility we explore is whether spatial correlations in gene expression fluctuations can be exploited rather than mitigated. By quantifying the amount of information that can be extracted from these correlations, we aim to shed light on potential mechanisms by which cells collectively enhance developmental precision.

In Section 4.1 I will review the information-theoretic framework of PI and how it links to ideas to optimal decoding. The quantities defined in this section will be useful for reading the remainder of the chapter. In Section 4.2 I define and characterize a quantity called correlational information (CI) [171] that measures how correlated in space are the fluctuations in gene expression for a certain morphogen. Then in Section 4.3 I explore how cells can infer their position using non-local decoding and how

the information gains granted by this strategy are linked to correlational information. Finally in Section 4.4 I compute effective linear filters that approximate the optimal non-local decoder. This will give us insights into possible molecular mechanisms that with a local or semilocal transformation can approximately implement the non-local decoding operation.

4.1 What does it mean for a cell to know its position? Overview on Positional Information, positional error, and decoding maps

4.1.1 Positional Information

To introduce the theoretical framework of Positional Information, let us consider a simple scenario: a single gene, denoted as $g(x)$, with its expression level varying according to the position x along a one-dimensional embryo. Experimentally $g(x)$ can be measured as the light intensity profile of fluorescently labeled antibodies against a particular gene product [160, 16]. Imagine performing this measurement for a batch of $\mathcal{N}$ embryos where each profiles is labeled as $g^{(\mu)}$ with $\mu = 1, \ldots, \mathcal{N}$. These profiles differ due to the intrinsic stochasticity of the molecular process involved in their establishment and because of extrinsic embryo-to-embryo variability. From these profiles, we can compute for each position the histogram $P(g|x)$, which represents the noisy encoding of position in the concentrations of morphogen gradients.

A useful approximation, which has been experimentally validated for some systems (e.g. *Drosophila*) [16, 18], is to assume that $P(g|x)$ is Gaussian. In this case, it suffices to estimate the average expression pattern across embryos $\bar{g}(x)$, and $\sigma_g^2(x)$, the position-dependent variance of the gene expression level at position x.

Figure 4.1 illustrates two examples mean profiles $\bar{g}(x)$ to build some intuition about the features of an ensemble of morphogen gradients that contribute to higher PI. In the step-function profile on the left, the morphogen gradient allows for the distinction of at most two states (before and after the step). In contrast, the graded profile on the right suggests the possibility of discerning more states, depending on noise levels. We can also observe that in the anterior part (small x), positional encoding is likely more precise, contributing more to PI, while in the posterior part, where the profile is almost flat, the contribution to PI is small like in the flat parts of the step-function.

To formalize these intuitive ideas, Dubuis et al [16], introduced a mathematical

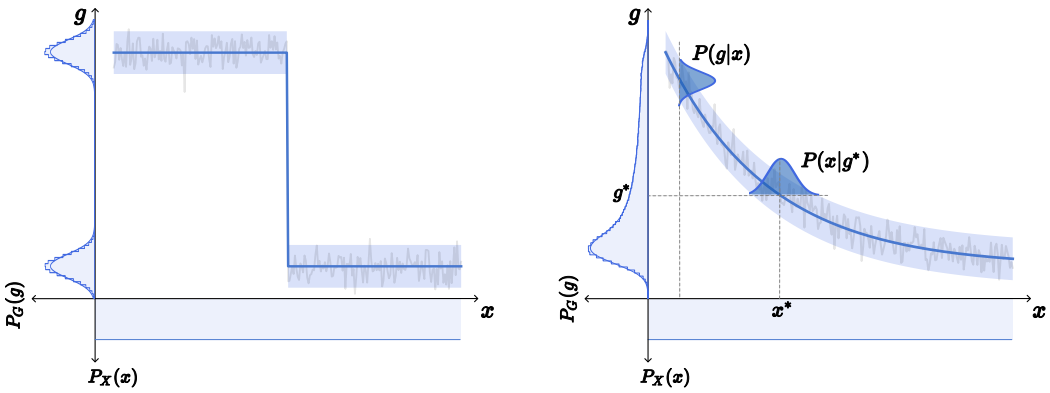


Figure 4.1: Examples of morphogen profiles with different levels of Positional Information (PI). The left panel shows a step-function profile, which can distinguish two states, resulting in a PI of 1 bit. The right panel shows a graded profile, which allows for more precise position discrimination, resulting in a PI of approximately 2.2 bits. The amount of positional information depends on the shape of the gradient and noise levels.

formulation of PI, defining it as the mutual information between the random variable X (representing position) and G (representing gene expression):

$$PI := I(X,G) = \int dx P_X(x) \int dg P(g|x) \log_2 \frac{P(g|x)}{P_G(g)} \tag{4.1}$$

PI measures the amount of information about the position x contained in the morphogen gradient and is expressed in *bits*. In Figure 4.1, the noisy step function carries 1 *bit* of positional information, indicating the ability to distinguish between two states. For the graded profile, PI is approximately 2.2 bits. If $PI = \log_2 N$, where N is the number of cells in the embryo, then the morphogen gradient provides enough information to distinguish all the cells.

An alternative expression for the mutual information which can be derived from Eq. 4.1 is:

$$I(X,G) = S[P_G(g)] - \langle S[P(g|x)] \rangle_x \tag{4.2}$$

Where $S[\cdot]$ denotes the Shannon entropy of a distribution [60]. This expression shows that PI decreases as the entropy of $P(g|x)$ increases, which formalizes the intuitive idea that higher noise levels lead to reduced positional information.

4.1.2 Decoding maps and positional error

Instead of a measure of statistical dependency, we can interpret PI in an alternative way if we re-write $I(X,G)$ as

$$I(X,G) = S[P_X(x)] - \langle S[P(x|g)] \rangle_g \tag{4.3}$$

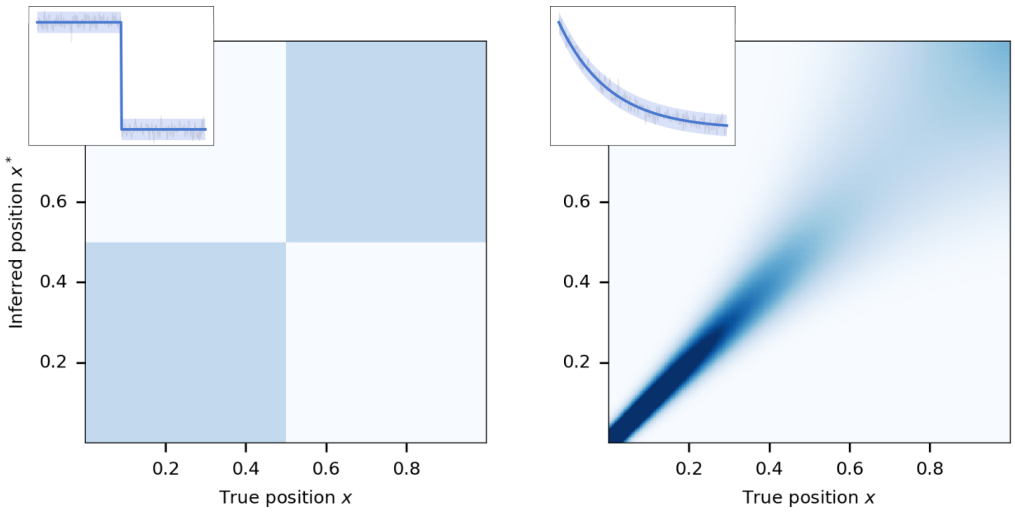


Figure 4.2: Decoding maps illustrating the relationship between actual position (x) and inferred position (x^*) based on morphogen concentrations. The heatmaps show the joint probability distribution $P(x, x^*)$, where the concentration around the diagonal indicates accurate decoding. For the step-function profile, the map shows two uniform regions, while the graded profile reveals more precise decoding in regions with steeper gradients.

Here, $S[P_X(x)]$ measures the initial uncertainty about position *prior* to measuring morphogen concentrations, while $S[P(x|g)]$ indicates the entropy of the posterior distribution of x once gene expression level g has been observed. $P_X(x)$ can be for instance a uniform distribution if there are no morphological differences between the different cells. This perspective highlights the so-called *decoding view*, for which cells infer or decode their position by reading concentrations of PI-carrying genes.

Until now we have seen PI as a single number that quantifies the reproducibility in a patterning system. It would be desirable to also have a local measure of reproducibility, which indicates whether PI is uniformly distributed along the anterior-posterior (AP) axis of the embryo. To summarize the decoding performance at each position x we can compute the posterior distribution $P(x^*|g)$ averaged over the possible measurements, i.e. with respect to the distribution $P(g|x)$:

$$P(x^*|x) = \int dg P(x^*|g) P(g|x) \tag{4.4}$$

which is the probability distribution of the decoded position x^* for a cell at the true location x. These distributions can be visualized in the form of decoding maps (Figure 4.2), which display the joint probability distribution $P(x, x^*)$ as heatmaps. Each vertical slice of the map contains the histogram of the distribution $P(x^*|x)$. For the step-function profile, the map shows two broad uniform blocks, indicating strong uncertainty in the decoding process. Instead, the exponential profile reveals a concentrated heatmap near the diagonal for the anterior regions, while for the

posterior part where the average profile is flatter, the heatmap is more blurred.

If a cell measures concentration g^*, the Bayes optimal estimate of its position is the Maximum A Posteriori (MAP) estimate: $x^* = \text{argmax}_x P(x|g^*)$, which reduces to a Maximum Likelihood estimate (ML) under a uniform prior.

By computing the mean squared error $\sigma_x^2(x^*) = \langle (x - x^*)^2 \rangle_{P(x|g^*)}$ for every x we obtain a function $\sigma_x^2(x)$ termed *positional error*. It can be proven [61], that the positional error cannot be smaller than the Cramer-Rao bound:

$$\sigma_x^2(x) \geq \mathcal{I}(x) \tag{4.5}$$

where $\mathcal{I}(x)$, the Fisher information, is given by:

$$\mathcal{I}(x) = \left\langle \frac{\partial^2 \log P(g|x)}{\partial x^2} \right\rangle_{P(g|x)} \tag{4.6}$$

Using this bound and some other weak assumptions (See [164]) it is possible to demonstrate that:

$$\sigma_x^2(x) \geq \mathcal{I}(x) \geq \left(\frac{d\bar{g}}{dx} \right)^{-2} \sigma_g^2(x) \tag{4.7}$$

and:

$$PI = I(X, G) \geq I(X, X^*) \geq \left\langle \log_2 \frac{1}{P_X(x)\sqrt{2\pi e \sigma_x^2(x)}} \right\rangle_x \tag{4.8}$$

Where $I(X, X^*)$ is the mutual information between the true and inferred position. These results link positional information and positional error, showing us that PI sets an upper bound to the average ability of cells to infer their positions and that the *local* inference accuracy depends on the noise amplitude $\sigma_g^2(x)$ and is inversely proportional to the local slope of the mean morphogen profile.

The idea of optimal decoding [162] has been particularly relevant in the context of mutations that affect patterning systems. In the optimal decoding framework a decoding map $P(x^*|x)$ is derived from the morphogen profile features of wild-type (WT) embryos, such as the shape of the average morphogen profile and the noise statistics. This is in some sense the mathematical formalization of the concept of fate maps [145, 172, 173]. The WT decoding maps can then be used to predict in a quantitatively precise manner shifts of positional markers in mutants in which the patterning gene regulatory networks have been perturbed.

The only assumption here is that the biological 'hardware' that must interpret morphogen gradients to commit to fate choices, has evolved to perform operations that very well approximate the optimal decoder. These predictions on mutant embryos can be seen as a test of the optimality assumption, which has been experimentally

verified in the Drosophila AP patterning system [18] and for the mammalian neural tube development along the dorsoventral axis [19].

4.2 Correlational information and patterning utility

In addition to Wolpert's idea of Positional Information (PI), another significant concept in developmental biology was introduced by Alan Turing in his landmark paper, '*The chemical basis of morphogenesis*' [146]. Historically, Turing's Reaction-Diffusion (RD) model and Wolpert's PI concept were seen as mutually exclusive mechanisms, leading to long-standing debates over which model provided the correct explanation for particular patterning events.

Turing's RD model describes how, starting from homogenous initial conditions, interactions between diffusing chemicals can spontaneously break symmetry, leading to the formation of regular and periodic patterns, such as stripes and spots, providing a model for self-organized patterning. In contrast, Wolpert's PI concept addresses how complex patterns can arise from simple asymmetries in a system. He reasoned that if the gradual differences in morphogen concentrations across space could define multiple positional identities reliably, then the morphogen could be used to create more complex patterns downstream.

The other big difference between RD models and the PI approach is that the regular dots and stripes in Turing patterns are thought to directly produce stripes or spots of different cell types. Instead, Wolpert's idea included an interpretation step, where cells read the information encoded in morphogen gradients to determine their position and fate. This allows a continuous morphogen gradient to specify complex spatial patterns, assuming it encodes enough positional information.

When Wolpert's idea was introduced, some classical experiments seemed to favor his view of patterning. Heterotypic (interspecies) grafts, such as the grafting of frog skin tissue into a newt mouth to obtain frog mouthparts [174], exemplify the idea that morphogens provide positional coordinates along an axis, followed by an additional interpretation step. In this example, the host organism's positional context provides spatial cues, which the transplanted tissue interprets to form structures appropriate for its new location.

Furthermore, in the 1980s and 1990s, research showed that the periodic segments of *Drosophila* larvae, an archetypical pattern for which RD might be expected to be involved, are generated by the stepwise subdivision of the embryo by multiple

overlapping gradients instead of a global periodic patterning process [175, 176, 177], reducing the popularity of RD models even more.

Recently, advances in molecular biology led to a reconsideration of Turing's ideas. A notable example is the patterning of digits during limb development [178], which was previously thought to be driven by the PI contained in the gradient of Sonic Hedgehog (Shh) emanating from the limb bud's posterior end. However, recent experiments showed that limb patterning involves a Turing-type RD system based on interactions between Wnt and Bmp proteins [179].

A particularly interesting case is the Nodal-Lefty reaction network involved in germ layer specification [180, 181] and left-right patterning [182] in vertebrates. In this system, Nodal and Lefty form a Turing RD system, where the resulting pattern manifests as a single morphogen gradient, which contains the PI used to establish more complex patterns downstream.

These examples show that self-organized patterning, as described by RD models and lateral inhibition [183], and directed patterning, as summarized by PI concepts, should not be seen as mutually exclusive. Instead, they represent two ends along a spectrum of patterning strategies [165, 184].

While the information-theoretic framework discussed earlier provided a robust understanding of the constraints imposed by stochasticity on morphogen-based patterning and for understanding how PI is encoded and transmitted in systems of directed patterning, there are still unresolved questions, especially regarding the origin and refinement of Positional Information (PI).

At the onset of the embryogenesis for a wide range of organisms (such as the mammalian embryo), the developing organism consists of a single fertilized egg that undergoes successive divisions. During these early stages, all cells are nearly identical, and there is no Positional Information. Understanding how PI emerges and becomes refined through subsequent developmental stages remains a significant challenge. This refinement likely involves processes that progressively increase the precision of spatial patterning. In many cases, initial pattern-forming processes, potentially based on mechanisms of self-organization, are thought to be subject to various sources of noise, including extrinsic and intrinsic fluctuations. To counter such noise, downstream operations such as diffusive couplings or cell-cell interactions, may allow cells to collectively reduce these fluctuations.

Recent work [171] proposed a new information-theoretic framework to quantify the patterning efficiency of self-organized systems. Notably, the paper showed that

the sequence of two archetypical self-organizing developmental steps of cell type proportioning and sorting can create PI in initially homogeneous embryos. Intriguingly, this process involves a phase where embryos exhibit correlational information (CI), a measure of spatial correlation in cell fate decisions, before fully converting it into PI. Based on these results we entertain the idea that the refinement and noise reduction discussed above may occur through the exploitation of spatial correlation in gene expression fluctuations, which are often spatially correlated [170].

In this chapter, I build upon these concepts and explore the idea that part of PI refinement may come from leveraging the correlation in gene expression fluctuations. To do this, I begin by generalizing the correlational information and other quantities, introduced for discrete cell fates in [171], to scenarios involving continuous-valued gene expression levels.

Consider a one-dimensional embryo consisting of N cells, where a single morphogen gradient is established. Assume that the expression levels of this morphogen, denoted by the vector $\vec{g} = (g_1, \ldots, g_N)$, exhibit spatially correlated fluctuations. To model this, we use a multivariate Gaussian distribution with mean vector $(\bar{g}(x_1), \ldots, \bar{g}(x_N))$ and covariance matrix C.

$$P(\vec{g}|\vec{x}) = \frac{1}{Z}\exp\left\{-\frac{1}{2}\chi^2(\vec{g}(\vec{x}))\right\} \propto \exp\left\{-\frac{1}{2}\sum_{i,j=1}^{N}(g_i - \bar{g}(x_i))[C^{-1}]_{ij}(g_j - \bar{g}(x_j))\right\} \tag{4.9}$$

Brückner et al. [171] defined a quantity named "*patterning utility*" U to measure the patterning capabilities of self-organizing systems:

$$U = \frac{1}{N}D_{KL}\left[P(\vec{g}|\vec{x})||Q(\vec{g})\right] \qquad \text{where} \qquad Q(\vec{g}) = \prod_{i=1}^{N} P_G(g_i) \tag{4.10}$$

Here, U measures the deviation of the distribution in Eq. 4.9 from the random ensemble where the gene expression level at each location is independently sampled from $P_G(g) = \sum_x P(g|x)P_X(x)$.

It can be shown (refer to Appendix C) that the total patterning utility can be decomposed in two additive components as $U = PI + CI$, where:

$$CI = \frac{1}{N}D_{KL}[P(\vec{g}|\vec{x})||\prod_{i=1}^{N}P(g_i|x_i)] = -\frac{1}{N}S[P(\vec{g}|\vec{x})] + \frac{1}{N}\sum_{i=1}^{N}S[P(g_i|x_i)] \tag{4.11}$$

Here, CI measures the deviation from an uncorrelated ensemble, a concept referred to as "multi-information" in information theory literature [185, 186]. The total patterning utility thus measures the reduction in entropy compared to the random ensemble

$Q(\vec{z})$. This reduction can be understood as the sum of two non-negative contributions: the local (positional) and the non-local (correlational) information, which jointly quantify the total information content of an ensemble of patterns.

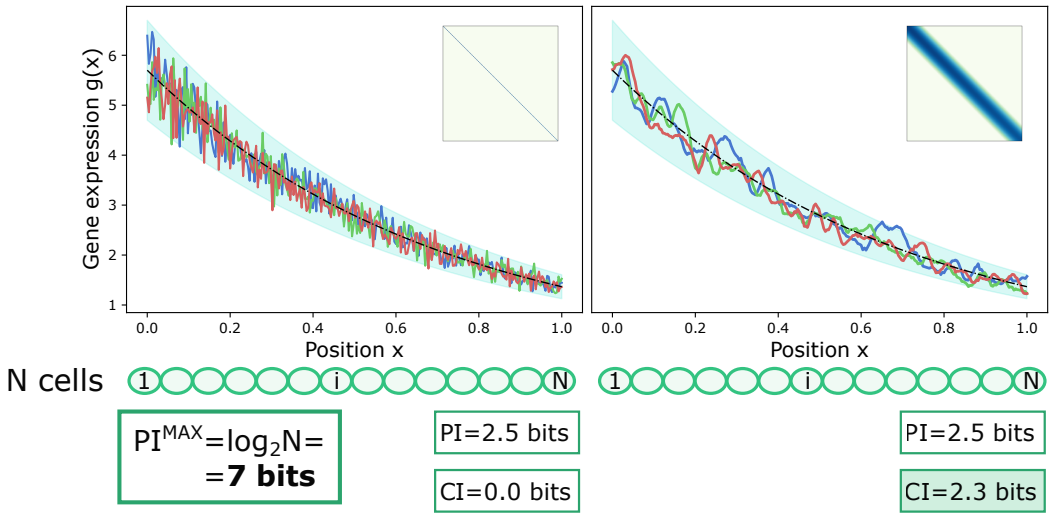


Figure 4.3: Both ensembles exhibit the same positional information (PI = 2.5 bits) but differ in their correlational information (CI). The left ensemble displays uncorrelated fluctuations, while the right ensemble features spatially correlated fluctuations, resulting in 2.3 bits of CI. Insets show the heatmaps of the covariance matrix C in logarithmic scale: the left inset represents a diagonal covariance matrix, indicating no correlation between different positions, while the right inset shows a covariance matrix with off-diagonal elements, highlighting the presence of spatial correlations in gene expression fluctuations.

To understand how different levels of correlation affect PI and CI, we examine ensembles of morphogen gradients while varying the covariance matrix C. From the definition in Eq. 4.2 we can see that fixing the mean profile $\bar{g}(x)$ and the local variances $\sigma_g^2(x)$ keeps PI constant. For instance, consider the two ensembles shown in Figure 4.3, both carrying 2.5 bits of Positional Information. What distinguishes these ensembles is that the one on the left has uncorrelated gene expression fluctuations showing no CI, whereas the one on the right is characterized by spatially correlated fluctuations, carrying 2.3 bits of CI.

To concretize the intuition that correlations could be exploited to enhance developmental precision, I generate ensembles of gene expression profiles with varying noise characteristics and monitor how PI and CI evolve. Specifically, I start with an exponentially decaying mean profile and add noise composed of two components: an uncorrelated noise component with amplitude σ_{ind} and a spatially correlated noise component with amplitude σ_s and correlation length L_{corr}. This simplified scenario helps clarify which aspects of the noise contribute to CI and PI. (Refer to Appendix C for a detailed description of how the fluctuations are sampled.)

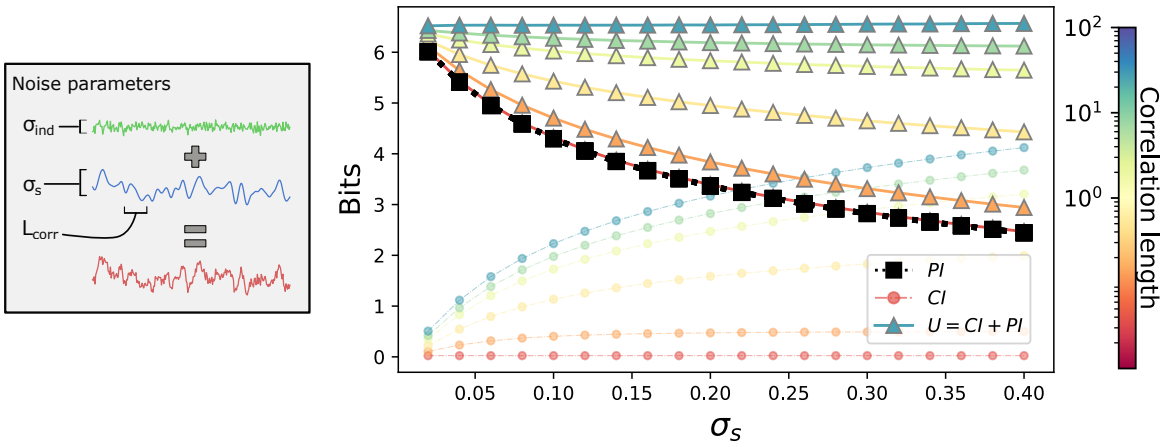


Figure 4.4: Plots showing how positional information, correlational information, and their sum as a function of the noise parameters. The behavior of PI is independent of the correlation length, as displayed by the unique black curve for all values of L_{corr}. Instead, we see that longer-range correlations make the total utility U more robust to the amplitude of fluctuations.

In Figure 4.4 we plot PI, CI and their sum $U = PI + CI$ as a function of σ_s and L_{corr}, keeping $\sigma_{ind} = 0.01$. As expected, PI decreases as the overall noise level increases, and all the lines representing different values of L_{corr} overlap. In contrast, the total information content U behaves differently: as L_{corr} increases, U decreases more slowly with increasing σ_s. For larger correlation length, U remains close to $\log_2 N$ even as the noise levels rise. This suggests that the information needed to specify each cell's identity might indeed be contained within the spatial correlations of morphogen fluctuations.

However, while PI can be clearly interpreted in the context of directed patterning as the amount of information that cells have about their position from a morphogen readout (and therefore about the fate they need to commit to), the practical use of correlational information by cells is less clear. In the following sections, I explore how adopting a non-local decoding strategy, instead of a local one, might be advantageous for cells, and how this advantage can be linked to the correlations present in the system.

4.3 Correlational information and non-local decoding

In Section 4.1 we introduced the concept of optimal decoding, which prescribes the best way to map a measured gene expression level to an inferred position, given the statistical properties of the ensemble of profiles. Following the optimal decoding pre-

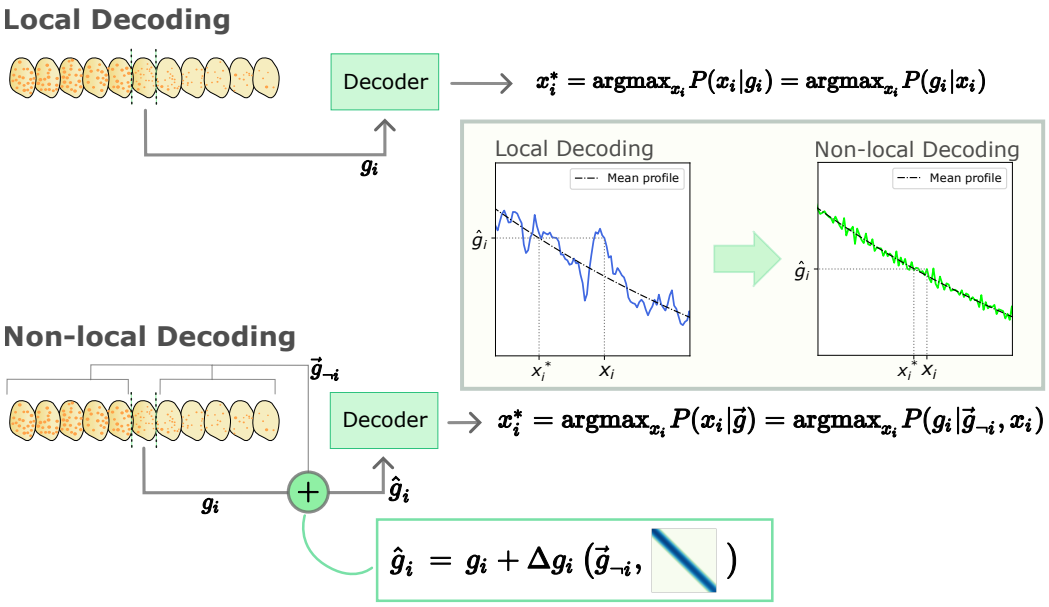


Figure 4.5: Schematic representation of optimal local and non-local decoding. The optimal local decoder takes as input the value of the gene expression level measured by cell i, g_i and makes an estimate of the location of the cell x_i. This operation is represented in the left plot within the grey box: the cell guesses its own position based on the location within the embryo where the mean profile $\bar{g}(\cdot)$ takes the measured value. For the non-local decoding also the measurements from the rest of the embryo $\vec{g}_{\neg i}$ are taken into consideration. The non-local decoder computes an effective expression level $\hat{g}_i$, on which then the same operation described for the local decoder is applied to make the positional inference. $\hat{g}_i$ is equal to g_i plus an additive correction that depends on the non-local measurements $\vec{g}_{\neg i}$ and on the covariance matrix C

scriptions we have found that the local decoding error (positional error) $\sigma_x^2(x)$ depends on the shape of the mean morphogen gradient and the local noise amplitudes $\sigma_g^2(x)$. These findings were derived by computing the inferred position as the Maximum A Posteriori estimate $x_i^* = \arg\max_{x_i} P(x_i|g_i)$, which relies on the gene expression level measured by cell i.

In this section, we derive the optimal decoding strategy under the assumption that each cell i has access not only to its own gene expression level g_i but also to measurements from all other locations $\vec{g}_{\neg i}$, to account for the possibility that the decoding strategy could include communication between different cells. In this scenario, optimal decoding involves maximizing the posterior probability $P(x_i|\vec{g})$, which can be expressed as:

$$P(x_i|\vec{g}) \propto P(g_i|\vec{g}_{\neg i}, x_i)P(x_i) \tag{4.12}$$

When the prior distribution $P(x_i)$ is uniform, finding the MAP estimate is equivalent to maximizing the conditional likelihood $P(g_i|\vec{g}_{\neg i}, x_i)$ with respect to x_i. Using properties of multivariate normal distributions (Eq. 4.9), we know that the conditional probability is normally distributed: $P(g_i|\vec{g}_{\neg i}, x_i) = \mathcal{N}(\tilde{g}(x_i), \tilde{\sigma}^2(x_i))$, where the mean $\tilde{g}(x_i)$ and the variance $\tilde{\sigma}^2(x_i)$ are given by:

$$\tilde{g}(x_i) = \bar{g}(x_i) + C_{i,\neg i}[C_{\neg i,\neg i}]^{-1}(\vec{g}_{\neg i} - \bar{g}(\vec{x}_{\neg i})) \tag{4.13}$$

$$\tilde{\sigma}^2(x_i) = \sigma^2(x_i) - C_{i,\neg i}[C_{\neg i,\neg i}]^{-1}C_{\neg i,i} \tag{4.14}$$

Here, $C_{\neg i,\neg i}$ is the submatrix of the full covariance matrix with the i-th row and column removed, and $C_{i,\neg i}$ is the i-th row of C without the i-th element. If the gene expression level of cell i does not co-vary with the expression levels at other locations *i.e.* $C_{i,\neg i} = \vec{0}$, then $\tilde{g}(x_i) = \bar{g}(x_i)$ and $\tilde{\sigma}^2(x_i) = \sigma^2(x_i)$. This implies that the conditional distribution $P(g_i|\vec{g}_{\neg i}, x_i)$ is equal to the marginal $P(g_i|x_i)$, meaning that the optimal non-local decoder for cell i will be the same as the optimal local decoder.

For the optimal decoder under the Gaussian approximation, the best inferred position $x^*_{i,loc}$ given a measured gene expression level g_i is computed by maximizing the log-likelihood $-(g_i - \bar{g}(x_i)^2)/\sigma_g^2(x_i)$. This can be simplified (assuming we are not at the embryo boundaries) to:

$$\bar{g}(x^*_i) = g_i \quad \longrightarrow \quad x^*_{i,local} = \bar{g}^{-1}(g_i) \tag{4.15}$$

The optimal decoder, therefore, infers the position x^*_i, based on where the mean profile $\bar{g}(\cdot)$ matches the measured value g_i, as illustrated in Figure 4.5.

For the optimal *non*-local decoder, we have that:

$$\tilde{g}(x^*_i) = g_i \quad \longrightarrow \quad x^*_{i,nonloc} = \bar{g}^{-1}\Big(g_i - C_{i,\neg i}[C_{\neg i,\neg i}]^{-1}(\vec{g}_{\neg i} - \bar{g}(\vec{x}_{\neg i}))\Big) \tag{4.16}$$

This effectively means that the optimal non-local decoder performs optimal local decoding with a *corrected* or effective gene expression level

$$\hat{g}_i = g_i - C_{i,\neg i}[C_{\neg i,\neg i}]^{-1}(\vec{g}_{\neg i} - \bar{g}(\vec{x}_{\neg i})) = g_i + \Delta g_i(\vec{g}_{\neg i}, C) \tag{4.17}$$

where the correction term depends on $\vec{g}_{\neg i}$ and the covariance matrix C.

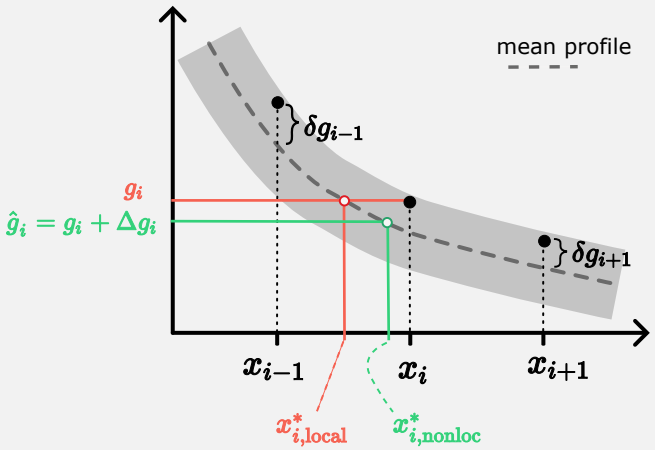
Figure 4.6 shows how the ensemble of the effective profiles $\hat{g}$ tightens around the mean morphogen gradient $\bar{g}(\cdot)$, thereby increasing positional information. Longer correlation lengths result in higher values of CI and smaller positional errors in the positions x^* inferred with non-local decoding.

Semi-nonlocal decoding: example with nearest neighbors correlations

The non-local decoding scheme discussed here assumes that a cell i can access the gene expression levels measured by all other cells, which may not always be a realistic scenario. One way to generalize this to cases where only partial information is available is to choose the x_i^* that maximizes $P(x_i|\vec{g}_{i,\text{avail}})$, where $\vec{g}_{i,\text{avail}}$ are the available measurements to cell i. For instance, if cell i only has access to measurements from its nearest neighbors, *i.e.* $\vec{g}_{i,\text{avail}} = (g_{i-1}, g_{i+1})$, the relevant portion of the covariance matrix C that cell i would use for the semi-nonlocal decoding is:

$$C_i = \begin{bmatrix} \sigma_{i-1}^2 & \rho & \alpha \\ \rho & \sigma_i^2 & \rho \\ \alpha & \rho & \sigma_{i+1}^2 \end{bmatrix} \tag{4.18}$$

This simple example helps build intuition for the case where complete non-local information is available. The figure on the next page visually represents how non-local decoding operates in this scenario.



In this case, the effective gene expression level $\hat{g}_i$ read $\hat{g}_i = g_i + \Delta g_i$ where:

$$\Delta g_i = -\frac{\rho}{\sigma_{i-1}^2\sigma_{i+1}^2 - \alpha^2}\left[\delta g_{i-1}(\sigma_{i+1}^2 - \alpha) + \delta g_{i+1}(\sigma_{i-1}^2 - \alpha)\right] \tag{4.19}$$

where δg_{i-1} and δg_{i+1} are the fluctuations around the mean profile of cells $i-1$ and $i+1$ (see figure). If there is positive covariance $\rho > 0$, these fluctuations contribute negatively (provided $\alpha < \sigma_{i-1}^2, \sigma_{i+1}^2$), and the weights of these contributions are inversely proportional to the noise levels of the measurements from those cells (i.e., the standard deviations at the denominator)

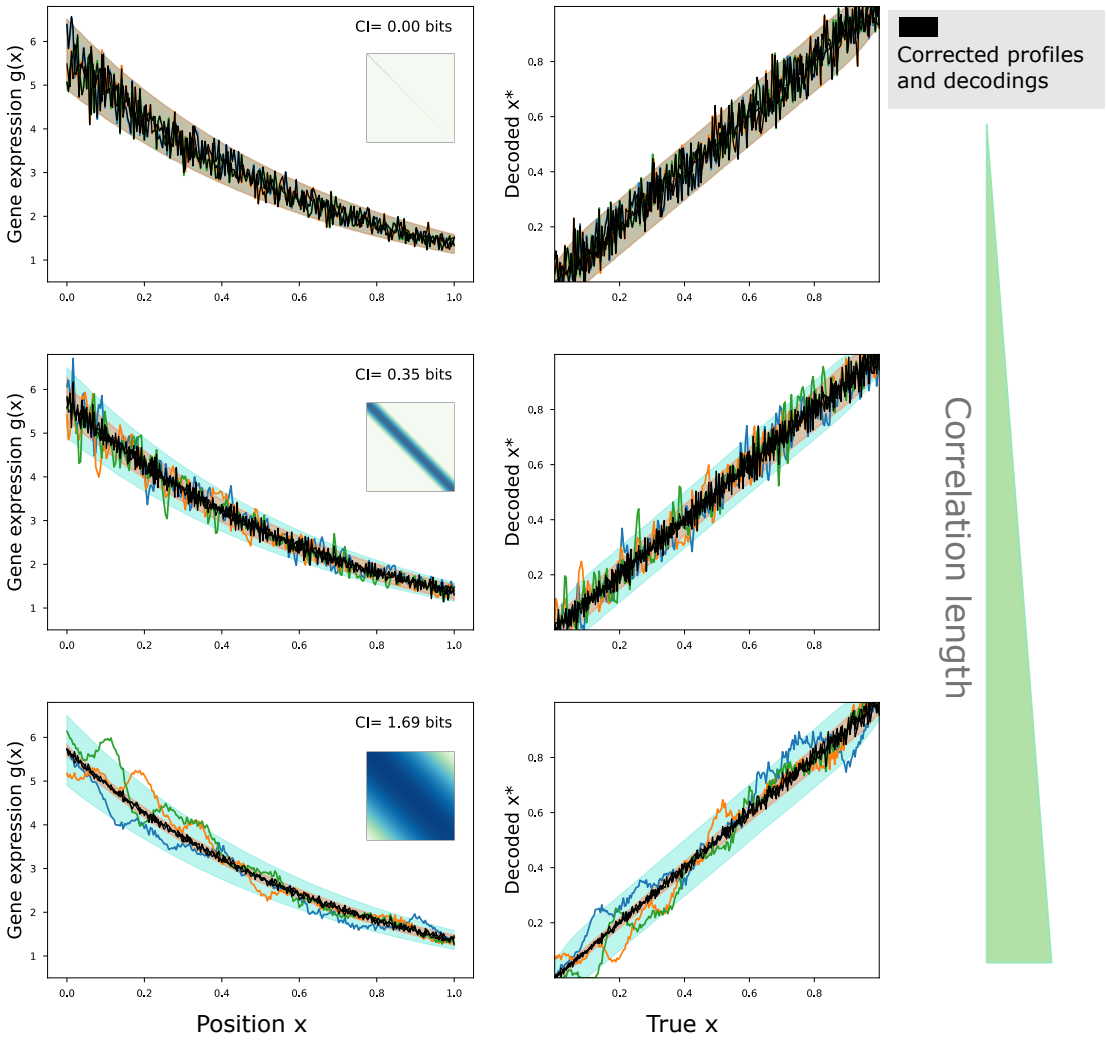


Figure 4.6: Examples of effective profiles $\hat{g}$ derived from the optimal non-local decoding rule. In each row, the ensemble of profiles has a different amount of CI depending on the structure of the covariance matrix C. The left column contains morphogen profiles, while the right one contains the MAP estimates x^* as a function of the true position x. Shaded green regions correspond to two standard deviations of $P(g|x)$ (left) and $P(x^*|x)$ respectively. Shaded orange regions refer to the effective profiles $\hat{g}$ and the associated decoded positions. The top row shows that in the absence of CI, non-local decoding does not have any effect.

4.3.1 Information content of non-local decoder

In Section 4.1 we established that positional information upper bounds the mutual information between true and inferred positions $I(X^*, X)$, which is a function of positional error (Eq. 4.8). When the noise is small enough and the distribution of positions consistent with the observed gene expression level $P(x|g)$, is approximately Gaussian, the bounds in Eq. 4.8 become tight [164]. Under these conditions, a useful approximation is:

$$PI \simeq I_{pos}^{loc} = \left\langle \log_2 \frac{L}{\sqrt{2\pi e \sigma_{x,loc}^2(x)}} \right\rangle \tag{4.20}$$

with L being the length of the embryo. Here, $\sigma_{x,loc}^2(x)$ is obtained by computing the Fisher Information as in Eq.4.6, because in the case of local decoding, x_i^* is obtained by maximizing the likelihood $P(g_i|x_i)$. In contrast, for optimal non-local decoding, we need to maximize the conditional likelihood $P(g_i|\vec{g}_{\neg i}, x_i)$. The variance of this estimate can be bounded by the following Fisher information:

$$I(x_i) = -\left\langle \frac{\partial^2 \log P(g_i|\vec{g}_{\neg i}, x_i)}{\partial x_i^2} \right\rangle_{P(g_i|\vec{g}_{\neg i})} \simeq \frac{(\tilde{g}'(x_i))^2}{\tilde{\sigma}^2(x_i)} \tag{4.21}$$

Therefore, the decoding information $I(x_{nonloc}^*, x) = I_{pos}^{nonloc}$ can be expressed as:

$$\begin{aligned}
I_{pos}^{nonloc} &= \left\langle \log_2 \frac{L}{\sqrt{2\pi e \sigma_{x,nonloc}^2(x)}} \right\rangle_x \simeq \\
&\simeq \log_2 \frac{L}{\sqrt{2\pi e}} + \frac{1}{2N} \sum_{i=1}^{N} \log_2 (\tilde{g}'(x_i))^2 - \frac{1}{2N} \sum_{i=1}^{N} \log_2 \tilde{\sigma}^2(x_i) \\
&= \left\langle \log_2 \frac{L}{\sqrt{2\pi e \sigma_{x,loc}^2(x)}} \right\rangle_x + \frac{1}{2N} \sum_{i=1}^{N} \left[\log_2 C_{ii} - \log_2 \tilde{\sigma}^2(x_i) \right] \\
&= I_{pos}^{loc} + \frac{1}{2N} \sum_{i=1}^{N} \left[\log_2 C_{ii} - \log_2 \tilde{\sigma}^2(x_i) \right]
\end{aligned} \tag{4.22}$$

This indicates that decoding positions from the effective profiles $\hat{g}$ leads to an information gain that we call ΔPI, which is a function of the covariance matrix of the gene expression level fluctuations. Recalling the expression for $\tilde{\sigma}^2(x_i)$:

$$\tilde{\sigma}^2(x_i) = C_{ii} - C_{i,\neg i}[C_{\neg i,\neg i}]^{-1} C_{\neg i,i} \tag{4.23}$$

we see that this is the Schur complement of C_{ii} in the matrix C. According to the properties of the Schur complement, we have $\tilde{\sigma}^2(x_i) = ([C^{-1}]_{ii})^{-1}$. Thus ΔPI can be re-written as:

$$\Delta PI = \frac{1}{2N} \sum_{i=1}^{N} \left[\log_2 C_{ii} + \log_2 [C^{-1}]_{ii} \right] \tag{4.24}$$

Now, take the expression for CI of Eq. 4.11 and write it explicitly for a multivariate normal distribution:

$$CI = -\frac{1}{N} S[P(\vec{g})] + \frac{1}{N}\sum_{i=1}^{N} S[P_i(g_i)] = -\frac{1}{2N}\left(\log_2 \det(C) - \sum_{i=1}^{N} \log_2 C_{ii}\right) \tag{4.25}$$

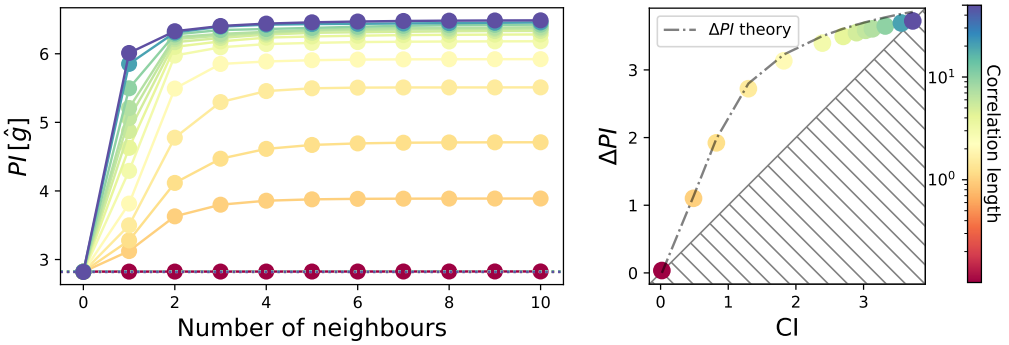


Figure 4.7: Information gains with non-local decoding. Non-local decoding warrants gains in PI that depend on the size of the neighborhood from which a cell receives information, and the correlation length of the gene expression level fluctuations. The plot on the right shows that the analytical expression of Eq. 4.24 well approximates values of ΔPI obtained numerically.

For a positive semi-definite matrix such as the covariance matrix C, Hadamard's inequality tells us that the determinant of the matrix is smaller than or equal to the product of its diagonal elements:

$$-\log_2 \det(C) = \log_2 \det(C^{-1}) \leq \sum_{i=1}^{N} \log_2 [C^{-1}]_{ii} \tag{4.26}$$

This implies that $\Delta PI \geq CI$. Therefore, Eq. 4.22 states that if all available non-local information is used optimally during the decoding process, the positional information gain should be at least equal to CI. This quantitatively supports the initial intuition that refining PI could benefit from leveraging spatially correlated fluctuations in gene expression levels.

The right panel in Figure 4.7 shows values of ΔPI as a function of CI for ensembles of profiles with varying correlation lengths, verifying the validity of our approximations in Eq. 4.22 and confirming the relationship between CI and ΔPI that we found analytically. The left panel shows how the PI of the effective profiles ensemble $\hat{g}$ changes with the size of the neighborhood from which a certain cell receives information. As expected, the information gain is larger for a bigger neighborhood, but the

curves quickly plateau at the information level attainable by interacting with all cells in the embryo. Particularly for highly correlated fluctuations, considering just a few neighboring cells is sufficient to achieve significant information gain.

4.4 Non-local decoding mechanisms

The results discussed thus far provide quantitative insights into how the refinement of positional information may arise by optimally exploiting the statistical regularities provided by spatial correlations in gene expression fluctuations. The amount of information that can potentially be gained is linked to correlational information, and we have proposed that from a computation point of view, this gain can be achieved by performing decoding of positions from non-local morphogen readouts. However, these findings do not directly clarify how such computations could be mechanistically implemented in biological systems. As shown in Eq. 4.17, to compute the effective $\hat{g}_i$ required for optimal non-local decoding, cell i must have access to information about the readout from other cells $\vec{g}_{\neg i}$, as well as the corresponding mean morphogen values $\bar{g}_{\neg i}$. It remains unclear how a cell would gather this information and perform the necessary computations to determine $\hat{g}_i$.

In this section, we explore whether there can be more interpretable operations that well approximate the computation of $\hat{g}$ implied by the optimal non-local decoding rule. This will hopefully help guide the exploration for a mechanistic implementation of the computations needed to exploit spatial correlations.

To begin gaining insights into this direction, we frame the mapping from $\vec{g}$ to $\hat{g}$ as a linear filtering problem. More explicitly, we assume that the vector $\hat{g}$ can be computed as:

$$\hat{g} = F\vec{g} + \epsilon \tag{4.27}$$

where F is a linear filter matrix and ϵ is an error vector. Assuming measured profiles for M embryos, we collect the morphogen profiles in the matrices $G = (\vec{g}_1, \ldots, \vec{g}_M)$ and $\hat{G} = (\hat{g}_1, \ldots, \hat{g}_M)$ with:

$$\hat{G} = FG + \mathcal{E} \tag{4.28}$$

We determine the best-fitting filter by minimizing the Frobenius norm of the error matrix $\mathcal{E}$:

$$F = \arg\min_F \|\hat{G} - FG\|_F^2 \tag{4.29}$$

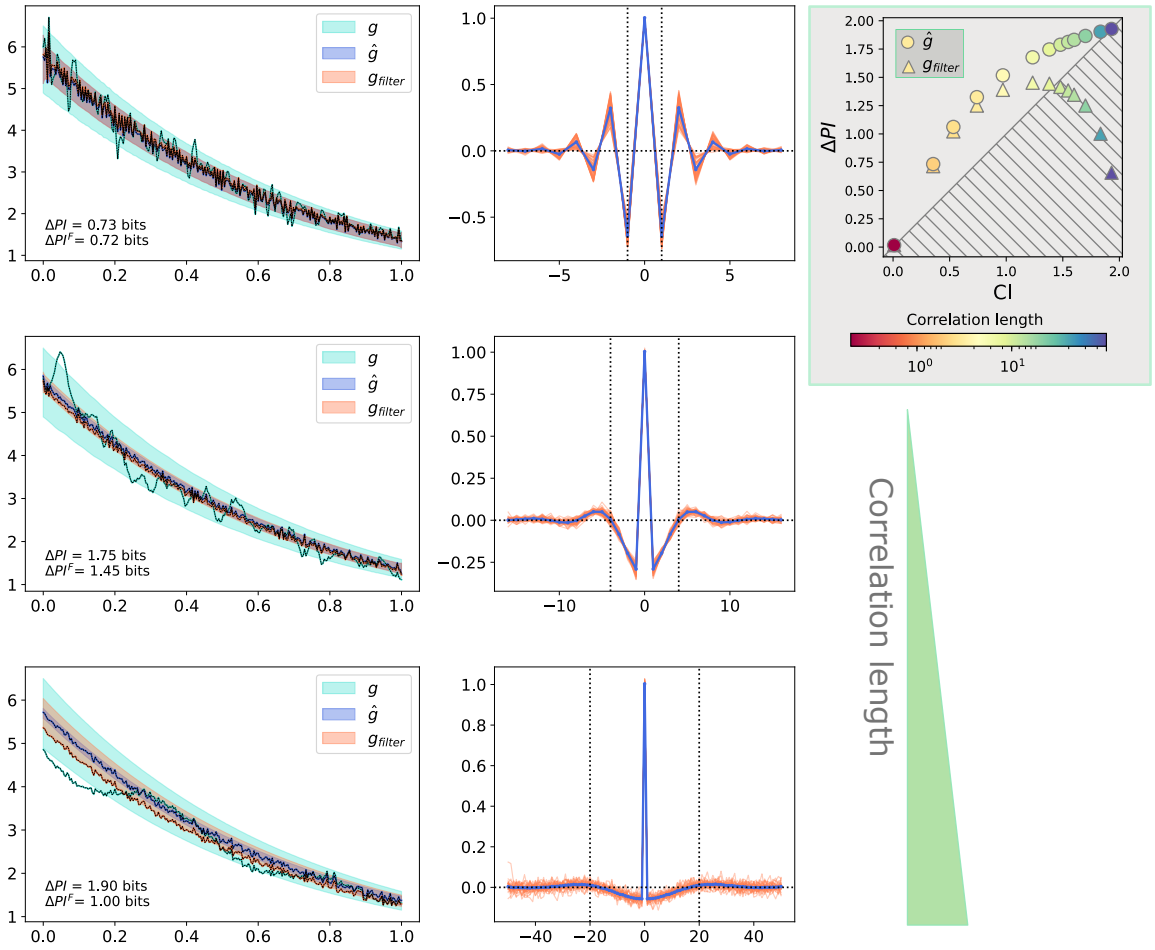


Figure 4.8: Linear filters that approximate the optimal non-local decoding. The left column displays example morphogen profiles $\vec{g}$ and the associated $\hat{g}$ obtained from non-local decoding and its linear filter approximation g_{filter}. The shaded areas of different colors represent $\pm$ two standard deviations computed on the difference ensembles of profiles. The values reported in each plot are the PI gains from optimal decoding and the associated linear filter approximations. The right column shows the best-fitting convolutive filter for various correlation lengths. The inset shows a comparison between the PI gains for non-local decoding and the associated filtering approximation. In this case, for correlations that decay over a distance of up to $\sim 4-5$ neighbors linear filtering represents a good approximation.

The Frobenius norm in the equation can be expressed as:

$$\|\hat{G} - FG\|_F^2 = \mathbf{Tr}\left((\hat{G} - FG)^T(\hat{G} - FG)\right) \tag{4.30}$$

$$= \mathbf{Tr}(\hat{G}^T\hat{G}) - 2\mathbf{Tr}(FG\hat{G}^T) + \mathbf{Tr}(FGG^TF^T) \tag{4.31}$$

From this equation, we compute the matrix derivative of the norm with respect to F and set it to zero:

$$\frac{\partial}{\partial F}\|\hat{G} - FG\|_F^2 = -2G\hat{G}^T + 2FGG^T = 0 \tag{4.32}$$

Therefore, the linear transformation that best approximates the mapping to $\hat{g}$ is:

$$F = \hat{G}G^T(GG^T)^{-1} \tag{4.33}$$

This equation gives us, given an ensemble of morphogen gradients, the recipe to compute the linear filter. We tested this for ensembles with fluctuations of varying correlation lengths, to check what is the regime (if any) of validity of the linear transformation approximation.

In the linear filtering framework, each element $\hat{g}_i$ of the filtered profile is a linear combination of the elements of $\vec{g}$, where the weighting coefficients are contained in the rows $F_{i,:}$ of the matrix F. Analysis of these matrices reveals that their significant values are concentrated near the main diagonal, with elements farther from the diagonal being, on average, close to zero. This pattern suggests that any mechanistic implementation would likely involve local or near-local interactions. To better understand this, we examined the averaged filter response around the diagonal by defining a window $\vec{f_i} = (F_{i,i-w}, \ldots, F_{i,i+w})$ for each row i and computing the average across rows. The resulting average filter $\bar{f}$ displays a clear mean structure with fluctuations added on top, indicating that the filters are largely consistent across different positions, effectively behaving as stationary convolution filters (see Figure 4.8).

We observed that for ensembles where fluctuations decay over short distances (a few neighbors), the filtered profiles $\vec{g}_{filt}$, obtained by applying the fitted filter F to the original profiles $\vec{g}$ closely approximate the optimal non-local decoding transformation, yielding similar gains in positional information. In contrast, for longer correlation lengths, the linear filter does not provide a good approximation (see inset in Figure 4.8), leading to the question of whether some non-linearities may be needed, and what their interpretations would be.

For shorter correlation lengths, when the linear filter or convolution effectively approximates the optimal non-local decoding, these fitted linear filters could serve as starting points for exploring potential biochemical network motifs that perform

such computations. For instance, in the second row of Figure 4.8, we observe a filter characterized by a central positive peak flanked by two negative lobes, extending up to a distance corresponding to the original fluctuations' correlation length. Beyond these, smaller positive lobes are observed. This pattern could suggest a reaction network where $\hat{G}$ is a chemical whose production is activated by G (the original morphogen), represented by the central peak. Simultaneously, G could activate two other chemicals, N_1 and N_2, where N_1 activates and N_2 inhibits the production of $\hat{G}$, both diffusing at different rates. N_1, with a higher diffusion constant, could diffuse from distant locations and activate $\hat{G}$ in cell i, whereas N_2, having a lower diffusion constant, inhibits $\hat{G}$ production nearer to its source. While this remains speculative, the shapes of these filters provide a quantitative foundation for investigating how cells in a developing embryo might achieve non-local decoding through biochemical interactions.

4.5 Discussions and Conclusions

In this chapter, we explored the role of spatial correlations in gene expression fluctuations and their contribution to refining positional information (PI) in developing tissues. The development of an optimal non-local decoding framework has shown that non-local interactions can significantly improve positional accuracy, particularly when fluctuations in gene expression are spatially correlated.

A critical insight we gained is the relationship expressed by the bound $\Delta PI \geq CI$, which quantitatively links the gain in positional information to the correlational information available. This suggests that cells can achieve substantial information gains by leveraging correlations, yet the fact that these gains can exceed the correlational information points to deeper, yet-to-be-understood mechanisms of information processing.

These insights open up several important questions for future research. One promising direction is the systematic exploration of biochemical networks that could implement the optimal non-local decoding strategies proposed. The speculative ideas presented in section 4.4 suggest that specific biochemical interactions could approximate non-local decoding. Developing quantitative models of these networks could help us understand how molecular mechanisms might facilitate the necessary non-local interactions.

Another area for further investigation is the surprising outcome that, in the absence of correlations, the optimal decoding strategy defaults to a purely local approach. This finding contradicts the intuitive expectation that spatial averaging might

offer a better strategy under such conditions. This discrepancy suggests that alternative formulations of the decoding problem could be explored to incorporate spatial averaging, potentially offering more robust decoding strategies when correlations are weak or absent.

The relationship between correlational information and positional information gain, encapsulated in the bound $\Delta PI \geq CI$, also warrants deeper exploration. While this bound provides a clear connection between correlation and information gain, the biological meaning of gaining more information than what is contained in the correlations remains unclear. Future work could focus on uncovering the mechanisms that enable such information gains.

In conclusion, this chapter has laid the groundwork for understanding how spatial correlations in gene expression can enhance developmental precision. However, the findings also highlight the complexity of decoding strategies and leave many open questions on how cells might effectively utilize non-local information.

5

Conclusions

In this concluding chapter, I summarize the key findings of my thesis, offer reflections on my work, and speculate about future directions. This thesis explored the relationship between spatial organization and biological function, initially focusing on chromatin structure and biochemistry, and later expanding to gene expression patterns during development.

At the nuclear and subnuclear scale, chromatin structure is believed to play a significant role in the regulation of transcription and other cellular processes, particularly through CTCF-mediated loops and topologically associated domains (TADs) [10, 9, 8]. These structures are thought to mediate the spatial interactions between promoters and the associated enhancer regions. This is a very attractive view, which however has been recently challenged by studies that show how the disruption of the 3D organization of chromatin does not necessarily affect transcription on large scales and that interactions between cis-regulatory elements can act also across TAD boundaries, which were previously thought to act as insulators for these interactions [12, 13, 14, 15]. These recent results suggested that the strength and direction of the link between biological function and 3D genome architecture still required further investigation.

In the first parts of this thesis, we studied this connection from both directions. We developed novel models, SEMPER and bioSBM, to study how the polymeric nature of chromatin influences biochemical processes and, conversely, how epigenetic markers contribute to chromatin's spatial organization. Both models emphasize interpretability, aiming to address the challenge of integrating noisy experimental data to reveal underlying biological mechanisms. We aimed to build models with a probabilistic formulation that allows for uncertainty quantification and the integration of data from different sources in a transparent way, and we developed approximate

posterior inference algorithms for these models.

In **Chapter 2**, a model was designed to quantify how the polymeric nature of chromatin influences nuclear biochemistry. SEMPER explores the relationship between chromatin structure and biochemical factors by integrating one-dimensional biochemical data into a three-dimensional polymer model. We designed an innovative method to quantify the extent to which the polymeric nature of chromatin contributes to the formation of epigenetic patterns, focusing on the histone modification H3K27me3. Discrepancies observed when performing inference with or without considering the polymer suggest that previous purely data-driven efforts [86] may have overestimated the importance of transcription factors in shaping the chromatin landscape. Our results indicate that many features of the histone modification we studied can largely emerge as a result of polymer mechanisms amplifying relatively subtle biochemical signals. Although the results were encouraging, particularly in demonstrating the significant role of polymeric properties in biochemical organization, SEMPER also highlighted the need for further refinement. Expanding the model to include more complex interactions, such as those involving additional histone marks or protein complexes, could deepen our understanding of the biophysical mechanisms underlying chromatin regulation.

In **Chapter 3**, we tackled the problem from the opposite angle, using bioSBM to predict chromatin structures from epigenetic features. This stochastic block model takes as input biochemical data, such as histone modification profiles, and predicts large-scale chromatin interactions and compartmentalization. bioSBM interprets these data in a low-dimensional, probabilistic framework, providing an interpretable model for understanding how biochemical signals are translated into physical chromatin structures. One of bioSBM's key strengths is its generalizability across different cellular contexts. It was shown to perform well not only on the cell line on which it was trained but also in predicting chromatin structure in unseen chromosomes and cell types. This robustness is significant because it suggests that the relationships between epigenetic modifications and chromatin folding are conserved across different conditions. Nonetheless, like SEMPER, bioSBM also simplifies certain aspects of chromatin physics, particularly by focusing on larger-scale interactions and neglecting finer-scale variations in chromatin structure. Future work should focus on extending the model to capture these finer details, potentially by integrating it with physically realistic information from polymer models or by including new biochemical datasets.

In the last part (**Chapter 4**) we then expanded the scope to examine how the spatial properties of gene expression profiles along some embryonic axis can influence patterning during development, shifting our focus from the study of chromatin to the

spatial organization of DNA at the multi-cellular level. Cells, during development, must interpret positional information (PI) encoded in the spatial patterns of gene expression levels, to correctly differentiate and organize into tissues. This chapter explored how cells acquire and decode this positional information, in the presence of spatially correlated noise, to increase developmental precision. We analyze the concept of correlational information (CI), which quantifies the degree of spatial correlations in gene expression fluctuations, and see how it can be used to enhance the precision of developmental processes. We propose the idea that cells can use information not just from their own gene expression level measurement but also from neighboring cells, a strategy we call non-local decoding. Our results suggested that cells can reduce positional errors by taking advantage of the spatial correlations in gene expression fluctuations using a non-local decoding strategy, thereby improving developmental precision. Importantly, starting from the assumption that *globally*, biological organisms can optimally exploit the correlations present in the system, we derive effective signal transformation operations (linear filters) that *individual cells* need to perform to extract as much information as possible. These derived operations can be a starting point to investigate microscopic biochemical mechanisms that can implement these computations in the real system. Although these results are promising, they remain theoretical, and more research is needed to validate these ideas experimentally.

While the thesis focused on different scales of spatial organization, chromatin structure in the earlier chapters and gene expression patterns in development in the later one, a connection between these scales remains unexplored. In this respect, it is worth noting not only the difference in scale between the two parts of the thesis but also the fundamental difference in approach. The first part operates under the *inference* framework, where one tries to gain some understanding of mechanisms by looking at statistical regularities in the data. The style of the second part could instead be labeled as *normative* or *optimality* theory approach [187, 188, 189]. In this approach, one assumes the existence of a quantity (in our case positional information), which is desirable for a functioning biological system, and that therefore has been optimized through evolution so that a functioning biological hardware should well approximate the *optimal* operations that theory predicts. While these approaches might seem at odds, they could be complementary in bridging the gap between scales. Recent work by Sokolowski et al (2023) [190], for instance, derived the parameters of a mechanistic and microscopic model of the Gap genes regulatory network involved in the early developmental stages of *Drosophila* by optimizing the positional information that these genes should have. Essentially, functional desiderata at the macroscopic scale (i.e. the ability of the embryo to precisely establish a body plan) was used to determine the interaction parameters at the microscopic level. In the same spirit, I

entertain the intriguing idea for future work that optimality results from macroscopic studies [191, 192, 193, 194] could be useful for constraining chromatin models at the microscopic scale.

One more aspect that was not touched upon in this thesis is that of dynamics. Alongside the static picture provided by chromosome conformation capture (structural) and ChIP-seq (biochemical), in recent years newly developed live imaging experiments have provided some insight into the dynamic nature of both loop-extrusion [195] and the relationship between enhancer-promoter interactions and transcriptional regulation [196]. One widespread idea is that one of the ways in which structure influences biological function is that spatial proximity between enhancers and promoters triggers transcription of the associated gene. However, recent results [197] found that there is a strong non-linear relationship between enhancer-promoter contact probability and transcriptional output, suggesting that merely considering static contact properties may not be enough. While in this thesis we only considered the link between structure and biological function, motivated by these recent results I think it will be exciting to also incorporate dynamics. Indeed, tracking of DNA loci with live imaging provides stochastic trajectories from which one can extract effective forces on cis-regulatory elements using force-inference techniques [198, 199]. An interesting avenue for future research could be to combine these inferred forces with models aimed at explaining the static properties of chromatin to build toward a more comprehensive understanding of chromatin physics and biology.

A

Supplementary material of Chapter 2

This appendix file contains a more detailed description of (i) the SEMPER model and the relative approximate Monte Carlo numerical scheme (Sec. A.1), (ii) the experimental data and how to include them in the model (Sec. A.2) and (iii) a validation test of the model (Sec. A.3). The file is accompanied by the supplemental table A.1 listing the source of experimental data employed in this work and by the supplemental figures A.1 to A.7.

A.1 Sequence Enhanced Magnetic PolymER (SEMPER): model and methods

A.1.1 The polymer model

We describe the chromatin filament and the experimentally observed epigenetic pattern in terms of a *magnetic polymer model* [52]. The polymeric nature of the chromatin fiber is modeled as a self-avoiding walk (SAW) on the $3D$ simple cubic lattice [94] of unit lattice spacing $= d$. The SAW is made of N monomers linearly connected by $N-1$ bonds of length equal to d, and each monomer $i = 1, ..., N$ bears a binary Ising-like spin variable $\sigma_i \in \{0, 1\}$. Then, the Hamiltonian $\mathcal{H}$ governing the physical behavior of the model chromatin filament is given by the expression:

$$\begin{aligned} \mathcal{H}(\{r\}, \{\sigma\}|\boldsymbol{\theta}) &= \mathcal{H}_{\mathrm{SAW}}(\{r\}) - \frac{J}{2}\sum_{i\neq j}\sigma_i\sigma_j\Delta(\boldsymbol{r}_i, \boldsymbol{r}_j) - \sum_i h_i(\boldsymbol{w})\sigma_i \\ &\equiv \mathcal{H}_{\mathrm{SAW}}(\{r\}) + \mathcal{H}_{\mathrm{Ising}}(\{r\}, \{\sigma\}|J) + \mathcal{H}_{\mathrm{ext}}(\{\sigma\}|\boldsymbol{w}), \end{aligned} \quad (A.1)$$

where the compact notations $\{r\} \equiv (\mathbf{r}_1, ..., \mathbf{r}_N)$ and $\{\sigma\} \equiv (\sigma_1, ..., \sigma_N)$ stand for the whole sets of monomers' spatial coordinates and the corresponding Ising-like spin variables. The meaning of the various terms in Eq. (A.1) is as the following:

- $\mathcal{H}_{\mathrm{SAW}}(\{r\})$ is the pure polymer contribution. It is $= +\infty$ if the polymer is self-crossing (*i.e.*, non self-avoiding) and zero otherwise [94].
- $J > 0$ is the coupling constant between two spins. The spin-spin interaction is limited to pairs of monomers that are nearest-neighbors on the lattice, *i.e.* $\Delta(\mathbf{r}_i, \mathbf{r}_j)$ takes value $= +1$ if $|\mathbf{r}_i - \mathbf{r}_j| = d$ and $= 0$ otherwise.
- h_i, $i = 1, ..., N$ are the local external fields modeling the effects of the sequence. In detail, their values are expressed in terms of the following "linear combinations" of f *sequence* and *transcription-factor-binding* "features" (suitably selected as described in Sec. A.2):

$$h_i(\boldsymbol{w}) = \sum_{\alpha=1}^{f} F_{i\alpha} w_\alpha \,, \tag{A.2}$$

  where $\boldsymbol{F}$ is the $N \times f$ matrix containing the features' values for all monomers' positions and $\boldsymbol{w}$ is the vector containing the "weights" that assign different importance to the different experimentally observed features. $F_{\alpha i}$ is the value of the α-th feature measured at the i-th genomic bin. The data for constructing the feature matrix come from the [99] database where data is provided as *log fold changes over control* in *.bigWig* file format. The accession codes for the utilized data can be found in the supplemental table A.1. For computational convenience one of the features is a constant ("bias") term for all monomers: by tuning the weight associated to it, it allows us to aptly modulate the overall external fields' magnitude.
- $\boldsymbol{\theta} \equiv (J, \boldsymbol{w}) = (J, w_1, ..., w_f)$ is a short-hand notation for all the free parameters featured in the model Hamiltonian (A.1).

A.1.2 Monte Carlo method

Magnetic polymer conformations $(\{r\}, \{\sigma\})$ at fixed values of the parameters $\boldsymbol{\theta} = (J, \boldsymbol{w})$ are obtained by sampling from the canonical distribution with the Hamiltonian (A.1),

$$P(\{r\}, \{\sigma\}|\boldsymbol{\theta}) \sim e^{-\mathcal{H}(\{r\},\{\sigma\}|\boldsymbol{\theta})/\kappa_B T} \,,$$

where T is the temperature and κ_B is the Boltzmann constant. In this work, the sampling is performed by using a standard Metropolis-Hastings [68] Monte Carlo (MC) algorithm where, similarly to [52, 53], we implement separate moves for the

polymer's and the spins' degrees of freedom. In particular, by following the recipe introduced in [94], for each polymer-like move we do combine together one (global) *pivot* move and $\sim \mathcal{O}(N)$ (local) *crankshaft* moves: if the proposed polymer conformation is not self-avoiding the move is rejected, otherwise it is accepted with the standard Metropolis-Hasting probability. Then, for the spin-like move we perform $\sim \mathcal{O}(N)$ straightforward single spin flips.

Quite unfortunately, in the high-J (*i.e.*, low-T) regime, this simple Monte Carlo algorithm gets unavoidably stuck into low-energy metastable polymer conformations resulting in highly inefficient sampling. In order to fix this issue, we have resorted to the so called *Multiple-Markov-Chain* MC method by Orlandini [98]. The method consists in the idea of running simultaneously a certain number p of simulations at different temperatures $T_1 > T_2 > ... > T_p$. Then, every once in a while, a pair of conformations from the ensembles at temperatures T_i and T_{i+1} are randomly picked and swapped with each other with a certain probability. In more detail, by denoting by $\pi_{\mathcal{S}}(T)$ the Boltzmann weight of the state $\mathcal{S}$ at temperature T, the swap between states "i" and "$i+1$" is taken with probability:

$$\alpha_{\rm swap}(i, i+1) = \min\left(1,\, \frac{\pi_{\mathcal{S}_{i+1}}(T_i)\,\pi_{\mathcal{S}_i}(T_{i+1})}{\pi_{\mathcal{S}_i}(T_i)\,\pi_{\mathcal{S}_{i+1}}(T_{i+1})}\right). \tag{A.3}$$

Now, if we call Γ the space of all possible states of the magnetic polymer, it can be shown that the procedure described above constitutes a Markov chain in Γ^p with joint stationary distribution given by the expression:

$$P(\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_p) \propto \prod_{i=1}^{p} \pi_{\mathcal{S}_i}(T_i)\,. \tag{A.4}$$

The factorization in (A.4) allows us to say that for each individual chain i the sequence of states is asymptotically distributed according to the canonical distribution with the Hamiltonian (A.1) and temperature T_i.

A.1.3 Inference method *via* approximate MC sampling

As mentioned in the main text the posterior distribution of the problem at hand,

$$P(\boldsymbol{\theta}|\{\tilde{\sigma}\}) \propto P(\{\tilde{\sigma}\}|\boldsymbol{\theta})P(\boldsymbol{\theta}) = \frac{e^{-\mathcal{H}_{\rm ext}(\{\tilde{\sigma}\}|\boldsymbol{w})}}{Z(\boldsymbol{\theta})}\Phi(J, \{\tilde{\sigma}\})P(\boldsymbol{\theta})\,, \tag{A.5}$$

is not explicitly tractable. In fact, if one attempts to sample from this probability distribution by resorting again to the Metropolis-Hastings MC algorithm [68], the corresponding acceptance ratio is given by:

$$\alpha_{\rm MC} = \min\left(1,\, \frac{e^{-\mathcal{H}_{\rm ext}(\{\tilde{\sigma}\}|\boldsymbol{w}+\boldsymbol{\delta w})}p(\boldsymbol{\theta}+\boldsymbol{\delta\theta})}{e^{-\mathcal{H}_{\rm ext}(\{\tilde{\sigma}\}|\boldsymbol{w})}p(\boldsymbol{\theta})} \cdot \frac{Z(\boldsymbol{\theta})}{Z(\boldsymbol{\theta}+\boldsymbol{\delta\theta})} \cdot \frac{\Phi(J+\delta J, \{\tilde{\sigma}\})}{\Phi(J, \{\tilde{\sigma}\})}\right), \tag{A.6}$$

where $\boldsymbol{\theta} \equiv (J, \boldsymbol{w})$ represent the current state of the parameters (see Sec. A.1.1), while $\boldsymbol{\theta} + \boldsymbol{\delta\theta} \equiv (J + \delta J, \boldsymbol{w} + \boldsymbol{\delta w})$ is the proposed new one. In order to compute the ratio (A.6), one needs to evaluate the two ratios $\frac{Z(\boldsymbol{\theta}+\boldsymbol{\delta\theta})}{Z(\boldsymbol{\theta})}$ and $\frac{\Phi(J+\delta J,\{\tilde{\sigma}\})}{\Phi(J,\{\tilde{\sigma}\})}$ between intractable normalization prefactors.

In order to deal with this issue, here we propose a method based on the fact that one can approximate these two ratios by writing the following Taylor expansions:

$$\ln \frac{Z(\boldsymbol{\theta}+\boldsymbol{\delta\theta})}{Z(\boldsymbol{\theta})} = \boldsymbol{\delta\theta} \cdot \nabla_{\boldsymbol{\theta}} \ln Z(\boldsymbol{\theta}) + \frac{\boldsymbol{\delta\theta}^T \cdot \boldsymbol{H}_{\ln Z}(\boldsymbol{\theta}) \cdot \boldsymbol{\delta\theta}}{2} + \mathcal{O}(\|\boldsymbol{\delta\theta}\|^3) \tag{A.7}$$

where $\boldsymbol{H}_{\ln Z}(\boldsymbol{\theta})$ is the Hessian matrix of the function $\ln Z$ evaluated in $\boldsymbol{\theta}$ and

$$\ln \frac{\Phi(J+\delta J,\{\tilde{\sigma}\})}{\Phi(J,\{\tilde{\sigma}\})} = \frac{\partial \ln \Phi(J,\{\tilde{\sigma}\})}{\partial J}\delta J + \frac{1}{2}\frac{\partial^2 \ln \Phi(J,\{\tilde{\sigma}\})}{\partial J^2}\delta J^2 + \mathcal{O}(\delta J^3)\,. \tag{A.8}$$

Then, by defining the array

$$\boldsymbol{s} \equiv \begin{bmatrix} \sum_i F_{i1}\sigma_i \\ \sum_i F_{i2}\sigma_i \\ \vdots \\ \vdots \\ \sum_i F_{if}\sigma_i \\ \frac{1}{2}\sum_{i\neq j} \sigma_i\sigma_j \Delta(\boldsymbol{r}_i, \boldsymbol{r}_j) \end{bmatrix} \in \mathbb{R}^{f+1}\,,$$

we have that

$$\begin{aligned} (\nabla_{\boldsymbol{w}} \ln Z(\boldsymbol{\theta}))_j &= \mathbb{E}_{\{r\},\{\sigma\}|\boldsymbol{w},J}\left[\sum_i F_{ij}\sigma_i\right] = \mathbb{E}_{\{r\},\{\sigma\}|\boldsymbol{w},J}\left[s_j\right]\,, \text{ for } j = 1\dots f \\ \frac{\partial \ln Z(\boldsymbol{\theta})}{\partial J} &= \mathbb{E}_{\{r\},\{\sigma\}|\boldsymbol{w},J}\left[\frac{1}{2}\sum_{i\neq j}\sigma_i\sigma_j\Delta(\boldsymbol{r}_i,\boldsymbol{r}_j)\right] = \mathbb{E}_{\{r\},\{\sigma\}|\boldsymbol{w},J}\left[s_{f+1}\right]\,, \end{aligned} \tag{A.9}$$

where $\mathbb{E}_{\{r\},\{\sigma\}|\boldsymbol{w},J}[\cdot]$ stands for the expectation value of the corresponding quantity w.r.t. the distribution function of the polymer and spins configurations at fixed $\boldsymbol{w}$ and J, *i.e.* w.r.t. the joint likelihood (see also Eq. (2) in the main text). Finally, the elements of the Hessian matrix read:

$$\begin{aligned} [\boldsymbol{H}_{\ln Z}(\boldsymbol{\theta})]_{ij} &= \frac{\partial}{\partial\theta_i\partial\theta_j}\ln Z(\boldsymbol{\theta}) = \frac{\partial}{\partial\theta_i}\left[\frac{1}{Z(\boldsymbol{\theta})}\sum_{\{r\},\{\sigma\}} s_j e^{-\mathcal{H}(\{r\},\{\sigma\}|\boldsymbol{\theta})}\right] \\ &= \frac{1}{Z(\boldsymbol{\theta})}\sum_{\{r\},\{\sigma\}}(s_j s_i)e^{-\mathcal{H}(\{r\},\{\sigma\}|\boldsymbol{\theta})} - \frac{1}{Z^2(\boldsymbol{\theta})}\sum_{\{r\},\{\sigma\}} s_j e^{-\mathcal{H}(\{r\},\{\sigma\}|\boldsymbol{\theta})}\sum_{\{r\},\{\sigma\}} s_i e^{-\mathcal{H}(\{r\},\{\sigma\}|\boldsymbol{\theta})} \\ &= \langle s_i s_j\rangle - \langle s_i\rangle\langle s_j\rangle \equiv \mathrm{cov}(\boldsymbol{s})_{ij}\,, \end{aligned} \tag{A.10}$$

where we use the shorthand notation $\langle\cdot\rangle$ for $\mathbb{E}_{\{r\},\{\sigma\}|\boldsymbol{w},J}[\cdot]$.

By using our simulator, one can obtain samples from the likelihood $P(\{r\},\{\sigma\}|\boldsymbol{\theta})$ and then use these samples to obtain unbiased estimates of the quantities needed to compute the above Taylor expansions. Indeed, the first ratio reads:

$$\frac{Z(\boldsymbol{\theta}+\boldsymbol{\delta\theta})}{Z(\boldsymbol{\theta})} = \exp\left(\boldsymbol{\delta\theta}\cdot\langle\boldsymbol{s}\rangle + \frac{\boldsymbol{\delta\theta}^T\cdot\mathrm{cov}(\boldsymbol{s})\cdot\boldsymbol{\delta\theta}}{2}\right)(1+\mathcal{O}(\|\boldsymbol{\delta\theta}\|^3))\,. \tag{A.11}$$

To evaluate the second Taylor expansion we notice that:

$$\frac{\partial \ln\Phi(J,\{\tilde\sigma\})}{\partial J} = \frac{1}{\Phi(J,\{\tilde\sigma\})}\sum_{\{r\}\in\mathrm{SAW}}\frac{1}{2}\sum_{i\neq j}\tilde\sigma_i\tilde\sigma_j\Delta(\boldsymbol{r}_i,\boldsymbol{r}_j)e^{\frac{J}{2}\sum_{i\neq j}\tilde\sigma_i\tilde\sigma_j\Delta(\boldsymbol{r}_i,\boldsymbol{r}_j)}\,, \tag{A.12}$$

corresponding to the expectation value of $\frac{1}{2}\sum_{i\neq j}\sigma_i\sigma_j\Delta(\boldsymbol{r}_i,\boldsymbol{r}_j)$ with respect to the distribution

$$P(\{r\}|J,\{\tilde\sigma\}) = \frac{e^{\frac{J}{2}\sum_{i\neq j}\tilde\sigma_i\tilde\sigma_j\Delta(\boldsymbol{r}_i,\boldsymbol{r}_j)}}{\Phi(J,\{\tilde\sigma\})}\,. \tag{A.13}$$

We can sample from this distribution using our simulator by fixing the spin configuration to $\{\tilde\sigma\}$ and by making only Monte Carlo moves on the polymer degrees of freedom. So in the end we have

$$\frac{\partial \ln\Phi(J,\{\tilde\sigma\})}{\partial J} = \mathbb{E}_{\{r\}|J,\{\tilde\sigma\}}\left[\frac{1}{2}\sum_{i\neq j}\tilde\sigma_i\tilde\sigma_j\Delta(\boldsymbol{r}_i,\boldsymbol{r}_j)\right]\,, \tag{A.14}$$

while, for the second derivative, we get

$$\begin{aligned}\frac{\partial^2 \ln\Phi(J,\{\tilde\sigma\})}{\partial J^2} &= \frac{1}{\Phi(J,\{\tilde\sigma\})}\sum_{\{r\}\in\mathrm{SAW}}\left(\frac{1}{2}\sum_{i\neq j}\tilde\sigma_i\tilde\sigma_j\Delta(\boldsymbol{r}_i,\boldsymbol{r}_j)\right)^2 e^{\frac{J}{2}\sum_{i\neq j}\tilde\sigma_i\tilde\sigma_j\Delta(\boldsymbol{r}_i,\boldsymbol{r}_j)} - \\ &\quad - \left(\frac{1}{\Phi(J,\{\tilde\sigma\})}\sum_{\{r\}\in\mathrm{SAW}}\frac{1}{2}\sum_{i\neq j}\tilde\sigma_i\tilde\sigma_j\Delta(\boldsymbol{r}_i,\boldsymbol{r}_j)e^{\frac{J}{2}\sum_{i\neq j}\tilde\sigma_i\tilde\sigma_j\Delta(\boldsymbol{r}_i,\boldsymbol{r}_j)}\right)^2 \\ &\equiv \mathbb{V}_{r|J,\{\tilde\sigma\}}\left[\frac{1}{2}\sum_{i\neq j}\tilde\sigma_i\tilde\sigma_j\Delta(\boldsymbol{r}_i,\boldsymbol{r}_j)\right]\,.\end{aligned} \tag{A.15}$$

The major problem with this approximated procedures arises when one has many data points. In fact, let us suppose $\{\{\tilde\sigma\}_m\}_{m=1}^M$ are M *i.i.d* observations, then the posterior reads:

$$P(\boldsymbol{\theta}|\{\{\tilde\sigma\}_m\}_{m=1}^M) \propto P(\boldsymbol{\theta})\prod_{m=1}^M P(\{\tilde\sigma\}_m|\boldsymbol{\theta}) = \frac{P(\boldsymbol{\theta})}{Z(\boldsymbol{\theta})^M}\prod_{m=1}^M e^{-\mathcal{H}_{\mathrm{ext}}(\{\tilde\sigma\}_m|\boldsymbol{w})}\,\Phi(J,\{\tilde\sigma\}_m)\,. \tag{A.16}$$

For the estimation of the log-ratio of the partition functions we just need to multiply M times the quantity obtained for the case of a single data point, whereas for $\Phi(J,\{\tilde\sigma\}_m)$ we need to run separate simulations to get different samples (one for each

observation). Although, in principle, this might increase considerably the cost of computing an accurate posterior sampling, for the kind of data we need the number of observations typically ranges from 1 to 3, therefore the increase in the computational burden is expected to be still affordable.

A.2 Inclusion of the experimental data in the model

In this work, we have focussed on a selection of different regions of human chromosome 2 from the H1 embryonic stem (H1-hESC) cell line. We examine the performance of our model at different levels of resolution: therefore, since we simulate polymers with fixed number of monomers $N = 1000$ (see Section 2.1 in the main text), the adopted resolutions (see Table 1 in the main text) correspond to chromatin filaments of contour lengths in the range from $5 \cdot 10^5$ to 10^7 basepairs. Moreover, due to the coarse-grained nature of the polymer model, the continuous ChIP-seq tracks relative to the two biological replicates for the histone modifications need to be converted to discrete-valued vectors of 0's and 1's. To this purpose, each monomer has been assigned a value $= 0$ or $= 1$ depending on whether the signal in the corresponding bin was found below or above the mean signal computed on the entire region (see Fig. A.1 for a schematic illustration of the procedure). Finally, the feature matrix $\boldsymbol{F}$ (Eq. (A.2)) is derived from experimental transcription-factor-binding data and CpG-density data: the latter, in particular, have been obtained from the DNA sequence from genome assembly GRCh38. All data went through a denoising pre-processing phase using a standard lowpass order one Butterworth filter with critical frequency $= 0.1$.

A.3 Validation test from synthetic data

In order to validate the SEMPER method, we test it first on synthetic data. Spins configurations are generated using the magnetic polymer simulator with *arbitrarily chosen* sets of parameter values as input. Then, these spin configurations are used as data points on which to perform posterior inference (see an example of the marginal posterior distributions in Fig. A.2).

Again, since the *true* posterior distributions for the problem at hand are fundamentally intractable, it is non-trivial to assess the goodness of our approximate method. On the other hand – as pointed out in the main text – in the case where the spin-spin coupling J is fixed to zero (*i.e.*, the polymer's and spins' degrees of freedom are not coupled to each other) the model reduces to the logistic regression

(LR) model proposed in [132, 86] for which, instead, exact posterior sampling is both tractable and fast. For this reason, and in order to assess the quality of our approximate method, in this Section we have compared its results to the ones obtained by the so called *Synthetic Likelihood* (SL) method by Wood [101] which has recently emerged as one of the reference approximate numerical schemes in the larger field of so called *likelihood-free* inference methods [96].

In detail, we have performed additional numerical experiments in which additional synthetic data are generated by setting $J = 0$. Then, posterior samples are computed with the two techniques (*i.e.*, our approximate method *and* the SL method), and the two results are compared with the ones obtained with the exact sampling of the LR model. The results shown in Fig. A.3 (top and central panels) and Fig. A.4 demonstrate the excellent agreement between the distributions generated with the exact and our approximate method while, for the same computational cost, the SL method is vastly outperformed (Fig. A.3 (bottom panel) and Fig. A.5).

Supplemental Table and Figures

Measured quantity	Experiment accession code	Data file code
H3K27me3 replicate 1	ENCSR928HYM	ENCFF417VQQ
H3K27me3 replicate 2	ENCSR928HYM	ENCFF780FNS
CTCF	ENCSR000AMF	ENCFF332TNJ
H3K4me3	ENCSR019SQX	ENCFF493QWY
ATF2	ENCSR000BQU	ENCFF031CBV
SIX5	ENCSR000BIQ	ENCFF724HLZ
SP1	ENCSR000BIR	ENCFF684HIL
TCF12	ENCSR000BIT	ENCFF717UNQ
TEAD4	ENCSR000BRY	ENCFF307KTY

Table A.1: *ENCODE* project [99] accession numbers for the *H3K27me3* histone modification data and the transcription factor (TF) binding data employed in this work.

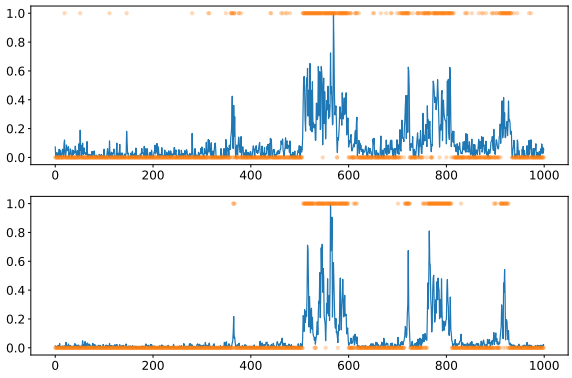


Figure A.1: The continuous ChIP-seq tracks relative to the histone modification *H3K27me3* are discretized in terms of binary spin variables = 0/1 representing, respectively, the absence/presence of the histone modification. The orange dots represent the spin value associated to a specific bin. The plots show data corresponding to the region *chr2:218,600,000-219,100,000*, discretized in 1000 bins of 500 bps each.

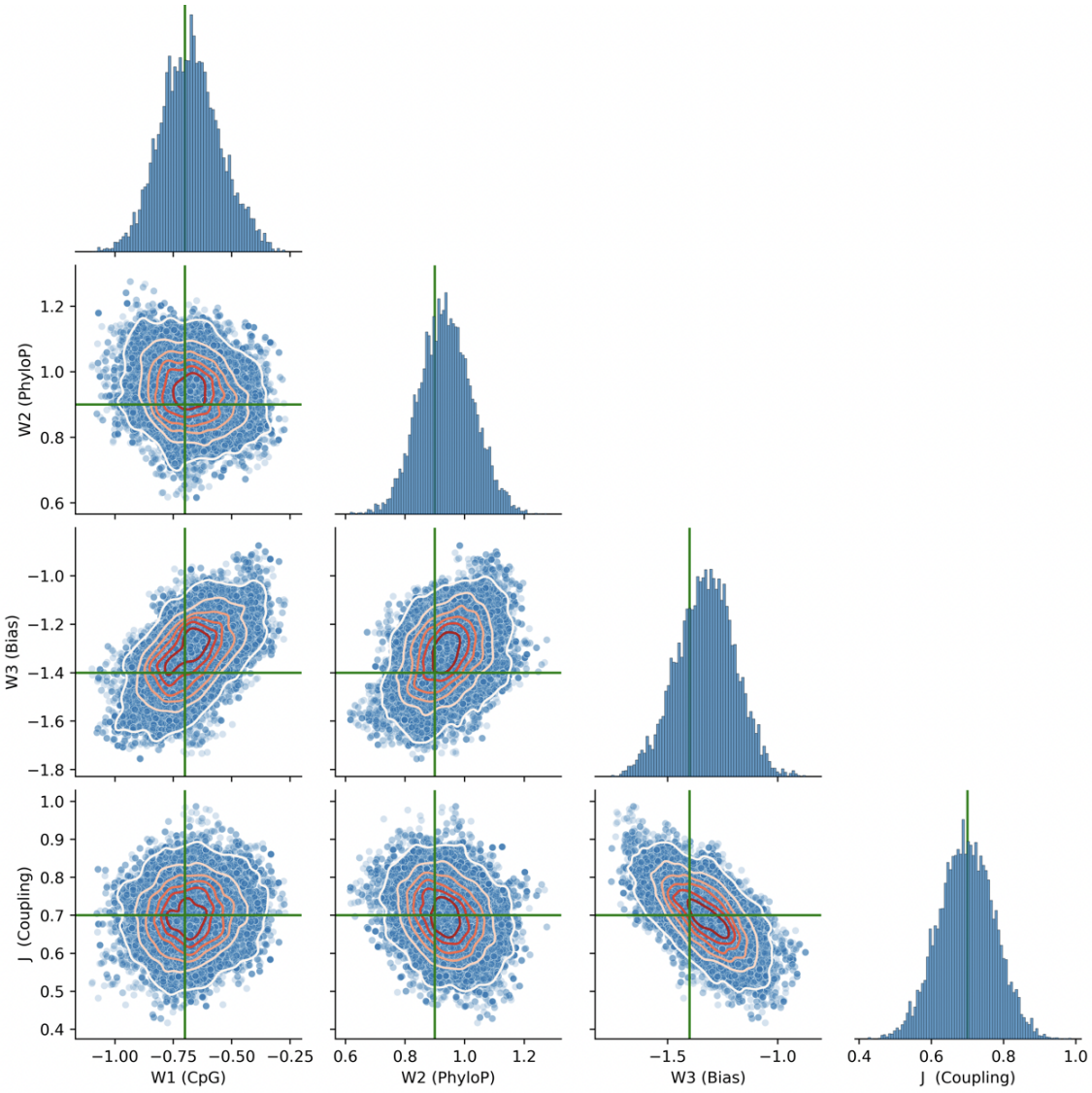


Figure A.2: Validating SEMPER through synthetic input data. The panels display the corresponding posterior distributions obtained with our approximate Monte Carlo procedure. Here, 15 spins configurations generated with the simulator were used as input. The green lines identify the *ground truth* parameter values used to generate the data.

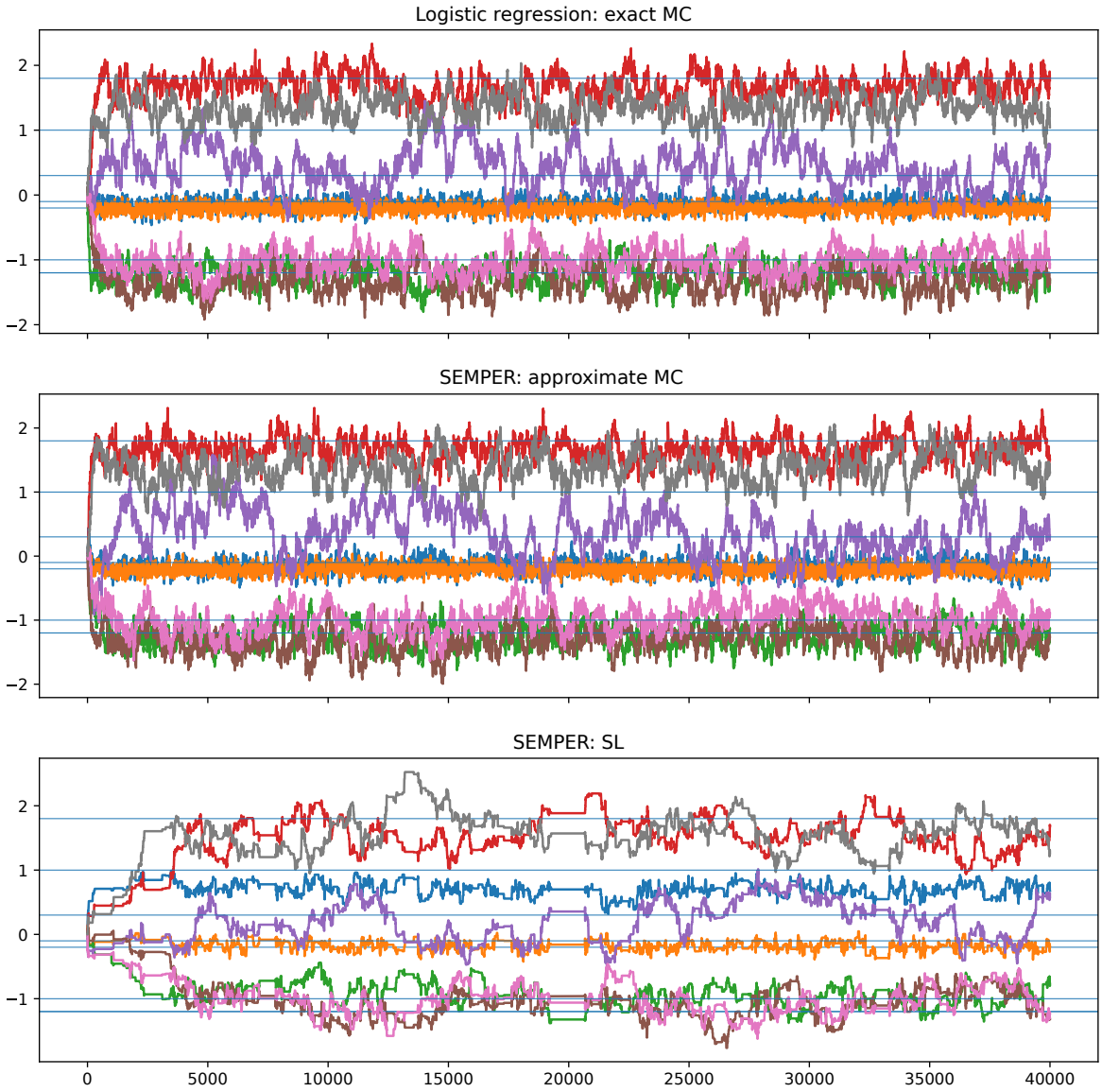


Figure A.3: Test of the model by using synthetic data generated at fixed coupling $J = 0$ between polymer and spin variables (i): MC trajectories for posterior sampling. For this particular case study, we have resorted to the same set of features employed in the real case (*i.e.*, $J \neq 0$) study. On the x- and y-axis we have reported, respectively, the MC run time and the values of the *ground-truth* parameters used to generate the data. (Top) Tracks for the *exact* sampling obtained through the logistic regression (LR) procedure. (Center) Tracks for the approximate Monte Carlo method developed in this work. Notice the remarkable agreement with the tracks obtained by exact LR. (Bottom) Tracks obtained by the likelihood-free method known as *Synthetic Likelihood* (SL) [101]. Notice the poor reproducibility of the tracks, due to the bad mixing of the Markov chains (Sec. A.1.2).

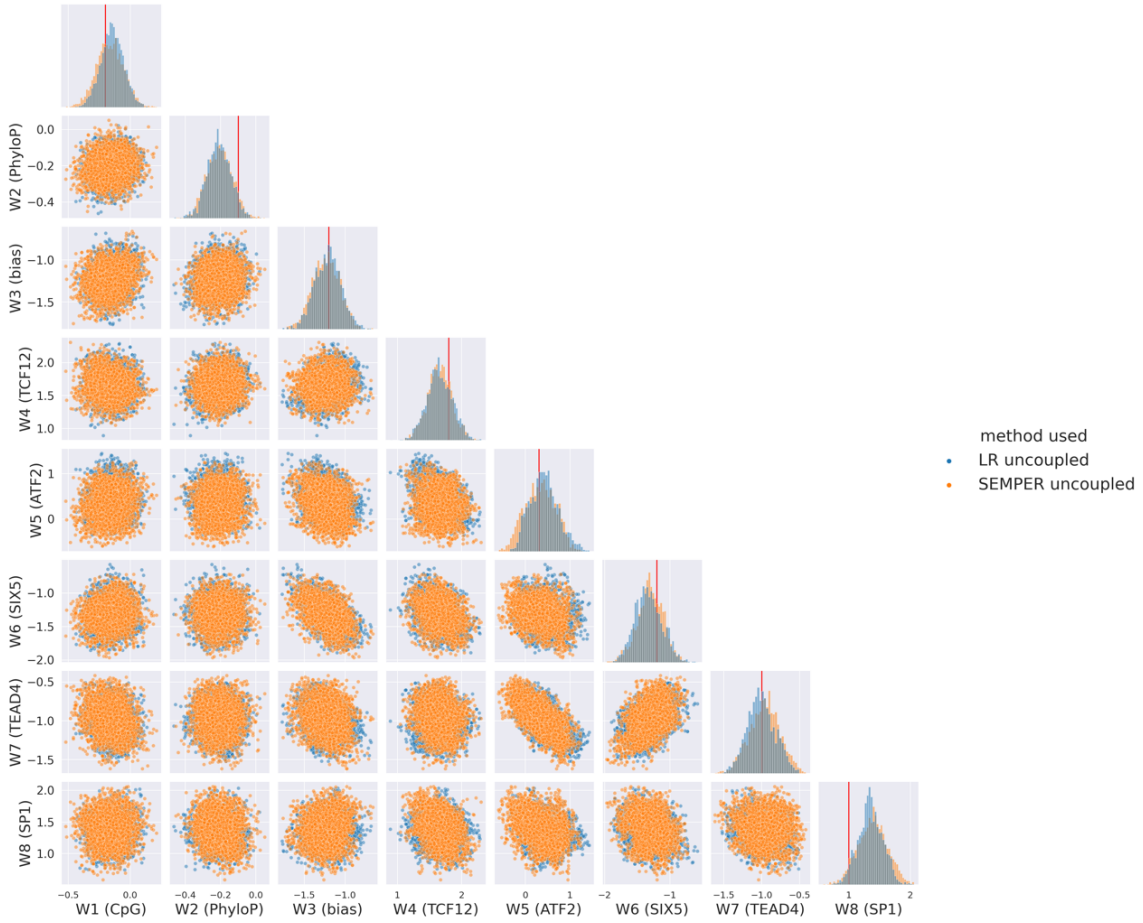


Figure A.4: Test of the model by using synthetic data generated at fixed coupling $J = 0$ between polymer and spin variables (ii): Comparison between the logistic regression (LR) and the SEMPER methods. Histograms and scatter plots are for the MC samples shown in the top and middle panels of Fig. A.3. One can appreciate the closeness of the posterior marginals obtained with our approximate inference method to the posteriors obtained with the asymptotically *exact* MC sampling.

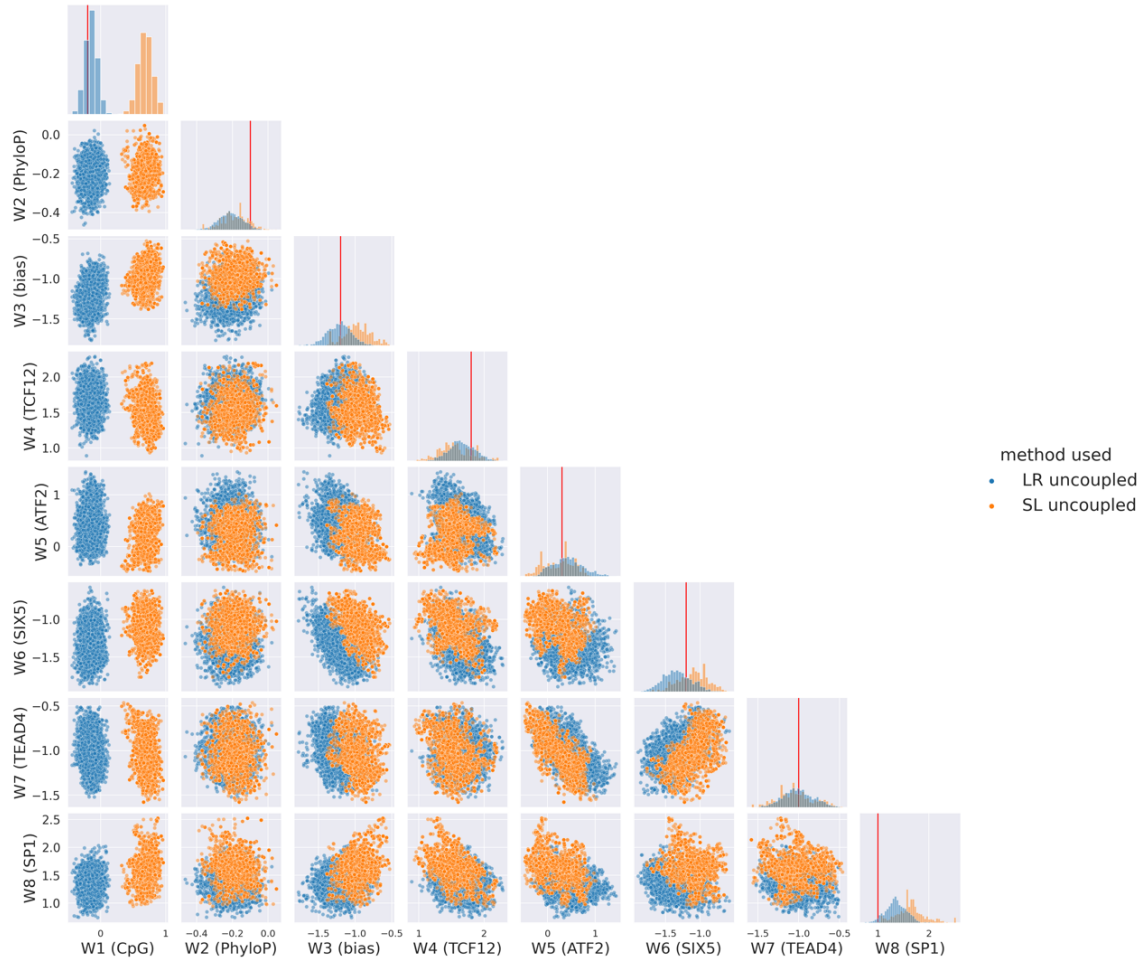
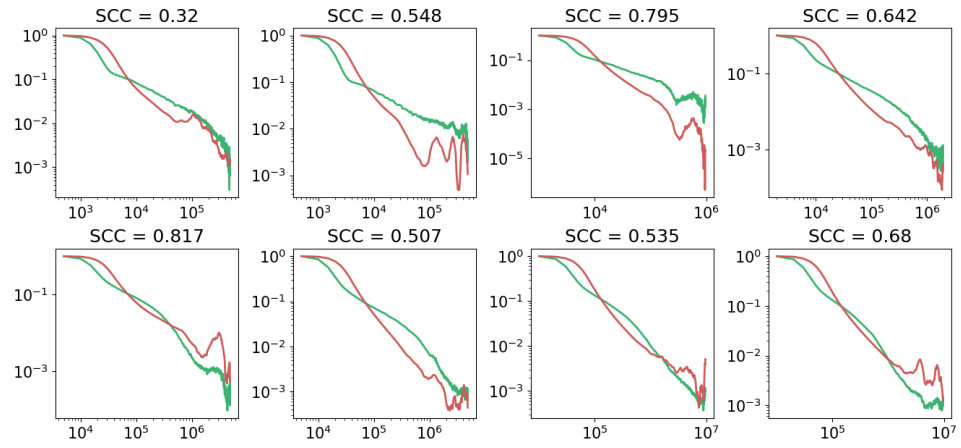


Figure A.5: Test of the model by using synthetic data generated at fixed coupling $J = 0$ between polymer and spin variables (iii): Comparison between the logistic regression (LR) and the synthetic likelihood (SL) methods. Histograms and scatter plots are for the MC samples shown in the middle and bottom panels of Fig. A.3. While in Fig. A.3 we have just appreciated the slow mixing of SL compared to our approximate method, here we notice that SL underperforms wrongly also in locating the mode of the posterior distributions (*i.e.*, the marginal of the weight w_1 associated to the CpG density covariate).

(A)



(B)

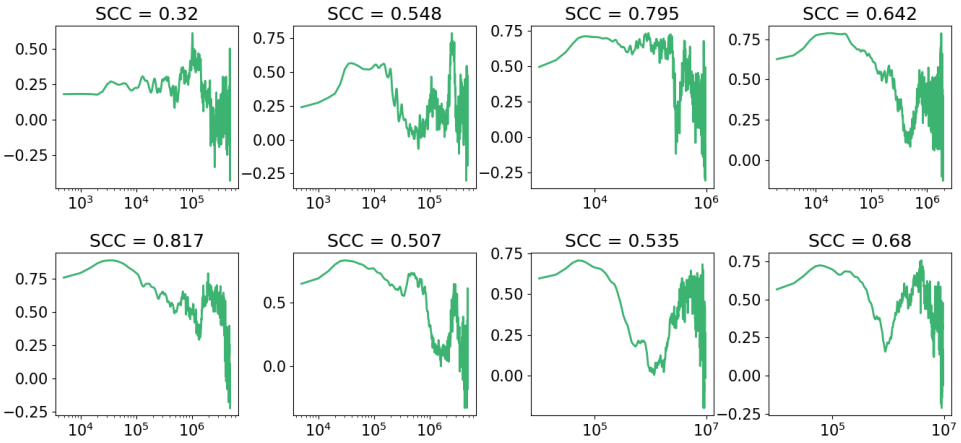


Figure A.6: Comparative analysis of Hi-C contacts (experiments and simulation). Genomic regions and corresponding *Stratum Correlation Coefficients* (SCC [105]) are as in Fig. 4 in the main text. (A) Mean contact probability between pairs of chromatin loci as a function of their genomic separation (in basepair units) along the chromatin fiber. Green and red lines are for, respectively, the experimental Hi-C data and our posterior averages. (B) Pearson correlation coefficients between experimental Hi-C and posterior-predicted contact probabilities as a function of the genomic separation.

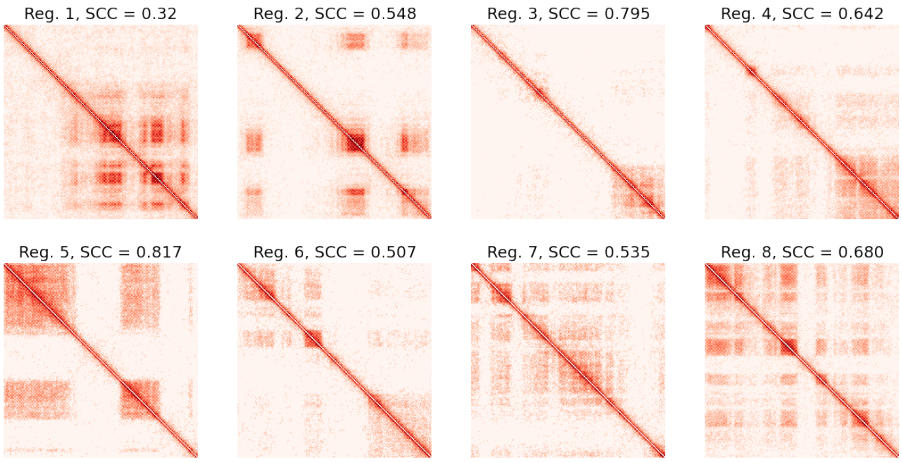


Figure A.7: Standard deviations of contact probabilities computed from the posterior predictive distributions. Genomic regions and corresponding *Stratum Correlation Coefficients* (SCC [105]) are as in Fig. 4 in the main text.

B

Supplementary material of Chapter 3

B.1 A VEM algorithm for posterior nference

In this section, I provide a more detailed derivation of the equations at the core of the Variational Expectation Maximization (VEM) algorithm for the model introduced and studied in Chapter 3.

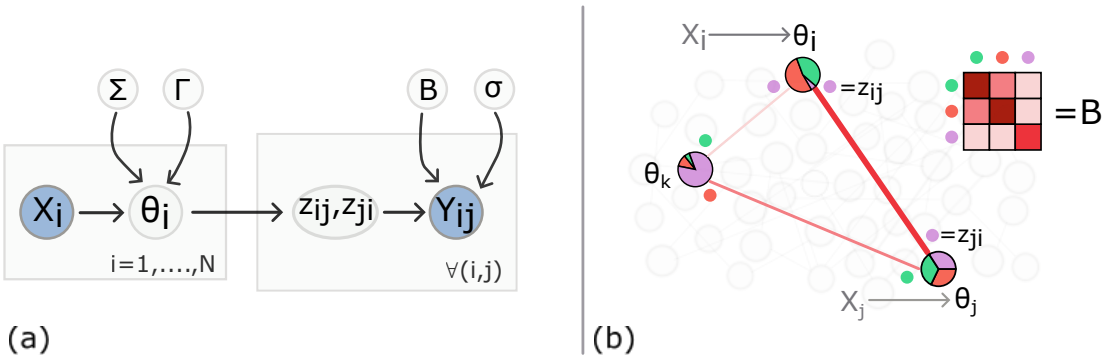


Figure B.1: Generative process of bioSBM.

The quantities in our problem are:

- Observed data $Y \in \mathbb{R}^{N \times N}$: matrix representing Hi-C maps with a log-observed-over-expected normalization.
- Known node covariates $X \in \mathbb{R}^{P \times N}$: each row x_i is the P-dimensional feature vector associated to node i.
- Latent variables: $\theta_{i=1\ldots N}$ and z_{ij} and z_{ji} for $i = 1 \ldots N, \;\; j = 1 \ldots i-1$, where θ_i

is the latent membership vector for node i and z_{ji} is the membership assignment of node i in the interaction with node j.

- Hyperparameters: $(\Sigma, \Gamma, B, \sigma) \equiv \Psi$.

A lower bound for the evidence $P(Y|\Sigma, \Gamma, B, X)$ is given by the ELBO (literally *Evidence Lower BOund*):

$$\begin{aligned}\mathcal{L}(q, \Psi) =& \mathbb{E}_q \left[\log P(Y, \eta_{1:N}, Z|\Psi, X) - \log q(\eta_{1:N}, Z)\right] = \\ =& \mathbb{E}_q \left[\log P(\eta_{1:N}|\Sigma, \Gamma, X) + \log P(Z|\eta_{1:N}) + \log P(Y|Z, B) - \log q(\eta_{1:N}, Z)\right] \\ =& \sum_{i=1}^{N} \mathbb{E}_q \left[\log P(\eta_i|\mu_i = \Gamma x_i, \Sigma])\right] + \sum_{i=1}^{N}\sum_{j=1}^{i-1} \mathbb{E}_q \left[\log P(z_{ij}|\theta_i) P(z_{ji}|\theta_j)\right] \qquad \text{(B.1)} \\ &+ \sum_{i=1}^{N}\sum_{j=1}^{i-1} \mathbb{E}_q \left[\log P(Y_{ij}|z_{ij}, z_{ji}, B)\right] + H(q) \leq P(Y|\Psi, X)\end{aligned}$$

The inequality becomes an equality if the distribution $q(\eta_{1:N}, Z)$ is exactly the posterior distribution. The Variational Bayes *Expectation-Maximization* (VBEM) method consists of iteratively tightening this bound to approximate the true posterior. In the E-step of the algorithm, one needs to maximize the ELBO with respect to the variational distribution $q(\eta_{1:N}, Z)$ while in the M-step one carries out the maximization with respect to the hyperparameters.

B.1.1 Variational E-step

To carry out the optimization task we take a mean-field assumption on the shape of the variational distribution $q(\eta, Z)$, where we assume the factorized form $q(\eta, Z) = q(\eta)q(Z) = \prod_i q(\eta_i) \prod_{i \neq j} q(z_{ij})$. By taking the functional derivatives of the ELBO with respect to the variational distribution and by fixing them to zero one finds the update equations for the E-step, which in the general form read:

$$q(\eta_i) \propto \exp\left\{\log P(\eta_i|\mu_i, \Sigma) + \mathbb{E}_{q(Z)}[\log P(Z|\eta_i)]\right\} \qquad \text{(B.2)}$$

$$q(z_{ij}) \propto \exp\left\{\mathbb{E}_{q(z_{ji})}[\log P(Y_{ij}|z_{ij}, z_{ji}, B)] + \mathbb{E}_{q(\eta_i)}[\log P(z_{ij}|\eta_i)]\right\} \qquad \text{(B.3)}$$

Because of the non-conjugacy of the logistic normal prior with the multinomial likelihood, we will resort to the approximation technique introduced by [200]. More explicitly:

$$q(\eta_i) \propto \exp\{\log P(\eta_i|\mu_i, \Sigma) + \sum_{j(\neq i)} \mathbb{E}_{q(z_{ij})}[\log P(z_{ij}|\eta_i)]\} := \exp\{f(\eta_i)\} \qquad \text{(B.4)}$$

Let $\hat{\eta}_i$ be the maximum a-posteriori value of the parameter. The method introduced by [200] consists in approximating the function $f(\eta_i)$ with its expansion up to second order around $\hat{\eta}_i$. This gives the following Gaussian approximation for the variational approximation $q(\eta_i)$:

$$q(\eta_i) \simeq \mathcal{N}(\eta_i | \, \hat{\eta}_i, -\nabla^2 f(\hat{\eta}_i)^{-1}) \tag{B.5}$$

Therefore we need the function at the exponent to be doubly differentiable and to find the maximum a posteriori (variational) estimate $\hat{\eta}_i$ we can use standard gradient ascent techniques. If we write out explicitly the function we have that

$$f(\eta_i) = -\frac{1}{2}(\eta_i - \mu_i)^T(\Sigma)^{-1}(\eta_i - \mu_i) + \sum_{j(\neq i)}^{N}\sum_{k=1}^{K}\left(\eta_{i,k} - \log\sum_{k'=1}^{K}\exp(\eta_{i,k'})\right)\mathbb{E}_{q(z_{ij})}[z_{ij,k}] \tag{B.6}$$

Defining $\theta(\eta)$ as the *softmax* of η i.e. $\theta(\eta)_k = \exp(\eta_k)/\sum_{k'}\exp(\eta_{k'})$ we can obtain the gradient and the Hessian matrix of $f(\eta_i)$.

$$\nabla f(\eta_i) = \sum_{j(\neq i)}^{N}\left(\mathbb{E}_{q(z_{ij})}[z_{ij}] - \theta(\eta_i)\sum_{k=1}^{K}\mathbb{E}_{q(z_{ij})}[z_{ij,k}]\right) - \Sigma^{-1}(\eta_i - \mu_i) \tag{B.7}$$

$$\nabla^2 f(\eta_i) = \sum_{j(\neq i)}^{N}\left((\mathrm{diag}(\theta(\eta_i)) - \theta(\eta_i)\theta(\eta_i)^T)\sum_{k=1}^{K}\mathbb{E}_{q(z_{ij})}[z_{ij,k}]\right) - \Sigma^{-1} \tag{B.8}$$

For the community assignments distributions, we have $q(z_{ij}) = \mathrm{Mult}(\phi_{ij})$ with ϕ_{ij} as variational parameters, then we can explicit the expectations in the above expressions as:

$$\nabla f(\eta_i) = -(N-1)\theta(\eta_i) + \sum_{j(\neq i)}^{N}\phi_{ij} - \Sigma^{-1}(\eta_i - \mu_i) \tag{B.9}$$

$$\nabla^2 f(\eta_i) = -(N-1)(\mathrm{diag}(\theta(\eta_i)) - \theta(\eta_i)\theta(\eta_i)^T) - \Sigma^{-1} \tag{B.10}$$

Where $\mathrm{diag}(\theta)$ is the diagonal matrix where the non-zero elements are the components of θ. It can be shown that $-\nabla^2 f(\eta_i)$ is always a positive definite ma-

trix, therefore the method always gives a valid covariance matrix for the variational approximation. We use the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm with second-order backtracking line-search [201]. The optimization procedure returns $(\hat{\eta}_i, -\nabla^2 f(\hat{\eta}_i)^{-1})$, thereby updating the variational distribution on η, $q(\eta_i) \simeq \mathcal{N}(\eta_i|\, \hat{\eta}_i, -\nabla^2 f(\hat{\eta}_i)^{-1})$. This is the E-step for the membership distributions per node. This step, as for most variational mean field methods for mixed membership block models, has a cost which is $\sim \mathcal{O}(N^2)$.

For the community assignments posteriors we explicitly write Eq. (B.3) and we get that for every (i, j) and community k

$$\begin{aligned}\phi_{ij,k} &\propto \exp\left\{\sum_{g=1}^{K} \phi_{ji,g} \log P(Y_{ij}|B_{kg}, \sigma_{kg}) + \mathbb{E}_{q(\eta_i)}[\log P(z_{ij}|\eta_i)]\right\} \\ &= \exp\{\mathbb{E}_{q(\eta_i)}[\log P(z_{ij}|\eta_i)]\} \prod_{g=1}^{K} P(Y_{ij}|B_{kg}, \sigma_{kg})^{\phi_{ji,g}}\end{aligned} \tag{B.11}$$

The term in the exponent reads:

$$\begin{aligned}\mathbb{E}_{q(\eta_i)}[\log P(z_{ij}|\eta_i)] &= \mathbb{E}_{q(\eta_i)}[\eta_{i,k} - \log \sum_{g=1}^{K} \exp(\eta_{i,g})] \\ &= \lambda_{i,k} - \int d\eta_i\, \mathcal{N}(\lambda_i, \nu_i) \log \sum_{g=1}^{K} \exp(\eta_{i,g})\end{aligned} \tag{B.12}$$

By imposing normalization $\sum_k \phi_{ij,k} = 1$ for all pairs (i, j) the updates for the community assignments posteriors become

$$\phi_{ij,k} \propto \exp(\lambda_{i,k}) \prod_{g=1}^{K} P(Y_{ij}|B_{kg}\sigma_{kg})^{\phi_{ji,g}} \tag{B.13}$$

Now, if we substitute in the Gaussian edge likelihood, we get the explicit update equations:

$$\log \phi_{ij,k} \;=\; \lambda_{i,k} - \sum_{g=1}^{K} \left[\frac{(Y_{ij} - B_{kg})^2}{2\sigma_{kg}} + \frac{\log \sigma_{kg}}{2}\right] \phi_{ji,g} + const. \tag{B.14}$$

Everything described until now represents the E-step of the VBEM algorithm and the updates are performed with fixed $(\Sigma, \Gamma, B, \sigma)$. The values for these parameters are updated in the M-step where instead we fix the variational parameters $\{\lambda_i, \nu_i\}$ for $i = 1 \dots N$ and ϕ_{ij} for $i \neq j$ that are optimized in the E-step.

B.1.2 Variational M-step

For the global covariance matrix, we need to take the matrix derivative of the terms in the ELBO that contain Σ. By exploiting some properties of matrix derivatives we find that:

$$\begin{aligned}\frac{\partial \mathcal{L}(q,\Psi)}{\partial \Sigma^{-1}} =& \frac{1}{2}\frac{\partial}{\partial \Sigma^{-1}}\sum_{i=1}^{N}\mathbb{E}_{q(\eta_i)}[(\eta_i-\mu_i)^T(\Sigma)^{-1}(\eta_i-\mu_i)+\log|\Sigma^{-1}|] = \\ =& \frac{1}{2}\frac{\partial}{\partial \Sigma^{-1}}\sum_{i=1}^{N}\log|\Sigma^{-1}| - (\lambda_i-\mu_i)^T\Sigma^{-1}(\lambda_i-\mu_i)+\mathrm{Tr}(\Sigma^{-1}\nu_i) = \\ =& \frac{N}{2}\Sigma^T - \frac{1}{2}\sum_{i=1}^{N}\nu_i^T+(\lambda_i-\mu_i)(\lambda_i-\mu_i)^T=0\end{aligned} \tag{B.15}$$

Then the covariance matrix that maximizes the ELBO is given by

$$\hat{\Sigma} = \frac{1}{N}\sum_{i=1}^{N}\nu_i^T+(\lambda_i-\Gamma(x_i))(\lambda_i-\Gamma(x_i))^T \tag{B.16}$$

Something additional that we can do is to add a regularization to the norm of Σ for numerical stability. In particular, something simple we can do is to regularize using the *nuclear norm*. For a generic matrix, the nuclear norm is given by the sum of the singular values; for a positive definite matrix such as Σ it is easy to show that the norm corresponds to the trace of the matrix $\mathrm{Tr}(\Sigma)$. We then introduce an additional term $-R\cdot\mathrm{Tr}(\Sigma)$ in the ELBO where $R=rN$ (selected with cross-validation) is the regularization strength and we write the explicit N dependence in such a way that all the terms of the ELBO containing Σ are *extensive*. The results analyzed in the paper are obtained using $r=0.1$ With the additional term, Eq. B.16 reads:

$$\begin{aligned}&\frac{N}{2}\Sigma - \frac{1}{2}\sum_{i=1}^{N}\nu_i^T+(\lambda_i-\mu_i)(\lambda_i-\mu_i)^T - R\frac{\partial \mathrm{Tr}(\Sigma)}{\partial\Sigma^{-1}} = \\ =&\frac{N}{2}\Sigma - \frac{1}{2}\sum_{i=1}^{N}\nu_i^T+(\lambda_i-\mu_i)(\lambda_i-\mu_i)^T + R\Sigma^2 = 0\end{aligned} \tag{B.17}$$

This equation is solved by the following matrix:

$$\hat{\Sigma} = \frac{1}{4r}\left[-\mathbb{I}+\sqrt{\mathbb{I}+\frac{8r}{N}\sum_{i=1}^{N}\nu_i^T+(\lambda_i-\Gamma(x_i))(\lambda_i-\Gamma(x_i))^T}\right] \tag{B.18}$$

Which is the M-step update concerning the covariance matrix Σ.

The maximization of the ELBO with respect to B_{kg} and σ_{kg} entails:

$$\frac{\partial \mathcal{L}(q,\Psi)}{\partial B_{kg}} = \sum_{i\neq j}^{N} \phi_{ij,k}\phi_{ji,g}\frac{Y_{ij}-B_{kg}}{\sigma_{kg}} \tag{B.19}$$

$$\frac{\partial \mathcal{L}(q,\Psi)}{\partial \sigma_{kg}} = \sum_{i\neq j}^{N} \phi_{ij,k}\phi_{ji,g}\left[\frac{(Y_{ij}-B_{kg})^2}{2\sigma_{kg}^2} - \frac{1}{2\sigma_{kg}}\right] \tag{B.20}$$

By setting the above derivatives to zero we obtain:

$$\hat{B}_{kg} = \frac{\sum_{i\neq j}^{N}\phi_{ij,k}\phi_{ji,g}Y_{ij}}{\sum_{i\neq j}^{N}\phi_{ij,k}\phi_{ji,g}} \tag{B.21}$$

$$\hat{\sigma}_{kg} = \frac{\sum_{i\neq j}^{N}\phi_{ij,k}\phi_{ji,g}(Y_{ij}-B_{kg})^2}{\sum_{i\neq j}^{N}\phi_{ij,k}\phi_{ji,g}} \tag{B.22}$$

These are the last update equations needed to specify the full VBEM algorithm for bioSBM.

Following these calculations, the explicit expression for the ELBO that we compute to monitor the progression of the optimization reads:

$$\begin{aligned}\mathcal{L} =& \frac{N}{2}\log|\Sigma^{-1}| - \frac{1}{2}\sum_{i=1}^{N}[(\lambda_i-\mu_i)^T\Sigma^{-1}(\lambda_i-\mu_i) + \mathrm{Tr}(\Sigma^{-1}\nu_i)] + \sum_{i\neq j}^{N}\sum_{k=1}^{K}\phi_{ij,k}\lambda_{i,k} \\ & - (N-1)\sum_{i=1}^{N}\left(\log\sum_{k'=1}^{K}\exp(\lambda_{i,k'}) + \frac{1}{2}\mathrm{Tr}\{(\mathrm{diag}(\theta(\lambda_i)) - \theta(\lambda_i)\theta(\lambda_i)^T)\nu_i\}\right) \\ & + \sum_{i=1}^{N}\sum_{j=1}^{i-1}\sum_{k,g=1}^{K}\phi_{ij,k}\phi_{ji,g}\log P(Y_{ij}|B_{k,g},\sigma_{k,g}) + \frac{1}{2}\sum_{i=1}^{N}\log|\nu_i| - \sum_{i\neq j}^{N}\sum_{k=1}^{K}\phi_{ij,k}\log\phi_{ij,k}\end{aligned} \tag{B.23}$$

B.2 Daset preprocessing

Chromatin contact files were downloaded from the 4DN consortium data portal [100], and log-observed-over-expected Hi-C maps were extracted from these files using the *hic-straw* [202] Python toolkit. To remove anomalously low reads from the Hi-C maps, we remove all rows and columns where the average value of the normalized Hi-C matrix has a modified Z-score smaller than -3.0.

ChIP-seq tracks for the input covariates were downloaded from the ENCODE Project database [99], and from these epigenetic data was extracted using the *pyBigWiG* [203] Python toolkit. These tracks were then denoised using a standard lowpass

order one Butterworth filter with critical frequency = 0.5 and normalized with a standard scaler.

B.3 Simulation of predicted maps

Starting from the global model parameter inferred from training data, we can generate predictions for an unseen chromosome with covariate matrix $X^{test} = (x_1^{test}, \ldots, x_N^{test})$. In particular, we compute the average membership proportions for each node as:

$$\theta_i^{test} = \Gamma x_i^{test} \tag{B.24}$$

Then for the prediction of edge weight Y_{ij}^{test}, we average over all possible pairs of community assignments of nodes i and j, weighting with the membership proportions:

$$Y_{ij}^{test} = \sum_{k,g}^{K} \theta_{ik} B_{kg} \theta_{jg} \tag{B.25}$$

The prediction accuracy is then computed as the Pearson correlation between the model-generated maps and the experimental logarithmic observed over expected Hi-C maps. For modeling the depletion of RAD21 we simply set the row of X^{test} corresponding to RAD21 to zero.

B.4 Supplemental Table and Figures

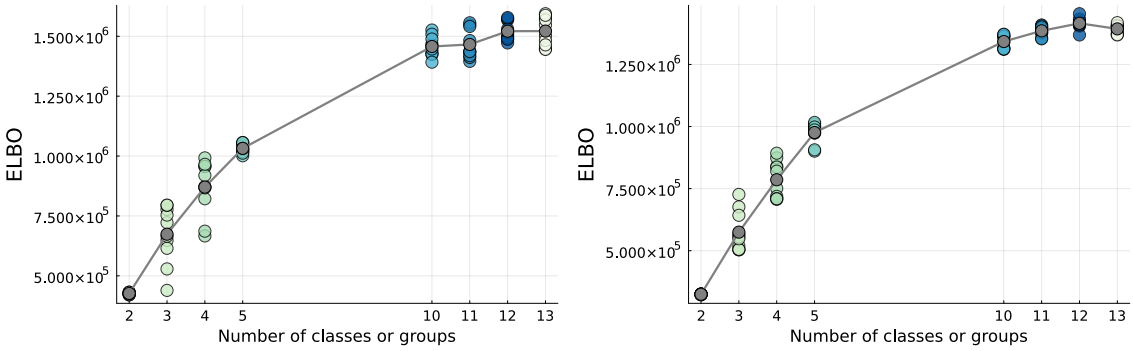


Figure B.2: The Evidence Lower BOund (ELBO) was used as a criterion for Bayesian model selection. The left plot shows ELBO as a function of the number of communities of the model for the training set comprising odd-numbered chromosomes from 11 to 21. The right plot refers to even-numbered chromosomes from 12 to 22. We identify the optimal number of communities as the value where the ELBO peaks or plateaus, leading to the choice $K = 12$

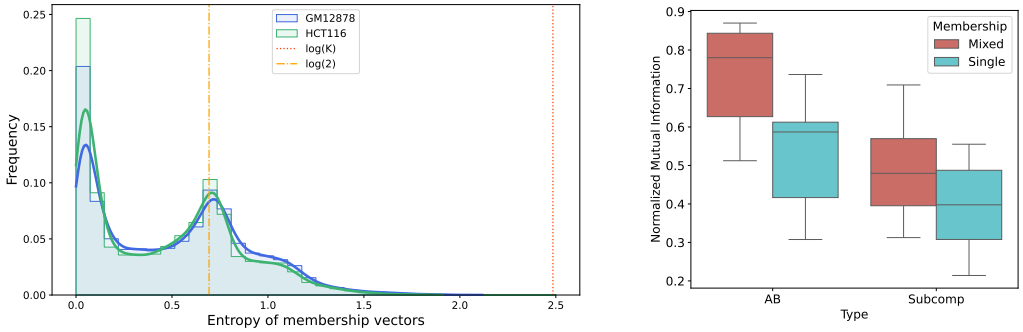


Figure B.3: **Left panel**: shows Shannon entropies of the MAP estimates of the membership vectors for all the genomic regions under study. The two cell lines that have been studied display a similar distribution of said entropy. In both cases the peak at zero shows that a big fraction of nodes displays a single membership behavior, however, there is still a considerable degree of *mixed-membershipness* if one looks at the entirety of the genomic regions under study. **Right panel**: Normalized Mutual Information (NMI) score between the clustering obtained from the model and the biological annotations. Allowing mixed membership leads to a better recapitulation of the biological annotations

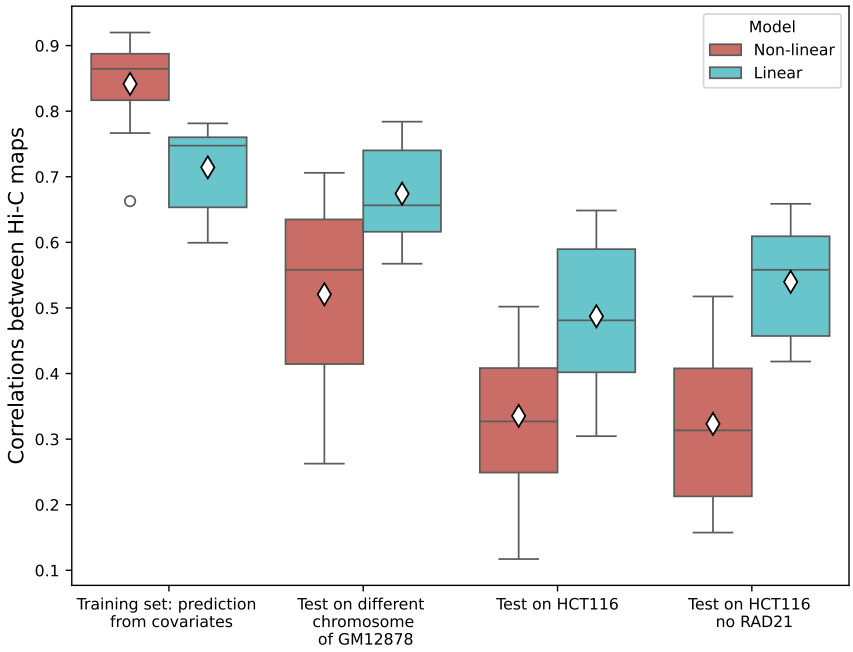


Figure B.4: Correlations between model-generated maps and experimental log O/E Hi-C maps for two models with a different map from biochemical covariates to membership vectors. The turquoise plots correspond to the case where Γ is the simple 14×12 matrix representing a linear mapping, discussed in the main text. The red boxes correspond to the case where $\Gamma()$ is a feed-forward deep neural network with 4 hidden layers of size $(64, 64, 64, 32)$ with the standard *ReLU* (rectified linear unit) non-linear activations. While on the training set, the non-linear model better reproduces experimental contact maps, already for test chromosomes on the same cell line the correlations decrease drastically.

Cell Line	Assay	Experiment accession code	Data file code
GM12878	*in situ* Hi-C	4DNES3JX38V5	4DNFI1UEG1HD
	H3K4me3 (ChIP-seq)	ENCSR057BWO	ENCFF287HAO
	H3K27ac (ChIP-seq)	ENCSR000AKC	ENCFF469WVA
	H3K27me3 (ChIP-seq)	ENCSR000AKD	ENCFF919DOR
	H3K4me1 (ChIP-seq)	ENCSR000AKF	ENCFF564KBE
	H3K36me3 (ChIP-seq)	ENCSR000AKE	ENCFF312MUY
	H3K9me3 (ChIP-seq)	ENCSR000AOX	ENCFF683HCZ
	H3K9ac (ChIP-seq)	ENCSR000AKH	ENCFF599TRR
	H3K4me2 (ChIP-seq)	ENCSR000AKG	ENCFF627OKN
	H4K20me1 (ChIP-seq)	ENCSR000AKI	ENCFF479XIQ
	H2AFZ (ChIP-seq)	ENCSR000AOV	ENCFF935EGN
	H3K79me2 (ChIP-seq)	ENCSR000AOW	ENCFF931USZ
	CTCF (ChIP-seq)	ENCSR000DZN	ENCFF485CGE
	POLR2A (ChIP-seq)	ENCSR000EAD	ENCFF328MMS
	RAD21 (ChIP-seq)	ENCSR000EAC	ENCFF571ZJJ
HCT116	*in situ* Hi-C	4DNESJV9TH8Q	4DNFILIM6FDL
	H3K4me3 (ChIP-seq)	ENCSR333OPW	ENCFF649ZLF
	H3K27ac (ChIP-seq)	ENCSR661KMA	ENCFF787LMI
	H3K27me3 (ChIP-seq)	ENCFF717ZKL	ENCSR810BDB
	H3K4me1 (ChIP-seq)	ENCFF239FXT	ENCSR161MXP
	H3K36me3 (ChIP-seq)	ENCFF024LGD	ENCSR091QXP
	H3K9me3 (ChIP-seq)	ENCFF254TIW	ENCSR179BUC
	H3K9ac (ChIP-seq)	ENCFF743PJP	ENCSR093SHE
	H3K4me2 (ChIP-seq)	ENCFF693HVA	ENCSR794ULT
	H4K20me1 (ChIP-seq)	ENCSR474DOV	ENCSR474DOV
	H2AFZ (ChIP-seq)	ENCFF863EHT	ENCSR227XNT
	H3K79me2 (ChIP-seq)	ENCFF631DLM	ENCSR494CCN
	CTCF (ChIP-seq)	ENCFF787LAV	ENCSR450DXU
	POLR2A (ChIP-seq)	ENCFF802CGI	ENCSR000EUU
	RAD21 (ChIP-seq)	ENCFF776IXR	ENCSR000BSB

Table B.1: Accession codes for experiments and data files used throughout Chapter 3

C

Supplementary material of Chapter 4

C.1 Derivation of U = PI + CI

Here, I provide a detailed derivation of the relationship linking the total patterning utility U to positional information (PI) and correlational information (CI). In the main text, the total patterning utility function for an ensemble of morphogen gradients was defined as:

$$U = \frac{1}{N} D_{KL}\left[P(\vec{g}|\vec{x}) \| Q(\vec{g})\right] \qquad \text{where} \qquad Q(\vec{g}) = \prod_{i=1}^{N} P_G(g_i) \tag{C.1}$$

Here, U measures the deviation of the distribution in Eq. 4.9 from the random ensemble where the gene expression levels at each location are independently sampled from $P_G(g) = \sum_x P(g|x)P_X(x)$. We can rewrite U as:

$$\begin{aligned} U &= \frac{1}{N}\int d\vec{g} P(\vec{g}|\vec{x}) \log_2 \frac{P(\vec{g}|\vec{x})}{Q(\vec{g})} = -\frac{1}{N} S[P(\vec{g}|\vec{x})] - \frac{1}{N}\int d\vec{g} P(\vec{g}|\vec{x}) \log_2 \prod_{i=1}^{N} P_G(g_i) = \\ &= -\frac{1}{N} S[P(\vec{g}|\vec{x})] - \frac{1}{N}\sum_{i=1}^{N}\int dg_i P(g_i|x_i) \log_2 \frac{P_G(g_i)P(g_i|x_i)}{P(g_i|x_i)} = \\ &= -\frac{1}{N} S[P(\vec{g}|\vec{x})] + \frac{1}{N}\sum_{i=1}^{N} S[P(g_i|x_i)] + \frac{1}{N}\sum_{i=1}^{N} D_{KL}\left[P(g_i|x_i)\|P_G(g_i)\right] \end{aligned} \tag{C.2}$$

The last term can be expressed as:

$$\frac{1}{N}\sum_{i=1}^{N} D_{KL}\left[P(g_i|x_i)\|P_G(g_i)\right] = \frac{1}{N}\sum_{i=1}^{N}\int dg_i P(g_i|x_i)\log_2\frac{P(g_i|x_i)}{P_G(g_i)}\frac{P_X(x_i)}{P_X(x_i)} = \\ = \sum_{i=1}^{N}\int dg_i P(g_i,x_i)\log_2\frac{P(g_i,x_i)}{P_G(g_i)P_X(x_i)} \equiv PI \quad \text{(C.3)}$$

where we have used $P_X(x_i) = 1/N$. The first two terms in the expression for U can be expressed as:

$$-\frac{1}{N}S[P(\vec{g}|\vec{x})] + \frac{1}{N}\sum_{i=1}^{N} S[P(g_i|x_i)] = \\ =\frac{1}{N}D_{KL}\left[P(\vec{g}|\vec{x})\middle\|\prod_{i=1}^{N}P(g_i|x_i)\right] \equiv CI \quad \text{(C.4)}$$

therefore proving the relationship:

$$U = PI + CI \quad \text{(C.5)}$$

C.2 Structure of covariance matrix of gene expression level fluctuations

Throughout Chapter 4 we used ensembles of morphogen profiles with an exponentially decaying mean profile $\bar{g}(x) = \mu_s \exp(-x/\lambda)$ where λ is the characteristic lengthscale of the gradient. The individual profiles are generated by adding spatially correlated fluctuations on top of the mean profile. Specifically, the covariance between the fluctuations at cell i and cell j reads:

$$C_{ij} = \sigma_s^2 \exp\left(-\frac{x_i + x_j}{\lambda}\right)\exp\left(-\frac{(x_i - x_j)^2}{2(L_{corr})^2}\right) + \sigma_{ind}^2\delta_{ij} \quad \text{(C.6)}$$

The first exponential models the variation in fluctuations amplitude from the anterior end to the posterior end of the embryo, the second exponential models correlations in the fluctuations that decay over a length scale of L_{corr}. Two additional parameters σ_s and σ_{ind} modulate the amplitudes of the correlated and uncorrelated components of the fluctuations. In practice, an instance of correlated fluctuations for one embryo is obtained by sampling from a Gaussian process [204] with radial basis function kernel with lengthscale L_{corr} and amplitude σ_s. Each element i of this vector is then rescaled by a factor $\exp(-x_i/\lambda)$ and independent Gaussian noise with standard deviation σ_{ind} is added on top.

Bibliography

[1] Eric S. Lander et al. "Initial sequencing and analysis of the human genome". en. In: *Nature* 409.6822 (Feb. 2001). Number: 6822 Publisher: Nature Publishing Group, pp. 860–921. ISSN: 1476-4687. DOI: 10.1038/35057062.

[2] Bruce Alberts et al. *Molecular biology of the cell*. eng. Seventh edition, international student edition. New York, NY London: W.W. Norton & Company, 2022. ISBN: 978-0-393-88485-2.

[3] Bradley E. Bernstein, Alexander Meissner, and Eric S. Lander. "The Mammalian Epigenome". English. In: *Cell* 128.4 (Feb. 2007). Number: 4 Publisher: Elsevier, pp. 669–681. ISSN: 0092-8674, 1097-4172. DOI: 10.1016/j.cell.2007.01.033.

[4] C. H. Waddington. "Canalization of Development and the Inheritance of Acquired Characters". en. In: *Nature* 150.3811 (Nov. 1942). Number: 3811 Publisher: Nature Publishing Group, pp. 563–565. ISSN: 1476-4687. DOI: 10.1038/150563a0.

[5] Jason Ernst et al. "Mapping and analysis of chromatin state dynamics in nine human cell types". en. In: *Nature* 473.7345 (May 2011). Number: 7345 Publisher: Nature Publishing Group, pp. 43–49. ISSN: 1476-4687. DOI: 10.1038/nature09906.

[6] Oliver J. Rando and Howard Y. Chang. "Genome-Wide Views of Chromatin Structure". en. In: *Annual Review of Biochemistry* 78.Volume 78, 2009 (July 2009). Number: Volume 78, 2009 Publisher: Annual Reviews, pp. 245–271. ISSN: 0066-4154, 1545-4509. DOI: 10.1146/annurev.biochem.78.071107.134639.

[7] David S. Johnson et al. "Genome-Wide Mapping of in Vivo Protein-DNA Interactions". In: *Science* 316.5830 (June 2007). Number: 5830 Publisher: American Association for the Advancement of Science, pp. 1497–1502. DOI: 10.1126/science.1141319.

[8] Elphège P. Nora et al. "Spatial partitioning of the regulatory landscape of the X-inactivation centre". eng. In: *Nature* 485.7398 (Apr. 2012). Number: 7398, pp. 381–385. ISSN: 1476-4687. DOI: 10.1038/nature11049.

[9] Darío G. Lupiáñez et al. "Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions". eng. In: *Cell* 161.5 (May 2015). Number: 5, pp. 1012–1025. ISSN: 1097-4172. DOI: 10.1016/j.cell.2015.04.004.

[10] Denes Hnisz et al. "Activation of proto-oncogenes by disruption of chromosome neighborhoods". eng. In: *Science (New York, N.Y.)* 351.6280 (Mar. 2016). Number: 6280, pp. 1454–1458. ISSN: 1095-9203. DOI: 10.1126/science.aad9024.

[11] William A. Flavahan et al. "Insulator dysfunction and oncogene activation in IDH mutant gliomas". en. In: *Nature* 529.7584 (Jan. 2016). Number: 7584 Publisher: Nature Publishing Group, pp. 110–114. ISSN: 1476-4687. DOI: 10.1038/nature16490.

[12] Rafael Galupa et al. "A Conserved Noncoding Locus Regulates Random Monoallelic Xist Expression across a Topological Boundary". eng. In: *Molecular Cell* 77.2 (Jan. 2020). Number: 2, 352–367.e8. ISSN: 1097-4164. DOI: 10.1016/j.molcel.2019.10.030.

[13] Alexandra Despang et al. "Functional dissection of the Sox9–Kcnj2 locus identifies nonessential and instructive roles of TAD architecture". en. In: *Nature Genetics* 51.8 (Aug. 2019). Number: 8 Publisher: Nature Publishing Group, pp. 1263–1271. ISSN: 1546-1718. DOI: 10.1038/s41588-019-0466-z.

[14] Elphège P. Nora et al. "Targeted Degradation of CTCF Decouples Local Insulation of Chromosome Domains from Genomic Compartmentalization". In: *Cell* 169.5 (May 2017). Number: 5, 930–944.e22. ISSN: 0092-8674. DOI: 10.1016/j.cell.2017.05.004.

[15] Yad Ghavi-Helm et al. "Highly rearranged chromosomes reveal uncoupling between genome topology and gene expression". en. In: *Nature Genetics* 51.8 (Aug. 2019). Number: 8 Publisher: Nature Publishing Group, pp. 1272–1282. ISSN: 1546-1718. DOI: 10.1038/s41588-019-0462-3.

[16] Julien O. Dubuis et al. "Positional information, in bits". In: *Proceedings of the National Academy of Sciences* 110.41 (Oct. 2013). Publisher: Proceedings of the National Academy of Sciences, pp. 16301–16308. DOI: 10.1073/pnas.1315642110.

[17] Gašper Tkačik and William Bialek. "Information Processing in Living Systems". In: *Annual Review of Condensed Matter Physics* 7.1 (2016). _eprint: https://doi.org/10.1146/annurev-conmatphys-031214-014803, pp. 89–117. DOI: 10.1146/annurev-conmatphys-031214-014803.

[18] Mariela D. Petkova et al. "Optimal Decoding of Cellular Identities in a Genetic Network". In: *Cell* 176.4 (Feb. 2019), 844–855.e15. ISSN: 0092-8674. DOI: 10.1016/j.cell.2019.01.007.

[19] Marcin Zagorski et al. "Decoding of position in the developing neural tube from antiparallel morphogen gradients". In: *Science* 356.6345 (June 2017). Publisher: American Association for the Advancement of Science, pp. 1379–1383. DOI: 10.1126/science.aam5887.

[20] James Fraser et al. "An Overview of Genome Organization and How We Got There: from FISH to Hi-C". In: *Microbiology and Molecular Biology Reviews : MMBR* 79.3 (July 2015). Number: 3, pp. 347–372. ISSN: 1092-2172. DOI: 10.1128/MMBR.00006-15.

[21] P R Langer-Safer, M Levine, and D C Ward. "Immunological method for mapping genes on Drosophila polytene chromosomes." In: *Proceedings of the National Academy of Sciences* 79.14 (July 1982). Number: 14 Publisher: Proceedings of the National Academy of Sciences, pp. 4381–4385. DOI: 10.1073/pnas.79.14.4381.

[22] T. Cremer and C. Cremer. "Chromosome territories, nuclear architecture and gene regulation in mammalian cells". en. In: *Nature Reviews Genetics* 2.4 (Apr. 2001). Number: 4 Publisher: Nature Publishing Group, pp. 292–301. ISSN: 1471-0064. DOI: 10.1038/35066075.

[23] Thomas Cremer and C. Cremer. "Rise, fall and resurrection of chromosome territories: a historical perspective. Part I. The rise of chromosome territories". eng. In: *European journal of histochemistry: EJH* 50.3 (2006). Number: 3, pp. 161–176. ISSN: 1121-760X.

[24] T. Cremer and C. Cremer. "Rise, fall and resurrection of chromosome territories: a historical perspective. Part II. Fall and resurrection of chromosome territories during the 1950s to 1980s. Part III. Chromosome territories and the functional nuclear architecture: experiments and models from the 1990s to the present". eng. In: *European journal of histochemistry: EJH* 50.4 (2006). Number: 4, pp. 223–272. ISSN: 1121-760X.

[25] Job Dekker et al. “Capturing Chromosome Conformation”. In: *Science* 295.5558 (Feb. 2002). Number: 5558 Publisher: American Association for the Advancement of Science, pp. 1306–1311. DOI: 10.1126/science.1067799.

[26] Erez Lieberman-Aiden et al. “Comprehensive mapping of long range interactions reveals folding principles of the human genome”. In: *Science (New York, N.Y.)* 326.5950 (Oct. 2009). Number: 5950, pp. 289–293. ISSN: 0036-8075. DOI: 10.1126/science.1181369.

[27] Stefan Schoenfelder et al. “The pluripotent regulatory circuitry connecting promoters to their long-range interacting elements”. en. In: *Genome Research* 25.4 (Apr. 2015). Number: 4 Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab, pp. 582–597. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.185272.114.

[28] James O. J. Davies et al. “Multiplexed analysis of chromosome conformation at vastly improved sensitivity”. en. In: *Nature Methods* 13.1 (Jan. 2016). Number: 1 Publisher: Nature Publishing Group, pp. 74–80. ISSN: 1548-7105. DOI: 10.1038/nmeth.3664.

[29] Borbala Mifsud et al. “Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C”. en. In: *Nature Genetics* 47.6 (June 2015). Number: 6 Publisher: Nature Publishing Group, pp. 598–606. ISSN: 1546-1718. DOI: 10.1038/ng.3286.

[30] Roland Jäger et al. “Capture Hi-C identifies the chromatin interactome of colorectal cancer risk loci”. en. In: *Nature Communications* 6.1 (Feb. 2015). Number: 1 Publisher: Nature Publishing Group, p. 6178. ISSN: 2041-1723. DOI: 10.1038/ncomms7178.

[31] Guillaume Andrey et al. “Characterization of hundreds of regulatory landscapes in developing limbs reveals two regimes of chromatin folding”. en. In: *Genome Research* 27.2 (Feb. 2017). Number: 2 Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab, pp. 223–233. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.213066.116.

[32] Eric Betzig et al. “Imaging Intracellular Fluorescent Proteins at Nanometer Resolution”. In: *Science* 313.5793 (Sept. 2006). Publisher: American Association for the Advancement of Science, pp. 1642–1645. DOI: 10.1126/science.1127344.

[33] Michael J. Rust, Mark Bates, and Xiaowei Zhuang. "Stochastic optical reconstruction microscopy (STORM) provides sub-diffraction-limit image resolution". In: *Nature methods* 3.10 (Oct. 2006), pp. 793–795. ISSN: 1548-7091. DOI: 10.1038/nmeth929.

[34] Brian J. Beliveau et al. "Single-molecule super-resolution imaging of chromosomes and in situ haplotype visualization using Oligopaint FISH probes". en. In: *Nature Communications* 6.1 (May 2015). Publisher: Nature Publishing Group, p. 7147. ISSN: 2041-1723. DOI: 10.1038/ncomms8147.

[35] Angelo Rosa and Christophe Zimmer. "Chapter Nine - Computational Models of Large-Scale Genome Architecture". In: *International Review of Cell and Molecular Biology*. Ed. by Ronald Hancock and Kwang W. Jeon. Vol. 307. New Models of the Cell Nucleus: Crowding, Entropic Forces, Phase Separation, and Fractals. Academic Press, Jan. 2014, pp. 275–349. DOI: 10.1016/B978-0-12-800046-5.00009-6.

[36] Masao Doi and Samuel F. Edwards. *The theory of polymer dynamics*. eng. Reprint. International series of monographs on physics 73. Oxford: Clarendon Press, 2013. ISBN: 978-0-19-852033-7.

[37] Michael Rubinstein and Ralph H. Colby. *Polymer physics*. eng. Oxford New York: Oxford university press, 2003. ISBN: 978-0-19-852059-7.

[38] Maxim V. Imakaev, Geoffrey Fudenberg, and Leonid A. Mirny. "Modeling chromosomes: beyond pretty pictures". In: *FEBS letters* 589.20 0 0 (Oct. 2015), pp. 3031–3036. ISSN: 0014-5793. DOI: 10.1016/j.febslet.2015.09.004.

[39] Daniel Jost, Cédric Vaillant, and Peter Meister. "Coupling 1D modifications and 3D nuclear organization: data, models and function". eng. In: *Current Opinion in Cell Biology* 44 (Feb. 2017), pp. 20–27. ISSN: 1879-0410. DOI: 10.1016/j.ceb.2016.12.001.

[40] Angelo Rosa and Ralf Everaers. "Structure and Dynamics of Interphase Chromosomes". en. In: *PLOS Computational Biology* 4.8 (2008). Number: 8 Publisher: Public Library of Science, e1000153. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1000153.

[41] Angelo Rosa, Nils B. Becker, and Ralf Everaers. "Looping Probabilities in Model Interphase Chromosomes". In: *Biophysical Journal* 98.11 (June 2010). Number: 11, pp. 2410–2419. ISSN: 0006-3495. DOI: 10.1016/j.bpj.2010.01.054.

[42] Fabrizio Benedetti et al. "Models that include supercoiling of topological domains reproduce several known features of interphase chromosomes". In: *Nucleic Acids Research* 42.5 (Mar. 2014). Number: 5, pp. 2848–2855. ISSN: 0305-1048. DOI: 10.1093/nar/gkt1353.

[43] Dusan Racko et al. "Transcription-induced supercoiling as the driving force of chromatin loop extrusion during formation of TADs in interphase chromosomes". In: *Nucleic Acids Research* 46.4 (Feb. 2018). Number: 4, pp. 1648–1660. ISSN: 0305-1048. DOI: 10.1093/nar/gkx1123.

[44] Jason Ernst and Manolis Kellis. "ChromHMM: automating chromatin-state discovery and characterization". en. In: *Nature Methods* 9.3 (Mar. 2012). Number: 3 Publisher: Nature Publishing Group, pp. 215–216. ISSN: 1548-7105. DOI: 10.1038/nmeth.1906.

[45] Davide Colì et al. "Magnetic polymer models for epigenetics-driven chromosome folding". In: *Physical Review E* 100.5 (Nov. 2019). Number: 5 Publisher: American Physical Society, p. 052410. DOI: 10.1103/PhysRevE.100.052410.

[46] Daniel Jost et al. "Modeling epigenome folding: formation and dynamics of topologically associated chromatin domains". In: *Nucleic Acids Research* 42.15 (Sept. 2014), pp. 9553–9561. ISSN: 0305-1048. DOI: 10.1093/nar/gku698.

[47] Chris A. Brackley et al. "Simulated binding of transcription factors to active and inactive regions folds human chromosomes into loops, rosettes and topological domains". eng. In: *Nucleic Acids Research* 44.8 (May 2016). Number: 8, pp. 3503–3512. ISSN: 1362-4962. DOI: 10.1093/nar/gkw135.

[48] Mariano Barbieri et al. "Complexity of chromatin folding is captured by the strings and binders switch model". In: *Proceedings of the National Academy of Sciences* 109.40 (Oct. 2012). Number: 40 Publisher: Proceedings of the National Academy of Sciences, pp. 16173–16178. DOI: 10.1073/pnas.1204799109.

[49] Steffen Rulands et al. "Genome-Scale Oscillations in DNA Methylation during Exit from Pluripotency". In: *Cell Systems* 7.1 (July 2018). Number: 1, 63–76.e12. ISSN: 2405-4712. DOI: 10.1016/j.cels.2018.06.012.

[50] Scott Berry, Caroline Dean, and Martin Howard. "Slow Chromatin Dynamics Allow Polycomb Target Genes to Filter Fluctuations in Transcription Factor Activity". In: *Cell Systems* 4.4 (Apr. 2017). Number: 4, 445–457.e8. ISSN: 2405-4712. DOI: 10.1016/j.cels.2017.02.013.

[51] Daniel Jost. "Bifurcation in epigenetics: Implications in development, proliferation, and diseases". In: *Physical Review E* 89.1 (Jan. 2014). Number: 1 Publisher: American Physical Society, p. 010701. DOI: 10.1103/PhysRevE.89.010701.

[52] T. Garel, H. Orland, and E. Orlandini. "Phase diagram of magnetic polymers". en. In: *The European Physical Journal B* 12.2 (Nov. 1999). Number: 2, pp. 261–268. ISSN: 1434-6028. DOI: 10.1007/s100510051003.

[53] Davide Michieletto et al. "Nonequilibrium Theory of Epigenomic Microphase Separation in the Cell Nucleus". In: *Physical Review Letters* 123.22 (Nov. 2019). Number: 22 Publisher: American Physical Society, p. 228101. DOI: 10.1103/PhysRevLett.123.228101.

[54] Mark E. J. Newman. *Networks: an introduction*. eng. Reprinted. Oxford: Oxford University Press, 2016. ISBN: 978-0-19-920665-0.

[55] Heidi K. Norton et al. "Detecting hierarchical genome folding with network modularity". en. In: *Nature Methods* 15.2 (Feb. 2018). Number: 2 Publisher: Nature Publishing Group, pp. 119–122. ISSN: 1548-7105. DOI: 10.1038/nmeth.4560.

[56] Asa Thibodeau et al. "Chromatin interaction networks revealed unique connectivity patterns of broad H3K4me3 domains and super enhancers in 3D chromatin". en. In: *Scientific Reports* 7.1 (Oct. 2017). Number: 1 Publisher: Nature Publishing Group, p. 14466. ISSN: 2045-2322. DOI: 10.1038/s41598-017-14389-7.

[57] Vera Pancaldi. "Chromatin Network Analyses: Towards Structure-Function Relationships in Epigenomics". English. In: *Frontiers in Bioinformatics* 1 (Oct. 2021). Publisher: Frontiers. ISSN: 2673-7647. DOI: 10.3389/fbinf.2021.742216.

[58] Yuchung J. Wang and George Y. Wong. "Stochastic Blockmodels for Directed Graphs". In: *Journal of the American Statistical Association* 82.397 (Mar. 1987). Number: 397 Publisher: Taylor & Francis _eprint: https://www.tandfonline.com/doi/pdf/10.1080/01621459.1987.10478385, pp. 8–19. ISSN: 0162-1459. DOI: 10.1080/01621459.1987.10478385.

[59] Tom A.B. Snijders and Krzysztof Nowicki. "Estimation and Prediction for Stochastic Blockmodels for Graphs with Latent Block Structure". en. In: *Journal of Classification* 14.1 (Jan. 1997). Number: 1, pp. 75–100. ISSN: 1432-1343. DOI: 10.1007/s003579900004.

[60] C. E. Shannon. "A Mathematical Theory of Communication". en. In: *Bell System Technical Journal* 27.3 (1948). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/j.1538-7305.1948.tb01338.x, pp. 379–423. ISSN: 1538-7305. DOI: 10.1002/j.1538-7305.1948.tb01338.x.

[61] Thomas M. Cover and Joy A. Thomas. *Elements of information theory*. eng. 2nd ed. Hoboken, N.J: Wiley-Interscience, 2006. ISBN: 978-0-471-24195-9.

[62] David Barber. In: *Bayesian reasoning and machine learning*. Cambridge ; New York: Cambridge University Press, 2012, pp. 49–62. ISBN: 978-0-521-51814-7.

[63] Andrew Roth et al. "PyClone: statistical inference of clonal population structure in cancer". en. In: *Nature Methods* 11.4 (Apr. 2014). Publisher: Nature Publishing Group, pp. 396–398. ISSN: 1548-7105. DOI: 10.1038/nmeth.2883.

[64] Catalina A. Vallejos, John C. Marioni, and Sylvia Richardson. "BASiCS: Bayesian Analysis of Single-Cell Sequencing Data". en. In: *PLOS Computational Biology* 11.6 (June 2015). Publisher: Public Library of Science, e1004333. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1004333.

[65] Carmen Bravo González-Blas et al. "cisTopic: cis-regulatory topic modeling on single-cell ATAC-seq data". en. In: *Nature Methods* 16.5 (May 2019). Publisher: Nature Publishing Group, pp. 397–400. ISSN: 1548-7105. DOI: 10.1038/s41592-019-0367-1.

[66] Nour El Kazwini and Guido Sanguinetti. "SHARE-Topic: Bayesian interpretable modeling of single-cell multi-omic data". In: *Genome Biology* 25 (Feb. 2024), p. 55. ISSN: 1474-7596. DOI: 10.1186/s13059-024-03180-3.

[67] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. "Latent Dirichlet Allocation". In: *Journal of Machine Learning Research* 3.Jan (2003), pp. 993–1022. ISSN: ISSN 1533-7928.

[68] Nicholas Metropolis et al. "Equation of State Calculations by Fast Computing Machines". In: *The Journal of Chemical Physics* 21.6 (June 1953). Number: 6, pp. 1087–1092. ISSN: 0021-9606. DOI: 10.1063/1.1699114.

[69] Matthew J Beal and Zoubin Ghahramani. "The Variational Bayesian EM Algorithm for Incomplete Data: With Application to Scoring Graphical Model Structures". In: *Bayesian Statistics 7: Proceedings of the Seventh Valencia International Meeting*. Ed. by V Lindley et al. Oxford University Press, July 2003. ISBN: 978-0-19-852615-5. DOI: 10.1093/oso/9780198526155.003.0025.

[70] Christopher M. Bishop. *Pattern recognition and machine learning*. eng. Information science and statistics. New York: Springer, 2006. ISBN: 978-0-387-31073-2.

[71] A. P. Dempster, N. M. Laird, and D. B. Rubin. "Maximum Likelihood from Incomplete Data Via the EM Algorithm". en. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 39.1 (1977). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.2517-6161.1977.tb01600.x, pp. 1–22. ISSN: 2517-6161. DOI: 10.1111/j.2517-6161.1977.tb01600.x.

[72] Rajesh Ranganath, Sean Gerrish, and David Blei. "Black Box Variational Inference". en. In: *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*. ISSN: 1938-7228. PMLR, Apr. 2014, pp. 814–822.

[73] Shelley Sazer and Helmut Schiessel. "The biology and polymer physics underlying large-scale chromosome organization". eng. In: *Traffic (Copenhagen, Denmark)* 19.2 (Feb. 2018). Number: 2, pp. 87–104. ISSN: 1600-0854. DOI: 10.1111/tra.12539.

[74] M. Jordan Rowley and Victor G. Corces. "Organizational principles of 3D genome architecture". en. In: *Nature Reviews Genetics* 19.12 (Dec. 2018). Number: 12 Publisher: Nature Publishing Group, pp. 789–800. ISSN: 1471-0064. DOI: 10.1038/s41576-018-0060-8.

[75] Elizabeth H. Finn and Tom Misteli. "Molecular basis and biological function of variability in spatial genome organization". In: *Science* 365.6457 (Sept. 2019). Number: 6457 Publisher: American Association for the Advancement of Science, eaaw9498. DOI: 10.1126/science.aaw9498.

[76] Tom Misteli. "The Self-Organizing Genome: Principles of Genome Architecture and Function". English. In: *Cell* 183.1 (Oct. 2020). Number: 1 Publisher: Elsevier, pp. 28–45. ISSN: 0092-8674, 1097-4172. DOI: 10.1016/j.cell.2020.09.014.

[77] Alessandra Merlotti, Angelo Rosa, and Daniel Remondini. "Merging 1D and 3D genomic information: Challenges in modelling and validation". In: *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*. Transcriptional Profiles and Regulatory Gene Networks 1863.6 (June 2020). Number: 6, p. 194415. ISSN: 1874-9399. DOI: 10.1016/j.bbagrm.2019.194415.

[78] Jesse R. Dixon et al. "Topological domains in mammalian genomes identified by analysis of chromatin interactions". en. In: *Nature* 485.7398 (May 2012). Number: 7398 Publisher: Nature Publishing Group, pp. 376–380. ISSN: 1476-4687. DOI: 10.1038/nature11082.

[79] Gaurav Bajpai et al. “Mesoscale phase separation of chromatin in the nucleus”. In: *eLife* 10 (May 2021). Ed. by Karsten Kruse, K VijayRaghavan, and Helmut Schiessel. Publisher: eLife Sciences Publications, Ltd, e63976. ISSN: 2050-084X. DOI: 10.7554/eLife.63976.

[80] Timothy Földes, Antony Lesage, and Maria Barbi. “Assessing the Polymer Coil-Globule State from the Very First Spectral Modes”. In: *Physical Review Letters* 127.27 (Dec. 2021). Number: 27 Publisher: American Physical Society, p. 277801. DOI: 10.1103/PhysRevLett.127.277801.

[81] Elnaz Alipour and John F. Marko. “Self-organization of domain structures by DNA-loop-extruding enzymes”. eng. In: *Nucleic Acids Research* 40.22 (Dec. 2012). Number: 22, pp. 11202–11212. ISSN: 1362-4962. DOI: 10.1093/nar/gks925.

[82] Geoffrey Fudenberg et al. “Formation of Chromosomal Domains by Loop Extrusion”. eng. In: *Cell Reports* 15.9 (May 2016). Number: 9, pp. 2038–2049. ISSN: 2211-1247. DOI: 10.1016/j.celrep.2016.04.085.

[83] Jonathan D. Halverson et al. “From a melt of rings to chromosome territories: the role of topological constraints in genome folding”. eng. In: *Reports on Progress in Physics. Physical Society (Great Britain)* 77.2 (2014). Number: 2, p. 022601. ISSN: 1361-6633. DOI: 10.1088/0034-4885/77/2/022601.

[84] Mark Ptashne. “The Chemistry of Regulation of Genes and Other Things”. In: *The Journal of Biological Chemistry* 289.9 (Feb. 2014). Number: 9, pp. 5417–5435. ISSN: 0021-9258. DOI: 10.1074/jbc.X114.547323.

[85] Mark Ptashne and Alexander Gann. *Genes & signals*. Cold Spring Harbor, New York: Cold Spring Harbor Laboratory Press, 2002. ISBN: 978-0-87969-631-3.

[86] Dan Benveniste et al. “Transcription factor binding predicts histone modifications in human cell lines”. In: *Proceedings of the National Academy of Sciences* 111.37 (Sept. 2014). Number: 37 Publisher: Proceedings of the National Academy of Sciences, pp. 13367–13372. DOI: 10.1073/pnas.1412081111.

[87] John W. Whitaker, Zhao Chen, and Wei Wang. “Predicting the human epigenome from DNA motifs”. en. In: *Nature Methods* 12.3 (Mar. 2015). Number: 3 Publisher: Nature Publishing Group, pp. 265–272. ISSN: 1548-7105. DOI: 10.1038/nmeth.3065.

[88] Jacob Schreiber et al. “Avocado: a multi-scale deep tensor factorization method learns a latent representation of the human epigenome”. In: *Genome Biology* 21.1 (Mar. 2020), p. 81. ISSN: 1474-760X. DOI: 10.1186/s13059-020-01977-6.

[89] Alex Hawkins-Hooker et al. "Getting personal with epigenetics: towards individual-specific epigenomic imputation with machine learning". en. In: *Nature Communications* 14.1 (Aug. 2023). Number: 1 Publisher: Nature Publishing Group, p. 4750. ISSN: 2041-1723. DOI: 10.1038/s41467-023-40211-2.

[90] Simona Bianco et al. "Polymer physics predicts the effects of structural variants on chromatin architecture". en. In: *Nature Genetics* 50.5 (May 2018). Number: 5 Publisher: Nature Publishing Group, pp. 662–667. ISSN: 1546-1718. DOI: 10.1038/s41588-018-0098-8.

[91] Michele Di Pierro et al. "Transferable model for chromosome architecture". In: *Proceedings of the National Academy of Sciences* 113.43 (Oct. 2016). Number: 43 Publisher: Proceedings of the National Academy of Sciences, pp. 12168–12173. DOI: 10.1073/pnas.1613607113.

[92] Michele Di Pierro et al. "De novo prediction of human chromosome structures: Epigenetic marking patterns encode genome architecture". In: *Proceedings of the National Academy of Sciences* 114.46 (Nov. 2017). Number: 46 Publisher: Proceedings of the National Academy of Sciences, pp. 12126–12131. DOI: 10.1073/pnas.1714980114.

[93] Geoff Fudenberg, David R. Kelley, and Katherine S. Pollard. "Predicting 3D genome folding from DNA sequence with Akita". en. In: *Nature Methods* 17.11 (Nov. 2020). Number: 11 Publisher: Nature Publishing Group, pp. 1111–1117. ISSN: 1548-7105. DOI: 10.1038/s41592-020-0958-x.

[94] Neal Madras and Alan D. Sokal. "The pivot algorithm: A highly efficient Monte Carlo method for the self-avoiding walk". en. In: *Journal of Statistical Physics* 50.1 (Jan. 1988). Number: 1, pp. 109–186. ISSN: 1572-9613. DOI: 10.1007/BF01022990.

[95] Shiv I. S. Grewal and Danesh Moazed. "Heterochromatin and Epigenetic Control of Gene Expression". In: *Science* 301.5634 (Aug. 2003). Number: 5634 Publisher: American Association for the Advancement of Science, pp. 798–802. DOI: 10.1126/science.1086887.

[96] S. A. Sisson, Y. Fan, and M. A. Beaumont. *Overview of Approximate Bayesian Computation*. Issue: arXiv:1802.09720 arXiv:1802.09720 [stat]. Feb. 2018. DOI: 10.48550/arXiv.1802.09720.

[97] Kyle Cranmer, Johann Brehmer, and Gilles Louppe. "The frontier of simulation-based inference". In: *Proceedings of the National Academy of Sciences* 117.48 (Dec. 2020). Publisher: Proceedings of the National Academy of Sciences, pp. 30055–30062. DOI: 10.1073/pnas.1912789117.

[98] Enzo Orlandini. "Monte Carlo Study of Polymer Systems by Multiple Markov Chain Method". en. In: *Numerical Methods for Polymeric Systems*. Ed. by Stuart G. Whittington, Avner Friedman, and Robert Gulliver. Vol. 102. Series Editors: _:n506 Series Title: The IMA Volumes in Mathematics and its Applications. New York, NY: Springer New York, 1998, pp. 33–57. ISBN: 978-1-4612-7249-6. DOI: 10.1007/978-1-4612-1704-6_3.

[99] *ENCODE*. June 2024.

[100] *4DN Data Portal*. June 2024.

[101] Simon N. Wood. "Statistical inference for noisy nonlinear ecological dynamic systems". en. In: *Nature* 466.7310 (Aug. 2010). Number: 7310 Publisher: Nature Publishing Group, pp. 1102–1104. ISSN: 1476-4687. DOI: 10.1038/nature09319.

[102] Quinn MacPherson, Bruno Beltran, and Andrew J. Spakowitz. "Bottom–up modeling of chromatin segregation due to epigenetic modifications". In: *Proceedings of the National Academy of Sciences* 115.50 (Dec. 2018). Number: 50 Publisher: Proceedings of the National Academy of Sciences, pp. 12739–12744. DOI: 10.1073/pnas.1812268115.

[103] James P. Reddington et al. "Redistribution of H3K27me3 upon DNA hypomethylation results in de-repression of Polycomb target genes". eng. In: *Genome Biology* 14.3 (Mar. 2013). Number: 3, R25. ISSN: 1474-760X. DOI: 10.1186/gb-2013-14-3-r25.

[104] Suhas S. P. Rao et al. "A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping". English. In: *Cell* 159.7 (Dec. 2014). Publisher: Elsevier, pp. 1665–1680. ISSN: 0092-8674, 1097-4172. DOI: 10.1016/j.cell.2014.11.021.

[105] Tao Yang et al. "HiCRep: assessing the reproducibility of Hi-C data using a stratum-adjusted correlation coefficient". eng. In: *Genome Research* 27.11 (Nov. 2017). Number: 11, pp. 1939–1949. ISSN: 1549-5469. DOI: 10.1101/gr.220640.117.

[106] D. Michieletto, E. Orlandini, and D. Marenduzzo. "Polymer model with Epigenetic Recoloring Reveals a Pathway for the de novo Establishment and 3D Organization of Chromatin Domains". In: *Physical Review X* 6.4 (Dec. 2016). Number: 4 Publisher: American Physical Society, p. 041047. DOI: 10.1103/PhysRevX.6.041047.

[107] Adam Buckle et al. "Polymer Simulations of Heteromorphic Chromatin Predict the 3D Folding of Complex Genomic Loci". eng. In: *Molecular Cell* 72.4 (Nov. 2018). Number: 4, 786–797.e11. ISSN: 1097-4164. DOI: 10.1016/j.molcel.2018.09.016.

[108] Michael Chiang et al. "Predictive Polymer Models for 3D Chromosome Organization". eng. In: *Methods in Molecular Biology (Clifton, N.J.)* 2301 (2022), pp. 267–291. ISSN: 1940-6029. DOI: 10.1007/978-1-0716-1390-0_14.

[109] H. E. Stanley. "Dependence of Critical Properties on Dimensionality of Spins". In: *Physical Review Letters* 20.12 (Mar. 1968). Number: 12 Publisher: American Physical Society, pp. 589–592. DOI: 10.1103/PhysRevLett.20.589.

[110] Alex Chen Yi Zhang, Angelo Rosa, and Guido Sanguinetti. *bioSBM: a random graph model to integrate epigenomic data in chromatin structure prediction.* arXiv:2409.14425 [physics, q-bio]. Sept. 2024. DOI: 10.48550/arXiv.2409.14425.

[111] A. Gregory Matera et al. "Nuclear Bodies: Random Aggregates of Sticky Proteins or Crucibles of Macromolecular Assembly?" In: *Developmental Cell* 17.5 (Nov. 2009), pp. 639–647. ISSN: 1534-5807. DOI: 10.1016/j.devcel.2009.10.017.

[112] Salman F. Banani et al. "Biomolecular condensates: organizers of cellular biochemistry". en. In: *Nature Reviews Molecular Cell Biology* 18.5 (May 2017). Publisher: Nature Publishing Group, pp. 285–298. ISSN: 1471-0080. DOI: 10.1038/nrm.2017.7.

[113] Liang Wang et al. "Histone Modifications Regulate Chromatin Compartmentalization by Contributing to a Phase Separation Mechanism". In: *Molecular Cell* 76.4 (Nov. 2019), 646–659.e6. ISSN: 1097-2765. DOI: 10.1016/j.molcel.2019.08.019.

[114] Anne-Marie Ladouceur et al. "Clusters of bacterial RNA polymerase are biomolecular condensates that assemble through liquid–liquid phase separation". In: *Proceedings of the National Academy of Sciences* 117.31 (Aug. 2020). Publisher: Proceedings of the National Academy of Sciences, pp. 18540–18549. DOI: 10.1073/pnas.2005019117.

[115] Michael H. Nichols and Victor G. Corces. "Principles of 3D compartmentalization of the human genome". In: *Cell Reports* 35.13 (June 2021), p. 109330. ISSN: 2211-1247. DOI: 10.1016/j.celrep.2021.109330.

[116] Alex Chen Yi Zhang, Angelo Rosa, and Guido Sanguinetti. "Bottom-up data integration in polymer models of chromatin organization". In: *Biophysical Journal* 123.2 (Jan. 2024), pp. 184–194. ISSN: 0006-3495. DOI: 10.1016/j.bpj.2023.12.006.

[117] Ron Schwessinger et al. "DeepC: predicting 3D genome folding using megabase-scale transfer learning". en. In: *Nature Methods* 17.11 (Nov. 2020). Publisher: Nature Publishing Group, pp. 1118–1124. ISSN: 1548-7105. DOI: 10.1038/s41592-020-0960-3.

[118] Jian Zhou. "Sequence-based modeling of three-dimensional genome architecture from kilobase to chromosome scale". en. In: *Nature Genetics* 54.5 (May 2022). Publisher: Nature Publishing Group, pp. 725–734. ISSN: 1546-1718. DOI: 10.1038/s41588-022-01065-4.

[119] Rui Yang et al. "Epiphany: predicting Hi-C contact maps from 1D epigenomic signals". In: *Genome Biology* 24.1 (June 2023), p. 134. ISSN: 1474-760X. DOI: 10.1186/s13059-023-02934-9.

[120] Irineo Cabreros, Emmanuel Abbe, and Aristotelis Tsirigos. *Detecting Community Structures in Hi-C Genomic Data*. arXiv:1509.05121 [cs, q-bio, stat]. Sept. 2015. DOI: 10.48550/arXiv.1509.05121.

[121] Vera Pancaldi et al. "Integrating epigenomic data and 3D genomic structure with a new measure of chromatin assortativity". In: *Genome Biology* 17.1 (July 2016). Number: 1, p. 152. ISSN: 1474-760X. DOI: 10.1186/s13059-016-1003-3.

[122] Vera Pancaldi. "Network models of chromatin structure". In: *Current Opinion in Genetics & Development* 80 (June 2023), p. 102051. ISSN: 0959-437X. DOI: 10.1016/j.gde.2023.102051.

[123] Alan Frieze and Michał Karoński. *Introduction to Random Graphs*. Cambridge: Cambridge University Press, 2015. ISBN: 978-1-107-11850-8. DOI: 10.1017/CBO9781316339831.

[124] Béla Bollobás. *Random Graphs*. 2nd ed. Cambridge Studies in Advanced Mathematics. Cambridge: Cambridge University Press, 2001. ISBN: 978-0-521-80920-7. DOI: 10.1017/CBO9780511814068.

[125] Emmanuel Abbe. "Community Detection and Stochastic Block Models: Recent Developments". In: *Journal of Machine Learning Research* 18.177 (2018), pp. 1–86. ISSN: 1533-7928.

[126] A. Yu. Grosberg. "How two meters of DNA fit into a cell nucleus: Polymer models with topological constraints and experimental data". en. In: *Polymer Science Series C* 54.1 (Sept. 2012), pp. 1–10. ISSN: 1555-614X. DOI: 10.1134/S1811238212070028.

[127] Clement Lee and Darren J. Wilkinson. "A review of stochastic block models and extensions for graph clustering". en. In: *Applied Network Science* 4.1 (Dec. 2019). Number: 1 Publisher: SpringerOpen, pp. 1–50. ISSN: 2364-8228. DOI: 10.1007/s41109-019-0232-2.

[128] J.-J. Daudin, F. Picard, and S. Robin. "A mixture model for random graphs". en. In: *Statistics and Computing* 18.2 (June 2008), pp. 173–183. ISSN: 1573-1375. DOI: 10.1007/s11222-007-9046-7.

[129] Mahendra Mariadassou, Stéphane Robin, and Corinne Vacher. "Uncovering latent structure in valued graphs: A variational approach". In: *The Annals of Applied Statistics* 4.2 (June 2010). Publisher: Institute of Mathematical Statistics, pp. 715–742. ISSN: 1932-6157, 1941-7330. DOI: 10.1214/10-AOAS361.

[130] Edoardo M. Airoldi et al. "Mixed Membership Stochastic Blockmodels". In: *Journal of machine learning research : JMLR* 9 (Sept. 2008), pp. 1981–2014. ISSN: 1532-4435.

[131] Greg M. Allenby, Peter E. Rossi, and Robert E. McCulloch. *Hierarchical Bayes Models: A Practitioners Guide*. en. SSRN Scholarly Paper. Rochester, NY, Jan. 2005. DOI: 10.2139/ssrn.655541.

[132] Andrew Gelman et al. *Bayesian data analysis*. eng. Third edition. Texts in statistical science series. Boca Raton London New York: CRC Press, Taylor and Francis Group, 2014. ISBN: 978-1-4398-4095-5.

[133] J. Aitchison. "The Statistical Analysis of Compositional Data". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 44.2 (Jan. 1982), pp. 139–160. ISSN: 0035-9246. DOI: 10.1111/j.2517-6161.1982.tb01195.x.

[134] David M. Blei and John D. Lafferty. "A correlated topic model of Science". In: *The Annals of Applied Statistics* 1.1 (June 2007). arXiv:0708.3601 [stat]. ISSN: 1932-6157. DOI: 10.1214/07-AOAS114.

[135] Andrea Esposito et al. "Polymer physics reveals a combinatorial code linking 3D chromatin architecture to 1D chromatin states". In: *Cell Reports* 38.13 (Mar. 2022), p. 110601. ISSN: 2211-1247. DOI: 10.1016/j.celrep.2022.110601.

[136] Suhas S. P. Rao et al. "Cohesin Loss Eliminates All Loop Domains". In: *Cell* 171.2 (Oct. 2017), 305–320.e24. ISSN: 0092-8674. DOI: 10.1016/j.cell.2017.09.026.

[137] Ivana Jerkovic´ and Giacomo Cavalli. "Understanding 3D genome organization by multidisciplinary methods". In: *Nature Reviews Molecular Cell Biology* 22.8 (Aug. 2021). Number: 8 Publisher: Nature Publishing Group, pp. 511–528. ISSN: 1471-0080. DOI: 10.1038/s41580-021-00362-w.

[138] Stefan Schoenfelder et al. "Promoter Capture Hi-C: High-resolution, Genome-wide Profiling of Promoter Interactions". In: *Journal of Visualized Experiments : JoVE* 136 (June 2018), p. 57320. ISSN: 1940-087X. DOI: 10.3791/57320.

[139] Guoliang Li et al. "ChIA-PET tool for comprehensive chromatin interaction analysis with paired-end tag sequencing". In: *Genome Biology* 11.2 (Feb. 2010), R22. ISSN: 1474-760X. DOI: 10.1186/gb-2010-11-2-r22.

[140] Guoliang Li et al. "Chromatin Interaction Analysis with Paired-End Tag (ChIA-PET) sequencing technology and application". In: *BMC Genomics* 15.12 (Dec. 2014), S11. ISSN: 1471-2164. DOI: 10.1186/1471-2164-15-S12-S11.

[141] Anne Ephrussi and Daniel St Johnston. "Seeing Is Believing: The Bicoid Morphogen Gradient Matures". In: *Cell* 116.2 (Jan. 2004), pp. 143–152. ISSN: 0092-8674. DOI: 10.1016/S0092-8674(04)00037-6.

[142] Lewis Wolpert. *Principles of development*. eng. 5th ed. Oxford: Oxford university press, 2015. ISBN: 978-0-19-870988-6.

[143] Theodor Boveri. *Die Polarität von Ovocyte, Ei und Larve des Stronglyocentrotus lividus*. German. 1901.

[144] Edmund B. Wilson. "Experimental studies on germinal localization". en. In: *Journal of Experimental Zoology* 1.1 (1904). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jez.1400010102, pp. 1–72. ISSN: 1097-010X. DOI: 10.1002/jez.1400010102.

[145] Edwing. Conklin. "Organ-forming substances in the eggs of ascidians". In: *The Biological Bulletin* 8.4 (Mar. 1905). Publisher: The University of Chicago Press, pp. 205–230. ISSN: 0006-3185. DOI: 10.2307/1535879.

[146] Alan Mathison Turing. "The chemical basis of morphogenesis". In: *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 237.641 (Jan. 1997). Publisher: Royal Society, pp. 37–72. DOI: 10.1098/rstb.1952.0012.

[147] Lewis Wolpert. "Chapter Thirty-Five - Positional Information and Pattern Formation". In: *Current Topics in Developmental Biology*. Ed. by Paul M. Wassarman. Vol. 117. Essays on Developmental Biology, Part B. Academic Press, Jan. 2016, pp. 597–608. DOI: 10.1016/bs.ctdb.2015.11.008.

[148] Lewis Wolpert. "One hundred years of positional information". In: *Trends in Genetics* 12.9 (Sept. 1996). Number: 9, pp. 359–364. ISSN: 0168-9525. DOI: 10.1016/S0168-9525(96)80019-9.

[149] L. Wolpert. "Positional information and the spatial pattern of cellular differentiation". In: *Journal of Theoretical Biology* 25.1 (Oct. 1969). Number: 1, pp. 1–47. ISSN: 0022-5193. DOI: 10.1016/S0022-5193(69)80016-0.

[150] Lewis Wolpert. "Positional information revisited". In: *Development* 107.Supplement (Apr. 1989). Number: Supplement, pp. 3–12. ISSN: 0950-1991. DOI: 10.1242/dev.107.Supplement.3.

[151] Wolfgang Driever and Christiane Nüsslein-Volhard. "A gradient of *bicoid* protein in Drosophila embryos". In: *Cell* 54.1 (July 1988), pp. 83–93. ISSN: 0092-8674. DOI: 10.1016/0092-8674(88)90182-1.

[152] Wolfgang Driever and Christiane Nüsslein-Volhard. "The bicoid protein is a positive regulator of hunchback transcription in the early Drosophila embryo". en. In: *Nature* 337.6203 (Jan. 1989). Publisher: Nature Publishing Group, pp. 138–143. ISSN: 1476-4687. DOI: 10.1038/337138a0.

[153] Carl Neumann and Stephen Cohen. "Problems and paradigms: Morphogens and pattern formation". en. In: *BioEssays* 19.8 (1997). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/bies.950190813, pp. 721–729. ISSN: 1521-1878. DOI: 10.1002/bies.950190813.

[154] Yu Chen and Alexander F. Schier. "The zebrafish Nodal signal Squint functions as a morphogen". en. In: *Nature* 411.6837 (May 2001). Publisher: Nature Publishing Group, pp. 607–610. ISSN: 1476-4687. DOI: 10.1038/35079121.

[155] Andrew P McMahon, Philip W Ingham, and Clifford J Tabin. "1 Developmental roles and clinical significance of Hedgehog signaling". In: *Current Topics in Developmental Biology*. Vol. 53. Academic Press, Jan. 2003, pp. 1–114. DOI: 10.1016/S0070-2153(03)53002-2.

[156] Rachel Kraut and Michael Levine. "Spatial regulation of the gap gene giant during Drosophila development". In: *Development* 111.2 (Feb. 1991), pp. 601–609. ISSN: 0950-1991. DOI: 10.1242/dev.111.2.601.

[157] Maria Capovilla, Elizabeth D. Eldon, and Vincenzo Pirrotta. "The giant gene of Drosophila encodes a b-ZIP DNA-binding protein that regulates the expression of other segmentation gap genes". In: *Development* 114.1 (Jan. 1992), pp. 99–112. ISSN: 0950-1991. DOI: 10.1242/dev.114.1.99.

[158] Rolando Rivera-Pomar et al. "Activation of posterior gap gene expression in the Drosophila blastoderm". en. In: *Nature* 376.6537 (July 1995). Publisher: Nature Publishing Group, pp. 253–256. ISSN: 1476-4687. DOI: 10.1038/376253a0.

[159] Johannes Jaeger, David Irons, and Nick Monk. "Regulative feedback in pattern formation: towards a general relativistic theory of positional information". In: *Development* 135.19 (Oct. 2008), pp. 3175–3183. ISSN: 0950-1991. DOI: 10.1242/dev.018697.

[160] Thomas Gregor et al. "Probing the Limits to Positional Information". In: *Cell* 130.1 (July 2007), pp. 153–164. ISSN: 0092-8674. DOI: 10.1016/j.cell.2007.05.025.

[161] Yoshihiro Morishita and Yoh Iwasa. "Accuracy of positional information provided by multiple morphogen gradients with correlated noise". In: *Physical Review E* 79.6 (June 2009). Publisher: American Physical Society, p. 061905. DOI: 10.1103/PhysRevE.79.061905.

[162] Yoshihiro Morishita and Yoh Iwasa. "Coding Design of Positional Information for Robust Morphogenesis". In: *Biophysical Journal* 101.10 (Nov. 2011), pp. 2324–2335. ISSN: 0006-3495. DOI: 10.1016/j.bpj.2011.09.048.

[163] Ken-ichi Hironaka and Yoshihiro Morishita. "Encoding and decoding of positional information in morphogen-dependent patterning". eng. In: *Current Opinion in Genetics & Development* 22.6 (Dec. 2012), pp. 553–561. ISSN: 1879-0380. DOI: 10.1016/j.gde.2012.10.002.

[164] Gašper Tkačik et al. "Positional Information, Positional Error, and Readout Precision in Morphogenesis: A Mathematical Framework". In: *Genetics* 199.1 (Jan. 2015), pp. 39–59. ISSN: 0016-6731. DOI: 10.1534/genetics.114.171850.

[165] Jeremy B. A. Green and James Sharpe. "Positional information and reaction-diffusion: two big ideas in developmental biology combine". In: *Development* 142.7 (Apr. 2015), pp. 1203–1211. ISSN: 0950-1991. DOI: 10.1242/dev.114991.

[166] Anaïs Bailles, Emily W. Gehrels, and Thomas Lecuit. "Mechanochemical Principles of Spatial and Temporal Patterns in Cells and Tissues". en. In: *Annual Review of Cell and Developmental Biology* 38.Volume 38, 2022 (Oct. 2022).

Publisher: Annual Reviews, pp. 321–347. ISSN: 1081-0706, 1530-8995. DOI: 10.1146/annurev-cellbio-120420-095337.

[167] Jan Rombouts, Jenna Elliott, and Anna Erzberger. "Forceful patterning: theoretical principles of mechanochemical pattern formation". In: *EMBO reports* 24.12 (Dec. 2023). Publisher: John Wiley & Sons, Ltd, e57739. ISSN: 1469-221X. DOI: 10.15252/embr.202357739.

[168] Shawn C. Little, Mikhail Tikhonov, and Thomas Gregor. "Precise Developmental Gene Expression Arises from Globally Stochastic Transcriptional Activity". In: *Cell* 154.4 (Aug. 2013), pp. 789–800. ISSN: 0092-8674. DOI: 10.1016/j.cell.2013.07.025.

[169] Benjamin Zoller, Shawn C. Little, and Thomas Gregor. "Diverse Spatial Expression Patterns Emerge from Unified Kinetics of Transcriptional Bursting". In: *Cell* 175.3 (Oct. 2018), 835–847.e25. ISSN: 0092-8674. DOI: 10.1016/j.cell.2018.09.056.

[170] Lauren McGough et al. "Finding the Last Bits of Positional Information". In: *PRX Life* 2.1 (Mar. 2024). Publisher: American Physical Society, p. 013016. DOI: 10.1103/PRXLife.2.013016.

[171] David B. Brückner and Gašper Tkačik. "Information content and optimization of self-organized developmental systems". In: *Proceedings of the National Academy of Sciences* 121.23 (June 2024). Publisher: Proceedings of the National Academy of Sciences, e2322326121. DOI: 10.1073/pnas.2322326121.

[172] Scott F. Gilbert and Scott F. Gilbert. *Developmental Biology*. 6th. Sinauer Associates, 2000. ISBN: 978-0-87893-243-6.

[173] Trudi Schüpbach and Eric Wieschaus. "Maternal-effect mutations altering the anterior-posterior pattern of the Drosophila embryo". en. In: *Roux's archives of developmental biology* 195.5 (July 1986), pp. 302–317. ISSN: 1432-041X. DOI: 10.1007/BF00376063.

[174] H. Spemann and O. Schotté. "Über xenoplastische Transplantation als Mittel zur Analyse der embryonalen Induktion". de. In: *Naturwissenschaften* 20.25 (June 1932), pp. 463–467. ISSN: 1432-1904. DOI: 10.1007/BF01504985.

[175] Michael Akam. "Making stripes inelegantly". en. In: *Nature* 341.6240 (Sept. 1989), pp. 282–283. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/341282a0.

[176] Michael Akam. "*Hox* and HOM: Homologous gene clusters in insects and vertebrates". In: *Cell* 57.3 (May 1989), pp. 347–349. ISSN: 0092-8674. DOI: 10.1016/0092-8674(89)90909-4.

[177] Peter A. Lawrence. *The making of a fly: the genetics of animal design*. Oxford [England] ; Cambridge, Mass., USA: Blackwell Science, 1992. ISBN: 978-0-632-03048-4.

[178] Stuart A. Newman and H. L. Frisch. "Dynamics of Skeletal Pattern Formation in Developing Chick Limb". In: *Science* 205.4407 (Aug. 1979). Publisher: American Association for the Advancement of Science, pp. 662–668. DOI: 10.1126/science.462174.

[179] J. Raspopovic et al. "Digit patterning is controlled by a Bmp-Sox9-Wnt Turing network modulated by morphogen gradients". In: *Science* 345.6196 (Aug. 2014). Publisher: American Association for the Advancement of Science, pp. 566–570. DOI: 10.1126/science.1252960.

[180] Alexander F. Schier. "Nodal Morphogens". In: *Cold Spring Harbor Perspectives in Biology* 1.5 (Nov. 2009), a003459. ISSN: 1943-0264. DOI: 10.1101/cshperspect.a003459.

[181] Patrick Müller et al. "Differential Diffusivity of Nodal and Lefty Underlies a Reaction-Diffusion Patterning System". In: *Science* 336.6082 (May 2012). Publisher: American Association for the Advancement of Science, pp. 721–724. DOI: 10.1126/science.1221920.

[182] Hidetaka Shiratori and Hiroshi Hamada. "The left-right axis in the mouse: from origin to morphology". In: *Development* 133.11 (June 2006), pp. 2095–2104. ISSN: 0950-1991. DOI: 10.1242/dev.02384.

[183] Yehuda Afek et al. "A Biological Solution to a Fundamental Distributed Computing Problem". In: *Science* 331.6014 (Jan. 2011). Publisher: American Association for the Advancement of Science, pp. 183–185. DOI: 10.1126/science.1193210.

[184] Patrick Hillenbrand, Ulrich Gerland, and Gašper Tkačik. "Beyond the French Flag Model: Exploiting Spatial and Gene Regulatory Interactions for Positional Information". en. In: *PLOS ONE* 11.9 (Sept. 2016). Publisher: Public Library of Science, e0163628. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0163628.

[185] Satosi Watanabe. "Information theoretical analysis of multivariate correlation". In: *IBM J. Res. Dev.* 4.1 (Jan. 1960), pp. 66–82. ISSN: 0018-8646. DOI: 10.1147/rd.41.0066.

[186] M. Studený and J. Vejnarová. "The Multiinformation Function as a Tool for Measuring Stochastic Dependence". In: Book Title: Learning in Graphical Models. Dordrecht: Springer Netherlands, 1998, pp. 261–297. DOI: 10.1007/978-94-011-5014-9_10.

[187] Rok Grah, Benjamin Zoller, and Gašper Tkačik. *Normative models of enhancer function*. en. Pages: 2020.04.08.029405 Section: New Results. Apr. 2020. DOI: 10.1101/2020.04.08.029405.

[188] Rok Grah, Benjamin Zoller, and Gašper Tkačik. "Nonequilibrium models of optimal enhancer function". In: *Proceedings of the National Academy of Sciences* 117.50 (Dec. 2020). Publisher: Proceedings of the National Academy of Sciences, pp. 31614–31622. DOI: 10.1073/pnas.2006731117.

[189] Benjamin Zoller, Thomas Gregor, and Gašper Tkačik. "Eukaryotic gene regulation at equilibrium, or non?" en. In: *Current Opinion in Systems Biology* 31 (Sept. 2022), p. 100435. ISSN: 24523100. DOI: 10.1016/j.coisb.2022.100435.

[190] Thomas R. Sokolowski et al. *Deriving a genetic regulatory network from an optimization principle*. arXiv:2302.05680 [physics, q-bio]. Feb. 2023. DOI: 10.48550/arXiv.2302.05680.

[191] Tamar Friedlander et al. "Intrinsic limits to gene regulation by global crosstalk". en. In: *Nature Communications* 7.1 (Aug. 2016). Publisher: Nature Publishing Group, p. 12307. ISSN: 2041-1723. DOI: 10.1038/ncomms12307.

[192] Tamar Friedlander et al. "Evolution of new regulatory functions on biophysically realistic fitness landscapes". en. In: *Nature Communications* 8.1 (Aug. 2017). Publisher: Nature Publishing Group, p. 216. ISSN: 2041-1723. DOI: 10.1038/s41467-017-00238-8.

[193] Claudia Igler et al. "Evolutionary potential of transcription factors for gene regulatory rewiring". en. In: *Nature Ecology & Evolution* 2.10 (Oct. 2018). Publisher: Nature Publishing Group, pp. 1633–1643. ISSN: 2397-334X. DOI: 10.1038/s41559-018-0651-y.

[194] I. Tomanek et al. "Gene amplification as a form of population-level gene expression regulation". en. In: *Nature Ecology & Evolution* 4.4 (Apr. 2020). Publisher: Nature Publishing Group, pp. 612–625. ISSN: 2397-334X. DOI: 10.1038/s41559-020-1132-7.

[195] Michele Gabriele et al. "Dynamics of CTCF- and cohesin-mediated chromatin looping revealed by live-cell imaging". In: *Science* 376.6592 (Apr. 2022). Publisher: American Association for the Advancement of Science, pp. 496–501. DOI: 10.1126/science.abn6583.

[196] David B. Brückner et al. "Stochastic motion and transcriptional dynamics of pairs of distal DNA loci on a compacted chromosome". In: *Science* 380.6652 (June 2023). Number: 6652 Publisher: American Association for the Advancement of Science, pp. 1357–1362. DOI: 10.1126/science.adf5568.

[197] Jessica Zuin et al. "Nonlinear control of transcription through enhancer–promoter interactions". en. In: *Nature* 604.7906 (Apr. 2022). Publisher: Nature Publishing Group, pp. 571–577. ISSN: 1476-4687. DOI: 10.1038/s41586-022-04570-y.

[198] Federica Ferretti et al. "Building General Langevin Models from Discrete Datasets". In: *Physical Review X* 10.3 (July 2020). Publisher: American Physical Society, p. 031018. DOI: 10.1103/PhysRevX.10.031018.

[199] David B. Brückner, Pierre Ronceray, and Chase P. Broedersz. "Inferring the Dynamics of Underdamped Stochastic Systems". In: *Physical Review Letters* 125.5 (July 2020). Publisher: American Physical Society, p. 058103. DOI: 10.1103/PhysRevLett.125.058103.

[200] Chong Wang and David M. Blei. *Variational Inference in Nonconjugate Models*. Issue: arXiv:1209.4360 arXiv:1209.4360 [stat]. Mar. 2013.

[201] Jorge Nocedal and Stephen J. Wright. *Numerical Optimization*. en. Springer Series in Operations Research and Financial Engineering. Springer New York, 2006. ISBN: 978-0-387-30303-1. DOI: 10.1007/978-0-387-40065-5.

[202] Neva C. Durand et al. "Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom". In: *Cell Systems* 3.1 (July 2016), pp. 99–101. ISSN: 2405-4712. DOI: 10.1016/j.cels.2015.07.012.

[203] *deeptools/pyBigWig*. original-date: 2015-07-31T12:07:36Z. Sept. 2024.

[204] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. en. The MIT Press, Nov. 2005. ISBN: 978-0-262-25683-4. DOI: 10.7551/mitpress/3206.001.0001.