Distribution Agreement

Signature:

Joseph L. Natale Date

Inference, Dynamics, and Coarse-Graining
of Large-Scale Biological Networks

By

Joseph L. Natale
Doctor of Philosophy

Physics

Ilya Nemenman, Ph.D.
Advisor

Gordon Berman, Ph.D.
Committee Member

Avani Gadani, Ph.D.
Committee Member

H. George E. Hentschel, Ph.D.
Committee Member

Daniel Weissman, Ph.D.
Committee Member

Accepted:

Lisa A. Tedesco, Ph.D.
Dean of the James T. Laney School of Graduate Studies

____
Date

Inference, Dynamics, and Coarse-Graining of Large-Scale Biological Networks

By

Joseph L. Natale
B.S., Stevens Institute of Technology, NJ, 2012
M.S., Stevens Institute of Technology, NJ, 2013

Advisor: Ilya Nemenman, Ph.D.

An abstract of
A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Physics
2020

Inference, Dynamics, and Coarse-Graining of Large-Scale Biological Networks

By Joseph L. Natale

Theoretical, experimental, and computational developments throughout the past three decades have rendered biological network modeling a powerful mainstay in the toolsets of physicists studying biology (and vice versa). Principal among experimental advancements are the multitude of so-called “-omics” techniques for gathering high-resolution, system-wide activity data at microscopic scales; on the computational side, they are the complementary abilities to manage and analyze far larger sets of data than ever before. The marriage of these endeavors, in the form of automated network inference, or “reverse-engineering”, has provided an unprecedentedly thorough characterization of small-scale biological systems, but remains costly and ill-equipped to predict the properties or behaviors of those same systems at larger scales. Here we review over two decades’ worth of work on network reconstruction, with an eye toward what new knowledge this exciting subfield has brought to modern biology, and then proceed to ask whether the typical products of the reverse-engineering endeavor might not be supplanted by more coarse-grained representations of biological data. Garnering inspiration from dynamical systems theory and statistical mechanics, we first study a random recurrent network model whose dynamics are amenable to a surprisingly compact description in terms of the system’s attractors. Then, following classical renormalization group methods in physics, we develop a general framework by which to pass from microscopic to macroscopic descriptions of a network even when the underlying interactions are not yet known. Our generic approach is able to extract appropriate large-scale degrees of freedom, and reproduce other previously established results, for a well-known system in physics. We describe an algorithm that can be applied directly to system-wide activity data, in the hope of obviating the need for explicit network inference as a preliminary step toward learning new biology.

Inference, Dynamics, and Coarse-Graining
# of Large-Scale Biological Networks

By

Joseph L. Natale
B.S., Stevens Institute of Technology, NJ, 2012
M.S., Stevens Institute of Technology, NJ, 2013

Advisor: Ilya Nemenman, Ph.D.

A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Physics
2020

Acknowledgments

Where do I start? There are two people I wouldn't be here without, who continue to make this all possible – and a third they sent along with me to college, despite never having attended themselves. *Grazie mille*, Mommy, "Sensei," and Alyssa.

Ilya, I have learned so much from you. It's hard to believe that our first meeting was seven full years ago – I knew then you'd be my advisor (our shared preference for espresso and proclivity for talking science on brisk walks were dead giveaways).

I do hope our adventure will continue.

To the other professors of my Committee... Hentschel, you've taught me so much about the art of precision in my thinking, and the value of never forgetting first principles. I will be forever indebted (and if you move to Umbria, do invite me for an espresso). Wildani, it was always a pleasure to brainstorm with you; I know we will explore more exciting avenues together in the future, and I am as flattered as I am grateful for your unwavering belief in my abilities. Weissman, you are a role model when it comes to the "human" side of science: I have always admired your demeanor when talking deep physics, how you stay rooted and candid when the basin of attraction for the "mad scientist" state is so large and inviting. Berman, you are so willing to share ideas that your enthusiasm for what you do is contagious; it has not gone unnoticed, and will benefit your students for years to come.

Vijay: How can I possibly thank you for all your years of guidance and friendship? All I can really say is ... Now, the Boys in Blue are both doctors.

Postdocs second, friends and companions first – you guys rock! Andrew, your technical repertoire and willingness to engage on any topic are astounding. I am so fortunate to have worked with you, and i hope we meet again soon. David, *compà* – I'll never forget our journey – somehow we moved house across the complex together, but never ended up in Italy at the same time. I still appreciate how our approach to

science is so unique. Damián, from coauthorship to our trip to watch a solar eclipse, I've been captivated by the way your mind works. Let's reunite soon! Itai, thank you for always being willing to talk neuroscience with me, and thank you for always swapping stories about our canines. Ze'eva has no idea what a great man helped me decide on (and spell) her name. Audrey, your poise and precise delivery in talking physics – from one-on-one discussions to formal presentations – is an inspiration. But above all, thank you for being my "thesis mom" and urging me to continue writing whenever unignorably interesting things were happening in the vicinity. Michael, you have the kindest heart, period. Your humility and prowess are undeniable, and I have been so very honored to begin our collaboration.

Other members of the group during my time – George, Xinxian, Martin, Caroline, Mia, Baohua, Ahmed, Maha, and anybody I've missed – I'm not sure there is anything quite like sharing a lab together. I thank you all for your support and wish you all the best going forward. Catalina, we came in together and now we're both nearing the end of our tenure. You have been such an important figure not only for my experiences in the Nemenman Lab, but my time at Emory. Exchanging graduate school woes, sharing lunches at Zoe's ... and further Italianizing my New Jersey accent so you could understand me better when we first met are going to be some of my most cherished memories. And yes, I forgive you for assuming in those first few months that I didn't *really* speak Italian.

Ben – what can I say to my "brother from another mother?" Congratulations on building not just an excellent resume at Emory, but a new career and family. I will be there to see you and Junior on the West Coast soon.

Michael, our "professional friendship" goes far – literally, as I'll never forget my surprise the day you picked me up for a bite to eat, all the way in NJ, while I was ill. Congratulations on the defense date!

Ryan, you too – congrats on entering the real world. It is unthinkable that we got to know each other before this all truly began. I will never tire of talking science, philosophy, or Dragonball Z with you. Some of our greatest hits were during our drives home on the first six days of school... when we found five unique routes to ger there (welcome to Atlanta!). Xinru, all the same for you – except perhaps the DBZ.

Other Emory people – Jason, Barbara, Susan, and Calvin, Justin (I cannot imagine what it would have been like teaching for anyone else, if for no other reason than the fact that your quick wit takes the cake), Skanda, Fereydoon (what great discussions about statistical physics!), Stefan (your dedication is unrivaled) and others within the physics department; Sensei Ikeda and the Shotokan karate club; the entire Emory Weightlifting team (I am so grateful for your ever-welcome, open invitation to come lift even today; if it were not for your urging, I may never have entered my first competition!); Simona, Angela, and the Italians at Emory; and my friend Elena, who helped me through my European homesickness for the "home I never had" (thanks, Niko) by indulging my efforts to bring our culture to campus – paragraphs can be written about your tenacious support. I thank you all.

To Professors Harry Lenzing, Ting Yu, Rainer Martini, and Chris Search of Stevens Institute of Technology – your encouragement and investment in both my scientific career and personal development I simply could not have gone without. Diane Gioia, ditto! Lisa Dolling – thanks to you, to this day I do not know whether I am a scientist or a philosopher. Garry Dobbins, thanks for believing in Joseph Natale. Susan Schept, thanks for reminding me to believe, period.

David Nicholson and Katasaur, High Rollers, Deep in Learning – no words.

Lysandra... I may not have made it, if not for your almost nightly provision of snacks as I neared the end of this phase. Your cookies kept me alive. Thank you.

Ricky – "*Revocate animos, maestumque timorem mittite: forsan et haec olim meminisse iuvabit.*" Yes, it is done. Best of all, our Theory worked, my friend.

Contents

List of Figures

List of Tables

List of Algorithms

Chapter 1

Old Problem, New Solutions; New Problems, Old Solution

Like all models, biological networks [1, 2] are abstractions. Built to capture select aspects of a system's properties or behavior, the most fitting measure of their success is not whether they are "correct," but whether they are useful [3, 4].

Like other decorated scientific frameworks [5, 6], biological networks have proven useful in many of the roles for which they were intended. Like many others [7, 8, 9], they have often proved even more successful in serving functions well outside their originally intended purposes. Yet network representations of living systems demand attention not due to their similitudes with other models, but because they are unique.

For instance, biological networks are children of the twenty-first century: only in the presence of modern experimental technology can one sustain ambitions of building comprehensive interaction maps for processes on molecular, cellular, organismal, and population levels. Moreover, such models are not mere databases; over the past few years, the systems biology community has figured out how to analyze their structures to enable powerful predictions about the underlying systems, sometimes with clinical relevance (as in the emerging subfield of network medicine [10]).

Of particular interest to the theoretical physics community is the possibility of using biological networks to make generalizable statements about the way information is stored and transmitted among the components of living systems. Such problems, poised at the multitudinous interfaces between physics and biology, also echo a deeper relationship between these two disciplines: while at present it might seem hyperbolic to parallel the distillation of Hooke's Law, or even Newtonian gravitational theory, with recent developments in systems biology, the refinement of an ability to quantify regularities from within the contemporary deluge of complex, seemingly disparate data is the precise object of pursuit in certain cross-disciplinary endeavors [11]. Some prominent figures in ongoing interdisciplinary efforts express hope that their work will unveil quantitative *theories*, or even universal *laws* for the biological sciences [12].

Theoretical physics and biology do share a history of distinguished collaborative success, including three joint Nobel awards in "Physiology or Medicine" [13, 14, 15]. Yet in utilizing large-scale network models to glean glimpses of how biological theories might look, much remains to be learned by both camps. The physicist, skilled with his array of mathematical hammers, has an onus to be cautious in pounding "nails" when exploring the myriad (ostensibly, unique [16]?) complexities of life. Likewise, specialists working in particular areas of biology must consider general patterns even where it requires them to loose some of the details that set their systems apart.

With time, the lines dividing the traditional experience of both parties continue to blur. The newer concerns, addressed by all parties, are about how to best describe or treat the system at hand; the range of applicability of the suite of tools available; and what new tools must be developed to answer previously unanswerable questions.

What specific kinds of difficulties does the study of large-scale networks bring? First, there is the problem of constructing or *inferring* an appropriate network model in the first place – and, having done so, knowing both how to interpret such a large-dimensional object and what it will be used for. In general, whole-network inference is also computationally costly; it is not always clear that reverse-engineering, for all the effort it entails, is necessarily the best approach for certain categories of problems.

A subsequent difficulty is that, while biological networks necessarily describe interactions on particular (typically, small) scale, the predictions we (as a community) are most interested in are often aspects of the underlying system's gross behavior – the output of some circuit, a decision which is reached by an animal – that manifest only at larger scales, or even span multiple length scales [17]. Such "macroscopic" properties and behaviors are often simpler, in the sense of affording lower-dimensional descriptions, than the "microscopic" interactions which comprise the individual-component, network-level account of the system... yet it is hard to tell when this is the case, and no general method exists by which to ascertain whether an inferred network might might encode low-dimensional collective behaviors or emergent properties.

Finally, there is the more difficult issue of bridging the gap *between* observation scales: even for systems which are known to admit "simple" macroscopic descriptions, how can we find an appropriate set of variables or features by which to summarize the large-scale behaviors from knowledge only of the microscopic details? Can (and should) large, complex network models themselves be systematically *coarse-grained*, as is sometimes done for statistical models in physics? This topic is – and promises to remain – of major interest in the machine learning community. Unsurprisingly, relationships between fundamental aspects of feature learning and model reduction in machine learning and their counterpart endeavors in theoretical physics are more than skin-deep [18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30]. Still, subtle differences in the guiding philosophies on each side prevent the naïve application or exchange of methods. While parallel developments may yet converge, another possibility is that some biological systems can be assigned simple models without ever having to infer full-scale networks in the first place. Can one adapt the abstraction level for biological models from the inception to match the intrinsic complexity of the data at hand [11]?

Thesis Statement and Summary of Dissertation Contents

> *Academic disciplines [...] define themselves either by their objects of study or by their style of inquiry. Physics [...] is firmly in the second camp* [12].

This Dissertation is an attempt to rethink aspects of the contemporary approach to "Big Data" biology, from the perspective of a theoretical physicist. Its core question is: can we divine simple descriptions of would-be complex network models in biology, without having to first understand all their constituent details? Or, do there exist coarse summaries, some combinations of these constituents, that serve better than the traditional fundaments for making certain types of predictions about living systems?

The following three Chapters comprise three interrelated projects, responding in turn to each of the broad challenges mentioned above – inferring complex network structures, approximating their dynamics with simple descriptors, and compressing biological data to obviate the need for full-scale inference – in the context of concrete physical or biological problems. In each, I emphasize a different way of thinking about biological networks, and tackle a different problem-solving scenario by borrowing tools and ideas from theoretical physics and related disciplines (these are, predominantly, statistical mechanics, information theory, and machine learning). I have focused primarily on problems in neuroscience, where increasingly large portions of *in vivo* cellular networks can be monitored [31]. This has prompted many scientists to try their hands at mapping large circuits in excruciating detail [32, 33, 34] and (somewhat ironically) a few others – myself included – to inquire whether single-cell activities are in fact the most logical choice for describing the rich functionalities of brains.

Our treatment begins by reviewing a resounding success. The technologies that have made network reconstruction possible have solved an old, outstanding problem: we can now measure and monitor thousands of biological species, simultaneously.

In Chapter 2, we survey the broad classes of modern, activity data-driven methods for learning the network structures of interactions between the (typically, microscopic) biological elements – who talks to whom, and who controls whom? Algorithms with a high degree of automation have generated significant excitement in all the so-called "-omics" subdisciplines, where reconstruction approaches often share similarities with previously established protocols, like system identification methods in engineering.

While the typical products or outputs of network inference tasks vary according to the computational equipment used and the prediction types intended, an astounding repertoire of data types, methodologies, and obstacles encountered by practitioners (both conceptual and practical) is highly conserved throughout fields. Thus, once we understand how relationships can be reliably inferred from data, we critically examine the progress and goals of this burgeoning, cross-concentration endeavor, grouping all examples collectively under the popular, umbrella term *reverse-engineering.*

A recurring theme within Chapter 2, both explicitly in the discussion of particular algorithms and implicitly in the general philosophy of biological network inference, is the notion of *locality.* Physicists will recognize this storied concept as one of their most basic assumptions about Nature. Sometimes locality assumes a concrete guise, as in the inference of site bonds between nearby amino acids from observable correlations between variable parts of the sequence that ultimately code for a particular protein shape. At other times it enters more abstractly. For example, one can consider interactions which are "local" in the sense that they couple only a handful of elements, regardless of distance in real (or any metric) space. For example, a certain gene might interact with only tens of others (out of tens of thousands of possible interactions). The locality of interactions recurs as an important theme throughout the Dissertation, playing a major role in Chapter 3 and forming the basis for discussion in Chapter 4.

The critical analysis in Chapter 2 concludes with the distinct assertion that the increasingly powerful network inference algorithms we see becoming more widespread in certain fields [35] are not always the best tools even for their own jobs. Where intuition or transparency about the motion of a system as a whole is required, whole-network reconstructions can become unwieldy hindrances that hamper progress. To replace full-scale reverse-engineering for such applications, we call for new approaches that are able to detect the functional units, or moving parts, that determine a system's behavior at large scales. Ideally, we would do this without partitioning or coarse-graining inferred network models directly, since (to paraphrase Vapnik in *The nature of statistical learning theory*), it makes little sense to solve this "hard" problem only to simplify it for the purpose of solving an "easy" one [36].

Thus to progress in our central problem of moving from network microscopy to macroscopic dynamics, our new goal will be to ascertain system's gross functionality without having to infer a network model "from scratch" in the first place. Two ways in which this might be done are 1) finding approximate, low-dimensional descriptions of a system's behavior in terms of its dynamical attractors and 2) coarse-graining data directly, prior to constructing a network representation, to infer an *effective model* at the desired scale of inquiry. These are the bases of Chapters 3 and 4, respectively.

In Chapter 3, we study a system whose constituent interactions are in no obvious way reducible, but whose dynamics across large scales may be nonetheless cast in a straightforward form involving groups of simultaneously active system components. Specifically, we examine a simple network model describing a neural system, comprised of synapses whose strengths are random and an overall firing activity that is regulated by global inhibition. This model is spatially extended, and its excitation modes have a natural geometrical interpretation that may be of interest in modeling neural aspects of spatial navigation in animals. When stimulated locally, the system attracts to one of a discrete set of constant-firing states that can be predicted with high accuracy from knowledge of the spatial region in which stimulation occurred.

Locality enters in two ways in Chapter 3. First, synaptic efficacies, representing strengths for the interactions between our model neurons, are chosen to be short-ranged and therefore spatially “local.” Second, our low-dimensional characterization of the system’s dynamical behavior – a simple decoder whose input-output relations realize one way of storing spatial memories – turns out to involve the participation of only a small subset of a system’s components for each attracting state. In particular, locating (or *decoding*) the region of stimulation entails tracking the firing rates of just a few highly active neurons, without careful calculation of their real-valued rates.

Finally, Chapter 4 addresses the problem of constructing simple, large-scale representations for complex systems in a more generic and formal environment. Inspired by *renormalization group* approaches from statistical physics, we develop an automated coarse-graining algorithm for learning effective summaries of “Big” activity data that do not depend on prior knowledge of the system’s topological network structure. Our framework generalizes the familiar real-space renormalization procedures, recovering a system’s behavior at large scales without any need to specify its microscopic dynamics in advance, nor even the subsets of activity variables which are are sufficiently “local” to engage in physical interactions. We test this novel method on a 2D Ising model, for which both the microscopic interactions and macroscopic degrees of freedom are known, and find that it reproduces key aspects of the established RG flow. Thus, at the climax, we propose to address the relatively new problem of how to sift through more data and more network complexity than ever in biology by reviving and generalizing into compatible form a classic, “old hat” methodology from physics.

Since the material in Chapter 2 is introductory, serving as the foundation for the subsequent projects, its length is greater (and its tone less technical) than the sequels. This academic review-style commentary was also published, with minimal differences, as a chapter in a recent textbook on Quantitative Biology [37]. Accompanying the main text (here included as an Apppendix) was a miniature, project-oriented problem that was developed and written in conjunction with my coauthors, D. Hofmann and D. G. Hernández, under the supervision of my Ph.D. Advisor, Ilya Nemenman.

Chapter 2

Reverse-Engineering Biological Networks from Large Data Sets

This contents of this Chapter initially appeared as an invited contribution to the text "Quantitative Biology: Theory, Computational Methods and Examples of Models" (edited by B. Munsky, W. S. Hlavacek, and L. S. Tsimring, MIT Press, 2018). *It was published there as Chapter 11, with my coauthors David Hofmann, Damián G. Hernández and Ilya Nemenman, under the same title; electronic pre-prints are also accessible as* Natale, Joseph L., et al. "Reverse-engineering biological networks from large data sets." *See* arXiv:1705.06370 (2017) *and bioRxiv* (DOI 10.1101/142034).

I was directly responsible for preliminary research on over 200 reverse-engineering algorithms, creating captioned figures, the structural organization of the sections, and delegating sections to my coauthors. Sections 2.1 (with I.N. contributing heavily to subsection 2.1.1 and 2.1.2), 2.2, the introductory paragraphs of Section 2.3, and 2.4 were my own; I established the vision and contents (including early drafts) for – and contributed editorially to the final drafts of – Sections 2.3.1 (D.G.H.) and 2.3.2 (D.H).

Much of contemporary systems biology owes its success to the abstraction of a network, the idea that diverse kinds of molecular, cellular, and organismal species and interactions can be modeled as relational nodes and edges in a graph of dependencies. Since the advent of high-throughput data acquisition technologies in fields such as genomics, metabolomics, and neuroscience, the automated inference and reconstruction of such interaction networks directly from large sets of activation data, commonly known as reverse-engineering, has become a routine procedure.

Whereas early attempts at network reverse-engineering focused predominantly on producing maps of system architectures with minimal predictive modeling, reconstructions now play instrumental roles in answering questions about the statistics and dynamics of the underlying systems they represent. Many of these predictions have clinical relevance, suggesting novel paradigms for drug discovery and disease treatment. While existing review articles have focused their attention predominantly on the implementation details and effectiveness associated with individual network inference algorithms, here we examine the emerging field as a whole.

We first summarize several key application areas in which inferred networks have made successful predictions. We then define and delineate the major classes of reverse-engineering methodologies, emphasizing that the type of prediction that one aims to make dictates the algorithms one should employ. We conclude by discussing whether recent breakthroughs justify the computational costs of large-scale reverse-engineering sufficiently to admit it as a mainstay in the quantitative analysis of living systems.

2.1 Lay of the land

Biological systems on all levels of organization, from cells to brains and to populations, are comprised of ensembles of interactions among smaller constitutive components [38, 39, 40]. These interactions are typically very specific, and highly coordinated spatially and temporally [41, 42, 43, 44, 45]. Involving not just pairs, but also larger groups of components acting in concert [46, 47, 48, 49, 50, 51], they are responsible for the rich diversity of complex phenomena and behaviors that make living systems work. Although often prohibitively numerous to model individually (though see [52]), these components and their corresponding interactions can be represented formally as graphs [53], known colloquially as *biological networks* [54, 55, 1, 56, 57, 58, 40, 59].

The variables in such networks (also called nodes) typically represent biochemical or ecological species, cells, or even amino acid residues when one is interested in the biophysics of proteins. The links among the nodes represent interactions, such as chemical transformations, catalysis, and binding; cooperative or predator-prey relations among species; electrical and chemical communication among cells; or geometric proximity among amino acid residues (Fig. 2.1).

To answer many questions in modern data-rich biology, an intermediate step often involves the reconstruction of such networks from empirical data. The data typically consist of joint samples of activities (often referred to as expressions, frequencies, abundances, or population sizes, depending on the context) of a large number of components measured in different biological contexts. Problems of this kind pervade the quantitative life sciences on all physical scales, even if they take different forms and use different languages across scientific disciplines.

At the smallest scale is the problem of inference of physical contacts for amino acids in a protein fold [61, 62, 63], which is a network representation of the 3D structure of the protein. Predicting such networks from the co-occurence of amino acids promises the ability to design proteins with specific functional properties. At the cellular level, different genes activate or suppress the activities of other genes, forming networks of genetic regulatory interactions [64, 65]. Similarly, metabolites transform into each other, catalyzed by various enzymes; these form metabolic networks [58, 66, 67], as well as networks that combine both protein and metabolic modalities. Protein signal-

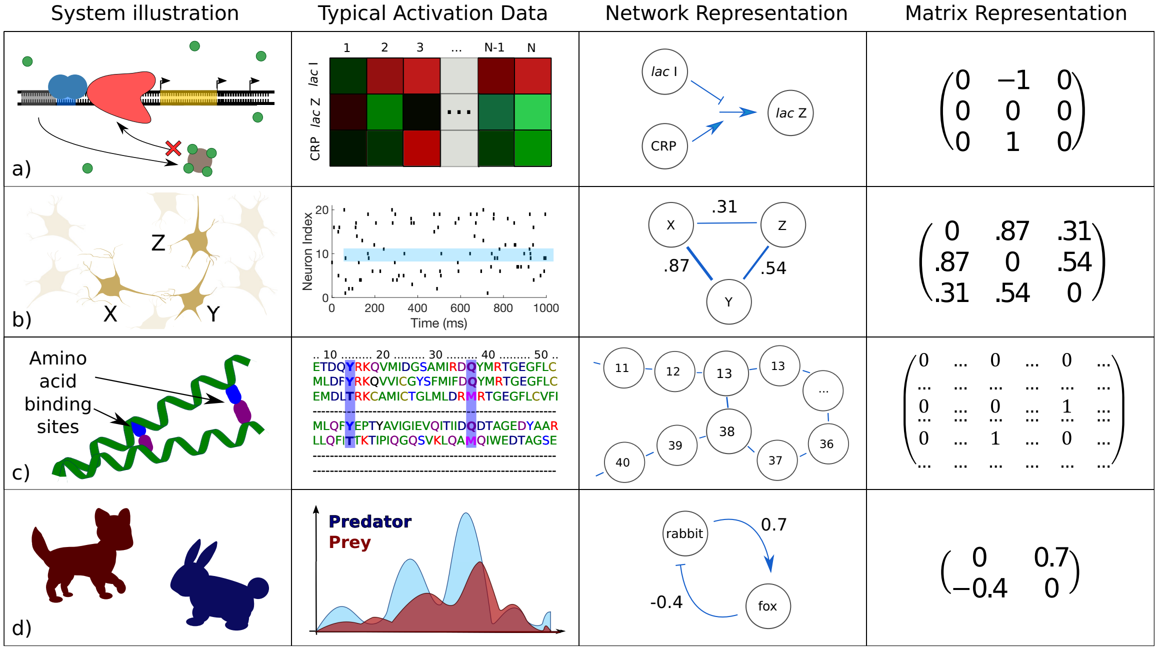

Figure 2.1: Examples of biological systems whose constituent interactions have been modeled using networks. (a) The regulation of gene transcription by transcription factors and other enzymes. For example, in the classic *lac* repressor circuit [60], when lactose concentration is high and glucose concentration is low, genes for metabolizing lactose are strongly activated. Activities in this case might consist of microarray data for each mRNA species; the network shows the logic (AND) of the system. (b) Neuronal co-activation networks can be measured by computing correlations between spike patterns. In this case, the network graphs are weighted, and weights may represent correlations. (c) Spatially proximal amino acids tend to co-evolve, as they often participate in bonds that are vital to the structure and function of the protein they form. Here activity values are discrete, assuming one of 20 values to identify the amino acid at each site; the network represents bonds that are inferred to exist by noting which site pairs are highly correlated across similar proteins in different organisms. (d) Complex predator-prey interaction dynamics can be cast in a network form as well. Here activation data represent the populations of each species, and connections are labeled with inferred parameter values for the governing population dynamics equations.

ing networks characterize the structure of decision-making and information processing in individual cells [68, 69, 70, 71, 72]. The accurate reconstruction of different types of these cellular networks is expected to lead to successful interventions that cure some of the most debilitating diseases [73].

On the scale of the nervous system, one often reverse-engineers neural circuits [74, 75, 76, 77, 32] and, on a larger scale, functional connectivity networks between brain regions [78, 79, 80, 81, 82]. The structure of the latter has been shown to be valuable as a diagnostic tool for some psychiatric diseases [83], and there is mounting evidence that the former can be "reprogrammed" via external interventions to repair damaged circuits [84]. Finally, on the largest scales, one can reconstruct networks of interactions among members of a particular species [85, 86, 87], or different species in an ecosystem [88, 89, 90, 91, 92, 93]. This knowledge may help in forecasting ecological catastrophes [94, 95] and addressing the spread of infectious disease [**?**] (or other epidemics [96]).

In all of these fields, data share similar properties, and data sets often have similar sizes. This imposes uniformity not only on the question of network inference itself, but also on the obstacles and algorithmic approaches that underlie reconstruction efforts across multiple biological domains. Inference methods designed for one system type ([97], [98], and [99]) can often be adapted to accommodate others ([100, 101, 102], [103, 68, 104], and [75, 105], respectively). Moreover, morally equivalent methods have been developed in nominally unrelated fields [106] – or else borrowed explicitly from established disciplines, such as systems identification techniques migrating to network biology from engineering [107, 108]).

An additional reason for the cross-pollination among the subfields of biological networks inference is that, like in other parts of bioinformatics, the field has benefited from advances in machine learning and related Big Data computational tools. In their turn, as is often true of mathematical approaches, these tools are applicable across multiple traditional biological subdisciplines, and hence provide for natural theoretical bridges not only among life-sciences subfields, but also to a "network" of other quantitative disciplines (physics, statistics, and computer science) [109].

However, one cannot embrace the unembraceable. Thus in this review, we will focus almost exclusively on applications of networks inference to the systems biology of the cell [110, 111, 112], and will mention bridges to other fields only briefly and haphazardly, leaving the reader certainly thirsty for more. Starting with a few of the references that we mention, as well as using Google Scholar (another network, this time of citations), is an easy way to quench this thirst!

Before proceeding any further, it is certainly worth warning the reader that the explosive growth of the field of biological network inference has covered with a thick blanket of journal articles some treacherous rocks. A few of them are very dangerous, and can, in principle, sink the field if not addressed thoughtfully. Specifically, while fully automated network inference has become a routine procedure, it is not immediately clear that the large-scale reconstruction of entire networks from high-throughput data will necessarily result in tangible insights or actionable understanding about biological systems. One reason is that most reconstructions are not experimentally verified, remaining in the literature as collections of information (or misinformation) of dubious quality. Another comes from the fact that it is still not clear what new knowledge entire-network inference yields, besides proposing potential interactions for experimental verification. If a goal of the field is to predict response of biological systems to yet-unseen exogenous perturbations, then the bridge between a network graph and such predictive knowledge will have to be built eventually, but it is not there yet in most practical applications. Most importantly, it is usually unclear what insights are delivered by large-scale networks, or how to interpret the typical product of the reconstruction enterprise – Lander's infamous "hairball" of decontextualized interactions [113]. One can even argue that exhaustively enumerating interactions is not inherently more insightful than cataloging the original experimental data, and both should give way to studying the system's emergent properties [114]. Having now warned the reader, we leave these important, foundational questions aside for the remainder of the review (save the *Discussion*).

2.1.1 Scale of the biological network inference problem

Network reverse-engineering is typically done in the "low-hanging fruit," Big Data *regime. Here the data sets are large, but the number of unknowns is even larger: not all the unknowns can be learned reliably.*

While reconstructions can be performed using different data types [115, 110], here we are concerned with approaches that are based exclusively on biological *activity measurements*. Suppose we have a network consisting of p nodes (e. g., a group of p interacting genes or neurons), and n simultaneous measurements of some activity variable for each of these nodes (which for our purposes fully characterizes the biological states of the nodes at a given moment in time). The activity variables can be binary (as in the characterization of whether a gene is on or off, or whether a neuron is spiking or not at a given time) or real-valued (gene expression levels, or firing rates for neurons). In other words, the total amount of available data is $\sim np$. The goal is to identify links between pairs of the p nodes (or more generally, higher order interaction structures) from patterns in their activities. If we focus on pairwise interactions among the nodes only, then the number of unknowns is $\sim p^2$. Thus the amount of data per unknown is $\alpha \sim np/p^2 = n/p$.

In the classical statistics regime, the amount of data is typically asymptotically large compared to the number of unknowns, $\alpha \gg 1$. In contrast, network inference usually proceeds in the regime where $p \gg 1$, with typical $p \sim 10^2 \dots 10^3$. For gene expression and other high throughput cellular data, in particular, it is not uncommon to have $p \sim 10^4$. Other fields are catching up [116, 117]. The number of measurements is also typically large, $n \gg 1$. We can consider $n < p$, as in most genetic data, or $n > p$ (but not $n \gg p$), as in many neuroscience applications. More generally, $n \sim p$, so that $\alpha \sim 1$, representing a qualitative departure from the classical statistics regime.

The situation gets even worse when we remember that the total number of parameters characterizing all (higher-order) interactions in a network scales as the total number of states that the network can be in (i. e., 2^p for binary nodes, or 2^{pS} for continuous ones, where S is the entropy of each node measured at the experimental resolution). Thus in the most general case, for biological network inference, $\alpha \ll 1$. It is clear then that, just like in most other Big Data applications, the problem cannot be solved completely, with all interactions identified. Thus networks inference necessarily is a "low-hanging fruit" problem, where the limited data allows us to focus only on the most salient features of the studied systems. This also means that, in any quantitative assessment of the quality of network reconstruction methods, we should focus a lot more on the precision (absence of false positives) of a method, rather than on its sensitivity (absence of false negatives), since the sensitivity of essentially any method on realistic data would be tiny.

2.1.2 Different ideologies for inference

In biological network inference, one can think of reconstructing actual physical interactions among the nodes or coarse-grained, phenomenological surrogates. We focus exclusively on the latter.

The notion of network inference may evoke the idea of reconstructing actual physical interactions among network nodes. For example, a regulatory interaction between two genes might mean the direct binding of a transcription factor protein, translated from one of these genes, to a specific part of the DNA sequence that controls the expression of the other gene [101]. We refer to the reconstruction of such physical, microscopically accurate interactions as the inference of *mechanistic* networks. In contrast, the majority of reconstruction methods focus (explicitly or not) on the inference of *effective* interaction networks [118], which keep track of purely phenomenological interactions. These may or may not be mechanistically accurate, but are sufficient to reproduce various statistics of the observed variables. Such effective interactions may correspond to subsets of the interactions in mechanistic networks. They may be compact, coarse-grained averages of some microscopic quantities. Or they may be entirely macroscopic properties that have remote and complicated relationships with the microscopic, mechanistic interactions.

One can focus on effective network inference for purely pragmatic reasons: as discussed above, even high-throughput data is insufficient to infer *all* the contributing actors in a complex system, and effective interactions may simply be the low-hanging, accessible fruit. In contrast to this pessimistic view, one may argue that every level of description requires its own proper degrees of freedom for efficient representation [114, 119, 11], and that the distinction between mechanistic and effective networks is not that clear-cut.

To wit, even mechanistic biophysical interactions are themselves effective interactions, just at a different scale. For example, the bonds between amino acids that form at protein-protein interfaces consist of electrostatic forces between constituent molecules. These forces can then be broken down in terms of quantum interactions between elementary particles, at which point the notion of an amino acid has long since disappeared. Likewise, the fact that communication between synapsing neurons requires the diffusion of neurotransmitters across the synapse undermines the notion that neurons can ever truly be in a direct, mechanistic contact. We are sympathetic with this viewpoint, which treats the distinction between mechanistic and effective networks less as a dichotomy than as a spectrum. In what follows, we cast the issue in terms of modeling assumptions: what is the appropriate set of nodes and interactions to answer *the specific questions being asked* while working at the *desired scale*?

Our perspective notwithstanding, a few authors have distinguished explicitly between these two ideologies (see [120] as the originator of the “physical” vs. “influence” network terminology, and [81] for a more fine-grained distinction among different types of effective networks in the brain). Many other sources refrain from making such explicit distinctions, presumably either for expedience in exposition or because they take seriously the aforementioned notion of pursuing the most efficient or useful description at a given level of study, regardless of the biological implementation details at other levels. While we remain agnostic to the particular reasons for the tendency of reverse-engineering literature to avoid making this distinction at the outset, we lament the absence of explicit declarations of the intended level of description when elaborating a new algorithm by the majority of publications. By default, in this Chapter, we focus on effective inference methods, for which authors do *not* make an effort to understand whether there is a mechanistic basis for inferred interactions, stating any exceptions at the outset when they appear.

2.1.3 Goals of this Chapter

We are now in a position to state our intended goals for this Chapter. In the following sections, we review relatively recent (within the last two decades) attempts at network inference, contending:

1. The aptness and success of a given inference method depend on the ultimate purpose of performing network reconstruction. One must first establish what kinds of predictions are desired (i. e., what does one seek to *learn* [1] using the network?), and only then decide which algorithm to use.

2. Large-scale network reverse-engineering has many fruitful applications, but it is not always the necessary – or not necessarily the best – approach for making certain kinds of predictions.

Note that we deal exclusively with inference methods that produce networks containing at most pairwise interactions. While the joint probability distribution for p discrete or continuous stochastic activation variables in a stationary state $\{g_i\}$ can be expanded [99] most generally as

$$P(\{g_i\}) \propto \exp\left[-\sum_{i}^{p} h_i(g_i) - \sum_{i,j}^{p} J_{ij}(g_i, g_j) - \sum_{i,j,k}^{p} \phi_{ijk}(g_i, g_j, g_k) - ...\right], \tag{2.1}$$

where functions h_i, J_{ij}, and ϕ_{ijk} denote first-, second-, and third-order interactions, respectively, it is clear from the considerations of Section 2.1.1 that reliable estimation for terms of higher order than J_{ij} is prohibitively difficult. In addition, we review only the algorithms that attempt to infer *static* values for J_{ij} under the assumption that the system is in (near-)stationary conditions, although some authors have attempted to estimate networks whose topologies are dynamically evolving [121, 122].

The progression of the Chapter is as follows. First, we examine highlights of the many places where network inference has been used to advance new knowledge in contemporary systems biology and establish novel paradigms in modern medicine. Then we proceed to delineate and explicate several types of inference methods, briefly describing the operation of several representative algorithms for each of the classes we name. We conclude with a brief outlook of where the field might be headed. However, these concluding comments should be taken with a lot of caution, since “it is difficult to make predictions, especially about the future.”

2.2 Roles for reverse-engineering in systems biology research

The reverse-engineering of large-scale networks by means of automated algorithms has become such a routine procedure that it has spawned a research field of its own. Why is the task of learning networks from data considered so important?

The modern imperative to generate comprehensive parts lists for large biological systems [58] is epitomized in what one author somewhat flippantly calls “the giant maps of metabolic pathways that many molecular biologists pin to their walls” [123]. Such diagrams encode and illustrate visually the entirety of observable interactions of a particular type in a specific system. Since the mid-2000s, attempts to generate such maps have been pursued vigorously by researchers in various disciplines, but the most prominent and systematic efforts have come from the network inference Challenges of the Dialogue on Reverse-Engineering Assessment and Methods (DREAM) initiative [124, 35]. Contestants participating in these ongoing Challenges submit network reconstructions, inferred by original algorithms operating on standardized data, for comparison against (experimentally) established sets of interactions in benchmark networks.

The top-scoring networks in early competitions achieved respectable accuracies, despite the difficulties associated with defining "gold standard" benchmarks and evaluation metrics [124, 125]. However, they also lacked the ability to provide intuition (beyond structural insights) about the systems they described. As static pictures of interaction architectures, they had limited ability to *predict* a system's behavior. The pattern of assembling a large, intricate network as the end goal, with no intention to use it as a tool for prediction – as in the iconic but largely uninformative hairball of Ref. [113] – thematized DREAM competitions roughly until 2014, nearly a decade after one reviewer declared the field to be "still in [its] 'natural history' phase" [39].

The emphasis of DREAM competitions has since shifted, mirroring changes in the attitude of the reverse-engineering community as a whole. Recent competitions have more strongly favored *predictive modeling*, with inferred networks serving not as ends in themselves, but as coarse summaries of high-dimensional data – a special type of statistic – to aid in projecting how the behavior or components of a system will change (as a function of time, due to changes in its environment, etc.).

This movement away from using learned topologies as a signal that the "work is done," and instead toward viewing the entire process of network inference as an intermediate step in an fully-fledged research pipeline [126], is also supported by theoretical work. In particular, it has been argued that structure alone provides insufficient information to achieve an adequate degree of control over the underlying system's dynamics [127]. In fact, the object of interest is not always a network's structural complexity (density of connections), but its *dynamical* complexity (the number of fixed points it can accommodate), which depends on other parameters beyond structure, such as its connection strengths. Indeed, only the latter is closely tied to the viability of a network architecture in the context of evolution [128].

The field's transition – from descriptive to predictive – is a natural one, and indeed reminiscent of the progression in other branches of science. While it is not completely clear why there was this prolonged period of exploration without modeling, it is plausible that reverse-engineers first needed to convince themselves that (1) networks can, indeed, be accurately reconstructed from activity data alone, and (2) the achieved reconstructions are statistically significant and reproducible. Furthermore, experimental tools for administering systematic perturbations to the networks under study took a while to develop, so that the need to predict dynamical responses to perturbations had not emerged for a while. As confidences in the statistical power of reverse-engineering grew, and new experimental tools were developed, the next level of questions naturally emerged. It is in answering this next level of questions that network reconstructions have found their broad spectrum of highly nontrivial, often unique, and even central roles in modern systems biology. For the remainder of this section, we survey several key application areas, focusing on the most impactful types of predictions that reconstructions are capable of generating.

2.2.1 Predictions regarding individual nodes or interactions

Reconstructions can help identify intervention targets or functionally similar cohorts of biological species.

The advent of modern, high-throughput data acquisition techniques transformed the enterprise of network reconstruction from a painstaking, often collaborative process into an exercise in algorithmic design. Verifying the existence of a single interaction no longer demands corroboration by multiple independent research efforts, and connections can now be inferred in parallel directly from a single set of data. An oft-cited consequence of this change of pace contends that modern reverse-engineering dramatically increases the rate at which hypotheses about potential interactions can be generated. To this end, whole-network reconstructions allow us to rapidly elucidate both the presence and nature of individual interactions, as well as predict the function of individual nodes from knowledge about their neighbors [129, 130].

Inference methods designed for the express purpose of proposing novel interactions for experimental verification [131, 99] have confirmed previously established gene targets [108] and identified novel targets for known transcription factors and drugs [132, 133]. Known broadly as statistical or *association* methods (see "Who talks to whom," Section 2.3.1), algorithms in this class have also discovered entirely new interactions [132, 134, 135, 47, 48, 136, 137], with previously unknown gene interactions often being verified experimentally [138, 139]. In a multi-algorithm litmus test, several of these methods were capable of inferring missing links in artificially corrupted, incomplete versions of established pathways [140].

Network-based strategies for the prediction of protein function [141] generalize more traditional approaches, such as clustering analyses [129], that have been used to classify genes and proteins according to their role at either the physiological or the network level. Individual gene clusters correspond to distinct functional groups in some systems [142]. They can be used to infer roles for unclassified elements according to the guilt-by-association (GBA) heuristic (i. e., assigning functions similar to those of nearby neighbors in the interaction space).

Clustering alone cannot produce a full interaction map, and its applicability is limited because its underlying assumptions are not universal among biological system types [143, 144] (GBA may be more valid for protein-protein interactions than gene-gene interactions, since the latter entail more latent or intervening steps). Nevertheless, clustering is still useful in modern reverse-engineering, predominantly in the data-processing phase that often precedes the inference of full interaction architectures [145]. Clustering the data prior to inference greatly restricts the search space by providing an effective prior to bias the set of candidate interactions. On the other hand, the same idea can also be used to *coarsen* inferred networks: “module-based” inference techniques [146] have identified entire groups of genes that are functionally related [147]. We will return to this idea of identifying coarse functional and conceptual (as opposed to simply structural) units in the *Discussion*.

2.2.2 Insights from the statistical properties of network ensembles

Certain structural statistics differentiate real biological systems from other kinds of complex networks.

While the rapid verification of microscopic interactions undeniably constitutes an improvement in the pace of discovery, it does not by itself generate categorically new kinds of knowledge. Systems biology is “more than an accelerated program of molecular biology” [113], and the relatively new tools of reverse-engineering must prove their worth by helping to play a part in that grander enterprise. This is reflected in the possibility of using reconstructions to make predictions not only about single nodes and individual connections, but about the statistical properties of network *ensembles*.

Work in this direction has produced various insights about what distinguishes biological systems and endows them with their unique characteristics among complex networks. For instance, it has been shown that the most highly connected nodes in protein networks are likely to be essential [148] for survival [149, 150]. Moreover, nodes with an exceptionally high degree (i. e., number of connections), called *hubs*, attach preferentially to nodes with low degree while tending to avoid one another [151]. This property, in part, underlies the widely observed *modular* organization of cellular systems: an efficient coding scheme in which network partitions include only components involved in related processes. This discourages overlap and ensures that (on average) no single node participates in too many processes [71]. This forms the basis for one type of biological robustness [152].

Certain modular structures recur with disproportionately high frequencies in biological systems (with respect to their chance rate of appearance in a random graph [53]). Known as *motifs* [57, 153], they can endow the network with vital control and design features, such as positive or negative feedback, and are often conserved throughout evolution [154, 155, 156]. Studying the appearance rates of motifs across different networks can help clarify the functional "purpose" they satisfy within a given network.

While a node's degree is its most fundamental attribute, studying other network parameters has also led to key insights. The betweenness centrality [53] for nodes in protein interaction networks has been observed to be even more highly correlated with protein essentiality than the degree [157]. Moving beyond individual nodes, it has been argued that the full degree *distribution* is approximately scale-free [56] for many systems, providing deep architectural support for the robustness of biological systems to noise and perturbations, at both environmental and genetic levels [158] (yet see [159] for a cautionary note about the associated power-law distributions).

In *network medicine* [10], clinically relevant predictions can often be made from such high-level statistics, irrespective of whether interactions can be enumerated exhaustively or determined at a fine-grained level. For instance, the aforementioned correlation between a node's degree and its essentiality for survival begets the notion that candidate drug targets can often be ruled out immediately if they are too highly connected, such that using them risks compromising the rest of the network [160].

While one should not focus exclusively on the architectural aspects of dynamically engaged networks [128], even microscopic statistics can sometimes go beyond structure to tell rich stories about the behavior of the underlying system. Maximum Entropy [161, 162] methods [106] (see Section 2.3.1) have been used to learn the effective coupling constants that connect neurons in the retina [74, 163], where the inferred values suggest that these networks naturally reside in the neighborhood of a *critical point* in their parameter spaces [164]. This might afford such networks an essentially optimal capacity for stimulus representation, as well as information storage and transmission [165, 166] (though see [167] for an alternative viewpoint). For the amino acid interaction networks that keep track of where bonds form during protein folding, the same methods corroborate the idea that geometrically proximal residues tend to coevolve [63] by showing that bond locations can be identified using a simple statistic on the ensemble of viable protein sequences (in this case, correlations between the activations of site pairs).

2.2.3 Using statistics to characterize or classify individual networks

Ensemble statistics can help identify defective or emergent properties in a network.

Sometimes, statistical surmises can be used to make statements about the typicality of a particular network. An approach known as *differential networking* (so named to contrast with *differential expression*, a popular type of approach to activation data in gene networks) has been increasingly used for this purpose.

For example, Refs. [168, 169] discuss the idea of using topological characteristics to solve supervised classification problems, such as determining whether a given network comes from a healthy or a pathological organism. This possibility is explored explicitly in [83], which nominates several criteria (reduced clustering and “small-worldness,” reduced probability of high-degree hubs, and increased robustness) as those which are markedly altered in patients with schizophrenia. The reconstruction method developed in [170] was able to identify genes that are either known tumor drivers, associated with biological processes relevant to disease, or correlated with patient prognosis for various types of cancer by examining how pathological networks differ from their counterparts in "normal" tissue. Changes in hub structure have also been used to forecast the survival outcome for breast cancer patients [171].

It is worth pointing out that the aforementioned Maximum Entropy methods [106] provide, in some sense, a complementary approach to ensemble statistics. Rather than addressing only aspects that networks have in common (or can be averaged over), these approaches are predicated on exploiting the intrinsic *variability* at the micro-scale in an attempt to reproduce what is conserved at the macro-scale. This is especially useful wherever diverse microscopic network connectivity structures are known to produce indistinguishable behavior at coarser resolutions, as in protein folding: there is no one-to-one mapping between amino acid sequence and tertiary structure, but an entire distribution of microscopic parameters – a wide variety of equally viable amino acids sequences – that code for roughly the same protein shape [172, 173, 174, 61]. Knowing this, one can easily imagine how running Maximum Entropy methods in reverse can help determine, for example, whether a given amino acid interaction network represents a viable protein. The same might be said for evaluating the typicality of an inferred retinal network, by measuring properties like criticality [175, 176] (NB: for a selection of competing viewpoints on the criticality of neuronal networks, consult the aforementioned [167], as well as [177, 178, 17, 179]).

2.2.4 Predicting how a given network will respond to perturbations

Reconstructions help identify and quantify response patterns in novel conditions.

Network models capture and summarize complex dependencies the among states of biological components, often allowing one to predict how a system will change its state or behavior with changes in the biological environment (i. e., modifications affecting the state of one or more nodes or interactions). Commonly studied perturbations can be local [180] (e. g., knockout of a single gene, as in the simulation of deleterious mutations), multifactorial (affecting many elements) [181], or fully global [182] (applying a drug to slightly suppress the firing of all neurons in a circuit), and the system's responses can be investigated at local or global levels as well. For instance, one might inquire about the effect of a drug or a mutation on the expression of a single gene, or the success or failure of signal propagation from start to end through a perturbed pathway.

The types of responses that are interesting to researchers vary widely, and range across a spectrum of detail. The simplest and the coarsest entail qualitative predictions: for example, is the activation state of a given node affected by a specific perturbation? Progressing to a more quantitative picture, one can try to predict the actual post-perturbation values for affected nodes, as in the prediction of gene expression levels following a knockout event [183]. At the finest granularity, models incorporating time-series measurements can be used to forecast the transient behavior for such a gene as it approaches a new steady-state expression level.

Recent DREAM Challenges have provided a testing ground for algorithms aiming to make these types of predictions. The DREAM4 Predictive Signaling Network Modeling Challenge [184] instructed contestants to predict phosphoprotein measurements "using an interpretable, predictive network"[1], and the bonus round of that year's *in-silico* Challenge [186, 187, 180] asked competitors explicitly to predict the system's

[1] `http://dreamchallenges.org/project/dream4-predictive-signaling-network-modeling`. The solution presented in [184] infers a network using Boolean truth tables, one of the most popular approaches during the early stages of automated network inference [185]. This approach has since fallen out of favor, yielding to the more sophisticated methods we discuss in Section 2.3, but Bayesian networks are often still discretized to Boolean values for convenience.

responses to "novel" perturbations that were not encountered in the training data. The DREAM7 Network Topology and Parameter Inference Challenge [140] specified the prediction of perturbation outcomes using gene regulatory network models as a separate step from inferring their topologies.

As we discuss later, prediction of time-course trajectories requires directed networks, but the converse is not true: directional links can sometimes be inferred from static data. On the level of qualitative predictions, the linear dynamical systems approach of [108] was able to deduce the targets of novel perturbations in a system of nine genes using only steady-state values of their expression levels, following a series of highly controlled perturbations (and the knowledge of which genes were targeted during the perturbations). We consider this result to be particularly important, for two reasons. First, it challenged previously expressed (and still later-held [188]) ideas by successfully determining a directed network, despite the fact that the applied perturbations elicited statistically significant changes in the activations of all nodes. Second, later improvements extended the abilities of the algorithm therein to determine which species were "hit" by applied perturbations even *without* specifying as inputs which genes were targeted during the data acquisition phase [189], reinforcing the idea that M static, independent, but carefully selected perturbation measurements can substitute for a series of time-course measurements taken at M intervals [190].

2.2.5 Representing the joint probability distribution for observables

A network model can be interpreted as shorthand for a joint probability distribution.

Activation values for each node depend on those of many others, rendering graphical models particularly convenient representations of their joint activities. Graphs can explicitly encode the statistical dependencies among different activation variables as connection *weights*, with the states of connected nodes given not by a stochastic transfer function, but by conditional probabilities.

A type of directed acyclic graph (DAG) known as a Bayesian network is a weighted construction whose connection strengths are typically learned [191] via Bayesian inference (i. e., computing the posterior probabilities for a set of candidate DAGs, and selecting the member with the highest value, etc.) Undirected variants, which communicate only binary dependency information via the presence or absence of symmetric links are popular in different applications. When activities are assumed to deviate normally from baseline values (an assumption that greatly simplifies the inference process), they are known as Gaussian graphical models [192].

Connection weights in a Bayesian network can be scaled so as to represent a proper, normalized probability distribution. Adjusted to match that of the observed data, the weights in such a dependency graph become an explicit encoding of the system's joint statistics. Bayesian networks satisfy a *Markov property*, such that the activity value distribution for a given node depends only on the values of its immediate predecessors (these activities are often discretized as binary variables for mathematical convenience, so the resulting graph neatly keeps track of the probability that a downstream node in the inferred network will be active if its predecessors are active). This directed conditional dependency structural arrangement offers a conceptually accessible and intuitive view of the system, although the presence of directed connections between two nodes does not mean there is a direct physical (i. e., mechanistic) or causal link between the corresponding species [193].

One of the most important and unique applications of network inference, this compact representation of probability distributions permits the quantitative prediction for nodal activity values, in both static and dynamic contexts. Probabilistic graphical models are particularly useful in putting numbers on answers to questions like “What is the probability of this protein being active, given that a particular stimulant is present?” or its converse: “What is the probability of the stimulant having been present, given that the expression level of this gene is high?” [194]. We discuss methods for inferring both types of probabilistic graphical models named here, and their limitations (including their ability to infer causality), in Section 2.3.

2.2.6 Reconstructions as a part of the Big Picture

Inferred network models can be combined with existing and new methods as one part of a larger repertoire for investigating many facets of living systems.

Reconstructions are increasingly combined with other tools and prior biological knowledge to form integrated frameworks for discovery. Some reverse-engineering approaches attempt to incorporate prior knowledge explicitly into the inference process for individual networks [195, 196, 197, 198, 199, 200], including one study which advocates the use of undirected gene networks (gleaned from functional association databases) as *priors* to enhance the inference of mechanistic, causal gene regulatory networks [201].

Other applications use networks to cross-reference, corroborate, or pre-screen evidence for predictions about specific systems. For example, the “network approach” to genome-wide association studies (GWAS) and disease gene prioritization is reviewed in [129], and the use of networks for the prediction of protein functions (in the general sense, not restricted to physical binding), evolutionary studies of pathogenic and non-pathogenic strains, and the bidirectional interactions between host and pathogen are reviewed for the specific context of infectious disease in [130].

We have already mentioned the work [126], which uses Bayesian networks in tandem with support vector machines to predict the toxicity of various chemicals in a supervised setting. Yet we believe the most pivotal roles to be played by reconstructed networks are those which completely change the way we think about biological phenomena, specifically by offering new ways to predict system-wide behaviors. Such a revolution is already underway in medicine: the treatment of various diseases is no longer unilaterally viewed from within the "one-gene, one-drug" paradigm, and it is gradually becoming the new standard to view related autoimmune disorders as emanating from a network of maladies with the same root causes [202, 203, 204].

2.3 Two different meanings of phenomenological "reconstruction"

We distinguish two principal categories for phenomenological network inference, accounting for methods that produce undirected and directed graphical models.

Algorithms in our **first category** define an inferred interaction as an *irreducible statistical dependency* among nodes, typically quantified by some measure of the similarity among the activation profiles of different nodes. This is a structure-only approach, and should be used when it is only necessary to reconstruct the overall network topology – in other words, for applications for which it is sufficient to know "who talks to whom." In some cases, topological maps can be augmented with weights that ascribe an effective strength or confidence level to the inferred interactions [205, 174].

Algorithms in our **second category** define interactions in terms of *asymmetric relations* capable of describing not only which nodes participate in an interaction, but also "who controls whom." Previous classification schemes have considered the inference of unweighted, directed links as a separate endeavor from discovering quantitative input-output relationships between nodal activities [206], or further distinguish algorithms that detect the sign of interactions without an explicit direction [207, 208]. However, since both the types of data and the processing techniques needed to infer all these kinds of graphs are similar, we treat them on equal footing.

2.3.1 Who talks to whom? *Presence, absence of undirected links*

The most basic question that one can answer in the course of network reconstruction is whether a given subset of nodes can be characterized as interacting – in other words, *who talks to whom*? Since our focus here is on the unsupervised inference of interaction networks directly from activation data, any notion of "interaction" that we consider must depend on these activations alone. A natural definition for the existence of an interaction among species is the presence of statistically significant correlations among their respective activation states. Such a choice results in an undirected network with symmetric (though possibly weighted) connections.

In practice, pairwise statistical dependencies are typically quantified by introducing a *similarity* metric, such as the first-order Pearson correlation. The Pearson correlation coefficient is a normalized, pairwise dependency measure bounded by the interval $[-1, 1]$. Positive (negative) values indicate an increasing (decreasing) linear relationship. While its value is always zero for statistically independent variables, a vanishing Pearson correlation cannot rule out nonlinear correlations. Conversely, in the absence of nonlinear effects, finite sampling can cause independent variables to appear correlated, so that connections can be inferred where no otherwise discernible interaction exists. To avoid inferring such spurious interactions, one must apply a threshold to filter raw correlation values.

When nonlinear effects cannot be ignored, one can quantify statistical dependencies using information-theoretic measures [209, 210, 211], which generalize the notion of correlation to such nonlinear cases. D'Haeseleer et al. [212] were the first to employ the mutual information to uncover gene-to-gene dependencies, while Butte et al. applied mutual information "relevance networks" [213] to propose single-gene determinants of anticancer agent susceptibility [214] for experimental verification. Mutual information-based methods must still contend with the same sampling and bias problems faced by linear correlation coefficients, and therefore also require thresholding.

Even under conditions of perfect sampling, neither Pearson correlations nor the mutual information can disambiguate so-called *direct* interactions from *indirect* interactions – statistical dependencies that are already accounted for by links involving other species. Note that this notion of "indirect" is distinct from its usage in the context of mechanistic networks. There, "direct" typically refers to physical contact, which often occurs between nodes whose activations are not included in the network model (unobserved, latent, or marginalized degrees of freedom in the system). Here instead we are concerned with statistical redundancies within the set of *observed* activation variables. For example, consider the case of three genes in a regulatory cascade: $X \to Y \to Z$. Inference methods based on measuring correlations between the associated activation variables would find a link between X and Z, which is *indirect*, in the sense that it is not actually needed to account for the joint statistics of X, Y, and Z.

While sometimes inconvenient, indirect links are not always superfluous. They are useful when probing the network at the single-node level, as when trying to discover a previously unknown member in an established pathway, propose a novel interaction for experimental verification, or predict the overall effect on the activation state of one node by perturbing another. On the other hand, in applications for which inferred networks must be treated as whole entities (e. g., when they encode normalized probability distributions; see MaxEnt methods described below), this sort of redundancy can be minimized by examining *conditional* dependency structures.

There exist several approaches to studying conditional dependencies. The most intuitive is to work explicitly with either partial correlation coefficients [131] or the conditional mutual information [215, 216, 47, 48, 217] between two activation variables

X and Y, given another variable (or set of variables) Z:

$$I(X;Y|Z) = I(X;Y,Z) - I(X;Z), \tag{2.2}$$

where $I(X,Z)$ is the mutual information between X and Z. In principle, one can refine a reconstruction by removing links between any pair of species X and Y that are associated with statistically insignificant values of $I(X;Y|Z)$. However, reliable estimation of this quantity is much more difficult than it is for the pairwise quantities, such as $I(X,Y)$, since it requires sufficiently dense concurrent sampling of at least three variables.

In order to dispose of indirect links without incurring the aforementioned estimation problems, some algorithms make additional assumptions and thus append ancillary filtering steps to the basic mutual information-based procedure. For instance, the Algorithm for the Reconstruction of Accurate Cellular Networks (ARACNe) [132, 99] invokes the Data Processing Inequality [210] to delete the weakest link in every closed triplet of nodes (this would be an exact step if the studied network was a tree). The Context Likelihood of Relatedness (CLR) method [134] determines the presence or absence of a link by assessing its strength against all other mutual information scores computed for that graph, as a background significance threshold. MRNET [218] builds a network iteratively, including a link between two variables if one is both a good predictor of the other and yields information that is non-redundant with that provided by the previously inferred links.

An alternative approach to solving the conditional independence problem is to use full probabilistic models that allow conditioning on the complete set of marginals, rather than requiring the progressive computation of higher-order partial correlations [217]. In particular, if a set of continuous, real-valued activation variables are (assumed to be) normally distributed, one can condition a single interaction on the full set of remaining variables. In this case the statistical independence of any two nodes can be ascertained by examining the elements of the inverse of the covariance matrix: $\Sigma_{ij}^{-1} = 0$ if and only if i and j are conditionally independent, given all other variables. An important facet of such multivariate Gaussian distributions is that they correspond to the least constrained, *maximum-entropy* models that satisfy the full set of first and second-order marginals for continuous variables [162, 106]. These first two moments correspond to the individual means and the pairwise correlations, which are usually well measured even in sparsely sampled data sets.

Beyond Gaussian variables, the Maximum Entropy principle has been a successful modeling approach in neuroscience [74, 219, 220, 221], natural images [222], the inference of gene networks (from expression data) [223] and signal transduction networks (from phosphorylation proteomics data) [224], and the prediction of amino acid contacts in proteins [172, 225, 173, 226], multidrug effects [227], protein structural attributes [62], antibody diversity [228], and even the dynamics of flocking birds [229]. The joint probability distribution for a Maximum Entropy model has a particular form, known in statistical mechanics as the Boltzmann distribution. If we ask to match only the empirical means $\langle x_i \rangle$ and pairwise correlations $\langle x_i x_j \rangle$ to those of the observed data, the distribution with maximal entropy is

$$P(\vec{x}) = \frac{1}{\mathcal{Z}} \exp\left(\sum_i h_i x_i + \sum_{ij} J_{ij} x_i x_j \right). \tag{2.3}$$

Here parameters h_i and J_{ij} are known as the fields and the couplings, respectively, and $\mathcal{Z}$ is the partition function (compare to the full expansion in Section 2.1.3).

For discrete variables, the Maximum Entropy model retains the form of Eq. (2.3), but is known as the Ising model (for binary variables) or Potts model (for categorical variables with more than two accessible states). In the discrete case, fitting the parameters $\{h_i, J_{ij}\}$ is highly nontrivial. Many methods exist, but their effectiveness depends on the system size and the density of its interactions, as well as on other properties [230, 231, 232, 233, 234]. One algorithm worth mentioning is the *adaptive cluster expansion*, which was developed in the context of the MaxEnt problem [234, 235]. It is closely related to information-theoretic approaches, being equivalent to relevance networks [213] for clusters of size two, and similar to conditional mutual information methods for clusters of size three.

Due to the limitations of finite sampling, both solving for the inverse of the covariance matrix and learning the parameters of an Ising model can constitute ill-posed problems. One way to avoid this is to impose a regularization [236], which invokes additional constraints on the interaction coefficients to ensure that the inference problem is well-defined – and moreover, that the inferred network generalizes well to unseen data. Regularization is often done in one of two common ways: either the interactions coefficients are assumed to be small (for example, using an L_2 norm) [235] or the interaction structure of the system is presumed to be sparse, so that the overall number of the interactions is small (this may be done explicitly by specifying the number of non-zero coefficients [108] or by invoking an L_1 norm [233]).

Frequently cited as the rationale behind these regularization procedures is the inherent sparsity of natural networks [237, 149, 238]. Indeed, for protein studies, the nodes in networks used to describe tertiary protein structure represent real amino acids in the three-dimensional space; they can therefore be connected to only a small subset of all possible neighbors. Similarly, the number of transcription factors that can influence a given gene's expression levels is limited by the physical extent and arrangement of its promoter sequence. While the general ubiquity of sparseness in biological systems is debated [103], the enforcement of sparsity constraints can be justified as a purely pragmatic measure in the "low-hanging fruit" inference regime.

2.3.2 Who controls whom? *Causal relations, directed links*

Directed network inference differs in an important way from that of undirected, symmetric, mutual-influence graphs: since questions of causality (or, more generally, the flow of information) are built not upon a single, universally agreeable concept like statistical correlation – but rather on more subtle, less straightforward notions like *control* – there exist many diverse criteria for establishing directed connections. Each method has its own operating definition of what counts as an interaction, and how to infer its direction.

Though disparate, the aforementioned definitions can be conveniently divided into approximately two subclasses, depending on the intended application of the inference procedure. In certain cases, it is enough to know the direction or causal *sense* of an inferred interaction. For example, will silencing a certain gene or disabling a particular neuron result in a collapse of the entire system? Can the intracellular concentration of a reactant be increased by introducing more of the product? Answers to questions like these do not require numbers, entailing purely qualitative predictions. On the other hand, if the goal is to use a reconstructed network to predict the amount by which one gene's expression level increases when two other genes are suppressed, directed connections must be weighted by quantitative values representing the effective *strengths* of interactions. We describe methods of both types, leaving it as an exercise for the reader to think about when a directed topology suffices, and when it is necessary to infer fully signed and weighted graphs.

Before we delve into specific methods, we advise the reader to tread with caution. The particular definitions of directed influence we explore in the following methods do not always correspond to our intuitive and/or mathematically formal notions of causality. As a result, producing a graph with directed links does not automatically satisfy a reverse-engineer's desire to uncover system-wide causes in an ontological sense, and should not be treated as such despite one's instincts. Instead, great care needs to be taken with each method in order to ensure that all idiosyncratic constraints are met, and to avoid generalizing or extrapolating beyond the predictive power of each algorithm.

To expound on this point, it is worth asking at the outset whether it is even possible to infer causal information from passive observations of activation variables [239, 240]. It has long been understood [241] that proximal causal relations can be inferred reliably when the observer is able to *interact* with the system in accordance with a principled protocol (as is done in many controlled experimental interventions [242, 243], including genetic knockouts [180, 183] and multifactorial perturbations [181, 244]). While this is old news to engineering audiences, it has also been shown that causal information (or at least a lower-bound estimate of causal effects) can be extracted from purely observational data when the equivalence class for the fully directed graph can be ascertained first [245, 246][2].

We mention again a surprising corollary of this result that directed influence (a less stringent condition – and slightly less nebulous concept – than causal influence) can often be established without time series data, using only static measurements. Where there was once a prevalent belief in the reverse-engineering community that the inference of directed edges requires temporal data [99], there is now a tradition of algorithms which accept static data as inputs [242, 103, 107, 108, 248, 249, 250, 251]. For coherence, we focus predominantly on methods that operate on time-series data.

We organize this subsection as follows. We first make a few general remarks about the inference of directed interaction patterns. We then explore a class of methods which presume that the measured activities can be treated as deterministic variables that change smoothly in accordance with a particular, predetermined quantitative

[2] Once the equivalence class is determined, formal causality detection methods can be subsequently applied to estimate the full causal graph. We refer curious readers to [241, 247] for a wealth of both philosophical explications and more rigorous treatments of algorithms designed explicitly to detect causality in its many guises.

law. Afterwards we switch to model-free deterministic methods, for which there is no need to specify a mathematical form or law in advance in order to detect interactions. We then treat the more general situation, in which activations are regarded as stochastic variables. Again we start with methods requiring a parameterized model and conclude with a discussion of stochastic model-free methods.

A naïve but conceptually intuitive approach to inferring directed connections is to take the presence of strong temporal correlations between the trajectories of different activity variables as evidence for a (casual) interaction between the corresponding species. It is common for changes in one activity variable to succeed that of another in time (consider a gene whose expression level is observed to increase consistently in response to the elevation of another), but the proxy of temporal precedence is not robust as a criterion for declaring control relations [92] because it also appears in the absence of causal influence. Despite its limitations, this strategy, combined with a projection method known as multidimensional scaling [97] in an algorithm entitled "Correlation Metric Construction," was originally used to infer the first steps of the glycolytic pathway [252] and more recently applied to study the pharmokinetics of the anticancer drug Gemcitabine [253].

In physics and engineering, signed and directed connections are often used to encode the weighted coupling constants that appear in systems of differential equations [254]. To write down such a system, one needs to first have in mind a particular quantitative form for a dynamical law, according to which activations will be presumed to interact. One then fits the model parameters, typically with some optimization or statistical learning technique that takes time series data as input, and reports the learned values as the weights for the corresponding connections, sometimes adding additional, unobserved, hidden variables in the process [11].

The inherent directionality of this method, which works best for small systems ($p \sim 10$), can be understood immediately by examining the matrix J_{ij} of pairwise interactions in Eq. (2.4) below: since this matrix is not constrained to be symmetric, couplings between two species can differ in the forward and backward directions. For continuous activation variables $\{x_i\}$, many popular models can be subsumed as special cases of the general form (though see [255, 11] for alternative forms):

$$\frac{dx_i(t)}{dt} = f_i\left(x_i + \sum_j^p J_{ij}x_j + u_i + \xi_i\right), \tag{2.4}$$

which includes at most pairwise interactions of strengths $\{J_{ij}\}$ between all element pairs i and j. Here the functions $\{f_i\}$ can be chosen according to the desired level of computational complexity (controlled by the amount of data available) or biochemical detail, or both. In the reverse-engineering of biological networks, many early applications were linear activation models [256, 257, 258, 259], for which $f_i(x) \propto x$. The sum determines the net (excitatory and inhibitory) effect on the activation of node i at time t, given its interactions with all other elements j. The next term accounts for external driving of the node, (i. e., any extrinsic perturbation that increases or decreases its activation value by an amount $u_i(t)$), and $\xi_i(t)$ represents noise.

Linear, “additive” regulatory models are based on the assumption that dynamical systems can be *linearized* about their steady-states. They are relatively easy to fit in sparsely sampled conditions, especially when the terms in Eq. (2.4) are discretized to form a linear difference equation [258, 260]. Early work countered undersampling by augmenting the number of data points for multilinear regression via nonlinear interpolation [257], or imposing sparsity constraints on singular decomposition algorithms [107]. Another approach to decreasing the number of interactions that must be inferred is to first cluster the nodes [145]. In any case, data are typically taken during the system’s approach to steady-state conditions (whether its natural equilibrium or another fixed point of its dynamics) after a perturbation.

A straightforward modification of the basic linear model, realized by overlaying the sum in Eq. (2.4) with a sigmoidal threshold function, leads to one version of the *artificial neural network* construction. Early methods based on neural networks were used to infer interactions between individual [261] and aggregate “genes” which encompass multiple degrees of freedom at the biological level [259]. Modern improvements use multilayer perceptrons [262]. Early neural-inspired architectures known as gene circuits [263] have also been used to infer mechanistic interactions [264].

Nonlinear models are attractive because they can capture more sophisticated dynamical behaviors than their linear counterparts (e. g., oscillations and multistability). Nonlinear reverse-engineering schemes based on mass-action kinetic laws like Michaelis-Menten or Hill equations [57] are also used in reconstruction [265, 266].

An important causal inference method based on the assumption of an underlying deterministic system, but which does not require the definition of an explicit dynamical model, is the convergent cross-mapping (CCM) approach [92]. As noted in [268], an essentially identical method had been developed earlier to study synchronization in chaotic dynamical systems [269]. The method draws from Takens' theorems [270], which provide both the conceptual framework and mathematical justification for a brand of state space reconstruction – reverse-engineering of the phase-space portrait for a dynamical system – known as *delay embedding.* Consider a multidimensional dynamical system, a special case of the general form (2.4) whose parameters are fixed, and whose temporal evolution $\mathbf{x}(t)$ is confined to a subspace determined by a d-dimensional attractor [271]. Under very general conditions, the attractor's state space can be reconstructed [270] from measurements of a single time series $\{x_t, x_{t+\tau}, x_{t+2\tau}, \ldots\}$, sampled at an interval τ. The number of consecutive time points needed to span the reconstruction space is given by the attractor dimension d; both τ and d are often found using Ragwitz' criterion [272], but alternative methods have been proposed as well [273, 274].

Delay embedding refers to the entire process of defining these two parameters

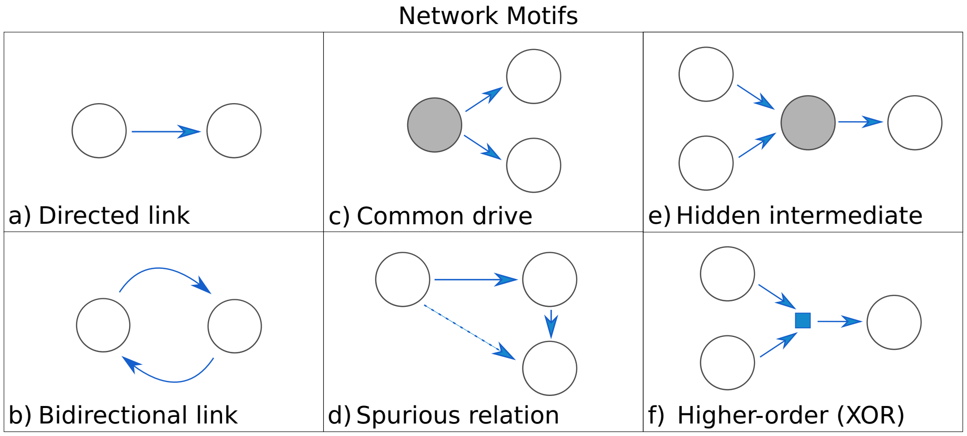

Figure 2.2: Simple directed network motifs help illustrate basic problems in directed network reconstruction. This list is not intended to be comprehensive, but to address some simple yet important scenarios. Links can represent mechanistic interactions or effective relations (i. e., information transfer). Nodes represent stochastic or deterministic activation variables, which can be either continuous or discrete. Here, dashed links represent spurious (erroneously inferred) interactions, dark nodes represent unobserved (hidden) variables, and the small square in f) refers to a computation that involves more than two nodes (in this case, a third-order interaction). **a)** The simplest scenario: a directed link between two nodes. **b)** A bidirectional coupling models a simple system with feedback (e. g., the predator-prey system of Fig. 2.1). **c)** A hidden common drive (dark node) to two observed nodes results in a correlative relation between those nodes. If care is not taken, this can be confounded with a direct causal interaction. **d)** A situation similar to that of c), with the difference that measurements of all three nodes are accessible. Naïve pairwise methods infer a spurious link between the initial and final nodes in the feedforward chain. Multivariate methods are required in this scenario to infer the correct links. **e)** In the case of a hidden node relaying the causal interaction, network reconstruction methods may infer the correct direction of interaction, but the inferred links will be effective rather than strictly causative since an intervention at the hidden node can disrupt the interaction. **f)** The logic gate XOR entails a higher-order interaction. The output is 0 if both input nodes carry the same value, and 1 if they are different: simultaneous knowledge of the states of both nodes is required to determine how each of the inputs affects the output. This is a classical example of a scenario where X and Y carry synergistic (as opposed to unique, or redundant) information [267].

and arriving at a reconstruction space onto which the time series can be mapped. It provides the substrate for causal inference via CCM as follows. For any two measured times series $\{x_t\}$ and $\{y_t\}$, the variables x and y are said to be causally linked if they belong to the same underlying dynamic system (i. e., the time series they represent are samples from the same attractor [270, 92, 271]). The direction of an interaction between x and y variables can be estimated by 1) using delay embedding to obtain reconstruction manifolds $\mathcal{M}_x$ and $\mathcal{M}_y$ for x_t and y_t, respectively [271]; 2) projecting one of the variables, say x, onto the other manifold – hence the name cross-mapping – and using the resulting, projected values to predict the values taken by the original time series (which *converge* to the measured values for a large enough number of samples); and 3) measuring (with any suitable measure, e. g. RMSE or correlation function) the deviation of the predicted values $\{\hat{x}_t\}$ from the actual values $\{x_t\}$. A causal interaction is declared if the prediction quality does not decay to zero for a growing number of samples.

In the original work, Sugihara et al. [92] did not analyze thoroughly the influence of noise on reconstruction. Indeed, Takens' original theorems allow for noise in the measurement procedure only (i. e., intrinsic stochasticity is prohibited; the breakdown of inference based on CCM in the presence of intrinsic noise has been demonstrated explicitly [275, 276, 277], and a thorough analysis of state space reconstruction in the presence of noise can be found in [272]). Nevertheless, artificially added measurement noise can actually improve the detection of causality [278].

Several other considerations must be taken into account when inferring causal relations by means of CCM. First, it seems that the outcome is quite sensitive to the sampling methods used to obtain training data (for example, eliminating nonstationarity on the way to the attractor is key) [268]. Second, CCM fails to infer the accurate coupling strengths and even the direction of causal interaction when time series are synchronous [277]. Third, it has been shown that the predictions made by CCM do not always conform to our intuitive notions of causality, even for certain rudimentary systems like a simple resistor-inductor (R-L) circuit with a sinusoidal driving voltage, where CCM does not unequivocally determine the causal dependence of the current on the voltage [275]. Finally, Cobey and Baskerville [276] provide a thorough numerical analysis of the limits of CCM, suggesting that the standard approach is generally prone to failure if the system dynamics are oscillatory and proposing a modification in the algorithm to alleviate this shortcoming [276].

For stochastic activations, early attempts to reconstruct the directionality of interactions included autoregressive models [279, 280], but autoregression by itself makes no assertions about causality. However, a method due to Granger [281] combines autoregression with the aforementioned notion of temporal precedence to infer quantify a robust stand-in for causality – namely, Weiner's predictability [282]. The framework for Granger Causality (GC) is built upon two central assumptions [283]:

1. The cause x occurs before the effect y.
2. The causal series $\{x_t\}$ contains unique information about the time series being caused $\{y_t\}$ that is not available in any other series $\{w_t\}$.

More generally, $\{w_t\}$ represents the entirety of processes that can influence $\{x_t\}$ and $\{y_t\}$. In the ideal scenario, for which these three variables together contain "all the information available in the universe at time t" [283] (i. e., in the closed system under investigation), GC guarantees that one can reconstruct the direction of the causal relationship between x and y. By definition, a variable x "Granger-causes" variable y if knowledge of past values of both x and y reduces the variance of the prediction error for y, in comparison with the history of y alone. Typically, these predictions are carried out via linear regression, and the direction of causality is decided by statistical tests on the variances of the respective residuals (prediction errors). However, this implicitly assumes (at most) linear relations between variables. Nonlinear extensions of GC exist, but these extensions can be more difficult to use in practice and their statistical properties are less well understood [284, 285, 286, 287, 288].

Granger causality can be extended to multivariate scenarios [289] as well, although finding Granger-causal links among all possible candidate interactions then becomes a combinatorially hard problem. For the particular case of inferring causal relations between the activity of distinct brain areas (using electroencephalograms or local field potential time series), it has been found to be of crucial importance to employ a multivariate approach rather than bivariate techniques [290].

A more general approach to the reverse-engineering of directed links between stochastic variables is to learn an explicit model for the joint probability distribution of the observed activities. This approach, based on *probabilistic graphical models*, was discussed earlier for undirected networks. For the directed case, one can define a class of models known as Bayesian networks [291, 292, 293, 294] which decompose the joint distribution into separate factors representing conditional probabilities. Edges are drawn starting from the nodes corresponding to variables being conditioned on (called the "parents") and ending on the conditioned variables (the "children") [294, 241]. Since the joint distribution of a Bayesian network is an exact product of conditional probabilities, the resulting graphical structure is a *directed acyclic graph* (DAG). Thus in order to be eligible for representation by a Bayesian network, systems need to satisfy the necessary criteria for forming a DAG. If the phenomenon in question is known to encompass cyclic dependencies (e. g., autoregulation pathways in gene regulatory networks, or autapses in neuronal networks), the only recourse is to "unroll" the cyclic dynamics in time, forming a *dynamic Bayesian network* [295, 193, 296, 297, 298]. The performance of dynamic Bayesian nets has been been compared directly against that of Granger causality [299], and favorably so when the observed time series are shorter than a certain length (NB: In general, findings like these should be taken with a grain of salt, since 1) they could be artifactual results that depend on idiosyncratic features of the data, and 2) notions of error and accuracy tend to rest on the existence of a reference network containing only the "correct" edges, which is in our opinion a dubious concept; see comments on evaluation metrics in the *Discussion.* In [299], the authors are clear in their admission that "the causal relationship derived from these two approaches could be different, in particular when we face the data obtained from experiments," in accordance with our introductory statements about the nonuniform definitions of causality that are assumed by different methods.).

With the conditional probability framework in place, one needs to select 1) a quantitative form for the underlying model that parameterizes the conditional probabilities, 2) a scoring or objective function that quantifies the quality of fit, and 3) an optimization or search routine by which to learn the parameters values that extremize the objective function. An example of such a parameterization, used quite frequently in the literature, is again that of linear regression [191, 294]. The choice of a specific parametric representation of conditional probabilities is often dictated by our knowledge or assumptions about the domain (prior knowledge) [300], or pragmatic principles favoring computationally simple models (Occam's razor). Standard objectives are the maximization of the likelihood function [295] or posterior probability distribution [191], as well as the Bayesian Information Criterion (BIC) [297], which penalizes for large numbers of parameters. Since the optimization search is an NP-hard problem [292, 294], exact methods are often computationally infeasible, so one often reverts to heuristics like greedy hill-climbing (which adds, deletes, or reverses edges to encourage maximal ascent in the objective score [301]), stochastic hill-climbing, or Monte Carlo methods [302].

An impressively comprehensive and thorough body of work regarding the concept of causality and its formal description via Bayesian nets has been provided by their originator, Judea Pearl [241]. Pearl introduces a conceptual framework called the *do-formalism* (known variously as the do-calculus, the intervention-calculus, etc.), which formally describes the use of experimental interventions to ascertain a causal structure. In the do-formalism, $p(y|do(x))$ denotes conditioning on a variable x that is experimentally controlled rather than simply measured (i. e., observed passively). In other words, this notation distinguishes the more familiar observational conditioning $p(y|x)$ from "interventional conditioning" [303, 182].

While correlation does not in general imply causal influence, Pearl reveals specific cases for which the conditional probability distribution – reflecting associative dependencies – is equivalent to that which denotes the corresponding mechanistic dependencies: in such situations, interventions which manipulate the values of parent nodes are clearly and unambiguously seen to have direct effects on the children, and the Bayesian graph is therefore also the correct casual graph.

It is often difficult to satisfy all the criteria for modeling a causal system with DAGs. In certain circumstances, it is easier to work with model-free stochastic frameworks, such as that of the transfer entropy (TE). TE was introduced twice independently, by the physicists Schreiber [304] and Paluš [305], and has since proven to be a versatile and useful tool for inferring the direction of information transfer in neuroscience [306, 307, 308], physiology [274, 309], climatology [310, 311, 312] and economics [313, 314]. TE is simply the conditional mutual information (2.2) between a target variable Y and the entire history of values assumed by a source variable X, given the history of the target:

$$\mathcal{T}(X \to Y) = I(\mathbf{X}_{t-}; Y_t | \mathbf{Y}_{t-}). \tag{2.5}$$

Here the arrow denotes the direction of information transfer (i. e., X informs Y) and $\mathbf{X}_{t-}$ and $\mathbf{Y}_{t-}$ respectively denote the histories of the corresponding stochastic processes up to, but not including, t; Y_t denotes the value taken by the target variable at time t. Conditioning on the history of the target ensures that only those bits of information that are unique (in the sense discussed earlier for Granger Causality; for a formal treatment see [315, 267]) to the source variable are considered.

Like all information-theoretic measures, TE and its surrogates [92] suffer from the curse of dimensionality because of the need to estimate entire probability distributions (discrete variables) or probability densities (continuous variables) for long time series and many variables. For discrete variables, the simplest estimation procedure entails simply counting frequencies to produce a histogram that approximates the desired distribution. A substantially more accurate estimation of information-theoretic quantities for discrete variables (especially if the data set is small) can be obtained by computing entropies directly with the NSB estimator [316, 317]. In the continuous case, a standard approach is to bin the data, rendering the distribution effectively discrete and therefore amenable to histogram methods. While less "data hungry" alternatives exist for continuous variables (such as kernel estimators [318]), they suffer from the same systematic estimation biases that are associated with histogram methods [319], and may even reverse the inferred direction of information flow [320]. Nearest neighbor estimators [319, 308] are some of the most commonly used in practice. In all cases, statistical testing against surrogate data or empirical control data [321] is recommended to help ameliorate the bias problem.

An approach to dimensionality reduction based on the concept of Markov chains has been proposed for the estimation of TE [311]. This approach is particularly useful in the case of delayed coupling between variables [322]: estimation of the delay time can prevent the inclusion of unnecessary time steps when tracking the history of the source variable (i. e., X_{t-} in Eq. (2.5)), which can clearly reduce the dimensionality of the latent representation. Finally, the curse of dimensionality can also be alleviated by first constructing an explicit, low-dimensional model of the time series (and hence, parameterizing the probability distribution). For the simplest case – linear dependence between X and Y with additive Gaussian noise – it has been shown analytically that TE will always recover the same network as Granger Causality, up to a constant factor [323].

Since some authors speak loosely about inferring causality when computing the TE or related quantities like the directed information [137], we reiterate that, although causal interaction is a necessity for information transfer, the converse is not true: information transfer, as quantified by TE and other information-theoretic functionals, does not imply underlying causal interactions. In fact, we caution readers that some methods for the detection of causal or directed influence have been routinely applied in ways that differ markedly from the intentions of their originators. For instance, the directed information was initially designed to infer achievable information rates on a known communication channel with feedback [324], rather than the inference of directed networks (for a thorough discussion, see [308]). However, TE specifically has been extended using the aforementioned do-formalism in a new procedure known as *information flow* [241], a more appropriate measure for inferring causality under certain constraints [303, 325]. Notably, this measure can correctly resolve the connectivities of an XOR circuit (see Fig. 2.2**f)**) even in special scenarios where the conditional mutual information fails [303], a fact overlooked by authors who have contended that conditional mutual information is sufficient for this purpose (see, for instance, the argument in [216]). Finally, we note that TE and similar methods have not achieved widespread implementation for large systems ($p \gg 1$) due to the aforementioned, intrinsic difficulty of estimating information theoretic measures in high dimensional spaces. Multivariate approaches to TE estimation and related methods are a subject of ongoing research.

2.4 Discussion

Since the year 2000, some thirty review articles that we know of have been published on the inference of gene networks alone (in addition to those referenced or mentioned throughout, see [326, 327, 328, 329, 330, 331, 332, 333, 334, 335, 336, 337, 338, 339, 340, 341, 342, 343, 344, 345, 346, 347, 348, 349, 350, 351, 352, 353]), and an increasing number have begun to specialize on the unique challenges faced by network reverse-engineers rather than merely listing different algorithms [354, 355, 356, 357, 129, 358, 65, 359]. One DREAM report [35] notes that the number of PubMed articles on reverse-engineering had doubled each year for over a decade through 2009, and "novel" algorithms (new twists on the same foundational principles we outline above) continue to emerge even as we write [360, 361].

Has this explosive growth in the number of reverse-engineering algorithms and studies helped carve out a niche for large-scale reverse-engineering in contemporary systems biology repertoires? Or has a staunch directive on the reconstruction of entire microscopic networks actually encumbered and obfuscated our understanding of the working principles that underlie these complex systems?

One major impediment to assessing the promise of reverse-engineering algorithms stems from the way in which they are assessed: we observe a rampant, pervasive, and potentially counterproductive tendency to draw direct, quantitative comparisons between reconstructions produced by different algorithms. In other words, despite the commoditization of network inference tools, there is still no consensus on the correct way to *evaluate* reconstruction results [124] – and perhaps for good reason! In the context of effective network inference, the notion that reconstructions can be *checked for accuracy* contradicts our very premise, that algorithms both among and within each of the classes we have described make diverse assumptions about what should count as an interaction. Recent work [125, 362] notwithstanding, we believe this issue continues to be confounded by a repeated mismatch between algorithms and metrics (as in the use of the area under receiver-operator characteristic curves, an intrinsically inconsistent measure [363] that presupposes the existence of a valid confusion matrix, to give an overall rank or "score" to effective reconstructions [355, 364, 365]).

The methods in different classes also differ in more concrete ways: they vary in the extent to which they can infer strengths, signs, and directions for the interactions they detect. This might be thought of as a "feature, not a bug" of reverse-engineering technologies: having a selection of versatile algorithms, each tailored to particular situations or designed with different inference goals in mind, increases the chances that researchers can make use of reverse-engineering algorithms. Yet the question of whether systems biologists should persist in pursuing whole-network reconstruction as a go-to modality or learning tool hinges not solely on whether the inference goals are achieved by the time the smoke clears, but on the attainment of a reasonable *tradeoff* between the computational effort consumed by inference algorithms and the (ideally, unique) benefits they afford to researchers.

Do the spectrum and short history of network inference successes live up to such high hopes? Along these lines, we have argued that reverse-engineering over the past two decades has played at least five distinct research roles – the acceleration of hypothesis generation and verification at the single-node/single-interaction level, the illumination of statistical properties that render biological networks unique among complex systems, the diagnosis of individual networks as either typical or perturbed (paralleled by the use of within-class variation to make theoretical statements about the system), the prediction of how the activities in a given network will respond to exogenous perturbations, and the compact encoding of joint probability distributions – that go far beyond the trivial task of piecing together which of a set of observed system elements engage in physical contacts or the transfer of biologically relevant information. The roles we have identified represent a far cry from the (three) uses of effective influence networks – identification of functional modules, probing the response to perturbations, and helping determine the underlying mechanistic interactions – named by the authors of Ref. [64] ten years ago.

While it is impossible to say which of recent attempts to use networks as compressed "statistics" to help make (quantitative or qualitative) predictions will have the biggest impact down the road, it is clear that new precedents for the prediction of drugs targets and systemic responses in network medicine [133] point to a significant departure from the more traditional, reductionist ways of thinking. The consequences here will almost certainly include dramatic impacts on the ways medicine is practiced in the lifetime of the reader. With this example in mind, we reiterate our assertion that reverse-engineering yields its most succulent fruit when it is used to augment other methods of expanding our understanding of how living systems work, rather than employed disposably as an end goal in itself. Indeed, changes in the ways network inference has been used over time seem to be in accordance with this sentiment: whereas in 2003 the field was still firmly entrenched in its "pattern-detection phase" [366] (to better understand the state of the art at that time, we recommend [1]), it was around the time of publication of [35] in 2009 that the DREAM4 Challenges first introduced predictive modeling tasks as part of the main competition.

Indeed, the DREAM competitions play a unique part in the reverse-engineering culture. They not only echo changes in the field's priorities but also inform them: they have helped set the precedent in establishing inferred networks as tools for making predictions (as in the DREAM8 prompt to anticipate the responses of cellular signals to yet-unseen perturbations [367]). More radically, some of the most recent Challenges go as far as skipping the hitherto-canonical intermediate step of network inference entirely, asking competitors to infer macroscopic properties or outcomes using wholly different types of data [368]. While we clearly do not advocate for the complete abandonment of automated, network-scale reverse-engineering from large data sets, we do view the foundation's decreasing reliance on methods which require the construction of a detailed microscopic model prior to making inference about the macroscopic system as a progressive step. In fact, we contend that, given suitable alternatives, whole-network reverse-engineering may not be justified in every case.

If the reverse-engineering of entire microscopic networks is not always the right tool for the job, what might be done instead? As a starting point, we suggest asking:

- Given a reverse-engineered network, can we find any further compressions of that network that still preserve information about (i. e., are equally good at predicting) the macroscopic properties and observables it encodes?

- Can we identify any coarse *functional units* (perhaps with their own set of interaction rules and dynamics) that might supplant individual nodes and edges as the elements of a common parlance for the study of living systems?

For instance, might more appropriate "parts lists" for biological systems consist not of individual species' activations, but of larger physical or conceptual elements (e. g., negative feedback loops and operons) with their own dynamical interaction laws? Alternatively, *attractors* of the dynamics of biological networks may serve as more laconic descriptors of the networks than interactions among the nodes themselves [369, 51]. These possibilities may also be motivated via historical analogy: *renormalization group* theory in physics [370] has offered a systematic way to deduce an appropriate new vocabulary (and the corresponding syntax) when one changes the physical scale at which a system is to be observed. The effective interaction rules which emerge (say, the interactions between groups of Ising spins) are not always easily reducible to the familiar dynamics of microscopic activation variables (the nearest-neighbor interactions associated with individual spins), but which nonetheless account accurately for their effects at the new scale.

A recent line of work, inspired directly by statistical physics, formalizes the argument that only a small subset of parameter combinations are easily learnable from data, and therefore that only certain (combinations of) microscopic parameters can be *relevant* in determining a complex system's macroscopic or emergent properties [371, 19, 372, 373]. By systematically integrating out "sloppy" parameters or parameter combinations, whose values remain relatively unconstrained, one can assemble coarse, parsimonious models in terms of the remaining "stiff" parameters that serve as effective, low-dimensional compressions of a system's microscopic statistics.

Answers to the second question – that of finding higher-level explanatory structures in terms of which system's behavior can be understood – have been explored since the inception of "module-based" inference [151, 154]. In fact, newer and more powerful tools have sparked a resurgence [374, 375, 376, 146] of this approach. Around the same time, it was demonstrated that the flow of information in development, from promoter sequence to expression, can be reliably understood in terms of coarse, multiple-sequence patterns called graph-mers [377] that encompass entire sequence motifs. Ultimately, we believe that it will be work in directions such as these, which involve gross reconceptualizations regarding the fundamental actors in the biological dynamics, that will supersede whole-network reverse-engineering.

If the end goal of emulating physics-style modeling is prediction, the penultimate is certainly intuition and conceptual understanding. We entertain *phenomenological* approaches like those which focus on attractor dynamics (Chapter 3) and renormalization (Chapter 4) because they promise to yield interpretable models, not intractably large sets of detailed equations. Yet we still stress that, while searching for modularity and simple descriptions entails an invocation of the engineering *mindset* that has informed systems biology since its inception, the principles of good biological design often differ markedly from what works in that context; an open mind is necessary to dream up fitting new constructs. Whatever the case, we are confident that it is only by focusing on phenomenological (rather than microscopic) accuracy that we can deliver a satisfying confutational blow to famous Rutherford's quip that "all sciences are either physics or stamp collecting" [378] and begin removing the major impediments to the advancement of formal theories in biology [12].

Chapter 3

Precise Spatial Memory in Local Random Networks

While developing the background on biological network models, their inference, and the diverse categories of predictive roles in which they have served, we paused briefly from discussing their typical use – recording the relationships between *microscopic* activity variables – to reflect on the idea of using the same kinds of models to describe the relationships between higher-level, phenomenological variables.

We concluded our reverse-engineering overview in Chapter 2 with an important observation: in order to maximally benefit from the network abstraction, we may need (counterintuitively) to discourage wholesale network inference *for its own sake* and focus instead on how to efficiently represent the underlying systems. For situations in which the objects of interest are predictions regarding the dynamical behavior of a system, or a subset of its activity variables, we noted that one way to do this would involve the *attractors* associated with the system's dynamics. The example we pursue here, in Chapter 3, is inspired by computational models of spatial working memory.

Self-sustained, elevated neuronal activity persisting on time scales of ten seconds or longer is thought to be vital for aspects of working memory, including brain representations of real space. The paradigm of continuous-attractor neural networks [379, 380], one of the most well-known modeling frameworks for persistent activity, have been able to model several crucial aspects of such spatial memory. Many of these models tend to require highly regular or structured synaptic architectures. In contrast, here we study a geometrically-embedded network model with a local but otherwise *random* connectivity profile which, when combined with a global regulation of the system's firing rate, produces localized, finely spaced discrete attractors that effectively span a 2D manifold. The main idea is that, although the random network has no obvious compression or simplifying features beyond its sparseness, it nonetheless exhibits a surprisingly low-dimensional dynamical input-output relation with few attractors.

Specifically, we demonstrate how the set of attracting states can reliably encode a succinct representation of the spatial locations at which the system receives external input, accomplishing spatial memory via attractor dynamics despite the lack of explicit fine-tuning or simplifying symmetries at the level of its (synaptic) interaction architecture. We measure the network's storage capacity and find that the retrievable positions are nearly equivalent to a full tiling of the plane, something typically achievable only with translationally invariant neuronal connections. Thus, despite emitting what would seem to be a complicated series of activity measurements – and, presumably, a rather complex effective network structure, according to many of the inference methods introduced in Chapter 2 – the system admits a coarse description.

> *The following, written under the supervision of H. George E. Hentschel and Ilya Nemenman, has previously appeared as the electronic pre-print* Natale, J.L., et al. "Precise Spatial Memory in Local Random Networks." arXiv preprint arXiv:1911.06921 (2019). *It is also accessible via BioRXiv, with pre-print identifier* 10.1101/845156. *It is currently in revision for publication in the American Physical Society journal* Physical Review E.

3.1 Introduction

Biological implementations of working memory bridge the gap between two fundamentally disparate time scales: single neurons process information in $\sim 10^{-3}$s, whereas organisms interact with their external environments over durations of $\sim$ 1s or longer. For species from fruit flies to primates, this extension of time scales is reflected at the neural level by elevated spiking activity that persists while a particular memory is being accessed [381].

These excitations tend to be highly localized: for various types of working memory tasks across brain regions, firing rates for only a subset of selectively receptive neurons appear to become elevated [382, 383, 384, 385]. Traditionally, these units are considered to be responsible for maintaining the memory, and their so-called *persistent activity*, which can last anywhere from tens of seconds to several minutes, is thought to underlie a multitude of well-studied neural computations [386] (see Ref. [387] for an alternative viewpoint). While the mechanistic drivers of persistent activity are not fully understood – both single-cell and network-level explanations have been proposed over the last several decades, but their relative contributions remain under debate [388] – attractor neural network models have provided phenomenological descriptions of persistent firing states as fixed points or stable manifolds of the neural dynamics [389, 390, 391].

Attractor neural networks were first developed within the context of discrete, long-term associative memory, where each attracting state in a multistable system represented a distinct, stored memory [392]. Continuous-valued variants have since been able to model transient memories, like the firing activity responsible for maintaining an animal's eye position between saccades in one dimension [389] or its heading direction in a 2D environment [393]. To be useful in this context, attractor networks must typically incorporate highly structured or precisely tuned connection topologies. For instance, the synaptic connectivity matrices in Ref. [389] satisfy stringent spectral tuning properties that allow certain firing patterns to persist indefinitely. This need for nontrivial structure is quite general: it allows models of persistent activity to ensure the requisite balance between excitation and inhibition, which in turn renders a circuit capable of memory [388].

Recently, a biological instance of continuous attractor dynamics was traced to a circuit in *Drosophila* that respects one version of these topological constraints [394]. It has been suggested that the fly computation derives from high-level network properties – topological configuration, local excitation, and long-range inhibition – rather than "fine-scale" details like synaptic weights [395]. Yet it is not clear that networks with *random* weights, or unstructured connectivities, can perform similar computations. Indeed, random excitatory-inhibitory networks have been shown to be capable of various complex computations, including conjunctive encoding for input classification [396] and, in the balanced case, emergent selectivity in the context of evidence integration tasks [397].

In this article, we ask how well a minimally structured, randomly weighted network model can perform a spatial memory task of the kind previously thought [398] to need tuned, regular topologies. To do this, we study the firing-rate dynamics of a system with local but otherwise random connections. The network is spatially extended, and we show that it is able to encode the locations of external stimuli as elevated firing activity in the region near stimulation. In other words, it is capable of spatial memory. We introduce this system in Section 3.2, and computationally measure its capacity for distinguishing different stimulation locations in Section 3.3. We conclude by discussing how the model relates to previous work, and how it might be extended, in Section 3.4. Our intent is not to model any specific biological system, but to demonstrate how computations similar to those of persistent, continuous attractors are theoretically possible in random networks whose overall excitation and inhibition are balanced at a global (not single-neuron [399, 400]) level.

3.2 Model and Methods

The network $\mathcal{G} = (\{i\}, \{J_{ij}\})$ consists of N excitatory rate neurons [389, 401], embedded on a two-dimensional manifold [402]. Specifically, we consider a square plane of side length L, and connections $\{J_{ij}\}$ pointing from neuron j to neighbor i ($i, j = 1 \dots N$). We choose a set of spatial point coordinates $X = \{(x_1, y_1), ..., (x_N, y_N)\}$, where each pair $\vec{x}_i = (x_i, y_i)$ is an independent random sample from the bivariate uniform distribution on the interval $[0, L]$. This system has uniform spatial density $\sigma = \frac{N}{L^2}$, which is equivalent to $\frac{L}{\sqrt{N}} \equiv \lambda$ as the average inter-neuron separation.

With matrix elements $\{d_{ij}\}$ representing the Euclidean distances between neurons i and j, we assign a nonzero value to the synapse strength J_{ij} if $d_{ij} < \xi$, where $\xi \ll L$. We prohibit autapses, or self-loops, and invoke periodic boundary conditions in the calculation of d_{ij}. For convenience and uniformity, we present all results using the reference plane $[0, L] \times [0, L]$. In all that follows, $L = 1$ and $\xi = 0.06L$ unless otherwise specified. We also choose $N = 2^{12}$, which fixes $\lambda \approx 0.016(L) \approx 0.26\xi$.

Choosing a value for ξ which is small relative to L ensures that connections remain short-ranged, and that the resulting network is sparse. We argue later that choosing a set of connections $\{J_{ij}\}$ that is too short or too long-ranged diminishes the ability of the network to support multiple nontrivial memory states. Quantitatively, since each neuron i interacts with $\sim \pi\xi^2\sigma$ downstream neighbors, a typical network realization $\mathcal{G}$ encompasses $\sim \pi N^2 \left(\xi/L\right)^2$ synapses, or about 1% of all possible connections.

The connection strengths, or synaptic efficiacies, are

$$J_{ij} = \begin{cases} \sim P(\mu, \sigma), & d_{ij} < \xi \text{ and } j \neq i, \\ 0, & d_{ij} \geq \xi \text{ or } j = i, \end{cases} \tag{3.1}$$

where each J_{ij} is an independent draw from $P(\mu, \sigma)$, representing a lognormal distribution (as argued for in Ref. [403] and elsewhere; we explored other distributions, but found no qualitative differences in the results). Since by definition lognormal random variables are positive definite, $J_{ij} > 0$ for all outgoing connections: all neurons are excitatory. In what follows, $\mu = -0.702$ and $\sigma = 0.8752$ (by convention, these parameters refer to the associated normal distribution). These values were taken from fits done during experimental investigations of neural circuit properties in the rat visual cortex [403].

As emphasized above, persistent activity typically demands a fine balance between excitation and inhibition, while our connectivities encompass no explicit inhibition. Therefore, we choose to model inhibition indirectly, imposing its main effect – which we assume is to stabilize the system's total firing activity to a constant value [404, 405] – directly. In particular, we insert a term into the usual nonlinear firing-rate equations [389, 401] to represent nonlocal inhibitory interactions. In summary, in the absence of synaptic or external inputs, the firing-rate activity $r_i(t)$ decays exponentially over the intrinsic time scale τ. Otherwise, $r_i(t + dt)$ is determined by integrating a nonlinear function of combined input currents $\sum_j J_{ij} r_j(t)$ the from upstream neighbors j and external drive $I_i(t)$ over the short interval $dt \ll \tau$. Thus, for constant $a > 0$,

$$\tau \frac{dr_i}{dt} = -r_i + aN \left(\frac{h_i}{\sum_j h_j} \right), \tag{3.2}$$

$$h_i = f\left(\sum_j J_{ij} r_j + I_i(t) \right). \tag{3.3}$$

This system will ultimately approach a steady state for which $\sum_i r_i(t \gg \tau) = aN$: global inhibitory interactions, implemented by the second, "activation," term in Eq. (3.2), create the desired balance. This can be verified by solving for the steady-state conditions $\frac{dr_i}{dt} = 0$. The parameter a in Eqs. (3.2-3.3) can be thought of as the system's baseline firing level (the rate at which all neurons would fire if they were to fire at equal rates in the steady state). A complementary interpretation, related to the fraction of active cells in the steady state, will be addressed in detail later. We set $a = 0.02$ and, without loss of generality, choose $\tau = 1$ so that time is measured in unit of τ.

Finally, we adopt for the nonlinearity a version of the firing-rate function introduced by Ref. [393],

$$f(x) = \alpha \cdot \left\{ \ln \left[1 + \ln \left(1 + e^{\beta(x-\gamma)} \right) \right] \right\}^{\delta}, \tag{3.4}$$

with $\alpha = 18$, $\beta = 0.5$, $\gamma = 16$, and $\delta = 1.5$. We selected these values to place activations $\{h_i\}$ in a biological range (tens or less, if measured in Hz) for arguments $x > 0$ spanning two orders of magnitude, with $f(0) \sim 10^{-4} \approx 0$. The reason for the choice given by Eq. (3.4) is that the gain of this curve increases at a value away from zero, and that its behavior in the limit of large inputs is nonsaturating over two orders of magnitude in x. These attributes are intended to better approximate the biological reality [406], as compared with the sigmoidal thresholding functions commonly used in artificial networks (which tend to feature inflection points near values corresponding to zero net input). We note that both of these properties are also satisfied by the ReLu (Rectified Linear unit) activation function [407], also commonly used in machine learning.

For a realization $\mathcal{G}$ with dynamics given by Eqs. (3.2-3.3), we would like to quantify how this system performs as a spatial memory architecture. In particular, if a group of neurons local to an arbitrary region of the plane is stimulated externally, can the system sustain a persistent representation of their coordinates? How many distinct stimulation sites can the system reliably encode?

To measure the number of resolvable sites, we perform n_{trials} "external stimulation" computational experiments, sequentially, in Matlab. First, we initialize the system, creating a network realization $\mathcal{G}$ by selecting values for the neuron positions X and connection strengths $\{J_{ij}\}$. We then set the firing rates of all neurons $i = 1 \ldots N$ to $r_i(0) = a$ and evolve Eqs. (3.2-3.3) from $t = 0$ to $t = 100\tau$, well beyond the point at which the individual firing rates stabilize, using the built-in Runge-Kutta (4,5) solver with $I_i(t) = 0$. The result can be a strong excitation, confined to a local region of the plane, or a fully *delocalized* firing state in which all neurons participate with rates near a. In either case, the rates do not change in time (this holds even if the system is initialized randomly, with rates that sum to the steady-state value aN, instead of uniformly).

To ensure that the system can switch out of this state, we perform a single external stimulation, abitrarily targeting the visual center of the plane, according to the following protocol. With the aformentioned state serving as our initial condition, we locate all neurons contained within an "input" patch of area $\pi\rho^2$ (for now, we choose $\rho = \xi = 0.06L$) centered at $\vec{x}_{\text{stim}} = (0.5, 0.5)$. For this subset of system elements only, we set

$$I_i(t) = A\,(1 - \Theta(t - \Delta t)) = \begin{cases} A, & t < \Delta t, \\ 0, & t \geq \Delta t, \end{cases} \tag{3.5}$$

where $\Theta(t)$ denotes the Heaviside step function, and $\Delta t = 5\tau$. We again solve Eqs. (3.2-3.3), integrating until $T = 40\tau$, sufficient time for the network to reach a persistent state.

We then repeat this protocol for n_{trials} iterations, each time sampling a random position $\vec{x}_{\text{stim}} = (x_{\text{stim}}, y_{\text{stim}})$ from a uniform grid of 10^4 finely-spaced points superimposed on the plane (that is, separated by $dL = 10^{-2}L$), to serve as the set of stimulation centers. The resulting state $\{r_i(t = T)\}$ then becomes the new initial condition for the following trial, representing re-stimulation and new memory formation. We set $n_{\text{trials}} = k \cdot (L/dL)^2$, partitioning stimulations into k successive groups of $(L/dL)^2$ trials that are each composed of independent random permutations of the full list of available gridpoints $\{x_{\text{stim}}\}$.

3.3 Results

3.3.1 Network supports multiple stable attractors

Upon stimulation, the system initialized as above tends to develop a localized excitation in the vicinity of $\vec{x}_{\text{stim}}$, which quickly coalesces into a roughly circular "bump" of activity [408, 391, 409]. Figure 3.1 depicts a representative bump in a system of size $N = 2^{12}$ at $T = 40\tau$. The inset reproduces the firing-rate trajectories for $t \leq T$, showing that all rates have stabilized to their final values by T.

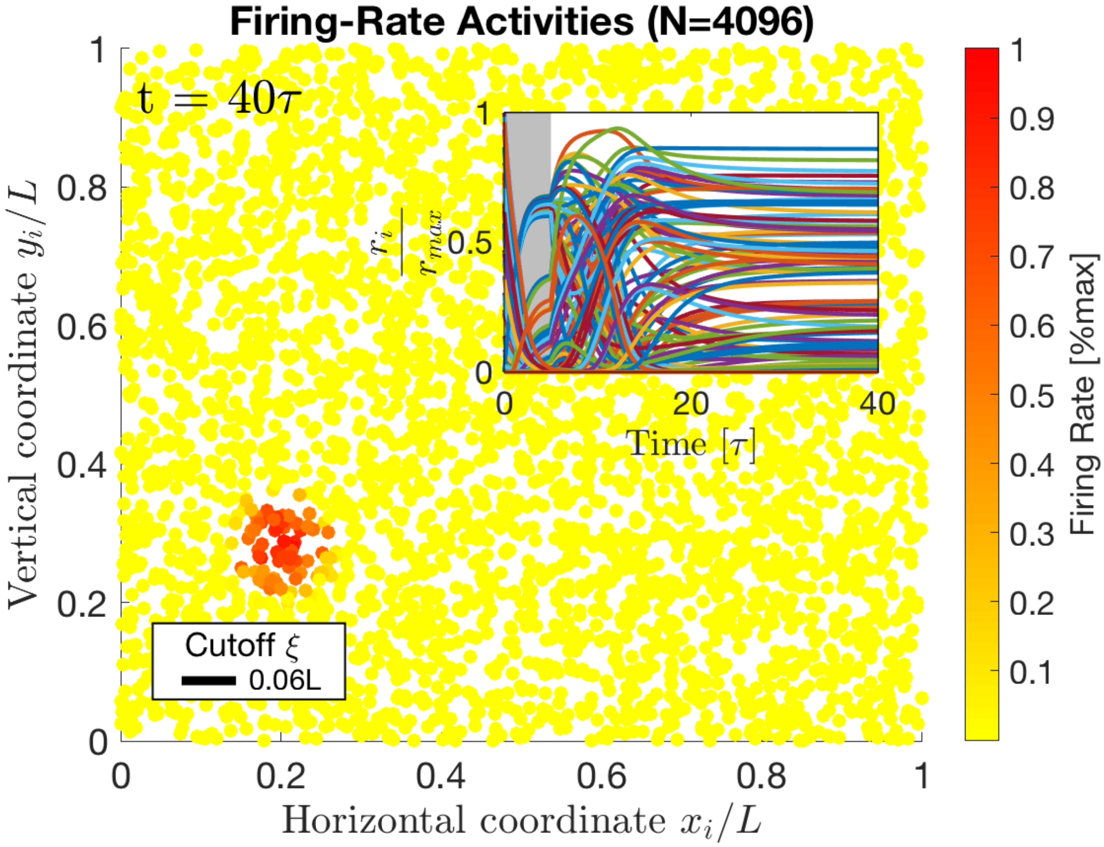

Figure 3.1: Sample bump state in a system with $N = 2^{12}$. The scale bar indicates the synaptic cutoff distance ξ, below which $\mathcal{G}$ appears fully connected. *Inset*: All the neural activities through time. Most of the trajectories remain near zero, and cannot be visually distinguished. Stimulation is shown as a gray block of width $\Delta t = 5\tau$.

While it is free to migrate or spread about $\vec{x}_{\text{stim}}$ during and after stimulation, this activity bump typically assumes a stable shape and location on the plane by the same time T. Analogous behaviors are observed when the system is stimulated from within a previously activated stable state. Then, activities associated with any preexisting bump are rapidly attenuated due to the global inhibition, typically returning to baseline activity values by Δt. Generally, given a sufficiently strong input current amplitude A and adequately long stimulation time Δt, an activity bump will form in any general region of the plane and remain thereafter in the vicinity of $\vec{x}_{\text{stim}}$.

In simulation, our model seems to support only one spatially localized excitation under steady-state conditions, even if stimulated briefly at two locations simultaneously. At least qualitatively, this might be understood by analogy with a simpler system consisting of just two units, representing distant regions of strong firing. If each unit acts according to Eqs. (3.2-3.3) – loosely, as a self-excitatory, positive-feedback system, with a global inhibition that enters via the normalization $h_i / \sum_j h_j$ – it is easy to imagine that their mutual feedback will lead to a single unit dominating (we ignore oscillations, since the feedback would need to be precisely tuned in order for these to appear). While it is not immediately clear from these equations that simultaneous activation at many locations will not lead inevitably to delocalized excitations or multiple small bumps, we are not focused on this here, precisely because we are interested in situations for which there is exactly one driving input at any given moment in time – and only one recent memory, as in the experimental system of Ref. [394]. Thus, as a rule of thumb, we say that the system supports a single bump at any given time [394], in any general spatial region of the plane.

How large are these activity bumps? Although they are not perfectly circular, we observe that excitations do take on a typical size for a fixed cutoff distance ξ. We can therefore speak about an effective bump radius R_{eff}. A simple way to measure R_{eff} would be to choose a firing-rate threshold above which neurons will be considered *active*, and compute the radius for the equivalent circular area πR_{eff}^2 occupied by this subset of system elements on the plane. Ideally, though, we would like to choose a criterion that is relatively insensitive to the cutoff distance. Fitting two-dimensional Gaussian curves to the spatial firing-rate distributions associated with each bump and measuring $2R_{\text{eff}}$ as the full width at half maximum, as done recently for the experimental system of Ref. [394], yields $R_{\text{eff}}(\xi = 0.06L) \approx 0.78\xi \approx 0.05L$. In other words, the bump radius is on the order of the cutoff distance. We expect this to be a generic result.

Taking the ratio $\frac{R_{\text{eff}}}{\lambda} \approx 2.99$, we see that typical activity bumps are also large in comparison with the inter-neuron separation λ, as well as the distance $dL = 10^{-2}L$ between adjacent gridpoints. This has an important consequence. If the system is stimulated at a point within (or too near) the area associated with an active bump, it may revert to the originally active bump state instead of evoking a new memory. This is particularly true if either the input time Δt or amplitude A are insufficiently large, but can occur more generally due to the fact that our random connectivity matrix lacks precise translational symmetry. This allows certain bumps to emerge as preferred states, which are more strongly favored than others (this limits the network representational capacity, as we determine quantitatively later). Nevertheless, the system does appear to select from a discrete, finite set of constant firing-rate states for the parameter values ($\lambda = N^{-\frac{1}{2}}$, $\xi \approx 3.84\lambda$) defined above.

In summary, for sufficiently strong input, we observe:

1. Local stimulation can cause the system to develop stable bumps in essentially any region of the plane;

2. The system seems able to transition, smoothly and repeatably, from sustaining one bump state to another (switch between multistable firing patterns);

3. Independent stimulations centered at different gridpoints can result in nearly indistinguishable memory bumps.

We take these observations together as the earmarks of dynamical attracting behavior – in particular, the system acts as a discrete approximation to a 2D plane attractor. We identify each achievable bump state with a stored, retrievable memory. By definition, an attracting state persists until stimulation evokes a new bump, so we say that the system stores *spatial memories* encoding the location at which it was most recently stimulated.

Since the basins of attraction (from within which stimulation at different $\vec{x}_{\mathrm{stim}}$ values consistently leads to the activation of specific memories) are not infinitely small but instead appear finite, the system cannot remember arbitrary positions on the plane. It is then natural to ask how many *unique* spatial locations can be distinguished by a given realization of the synaptic structure. That is, the resolution with which $\vec{x}_{\mathrm{stim}}$ can be decoded requires quantification.

3.3.2 Spatial memories span the entire plane

How many distinct stimulation locations $\vec{x}_{\mathrm{stim}}$ might we anticipate a realization $\mathcal{G}$ to resolve? We expect this *capacity* to depend largely on gross statistics like the average size of the attracting basins, rather than on details of the instantial arrangement of neuron positions and synaptic connections associated with a given system configuration.

Since the dynamical equations (3.2-3.3) are deterministic, the attracting state evoked by stimulation at a given site should be unique, apart from the aforementioned dependencies on the initial state and input-current parameters. This variation can even be minimized: the stronger the external inputs, the more reliably we can anticipate that the system will find an attractor in the vicinity of the stimulation location, independent of where it is currently excited. Thus all that remains to determine the exact set of attractors supported by a given configuration $\mathcal{G}$ are the the coupling strengths. Accordingly, we expect that the bumps to which excitations attract will be almost exclusively a function of the (quenched) random variable J_{ij}.

We coarsely estimate the system's capacity as follows. Assuming homogeneous basins of attraction and one-to-one retrieval within a basin, the number of reliably stored memories will be equal to the number of basins that fit on the plane. Dividing the $L \times L$ space into equally-sized square sections of width $2R_{\text{eff}}^{-2}$ implies, for our parameter values, $\sim 10^2$ distinct, nonoverlapping basins that span the 2D space. Thus our baseline will be ~ 100 bumps, touching tangentially.

A preliminary step towards more accurately quantifying the number of stimulation locations that the system can reliably encode is simply enumerating all the unique attractors activated during a given series of n_{trials} stimulations. This allows us to conceptualize the capacity in terms of input (stimulation site) to output (bump location) relations. For each stimulation, we track the *center of excitation* $\vec{x}_{\text{COE}}(t) = \sum_{i'} \frac{r_{i'}(t)\vec{x}_{i'}}{aN}$ among cells i' which we identify as actively participating. Instead of accommodating for the uncertainties associated with Gaussian fits, here we employ simple thresholding to identify active units, for two principal reasons. First, even the fixed-threshold criterion $r_i > 10a$ predicts the number of active neurons to within 10 units of the amount given by the *participation number* $p_\nu = (\sum_{i=1}^{N} r_i^\nu)^2 / \sum_{i=1}^{N} r_i^{2\nu}$, and it exhibits similar qualitative behavior across the surprisingly large range of cutoffs from roughly zero to 10λ. In addition, this criterion was found to predict coordinates for the excitations that coincide well with the measured Gaussian peaks.

For large cutoffs, it is possible that even a fairly nonrestrictive threshold can exclude relatively strongly firing neurons: our constraint $\sum_i r_i(t) = a$ implies that firing activity within a given bump decreases as bumps increase in size, which is precisely what we observed to happen as we increase ξ. Excitations encompassing zero active neurons were to be assigned a special value of $\vec{x}_{\text{COE}}(t)$, allowing us to count them separately toward the capacity, but this was not observed for the $\xi = 0.06L$ presented below. We enumerate all distinct bumps by counting the unique values of $\vec{x}_{\text{COE}}(T)$ observed, to within a specific resolution (we discuss the importance of this resolution below). For n_{trials} large, this number should approach the cardinality of the set of possible memories. The next step will be to quantify how many – or with what fidelity – distinct values of the gridpoint coordinates $\vec{x}_{\text{stim}}$ can be discriminated by these enumerated attractors.

We measure the capacity for a given realization $\mathcal{G}$ as follows. Although each site in the set of $(L \cdot dL)^{-2} = 10^4$ available stimulation gridpoints is visited $\frac{n_{\text{trials}}}{L \cdot dL^{-2}} = k$ times each in each series of stimulation events, averaging over all possible initial conditions for each gridpoint would require too much time. Here we choose $k = 10$ to further mitigate finite-sampling errors due to the situation described above, in which stimulation near a highly active bump simply reverts the system back to that previous attractor after a transient. We also choose to work with an information-theoretic capacity metric, to treat the inherently nonuniform stochasticity associated with the "stimulus-response" records in a natural framework.

Specifically, we measure the mutual information [410] between random variables $\vec{x}_{\text{stim}}$ and $\vec{x}_{\text{COE}}(T)$ for a realization $\mathcal{G}$. To do this, we obtain the frequencies of occurrence for all observed stimulation locations $\{\vec{x}_{\text{stim}}\}$ and bump centers $\{\vec{x}_{\text{COE}}(T)\}$, over a set of n_{trials} stimulation events. We then use these frequencies as the maximum-likelihood estimates of the corresponding probabilities to form the "plug-in" or naïve estimators for the relevant entropies [411, 412, 413], from which we can calculate the mutual information $MI\left(\{\vec{x}_{\text{stim}}\}, \{\vec{x}_{\text{COE}}\right)(T)\}$. Since asking how many different attractors were observed for each stimulation position is equivalent to asking how many different stimulation positions lead to the same attractor (i.e., the mutual information is symmetric), we choose the latter. Finally, from the mutual information, we define the capacity

$$C = 2^{MI(\{\vec{x}_{\text{stim}}\},\{\vec{x}_{\text{COE}}(T)\})}. \tag{3.6}$$

Since the information is measured over discrete states, we must discretize the the values of $\vec{x}_{\text{COE}}(T)$ by rounding them to an appropriate resolution. As seen in Fig. 3.2, truncating $\vec{x}_{\mathrm{COE}}(T)$ to two decimal places still represents 87.5% of the maximum information, or ≈ 6.25 bits. Assuming that the system cannot track bump centers to a precision better than these two decimal places – roughly the theoretical separation between neurons – we arrive at $C \approx 76$ distinct stimulation regions for the values of L, λ, ξ and ρ used throughout.

In other words, on average, $\mathcal{G}$ is able to store and reliably retrieve a number of memories approximately equal to our naïve, baseline estimate. Unlike in that coarse estimation, we did not require bumps to be nonoverlapping in measuring the capacity – yet the system's recall ability turns out to be nearly as accurate as a fully deterministic discriminator that simply decides in which $R_{\mathrm{eff}} \times R_{\mathrm{eff}}$-sized, homogeneous division of the plane the last stimulation occurred. Thus the information-theoretic capacity, measured to two decimal digits precision in $\vec{x}_{\mathrm{COE}}(T)$, is also consistent with a typical size for the attracting basins which matches R_{eff} for stable bumps. Furthermore, we observe that the retrievable memories span more or less the entire spatial extent of the $L \times L$ plane. This can be readily observed in Fig. 3.3, which depicts the set $\{\vec{x}_{\mathrm{COE}}\}$ of unique bumps accounted for over a course of n_{trials} stimulations for one network realization.

3.3.3 Mutual information is near-optimal for a broad range of parameter values

The cutoff distance is an important length scale in the system. The structure of the network depends crucially on ξ, allowing us to go from completely unconnected neurons in the extreme of $\xi = 0$ to the fully-connected network for $\xi = L$. It is important to understand how ξ affects our main findings – in particular, the existence of localized excitations, and the number of memories $\mathcal{G}$ can support.

For the unconnected case $\xi = 0$, we have $\{J_{ij}\} = 0$. In the absence of recurrent connections (besides the implicit inhibition), all neurons respond independently to

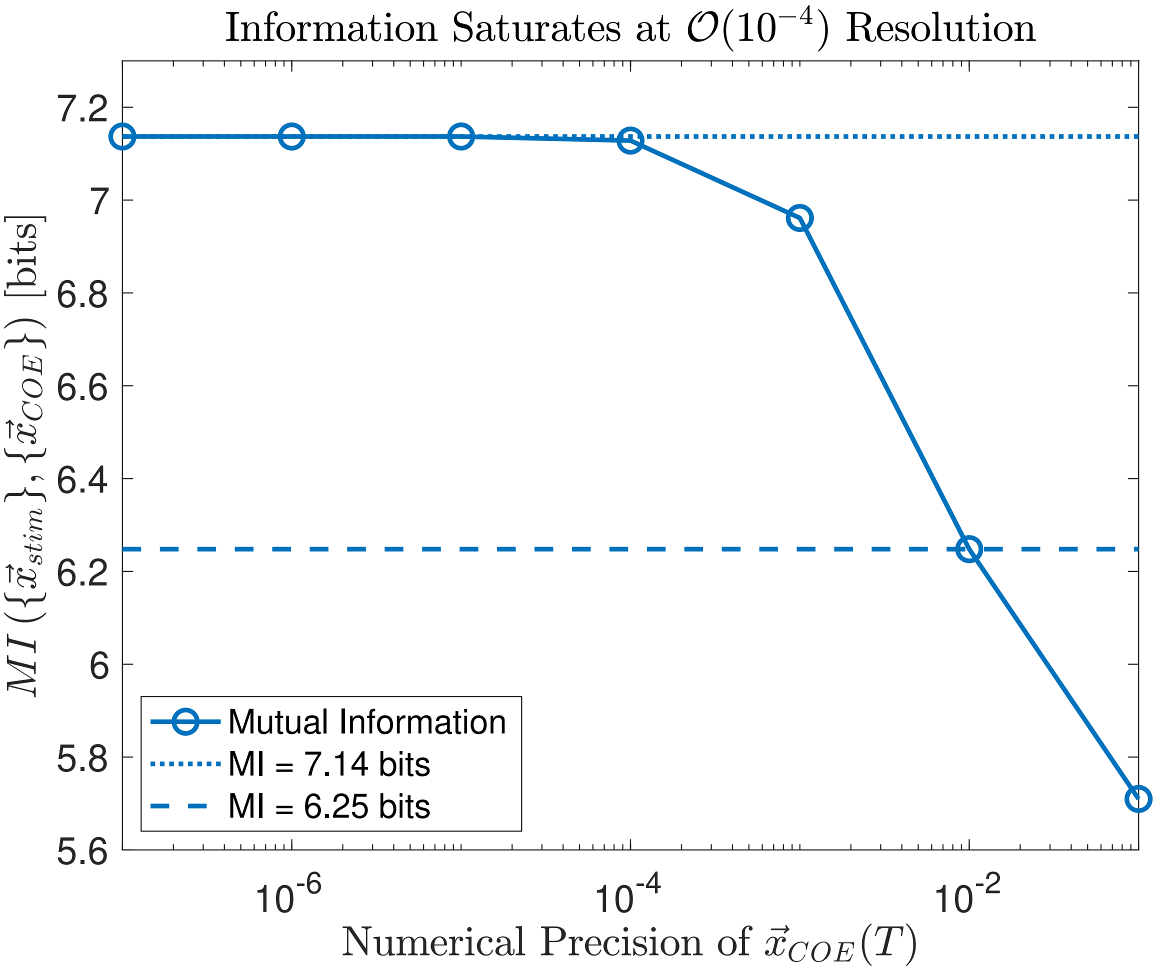

Figure 3.2: Mutual information as a function of rounding precision in the center-of-excitation values $\{\vec{x}_{\text{COE}}(T)\}$. Saturation occurs by four decimal places, but in what follows we keep two places to ensure the precision of $\vec{x}_{\text{COE}}$ is not finer than the inter-neuron separation λ. The changes the capacity by less than a factor of 2.

their respective external inputs $I_i(t)$: that is, the $\{r_i\}$ obey a simplified version of Eqs. (3.2-3.3). In order to write down the dynamics in this case, we first note that neurons outside the stimulation patch have activations $h_i = f(0) \approx 0$ for both $t < \Delta t$ and $t \geq \Delta t$. These units at first experience an exponential decay in their firing activities and then approach the steady-state value $r_i(t \gg \Delta t) = a$. The $\sim \pi N(\rho/L)^2$ neurons encompassed by the stimulation patch also approach a constant value. To show this, we note that each of the units in this latter subset sees the same input $h_i = f\left[A\left(1 - \Theta(t - \Delta t)\right)\right]$, so that the ratio $h_i(t)/\sum_j h_j(t)$ stays constant. Therefore

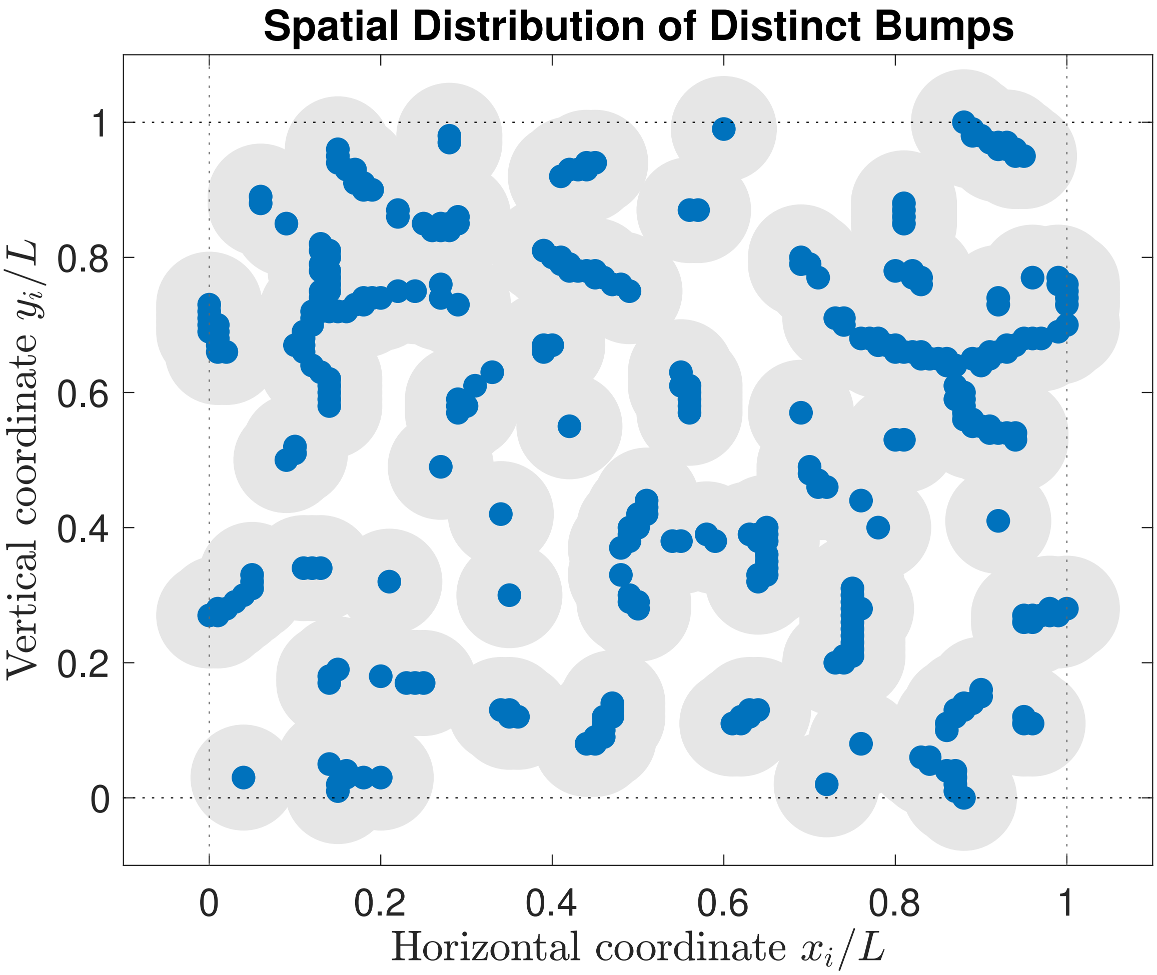

Figure 3.3: The different attracting bumps observed over the $k \cdot n_{\text{trials}}$ computational experiments are distributed in such a way that they span the majority of the plane. Bump centers are shown as blue dots; radii for their surrounding gray circles are $\approx R_{\text{eff}}$. Dotted lines are periodic boundaries.

we can remove the nonlinearities entirely and write

$$\frac{dr_i}{dt} = -r_i + I_i'(t), \tag{3.7}$$

$$I_i'(t) = \begin{cases} \frac{a}{\pi\rho^2}, & t < \Delta t, \\ a, & t \geq \Delta t. \end{cases} \tag{3.8}$$

Then, in the long-time limit, the unconnected system relaxes to the trivial stable state $\{r_i(t \gg \tau)\} = a$, in which all neurons fire at the same, baseline rate. It cannot sustain any excitations that can be decoded as memories. In the other extreme, $\xi \to L$, it seems unlikely that a fully-connected network can support any *localized* excitations.

We quantify the precise dependence of our findings on the value of the cutoff distance in Fig. 3.4. We generated this plot by progressively decreasing ξ for an initial, fully-connected realization $\mathcal{G}$. Here we chose $k = 1$, stimulating at the first 10^4 of the 10^5 sites used to generate Fig. 3.2, and rounded the measured information values to a precision of two decimal places in $\vec{x}_{\mathrm{COE}}$ as decided above. Clearly, the mutual information quickly drops to zero below the inter-neuron separation λ. This means that the system attains only states that are delocalized – effectively all neurons contribute to the excitation, but none exceed the threshold $r_i > 10a$ to be considered "active" – which we identify as the single, trivial state.

At the other extreme, the mutual information returns to zero for large values of ξ. This can be explained in terms of the circumstances discussed in Section 3.3.1, in which it becomes difficult for the network to switch out of its preferred states. As the cutoff distance increases above $\xi \approx 7\lambda$ (or $\approx 5\lambda$ for the stricter threshold of $r_i = 50a$), more neurons are directly involved in sustaining a given excitation, and the structure of the basins of attractions changes so as to accommodate fewer feasible memories. As in the case of insufficient stimulation time or amplitude, the success or failure of a given stimulation in evoking a nearby bump is somewhat history-dependent (in the sense that some memories might be retrievable from some initial states but not certain preferred states), but invariably the system comes to favor a single state in the limit that the network becomes fully connected. For the $10a$ threshold, the network cannot reliably store any spatial memories for roughly $\xi > 0.16L \approx 10\lambda$.

Between these two extremes, there is an optimal value $\xi^* \approx 0.02L$, for which the greatest number of stimulation gridpoints can be distinguished. Moreover, starting at this value, there is a plateau in the system's accuracy from roughly $\xi = 0.02L \ldots 0.11L \approx \lambda \ldots 7\lambda$, across which the mutual information varies by only ~ 1 bit. More precisely, the gap between the highest and lowest points on the $10a$-threshold curve of Fig. 3.4 corresponds to the difference between resolving $C \approx 156$ and $C \approx 72$ distinct stimulation sites. These values are of the same rough order of magnitude, and their average is nearly equal to our very first baseline estimate of 100 distinct, homogeneous basins. We note in particular that the cutoff distance $\xi = 0.06L$ used throughout the rest of the paper is nominally three times larger than ξ^*, but different by less than the aforementioned bit in terms of information.

In principle, the capacity should also depend on how reliably the system accesses its attractors for (or indeed, whether the set of accessible attractors changes with) different values of the size of the input patch, ρ. Figure 3.5 records the dependence of the mutual information on ρ. Outside this range, the system will attract to (possibly different) preferred states, but between roughly 2λ and 6λ we observe that the system attracts to the same bump state regardless of the specific value of ρ (not explicitly depicted). This gives the appearance that the system really is tracking the stimulation centers in computing its final states, at least for input patch sizes in this range.

To the extent that different proxies for $\vec{x}_{\text{COE}}$ agree, this suggests that the system does in fact encode a coarse representation of the stimulation location – the bump centers of excitations – rather than tracking high-dimensional quantities like the real-valued firing rates. That is, although an experimental system wired according to our prescription for $\{J_{ij}\}$ could indeed store information in individual firing rates for other purposes, we are not merely imposing but discovering that the low-dimensional summary variable $\vec{x}_{\text{COE}}$ is sufficient to predict the stimulation region to a considerable accuracy. Another step toward testing this hypothesis would be to systematically map the basins of attraction for a given realization $\mathcal{G}$, and check whether the steep decrease shown in Fig. 3.5 occurs when the stimulation patch grows large enough to extend into multiple basins besides that of the targeted memory.

Together, the above results suggest that our randomly-weighted network can sustain local excitations for a range of parameter values. In general, these excitations can serve reliably as spatial memories encoding the system's most recent stimulation location if the number of neurons activated via stimulation and local synaptic input is small relative to the system size N. This can be achieved by choosing ξ less than approximately $\mathcal{O}(10\lambda)$, which ensures that a given neuron synapses with anywhere from roughly $\pi(\lambda)^2\sigma \ldots \pi(10\lambda)^2\sigma \approx 10^0 \ldots 10^2$ neighbors.

3.4 Discussion

We have showed that short-range, but otherwise unstructured connectivities can support spatial memory via persistent firing if the overall activity of the network is constrained through excitation-inhibition balance. The spatial regions that can be remembered (discriminated) with high-fidelity effectively tile our $L \times L$ planar section, with a resolution of $\mathcal{O}(\lambda^{-1})$ distinct sites, roughly equivalent to the number of nonoverlapping memories that span the same area. This performance corresponds to an information-theoretic capacity that scales as $C \propto \sqrt{N}/L$, or $C \propto \sqrt{\sigma}$ in terms of the neuron density, which can be checked experimentally by testing larger system sizes.

Since the inter-neuron separation sets the scale of the problem at the outset, it is not necessarily surprising that the optimal cutoff distance $\xi^* \approx \lambda$. What is unexpected in our results is the fact that a spatial memory spanning a two-dimensional manifold can be achieved without explicit tuning of synaptic connections. This is reinforced by the fact that we observe not just an isolated peak at ξ^*, but a broad plateau of near-optimal cutoff distances.

While it is traditionally maintained [388] that only tuned connectivity profiles can produce continuous attractors, the idea that random networks support memory on short time scales is not altogether new [414, 415, 416, 398]. Indeed, recent work argues that quasi-random topologies, refined via a non-linear Hebbian learning rule, can give rise to attractor dynamics in the specific context of persistent neural activity as a substrate for working memory [417]. Here, we are interested in using such random networks to store spatial memories that effectively span a continuous manifold [418]. In addition, we accomplish spatial memory using a random network, which emphatically requires no learning.

Similarly, distance-dependent topologies [419] have been implemented in previous models, including the seminal work on continuous neural attractors [408], yet we are aware of only two related studies that link sparse, short-range (1D nearest-neighbor) connections formally to the localization of firing-rate excitations [420, 421]. As we do, both respect Dale's Principle [422] for the signs of synaptic connections only indirectly [423] and explore random weights. While it may be interesting to explore the spectra of our $\{J_{ij}\}$ in the context of Anderson localization or the notion of "spatially structured" disorder developed in [421], a more obvious generalization of our model would be to relax the hard-threshold cutoff condition to a connection probability. For example, we could set $J_{ij} \propto e^{-|\vec{x}_i - \vec{x}_j|/\xi}$, or another function of $d_{ij} = \|\vec{x}_i - \vec{x}_j\|$ (see, for example, the related work of Refs. [424, 425, 426, 427]).

A drawback to our model, in the form presented here, is that the system of Eqs. (3.1-3.5) incorporates no explicit noise terms. Fundamental to our results is the firing-rate constraint $\sum_i r_i(t) = aN$, an imposition which corresponds only approximately to the biological reality for real circuits (as in [394]). In our future work, we propose to replace the constant parameter a by a Gaussian process $\alpha(t) = a + \eta(t)$. We expect that, for small amounts of noise, the system will retain its qualitative behavior, but with a reduced capacity. On the other hand, for $\eta(t)$ with large variance, it is possible that the system will fail to store memories with high fidelity due to longer bump excursions or delocalization, or entirely as with ξ and ρ.

If these assumptions regarding the inclusion of noise are found to hold, it would be interesting to explore noise parameters that place the firing-rate variability in a regime consistent with previous experiments [406, 388] while respecting our sparsity constraints. Yet we reiterate that our goal is not to model any known experimental system. Indeed, whether or not our model relates to specific, observable experimental systems remains to be seen. In anticipation of such *in vivo* analogs, we offer the following predictions regarding which features of our model might be used to infer whether short-range, randomly weighted connections drive a given instance of persistent activity.

First, in the best case scenario, novel technologies may allow researchers to probe structural properties directly. This promises a trivial way of checking whether synaptic matrices are untuned, as in Eq. (3.1), and is already underway for the fly [428, 33, 429]. While the emerging picture for *Drosophila* is one of decidedly nonrandom connectivity, this may not hold for significantly larger organisms. Indeed, the number of possible synapses in a neural system scales as $\mathcal{O}(N^2)$. Thus genetic encoding of precise values for some billions of pairwise connections even in modestly sized vertebrates is simply not feasible. On the other hand, it is plausible that regularity appears at the level of local rules superimposed on essentially random connectivities, as in canonical microcircuit models [430], which would be consistent with our setup.

In the absence of structural information, the firing-rate activities themselves can also help support or reject our model. Since most classic continuous-attractor architectures have translationally invariant connections, they are able to host bumps at virtually any location [431]. Our $\{J_{ij}\}$, on the other hand, lack such a symmetry. This leads to discrete attractors [432] with variable spacing and portions of the plane that cannot be reliably encoded. Such "discrete approximations" to attracting manifolds have even been touted as more robust than their continuous counterparts, for example to perturbations in the synaptic weights [433]. It would be interesting to quantify the fraction or extent of the plane that the system can remember in the presence of the aforementioned noise.

In addition, while continuous attractor models accommodate a degree of drift or diffusion for activity bumps following their settlement upon the manifold [434], tracking $\vec{x}_{\rm COE}(t)$ reveals that excursions in our random networks occur predominantly *before* $t = T$; see the inset of Fig. 3.1. Thus, comparing the observed distribution of displacements, between the tested $\vec{x}_{\rm stim}$ values and the corresponding $\vec{x}_{\rm COE}(t)$ could also distinguish our model.

Finally, the raw activity measurements $\{r_i(t)\}$ are also subject to what is known as network reverse-engineering, or automated inference methods that operate directly on data to reconstruct network interaction structures [37]. Although we do not advocate applying out-of-the-box algorithms to glean structural information in general, there do exist certain signatures and gross statistics which can be used to differentiate truly random graphs from more complex or subtle architectures at a coarse level [53].

Our model is one of many that attempt to capture the ability of different neural systems to support localized excitations that encode real-valued quantities. Here, we eschew structured topographic mappings [394] in favor of a random connectivity that we find to be capable of storing similar neural representations. Whether or not *in vivo* circuits conforming to the specifications of our model are found experimentally to underlie one of these interesting systems, in our view such random, balanced excitatory-inhibitory networks should still be taken seriously as null models for recurrent neural computation [397].

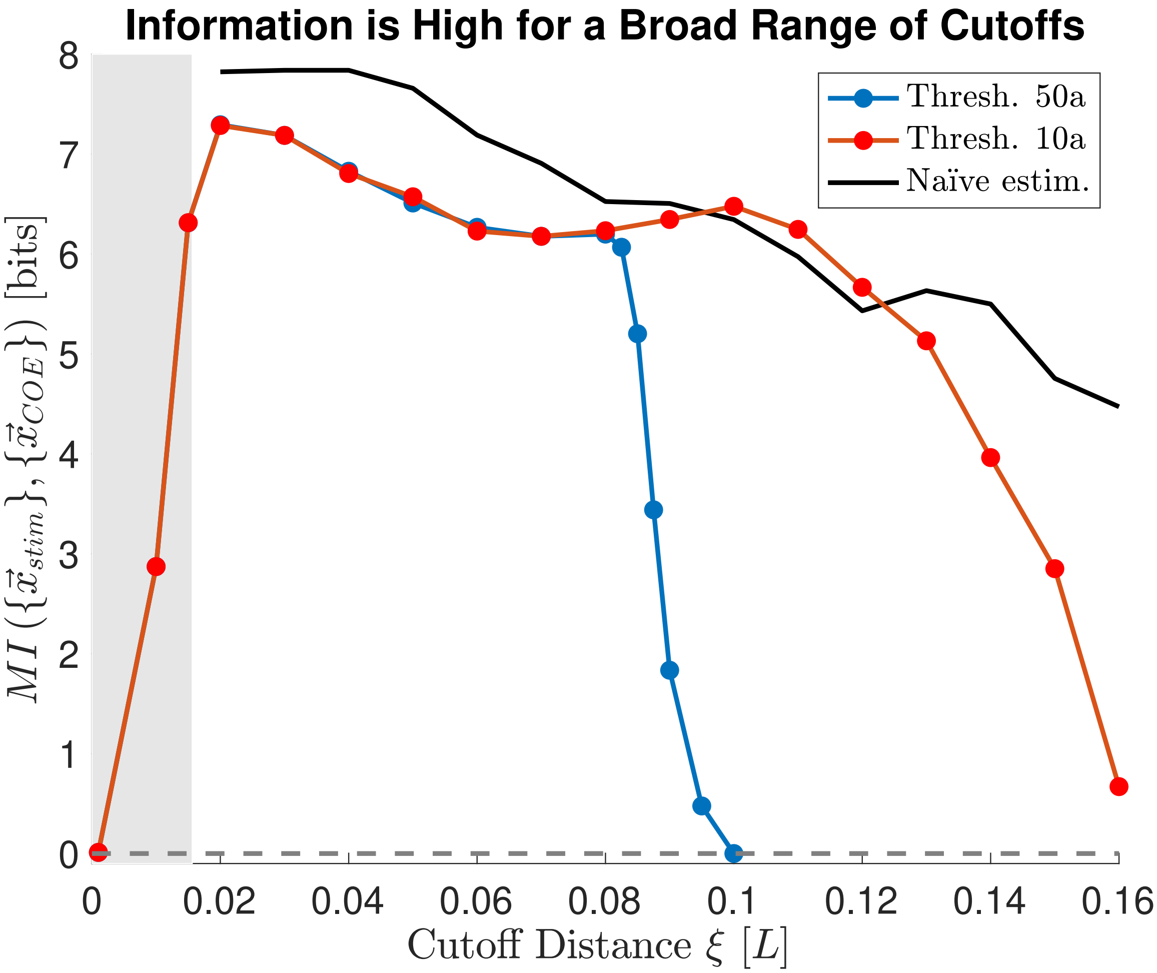

Figure 3.4: Varying ξ reveals a broad plateau over which the mutual information remains within a single bit of its maximum value. At either extreme of ξ the information falls to zero as connectivities become too sparse or too dense to support the type of spatial memory discussed throughout. The black curve represents the information $\log_2 \frac{L^2}{\pi R_{\text{eff}}}$ corresponding to our original, naïve estimate of C, with $R_{\text{eff}}(\xi)$ adjusted to match the typical values given by Gaussian fits to $\sim$ 1000 bumps. Note that the black curve, representing $\xi < \lambda$, exists only outside the shaded gray box because the bumps that did localize for small ξ were too few to measure $R_{\text{eff}}(\xi)$ accurately.

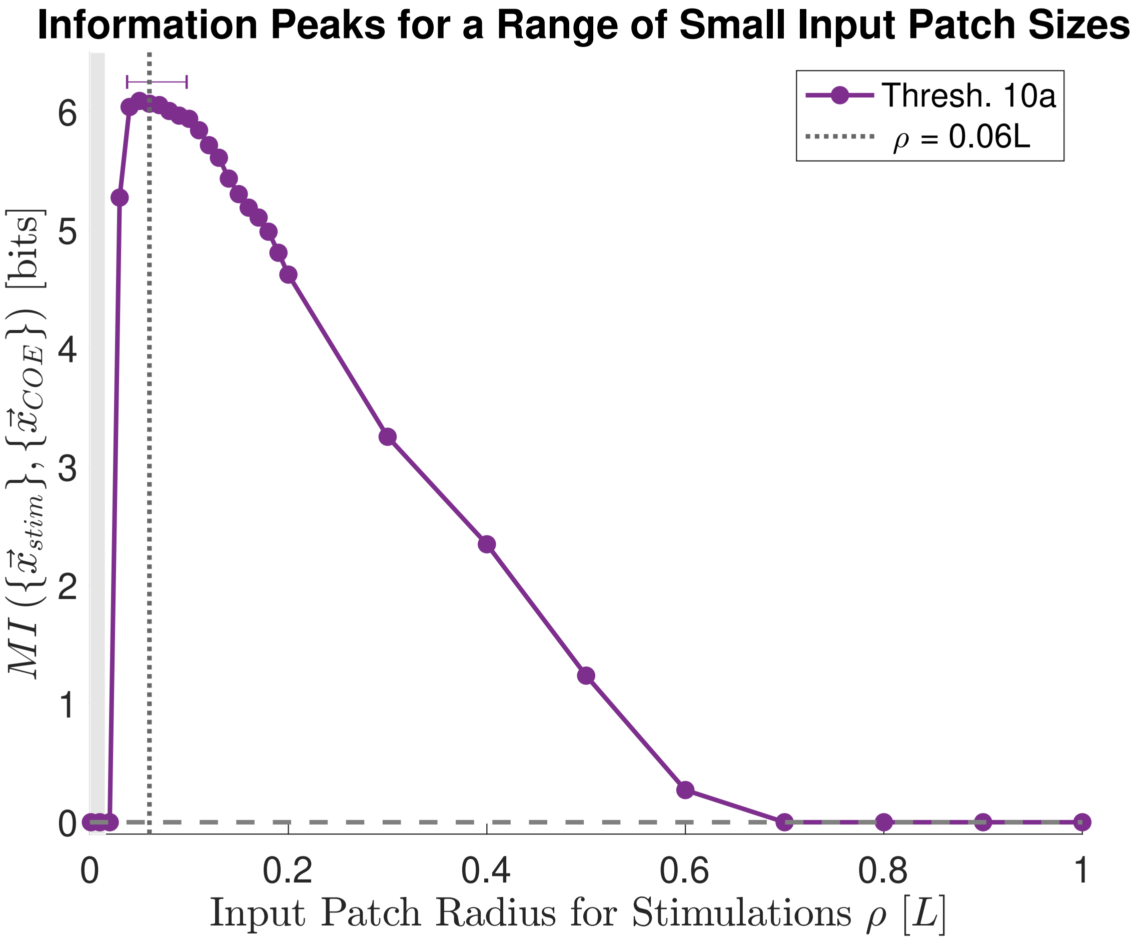

Figure 3.5: In the neighborhood of $\rho = \xi = 0.06L$, the mutual information does not vary significantly. We verified that the system tends to fall into the same attractor regardless of the specific value of ρ until a large percentage of neurons are stimulated, thereby activating the aforementioned "preferred" or global states. At roughly the same value after which see a decrease in information with the cutoff distance, we observe a drop in information with ρ. This continues monotonically until $\rho > 50\lambda$, after which stimulations leads only to excitations below the activity threshold.

Chapter 4

Coarse-Graining and Renormalization without Locality

In Chapter 3, we explored a network model that affords a compact reduced description by virtue of encompassing only a few dynamical attractors, all of which live on a low-dimensional manifold. Far from a situation in which the firing activity of every neuron is of equal importance in determining the system's large-scale behavior, knowing the region of stimulation to a reasonable precision (determined by the system's intrinsic length scales) gave an accurate summary of the resulting spatial activity pattern.

We treated in the underlying network model in Chapter 2 as if it came to us via reverse-engineering: its randomly distributed, effective connection strengths suggested no obvious compressibility, but its dynamical constraints, originating from unobserved elements, ensured that the system could only exist in a handful of states, in the large (time) scale limit. In theoretical physics, the constraints on a system's Hamiltonian that are provided by symmetries ("invariances" to translation, rotation, etc.) and locality (in Chapter 3, the short interaction range and small numbers of actively participating neurons in a given memory state) are what limit the number of possible interactions and behaviors. Without such constraints, there is no guarantee that most traditional physical models would even be tractable.

Thus, even the most intricate physical models in physics are expected to attain a degree of elegance or simplicity merely by virtue of accommodating the underlying symmetries of the systems they describe. Yet complex systems generally – including biological networks – do not as readily admit simplification in this manner [435, 372, 11]. What might be learned from traditional approaches to dimensionality reduction in theoretical physics to inform how to coarse-grain systems that do not as readily admit simplification merely by means of invariance and locality arguments?

In this final Chapter, we attempt to realize our goal of reducing the complexity associated with models of biological systems in a more general and formal manner. Inspired by classical results from theoretical physics, and again with applications like neural activities in mind, here we examine one way of deducing, directly from data, a compressed descriptions of the large-scale behavior associated with a complex system, without first supposing an explicit form for the underlying interactions or inferring them in advance. We outline how one might expect to obtain a satisfying reduction for a well-studied model, without deferring to its known symmetries or spatial structure for guidance. The idea we espouse is to progressively combine or *coarsen* a system's short-range, microscopic degrees of freedom while preserving – indeed, rendering more salient – those features that are *relevant* to determining its statistics at macroscopic scales (here, "microscopic" refers to a resolution on the level of the individual activity variables of Chapter 2, while "macroscopic" corresponds to coarsened variables that each summarize a number of microscopic variables on the order of the system size). This practice was introduced to statistical physics some four decades ago [436, 437], and continues to be developed [438], under the name *renormalization group.*

Renormalization group (RG) theory describes how to systematically coarse-grain physical models and calculate, among other quantities, critical scaling behaviors. The RG approach has been immensely successful for systems where interactions are known to possess a high degree of symmetry at the microscopic level that effectively reduces the number of parameters that can contribute to the system's behavior in the ultraviolet limit. In this final Chapter, we develop a method, analogous to real-space RG techniques and inspired in part by recent demonstrations that low-dimensional descriptions of of certain biological systems may be within reach [371, 373, 439], that is capable of coarse-graining and detecting infra-red behavior directly from experimental data – without explicit reference to known symmetries or (spatial) locality.

Specifically, our information-theoretic approach replaces the usual variational RG objective of minimizing the change in free energy between the microscopic and coarsened systems with an equivalent optimization that preserves a system large-scale statistics. In order to retain the original philosophy and operational criteria of the variational renormalization group [440], we coarse-grain (compress) subsets of system elements that are maximally proximal in information space, sequentially, until all identifiably "short" scales in that space are eliminated. We demonstrate that a coarse-graining transformation defined by an Information Bottleneck compression is fully equivalent to the usual RG map that preserves the information, at each coarse-graining step, about the system's statistics next-shortest scale.

Through repeated applications of the aforementioned transformation, we observe a qualitative analog of the known renormalization group flow for measurable quantities associated a "network" of Ising spins embedded on a two-dimensional, square lattice [441] – a canonical model for the real-space RG. After showing viability of our data-driven approach, we mention how one could have used this generic method to infer the correlation length (i.e., effective temperature) and the corresponding, "large-scale" degrees of freedom for this system from data alone.

> *The material in this Chapter, written under the supervision of K. Michael Martini and Ilya Nemenman, represents a work in preparation for submission to the American Physical Society journal* Physical Review E.

4.1 Introduction

The properties attributed to a physical system depend on the *scale*, at which that system is observed. In many familiar cases, the dominating forces or effects that govern a system at microscopic scales tend to cancel out [370], or else become unobservable [371], at some larger scale. For instance, minute fluctuations in the local magnetization of a bulk ferromagnetic material far below its Curie point do not destroy the crystal's long-range order. Conversely, collective phenomena and emergent complexity [114, 442] ensure that novel interactions and behaviors can yet appear in larger-scale representations of systems whose microscopic dynamics are governed by seemingly simple rules. These two intimately related observations together lead to Kadanoff and Goldenfeld's maxim [119]: one should not, in general, attempt to model bulldozers with quarks, but instead tailor descriptions of nature to accommodate the specific scales at which one intends to make predictions.

In statistical mechanics and quantum field theory, the *renormalization group (RG)* [443, 444] provides systematic ways to interpolate between different observation scales, identifying the phenomenological variables (i.e., relevant interactions or scaling fields and coarse-grained degrees of freedom) that capture a system's dynamics or behavior at new resolutions. Real-space RG techniques [445, 438] proceed from sets of finely resolved, spatially-dependent degrees of freedom or statistical fields that comprise a system's microscopic Hamiltonian to appropriate sets of macroscopic variables via iterative transformations.

Formally, the first step in such an RG procedure entails "integrating out" or combining those degrees of freedom that are associated with the shortest scale of interaction in the system. These deprecated degrees of freedom are replaced, according to some *coarse-graining rule*, with a lower-dimensional set of summary variables. Such a rule allows one to preserve information about the system's long-wavelength statistics – macroscopic thermodynamic observables – by ascribing *effective interac-*

tion strengths to those coarse-grained variables, with values that leave unchanged the free energy of the original system.

This compression-like process constitutes one RG step, and can be repeated until all length scales shorter than the system's intrinsic correlation length have been eliminated. Ideally, it is iterated only until functional forms for the RG *recursion relations*, or "flow" equations (which describe how the Hamiltonian transforms under repeated compressions) can be found. Then, the RG equations can be solved in order to characterize their dynamical *fixed points*, which here represent models that remain unchanged when subject to further coarse-graining.

Even where exact recursion relations cannot be established, *variational* RG approaches [440] can, in principle, be used to identify large-scale degrees of freedom via approximate transformations. Indeed, the variational task of minimizing changes to the free energy has been shown [19] to be equivalent to the optimizations performed by deep neural network architectures based on Restricted Boltzmann Machines [446]. Yet, while the latter have been used to perform a broad range of pattern recognition and generative modeling tasks for diverse systems [447], renormalization group methods, in their traditional forms, remain largely unused in studies of complex systems outside the specific domains of quantum and statistical physics.

There are technical reasons for this domain specificity. First, since the very notion of coarse-graining presumes the existence of an initial model, a necessary preliminary step toward applying RG procedures is the specification of the most general Hamiltonian that is consistent with the known symmetries of the system under study [448]. In physics, known symmetries (such as translational and rotational invariance) serve along with locality constraints to severely limit the number of parameters that can enter such a Hamiltonian; for an arbitrary complex system, this dimensionality-reducing information may not be readily available. Moreover, inferring mechanistic models is computationally expensive (as seen in Chapter 2), and even where effective interaction structures can be “reverse-engineered” from system-wide activity data, such network reconstructions tend to be intractably large or else difficult to interpret – much less coarse-grain systematically. In a sense, then, the automated network inference idea explored in Chapter 2 runs somewhat counter to the RG task of isolating only a few relevant variables.

Despite these challenges, several modern lines of work, inspired by the RG in physics, suggest that it is indeed possible to separate scales and arrive at low-dimensional, fixed-point descriptions for certain systems in biology and social science [371, 373]. Both approaches pursued therein introduce alternative ways to perform the coarse-graining step for systems that were not amenable to standard RG methods, but they are not capable of reducing the space of potentially relevant interactions to a tractable ($\mathcal{O}(N)$ rather than the $\mathcal{O}(2^N)$ total connections in a network of N interacting degrees of freedom) subset without invoking the aforementioned symmetry considerations [449]. Meanwhile, a recent coarse-graining of neural activity data in the murine hippocampus [439] identified various power-law scaling behaviors, indicative of the possibility of attaining a simple description of one network’s dynamics in terms of a nontrivial fixed point of the RG flow.

This latter study performed the coarse-graining step by combining degrees of freedom according to their distance in an abstract, “correlation” space instead of real space, since the connections between neurons in their system are not organized according to geometrical proximity, raising the question of whether it is possible – and then, how – to construct RG-style transformations for general complex systems in the complete absence of knowledge regarding “who interacts with whom” (i.e., the Hamiltonian) [37].

In this Chapter, we aim to close the gap between the profound successes of renormalization techniques within their native branches of condensed matter and high energy physics and the lack of equivalent tools for discerning an appropriate vocabulary of large-scale variables for microscopically complex systems generally. We recast the variational renormalization group objective function as an information-theoretic optimization to develop a data-driven method of coarse-graining that operates directly on the network's joint probability distribution. We define "local" interaction neighborhoods in *information space* that can be iteratively removed, as in the real-space RG. Using this method, we identify, from simulated data, sets of large-scale degrees of freedom for a 2D Ising model [441] – the simplest network of spin-spin interactions known to exhibit a phase transition, with an analytical solution [450] – that resemble the traditional "block spin" variables [445]. In addition, our algorithm recovers the qualitative RG flow for the nearest-neighbor correlations associated with pairs of these "block" variables, at various scales. Our results suggest that a fully quantitative mapping between our outputs and previous analytical solutions is also within reach; we discuss the concrete, remaining steps toward achieving this goal.

In the following section, we review the basic aspects of variational renormalization group theory that are needed to motivate our algorithm. Then, we relate the standard RG parlance to our new terminology, showing that minimization of a free energy difference can be thought of in terms of a well-known information-theoretic problem in Section 4.2.2. We describe the general data requirements and specific example system to be coarse-grained in Section 4.2.3, and summarize our specific algorithmic choices for that system in Section 4.2.4. Our main results are presented in Section 4.3, with a commentary on their interpretation and agreement with predictions from standard RG theory following in Section 4.4. We conclude with a discussion of the aforementioned, concrete steps that are needed to complete our future work in Section 4.5.

4.2 Motivation and Methodology

We start here with an abbreviated review of real-space renormalization group theory, as it was conceived in statistical physics, for completeness and to foster intuition that is useful for understanding the remainder of this work. We then depart from this traditional formulation, building on this intuition, to develop the fundamentals of our information-theoretic approach. Once we have motivated our own method, we describe both the general type of data we wish to coarse-grain and our Ising "test" system, to which we will apply our algorithm. Finally, we commit to the specific algorithmic details for this system, wherever there were freedoms of choice in implementation, summarize the iterative algorithm itself, and discuss additional details regarding its validation.

4.2.1 Real-Space RG within Statistical Physics

We begin with known Hamiltonian $\mathcal{H} = \mathcal{H}\left(\{\sigma_i\}, \{K\}\right)$ describing the behavior of a physical system in terms of N "micro" degrees of freedom $\{\sigma_i\}$ and parameters $\{K\}$. Each realizable microscopic state $\{\sigma_i\} = \{\sigma_1, \sigma_2, \ldots \sigma_N\}$ is associated with an appearance probability or *weight* in some statistical ensemble $P\left(\{\sigma_i\}\right)$, which is a function of $\mathcal{H}$. For the canonical ensemble in particular, we have the customary Boltzmann-Gibbs measure defined by $P\left(\{\sigma_i\}\right) = e^{-\beta\mathcal{H}}/Z_\beta$, where β is the inverse temperature and Z_β is the partition function (we set $\beta = 1$ wherever it appears, for the remainder of this Section).

By defining a suitable projection operator $\mathcal{T}(\sigma_i, \sigma'_{i'})$, we can then build a set of new, *coarse-grained* variables $\{\sigma'_{i'}\}$ that summarize those original degrees of freedom, whose weights $P'\left(\{\sigma'_{i'}\}\right)$ are sums of the probability weights corresponding to the microstates they replace:

$$P'\left(\{\sigma'_{i'}\}\right) = \frac{1}{Z_\beta} e^{-\sum_i \mathcal{T}(\sigma_i,\sigma'_{i'})\mathcal{H}(\{\sigma_i\},\{K\})}. \tag{4.1}$$

Here, $\mathcal{T}(\sigma_i, \sigma'_{i'})$ is some local operator (that is, it relates subsets of interacting system elements) whose form is expected to conform to the general symmetries obeyed by the system. If its values are positive-definite, it makes intuitive sense to define an *effective Hamiltonian* for the coarse-grained degrees of freedom. By analogy with $\mathcal{H}$,

$$\mathcal{H}'\left(\{\sigma'_{i'}\}, \{K'\}\right) \propto \log P'\left(\{\sigma'_{i'}\}\right), \tag{4.2}$$

where the effective parameters $\{K'\}$ are yet unknown.

In fact, the fundamental assumption is that $\mathcal{H}$ and $\mathcal{H}'$ are special cases of the most general Hamiltonian that still reflects all the symmetries we expect of the system itself: their differences can only be found in the values of the couplings – including the possibility that a coupling previously missing from $\{K\}$ can assume nonzero values in $\{K'\}$ – and in rescalings of the degrees of freedom. This is characteristic of any system near its critical point. If in addition we require $\sum_{\{\sigma'_{i'}\}} \mathcal{T}(\sigma_i, \sigma'_{i'}) = 1$, we see that

$$Z \equiv Tr_{\{\sigma_i\}} e^{-\mathcal{H}(\{\sigma_i\},\{K\})} = \mathrm{Tr}_{\{\sigma'_{i'}\}}\, e^{-\mathcal{H}'(\{\sigma'_{i'}\},\{K'\})}. \tag{4.3}$$

In other words, the system's partition function remains invariant under the transformation $\mathcal{H} \to \mathcal{H}'$. Indeed, this transformation also preserves the various derivatives of the partition function, including the thermodynamic free energy $\mathcal{F} = -\ln Z_\beta$. Since the coarse-graining represents some averaging of the degrees of freedom over some small spatial scale, we can say that the "new" Hamiltonian $\mathcal{H}'$ retains the large-scale (macroscopic) behavior of $\mathcal{H}$.

The coarse-graining transformation $\{\sigma_i\} \to \{\sigma'_{i'}\}$ can be iterated, so that we have

$\{\sigma'_{i'}\} \to \{\sigma''_{i''}\}$, and so forth. At each successive stage, the coupling parameters $\{K'\}$, $\{K''\}$, etc., are chosen to ensure the constancy of the free energy. In the ideal case, the couplings at each stage of coarse-graining will map onto those of the next according to a precise functional form $\mathcal{R}(K) : \{K\} \to \{K'\}$, known as the RG *recursion relation(s)* for the system. Then the so-called RG *flow* that describes the trajectory of models realized by the RG procedure in parameter space.

Identifying the dynamical *fixed points* of the RG flow is equivalent to finding a set of simplified or coarsened set of descriptors for the system's behavior at large scales. Formally, linearizing the RG flow equations $K' = \mathcal{R}(K)$ about a given fixed point $K^* = \mathcal{R}(K^*)$ allows one to read off eigenvalues for different scaling variables (i.e., linear combinations of the parameter deviations from their fixed-point values); these eigenvalues are either greater, less than, or equal to unity. If they are greater, the repeated action of (the linearized version of) $\mathcal{R}(K)$ causes the scaling variables to increase, leading them further away from their fixed-point values. Such eigenvalues are known as *relevant* because their associated eigenvectors represent a flow directed away from the critical manifold. *Irrelevant* eigenvalues, with values below unity, denote the opposite flow, toward the fixed point (i.e., they span the critical manifold). Only relevant eigenvalues are associated with the experimental "knobs" that must be adjusted to tune a system to its critical point; *marginal* eigenvalues, equal to unity, are not directly informative about the approach to a fixed point. They are instead associated with logarithmic corrections to scaling [444].

Unfortunately, renormalization group equations which preserve the partition function exactly are known for only a handful of systems [451]. Where solving for $\mathcal{R}(K)$ is infeasible, the *variational RG* approach [440] can often be used to approximate it. The idea is to minimize the free energy *difference* between the original and coarse-grained systems, rather than attempting to preserve the partition function identically.

This difference is defined by

$$\Delta\mathcal{F} = (\mathcal{F}' - \mathcal{F}) = \left(\ln \frac{\mathrm{Tr}_{\{\sigma_i\}}\, e^{-\mathcal{H}}}{\mathrm{Tr}_{\{\sigma'_{i'}\}}\, e^{-\mathcal{H}'}} \right). \tag{4.4}$$

The free energy calculations that lead to Eq. (4.4) above are, of course, only possible because we assumed at the outset to be in possession of the Hamiltonian $\mathcal{H}$. Without such detailed knowledge about the interactions that comprise a system, not even variational RG methods can be used to determine an appropriate set of large-scale degrees of freedom and effective coupling constants $\{K'\}$.

4.2.2 Algorithm Motivation

In the absence of detailed, prior knowledge about the interaction structure of the system (i.e., the form of $\mathcal{H}$), it is not straightforward to choose a coarse-graining rule $T : \{\sigma_i\} \to \{\sigma'_{i'}\}$ by which to combine and compress the $\{\sigma_i\}$: without a way to measure the global free energy difference $\Delta\mathcal{F}$, one cannot quantify the contribution of individual, microscopic features to a system's large-scale behavior. For biological networks and many other complex systems, it is the prerequisite step of writing down a mechanistic form for $\mathcal{H}$ that is problematic.

Is there some way to *construct* $\mathcal{H}$, approximately? We have seen in Chapter 2 that, under a broad range of conditions, network interaction architectures can be reconstructed, or "reverse-engineered," from abundant activity measurements. Yet the very notion of inferring a large, high-dimensional object as an intermediate step, only to coarse-grain (reduce the dimensionality) of that object, is something we have argued vehemently against. To wit, performing tasks in this order seems to violate a well-known heuristic from statistical learning theory: one should avoid solving a "hard" problem as an intermediate step toward solving an "easy" (that is, more direct) one [36]. What is needed, then, is a method of "renormalization" that – like the network inference algorithms of Chapter 2 – operates directly and exclusively on *data*.

Indeed, it is possible, and arguably more transparent, to cast the RG in terms of joint probability distributions. This can be done because the preservation of the free energy is actually an ancillary statement about $P\left(\{\sigma_i\}\right)$ and $P'\left(\{\sigma'_{i'}\}\right)$ for the original and transformed systems: Equations (4.1) and (4.2) together imply that the full probability measure for quantities that depend only on higher-order "coarsenings" ($\{\sigma'_{i'}\}$, $\{\sigma''_{i''}\}$, $\{\sigma'''_{i'''}\}$, etc.) – such as the thermodynamic observables – are preserved (exactly, at least in the case of exact RG transformations) [444].

Here, we explore the possibility of finding a new coarse-graining rule, without explicit reference to the interaction structure of the system. This lack of a microscopic model necessarily places us in the variational regime; we now try to build intuition regarding how to preserve the system's large-scale behavior in terms of $P\left(\{\sigma_i\}\right)$ and $P'\left(\{\sigma'_{i'}\}\right)$.

Let X denote the subset of system elements that is to be replaced via some coarse-graining transformation, and X' its respective coarse-grained variable(s). The remaining system elements $Y = \{\sigma_i\}\backslash X$, where the symbol $\backslash$ denotes the set difference operation, each then serve as their own coarse-grained variables $\{\sigma'_{i'}\}$. We now seek a coarse-graining rule $\mathcal{T} : X \rightarrow X'$ whose output must respect the large-scale features of the system by leaving the distribution $P'(X', Y)$ as close as possible to $P(X, Y)$.

In principle, two probability distributions can be compared by measuring their *statistical distance*, according to information-theoretic metrics (and pseudo-metrics) [210]. Since the above distributions have different supports (the cardinalities of X and X' differ by definition), it is not possible to quantify their statistical distance directly. Instead, we find the Kullback-Leibler (KL) divergence [452] between the initial distribution over Y, conditioned on X, and the new distribution over Y, conditioned on

X':

$$D_{KL}\left[P\left(Y|X\right)||P'\left(Y|X'\right)\right] \\ = \sum_{y\in Y} P\left(Y|X\right)\log\frac{P\left(Y|X\right)}{P\left(Y|X'\right)}, \tag{4.5}$$

which is in this context a function of the values $x \in X$ and their corresponding $x' \in X'$, as defined by the coarse-graining rule $\mathcal{T}$. Then, we propose to minimize the average of this quantity over all input-output pairs (X, X'),

$$\Big\langle D_{KL}\left[P\left(Y|X\right)||P'\left(Y|X'\right)\right]\Big\rangle_{X,X'}, \tag{4.6}$$

as a surrogate objective for the free energy of Eq. (4.4).

That Eq. (4.5) constitutes a natural quantity with which to encode the large-scale behavior of the system can be seen by acknowledging the following. If, without loss of generality, and in the spirit of Kadanoff's variational prescription, the couplings among the $\sigma_i \in X$ represent the shortest-scale interactions in the system, the elements of Y necessarily contain the information about the system at larger scales. Therefore, the probability measure that needs to be preserved (i.e., minimally modified) under $X \to X'$ is the conditional distribution $P\left(Y|X\right)$.

The variational problem of finding a (stochastic) map $X \to X'$ that minimizes Eq. (4.5) is related intimately [453, 454] to the *Information Bottleneck* optimization [455]

$$\min_{P(X'|X)} \{I(X;X') - \Lambda I(X';Y)\}, \tag{4.7}$$

with

$$\Lambda \to \infty.$$

In Eq. (4.7), the mutual information $I(X;X')$ measures the similarity, or degree of compression, between the subset of elements X to be coarse-grained and their coarsened version, X'. Meanwhile, $I(X', Y)$ quantifies the information that X' contains about the set of reference variables Y; the value of this quantity is bounded from above by $I(X, Y)$. Minimizing the difference of these two terms, with the Lagrange multiplier Λ acting as a tradeoff parameter, means seeking a distribution $P(X'|X)$ such that the latter approximates $I(X, Y)$ as closely as possible, with X' being constrained to be as different as possible – in other words, maximally compressed – from X itself.

The demand that $\Lambda \to \infty$ amounts to neglecting the first, "compression," term in favor of maximizing the relevant information about degrees of freedom located at distances larger than the separation between those $\sigma_i \in X$. If the dimensionality of (number of possible values taken by) X' is much smaller than that of X, we assume that this reduction in dimensionality represents the most significant compression that X will undergo. Then, dropping the first term completely will not affect the output distribution $P(X'|X)$, and the limit is consistent. In what follows, we maintain this assumption whenever the dimensionality of X' is smaller than X, with the simple substitution of a large, finite value for the parameter Λ.

Note:

The relation between these two optimization problems can be illuminated by rewriting the mutual information variables of Eq. (4.7) in terms of the constituent entropies:

$$\begin{aligned} I(X;Y) = I(Y;X) &= S(X) - S(X|Y) \\ &= S(Y) - S(Y|X) \end{aligned} \tag{4.8}$$

$$I(X';Y) = I(Y;X') = S(Y) - S(Y|X') \tag{4.9}$$

Consider the identity formed by subtracting and adding the quantity $I(X;Y)$ from

$I(Y;X')$. That is, we study

$$I(X';Y) = I(X';Y) - I(X;Y) + I(X;Y), \tag{4.10}$$

the first two terms of which can themselves be rewritten in terms of entropies by using Eq. (4.8) and Eq. (4.9) above:

$$I(X';Y) - I(X;Y) = S(Y|X) - S(Y|X'). \tag{4.11}$$

This difference of conditional entropies can be written in a more compact form. To do this, we must rewrite both terms of Eq. (4.11) in new forms that encompass sums over all three random variables X, Y, and X'. We have

$$\begin{aligned} S(Y|X) &= -\sum_{\substack{x\in X,\\ y\in Y}} P(X,Y)\log P(Y|X) \\ &= -\sum_{\substack{x\in X,\\ y\in Y,\\ x'\in X'}} P(X,Y,X')\log P(Y|X) \end{aligned} \tag{4.12}$$

for the first entropy, and

$$\begin{aligned} S(Y|X') &= -\sum_{\substack{x'\in X',\\ y\in Y}} P(X',Y)\log P(Y|X') \\ &= -\sum_{\substack{x\in X,\\ y\in Y,\\ x'\in X'}} P(X,Y,X')\log P(Y|X') \end{aligned} \tag{4.13}$$

for the second. Then we can write

$$\begin{aligned}
&S(Y|X) - S(Y|X') \\
&\quad = \sum_{\substack{x\in X \\ z\in X'}} P(X,X') \sum_{y\in Y} P(Y|X,X') \log \frac{P(Y|X')}{P(Y|X)} \\
&\quad = - \sum_{\substack{x\in X \\ z\in X'}} P(X,X') D_{KL} \left[P(Y|X)||P(Y|X')\right], \qquad (4.14)
\end{aligned}$$

where the equation $P(Y|X,X') = P(Y|X)$ in last line above is valid only if the variables form the Markov chain $X' \leftarrow X \leftarrow Y$, as in the Information Bottleneck [455].

We assume the Markov chain requirement will be satisfied in our RG context, since this is essentially the purpose of the projection operator $\mathcal{T}(\sigma_i, \sigma'_{i'})$. This allows us to collapse the right-hand side of Eq. (4.10) to the form

$$I(X;Y) - \langle D_{KL} \left[P(Y|X)||P(Y|X')\right]\rangle_{X,X'}. \qquad (4.15)$$

Since $I(X;Y)$ is fixed, it does not contribute (apart from a global offset), to the minimization in Eq. (4.7). We can then recast our RG problem in the equivalent form

$$\min \{I(X;X') + \Lambda \langle D_{KL} \left[P(Y|X)||P(Y|X')\right]\rangle\}, \qquad (4.16)$$

where, again, the "compression" term, $I(X;X')$, is to be neglected (asymptotically, in the limit that $\Lambda \to \infty$).

Thus the Information Bottleneck optimization, Eq. (4.7), in the aforementioned limit, is completely equivalent to the minimization of our desired KL divergence, averaged over all realizable input-output pairs (X,X') of the coarse-graining transformation. The mapping $X \rightarrow X'$ via the conditional distribution $P(X'|X)$ thus generalizes the usual coarse-graining projection operator $\mathcal{T}(\sigma_i, \sigma_i')$.

The Information Bottleneck optimization can solved by an iterative procedure that generalizes the Blahut-Arimoto algorithm in information-theoretic rate distortion theory. Whereas a search through the space of possible coarse-graining rules to find an optimal $P(X'|X)$ could quickly become prohibitively expensive for large systems, the Information Bottleneck's iterative optimization procedure inherits convergence properties from that original algorithm, guaranteeing a local minimum solution, justifying its use even for the special case $\Lambda \rightarrow \infty$.

Moreover, this implies that the task of finding an appropriate, stochastic coarse-graining transformation that preserves, as closely as possible, the probability weights over the remaining activity variables – in other words, the information about all longer "scales" in the system – can be reduced to the optimization of various mutual information values (nonlinear correlations) between subsets of activity trajectories. In particular, working at the level of probability distributions has ensured that the free energy difference ΔF can be minimized simply by finding a compression X' of the subset X that reproduces, as accurately as possible, the mutual information between the trajectories X and the rest of the trajectories Y – with no need to enumerate the intractably large number of terms that could comprise the partition function.

In systems with a well-defined range of interaction, it is in principle possible to to simplify this optimization yet further. Namely, we can take for Y not "all remaining" trajectories $\{\sigma_i\} \setminus \{X\}$, but a smaller subset representing the *local neighborhood* of X. This can be done whether there is a sharp cutoff (as in, say, a 2D Ising model with nearest-neighbor interactions) or longer-range influences that diminish over some characteristic length scale.

Indeed, in developing our method, we would also like to consider more general notions of locality, such as having a small number of neighbors in some abstract space.

This type of locality is exhibited by certain genetic networks (see Chapter 2) that incorporate only a handful of neighbors per node. One can also speak about “local” interactions in correlation space, as in the case of inferring spatial contact among amino acid residues from pairwise correlation within protein sequences ([172], or see Chapter 2), or the neighbors of a given pixel in natural images [1]. While the basis for these two endeavors relies on an explicit correspondence between the neighborhoods in geometric and correlation space, we are interested in coarse-graining systems even for which we have no access to geometric information – or, where the geometric and correlation “spaces” do not map neatly into one another. This latter kind of locality was recently used, successfully, to define neighborhoods for interacting neurons [439].

The extent to which a system's constituent interactions are local, in either sense, may also affect the cardinality of X, as we will see later. In the following sections, we ask how the variational approach motivated in this section can be used in an algorithmic pipeline to determine the large-scale, effective degrees of freedom $\{\sigma'_{i'}\}$ from data alone, and without explicit reference to spatial structure.

4.2.3 Description of the Data to be Coarse-Grained

In motivating our approach, we assumed no access to the system's (local or nonlocal) interaction structure. We will work directly with data, which presumes, at most, knowledge of the joint probability distribution $P(X, Y)$. In practice, this $P(X, Y)$ is never known exactly. We will estimate the needed marginal distributions empirically, from observations or *activity measurements* $m = 1 \dots M$ on the N elements, sometimes called nodes, that comprise the system at the microscopic level. These activity variables, $\{\sigma_i(m)\}$, can be either discrete or continuous.

For convenience, we arrange our activity measurements in an $N \times M$ matrix, so that each column will represent one possible realization of the system's microstate

[1] Private communication with Mahajabin Rahman and Ilya Nemenman, regarding original work in preparation for publication.

(i.e., configuration $\{\sigma_i\}$), and each row the set of activity samples for one individual element. We assume these samples reflect simultaneous measurements across all elements in the system, but make no distinction regarding how they are obtained: they might be independent or consecutive samples, from a system in equilibrium or a time series.

Until now, our discussion has remained generic, since we desire to develop a method that will work for many different systems. Since we will be working directly at the level of probabilities, the only available input will be the system's empirical joint distribution, and this should be possible in principle. Yet at this point, for concreteness, and in order to test our algorithm on a system for which an appropriate set of large-scale degrees of freedom are known, we shall focus on a specific model system.

For the remainder of the present work, we study the $d = 2$ Ising model on the square lattice. We consider N interacting spin-$\frac{1}{2}$ particles, or $\sigma_i(m) = \{\pm 1\}$. Our spins occupy a lattice of side length $L = \sqrt{N}$, and remain in thermal contact with a heat bath at inverse temperature $\beta \equiv \frac{1}{k_B T}$. If only pairwise, nearest-neighbor interactions and coupling with a uniform, external magnetic field H are permitted, we can define the Hamiltonian

$$\mathcal{H}\left(\{\sigma\},\{K\}\right) = -J \sum_{\langle i,j \rangle} \sigma_i \sigma_j + H \sum_i \sigma_i, \tag{4.17}$$

where the symbol $\sum_{\langle \cdot \rangle}$ refers to summation over nearest-neighbor pairs. For simplicity, we ignore the second term (set the external field $H = 0$) and consider only $J > 0$.

Spin-spin interactions $\sigma_i \sigma_j$ in this system exist at short range exclusively, with a coupling strength $(-J)$ that is uniform for all interacting pairs. That is, the interactions are isotropic and spatially homogeneous, so that each σ_i interacts in the same manner with all four neighbors.

Whereas our choice above represents the ferromagnetic interaction, it can be shown that the free energy density is invariant to the sign of J. Since $\sigma_i \in \{-1, 1\}$, $\mathcal{H}$ is also clearly invariant to the transformation $\sigma_i \to -\sigma_i$. Given $H = 0$, $\mathcal{H}$ possesses the "sub-lattice symmetry," which allows one to split the lattice into "even" and "odd" sub-lattices that do not interact [443]. We exploit this symmetry to increase the efficiency of our sampling process [456].

This model exhibits a second-order phase transition at $(T = T_c,\ H = 0)$, with a critical temperature given by [450] $\sinh^2\left(\frac{2}{k_B T_c}\right) = 1$ (here we work in units such that $\frac{J}{k_B} = 1$, and therefore $T_c = \frac{2}{1+\sqrt{2}} \approx 2.2692$). Among this system's critical properties, there can be observed the power-law decay of the spin-spin correlation function with distance, $G(r, T)$, at precisely T_c, with *critical exponent* $\eta = \frac{1}{4}$ [445]:

$$G(r, T = T_c) \propto \frac{1}{r^\eta}. \tag{4.18}$$

Taking into account the symmetries mentioned above, RG methods predict just two relevant parameters for the Ising universality class, which turn out to be the *reduced* temperature $t = \frac{T-T_c}{T_c}$ and magnetic field $h = \frac{H}{k_B T}$.

The scaling fields, t and h, control the statistics of the large-scale degrees of freedom $\{\sigma'_{i'}\}$, which for the Ising model are simply local spatial averages of the microscopic spins. The *de facto* implementation of these averages is Kadanoff's "block spin" transformation [445], which can be performed by a projection operator $\mathcal{T}(\sigma_i, \sigma'_{i'})$ which takes the "majority rule," or sign of the mean spin value.

In order to sample from the canonical equilibrium distribution $P(\{\sigma_i\}) = \frac{e^{-\beta\mathcal{H}(\{\sigma\},\{K\})}}{Z_\beta}$ with $\{K\} = \{J, 0\}$, we perform Monte Carlo simulations in Matlab. In particular, we implemented the parallelized version of the heat bath algorithm [457], or "Gibbs sampler," that appears in Ref. [456]. Simulation results for select thermodynamic observables are shown below for a system of 64×64 spins, produced with 10^6 flips/spin for each Monte Carlo step.

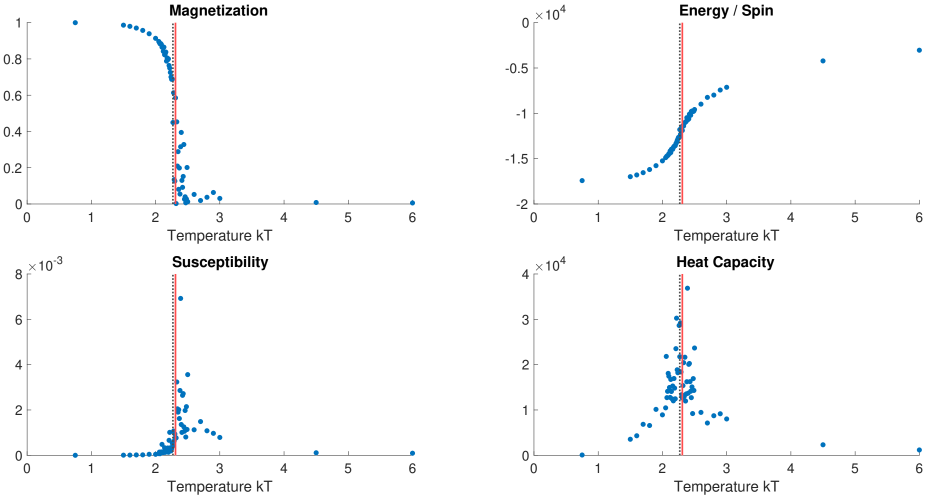

Figure 4.1: *Upper left*: Magnetization per spin; *Upper right*: Energy per spin; *Lower left*: Susceptibility; *Lower right*: Heat Capacity; Red vertical line: T_c.

This sampling protocol, consisting of M Monte Carlo steps in a given run, results in a series of M whole-lattice spin configurations. Since each such step results in a fully thermalized lattice state, we can draw individual samples at either the same (inverse) temperature β or at different temperatures $\{\beta_1, \beta_2, \ldots, \beta_M\}$.

4.2.4 Algorithm Outline

The end goal of any RG program is to extract certain *relevant* features – those which determine a system's behavior at macroscopic scales – from a microscopic model. Traditionally, these features include an appropriate set of "effective" degrees of freedom, as we have mentioned, as well as the parameters (or, technically, their combinations in the form of scaling fields) that summarize their statistics. So far, we have argued that the defining objective of the (variational) renormalization group – the preservation of the free energy – can be given a precise interpretation in term of information-theoretic concepts. In this section, we begin our conversion of this equivalent view into an algorithm by which to detect the effective degrees of freedom that describe a system at a particular, larger length scale, directly from data.

In developing our algorithm, we will encounter several freedoms in implementation – such as how to choose the element subset X to be coarse-grained, or what subset of reference elements Y will to represent the interactions or statistics at larger scales – that depend on the data being analyzed. We emphasize clearly throughout where our specific choices could have differed, postponing for this work any discussions of their general validity.

How can we choose the subset of spins to coarse-grain at a given RG step? As usual, we must somehow choose X to represent those degrees of freedom which interact on the shortest scales encompassed by the system. Since we have eschewed any real-space notion of locality from the outset, we determine the "neighborhood" of a given spin σ_i by measuring its distance from each other spin $\sigma_{j\neq i}$ in *information space*. Stronger correlations are taken as signatures of more local, "shorter-scale," interactions.

The logic for this substitution is as follows. In many physical systems spatially local interactions will produce strong spatial correlations in the vicinity of a particular element. For example, this is trivially true for networks of interacting spins [441], but has also been observed in various biological contexts ([75, 174], or see Chapter 2).

As mentioned earlier, the mutual information is our natural choice for measuring these correlations, since the Information Bottleneck problem of Eq. (4.7) is written in terms of this same measure of statistical dependency. In addition, for large numbers of samples and variables $\{\sigma_i\}$ that take on a small number of different values, this quantity is not much harder to estimate reliably than the familiar linear (Pearson) correlation [316]. Thus, in order to select members for the subset X to be coarse-graining, we compute

$$I\left(\sigma_i;\sigma_j\right) = S\left(P\left(\sigma_i\right)\right) - S\left(P\left(\sigma_i|\sigma_j\right)\right) \tag{4.19}$$

for all $\binom{N}{2}$ unique pairs (i,j), and rank them in descending order, beginning with the value $I(\sigma_{i^*};\sigma_{j^*(i^*)})$ associated with the maximally dependent sample sets $\sigma_{i^*}(m)$ and $\sigma_{j^*(i^*)}(m)$. Then, we establish the neighborhoods for each spin by enumerating those "surrounding" elements for which the value of Eq. (4.19) exceeds some threshold.

There are several ways to define such a threshold. Our strategy here is never intended to reproduce a system's precise interaction architecture (this is known as inferring "mechanistic" interactions, and was discussed in Chapter 2), but rather to isolate the most strongly coupled subset of elements, giving a precise meaning to the standard RG notion of "local" averaging. Thus we employ a strict cutoff that prevents aggregating all but a small number of "nearby" neighbors for coarse-graining. Managing only a few neighbors at a time also keeps the problem tractable.

For the ferromagnetic Ising system of Eq. (4.17), it is clear that the strongest couplings at the microscopic level are local (indeed, the direct interactions involve nearest neighbors exclusively). Due to this inherent *locality*, we expect to find the highest values of $I\left(\sigma_i;\sigma_j\right)$ for nearby neighbors on the lattice. In language of correlations, this also follows from the well-known exponential decay of the spatial correlation function, $G(r,T) \propto r^{-\eta}\ e^{-\frac{r}{\xi(T)}}$, for $T \neq T_c$, where $\xi(T)$ is the correlation length and η is the critical exponent describing the power-law decay of pairwise correlations at $T = T_c$ exactly [443] . We observe in simulation that this holds even if correlations are computed using independent Ising samples that are drawn at a range of different temperatures (not depicted here), although it may not be true in general. For now, we will not discuss the general validity of exploiting across-sample correlations as a proxy for locality for other systems.

For the Ising case, we define our threshold such that the subset to be coarse-grained encompasses only the single, maximally correlated pair; that is $X = \{\sigma_{i^*}, \sigma_{j^*(i^*)}\}$. By this extreme choice, we mean to take literally Kadanoff's prescription to coarse-grain or "integrate out" only those degrees of freedom associated with the smallest scale of interaction present in the system at each RG step. This allows us to liberally and intentionally discard, in making each successive compression, any information which is not relevant to the network's large-scale statistics. In contrast, previous efforts to solidify connections between the RG and (deep) machine learning [19, 25] have focused predominantly on architectures like Restricted Boltzmann Machine-based autoencoders, which attempt to compress all aspects of input data on equal footing.

With our selection of X complete for a given RG step, we can proceed to revisit and refine our choice for Y. For computational tractability, we would like to invoke our earlier assumption that a comparison with "all remaining" spins in Eq. (4.7) can be supplanted by comparison with some representative subset. Although this is guaranteed to be valid only if the interactions are sufficiently short-ranged, this is the case for the Ising model. We keep the single pair of next-most highly correlated spins, respectively, to each of the elements in X: $Y = \{\sigma_{k^*(i^*)}, \sigma_{l*(j^*(i^*))}\}$, with $I(\sigma_{i^*}; \sigma_{k^*(i^*)}) \geq I(\sigma_{j^*(i^*)}; \sigma_{l*(j^*(i^*))})$ by convention. This choice also reflects Kadanoff's prescription, but instead of aiming to preserve the information *all* higher scales, it preserves only the "next" interaction scale in the system.

Once we have designated values for both X and Y, and estimated their respective distributions $P(X)$ and $P(Y)$, we use the iterative optimization procedure (extension of the Blahut-Arimoto algorithm) of Ref. [455] to determine a minimizing distribution $P(X'|X)$. In practice, we set the parameter $\Lambda = 100$. This strongly de-emphasizes the "compression" term of Eq. (4.7), under the aforementioned assumption that further compression – beyond changing from the quaternary alphabet

$\{\{-1,-1\},\{-1,+1\},\{+1,-1\},\{+1,+1\}\}$ of X to the binary alphabet $\{\{-1\},\{+1\}\}$ – is not needed [2].

Using the conditional distribution $P(X'|X)$ and the trajectories for both (hyper-)spins in the set X, we create a set of samples for new hyperspin X'. This is done simply by flipping a coin (generating a uniformly distributed random number) weighted by $P(X'|X)$ for each $m = 1 \ldots M$. As a convention, if a randomly number is less than $P(X'|X)$, we assign a value of $+1$ to the corresponding (mth) sample of X'. Then, once all M binary samples are drawn, we adjust the overall sign in for this set of samples such that its linear correlation with the quantity $\left(\sigma_{i^*} + \sigma_{j^*(i^*)}\right)$ is higher than its correlation with $-\left(\sigma_{i^*} + \sigma_{j^*(i^*)}\right)$. This last adjustment effects the "renormalization" part of the RG by ensuring that the new hyperspin X takes on the same range of values as the spins it replaced, in the same circumstance. In other words, the compression *behaves* like a spin subject to ferromagnetic (and not antiferromagnetic influences.

Finally, we add X' to the system, while removing X. This entails modifying the (originally $N \times M$) activity matrix by deleting the two rows associated with $\sigma_{i^*}(m)$ and $\sigma_{j^*(i^*)(m)}$, and appending a one to contain the newly generated set of samples for X'. We begin the next iteration by selecting values for X and Y from this reduced matrix, which will contain $N - \alpha$ rows following coarse-graining iteration α. This process can continue, in principle, until some small number ω of rows remains. We

[2] Matlab implementation was adapted from a function written by C. Wiggins and I. Nemenman, ©2002 (used with permission)

summarize the overall operation of our algorithm below.

Algorithm 1: Coarse-Graining Procedure

```
Input  : $\{\sigma_i(m)\}$, $i = 1 \ldots N$, $m = 1 \ldots M$
Output: $\{\sigma'_{i'}(m)\}$, $i = 1 \ldots \omega$,  $m = 1 \ldots M$

repeat
    Compute $I(\sigma_i; \sigma_j)$ $\forall$ $i, j$ ;
    Select $X = \{\sigma_{i^*}, \sigma_{j^*(i^*)}\}$
      where $^*$ denotes "max" ;
    Select "reference" subset $Y$
      as next-highest info. spins,
      respectively, with $\sigma_{i^*}$ & $\sigma_{j^*(i^*)}$;
    Estimate $P(X, Y), P(X), P(Y)$
      by counting appearance frequencies
      (i.e., maximum-likelihood estimate) ;
    Initialize $P(X'|X)$ randomly, as
      matrix from uniform distribution ;
    Find $P(X'|X)$ via minimization of:
      $\min_{P(X'|X)} \{I(X;Y) - \Lambda I(X';Y)\}$
      using Blahut-Arimoto iterative solution [455] ;
    Use $P(X'|X)$ & values of $X$ to draw set
      of $M$ samples for the new hyperspin $X'$ ;
    Re-define system: $\left\{ [\{\sigma'_{i'}\} \setminus X] , X' \right\} \leftarrow \{\sigma'_{i'}\}$
until less than $\omega$ system elements remain
  OR  $I(\sigma_{i^*}; \sigma_{j^*(i^*)}) <$ significance threshold;
```

4.2.5 Note About Validation Procedures

Our algorithm, as outlined in the previous section, can in principle be applied, in *ad hoc* fashion, to any given set of Ising trajectories. Since the correlation structure of Ising data itself differs markedly with the external temperature T, our information-theoretic "neighborhood" and selection criteria will, themselves, depend on T. For example, for $T \to 0$ and $T \to \infty$, the maximal information value $I\left(\sigma_i; \sigma_j\right)$ can refer to spins at arbitrary distances from one another on the lattice.

Is this desirable? Before answering this question, we must first consider that the knowledge of T as the (sole, in this case) macroscopic parameter required to determine the phase space structure and statistics for the system is something that *emerges* from an RG treatment of our Ising model. Without such a prior knowledge, we could not have anticipated which external parameters to vary; ultimately, this type of knowledge must be a *product* of our approach, not an input. Indeed, we discuss later how this very same lack of knowledge allows our method to characterize large-scale properties in unfamiliar systems.

Nonetheless, in order to ensure that the program laid out in Algorithm 1 results in a set of large-scale degrees of freedom that are consistent with the correct macroscopic variables, we will test our approach preliminarily on data that expresses separately each distinct part of the system's phase space. For our Ising spins, this means testing data taken at temperature values $T < T_c$, $T > T_c$, and near the critical point. With each of the three cases is associated a distinct qualitative and quantitative RG flow, or approach to fixed-point behavior. It is this flow, as well as the way in which the microscopic spin variables are found to combine, yielding some set of macroscopic degrees of freedom, that we wish to observe for each case.

In the context of the usual RG transformation, such as Kadanoff's "block spin" [445] technique, we can expect to apply an operator $\mathcal{T}(\sigma_i, \sigma'_{i'})$ of fixed form to any system sample (that is, realization of the microstate $\{\sigma_i\}$), at any temperature, with the result that the system will flow to the correct fixed point under repeated transformations. The action of $\mathcal{T}(\sigma_i, \sigma'_{i'})$ is the same at all temperatures.

This temperature-independence is easy to ensure when "nearest-neighbor" interactions are defined in real space, since Ising spins change only their values, and not their positions, with temperature. In order to ensure that our notion of locality remains independent of temperature during the validation of our procedure, we first "train" our algorithm by learning a sequence of subsets $\{X, Y\}^\alpha$, or *merge order* – according to which the (hyper-) spins at coarse-graining iteration α will be compressed – on data taken at various mixed temperatures, centered about T_c.

Once this sequence of the "most local" and next-most closely interacting "reference spins" across all included temperatures has been established, we switch to a data set taken at particular values of T where, skipping the establishment of the information-space "neighborhood" and selection of X and Y (steps 1-3 in Algorithm 1), we coarse-grain according to the merge order instead. This also helps us to respect, at least approximately, the known translational symmetry of the Ising system (only temporarily, during our validation procedures) by building distinct neighborhoods of roughly equivalent information content, over which we can later take averages.

For the results presented in the following section, we learned separate merge orders for 10 data subsets or *pools* of 1000 samples each, and then applied these to coarse-grain multiple different sets of fixed-temperature data. Each of these fixed-temperature data sets contained 400 samples, chosen at random from the full availability of samples taken at the desired value of T across all 10 pools. The direct application of our Algorithm 1 to fixed-temperature data in an *ad hoc* manner, without learning a merge order, will be addressed later (see Section 4.5).

4.3 Results

We first examine typical outputs for our Information Bottleneck compressions. Figure 4.2 reiterates the action of steps 6-8 in Algorithm 1: the $2 \times M$ vector of activities associated with the maximally correlated spin pair X is compressed to form a new trajectory for hyperspin X'.

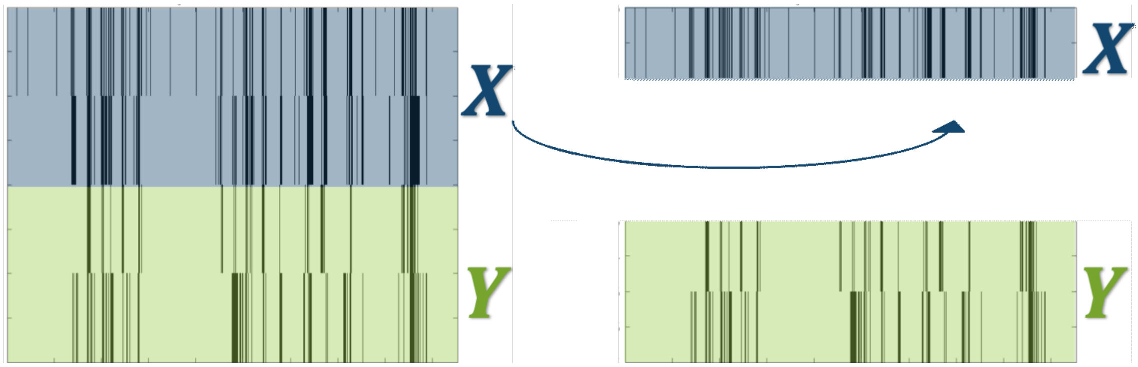

Figure 4.2: In this schematic, the binary two-vector of activities for X is compressed to a one-vector trajectory for the new hyperspin X'; those for Y remain as before.

This algorithm then introduces hyperspin X' into the system, with distribution $P(X')$, for which the mutual information with the reference spins, $I(X';Y)$ is intended to approximate $I(X;Y)$. The relationships between these and other information quantities are illustrated in Fig. 4.3. At each "RG" iteration, the mutual information is highest between X' and one of the spins in X. Sometimes, the Bottleneck samples a trajectory for X' that is identical to the trajectory for one of these spins; other times, the compression is almost equally similar to the trajectories for both spins in X . Meanwhile, the magnitudes of the next-nearest neighbor information values, $I(X_1, Y_1)$ and $I(X_2, Y_2)$, and the opposite pairings, $I(X_1, Y_2)$ and $I(X_2, Y_1)$, can be seen relative to the "shortest-scale" information $I(X_1, X_2)$. For mixed-temperature data, we expect the latter to decrease over successive iterations, as this quantity set our effective length scale in the absence of real-space structure.

We would like to somehow compare our coarse-grained systems, consisting of many

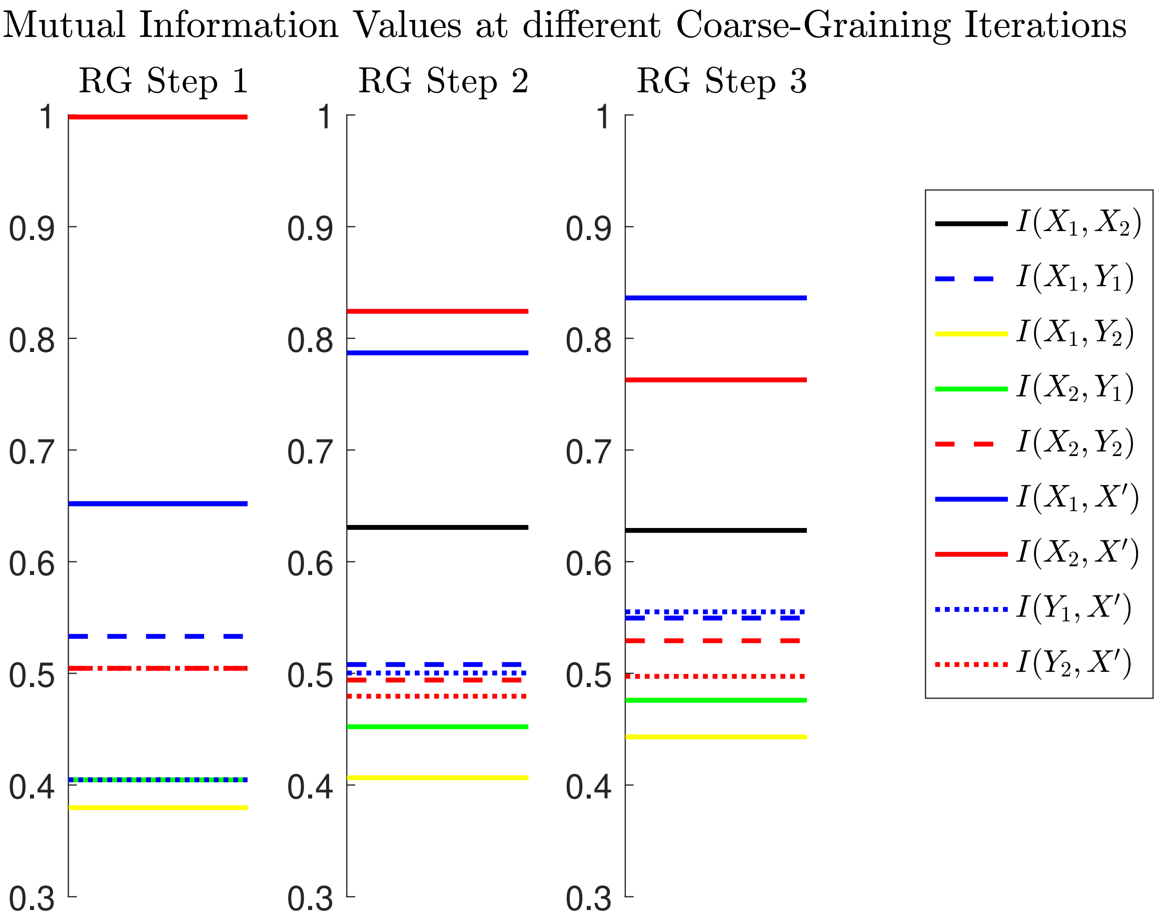

Figure 4.3: For "RG Step 1," the black line indicating $I(X_1, X_2)$ lies under the blue line for $I(X_1, X')$ and therefore the dotted lines for $I(X_2, Y_1)$ and $I(Y_1, X')$ overlap, as do the lines for $I(X_2, Y_2)$ and $I(Y_2, X')$.

such hyperspins after many such successive compressions, to the classic results of the real-space RG for the square-lattice Ising model. For our method to serve as a viable replacement for traditional RG techniques where knowledge of (spatial) interaction structures is unavailable, it is crucial in particular that the large-scale degrees of freedom $\{\sigma'_{i'}\}$ – the X' of later iterations – bear an identifiable relation to the usual, local spin averages that characterize this system.

One way to accomplish this would be to study in some detail the *receptive fields* at various stages of coarse-graining. These are the full sets of original spins accounted for by a given hyperspin X' – including those already replaced by previously-added hyperspins that have X' as their compression, and so on, recursively. We would hope to verify that our algorithm generates objects similar to Kadanoff's block spins (that is, hierarchical sets of geometrically proximal hyperspins; see Appendix 2). Given that Algorithm 1 operates without any reference to the geometrical structure of the 2D lattice, we cannot expect our receptive fields to correspond exactly to the familiar block structure. For instance, we will not necessarily observe equally-sized receptive fields across the lattice at n-fold reductions in the number of spins, for integer values of n. Yet, the "blocks" should grow in size predictably with the effective (spatial) length scale $\ell = \frac{L}{\sqrt{N_{\rm rem}}}$, where $N_{\rm rem}$ is the number N of original spins in the system, minus the number of coarse-graining iteration; recall that each iteration removes two (hyper-) spins and adds X'.

Several representative examples of the receptive fields that coalesce at different stages of coarse-graining are reproduced in Fig. 4.4. Beyond $\ell \approx 5.5$ (corresponding to 3961 iterations of the algorithm, with few of the original spins remaining), the hyperspins shown begin to take the shape of large, delocalized clusters that span the lattice. By $\ell = 10$, some fields contain spins that are entirely nonlocal to one another in real space.

What is happening to the information at these stages? We plot the values $I(X_1, X_2)$ associated with the shortest interaction scale in information space at each iteration in Fig. 4.5. The mutual information drops significantly in the vicinity of $\ell = 10$, where the spin neighborhoods inferred by our algorithm were seen to become nonlocal. We find that this value corresponds with reasonable precision to the maximal correlation length $\xi(\tilde{T}_c)$ in the finite system studied in Ref. [458], in addition to our own numerical measurements of the same quantity (not shown), where $\tilde{T}_c$ refers to the shifted, "critical" temperature at which all observables that diverge in the thermodynamic ($L \to \infty$) limit experience a rounded peak for L finite.

In particular, $I(X_1, X_2)$ begins rapidly approaching, starting around $\ell = 10$, a value nearly indistinguishable from the average (mean or median) pairwise informa-

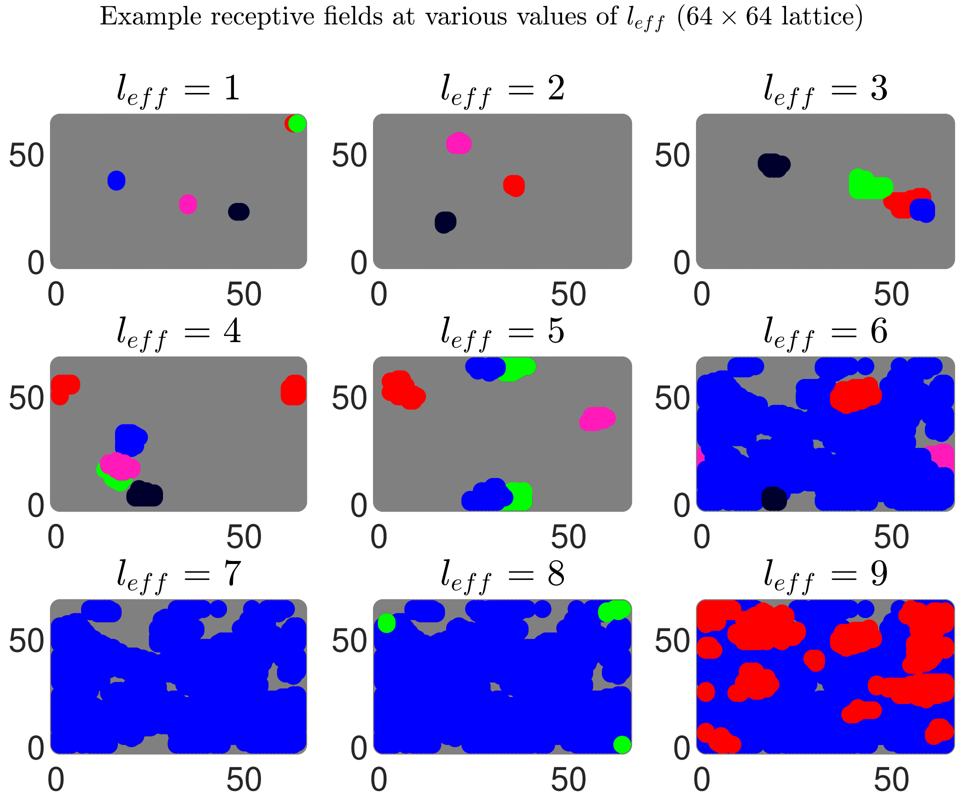

Figure 4.4: Already by $\ell \approx 5.5$, some of the receptive fields have grown large. The above represent five consecutive values of X', at different stages of the coarse-graining, at approximately the indicated value of ℓ. *Note*: Fields are overlaping where the "highest information" value is between a small (hyper)-spin and an existing, adjacent hyperspin of comparable or even larger size. Growing clusters of hyperspins in this manner is not desired at small ℓ, and shall be discussed in more detail below.

tion between elements in the system. In other words, beyond this point, all interactions in the system are viewed by our Algorithm 1 as equally or indistinguishably "local". This is essentially the same as reaching an effective lattice on the order of the correlation in the usual RG, at which point the hyperspins become effectively decoupled, and suggests a natural stopping point for our coarse-graining procedure: by analogy with the role usually played by the correlation length, we can iterate our algorithm until the "shortest-scale" mutual information $I(X_1, X_2)$ begins to saturate or becomes statistically insignificant.

The quantity $I(X_1, X_2)$ can be related in an even more precise manner to standard

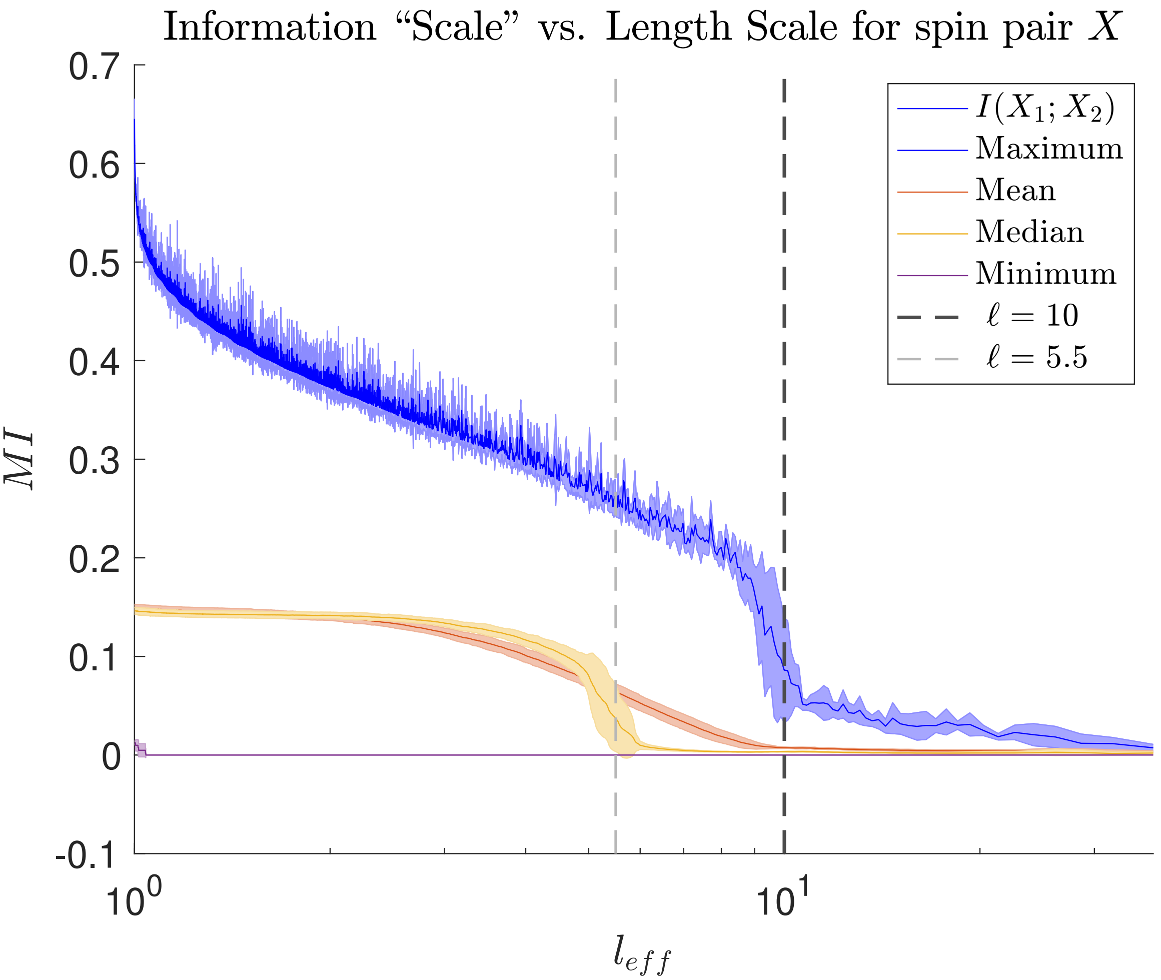

Figure 4.5: This plot shows the maximal information value $I(\sigma_i^*; \sigma_j^*(i^*))$ used to select X across all iterations of the algorithm on a mixed-temperature data set, relative to the the bounding and average information values at the corresponding values of ℓ. Specifically, the “Maximum” curve records the highest mutual information value observed at the corresponding scale (it is computed from the same matrix $I(\sigma_i, \sigma_j)$, but using a different Matlab function); the “Minimum” curve records the lowest value of $I(\sigma_i, \sigma_j)$ at a particular scale, using a Matlab function analogous to that for the “Maximum” curve. Similarly, the “Mean” and “Median” curves are different averages computed using the same data. The behavior of $I(X_1, X_2)$ between $\ell \approx 0.5$ and $\ell \approx 8$ suggests a power-law scaling.

analyses of the Ising model. For binary variables, mutual information is closely tied to the linear correlation between hyperspins $\sigma_{i'}$ and $\sigma_{j'}$:

$$C_{i'j'} = \frac{\langle \sigma_{i'} \sigma_{j'} \rangle_m - \langle \sigma_{i'} \rangle_m \langle \sigma_{j'} \rangle_m}{\sqrt{\mathrm{Var}\left[\sigma_{i'}\right] \mathrm{Var}\left[\sigma_{j'}\right]}}. \tag{4.20}$$

Since $I(X_1, X_2)$ represents the degree of statistical dependence at the shortest interaction (here, length) scale present in each coarsened version of the system, the corresponding value of $C_{i'^* j'^*}$is itself related to the nearest-neighbor correlation between "block" (hyper-)spins.

The nearest-neighbor correlation, or pairwise average $\langle \sigma_i'^* \sigma_j'^* \rangle_m$ for i'^* and j'^* representing hyperspins which are adjacent (in information space), has a known dependence on the temperature (see Fig. B.1). Renormalization group theory predicts that repeated RG transformations will cause this quantity to "flow" to either the extreme of $\langle \sigma_i \sigma_j \rangle_m = 1$ or $\langle \sigma_i \sigma_j \rangle_m = 0$, depending on whether the initial temperature is above or below T_c.

Equally simple and demonstrative of the RG flow the behavior of the net (squared) magnetization, $M^2 = \frac{1}{N^2} \left(\sum_i \sigma_i \right)$, which approaches $M^2 = 1$ if the starting temperature is below T_c and tends toward zero (apart from the usual $\frac{1}{\sqrt{N}}$ fluctuations associated with a sum uniformly distributed, binary random numbers for a finite system) for starting temperatures above T_C. Figure 4.6 records observed values of M^2 across the effective length scales corresponding to successive coarse-graining iterations, and Fig. 4.7 the nearest-neighbor hyperspin correlations. Each curve in both plots represents a subset of Ising samples taken at a different temperature surrounding T_c; all were formed by applying the merge order learned on mixed-temperature data to these fixed-temperature data subsets. That all curves (excepting $T = 2.27$) gather near one of two specific values in these plots suggests that our algorithm has, correctly, recovered the existence of our Ising model's distinct low- and high-temperature phases.

We study also the covariance, or *connected correlation* $\langle \sigma_i' \sigma_j' \rangle - \langle \sigma_i' \rangle \langle \sigma_j' \rangle$ associated with activity trajectories for each realization of the coarse-grained pair X. This "truncated" version of the spin-spin correlation function does not flow to separate fixed-point values, but has the advantageous property of decaying toward zero at

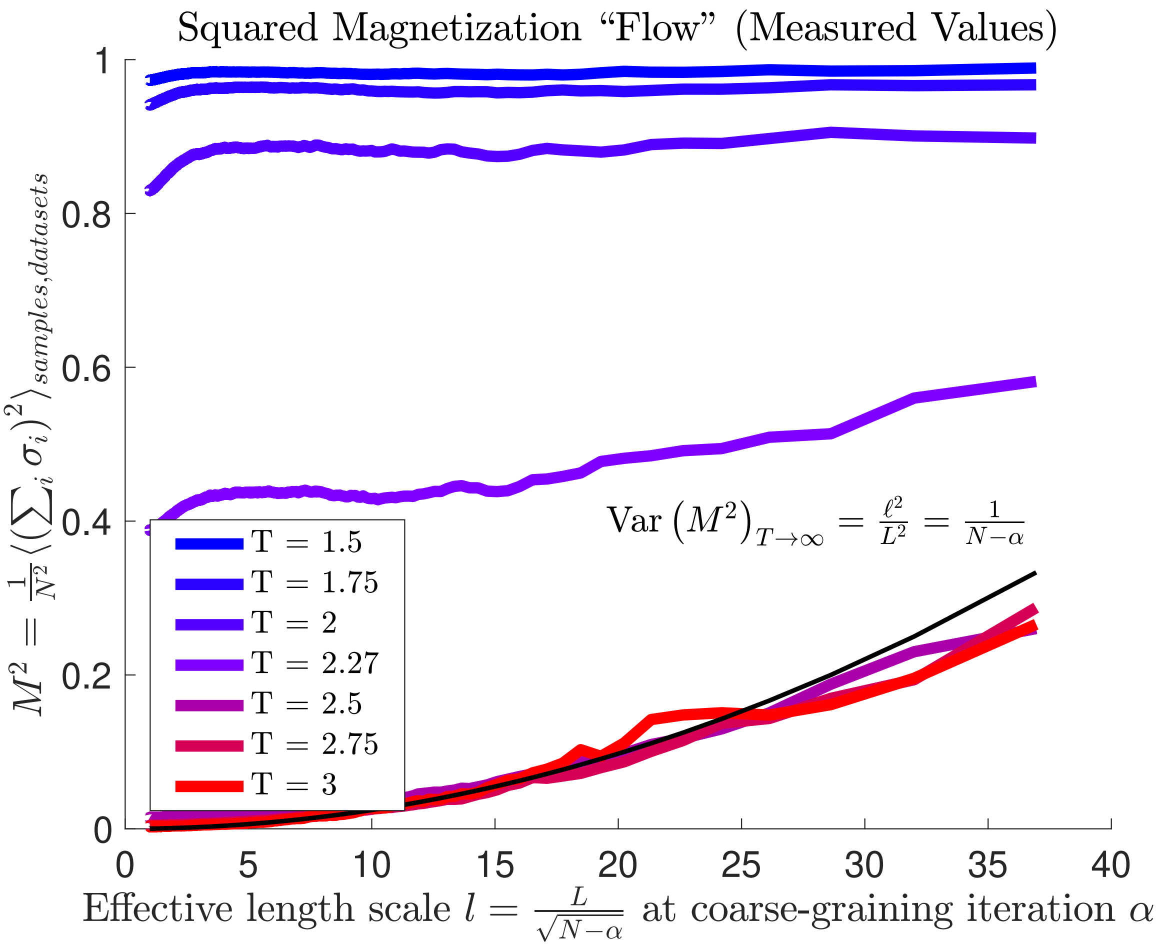

Figure 4.6: "Flow" of the squared magnetization. For temperatures $T < T_c$, the flow is toward the maximal value of $\langle \sigma_i \sigma_j \rangle = 1$; for $T > T_c$ it is toward a separate value, which is given by the variance of an unbiased random walker. That is, one starts at position "zero" and, flipping a coin to take successive steps to either the left or right, finishes a distance $\frac{1}{\sqrt{N}}$ units away from zero after N steps. In general, these results suggests that our algorithm is reccovering the existence of distinct low- and high-temperature phases. We discuss later the finite-size and symmetry-breaking effects that prevent the $T = 2$ line from reaching $M^2 = 1$.

long range. Since X is to represent the shortest length scale in the system, which in the real-space RG is always rescaled to restore the unit lattice spacing, our measured covariances correspond to *nearest-neighbor (block) spin correlations.*

The nearest-neighbor, connected correlations for block spins are predicted (see [443]

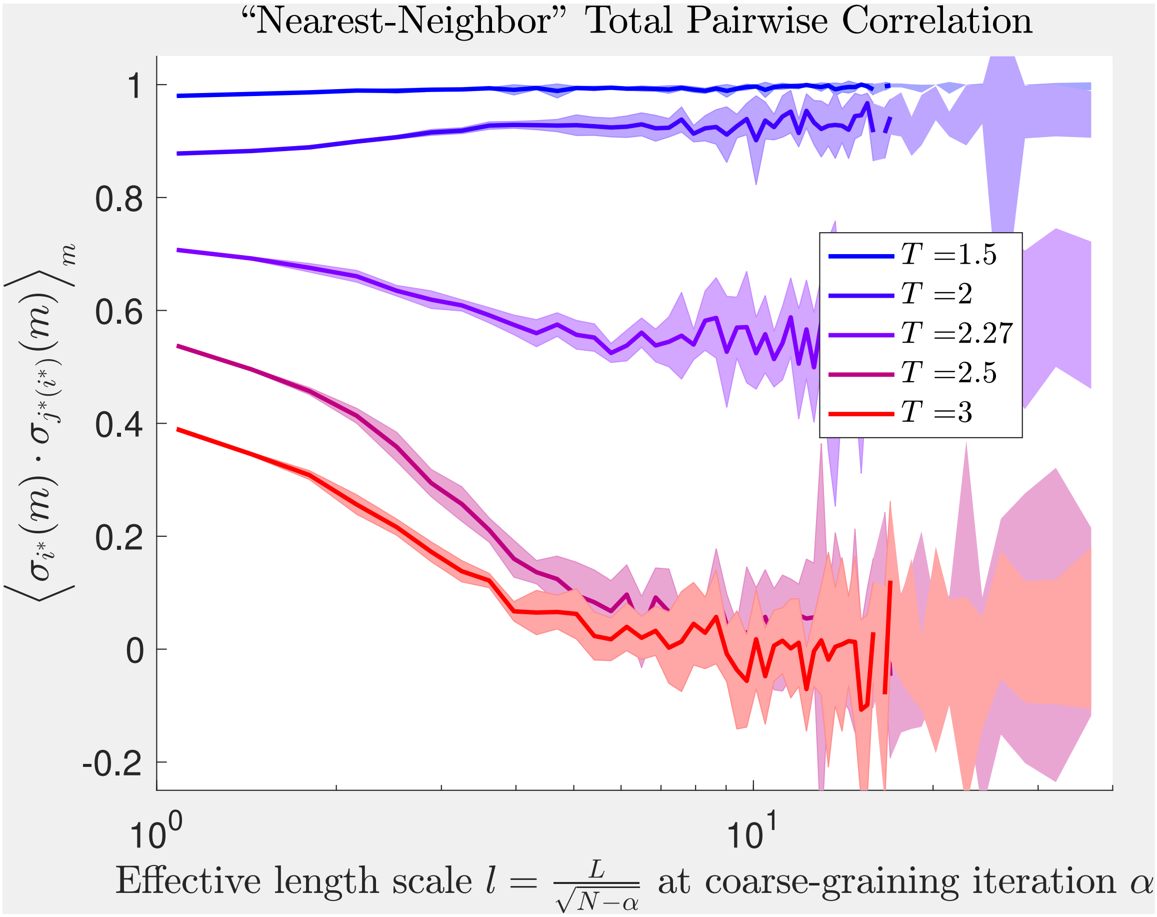

Figure 4.7: "Flow" of the total pairwise, nearest-neighbor correlations. For temperatures $T < T_c$ flow is toward the maximal value of $\langle \sigma_i \sigma_j \rangle = 1$; for $T > T_c$ it is toward $\langle \sigma_i \sigma_j \rangle = 0$. These results serve as an independent confirmation that the algorithm is able to distinguish the low- and high-temperature phases of the model.

and Appendix B for details) to transform as

$$G(r = 1, t) = \ell^{-\eta} G\left(t\ell\right), \tag{4.21}$$

where ℓ is the block size, t is the reduced temperature, and $\eta = \frac{1}{4}$ (for the infinite-size system). Note that the effective temperature responsible for the statistics of the transformed system is given by $t' = t\ell^{y_t} = t\ell$. Since the total pairwise nearest-neighbor correlation $\langle \sigma_i' \sigma_j' \rangle$ for this system is known analytically (see Appendix B) and $\langle \sigma_i' \rangle \langle \sigma_j' \rangle$ is just the squared magnetization, we can combine the previous relation for the effective temperature with Eq. (4.21) to predict the nearest-neighbor connected correlation after a block transformation of any size ℓ.

While for block spins ℓ typically takes on integer values (multiples of the microscopic lattice spacing), there is no such constraint in our case. Therefore we interpolate the effective system size as $\ell = \frac{L}{\sqrt{N_\alpha}}$ to plot our results, with

$$N_\alpha = N - \alpha \tag{4.22}$$

the number of constituent hyperspins at a given iteration α (since one coarse-graining step removes two hyperspins X and adds the single, new hyperspin X' to the system).

In theory, the connected correlation $G(r, T)$ reduces to the pairwise product $\langle \sigma_i \sigma_j \rangle$ for for $T > T_c$, because the spontaneous magnetization M vanishes there. Yet for a system of finite size, as for our 64 × 64 lattice, there can in principle be a nonzero magnetization everywhere, since a true phase transition and critical behavior exist only in the thermodynamic limit. Although the RG blocking transformation itself (being a local operation that acts on only a small subset of spins) remains agnostic to whether the system is infinite or not, the magnetization values $M(T)$ must be corrected for finite-size effects.

Details regarding finite-size corrections to the magnetization can be found in Appendix B; there, we argue that corrections to the total pairwise correlations are small, and neglect them in what follows. We plot the measured connected correlations $G(1, T)$ associated with the coarse-grained spin pair X at each iteration of the algorithm, along with their predicted values for the finite-size system, against the effective length scale ℓ in Fig. 4.8. There, the shaded curves of various colors represent our measured results for data of different temperatures, while the dots of corresponding colors are the predictions.

Immediately it is evident that, despite some quantitative agreement (i.e., the start

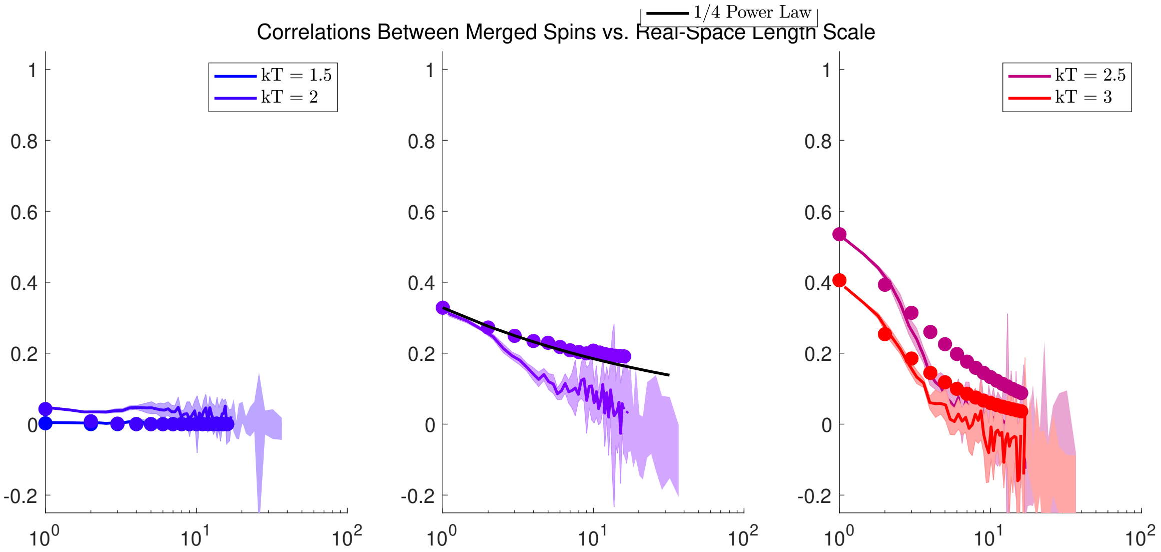

Figure 4.8: Nearest-neighbor connected correlations, separated into distinct phases for visual clarity. Finite-size corrections allow us to predict the start points for curves of different temperatures, as well as their general shapes, but fail to predict their exact rates of decay accurately beyond a scale of $\ell \approx 5$. Further work is needed to determine how to predict correlations in this finite-size system. For example, how must infinite-size scaling exponent value $\eta = \frac{1}{4}$ be modified?

points and early decays for high and low temperature curves), there are disparities between our results and our theoretical predictions. The disagreement is exacerbated in the neighborhood of the critical point, which strongly suggests that $\eta = \frac{1}{4}$ is not the correct exponent to describe the decay of correlations with the effective length scale in this system of finite size.

The most significant quantitative disagreements occur, as might be expected, at scales ℓ beyond the point at which we have showed that the algorithm begins learning nonlocal receptive fields. In addition, our algorithm recovers many qualitative features – such as the existence of distinct high- and low-temperature phases – with remarkable robustness, given that it was imparted with no knowledge of the system's local spatial structure, or any other information that could be used to identify our data as having come from an Ising model.

In the following section, we discuss several reasons for these discrepancies between our results and predictions, particularly in the region corresponding to $\ell > 5$, and what changes are needed to see them coincide.

4.4 Improving the Quantitative Agreeement

We have seen that the receptive fields generated for our hyperspins on mixed-temperature data resemble those of traditional "block spins," at least qualitatively, for small values of ℓ. This suggests that our algorithm is learning an information-space representation of the spin neighborhoods, and "compressed" large scale degrees of freedom, that both recover the known, spatially-informed interaction structure of the system. This structure is detected implicitly, without explicit reference to space.

In addition, we have been able to reproduce certain key properties and statistics of these large-scale degrees of freedom, with varying degrees of success. These included a qualitative reproduction of the RG flow for the (unconnected) spin-spin correlation, as well as a semi-quantitative prediction for the decay of connected correlations. In all cases, our measured values for $G(r = 1, T)$ fall to zero with different rates than do our predictions. This requires explanation; in this section, we attempt to elucidate why we could have expected such discrepancies, and which modifications can facilitate better agreement between our results and quantitative predictions.

4.4.1 Symmetry Breaks in Current Implementation

In classic block spin approaches to renormalization of the Ising model, the system is first partitioned into multiple subsets to be coarse-grained, representing different groups of spins that each interact on a particular (length) scale. Then, all these subsets are transformed simultaneously according to the coarse-graining rule, and the lattice is *rescaled* to restore the system's local geometry. Such a strategy preserves the translational invariance of the system, since the receptive fields for the block spins are free to "slide" across the lattice along any spatial dimension, and all interactions on a given scale are removed at once.

Here, our notion of interaction *scale* and its restoration by *rescaling* exist in information space, where there may be a clear ordering of "shortest" distances, but on a continuum rather than the discrete set of possible lengths that arise naturally as multiples of the lattice spacing. In selecting only a single, short scale to coarse-grain out, we break this symmetry: once X is removed and replaced with X', the $N-\alpha-1$ interactions of X' with all other (hyper-) spins is considered on equal footing with the rest of the interactions that were "next in line" if not for X.

Renormalization approaches are predicated on respecting the symmetries of the system at hand, and without this requirement there is no guarantee of reproducing the correct RG flow [443, 444]. Still, we expect our Algorithm 1 to respect the aforementioned translation invariance at least approximately for a given set of data, which we can illustrate as follows. At the inception of coarse-graining, we select the single pair X of spins corresponding to the maximally informative pair of trajectories to coarse-grain out. To within the precision with which we measure the mutual information, this maximally informative pair can be any one of several possible candidates. When this first pair is replaced with X', the mutual information between the sampled trajectory for X' and the remaining spins is bound by the Data Processing Inequality [210] to be of a value no higher than the already-measured information content between X and those same spins. Therefore, the subsequent steps will entail, at worst, a choice between removing the effective interaction of X' with one of the remaining spins and coarse-graining another short scale, associated with another pair of the remaining spins. Ideally, our algorithm would continue, at each iteration, to select X in such a way that all interactions at ranges close to that of the initially removed scale are replaced (on average) before revisiting the first collection of hyperspins produced, and so on for "longer" scales later.

This last condition can, of course, be mandated, but instead it emerges, in an approximate form, as a feature of our algorithm in the early stages of coarse-graining. Yet, as noted earlier, our measured receptive fields grow larger than predicted for standard block spins at the equivalent reduction in scale, with fully delocalized clusters of appearing already by the time $\ell \approx 5.5$. These clusters form because, beyond a certain stage of coarse-graining, the mutual information between some large hyperspin, representing the X' of a previous iteration, and some other (hyper)-spin is found to be higher than the information values between comparably sized hyperspins. Many of these events are consecutive in the later stages of coarse-graining, with large, "hoarder" hyperpins simply assuming neighbors in their periphery.

Thus coarse-graining (and information loss) goes faster in some localized spatial sections than others, meaning that our "renormalized" lattices exist at multiple different effective temperatures, rather than a uniform t', which explains why the usual argument leading to Eq. (4.21) fails to predict the rates of decay in Fig. 4.8. What causes such a bias in mutual information values that leads to cluster formation and a faster compression in localized regions of the lattice?

The information content of new hyperspins about the states of those already in the system is controlled by the Information Bottleneck compression process. As we have seen, $I(X', Y)$ is bounded by $I(X, Y)$, but it is how the former compares to the "next" scales in the system – in other words, the mutual information values associated with the interactions between all other spin pairs $\{\sigma_i, \sigma_j\}$, $\forall\ i, j \notin X$ – that determines whether X' is soon likely to participate in a coarse-graining event.

Since the Blahut-Arimoto solution of the Information Bottleneck problem posed in Eq. (4.7) guarantees only a locally, not globally, optimal encoding $P(X'|X)$, it is possible that sampled trajectory for hyperspin X' at a given iteration can be "ranked" at an inappropriate information-theoretic distance from the rest of the system's constituent hyperspins for subsequent iterations. In reviewing and re-computing the output values of $I(X', Y)$ for specific iterations, we have observed that multiple suboptimal compressions can be found with probabilities comparable to the of the maximal value of $I(X', Y)$. While the effect of chosing suboptimal compressions has yet to quantified, their ability to distort the lattice geometry and effective temperature should not be overlooked. We comment on how to avoid this problem in the following section, among other directions for future work.

In addition to setting a new hyperspin at an inappropriate scale, the version of our algorithm used to generate the plots of Section 4.3 did not distinguish between *symmetric* and *asymmetric* solutions $P(X'|X)$. Since, by our convention, the first elements of X and Y (σ_{i^*} and $\sigma_{k^*(i^*)}$ in the notation of Section 4.2.4) were demarcated as "next-shortest" scale of interaction, $P(X'|X)$ could have had different values for the two cases $X = \{-1, +1\}$ and $X = \{+1, -1\}$. This, too, has the potential to destroy translational invariance, but be ameliorated by the corrective strategies discussed below.

4.4.2 Simple Modifications May Restore Symmetry

We have argued that several disparities in physical observables between our measurements and theoretical predictions are due principally to to a violation of the inherent symmetries in our test case of 2D Ising data. Namely, our algorithm fails to predict a form for our large-scale degrees of freedom that resemble the usual block-average variables at large ℓ, and their decay rate of their associated connected correlations, due to aspects of our compression process that encourage favored solutions at these later scales. Again we stress that, while not all data will incorporate such symmetries, it is important to show that our method is capable of finding solutions consistent with them, where they do exist.

In order to ensure that the chosen value of $P(X'|X)$ represents the best possible solution for preserving the information about all longer interaction scales in the system beyond that associated with X, we so far have modified the Information Bottleneck step (7) of Algorithm 1 for all future work to repeat the compression step until a clear, maximum $I(X', Y)$ is identified. For our chosen Bottleneck parameter value $\Lambda = 100$, we observe that that the matrix $P(X'|X)$ takes one of $\sim \mathcal{O}(10)$ possible values, so we expect that running the Information Bottleneck algorithm until convergence $\sim \mathcal{O}(10^2)$ times exhaust these values and find the global optimum. Then the resulting matrix of $P(X'|X)$ is symmetrized by replacing the columns for which the spins that comprise X assume opposite values – $P(X'|X = \{-1, +1\})$ and $P(X'|X = \{+1, -1\})$ – with their arithmetic average.

A viable alternative would be to average the realized values of $P(X'|X)$, weighting realizations in proportion to the proximity of $I(X', Y)$ to its maximum observed value. A more rigorous way to do this would be to "bootstrap" our estimation of $I(X', Y)$ by subsampling the trajectories X and Y with replacement and repeating the compression for each subsample as described above.

If these changes do not restore the prerequisite translational symmetry, additional measures can be taken to enforce it. An extreme example would be keep track of which pairs of the system have been coarse-grained and mandate that all the original spins must be partnered, in information-rank order, before coarse-graining the newly formed hyperspins – and so on for higher levels (scales) of coarse-graining. In the special case that these hyperspins are not allowed to serve even as reference spins, this paradigm corresponds roughly to the approach taken in Ref. [439], although the pair-selection rule was based on linear correlation and the compression rule identical for all spins therein. Other variations that interpolate between this extreme and the precise version of algorithm presented here exist as well – for example, we can introduce an artificial-temperature "noise" to the selection of X and Y to allow coarse-graining away from the "most informative" pair (which can later be annealed), but this could ruin our attempt to emulate the Kadanoff-Wilson strategy of removing the shortest scales first. Moreover, since all of these "corrections" entail ways to hard-code translational invariance, they do not solve the problem of *discovering* symmetries directly from the data.

Furthermore, not all systems will exhibit translational symmetry. For our future plans, we propose a way to use method to coarse-grain more general, unfamiliar systems.

4.5 Learning New Physics – or Biology (Future Work)

So far, we have developed a data-driven coarse-graining procedure, derived from an information-theoretic recasting of the variational RG for physical models, that recovers several key aspects of the long-length scale behavior of a 2D Ising system. In the previous section, we discussed how our algorithm could be modified to better account for the symmetries of this and other systems, and briefly sketched how we would expect our results to change.

These modifications were introduced in the context of applying our algorithm to mixed-temperature data, where we demanded that our notion of information-theoretic neighborhoods – our implicit proxy for locality – remain independent of T. As stated in Section 4.2.5, this requirement was deemed necessary for our test case, but not realizable for generic data sets: this would, in turn, require knowledge of the temperature as a relevant parameter in a correct large-scale description of the model.

How can we use the methods developed here to build a pipeline capable of isolating an appropriate set of large-scale degrees of freedom for general complex data? One central finding in Section 4.3 was that the onset of full breakdown for our predictions occurred at roughly $\ell \approx 8.8$, which corresponds to the rough size of the maximal ("critical") value of the Ising correlation length $\xi(T)$ on a 64×64 lattice [458]. Our algorithm appears to have "learned" to stop coarse-graining according to a criterion – the inability to distinguish outstandingly high values of the mutual information – that is in essence equivalent to that of the standard (real-space) RG.

What might happen, then, if we apply our algorithm directly to Ising data at a fixed temperature T, so that the theoretical correlation length (and ostensibly, our notion of locality) can vary across independent runs? Preliminary work (not depicted) indicates that the information $I(\sigma_{i^*}, \sigma_{j^*})$ will saturate at an iteration corresponding to a length scale ℓ that corresponds (roughly) to the known correlation length at T. This intuitive results suggests that our algorithm should indeed detect, automatically, the scale at which the independent, macroscopic degrees of freedom "live" for a given system.

More precisely, applying our algorithm without first training as done here throughout, will not permit us to reconstruct the entire phase-space structure of the macroscopic system, but may nonetheless uncover correct, compact description *within* local regions of the phase space. Thus, based on the statistical simplicity of the data, as ascertained by our algorithm, and typical values of $I(\sigma_{i^*}, \sigma_{j^*})$, we should be able to distinguish low-temperature, high-temperature, and near-critical Ising systems – and in general, say whether any new system is possessed by a nontrivial cascade of scales, indicative of being poised a critical point – without first averaging over a preestablished "merge order." In the future, we will continue to develop both modes of operation.

Chapter 5

Outlook: Where do we stand?

In the Age of Big Data, it is sometimes tempting to dive, full-force, into the nuances of multitudinous measurements without clear predictive goals. Here, we have first asked whether this inclination should be resisted for the specific case of large-scale biological networks, and then explored alternative strategies, rooted in theoretical physics, that are consistent with the goal of generating interpretable, predictive, and quantitative frameworks for conveying the rich complexity of living systems – all while escaping the diminishing returns associated with detailed, whole-network inference.

Specifically, it was argued from the outset, in Chapter 2, that the reconstructive inference of large-sale biological networks without regard or prior reflection about how the product is to be *used* not only distracts, but in certain cases detracts, from characterizing any collective or emergent aspects of the underlying biological processes. In particular, reconstructions can answer many questions, provided they are isolated to a given, typically small scale; yet many interesting behaviors, some clinically relevant, appear only when aggregating over (or, emphatically span multiple) scales. Reverse-engineering is ill-equipped to determine at which level such dynamics "live."

For Chapter 3, we focused on a system for which reverse-engineering would be expected to show no specific insights about the macroscopic behavior, but whose dynamics nonetheless at the whole-circuit scale could be summarized in terms of simple input-output relations defined over a two-dimensional manifold. While the components of this system that determined this behavior at a functional level could be divided into active and inactive subsets at the microscopic scale, it was not clear that such a partitioning can be found it general. We then switched our focus the main question of how to derive an appropriate set of coarse, functional variables with which to describe a system's macroscopic behavior, directly from microscopic activity data (of the same type used in reverse-engineering).

The data-driven coarse-graining method developed in Chapter 4 to answer this question, based on renormalization group ideas, was able to rediscover key aspects of the large-scale properties associated with a "network" of Ising spins. Our information-theoretic approach, which emphasized preserving the mutual information between the subset of system elements to be removed via coarse-graining and the rest of the system's constituent spins, bears certain overall similarities to the approaches outlined in Refs. [26, 439]. In the former, the goal was to maximize the information between the compression of some "system" of interest and its "environment" of surrounding variables representing the larger scales; the statistics of both are represented implicitly, by neural network architectures, something we would discourage here since training the machine seems to violate the statistical learning principle of trying to avoid solving a "hard" problem as an intermediate step toward solving an "easy" one [36].

The latter [439] applied a simple coarse-graining rule (summation with rescaling) to time series of neuronal activities, even suggesting that the system exhibits certain nontrivial (critical) scalings. As done here, trajectories were compressed in pairs, but with a selection criterion based on linear correlation. It is unclear how this criterion, as well as the static coarse-graining rule, could be generalized to other systems.

While other approaches to data compression and model reduction applicable to biological systems have been developed in the machine learning community, our renormalization-inspired approach offers a distinct advantage. Namely, whereas popular deep architectures based on Variational Autoencoders [459] and Restricted Boltzmann Machines [446] concentrate their efforts on compressing the totality of input data, with the hope that high-level features emerge naturally due to the bottleneck formed by the smaller number of hidden units comprising successive layers, our method selects for "large-scale" features from the start. This is accomplished by shifting the objective from maximixing (the equivalent of) $I(X', X)$, the information held by a compression about the original data itself, to optimizing $I(X', Y)$ – in fact, with a minor penalty for preserving $I(X', X)$ – in our approach.

In fact, our information-theoretic compression can be thought of as a generalization of the "pooling" operation introduced in the common in many deep machine learning architectures. Pooling techniques originated in the context of convolutional architectures – neural networks that are regularized by imposing translational invariance – where they serve to "sub-sample" the outputs of previous layers in such a way that still preserves certain meaningful statistics. Our Bottleneck output reduces trivially to "max" pooling when it leads to a set of samples X' that identically matches that of one element in X (i.e., decimation) and is closer to"average" pooling when the set of samples for X' is formed by a majority rule. Yet, whereas the regularization methods for convolutional networks are tailored to a specific type of input data (typically, images in which the same types of objects are not tied to a particular location, or even orientation), our method of compressing information from previous "layers," or length scales, makes no such assumptions. Rather than mandating in advance what counts as meaningful information, we exploit the Information Bottleneck notion of using relevance to the statistics of a reference variables (here, the variable Y containing information about longer length scales) as its own distortion measure. We favor the latter approach in the hopes that it can inform future attempts to *learn* the symmetries underlying data sets of interest.

Eventually, we intend our method to be applied in the context of massively parallel biological data, but it is sufficiently general that it can be applied to other systems for which large amounts of "microscopic" activity data can be obtained as well. The crucial requirement is that one be able to reliably estimate the joint entropies or mutual information values (if not the entire joint probability distributions) associated with pairs of activity variables. If done well [316], this should not require significantly more data points than measuring the first two standard moments (means and variances) of the corresponding distributions. Provided activity measurements are system-wide (i.e., taken in parallel for all elements), we can coarse-grain trajectories consisting of sequential time series data (as in mRNA expression levels [257], neuronal firing [439], or even economics [373]), "spot" measurements taken under different conditions (as in genetic "knockout" and other perturbative experiments [212, 181, 180, 182]), or independent samples from "equilibrium" distributions [228, 229], as done here.

Having laid the foundation over the last four Chapters for the application of this method to such diverse systems in the future, I now conclude with a brief anecdote.

I presented an early form of the project in Chapter 3 at the 2016 March Meeting of the American Physical Society. While some basic motivation to pursue the study of biological systems using tools and perspectives from theoretical physics was latent in my decision to attend Emory's graduate program, I was particularly inspired at a session of this conference, where I encountered a particularly lucid articulation of the inadequacy of traditional thinking for certain imminent problems in human health.

In an invited talk, Robert Austin motivated part of his presentation (entitled "Evolution, Physics, and Cancer: Disrupting Traditional Approaches" [460]) by referencing a news article, published by Nature, on the status of The Cancer Genome

Atlas (TCGA) project. As reported in the article [461], the collaboration claimed to have identified some 10^7 “cancer-related” mutations. Aggressive cataloguing had revealed “little commonality between tumors,” and, moreover, had not – as Austin’s fellow presenter Chris Adami argued should be possible, within the hour at the very same session [462] – wrought any satisfying *functional* theories about the disease.

Can it be that such an achievement went unobtained – and perhaps even remains unattainable – not due to insufficient time or effort, or because the collaborators were looking at the ”wrong” set of 10,000 tumors, but due instead to a mismatch between the *scales* of observation and those of the desired functional understanding?

I hope by this point that I have rendered my own position in a clear and defensible way. Namely, I am pessimistic that the TCGA discovering a ten million-and-first mutation would add categorically new or predictive knowledge of the types called for in the throughout this Dissertation – much less a functional, intuitive, or actionable understanding of oncogenic processes and pathologies (I have the impression that Austin agrees). My own contribution to resolving the tension between competing approaches, at least on the boundaries between physics, biology, statistics, and machine learning, has been to offer a serious critique and several modest alternatives by which my successors might afford a new means of looking at living systems.

There is much left to do, but this is an exciting time for my field(s) of study. Far more now than even when I began my Ph.D., I believe we can hope for true harmony and progress among various factions – I am honored to join them, as a professional.

Appendix A

Student Research Problem

Try On Your Own: Become a Reverse-Engineer

By now we hope to have made a convincing case for our contention that different reverse-engineering methodologies are, in general, best-suited for answering different types of questions. We have reviewed the most prominent such questions, and illustrated how the "goals" fulfilled by specific algorithms are really manifestations of their underlying assumptions about what should count as an interaction.

Since no one definition of biological interaction can be considered more "correct" than the others in all contexts (different algorithms merely capture different aspects of the same system), a diversity of goals and operational idiosyncracies might be viewed as a blessing rather than a curse. Yet choices should be made at the outset regarding what one wishes to learn by doing reverse-engineering, because these choices inform which algorithms are best suited for the job.

In this section, we simulate the conditions under which the need for such choices arises. Imagine that you have just been handed a set of high-throughput data, for a system whose interaction architectures have not yet been fully mapped. Follow the series of prompts in the box to embark on an exploratory challenge with a representative set of actual experimental data.

Consider a set of multi-electrode recordings from the retina of a salamander (we thank M. Berry for providing us with data from [221]; download link at `https://figshare.com/articles/bint_fishmovie32_100_mat/5009840`). As explained in detail in the `README.txt` file, the data consists of the responses from $p = 160$ ganglion cells to the presentation of a naturalistic stimulus – in this case, a short (~ 20 sec) movie of a fish tank, repeated $n = 297$ times. The activity of each neuron is binarized as 0 (when the neuron is not firing an action potential) and 1 (when it is firing an action potential) within discrete time bins of length 20 ms.

1. Of the methods discussed in this Chapter, which are clearly applicable to this particular set of data? Are there any which are not?

2. What kinds of predictions might a researcher want to make using this data?

   Consider multiple levels of analysis, from single nodes in the neuronal network (Will removing a single node cause the network to collapse? Can we predict a future value for a given neuron, given the values of certain others?) to multiple nodes (Are there any functional groups that seem to be operating as a unit? Are there hub structures present?) to the entire system as an emergent whole (What can we say about the percentage of time the system is silent, versus when it is spiking? What other information would we need to say something about the "typicality" of the recorded networks, with respect to their structural and dynamical properties?).

3. Crowdsourcing [463] – the idea that conglomerate predictions, made by combining the wisdom of many independent thinkers, are more accurate than those of any individual – is a popular strategy in DREAM competitions [464, 465] (for recent examples, see the closed Sage Bionetworks-DREAM Breast Cancer Prognosis (DREAM7, 2012), NIEHS-NCATS-UNC DREAM Toxicogenetics (DREAM8, 2013), and ICGC-TCGA DREAM Somatic Mutation Calling

(DREAM 8.5-9, 2013-2014) Challenges). Yet we have seen that different reverse-engineering methods often yield disparate – even antagonistic or contradictory – predictions. For which combination of the following algorithms would you feel comfortable following the "wisdom of crowds" (say, averaging the results, or taking majority rules)?

Think about ARACNe, CLR, Bayesian networks (static and dynamic), MaxEnt approaches, and possibly other methods. Given the assumptions these methods make, would you take the union or intersection of the set of results produced by Bayesian methods and ARACNe? MaxEnt and CLR? Other combinations? When do you think crowdsourcing in general is a good strategy?

Appendix B

Ising Model Details for Coarse-Graining Analysis

Connected Correlations: Analytics

For Kadanoff-Wilson "block spins" of linear dimension ℓ, the general connected correlation function $G \equiv \langle \sigma_i \sigma_j \rangle - \langle \sigma_i \rangle \langle \sigma_j \rangle$ transforms as

$$G(r, t, h) = \ell^{-2(d-y_h)} G\left(\frac{r}{\ell}, t\ell^{y}_{t}, h\ell^{y}_{h}\right), \tag{B.1}$$

under coarse-graining [443]. Here, d is the lattice dimension, exponents y_t and y_h are to be determined, $h = 0$ due to the modeling choices made Sec. 4.2.3. For the 2D Ising model we study here [443], $y_t = \frac{1}{\nu} = 1$ and $y_h = \frac{15}{8}$, so that we have $2(d - y_h) = \eta = \frac{1}{4}$. Dropping the reduced magnetic field h from Eq. (B.1) and restricting ourselves to nearest-neighbor correlations only simplifies the desired scaling relation to

$$G(r = 1, t) = \ell^{-\eta} G\left(t\ell^{y}_{t}\right), \tag{B.2}$$

where we have dropped the dependence on $r = \frac{1}{\ell}$ (i.e., the separation, in the original units, between the center points of nearest-neighbor *blocks*) because it is a constant.

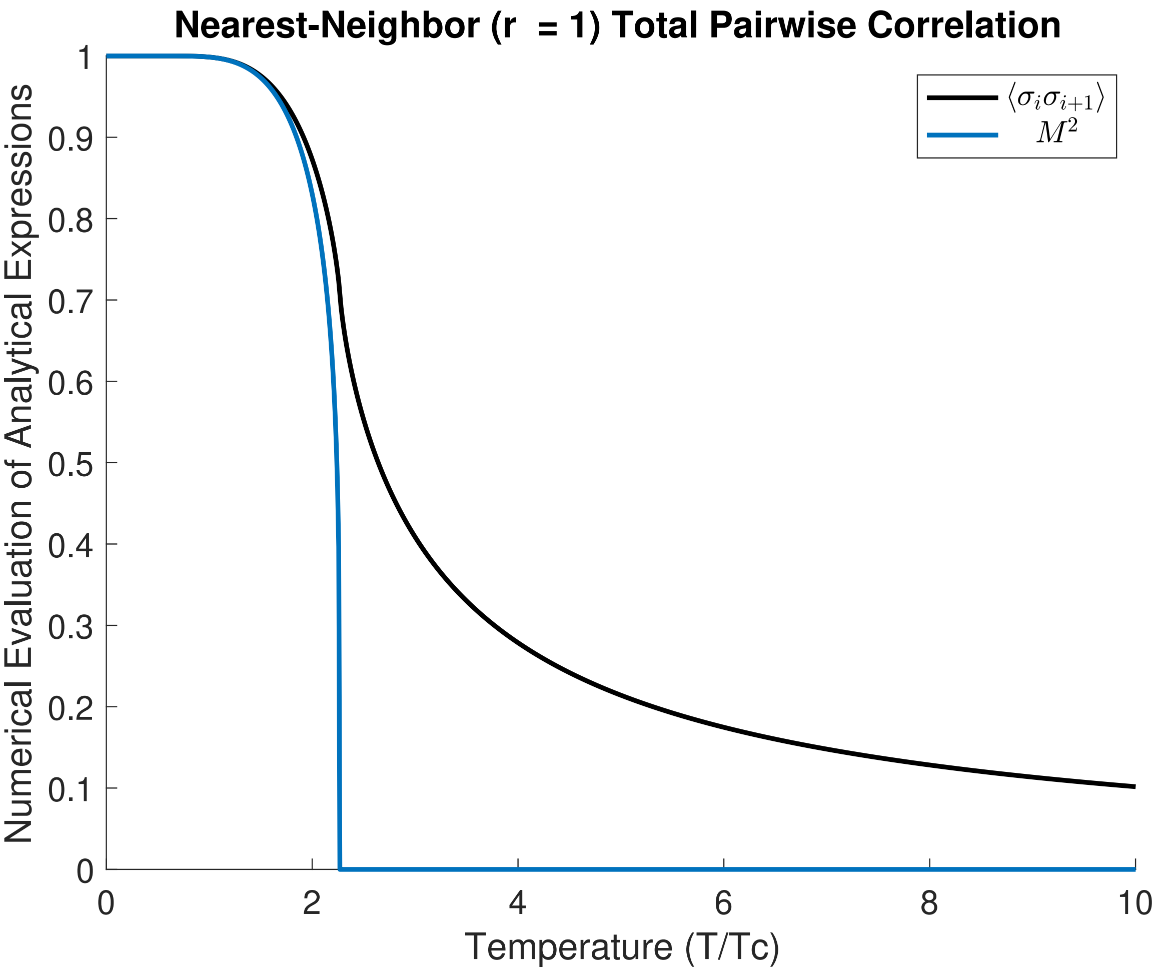

Figure B.1: The correlation function for nearest neighbors, $G(r = 1, T)$, as given in Eq. (B.3). For spins which are not nearest neighbors, but separated by distance r, the square of Onsager's magnetization, given by Eq. (B.5) is the limit of the pairwise correlation $\langle\sigma_i\sigma_j\rangle$. Here, where $r = 1$, the two curves coincide for small T. These are numerical evaluations of the known analytical results.

The total pairwise correlation has a known analytical form for this system. Using the notation $\langle i, j\rangle$ to denote nearest-neighbor pairs, as in Eq. (4.17), we can simplify the general results of Refs. [466, 467] to write

$$\langle\sigma_i\sigma_j\rangle = \begin{cases} \frac{1}{2}\left(1+s^{-2}\right)^{\frac{1}{2}}\left[\left(1-s^{-2}\right)\tilde{K}(s^{-2})+1\right], & T < T_c, \\ \frac{1}{2}\left(1+s^{-2}\right)^{\frac{1}{2}}\left[\left(s^{-2}-1\right)\tilde{K}(s^{2})+1\right], & T > T_c, \end{cases} \tag{B.3}$$

where $s \equiv \sinh 2\beta J$ and $\tilde{K}(\cdot)$ is an elliptic integral of the first kind, written in terms of the elliptic modulus and scaled by a constant factor so that $\tilde{K}(0) = 1$:

$$\tilde{K}(\cdot) = \frac{2}{\pi} \int_0^{\frac{\pi}{2}} \frac{d\phi}{\left(1 - (\cdot)^2 \sin^2 \phi\right)}. \tag{B.4}$$

Numerical evaluations of Eq. (B.3) are depicted in Fig. B.1.

We can build the connected correlation $G(r, t)$ by subtracting from Eq. (B.3) the squared magnetization per spin [466], which we evaluate here using Onsager's famously unpublished [468] solution for the spontaneous magnetization:

$$M_0(T) = \left(1 - k^2\right)^{\frac{1}{8}}, \tag{B.5}$$

with $k \equiv \frac{1}{\sinh^2 2\beta J} = s^{-2}$ in our isotropic Ising model. It is well known that the square of Eq. (B.5), which is also evaluated in Fig. B.1, is the limit of $\langle \sigma_i \sigma_j \rangle$, for spins σ_i and σ_j at distance $r \to \infty$ [468] from each other.

Since the correlation length ξ_ℓ measured in units of the spacing between block spins on the rescaled lattice, is shorter than the correlation length ξ, measured in terms of the original lattice spacing, the renormalized system will appear further from criticality after a coarse-graining transformation ℓ. If we assume [443] that, at least inside the critical region, the new, "effective" reduced temperature can be computed as $t_\ell = t\ell^{y_t}$, we can predict the nearest-neighbor correlation at a given coarse-graining step from the initial temperature and effective length scale.

Finite-Size Corrections

In order to take finite-size effects into account [444], we assume that the correlation length at $T = T_c$ in the finite system cannot exceed the lattice size L. Near the critical point (i.e., $t,\ h = 0$ for $L \to \infty$), we expect the behavior

$$\xi \sim |t|^{-\nu}, \tag{B.6}$$

that is, a diverging correlation length. This suggests

$$T_{\max} - T_c \sim \xi^{-\frac{1}{\nu}} \sim L^{-\frac{1}{\nu}} \tag{B.7}$$

for a lattice of finite size L, where $T_{\max}$ is the temperature at which the diverging thermodynamic quantities exhibit a rounded peak rather than a singularity [469]. Rewriting the scaling relations for these thermodynamic observables with the help of Eq. (B.7) yields for the magnetization

$$M = L^{-\frac{\beta}{\nu}} \cdot M_{t,L}(tL^{\frac{1}{\nu}}), \tag{B.8}$$

where $\beta = \frac{1}{8}$ and $M_{t,L} \sim (tL^{\frac{1}{\nu}})^{-\frac{7}{8}}$ for $T > T_c$ [458].

The above form for $M_{t,L}$ is valid in the region where $tL > 1$ (at $T_c \approx 2.2692$, this shape function becomes a constant). Near the critical point, it is not guaranteed to predict accurate values for M; see Fig. (B.2). Here, To correct the magnetization at arbitrary temperatures, we plot $ML^{\frac{\beta}{\nu}}$ against $tL^{\frac{1}{\nu}}$ to observe the data collapse anticipated by Eq. (B.8) and correct for finite L numerically.

Specifically, we first measure $M(T) = \frac{1}{N^2}\langle\sqrt{\sum_i \sigma_i}\rangle_s$, where $s = 1, 2, \ldots$ denotes a given set of samples recorded at a particular value of T. We then generate a scatter plot for the "collapsed" curve $ML^{\frac{\beta}{\nu}}$ vs. $tL^{\frac{1}{\nu}}$ (see Fig. B.3), create a lookup table for this curve by binning individual data points along the tL axis. Finally, we use this average curve to represent $M_{t,L}(tL^{\frac{1}{\nu}})$ numerically, with $L = 64$.

The result of such a numerical approximation to the Finite-size scaling of the magnetization is illustrated in Fig. B.2, where we plot numerical evaluations of Eq. (B.8) for $T > T_c$ against the aforementioned, measured values of M, for different lattice sizes L. It is clear that there is a substantial net magnetization even beyond $T = 2T_c$.

In principle, the total pairwise correlation $\langle\sigma_i\sigma_j\rangle$ must also be corrected for finite

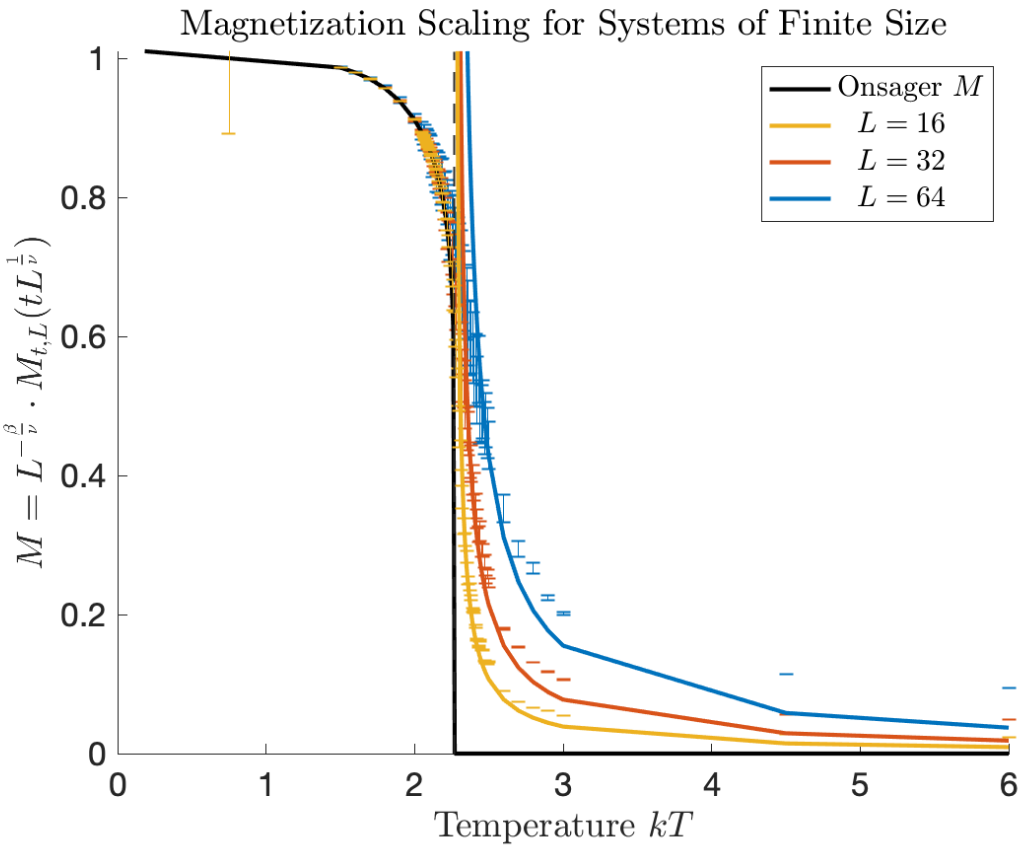

Figure B.2: An illustration of the result printed as Eq. (B.8), for four different values of L; the points with error bars represent the mean and standard deviation of our data. In the neighborhood of T_c, this form cannot be used to predict M accurately, so we approximate it numerically.

size. In general, we have [443]

$$G\left(r,t,L\right) \sim |r|^{-(d-2+\eta)} \cdot G\left(\frac{1}{L}, tL^{\frac{1}{\nu}}\right). \tag{B.9}$$

In practice, we find that this quantity does not require significant corrections (see Fig. B.4). By combining our numerical finite-size corrections to M and subtracting the corresponding values of $M^2_{L=64}(T)$ from the unconnected correlation, where the effective T is given by the block spin argument above, we can predict the curve traced by the connected correlation over different "length" scales ℓ.

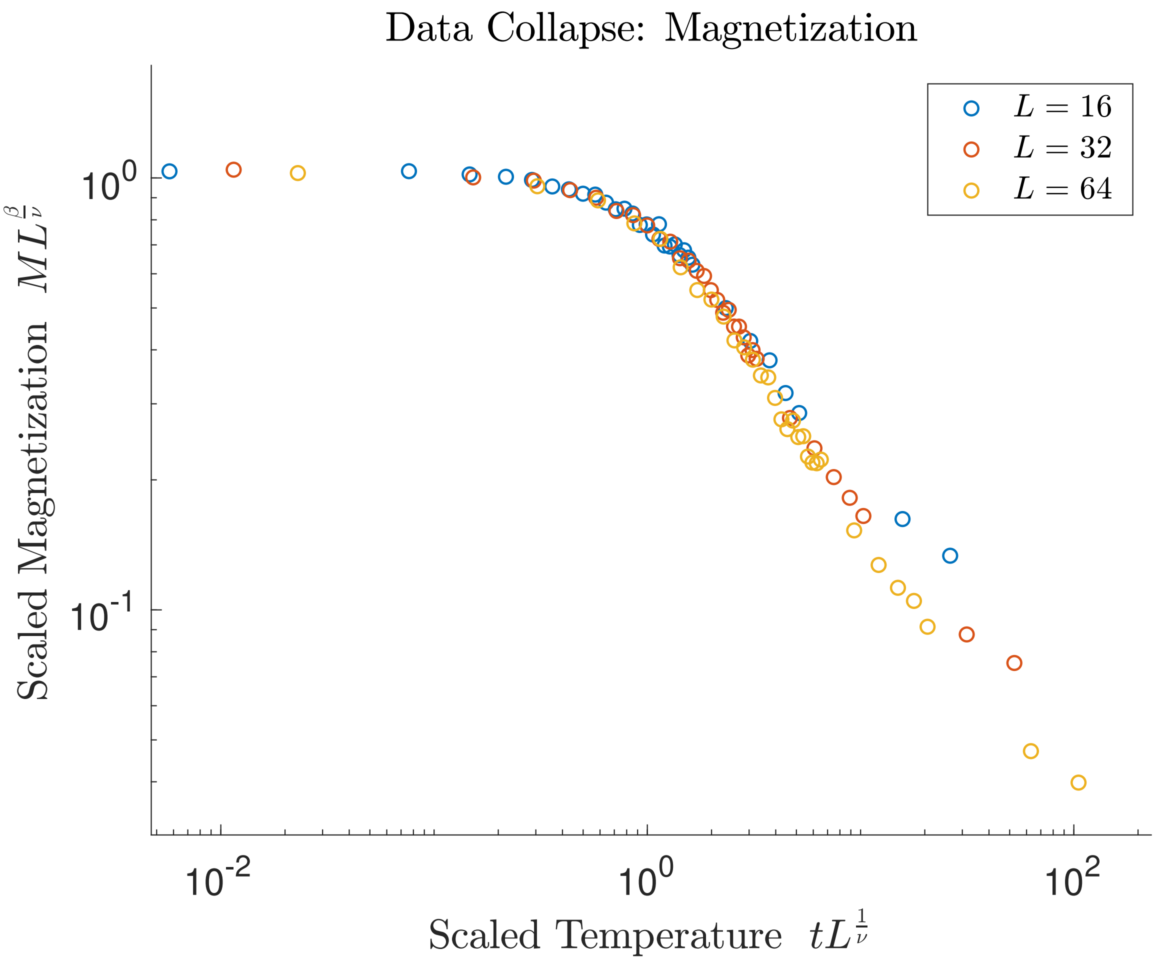

Figure B.3: Data collapse for finite magnetization: the power of $tL^{1\nu}$ in $M_{t,L}$ differs for $T > T_c$ and $T < T_c$. Open-circle markers are used to emphasize overlap.

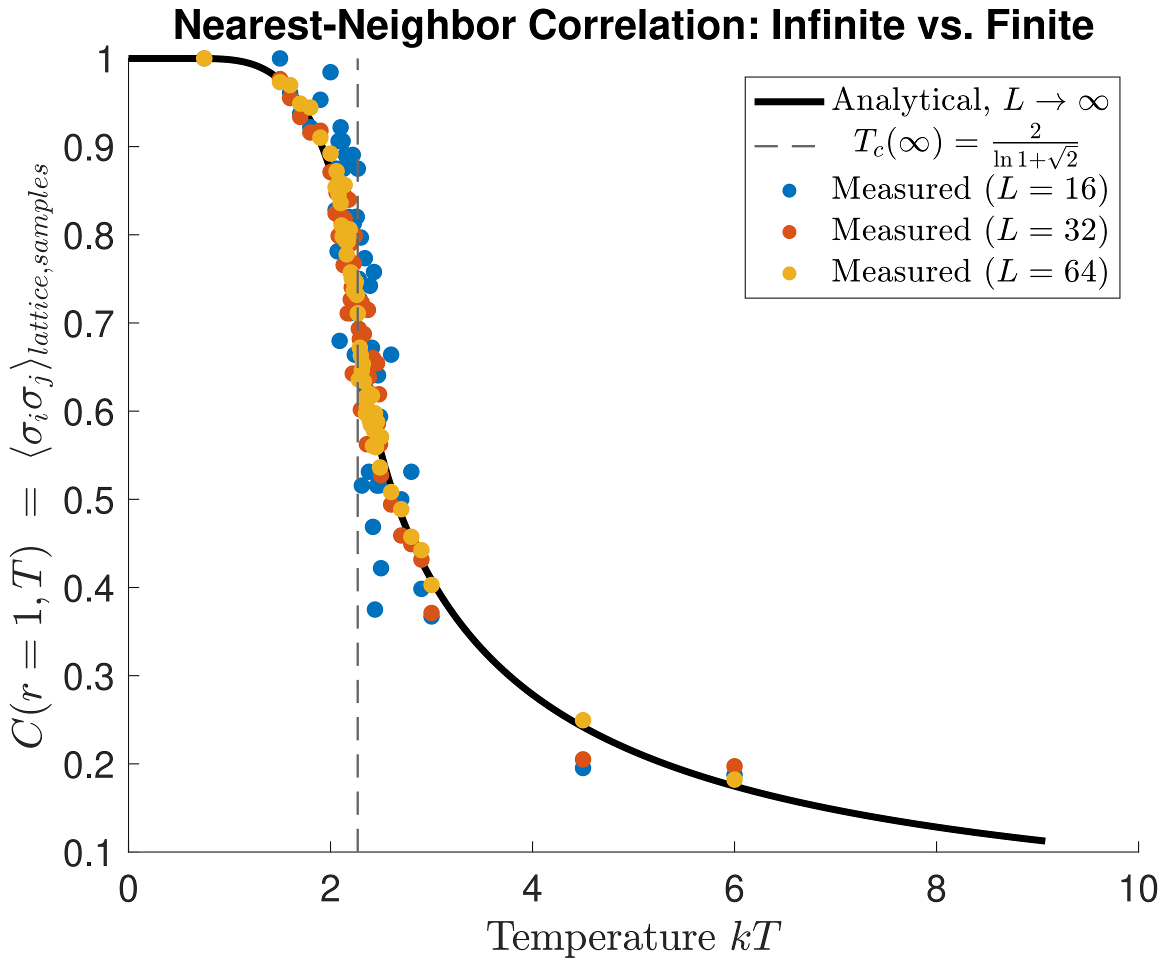

Figure B.4: The mean pairwise product between every spin on the lattice with its four nearest neighbors, averaged over 400 lattice samples each. The significantly smaller corrections required for $G(r = 1, T)$, in contrast to M, are also consistent with the observations in Ref. [458]. Lattice sizes $L = 32$ and $L = 16$ were arrived at by implementing Kadanoff-Wilson block spins transformations (majority rule, takign the sign of the average) numerically.

Bibliography

[1] E Alm and Adam Arkin. Biological networks. *Curr Opin Struct Biol*, 13(2):193–202, 2003.

[2] Francois Kepes. *Biological networks*, volume 3. World Scientific, 2007.

[3] Arturo Rosenblueth and Norbert Wiener. The role of models in science. *Philosophy of science*, 12(4):316–321, 1945.

[4] George EP Box. Science and statistics. *Journal of the American Statistical Association*, 71(356):791–799, 1976.

[5] Somendra M Bhattacharjee and Avinash Khare. Fifty years of the exact solution of the two-dimensional ising model by onsager. *Current science*, 69(10):816–821, 1995.

[6] T Ising, R Folk, R Kenna, B Berche, and Yu Holovatch. The fate of ernst ising and the fate of his model. *Journal of Physical Studies*, 21(3):3002, 2017.

[7] Tsung-Dao Lee and Chen-Ning Yang. Statistical theory of equations of state and phase transitions. ii. lattice gas and ising model. *Physical Review*, 87(3):410, 1952.

[8] Dietrich Stauffer. Social applications of two-dimensional ising models. *American Journal of Physics*, 76(4):470–473, 2008.

[9] Yi-Ping Ma, Ivan Sudakov, Courtenay Strong, and Kenneth M Golden. Ising model for melt ponds on arctic sea ice. *New Journal of Physics*, 21(6):063029, 2019.

[10] Albert-László Barabási, Natali Gulbahce, and Joseph Loscalzo. Network medicine: a network-based approach to human disease. *Nature Rev Genet*, 12(1):56–68, January 2011.

[11] Bryan C Daniels and Ilya Nemenman. Automated adaptive inference of phenomenological dynamical models. *Nature Comm*, 6, 2015.

[12] William Bialek. *Biophysics: Searching for principles.* Princeton University Press, Princeton, 2012.

[13] James Watson, Francis Crick, and Maurice Wilkins. The nobel prize in physiology or medicine 1962.

[14] Alan Hodgkin, John Eccles, and Andrew Huxley. The nobel prize in physiology or medicine 1963.

[15] Max Delbrück, Alfred Hershey, and Salvador Luria. The nobel prize in physiology or medicine 1969.

[16] Stuart A Kauffman. *A world beyond physics: the emergence and evolution of life.* Oxford University Press, 2019.

[17] Thierry Mora and William Bialek. Are biological systems poised at criticality? *J Stat Phys*, 144(2):268–302, 2011.

[18] Cédric Bény. Deep learning and the renormalization group. arxiv, 2013. *URL http://arxiv.org/abs/1301.3124.*

[19] Pankaj Mehta and David J Schwab. An exact mapping between the variational renormalization group and deep learning. *arXiv preprint arXiv:1410.3831*, 2014.

[20] Cédric Bény and Tobias J Osborne. Information-geometric approach to the renormalization group. *Physical Review A*, 92(2):022330, 2015.

[21] Cédric Bény and Tobias J Osborne. The renormalization group via statistical inference. *New Journal of Physics*, 17(8):083005, 2015.

[22] Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle. In *2015 IEEE Information Theory Workshop (ITW)*, pages 1–5. IEEE, 2015.

[23] Henry W Lin, Max Tegmark, and David Rolnick. Why does deep and cheap learning work so well? *Journal of Statistical Physics*, 168(6):1223–1247, 2017.

[24] Ravid Shwartz-Ziv and Naftali Tishby. Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810*, 2017.

[25] Satoshi Iso, Shotaro Shiba, and Sumito Yokoo. Scale-invariant feature extraction of neural network and renormalization group flow. *Physical Review E*, 97(5):053304, 2018.

[26] Maciej Koch-Janusz and Zohar Ringel. Mutual information, neural networks and the renormalization group. *Nature Physics*, 14(6):578, 2018.

[27] Shotaro Shiba Funai and Dimitrios Giataganas. Thermodynamics and feature extraction by machine learning. *arXiv preprint arXiv:1810.08179*, 2018.

[28] Shuo-Hui Li and Lei Wang. Neural network renormalization group. *Physical review letters*, 121(26):260601, 2018.

[29] Ellen de Mello Koch, Robert de Mello Koch, and Ling Cheng. Is deep learning an rg flow? *arXiv preprint arXiv:1906.05212*, 2019.

[30] Patrick M Lenggenhager, Zohar Ringel, Sebastian D Huber, and Maciej Koch-Janusz. Optimal renormalization group transformation from information theory. *arXiv preprint arXiv:1809.09632v2*, 2019.

[31] Carsen Stringer, Marius Pachitariu, Nicholas Steinmetz, Charu Bai Reddy, Matteo Carandini, and Kenneth D Harris. Spontaneous behaviors drive multidimensional, brainwide activity. *Science*, 364(6437):255–255, 2019.

[32] Jonathan R. Epp, Yosuke Niibori, Hwa-Lin Liz Hsiang, Valentina Mercaldo, Karl Deisseroth, Sheena A. Josselyn, and Paul W. Frankland. Optimization of CLARITY for Clearing Whole-Brain and Other Intact Organs. *eNeuro*, 2(3):ENEURO.0022–15.2015, April 2015.

[33] Romain Franconville, Celia Beron, and Vivek Jayaraman. Building a functional connectome of the drosophila central complex. *eLife*, 7:e37017, 2018.

[34] Steven J Cook, Travis A Jarrell, Christopher A Brittin, Yi Wang, Adam E Bloniarz, Maksim A Yakovlev, Ken CQ Nguyen, Leo T-H Tang, Emily A Bayer, Janet S Duerr, et al. Whole-animal connectomes of both caenorhabditis elegans sexes. *Nature*, 571(7763):63–71, 2019.

[35] Gustavo Stolovitzky, Robert J. Prill, and Andrea Califano. Lessons from the DREAM2 Challenges. *Ann New York Acad Sci*, 1158:159–195, March 2009.

[36] Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 2013.

[37] Joseph L Natale, David Hofmann, Damián G Hernández, and Ilya Nemenman. Reverse-engineering biological networks from large data sets. In B Munsky, WS Hlavacek, and LS Tsimring, editors, *Quantitative Biology: Theory, Computational Methods and Examples of Models*. Cambridge, MA: MIT Press)(https://mitpress.mit.edu/books/quantitative-biology), 2018.

[38] L. H. Hartwell, J. J. Hopfield, Stanislas Leibler, and Andrew W. Murray. From molecular to modular cell biology. *Nature*, 402(6761 Suppl):C47–52, December 1999.

[39] Stephen R. Proulx, Daniel E. L. Promislow, and Patrick C. Phillips. Network thinking in ecology and evolution. *Trends Ecol Evol*, 20(6):345–353, 2005.

[40] Xiaowei Zhu, Mark Gerstein, and Michael Snyder. Getting connected: analysis and principles of biological networks. *Genes Dev*, 21(9):1010–1024, May 2007.

[41] Boris N. Kholodenko. Cell-signalling dynamics in time and space. *Nature Rev Mol Cell Biol*, 7(3):165–176, March 2006.

[42] Smadar Ben-Tabou de Leon and Eric H. Davidson. Gene Regulation: Gene Control Network in Development. *Annual Rev Biophys Biomol Struct*, 36(1):191–212, 2007.

[43] Boris N. Kholodenko, John F. Hancock, and Walter Kolch. Signalling ballet in space and time. *Nature Rev Mol Cell Biol*, 11(6):414–426, June 2010.

[44] Julien O. Dubuis, Gašper Tkačik, Eric F. Wieschaus, Thomas Gregor, and William Bialek. Positional information, in bits. *Proc Natl Acad Sci (USA)*, 110(41):16301–16308, October 2013.

[45] Xiujuan Wang, Xiaomu Wei, Bram Thijssen, Jishnu Das, Steven M. Lipkin, and Haiyuan Yu. Three-dimensional reconstruction of protein networks provides insight into human genetic disease. *Nature Biotechn*, 30(2):159–164, February 2012.

[46] Nicolas E. Buchler, Ulrich Gerland, and Terence Hwa. On schemes of combinatorial transcription logic. *Proc Natl Acad Sci (USA)*, 100(9):5136–5141, April 2003.

[47] Kai Wang, Masumichi Saito, Brygida C. Bisikirska, Mariano J. Alvarez, Wei Keat Lim, Presha Rajbhandari, Qiong Shen, Ilya Nemenman, Katia Basso, Adam A. Margolin, Ulf Klein, Riccardo Dalla-Favera, and Andrea Califano. Genome-wide identification of post-translational modulators of transcription factor activity in human B cells. *Nature Biotechn*, 27(9):829–837, September 2009.

[48] Adam A Margolin, Kai Wang, Andrea Califano, and Ilya Nemenman. Multivariate dependence and genetic networks inference. *IET Syst Biol*, 4(6):428–440, 2010.

[49] Elad Ganmor, Ronen Segev, and Elad Schneidman. Sparse low-order interaction network underlies a highly correlated and learnable neural population code. *Proc Natl Acad Sci (USA)*, 108(23):9679–9684, June 2011.

[50] Dong-Yeon Cho, Yoo-Ah Kim, and Teresa M. Przytycka. Network Biology Approach to Complex Diseases. *PLoS Comp Biol*, 8(12), December 2012.

[51] Lina Merchan and Ilya Nemenman. On the sufficiency of pairwise interactions in maximum entropy models of networks. *J Stat Phys*, 162(5):1294–1308, 2016.

[52] William S Hlavacek. How to deal with large models? *Mol Syst Biol*, 5(1):240, 2009.

[53] Ernesto Estrada. *The Structure of Complex Networks: Theory and Applications.* Oxford University Press, Oxford, UK, October 2011.

[54] Hiroaki Kitano. *Foundations of Systems Biology.* MIT Press, 2001.

[55] U. Alon. Biological Networks: The Tinkerer as an Engineer. *Science*, 301(5641):1866–1867, September 2003.

[56] Albert-László Barabási and Zoltán N. Oltvai. Network biology: understanding the cell's functional organization. *Nature Rev Genet*, 5(2):101–113, February 2004.

[57] U. Alon. *An Introduction to Systems Biology: Design Principles of Biological Circuits*. CRC Press, 2006.

[58] Bernhard Ø Palsson. *Systems Biology: Properties of Reconstructed Networks*. Cambridge University Press, April 2006.

[59] Naeha Subramanian, Parizad Torabi-Parizi, Rachel A. Gottschalk, Ronald N. Germain, and Bhaskar Dutta. Network representations of immune system complexity. *Wiley Interdiscip Rev: Syst Biol Med*, 7(1):13–38, January 2015.

[60] François Jacob and Jacques Monod. Genetic regulatory mechanisms in the synthesis of proteins. *J Mol Biol*, 3(3):318–356, 1961.

[61] Faruck Morcos, Terence Hwa, José N. Onuchic, and Martin Weigt. Direct Coupling Analysis for Protein Contact Prediction. In Daisuke Kihara, editor, *Protein Structure Prediction*, number 1137 in Methods in Molecular Biology, pages 55–70. Springer New York, 2014.

[62] Najeeb Halabi, Olivier Rivoire, Stanislas Leibler, and Rama Ranganathan. Protein sectors: evolutionary units of three-dimensional structure. *Cell*, 138(4):774–786, 2009.

[63] Christoph Feinauer, Hendrik Szurmant, Martin Weigt, and Andrea Pagnani. Inter-Protein Sequence Co-Evolution Predicts Known Physical Interactions in Bacterial Ribosomes and the Trp Operon. *PLoS One*, 11(2):e0149166, February 2016.

[64] Mukesh Bansal, Vincenzo Belcastro, Alberto Ambesi-Impiombato, and Diego di Bernardo. How to infer gene networks from expression profiles. *Mol Syst Biol*, 3(1), February 2007.

[65] Jörg Linde, Sylvie Schulze, Sebastian G. Henkel, and Reinhard Guthke. Data- and knowledge-based modeling of gene regulatory networks: an update. *EXCLI J*, 14:346–378, March 2015.

[66] Tunahan Çakır, Margriet M. W. B. Hendriks, Johan A. Westerhuis, and Age K. Smilde. Metabolic network discovery through reverse engineering of metabolome data. *Metabolomics*, 5(3):318–329, February 2009.

[67] Tunahan Çakır and Mohammad Jafar Khatibipour. Metabolic Network Discovery by Top-Down and Bottom-Up Approaches and Paths for Reconciliation. *Frontiers Bioeeng Biotechn*, 2, December 2014.

[68] Boris N. Kholodenko, Anatoly Kiyatkin, Frank J. Bruggeman, Eduardo Sontag, Hans V. Westerhoff, and Jan B. Hoek. Untangling the wires: A strategy to trace functional interactions in signaling and gene networks. *Proc Natl Acad Sci (USA)*, 99(20):12841–12846, October 2002.

[69] Jaroslav Stark, Robin Callard, and Michael Hubank. From the top down: towards a predictive biology of signalling networks. *Trends Biotechn*, 21(7):290–293, July 2003.

[70] Robert J. Prill, Julio Saez-Rodriguez, Leonidas G. Alexopoulos, Peter K. Sorger, and Gustavo Stolovitzky. Crowdsourcing Network Inference: The DREAM Predictive Signaling Network Challenge. *Sci Sign*, 4(189):mr7, 2011.

[71] Raymond Cheong, Alex Rhee, Chiaochun Joanne Wang, Ilya Nemenman, and Andre Levchenko. Information Transduction Capacity of Noisy Biochemical Signaling Networks. *Science*, 334(6054):354–358, October 2011.

[72] Qiang Ni, Ambhighainath Ganesan, Nwe-Nwe Aye-Han, Xinxin Gao, Michael D Allen, Andre Levchenko, and Jin Zhang. Signaling diversity of pka achieved via a ca2+-camp-pka oscillatory circuit. *Nature Chem Biol*, 7(1):34–40, 2011.

[73] Susan J Little, Sergei L Kosakovsky Pond, Christy M Anderson, Jason A Young, Joel O Wertheim, Sanjay R Mehta, Susanne May, and Davey M Smith. Using hiv networks to inform real time prevention interventions. *PLoS One*, 9(6):e98443, 2014.

[74] Elad Schneidman, Michael J. Berry, Ronen Segev, and William Bialek. Weak pairwise correlations imply strongly correlated network states in a neural population. *Nature*, 440(7087):1007–1012, 2006.

[75] Luis MA Bettencourt, Greg J Stephens, Michael I Ham, and Guenter W Gross. Functional structure of cortical neuronal networks grown in vitro. *Phys Rev E*, 75(2):021915, 2007.

[76] Rainer Friedrich, Christel Genoud, and Adrian A. Wanner. Analyzing the structure and function of neuronal circuits in zebrafish. *Frontiers Neur Circ*, 7:71, 2013.

[77] Kwanghun Chung and Karl Deisseroth. CLARITY for mapping the nervous system. *Nature Meth*, 10(6):508–513, 2013.

[78] Ian H. Stevenson, James M. Rebesco, Lee E. Miller, and Konrad P. Körding. Inferring functional connections between neurons. *Curr Opin Neurobiol*, 18(6):582–588, December 2008.

[79] Edward T. Bullmore and Olaf Sporns. Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature Rev Neurosci*, 10(3):186–198, March 2009.

[80] Edward T. Bullmore and Danielle S. Bassett. Brain Graphs: Graphical Models of the Human Brain Connectome. *Annual Rev Clin Psych*, 7(1):113–140, April 2011.

[81] Karl J. Friston. Functional and Effective Connectivity: A Review. *Brain Connect*, 1(1):13–36, January 2011.

[82] Cristiano Capone, Carla Filosa, Guido Gigante, Federico Ricci-Tersenghi, and Paolo Del Giudice. Inferring Synaptic Structure in Presence of Neural Interaction Time Scales. *PLoS One*, 10(3):e0118412, 2015.

[83] Mary-Ellen Lynall, Danielle S. Bassett, Robert Kerwin, Peter J. McKenna, Manfred Kitzbichler, Ulrich Muller, and Edward T. Bullmore. Functional connectivity and brain networks in schizophrenia. *J Neurosci*, 30(28):9477–9487, 2010.

[84] Rebecca G Canter, Jay Penney, and Li-Huei Tsai. The road to restoring neural circuits for the treatment of alzheimer's disease. *Nature*, 539(7628):187–196, 2016.

[85] Nathan Eagle, Alex (Sandy) Pentland, and David Lazer. Inferring friendship network structure by using mobile phone data. *Proc Natl Acad Sci (USA)*, 106(36):15274–15278, September 2009.

[86] Dana Angluin, James Aspnes, and Lev Reyzin. Inferring Social Networks from Outbreaks. In Marcus Hutter, Frank Stephan, Vladimir Vovk, and Thomas. Zeugmann, editors, *Proc 21st Intern Conf on Algorithmic Learning Theory*, ALT'10, pages 104–118, Berlin, Heidelberg, 2010. Springer-Verlag.

[87] Ioannis Psorakis, Stephen J. Roberts, Iead Rezek, and Ben C. Sheldon. Inferring social network structure in ecological systems from spatio-temporal data streams. *J R Soc Interf*, 9(76):3055–3066, November 2012.

[88] Robert M. May. Stability in multispecies community models. *Math Biosci*, 12(1-2):59–79, 1971.

[89] Robert M. May. Will a large complex system be stable? *Nature*, 238(5364):413–414, 1972.

[90] Robert M. May. Network structure and the biology of populations. *Trends Ecol Evol*, 21(7):394–399, July 2006.

[91] Michael W McCoy, Benjamin M Bolker, Karen M Warkentin, and James R Vonesh. Predicting predation through prey ontogeny using size-dependent functional response models. *Amer Naturalist*, 177(6):752–766, 2011.

[92] George Sugihara, Robert M. May, Hao Ye, Chih-hao Hsieh, Ethan R. Deyle, Michael Fogarty, and Stephan Munch. Detecting Causality in Complex Ecosystems. *Science*, 338:496–500, 2012.

[93] Ethan R. Deyle, Robert M. May, Stephan B. Munch, and George Sugihara. Tracking and forecasting ecosystem interactions in real time. *Proc R Soc B*, 283(1822):20152258, January 2016.

[94] Hankyu Moon and Tsai-Ching Lu. Network Catastrophe: Self-Organized Patterns Reveal both the Instability and the Structure of Complex Networks. *Scientific Reports*, 5:9450, March 2015.

[95] Ryan Compton, Hankyu Moon, and Tsai-Ching Lu. Catastrophe prediction via estimated network autocorrelation, 2015. US Patent 9,020,875.

[96] David B Bahr, Raymond C Browning, Holly R Wyatt, and James O Hill. Exploiting social networks to mitigate the obesity epidemic. *Obsesity*, 17(4):723–728, 2009.

[97] Adam Arkin and John Ross. Statistical Construction of Chemical Reaction Mechanisms from Measured Time-Series. *J Phys Chem*, 99(3):970–979, January 1995.

[98] Jan-Hendrik S Hofmeyr and Athel Cornish-Bowden. Co-response analysis: a new experimental strategy for metabolic control analysis. *J Theor Biol*, 182(3):371–380, 1996.

[99] Adam A. Margolin, Ilya Nemenman, Katia Basso, Chris Wiggins, Gustavo Stolovitzky, Riccardo D Favera, and Andrea Califano. Aracne: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinf*, 7(Suppl 1):S7, 2006.

[100] William A. Schmitt, R. Michael Raab, and Gregory Stephanopoulos. Elucidation of Gene Interaction Networks Through Time-Lagged Correlation Analysis of Transcriptional Data. *Genome Res*, 14(8):1654–1663, August 2004.

[101] Mary J Dunlop, Robert Sidney Cox, Joseph H Levine, Richard M Murray, and Michael B Elowitz. Regulatory activity revealed by dynamic correlations in gene expression noise. *Nature Gen*, 40(12):1493–1498, 2008.

[102] Alejandro F Villaverde, John Ross, Federico Morán, and Julio R Banga. Mider: network inference with mutual information distance and entropy reduction. *PLoS One*, 9(5):e96732, 2014.

[103] Alberto de la Fuente, Paul Brazhnik, and Pedro Mendes. Linking the genes: inferring quantitative gene networks from microarray data. *Trends Genet*, 18(8):395–398, August 2002.

[104] Silvia DM Santos, Peter J Verveer, and Philippe IH Bastiaens. Growth factor-induced mapk network topology shapes erk response determining pc-12 cell fate. *Nature Cell Biol*, 9(3):324–330, 2007.

[105] Pradeep Bandaru, Mukesh Bansal, and Ilya Nemenman. Mass conservation and inference of metabolic networks from high-throughput mass spectrometry data. *J Comp Biol*, 18(2):147–154, February 2011.

[106] Richard R. Stein, Debora S. Marks, and Chris Sander. Inferring Pairwise Interactions from Biological Data Using Maximum-Entropy Probability Models. *PLoS Comp Biol*, 11(7):e1004182, 2015.

[107] M. K. Stephen Yeung, Jesper Tegnér, and James J. Collins. Reverse engineering gene networks using singular value decomposition and robust regression. *Proc Natl Acad Sci (USA)*, 99(9):6163–6168, April 2002.

[108] Timothy S. Gardner, Diego di Bernardo, David Lorenz, and James J. Collins. Inferring Genetic Networks and Identifying Compound Mode of Action via Expression Profiling. *Science*, 301(5629):102–105, 2003.

[109] Alejandro F. Villaverde and Julio R. Banga. Reverse engineering and identification in systems biology: strategies, perspectives and challenges. *J R Soc Interf*, 11(91):20130505, February 2014.

[110] C. Christensen, J. Thakar, and R. Albert. Systems-level insights into cellular regulation: inferring, analysing, and modelling intracellular networks. *IET Syst Biol*, 1(2):61–77, March 2007.

[111] Marc Vidal, Michael E Cusick, and Albert-László Barabási. Interactome Networks and Human Disease. *Cell*, 144(6):986–998, March 2011.

[112] Chang F. Quo, Chanchala Kaddi, John H. Phan, Amin Zollanvari, Mingqing Xu, May D. Wang, and Gil Alterovitz. Reverse engineering biomolecular systems using omic data: challenges, progress and opportunities. *Brief Bioinf*, 13(4):430–445, July 2012.

[113] Arthur D. Lander. The edges of understanding. *BMC Biol*, 8:40, 2010.

[114] P. W. Anderson. More Is Different. *Science*, 177(4047):393–396, August 1972.

[115] Michael Hecker, Sandro Lambeck, Susanne Toepfer, Eugene P van Someren, and Reinhard Guthke. Gene regulatory network inference: Data integration in dynamic models—A review. *Biosyst*, 96(1):86–103, April 2009.

[116] Misha B Ahrens, Michael B Orger, Drew N Robson, Jennifer M Li, and Philipp J Keller. Whole-brain functional imaging at cellular resolution using light-sheet microscopy. *Nature Meth*, 10(5):413–420, 2013.

[117] David A Schwarz, Mikhail A Lebedev, Timothy L Hanson, Dragan F Dimitrov, Gary Lehew, Jim Meloy, Sankaranarayani Rajangam, Vivek Subramanian, Peter J Ifft, Zheng Li, Arjun Ramakrishnan, Andrew Tate, Katie Z Zhuang, and Miguel A L Nicolelis. Chronic, wireless recordings of large-scale brain activity in freely moving rhesus monkeys. *Nature Meth*, 11(6):670–676, 2014.

[118] Marc Timme and Jose Casadiego. Revealing networks from dynamics: an introduction. *J Phys A*, 47(34):343001, 2014.

[119] Nigel Goldenfeld and Leo P Kadanoff. Simple lessons from complexity. *Science*, 284(5411):87–89, 1999.

[120] Timothy S. Gardner and Jeremiah J. Faith. Reverse-engineering transcription control networks. *Phys Life Rev*, 2(1):65–88, March 2005.

[121] Ankur P Parikh, Wei Wu, Ross E Curtis, and Eric P Xing. Treegl: reverse engineering tree-evolving gene networks underlying developing biological lineages. *Bioinf*, 27(13):i196–i204, 2011.

[122] Le Song, Mladen Kolar, and Eric P Xing. Keller: estimating time-varying interactions between genes. *Bioinf*, 25(12):i128–i136, 2009.

[123] M. Newman. The Structure and Function of Complex Networks. *SIAM Rev*, 45(2):167–256, January 2003.

[124] Gustavo Stolovitzky, Don Monroe, and Andrea Califano. Dialogue on reverse-engineering assessment and methods: the DREAM of high-throughput pathway inference. *Ann New York Acad Sci*, 1115(1):1–22, December 2007.

[125] Paul C. Boutros, Adam A. Margolin, Joshua M. Stuart, Andrea Califano, and Gustavo Stolovitzky. Toward better benchmarking: challenge-based methods assessment in cancer genomics. *Genome Biol*, 15(9):462, 2014.

[126] Junko Yamane, Sachiyo Aburatani, Satoshi Imanishi, Hiromi Akanuma, Reiko Nagano, Tsuyoshi Kato, Hideko Sone, Seiichiroh Ohsako, and Wataru Fujibuchi. Prediction of developmental chemical toxicity based on gene networks of human embryonic stem cells. *Nucl Acids Res*, 44(12):5515–5528, July 2016.

[127] Alexander J Gates and Luis M Rocha. Control of complex networks requires both structure and dynamics. *Scientific Reports*, 6, 2016.

[128] Mikhail Tikhonov and William Bialek. Complexity of generic biochemical circuits: topology versus strength of interactions. *Phys Biol*, 13(6), 2016.

[129] Donghyeon Yu, MinSoo Kim, Guanghua Xiao, and Tae Hyun Hwang. Review of Biological Network Data and Its Applications. *Genomics and Informatics*, 11(4):200–210, December 2013.

[130] Nicola J. Mulder, Richard O. Akinola, Gaston K. Mazandu, and Holifidy Rapanoel. Using biological networks to improve our understanding of infectious diseases. *Comput Struct Biotechn J*, 11(18):1–10, August 2014.

[131] Alberto de la Fuente, Nan Bing, Ina Hoeschele, and Pedro Mendes. Discovery

of meaningful associations in genomic data using partial correlation coefficients. *Bioinf*, 20(18):3565–3574, December 2004.

[132] Katia Basso, Adam A. Margolin, Gustavo Stolovitzky, Ulf Klein, Riccardo Dalla-Favera, and Andrea Califano. Reverse engineering of regulatory networks in human B cells. *Nature Gen*, 37(4):382–390, April 2005.

[133] Peter Csermely, Tamás Korcsmáros, Huba J.M. Kiss, Gábor London, and Ruth Nussinov. Structure and dynamics of molecular networks: A novel paradigm of drug discovery: A comprehensive review. *Pharmacology and Therapeutics*, 138(3):333–408, June 2013.

[134] Jeremiah J. Faith, Boris Hayete, Joshua T. Thaden, Ilaria Mogno, Jamey Wierzbowski, Guillaume Cottarel, Simon Kasif, James J. Collins, and Timothy S. Gardner. Large-Scale Mapping and Validation of Escherichia coli Transcriptional Regulation from a Compendium of Expression Profiles. *PLoS Biol*, 5(1):e8, 2007.

[135] Arvind Rao, Alfred O. Hero, David J. States, and James Douglas Engel. Using directed information to build biologically relevant influence networks. *Comput Syst Bioinf Conf*, 6:145–156, 2007.

[136] Gökmen Altay and Frank Emmert-Streib. Inferring the conservative causal core of gene regulatory networks. *BMC Syst Biol*, 4:132, 2010.

[137] Christoph Kaleta, Anna Göhler, Stefan Schuster, Knut Jahreis, Reinhard Guthke, and Swetlana Nikolajewa. Integrative inference of gene-regulatory networks in Escherichia coli using information theoretic concepts and sequence analysis. *BMC Syst Biol*, 4:116, 2010.

[138] Sapna Kumari, Jeff Nie, Huann-Sheng Chen, Hao Ma, Ron Stewart, Xiang Li, Meng-Zhu Lu, William M. Taylor, and Hairong Wei. Evaluation of Gene

Association Methods for Coexpression Network Construction and Biological Knowledge Discovery. *PLoS One*, 7(11):e50411, November 2012.

[139] Sisi Ma, Patrick Kemmeren, David Gresham, and Alexander Statnikov. De-Novo Learning of Genome-Scale Regulatory Networks in S. cerevisiae. *PLoS One*, 9(9):e106479, 2014.

[140] Pablo Meyer, Thomas Cokelaer, Deepak Chandran, Kyung Hyuk Kim, Po-Ru Loh, George Tucker, Mark Lipson, Bonnie Berger, Clemens Kreutz, Andreas Raue, Bernhard Steiert, Jens Timmer, Erhan Bilal, Herbert M. Sauro, Gustavo Stolovitzky, and Julio Saez-Rodriguez. Network topology and parameter estimation: from experimental design methods to gene regulatory network kinetics using a community based approach. *BMC Syst Biol*, 8:13, 2014.

[141] Roded Sharan, Igor Ulitsky, and Ron Shamir. Network-based prediction of protein function. *Mol Syst Biol*, 3:88, March 2007.

[142] Xiling Wen, Stefanie Fuhrman, George S. Michaels, Daniel B. Carr, Susan Smith, Jeffery L. Barker, and Roland Somogyi. Large-scale temporal gene expression mapping of central nervous system development. *Proc Natl Acad Sci (USA)*, 95(1):334–339, January 1998.

[143] Cecily J. Wolfe, Isaac S. Kohane, and Atul J. Butte. Systematic survey reveals general applicability of ”guilt-by-association” within gene coexpression networks. *BMC Bioinf*, 6:227, 2005.

[144] Jesse Gillis and Paul Pavlidis. “guilt by association” is the exception rather than the rule in gene networks. *PLoS Comp Biol*, 8(3):e1002444, March 2012.

[145] Richard Bonneau, David J. Reiss, Paul Shannon, Marc Facciotti, Leroy Hood, Nitin S. Baliga, and Vesteinn Thorsson. The Inferelator: an algorithm for

learning parsimonious regulatory networks from systems-biology data sets de novo. *Genome Biol*, 7:R36, 2006.

[146] Riet De Smet and Kathleen Marchal. Advantages and limitations of current network inference methods. *Nature Rev Microbiol*, 8(10):717–729, October 2010.

[147] Eran Segal, Michael Shapira, Aviv Regev, Dana Pe'er, David Botstein, Daphne Koller, and Nir Friedman. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nature Gen*, 34(2):166–176, June 2003.

[148] Nataly Kravchenko-Balasha, Alexander Levitzki, Andrew Goldstein, Varda Rotter, A Gross, Françoise Remacle, and RD Levine. On a fundamental structure of gene networks in living cells. *Proc Natl Acad Sci (USA)*, 109(12):4702–4707, 2012.

[149] H. Jeong, S. P. Mason, Albert-László Barabási, and Z. N. Oltvai. Lethality and centrality in protein networks. *Nature*, 411(6833):41–42, May 2001.

[150] Xionglei He and Jianzhi Zhang. Why Do Hubs Tend to Be Essential in Protein Networks? *PLoS Genet*, 2(6):e88, 2006.

[151] Sergei Maslov and Kim Sneppen. Specificity and Stability in Topology of Protein Networks. *Science*, 296(5569):910–913, May 2002.

[152] Ney Lemke, Fabiana Herédia, Cláudia K. Barcellos, Adriana N. dos Reis, and José C. M. Mombach. Essentiality and damage in metabolic networks. *Bioinf*, 20(1):115–119, January 2004.

[153] Shai S Shen-Orr, Ron Milo, Shmoolik Mangan, and Uri Alon. Network motifs in the transcriptional regulation network of escherichia coli. *Nature Gen*, 31(1):64–68, 2002.

[154] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network Motifs: Simple Building Blocks of Complex Networks. *Science*, 298(5594):824–827, October 2002.

[155] Roded Sharan, Silpa Suthram, Ryan M. Kelley, Tanja Kuhn, Scott McCuine, Peter Uetz, Taylor Sittler, Richard M. Karp, and Trey Ideker. Conserved patterns of protein interaction in multiple species. *Proc Natl Acad Sci (USA)*, 102(6):1974–1979, February 2005.

[156] Joshua M. Stuart, Eran Segal, Daphne Koller, and Stuart K. Kim. A Gene-Coexpression Network for Global Discovery of Conserved Genetic Modules. *Science*, 302(5643):249–255, October 2003.

[157] Matthew W. Hahn and Andrew D. Kern. Comparative Genomics of Centrality and Essentiality in Three Eukaryotic Protein-Interaction Networks. *Mol Biol Evol*, 22(4):803–806, April 2005.

[158] Hiroaki Kitano. Biological robustness. *Nature Rev Genet*, 5(11):826–837, 2004.

[159] Michael PH Stumpf and Mason A Porter. Critical truths about power laws. *Science*, 335(6069):665–666, 2012.

[160] Paola Lecca and Corrado Priami. Biological network inference for drug discovery. *Drug Discovery Today*, 18(5–6):256–264, March 2013.

[161] Edwin T Jaynes. Information theory and statistical mechanics. *Phys Rev*, 106(4):620, 1957.

[162] Edwin T Jaynes. On the rationale of maximum-entropy methods. *Proc IEEE*, 70(9):939–952, 1982.

[163] Gašper Tkačik, Olivier Marre, Thierry Mora, Dario Amodei, Michael J Berry II,

and William Bialek. The simplest maximum entropy model for collective behavior in a neural network. *J Stat Mech: Thy Exp*, 2013(03):P03011, 2013.

[164] John M Beggs and Dietmar Plenz. Neuronal avalanches in neocortical circuits. *J Neurosci*, 23(35):11167–11177, 2003.

[165] John M Beggs. The criticality hypothesis: how local cortical networks might optimize information processing. *Philosophical Trans R Soc London A: Math, Phys, Eng Sci*, 366(1864):329–343, 2008.

[166] Woodrow L Shew and Dietmar Plenz. The functional benefits of criticality in the cortex. *Neuroscientist*, 19(1):88–100, 2013.

[167] David J Schwab, Ilya Nemenman, and Pankaj Mehta. Zipf's law and criticality in multivariate data without fine-tuning. *Phys Rev Lett*, 113(6):068102, 2014.

[168] Alberto de la Fuente. From 'differential expression' to 'differential networking' – identification of dysfunctional regulatory networks in diseases. *Trends Genet*, 26(7):326–333, July 2010.

[169] Trey Ideker and Nevan J Krogan. Differential network biology. *Mol Syst Biol*, 8:565, January 2012.

[170] Maxim Grechkin, Benjamin A. Logsdon, Andrew J. Gentles, and Su-In Lee. Identifying Network Perturbation in Cancer. *PLoS Comp Biol*, 12(5):e1004888, 2016.

[171] Ian W. Taylor, Rune Linding, David Warde-Farley, Yongmei Liu, Catia Pesquita, Daniel Faria, Shelley Bull, Tony Pawson, Quaid Morris, and Jeffrey L. Wrana. Dynamic modularity in protein interaction networks predicts breast cancer outcome. *Nature Biotechn*, 27(2):199–204, February 2009.

[172] Martin Weigt, Robert A. White, Hendrik Szurmant, James A. Hoch, and Terence Hwa. Identification of direct residue contacts in protein–protein interaction by message passing. *Proc Natl Acad Sci (USA)*, 106(1):67–72, January 2009.

[173] Debora S Marks, Lucy J Colwell, Robert Sheridan, Thomas A Hopf, Andrea Pagnani, Riccardo Zecchina, and Chris Sander. Protein 3d structure computed from evolutionary sequence variation. *PLoS One*, 6(12):e28766, 2011.

[174] Lucy J Colwell, Michael P Brenner, and Andrew W Murray. Conservation weighting functions enable covariance analyses to detect functionally important amino acids. *PLoS One*, 9(11):e107723, 2014.

[175] Woodrow L Shew, Hongdian Yang, Thomas Petermann, Rajarshi Roy, and Dietmar Plenz. Neuronal avalanches imply maximum dynamic range in cortical networks at criticality. *J Neurosci*, 29(49):15595–15600, 2009.

[176] Gašper Tkačik, Thierry Mora, Olivier Marre, Dario Amodei, Stephanie E Palmer, Michael J. Berry, and William Bialek. Thermodynamics and signatures of criticality in a network of neurons. *Proc Natl Acad Sci (USA)*, 112(37):11508–11513, 2015.

[177] Dante R Chialvo. Emergent complex neural dynamics. *Nature Phys*, 6(10):744–750, 2010.

[178] Iacopo Mastromatteo and Matteo Marsili. On the criticality of inferred models. *J Stat Mech: Thy Exp*, 2011(10):P10012, 2011.

[179] John M Beggs and Nicholas Timme. Being critical of criticality in the brain. *Frontiers Physiol*, 3:163, 2012.

[180] Andrea Pinna, Nicola Soranzo, and Alberto de la Fuente. From Knockouts

to Networks: Establishing Direct Cause-Effect Relationships through Graph Analysis. *PLoS One*, 5(10):e12912, 2010.

[181] Ritsert C. Jansen. Studying complex biological systems using multifactorial perturbation. *Nature Rev Genet*, 4(2):145–151, February 2003.

[182] Frank Emmert-Streib. Influence of the experimental design of gene expression studies on the inference of gene regulatory networks: environmental factors. *PeerJ*, 1:e10, 2013.

[183] S. M. Minhaz Ud-Dean and Rudiyanto Gunawan. Optimal design of gene knock-out experiments for gene regulatory network inference. *Bioinf*, 32(6):875–883, March 2016.

[184] Federica Eduati, Alberto Corradin, Barbara Di Camillo, and Gianna Toffolo. A boolean approach to linear prediction for signaling network modeling. *PLoS One*, 5(9):e12789, 2010.

[185] Shoudan Liang, Stefanie Fuhrman, and Roland Somogyi. Reveal, A General Reverse Engineering Algorithm for Inference of Genetic Network Architectures. January 1998.

[186] Alex Greenfield, Aviv Madar, Harry Ostrer, and Richard Bonneau. DREAM4: Combining Genetic and Dynamic Information to Identify Biological Networks and Dynamical Models. *PLoS One*, 5(10):e13397, 2010.

[187] Vân Anh Huynh-Thu, Alexandre Irrthum, Louis Wehenkel, and Pierre Geurts. Inferring Regulatory Networks from Expression Data Using Tree-Based Methods. *PLoS One*, 5(9):e12776, 2010.

[188] Andreas Wagner. Reconstructing pathways in large genetic networks from genetic perturbations. *J Comp Biol*, 11(1):53–60, 2004.

[189] Diego di Bernardo, Michael J. Thompson, Timothy S. Gardner, Sarah E. Chobot, Erin L. Eastwood, Andrew P. Wojtovich, Sean J. Elliott, Scott E. Schaus, and James J. Collins. Chemogenomic profiling on a genome-wide scale using reverse-engineered gene networks. *Nature Biotechn*, 23(3):377–383, March 2005.

[190] Mukesh Bansal, Giusy Della Gatta, and Diego di Bernardo. Inference of gene regulatory networks and compound mode of action from time course gene expression profiles. *Bioinf*, 22(7):815–822, 2006.

[191] Nir Friedman, Michal Linial, Iftach Nachman, and Dana Pe'er. Using Bayesian Networks to Analyze Expression Data. *J Comp Biol*, 7(3-4):601–620, August 2000.

[192] Juliane Schäfer, Korbinian Strimmer, José FF Mendes, SN Dorogovtsev, A Povolotsky, FV Abreu, and JG Oliveira. Learning large-scale graphical gaussian models from genomic data. In *AIP Conf Proc*, volume 776, pages 263–276, La Herradura, Spain, 2005. AIP.

[193] Alexander J Hartemink, David K Gifford, Tommi S Jaakkola, and Richard A Young. Combining location and expression data for principled discovery of genetic regulatory network models. In *Pac Symp Biocomp*, volume 6, pages 437–449, Stanford, CA, 2001. World Scientific Pub Co Inc.

[194] Chris J Needham, James R Bradford, Andrew J Bulpitt, and David R Westhead. Inference in bayesian networks. *Nature Biotechn*, 24(1):51–54, 2006.

[195] Adriano V Werhli and Dirk Husmeier. Reconstructing gene regulatory networks with bayesian networks by combining expression data with multiple sources of prior knowledge. *Stat Appl Genet Mol Biol*, 6(1):15, 2007.

[196] Florian Geier, Jens Timmer, and Christian Fleck. Reconstructing gene-regulatory networks from time series, knock-out data, and prior knowledge. *BMC Syst Biol*, 1(1):11, 2007.

[197] Sach Mukherjee and Terence P Speed. Network inference using informative priors. *Proc Natl Acad Sci (USA)*, 105(38):14313–14318, 2008.

[198] Alex Greenfield, Christoph Hafemeister, and Richard Bonneau. Robust data-driven incorporation of prior knowledge into the inference of dynamic regulatory networks. *Bioinf*, 29(8):1060–1067, 2013.

[199] Yupeng Li and Scott A Jackson. Gene network reconstruction by integration of prior biological knowledge. *G3: Genes, Genomes, Genetics*, 5(6):1075–1079, 2015.

[200] Mahsa Ghanbari, Julia Lasserre, and Martin Vingron. Reconstruction of gene networks using prior knowledge. *BMC Syst Biol*, 9(1):84, 2015.

[201] Matthew E. Studham, Andreas Tjärnberg, Torbjörn E.M. Nordling, Sven Nelander, and Erik L. L. Sonnhammer. Functional association networks as priors for gene regulatory network inference. *Bioinf*, 30(12):i130–i138, June 2014.

[202] Kwang-Il Goh, Michael E Cusick, David Valle, Barton Childs, Marc Vidal, and Albert-László Barabási. The human disease network. *Proc Natl Acad Sci (USA)*, 104(21):8685–8690, 2007.

[203] Joseph Loscalzo and Albert-László Barabási. Systems biology and the future of medicine. *Wiley Interdiscip Rev: Syst Biol Med*, 3(6):619–627, 2011.

[204] XueZhong Zhou, Jörg Menche, Albert-László Barabási, and Amitabh Sharma. Human symptoms–disease network. *Nature Comm*, 5, 2014.

[205] Steve Horvath. *Weighted network analysis: applications in genomics and systems biology.* Springer Science & Business Media, 2011.

[206] John Jeremy Rice, Yuhai Tu, and Gustavo Stolovitzky. Reconstructing biological networks using conditional correlation analysis. *Bioinf*, 21(6):765–773, March 2005.

[207] Bin Zhang and Steve Horvath. A General Framework for Weighted Gene Co-Expression Network Analysis. *Stat Appl Genet Mol Biol*, 4(1):1128, 2005.

[208] Pegah Khosravi, Vahid H. Gazestani, Leila Pirhaji, Brian Law, Mehdi Sadeghi, Bahram Goliaei, and Gary D. Bader. Inferring interaction type in gene regulatory networks using co-expression data. *Algorithms for Molecular Biology*, 10:23, 2015.

[209] CE Shannon. A mathematical theory of communication. *Bell Syst Techn J*, 27:379–423, 1948.

[210] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. John Wiley & Sons, November 2012.

[211] Andre Levchenko and Ilya Nemenman. Cellular noise and information transmission. *Curr Opin Biotechn*, 28:156–164, 2014.

[212] Patrik D'haeseleer, Xiling Wen, Stefanie Fuhrman, and Roland Somogyi. Mining the Gene Expression Matrix: Inferring Gene Relationships from Large Scale Gene Expression Data. In Mike Holcombe and Ray Paton, editors, *Information Processing in Cells and Tissues*, pages 203–212. 1998.

[213] A. J. Butte and I. S. Kohane. Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. In *Pac Symp Biocomp*, pages 418–429, Stanford, CA, 2000. World Scientific Pub Co Inc.

[214] Atul J Butte, Pablo Tamayo, Donna Slonim, Todd R Golub, and Isaac S Kohane. Discovering functional relationships between rna expression and chemotherapeutic susceptibility using relevance networks. *Proc Natl Acad Sci (USA)*, 97(22):12182–12186, 2000.

[215] Kai Wang, Ilya Nemenman, Nilanjana Banerjee, Adam A. Margolin, and Andrea Califano. Genome-Wide Discovery of Modulators of Transcriptional Interactions in Human B Lymphocytes. In Alberto Apostolico, Concettina Guerra, Sorin Istrail, Pavel A. Pevzner, and Michael Waterman, editors, *Res Comput Mol Biol (RECOMB)*, number 3909 in Lecture Notes in Computer Science, pages 348–362. Springer, Berlin-Heidelberg, April 2006.

[216] Kuo-Ching Liang and Xiaodong Wang. Gene Regulatory Network Reconstruction Using Conditional Mutual Information. *EURASIP J Bioinf Syst Biol*, 2008(1):253894, June 2008.

[217] Xiujun Zhang, Xing-Ming Zhao, Kun He, Le Lu, Yongwei Cao, Jingdong Liu, Jin-Kao Hao, Zhi-Ping Liu, and Luonan Chen. Inferring gene regulatory networks from gene expression data by path consistency algorithm based on conditional mutual information. *Bioinf*, 28(1):98–104, January 2012.

[218] Patrick E. Meyer, Kevin Kontos, Frederic Lafitte, and Gianluca Bontempi. Information-theoretic inference of large transcriptional regulatory networks. *EURASIP J Bioinf Syst Biol*, 2007(1):79879, 2007.

[219] Ifije E Ohiorhenuan, Ferenc Mechler, Keith P Purpura, Anita M Schmid, Qin Hu, and Jonathan D Victor. Sparse coding and high-order correlations in fine-scale cortical networks. *Nature*, 466(7306):617–621, 2010.

[220] Greg D Field, Jeffrey L Gauthier, Alexander Sher, Martin Greschner, Timothy A Machado, Lauren H Jepson, Jonathon Shlens, Deborah E Gunning,

Keith Mathieson, Wladyslaw Dabrowski, Liam Paninski, A. M. Litke, and E. J. Chichilnisky. Functional connectivity in the retina at the resolution of photoreceptors. *Nature*, 467(7316):673–677, 2010.

[221] Gašper Tkačik, Olivier Marre, Dario Amodei, Elad Schneidman, William Bialek, and Michael J. Berry. Searching for Collective Behavior in a Large Network of Sensory Neurons. *PLoS Comp Biol*, 10(1):e1003408, 2014.

[222] Matthias Bethge and Philipp Berens. Near-maximum entropy models for binary neural representations of natural images. In *NIPS 20*, pages 97–104, Vancouver, Canada, 2007. MIT Press.

[223] Timothy R. Lezon, Jayanth R. Banavar, Marek Cieplak, Amos Maritan, and Nina V. Fedoroff. Using the principle of entropy maximization to infer genetic interaction networks from gene expression patterns. *Proc Natl Acad Sci (USA)*, 103(50):19033–19038, December 2006.

[224] Jason W. Locasale and Alejandro Wolf-Yadlin. Maximum Entropy Reconstructions of Dynamic Signaling Networks from Quantitative Proteomics Data. *PLoS One*, 4(8):e6522, 2009.

[225] Faruck Morcos, Andrea Pagnani, Bryan Lunt, Arianna Bertolino, Debora S. Marks, Chris Sander, Riccardo Zecchina, José N. Onuchic, Terence Hwa, and Martin Weigt. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc Natl Acad Sci (USA)*, 108(49):E1293–E1301, 2011.

[226] David T Jones, Daniel WA Buchan, Domenico Cozzetto, and Massimiliano Pontil. Psicov: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinf*, 28(2):184–190, 2012.

[227] Kevin Wood, Satoshi Nishida, Eduardo Sontag, and Philippe Cluzel. Mechanism-independent method for predicting response to multidrug combinations in bacteria. *Proc Natl Acad Sci (USA)*, 109(30):12254–12259, 2012.

[228] Thierry Mora, Aleksandra M Walczak, William Bialek, and Curtis G Callan. Maximum entropy models for antibody diversity. *Proc Natl Acad Sci (USA)*, 107(12):5405–5410, 2010.

[229] William Bialek, Andrea Cavagna, Irene Giardina, Thierry Mora, Edmondo Silvestri, Massimiliano Viale, and Aleksandra M Walczak. Statistical mechanics for natural flocks of birds. *Proc Natl Acad Sci (USA)*, 109(13):4786–4791, 2012.

[230] David H Ackley, Geoffrey E Hinton, and Terrence J Sejnowski. A learning algorithm for boltzmann machines. *Cogn Sci*, 9(1):147–169, 1985.

[231] Yasser Roudi, Joanna Tyrcha, and John Hertz. Ising model for neural data: model quality and approximate methods for extracting functional connectivity. *Phys Rev E*, 79(5):051915, 2009.

[232] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *Unsupervised Learning*. Springer, 2009.

[233] Pradeep Ravikumar, Martin J Wainwright, and John D Lafferty. High-dimensional ising model selection using $\ell 1$-regularized logistic regression. *Ann Stat*, 38(3):1287–1319, 2010.

[234] Simona Cocco and Rémi Monasson. Adaptive cluster expansion for inferring boltzmann machines with noisy data. *Phys Rev Lett*, 106(9):090601, 2011.

[235] S. Cocco and R. Monasson. Adaptive Cluster Expansion for the Inverse Ising Problem: Convergence, Algorithm and Tests. *J Stat Phys*, 147(2):252–314, March 2012.

[236] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[237] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and Albert-László Barabási. The large-scale organization of metabolic networks. *Nature*, 407(6804):651–654, October 2000.

[238] Robert D Leclerc. Survival of the sparsest: robust gene networks are parsimonious. *Mol Syst Biol*, 4(1):213, 2008.

[239] Juha Karvanen. Estimating complex causal effects from incomplete observational data. *arXiv preprint arXiv:1403.1124*, 2014.

[240] Mor Nitzan, Jose Casadiego, and Marc Timme. Revealing physical interaction networks from statistics of collective dynamics. *Sci Adv*, 3(2):e1600396, 2017.

[241] Judea Pearl. *Causality*. Cambridge University Press, 2nd edition, September 2009.

[242] Andreas Wagner. How to reconstruct a large genetic network from n gene perturbations in fewer than n^2 easy steps. *Bioinf*, 17(12):1183–1197, 2001.

[243] Evan J. Molinelli, Anil Korkut, Weiqing Wang, Martin L. Miller, Nicholas P. Gauthier, Xiaohong Jing, Poorvi Kaushik, Qin He, Gordon Mills, David B. Solit, Christine A. Pratilas, Martin Weigt, Alfredo Braunstein, Andrea Pagnani, Riccardo Zecchina, and Chris Sander. Perturbation Biology: Inferring Signaling Networks in Cellular Systems. *PLoS Comp Biol*, 9(12), December 2013.

[244] Jesper Tegnér and Johan Björkegren. Perturbations to uncover gene networks. *Trends in genetics: TIG*, 23(1):34–41, January 2007.

[245] Marloes H Maathuis, Markus Kalisch, and Peter Bühlmann. Estimating

high-dimensional intervention effects from observational data. *Ann Stat*, 37(6A):3133–3164, 2009.

[246] Marloes H Maathuis, Diego Colombo, Markus Kalisch, and Peter Bühlmann. Predicting causal effects in large-scale systems from observational data. *Nature Meth*, 7(4):247–248, 2010.

[247] Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*. Springer Science & Business Media, December 2012.

[248] Jesper Tegnér, M. K. Stephen Yeung, Jeff Hasty, and James J. Collins. Reverse engineering gene networks: Integrating genetic perturbations with dynamical modeling. *Proc Natl Acad Sci (USA)*, 100(10):5944–5949, May 2003.

[249] Eduardo Sontag, Anatoly Kiyatkin, and Boris N. Kholodenko. Inferring dynamic architecture of cellular networks using time series of gene expression, protein and metabolite data. *Bioinf*, 20(12):1877–1886, 2004.

[250] Chris J Oates, Bryan T Hennessy, Yiling Lu, Gordon B Mills, and Sach Mukherjee. Network inference using steady-state data and goldbeter–koshland kinetics. *Bioinf*, 28(18):2342–2348, 2012.

[251] Djordje Djordjevic, Andrian Yang, Armella Zadoorian, Kevin Rungrugeecharoen, and Joshua W. K. Ho. How Difficult Is Inference of Mammalian Causal Gene Regulatory Networks? *PLoS One*, 9(11):e111661, November 2014.

[252] Adam Arkin, Peidong Shen, and John Ross. A Test Case of Correlation Metric Construction of a Reaction Pathway from Measurements. *Science*, 277(5330):1275–1279, August 1997.

[253] Paola Lecca, Daniele Morpurgo, Gianluca Fantaccini, Alessandro Casagrande,

and Corrado Priami. Inferring biochemical reaction pathways: the case of the gemcitabine pharmacokinetics. *BMC Syst Biol*, 6(1):51, 2012.

[254] Lennart Ljung. *System identification.* Wiley Online Library, 1999.

[255] Michael Schmidt and Hod Lipson. Distilling free-form natural laws from experimental data. *Science*, 324(5923):81–85, 2009.

[256] T. Chen, H. L. He, and G. M. Church. Modeling gene expression with differential equations. In *Pac Symp Biocomp*, pages 29–40, Stanford, CA, 1999. World Scientific Pub Co Inc.

[257] P. D'haeseleer, X. Wen, S. Fuhrman, and R. Somogyi. Linear modeling of mRNA expression levels during CNS development and injury. In *Pac Symp Biocomp*, pages 41–52, Stanford, CA, 1999. World Scientific Pub Co Inc.

[258] D. C. Weaver, C. T. Workman, and G. D. Stormo. Modeling regulatory networks with weight matrices. In *Pac Symp Biocomp*, pages 112–123, Stanford, CA, 1999. World Scientific Pub Co Inc.

[259] Eric Mjolsness, Tobias Mann, Rebecca Castaño, and Barbara J. Wold. From Coexpression to Coregulation: An Approach to Inferring Transcriptional Regulation among Gene Classes from Large-Scale Expression Data. In S. A. Solla, T. K. Leen, and K. Müller, editors, *Adv Neural Inf Proc Syst 12*, pages 928–934. MIT Press, 2000.

[260] Eugene P van Someren, Lodewyk FA Wessels, and Marcel JT Reinders. Linear modeling of genetic networks from experimental data. In *Ismb*, pages 355–366, San Diego, CA, 2000. Eighth International Conference on Intelligent Systems for Molecular Biology (ISMB-2000).

[261] Mattias Wahde and John Hertz. Coarse-grained reverse engineering of genetic regulatory networks. *Biosyst*, 55(1–3), 2000.

[262] Marco Grimaldi, Roberto Visintainer, and Giuseppe Jurman. Regnann: reverse engineering gene networks using artificial neural networks. *PLoS One*, 6(12):e28646, 2011.

[263] Eric Mjolsness, David H. Sharp, and John Reinitz. A connectionist model of development. *J Theor Biol*, 152(4):429–453, October 1991.

[264] Anton Crombach, Karl R Wotton, Damjan Cicin-Sain, Maksat Ashyraliyev, and Johannes Jaeger. Efficient reverse-engineering of a developmental gene regulatory network. *PLoS Comp Biol*, 8(7):e1002589, 2012.

[265] Chris J Oates, Frank Dondelinger, Nora Bayani, James Korkola, Joe W Gray, and Sach Mukherjee. Causal network inference using biochemical kinetics. *Bioinf*, 30(17):i468–i474, 2014.

[266] Niall M Mangan, Steven L Brunton, Joshua L Proctor, and J Nathan Kutz. Inferring biological networks by sparse identification of nonlinear dynamics. *IEEE Trans Mol Biol Multi-Scale Commun*, 2(1):52–63, 2016.

[267] Michael Wibral, Joseph T. Lizier, and Viola Priesemann. Bits from Brains for Biologically Inspired Computing. *Frontiers Robotics AI*, 2, 2015.

[268] Ming Luo, Holger Kantz, Ngar-Cheung Lau, Wenwen Huang, and Yu Zhou. Questionable dynamical evidence for causality between galactic cosmic rays and interannual variation in global temperature. *Proc Natl Acad Sci (USA)*, page 201510571, 2015.

[269] Nikolai F Rulkov, Mikhail M Sushchik, Lev S Tsimring, and Henry DI Abar-

banel. Generalized synchronization of chaos in directionally coupled chaotic systems. *Phys Rev E*, 51(2):15, 1995.

[270] Floris Takens. Detecting strange attractors in turbulence. In D. A. Rand and L. S. Young, editors, *Symposium on Dynamical Systems and Turbulence*, pages 366–381. Springer, 1981.

[271] Steven H. Strogatz. *Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry, and Engineering.* Westview Press, Boulder, CO, 2nd edition, 2015.

[272] Mario Ragwitz and Holger Kantz. Markov models from data by simple nonlinear time series predictors in delay embedding spaces. *Phys Rev E*, 65(5):12, 2002.

[273] Michael Small and C. K. Tse. Optimal embedding parameters: A modelling paradigm. *Physica D: Nonlinear Phenomena*, 194(3–4):283–296, July 2004.

[274] Luca Faes, Giandomenico Nollo, and Alberto Porta. Non-uniform multivariate embedding to assess the information transfer in cardiovascular and cardiorespiratory variability series. *Computers Biol Med*, 42(3):290–297, March 2012.

[275] James M McCracken and Robert S Weigel. Convergent cross-mapping and pairwise asymmetric inference. *Phys Rev E*, 90(6):062903, 2014.

[276] Sarah Cobey and Edward B. Baskerville. Limits to Causal Inference with State-Space Reconstruction for Infectious Disease. *PLoS One*, 11(12):e0169050, December 2016.

[277] Dan Mønster, Riccardo Fusaroli, Kristian Tylén, Andreas Roepstorff, and Jacob F Sherson. Inferring causality from noisy time series data. In V Muñoz, Oleg Gusikhin, and Victor Chang, editors, *Proceedings of the 1st International*

Conference on Complex Information Systems (COMPLEXIS 2016), pages 48–56, Funchal, Madeira, Portgual, 2016. SCITEPRESS.

[278] Jun-Jie Jiang, Zi-Gang Huang, Liang Huang, Huan Liu, and Ying-Cheng Lai. Directed dynamical influence is more detectable with noise. *Scientific Reports*, 6, 2016.

[279] CH Wiggins and Ilya Nemenman. Process pathway inference via time series analysis. *Exp Mech*, 43(3):361–370, 2003.

[280] Rainer Opgen-Rhein and Korbinian Strimmer. Learning causal networks from systems biology time course data: an effective model selection procedure for the vector autoregressive process. *BMC Bioinf*, 8(2):S3, 2007.

[281] C. W. J. Granger. Investigating Causal Relations by Econometric Models and Cross-spectral Methods. *Econometrica*, 37(3):424–438, 1969.

[282] Norbert Wiener. The theory of prediction. *Modern Math for Engineers*, 1:125–139, 1956.

[283] C. W. J. Granger. Some recent development in a concept of causality. *J Econometrics*, 39(1–2):199–211, September 1988.

[284] Winrich A. Freiwald, Pedro Valdes, Jorge Bosch, Rolando Biscay, Juan Carlos Jimenez, Luis Manuel Rodriguez, Valia Rodriguez, Andreas K. Kreiter, and Wolf Singer. Testing non-linearity and directedness of interactions between neural groups in the macaque inferotemporal cortex. *J Neurosci Meth*, 94(1):105–119, December 1999.

[285] Nicola Ancona, Daniele Marinazzo, and Sebastiano Stramaglia. Radial basis function approach to nonlinear Granger causality of time series. *Phys Rev E*, 70(5):056221, November 2004.

[286] Yonghong Chen, Govindan Rangarajan, Jianfeng Feng, and Mingzhou Ding. Analyzing multiple nonlinear time series with extended Granger causality. *Phys Lett A*, 324(1):26–35, April 2004.

[287] Boris Gourévitch, Régine Le Bouquin-Jeannès, and Gérard Faucon. Linear and nonlinear causality between signals: Methods, examples and neurophysiological applications. *Biological Cybern*, 95(4):349–369, October 2006.

[288] Daniele Marinazzo, Mario Pellicoro, and Sebastiano Stramaglia. Kernel Method for Nonlinear Granger Causality. *Phys Rev Lett*, 100(14):144103, April 2008.

[289] Mingzhou Ding, Yonghong Chen, and Steven L. Bressler. Granger causality: Basic theory and application to neuroscience. In *Handbook of Time Series Analysis: Recent Theoretical Developments and Applications*. Wiley, Wienheim, 2006.

[290] Katarzyna J. Blinowska, Rafał Kuś, and Maciej Kamiński. Granger causality and information flow in multivariate processes. *Phys Rev E*, 70(5):050902, November 2004.

[291] Judea Pearl. Bayesian networks: A model of self-activated memory for evidential reasoning. In *Proceedings, Cognitive Science Society, UC Irvine*, pages 329–334, Los Angeles, CA, 1985. UCLA Computer Science Department Technical Report 850021 (R-43).

[292] Eugene Charniak. Bayesian Networks without Tears. *AI Magazine*, 12(4):50, December 1991.

[293] Z Ghahramani. An introduction to hidden markov models and bayesian networks. *Int J Pattern Recogn Artificial Intelligence*, 15(01):9–42, February 2001.

[294] Nir Friedman. Inferring Cellular Networks Using Probabilistic Graphical Models. *Science*, 303(5659):799–805, February 2004.

[295] Nir Friedman, Kevin Murphy, and Stuart Russell. Learning the structure of dynamic probabilistic networks. In Gregory F. Cooper Cooper and Serafín Moral, editors, *Proc Fourteenth Conf Uncertainty in Artificial Intelligence (UAI-98)*, pages 139–147, San Francisco, CA, 1998. Morgan Kaufmann Publishers Inc.

[296] V Anne Smith, Erich D Jarvis, and Alexander J Hartemink. Evaluating functional network inference using simulations of complex biological systems. *Bioinf*, 18(suppl 1):S216–S224, 2002.

[297] Iftach Nachman, Aviv Regev, and Nir Friedman. Inferring quantitative models of regulatory networks from expression data. *Bioinf*, 20(suppl 1):i248–i256, 2004.

[298] Enzo Acerbi, Teresa Zelante, Vipin Narang, and Fabio Stella. Gene network inference using continuous time bayesian networks: a comparative study and application to th17 cell differentiation. *BMC Bioinf*, 15(1):387, 2014.

[299] Cunlu Zou and Jianfeng Feng. Granger causality vs. dynamic bayesian network inference: a comparative study. *BMC Bioinf*, 10(1):122, 2009.

[300] Jun Zhu, Bin Zhang, Erin N. Smith, Becky Drees, Rachel B. Brem, Leonid Kruglyak, Roger E. Bumgarner, and Eric E. Schadt. Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks. *Nature Gen*, 40(7):854–861, July 2008.

[301] Iosifina Pournara and Lorenz Wernisch. Reconstruction of gene networks using Bayesian learning and manipulation experiments. *Bioinf*, 20(17):2934–2942, November 2004.

[302] Nir Friedman and Daphne Koller. Being bayesian about network structure. In Craig Boutilier and Moises Goldszmidt, editors, *Proc Sixteenth Conf Uncertainty in Artificial Intelligence (UAI-00)*, pages 201–210, San Francisco, CA, 2000. Morgan Kaufmann Publishers Inc.

[303] Nihat Ay and Daniel Polani. Information flows in causal networks. *Adv Compl Syst*, 11(01):17–41, 2008.

[304] Thomas Schreiber. Measuring Information Transfer. *Phys Rev Lett*, 85(2):461–464, 2000.

[305] Milan Paluš, Vladimír Komárek, Zbyněk Hrnčíř, and Katalin Štěrbová. Synchronization as adjustment of information rates: detection from bivariate time series. *Phys Rev E*, 63(4):6, 2001.

[306] Christopher J. Honey, Rolf Kötter, Michael Breakspear, and Olaf Sporns. Network structure of cerebral cortex shapes functional connectivity on multiple time scales. *Proc Natl Acad Sci (USA)*, 104(24):10240–10245, 2007.

[307] Olav Stetter, Demian Battaglia, Jordi Soriano, and Theo Geisel. Model-free reconstruction of excitatory neuronal connectivity from calcium imaging signals. *PLoS Comp Biol*, 8(8):1–25, 2012.

[308] Michael Wibral, Raul Vicente, and Michael Lindner. Transfer Entropy in Neuroscience. In Michael Wibral, Raul Vicente, and Joseph T. Lizier, editors, *Directed Information Measures in Neuroscience*, Understanding Complex Systems, pages 3–36. Springer Berlin Heidelberg, 2014.

[309] Luca Faes, Giandomenico Nollo, and Alberto Porta. Compensated Transfer Entropy as a Tool for Reliably Estimating Information Transfer in Physiological Time Series. *Entropy*, 15(1):198–219, January 2013.

[310] Bernd Pompe and Jakob Runge. Momentary information transfer as a coupling measure of time series. *Phys Rev E*, 83(5):051122, May 2011.

[311] Jakob Runge, Jobst Heitzig, Vladimir Petoukhov, and Jürgen Kurths. Escaping the Curse of Dimensionality in Estimating Multivariate Transfer Entropy. *Phys Rev Lett*, 108(25):5, 2012.

[312] Jakob Runge, Jobst Heitzig, Norbert Marwan, and Jürgen Kurths. Quantifying causal coupling strength: A lag-specific measure for multivariate time series related to transfer entropy. *Phys Rev E*, 86(6):061121, December 2012.

[313] O. Kwon and J.-S. Yang. Information flow between stock indices. *EPL (Europhys Lett)*, 82(6):68003, 2008.

[314] Jinkyu Kim, Gunn Kim, Sungbae An, Young-Kyun Kwon, and Sungroh Yoon. Entropy-Based Analysis and Bioinformatics-Inspired Integration of Global Economic Information Transfer. *PLoS One*, 8(1):e51986, January 2013.

[315] Nils Bertschinger, Johannes Rauh, Eckehard Olbrich, Jürgen Jost, and Nihat Ay. Quantifying Unique Information. *Entropy*, 16(4):2161–2183, April 2014.

[316] Ilya Nemenman, Fariel Shafee, and William Bialek. Entropy and inference, revisited. In Thomas G. Dietterich, Suzanna Becker, and Zoubin Ghahramani, editors, *Advances in Neural Information Processing Systems 14*. MIT Press, 2002.

[317] Ilya Nemenman. Coincidences and estimation of entropies of random variables with large cardinalities. *Entropy*, 13:2013–2023, 2011.

[318] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The Elements of Statistical Learning*. Springer series in statistics, Springer Berlin, 2 edition, 2001.

[319] Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Phys Rev E*, 69(6):066138, June 2004.

[320] Daniel W. Hahs and Shawn D. Pethel. Distinguishing Anticipation from Causality: Anticipatory Bias in the Estimation of Information Flow. *Phys Rev Lett*, 107(12):128701, 2011.

[321] Michael Lindner, Raul Vicente, Viola Priesemann, and Michael Wibral. TRENTOOL: A Matlab open source toolbox to analyse information flow in time series data with transfer entropy. *BMC Neuroscience*, 12:119, 2011.

[322] Michael Wibral, Nicolae Pampu, Viola Priesemann, Felix Siebenhühner, Hannes Seiwert, Michael Lindner, Joseph T. Lizier, and Raul Vicente. Measuring Information-Transfer Delays. *PLoS One*, 8(2):19, 2013.

[323] Lionel Barnett, Adam B. Barrett, and Anil K. Seth. Granger Causality and Transfer Entropy Are Equivalent for Gaussian Variables. *Phys Rev Lett*, 103(23):4, 2009.

[324] James Massey. Causality, feedback and directed information. In *Proc. Int. Symp. Inf. Theory Applic.(ISITA-90)*, pages 303–305, 1990.

[325] J. T. Lizier and M. Prokopenko. Differentiating information transfer and causal effect. *Eur Physical J B*, 73(4):605–615, 2010.

[326] Patrik D'haeseleer, Shoudan Liang, and Roland Somogyi. Genetic network inference: from co-expression clustering to reverse engineering. *Bioinf*, 16(8):707–726, August 2000.

[327] Lodewyk FA Wessels and Marcel JT Reinders. A comparison of genetic network models. In *Pac Symp Biocomp*, volume 6, pages 508–519, Stanford, CA, 2001. World Scientific Pub Co Inc.

[328] Paul Brazhnik, Alberto de la Fuente, and Pedro Mendes. Gene networks: how to put the function in genomics. *Trends Biotechn*, 20(11):467–472, November 2002.

[329] Hidde De Jong. Modeling and simulation of genetic regulatory systems: a literature review. *J Comp Biol*, 9(1):67–103, 2002.

[330] Eugene P van Someren, Lodewyk FA Wessels, Eric Backer, and Marcel JT Reinders. Genetic network modeling. *Pharmacogenomics*, 3(4):507–525, 2002.

[331] Réka Albert. Boolean modeling of genetic regulatory networks. In *Complex networks*, pages 459–481. Springer Berlin Heidelberg, 2004.

[332] Natal AW van Riel. Dynamic modelling and analysis of biochemical networks: mechanism-based models and model-based experiments. *Brief Bioinf*, 7(4):364–374, 2006.

[333] Réka Albert. Network inference, analysis, and modeling in systems biology. *The Plant Cell*, 19(11):3327–3338, 2007.

[334] K-H Cho, S-M Choo, SH Jung, J-R Kim, H-S Choi, and J Kim. Reverse engineering of gene regulatory networks. *IET Syst Biol*, 1(3):149–163, 2007.

[335] John Goutsias and NH Lee. Computational and experimental approaches for modeling gene regulatory networks. *Curr Pharmaceutical Design*, 13(14):1415–1436, 2007.

[336] Florian Markowetz and Rainer Spang. Inferring cellular networks – a review. *BMC Bioinf*, 8(6), 2007.

[337] Lars Kaderali and Nicole Radde. Inferring gene regulatory networks from expression data. In *Computational Intelligence in Bioinformatics*, pages 33–74. Springer, 2008.

[338] Guy Karlebach and Ron Shamir. Modelling and analysis of gene regulatory networks. *Nature Rev Mol Cell Biol*, 9(10):770–780, 2008.

[339] Richard Bonneau. Learning biological networks: from modules to dynamics. *Nature Chemical Biology*, 4(11):658–664, November 2008.

[340] Wei-Po Lee and Wen-Shyong Tzou. Computational methods for discovering gene networks from expression data. *Brief Bioinf*, 10(4):408–423, 2009.

[341] Ilya Shmulevich and John D Aitchison. Deterministic and stochastic models of genetic regulatory networks. *Meth Enzymol*, 467:335–356, 2009.

[342] Chao Sima, Jianping Hua, and Sungwon Jung. Inference of gene regulatory networks using time-series data: a survey. *Curr Genomics*, 10(6):416–429, 2009.

[343] Ahmet Ay and David N Arnosti. Mathematical modeling of gene expression: a guide for the perplexed biologist. *Critical Rev Biochem Mol Biol*, 46(2):137–151, 2011.

[344] Christopher A Penfold and David L Wild. How to infer gene networks from expression profiles, revisited. *Interface Focus*, 1(6):857–870, 2011.

[345] Diogo FT Veiga, Bhaskar Dutta, and Gábor Balázsi. Network inference and network response identification: moving genome-scale data to the next level of biological discovery. *Mol bioSyst*, 6(3):469–480, 2010.

[346] Enrique Hernández-Lemus and Jesús M Siqueiros-García. Information theoretical methods for complex network structure reconstruction. *Complex Adaptive Systems Modeling*, 1(1):8, 2013.

[347] Blagoj Ristevski. A survey of models for inference of gene regulatory networks. *Nonlinear Anal Model Control*, 18(4):444–465, 2013.

[348] Nedumparambathmarath Vijesh, Swarup Kumar Chakrabarti, and Janardanan Sreekumar. Modeling of gene regulatory networks: A review. *J Biomed Sci Eng*, 6(02):223, 2013.

[349] Alejandro F. Villaverde, John Ross, and Julio R Banga. Reverse engineering cellular networks with information theoretic methods. *Cells*, 2(2):306–329, 2013.

[350] Frank Emmert-Streib, Matthias Dehmer, and Benjamin Haibe-Kains. Gene regulatory networks and their applications: understanding biological and medical problems in terms of networks. *Frontiers Cell Devel Biol*, 2:38, 2014.

[351] Y. X. Rachel Wang and Haiyan Huang. Review on statistical methods for gene network reconstruction using expression data. *Journal of Theoretical Biology*, 362:53–61, December 2014.

[352] Xiaoxi Dong, Anatoly Yambartsev, Stephen A Ramsey, Lina D Thomas, Natalia Shulzhenko, and Andrey Morgun. Reverse engeneering of regulatory networks from big data: a roadmap for biologists. *Bioinf and Biol Insights*, 9:61, 2015.

[353] Christophe Liseron-Monfils and Doreen Ware. Revealing gene regulation and associations through biological networks. *Curr Plant Biol*, 3:30–39, 2015.

[354] Daniel Marbach, Robert J. Prill, Thomas Schaffter, Claudio Mattiussi, Dario Floreano, and Gustavo Stolovitzky. Revealing strengths and weaknesses of methods for gene network inference. *Proc Natl Acad Sci (USA)*, 107(14):6286–6291, 2010.

[355] Chris J. Oates and Sach Mukherjee. Network inference and biological dynamics. *Ann Stat*, 6(3):1209–1235, September 2012.

[356] Amina Noor, Erchin Serpedin, Mohamed Nounou, Hazem Nounou, Nady Mohamed, and Lotfi Chouchane. An Overview of the Statistical Methods Used for

Inferring Gene Regulatory Networks and Protein-Protein Interaction Networks. *Advances in Bioinformatics*, 2013:e953814, February 2013.

[357] Jimmy Omony. Biological Network Inference: A Review of Methods and Assessment of Tools and Techniques. *Ann Res Rev Biol*, 4(4):577–601, November 2013.

[358] Spencer Angus Thomas and Yaochu Jin. Reconstructing biological gene regulatory networks: where optimization meets big data. *Evolutionary Intelligence*, 7(1):29–47, 2014.

[359] Richard Bonneau and Tarmo Aijo. Biophysically motivated regulatory network inference: progress and prospects. *bioRxiv*, page 051847, May 2016.

[360] David M Budden and Edmund J Crampin. Information theoretic approaches for inference of biological networks from continuous-valued data. *BMC Syst Biol*, 10(1):89, 2016.

[361] Yasser Abduallah, Turki Turki, Kevin Byron, Zongxuan Du, Miguel Cervantes-Cervantes, and Jason TL Wang. Mapreduce algorithms for inferring gene regulatory networks from time-series microarray data using an information-theoretic approach. *BioMed research international*, 2017, 2017.

[362] Caroline Siegenthaler and Rudiyanto Gunawan. Assessment of network inference methods: how to cope with an underdetermined problem. *PLoS One*, 9(3):e90481, 2014.

[363] David J. Hand. Measuring classifier performance: a coherent alternative to the area under the ROC curve. *Machine Learning*, 77(1):103–123, June 2009.

[364] Stefan R. Maetschke, Piyush B. Madhamshettiwar, Melissa J. Davis, and

Mark A. Ragan. Supervised, semi-supervised and unsupervised inference of gene regulatory networks. *Brief Bioinf*, 15(2):195–211, May 2013.

[365] P Xenitidis, I Seimenis, S Kakolyris, and A Adamopoulos. Evaluation of artificial time series microarray data for dynamic gene regulatory network inference. *J Theor Biol*, 2017.

[366] Laurie Goodman. Hypothesis-limited research. *Genome Res*, 9(8):673, 1999.

[367] Fan Zhu and Yuanfang Guan. Predicting dynamic signaling network response under unseen perturbations. *Bioinf*, 30(19):2772–2778, 2014.

[368] D. P. Noren, B. L. Long, Raquel Norel, Kahn Rrhissorrakrai, Kenneth Hess, Chenyue Wendy Hu, Alex J. Bisberg, Andre Schultz, Erik Engquist, Li Liu, Xihui Lin, Gregory M. Chen, Honglei Xie, Geoffrey A. M. Hunter, Paul C. Boutros, Oleg Stepanov, DREAM 9 AML-OPC Consortium, Thea Norman, Stephen H. Friend, Gustavo Stolovitzky, Steven Kornblau, and Amina A. Qutub. A crowdsourcing approach to developing and assessing prediction algorithms for aml prognosis. *PLoS Comp Biol*, 12(6):e1004890, 2016.

[369] A. H. Lang, H. Li, J. J. Collins, and P. Mehta. Epigenetic landscapes explain partially reprogrammed cells and identify key reprogramming genes. *PLoS Comp Biol*, 10:e1003734, 2014.

[370] Kenneth G Wilson. Problems in physics with many scales of length. *Scient Amer*, 241(2):158–179, 1979.

[371] Benjamin B Machta, Ricky Chachra, Mark K Transtrum, and James P Sethna. Parameter space compression underlies emergent theories and predictive models. *Science*, 342(6158):604–607, 2013.

[372] Mark K Transtrum, Benjamin B Machta, Kevin S Brown, Bryan C Daniels, Christopher R Myers, and James P Sethna. Perspective: Sloppiness and emergent theories in physics, biology, and beyond. *J Chem Phys*, 143(1):07B201_1, 2015.

[373] Serena Bradde and William Bialek. Pca meets rg. *arXiv preprint arXiv:1610.09733*, 2016.

[374] Anagha Joshi, Riet De Smet, Kathleen Marchal, Yves Van de Peer, and Tom Michoel. Module networks revisited: computational assessment and prioritization of model predictions. *Bioinf*, 25(4):490–496, February 2009.

[375] Tom Michoel, Riet De Smet, Anagha Joshi, Yves Van de Peer, and Kathleen Marchal. Comparative analysis of module-based versus direct methods for reverse-engineering transcriptional regulatory networks. *BMC Syst Biol*, 3(1):49, 2009.

[376] Eric Bonnet, Marianthi Tatari, Anagha Joshi, Tom Michoel, Kathleen Marchal, Geert Berx, and Yves Van de Peer. Module Network Inference from a Cancer Gene Expression Data Set Identifies MicroRNA Regulated Modules. *PLoS One*, 5(4):e10162, April 2010.

[377] Xuejing Li, Casandra Panea, Chris H Wiggins, Valerie Reinke, and Christina Leslie. Learning “graph-mer” motifs that predict gene expression trajectories in development. *PLoS Comp Biol*, 6(4):e1000761, 2010.

[378] John Betteley Birks. *Rutherford at Manchester*. Heywood, London, 1962.

[379] Carlos D Brody, Ranulfo Romo, and Adam Kepecs. Basic mechanisms for graded persistent activity: discrete attractors, continuous attractors, and dynamic representations. *Curr Opin Neurobiol*, 13(2):204–211, 2003.

[380] Si Wu and Shun-Ichi Amari. Computing with continuous attractors: stability and online aspects. *Neural Comput*, 17(10):2215–2239, 2005.

[381] Patricia S Goldman-Rakic. Cellular basis of working memory. *Neuron*, 14(3):477–485, 1995.

[382] Joaquin M Fuster and Garrett E Alexander. Neuron activity related to short-term memory. *Science*, 173(3997):652–654, 1971.

[383] Shintaro Funahashi, Charles J Bruce, and Patricia S Goldman-Rakic. Mnemonic coding of visual space in the monkey's dorsolateral prefrontal cortex. *J. Neurophysiol*, 61(2):331–349, 1989.

[384] Joaquin M Fuster and John P Jervey. Inferotemporal neurons distinguish and retain behaviorally relevant features of visual stimuli. *Science*, 212(4497):952–955, 1981.

[385] James W Gnadt and Richard A Andersen. Memory related motor planning activity in posterior parietal cortex of macaque. *Experimental Brain Research*, 70(1):216–220, 1988.

[386] David MacNeil and Chris Eliasmith. Fine-tuning and the stability of recurrent neural networks. *PLoS One*, 6(9):e22885, 2011.

[387] Mikael Lundqvist, Pawel Herman, and Earl K Miller. Working memory: Delay activity, yes! persistent activity? maybe not. *J Neurosci*, 38(32):7013–7019, 2018.

[388] Joel Zylberberg and Ben W Strowbridge. Mechanisms of persistent activity in cortical circuits: possible neural substrates for working memory. *Annu. Rev. Neurosci.*, 40:603–627, 2017.

[389] H Sebastian Seung. How the brain keeps the eyes still. *Proc Natl Acad Sci (USA)*, 93(23):13339–13344, 1996.

[390] Daniel J Amit and Nicolas Brunel. Model of global spontaneous activity and local structured activity during delay periods in the cerebral cortex. *Cerebral Cortex*, 7(3):237–252, 1997.

[391] Albert Compte, Nicolas Brunel, Patricia S Goldman-Rakic, and Xiao-Jing Wang. Synaptic mechanisms and network dynamics underlying spatial working memory in a cortical network model. *Cerebral Cortex*, 10(9):910–923, 2000.

[392] John J Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proc Natl Acad Sci (USA)*, 79(8):2554–2558, 1982.

[393] Kechen Zhang. Representation of spatial orientation by the intrinsic dynamics of the head-direction cell ensemble: a theory. *J Neurosci*, 16(6):2112–2126, 1996.

[394] Sung Soo Kim, Hervé Rouault, Shaul Druckmann, and Vivek Jayaraman. Ring attractor dynamics in the drosophila central brain. *Science*, 356(6340):849–853, 2017.

[395] Kyobi S Kakaria and Benjamin L de Bivort. Ring attractor dynamics emerge from a spiking model of the entire protocerebral bridge. *Front Behav Neurosci*, 11:8, 2017.

[396] Jude Baby George, Grace Mathew Abraham, Zubin Rashid, Bharadwaj Amrutur, and Sujit Kumar Sikdar. Random neuronal ensembles can inherently do context dependent coarse conjunctive encoding of input stimulus without any specific training. *Scientific Reports*, 8(1):1403, 2018.

[397] Audrey J Sederberg and Ilya Nemenman. Randomly connected networks generate emergent selectivity and predict decoding properties of large populations of neurons. *arXiv preprint arXiv:1909.10116*, 2019.

[398] Birgit Kriener, Håkon Enger, Tom Tetzlaff, Hans E Plesser, Marc-Oliver Gewaltig, and Gaute T Einevoll. Dynamics of self-sustained asynchronous-irregular activity in random networks of spiking neurons with strong synapses. *Front Comput Neurosc*, 8:136, 2014.

[399] Carl Van Vreeswijk and Haim Sompolinsky. Chaos in neuronal networks with balanced excitatory and inhibitory activity. *Science*, 274(5293):1724–1726, 1996.

[400] Carl van Vreeswijk and Haim Sompolinsky. Chaotic balanced state in a model of cortical circuits. *Neural Comput*, 10(6):1321–1371, 1998.

[401] Shaul Druckmann and Dmitri B Chklovskii. Neuronal circuits underlying persistent representations despite time varying activity. *Current Biology*, 22(22):2095–2103, 2012.

[402] Francesco P Battaglia and Alessandro Treves. Attractor neural networks storing multiple space representations: a model for hippocampal place fields. *Physical Review E*, 58(6):7738, 1998.

[403] Sen Song, Per Jesper Sjöström, Markus Reigl, Sacha Nelson, and Dmitri B Chklovskii. Highly nonrandom features of synaptic connectivity in local cortical circuits. *PLoS Biol*, 3(3):e68, 2005.

[404] Yasser Roudi and Alessandro Treves. Representing where along with what information in a model of a cortical patch. *PLoS Comp Biol*, 4(3):e1000012, 2008.

[405] Rémi Monasson and Sophie Rosay. Crosstalk and transitions between multiple spatial maps in an attractor neural network model of the hippocampus: Phase diagram. *Phys Rev E*, 87(6):062813, 2013.

[406] Francesca Barbieri and Nicolas Brunel. Can attractor network models account for the statistics of firing during persistent activity in prefrontal cortex? *Frontiers in Neuroscience*, 2:3, 2008.

[407] Richard HR Hahnloser and H Sebastian Seung. Permitted and forbidden sets in symmetric threshold-linear networks. In *Advances in Neural Information Processing Systems*, pages 217–223, 2001.

[408] Shun-ichi Amari. Dynamics of pattern formation in lateral-inhibition type neural fields. *Biological Cybernetics*, 27(2):77–87, 1977.

[409] Xiao-Jing Wang. Synaptic reverberation underlying mnemonic persistent activity. *Trends Neurosci*, 24(8):455–463, 2001.

[410] Claude Elwood Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423, 1948.

[411] András Antos and Ioannis Kontoyiannis. Convergence properties of functional estimates for discrete distributions. *Random Structures & Algorithms*, 19(3-4):163–193, 2001.

[412] Steven P Strong, Roland Koberle, Rob R de Ruyter van Steveninck, and William Bialek. Entropy and information in neural spike trains. *Phys Rev Lett*, 80(1):197, 1998.

[413] Liam Paninski. Estimation of entropy and mutual information. *Neural Comput*, 15(6):1191–1253, 2003.

[414] JS Griffith. On the stability of brain-like structures. *Biophys. J.*, 3(4):299–308, 1963.

[415] Wolfgang Maass, Thomas Natschläger, and Henry Markram. Real-time computing without stable states: A new framework for neural computation based on perturbations. *Neural Comput*, 14(11):2531–2560, 2002.

[416] Dean V Buonomano and Wolfgang Maass. State-dependent computations: spatiotemporal processing in cortical networks. *Nature Reviews Neuroscience*, 10(2):113, 2009.

[417] Ulises Pereira and Nicolas Brunel. Attractor dynamics in networks with learning rules inferred from in vivo data. *Neuron*, 99(1):227–238, 2018.

[418] R Monasson and S Rosay. Transitions between spatial attractors in place-cell networks. *arXiv preprint arXiv:1507.05725*, 2015.

[419] Wulfram Gerstner, Werner M Kistler, Richard Naud, and Liam Paninski. *Neuronal dynamics: From single neurons to networks and models of cognition*. Cambridge University Press, 2014.

[420] Ariel Amir, Naomichi Hatano, and David R Nelson. Non-hermitian localization in biological networks. *Phys Rev E*, 93(4):042310, 2016.

[421] Hidenori Tanaka and David R Nelson. Non-hermitian quasilocalization and ring attractor neural networks. *Phys Rev E*, 99(6):062406, 2019.

[422] Eleonora Catsigeras. Dale's principle is necessary for an optimal neuronal network's dynamics. *arXiv preprint arXiv:1307.0597*, 2013.

[423] Mengchen Zhu and Christopher J Rozell. Modeling inhibitory interneurons in efficient sensory coding models. *PLoS Comp Biol*, 11(7):e1004353, 2015.

[424] Alfonso Renart, Pengcheng Song, and Xiao-Jing Wang. Robust spatial working memory through homeostatic synaptic scaling in heterogeneous cortical networks. *Neuron*, 38(3):473–485, 2003.

[425] Joshua A Goldberg, Uri Rokni, and Haim Sompolinsky. Patterns of ongoing activity and the functional architecture of the primary visual cortex. *Neuron*, 42(3):489–500, 2004.

[426] Robert Rosenbaum and Brent Doiron. Balanced networks of spiking neurons with spatially dependent recurrent connections. *Physical Review X*, 4(2):021039, 2014.

[427] Chengcheng Huang, Douglas A Ruff, Ryan Pyle, Robert Rosenbaum, Marlene R Cohen, and Brent Doiron. Circuit models of low-dimensional shared variability in cortical networks. *Neuron*, 101(2):337–348, 2019.

[428] Daniel Turner-Evans, Stephanie Wegener, Herve Rouault, Romain Franconville, Tanya Wolff, Johannes D Seelig, Shaul Druckmann, and Vivek Jayaraman. Angular velocity integration in a fly heading circuit. *Elife*, 6:e23496, 2017.

[429] Peter H Li, Larry F Lindsey, Michał Januszewski, Zhihao Zheng, Alexander Shakeel Bates, István Taisz, Mike Tyka, Matthew Nichols, Feng Li, Eric Perlman, et al. Automated reconstruction of a serial-section em drosophila brain with flood-filling networks and local realignment. *bioRxiv*, page 605634, 2019.

[430] Olaf Sporns. The non-random brain: efficiency, economy, and complex dynamics. *FCN*, 5:5, 2011.

[431] Si Wu, KY Michael Wong, CC Alan Fung, Yuanyuan Mi, and Wenhao Zhang. Continuous attractor neural networks: candidate of a canonical model for neural information representation. *F1000Research*, 5, 2016.

[432] Hidehiko K Inagaki, Lorenzo Fontolan, Sandro Romani, and Karel Svoboda. Discrete attractor dynamics underlies persistent activity in the frontal cortex. *Nature*, 566(7743):212, 2019.

[433] Zachary P Kilpatrick, Bard Ermentrout, and Brent Doiron. Optimizing working memory with heterogeneity of recurrent cortical excitation. *J Neurosci*, 33(48):18999–19011, 2013.

[434] Klaus Wimmer, Duane Q Nykamp, Christos Constantinidis, and Albert Compte. Bump attractor dynamics in prefrontal cortex explains behavioral precision in spatial working memory. *Nature Neuroscience*, 17(3):431, 2014.

[435] Giuseppe Longo and Maël Montévil. From physics to biology by extending criticality and symmetry breakings. In *Perspectives on Organisms*, pages 161–185. Springer, 2014.

[436] Kenneth G Wilson. Renormalization group and critical phenomena. i. renormalization group and the kadanoff scaling picture. *Physical review B*, 4(9):3174, 1971.

[437] Kenneth G Wilson. Renormalization group and critical phenomena. ii. phase-space cell analysis of critical behavior. *Physical Review B*, 4(9):3184, 1971.

[438] Efi Efrati, Zhe Wang, Amy Kolan, and Leo P Kadanoff. Real-space renormalization in statistical mechanics. *Reviews of Modern Physics*, 86(2):647, 2014.

[439] Leenoy Meshulam, Jeffrey L Gauthier, Carlos D Brody, David W Tank, and William Bialek. Coarse graining, fixed points, and scaling in a large population of neurons. *Physical Review Letters*, 123(17):178103, 2019.

[440] Leo P Kadanoff. Variational principles and approximate renormalization group calculations. *Physical Review Letters*, 34(16):1005, 1975.

[441] Ernst Ising. Beitrag zur theorie des ferromagnetismus. *Zeitschrift für Physik A Hadrons and Nuclei*, 31(1):253–258, 1925.

[442] Per Bak. *How nature works: the science of self-organized criticality.* Springer Science & Business Media, 2013.

[443] Nigel D Goldenfeld. Lectures on phase transitions and the renormalization group. 1992.

[444] John Cardy. *Scaling and renormalization in statistical physics*, volume 5. Cambridge university press, 1996.

[445] Leo P Kadanoff. Scaling laws for ising models near t c. *Physics Physique Fizika*, 2(6):263, 1966.

[446] Paul Smolensky. Information processing in dynamical systems: Foundations of harmony theory. Technical report, Colorado Univ at Boulder Dept of Computer Science, 1986.

[447] Asja Fischer and Christian Igel. Training restricted boltzmann machines: An introduction. *Pattern Recognition*, 47(1):25–39, 2014.

[448] Mehran Kardar. *Statistical physics of fields.* Cambridge University Press, 2007.

[449] I Nemenman. Renormalizing complex models: It is hard without landau. *J. Club Condens. Matter Phys*, 2017.

[450] Lars Onsager. Crystal statistics. i. a two-dimensional model with an order-disorder transition. *Physical Review*, 65(3-4):117, 1944.

[451] Rodney J Baxter. *Exactly solved models in statistical mechanics.* Elsevier, 2016.

[452] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

[453] Shayan Hassanpour, Dirk Wübben, and Armin Dekorsy. Overview and investigation of algorithms for the information bottleneck method. In *SCC 2017; 11th International ITG Conference on Systems, Communications and Coding*, pages 1–6. VDE, 2017.

[454] Jiuyang Alan Zhang and Brian M Kurkoski. Low-complexity quantization of discrete memoryless channels. In *2016 International Symposium on Information Theory and Its Applications (ISITA)*, pages 448–452. IEEE, 2016.

[455] Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

[456] Paul Fieguth. Fast matlab for ising. `http://ocho.uwaterloo.ca/Software/Ising/ising.html`, 2002. Private Communication, 2019 - 2020.

[457] David P Landau and Kurt Binder. A guide to monte carlo simulations in statistical physics. *A Guide to Monte Carlo Simulations in Statistical Physics, by David P. Landau, Kurt Binder, Cambridge, UK: Cambridge University Press, 2009*, 2009.

[458] DP Landau. Finite-size behavior of the ising square lattice. *Physical Review B*, 13(7):2997, 1976.

[459] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[460] Robert Austin. Evolution, physics, and cancer: Disrupting traditional approache. *Bulletin of the American Physical Society*, 61, 2016.

[461] Heidi Ledford. End of cancer atlas prompts rethink: geneticists debate whether focus should shift from sequencing genomes to analysing function. *Nature*, 517(7533):128–130, 2015.

[462] Chris Adami. Information, physics, and cancer. *Bulletin of the American Physical Society*, 61, 2016.

[463] James Surowiecki. *The wisdom of crowds*. Knopf Doubleday Publishing Group, 2005.

[464] Daniel Marbach, James C Costello, Robert Küffner, Nicole M Vega, Robert J Prill, Diogo M Camacho, Kyle R Allison, Manolis Kellis, James J Collins, Gustavo Stolovitzky, et al. Wisdom of crowds for robust gene network inference. *Nature Meth*, 9(8):796–804, 2012.

[465] Julio Saez-Rodriguez, James C Costello, Stephen H Friend, Michael R Kellen, Lara Mangravite, Pablo Meyer, Thea Norman, and Gustavo Stolovitzky. Crowd-sourcing biomedical research: leveraging communities as innovation engines. *Nature Rev Genet*, 17(8):470–486, 2016.

[466] Barry M McCoy. The two-dimensional ising model. 1973.

[467] BM McCoy and JM Maillard. The anisotropic ising correlations as elliptic integrals: duality and differential equations. *Journal of Physics A: Mathematical and Theoretical*, 49(43):434004, 2016.

[468] RJ Baxter. Onsager and kaufman's calculation of the spontaneous magnetization of the ising model. *Journal of Statistical Physics*, 145(3):518–548, 2011.

[469] Richard J Creswick, Charles P Poole, and Horacio A Farach. Introduction to renormalization group methods in physics. 1992.