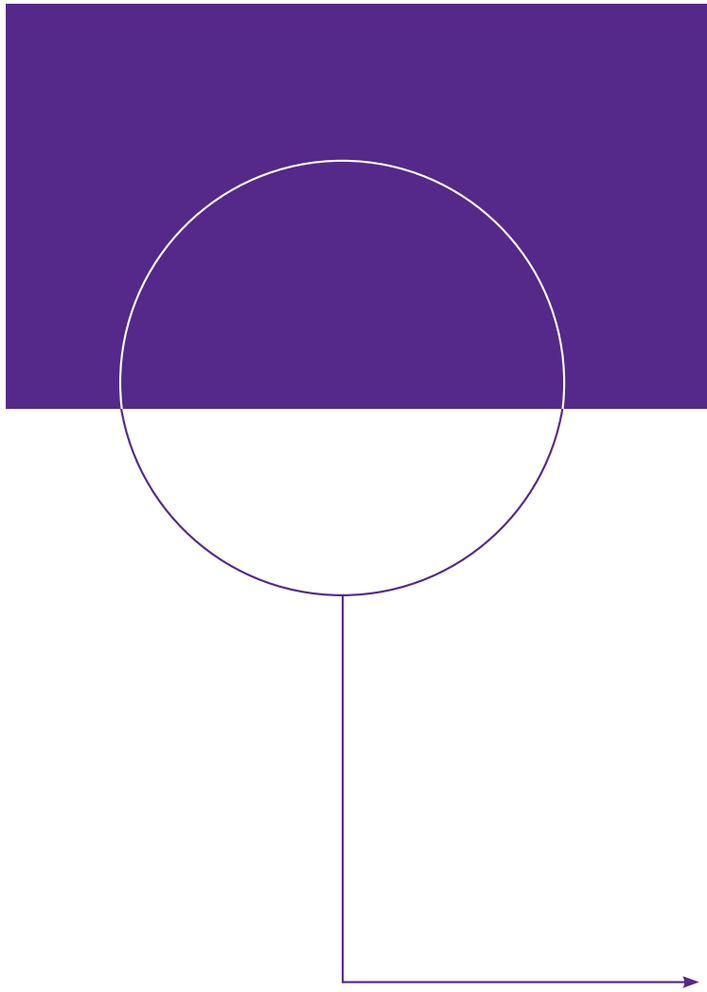


Master's thesis

2021

Michael Tarlton

NTNU
Norwegian University of
Science and Technology
Faculty of Medicine and Health Sciences
Kavli Institute for Systems Neuroscience

Michael Tarlton

Novel Model Selection Criterion for Inference of Ising Models

January 2021

NTNU
Norwegian University of
Science and Technology

NTNU
Norwegian University of
Science and Technology

NTNU

Norwegian University of
Science and Technology

Novel Model Selection Criterion for Inference of Ising Models

Michael Tarlton

Master of Science in Neuroscience
Submission date: January 2021
Supervisor: Yasser Roudi
Co-supervisor: Nicola Bulso

Norwegian University of Science and Technology
Kavli Institute for Systems Neuroscience

Master's thesis

NTNU
Norwegian University of Science and Technology
Faculty of Medicine and Health Sciences
Kavli Institute for Systems Neuroscience

Michael Tarlton

Novel Model Selection Criterion for Inference of Ising Models

Master's thesis in Neuroscience
Supervisor: Nicola Bulso, Yasser Roudi
January 2021

NTNU
Norwegian University of
Science and Technology

Michael Tarlton

Novel Model Selection Criterion for Inference of Ising Models

Master's thesis in Neuroscience
Supervisor: Nicola Bulso, Yasser Roudi
January 2021

Norwegian University of Science and Technology
Faculty of Medicine and Health Sciences
Kavli Institute for Systems Neuroscience

NTNU
Norwegian University of
Science and Technology

Thanks to Nicola and Yasser for their supervision,
and the support of everyone in the Spinor lab.
I'd like to acknowledge the Kavli Institute and NTNU,
as well as my fellow Neuroscience class of 2018.
The Coatney family for their critical support.
And everyone I've had the pleasure of knowing here,
on this crazy adventure to the North.

Summary

In this thesis we evaluate the performance of the novel Model Selection criteria proposed in Bulso et al. 2019, for inference of network topologies. To this purpose, we consider networks of binary nodes whose probability of activation is modelled by Ising models and generate data by simulating the network dynamics. After which, we infer the network topology by implementing the proposed criterion in a Bayesian model selection framework and compare the inferred topology with the ground truth model. The performance of the proposed method in recovering the network structure is contrasted with that of other popular model selection criteria in varied configurations of Ising parameters, network topologies, and sample size.

We begin by introducing the Equilibrium Ising model and proceed by describing the approximate solutions for making inferences in Ising models. The novel criteria is one of a class of selection methods adapting concepts from information theory, namely the *Minimum Description Length*; We will also discuss the nonscientific applications and parallels suitable to our approach.

Our results reinforce those found in Bulso et al. 2019. The novel criteria performs similarly to other selection criteria in the experiment regimes tested, with certain exceptions that will be addressed. Unique behaviors identified in the larger regimes may propose further avenues of investigation in networks of larger size and diversity.

Contents

1 Introduction

As above, so below.

1.1 Complex Dynamical Systems

Everything exists in a system, taking part in a greater gestalt; members of the system, interacting with the others in tightly interwoven connections. In disorder, or isolation, these pieces of the whole are limited, aimless, chaotic, but when part of an ordered network, they can give rise to complex behaviors. This phenomenon is reflected across all scales. A lone cell may only perform single tasks. Arranged properly, 10^{16} [14] more cells integrate to become as a whole, which in return may organize itself as part of collective, a colony, a society. In all aspects of the natural order, simple unit interactions merge giving rise to complex properties [120]. This occurs in biology (Figure 1), politics [35], markets [20] [71] [15], and sociology [46]; any abstracted network of interactions may be described in this manner. Lately, this abstraction of interactions is witnessed in machine learning and artificial neural networks. What are the mechanics underlying these systems, how are they defined?

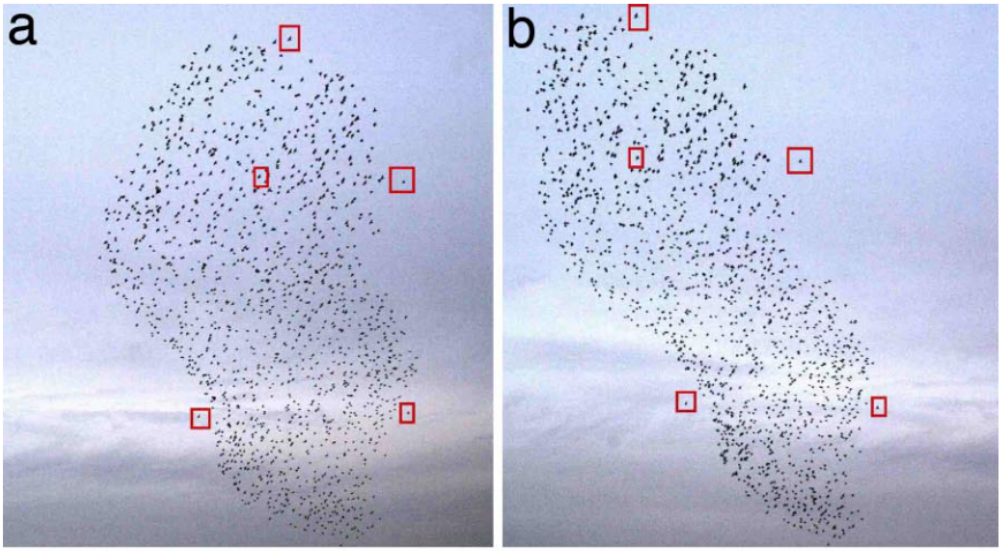


Figure 1
A flock of starlings whose collective flocking behavior was analyzed with a maximum entropy model of interactions between individual starlings and their "nearest neighbors": other birds whose behavior is closely paired with the individual. The stereographic photo allows 3D tracking of the birds' movement. The red squares highlight five matched pairs of birds. Adapted from Ballerini et al 2007, Figure 1 [9].

1.2 Statistical Mechanics and Systems Modeling

Statistical mechanics is the methods used to model the dynamics of complex systems such as the behaviors of gases, liquids, and other large particle bodies. These methods eventually expanded into describing the interactions in other complex dynamical systems.

Statistical mechanics reduces high dimensional problems to the behavior of a volume, or *field*, of particles; first by describing the particle-to-particle interactions, then scaling those descriptions to a statistical summary of the interactions underlying the whole. This system of relationship between particles is interpreted by a map of the system and its states as a network of nodes.

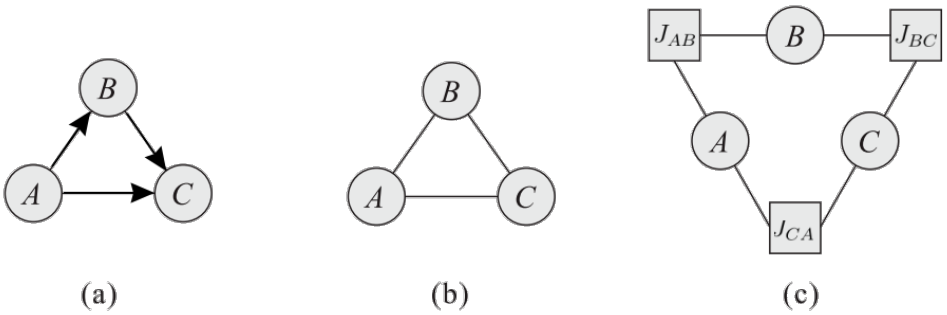


Figure 2
Nodes are visualized as having some connection to each other by the *edges* in the graph. **(A)** A *directed graph* where a connection is one-way between units. **(B)** An *undirected graph* where the connection between two nodes is unidirectional. **(C)** A *directed weighted graph* where some level of connection strength is set but is still unidirectional $J_{AB} = J_{BA}$. Adapted from Koller 2009 [28].

The *Markov network* [77], an *undirected graphical model*, maps these relationships between elements as a set of parameters in a graph (Figures 2 and 3). These parameters can represent the state of an node and its interactivity with other elements. The collective interactions between nodes creates an ongoing stochastic change in the states of the nodes (e.g. the on or off firing of a neuron) continuing over time. The states of an element at one moment in time, directly causing the change in states at the next moment in time. A *Markov process.*

Similarly, neurons do not activate independently but rely on a highly interconnected set of relationships between neuronal units, firing in correlated, redundant patterns [3]. The Hopfield model [54] is a Markov network designed to replicate the spiking dynamics found in the neural ensemble, which is believed to be the basis for encoding information and particular brain states [115], using the terms of the *Ising spin model.*

If one were to take a "snapshot" of the Markov process of the network, they would observe the *system state distribution:* the state on all individual elements at a discrete step in time. In a system of binary spin states this would be represented by a string of binary values, e.g. $[0, 1, 0, 1, 1, 0]$ each binary variable representing the state on an individual element. In a neuro-anatomy sense, this would be analogous to the firing state of all neurons in an observed assembly during a discrete time-bin: $firing = 1$, and $not\ firing = 0$.

This string of binary variables is akin to a "code-word", which may encode information such as a stimulus, a memory, or resting state. Subsequently, this configuration may also be titled a *spike-word* as it describes the spiking state of a neuronal network.

The time-series of states produced by the Markov process is the *Markov chain*, i.e. the output of the system process and its distribution of states over some amount of time, analogous to the electrophysiological *spike trains* observed in biological systems [86] [50].

We arrive at the core problem: often we are able to observe the change of states in some system without meaningful access to the underlying causal structure in the system. How then can a representation of the system be reconstructed from its observed output states?

If the observed outputs are dictated by statistical rules governing the interactions of the system, then underlying statistical dependencies in the system should be inferable if given sufficient observation of the system's process. The methods used in statistical mechanics to describe the functions in a markov process is the *forward process.* For a system like a Markov network that is governed by such functions, inversions of the functions can be developed to create an *inverse solution.*

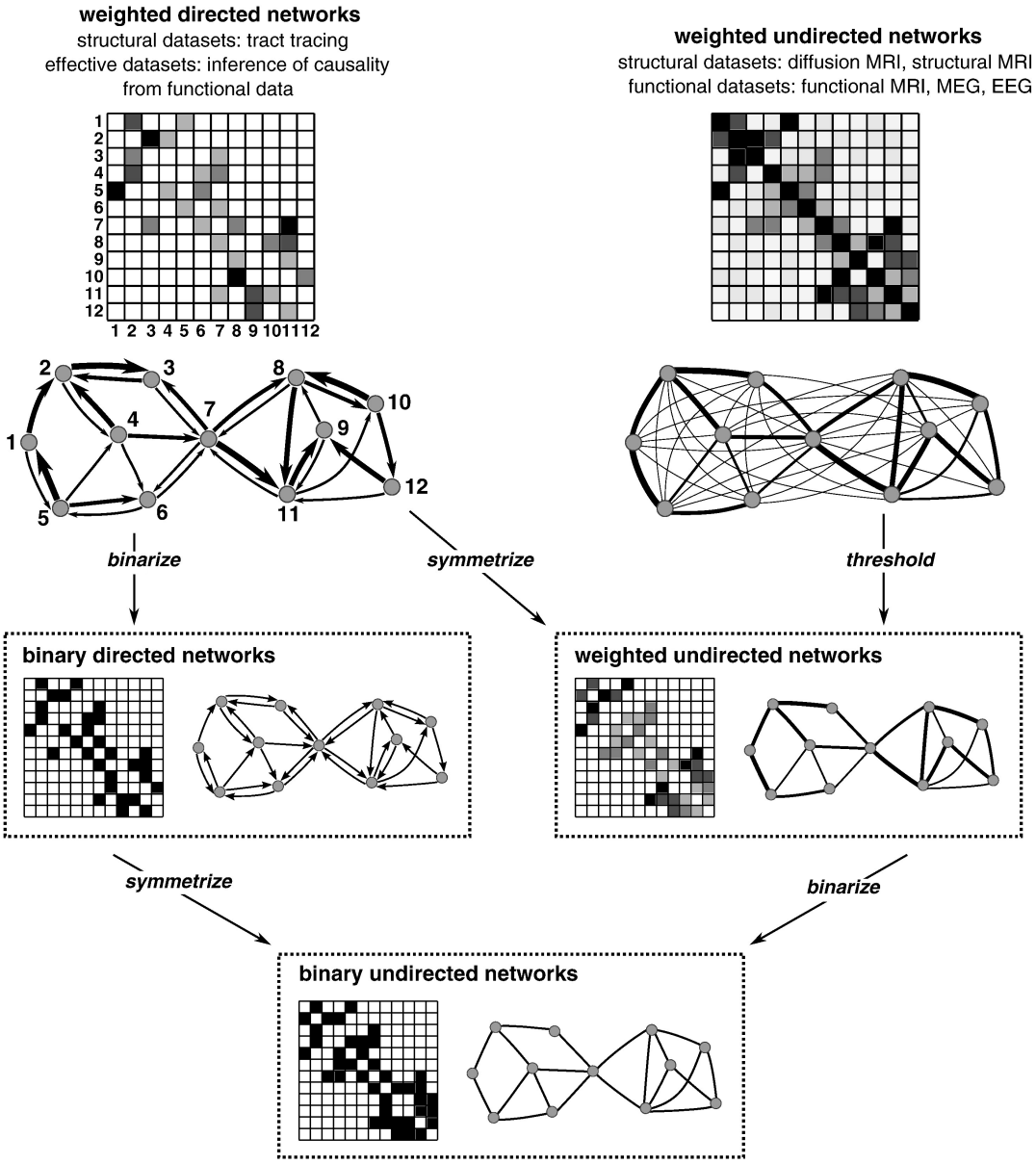


Figure 3
Further illustration of graph models and their representation in a graph array. Sporns notes the biophysical data types analogous to these network models where the connection where each row and column represent the nodes and the individual cells of the matrix represent their mutual connectivity. Here the term *binarize* refers to reducing continuous weight values into two discrete terms; *symmetrize* refers to converting directional connections to a unidirectional model (shown here as done by a logical OR decider, where only the bidirectional connections are kept), and finally, *thresholding* where weighted terms below a certain connection strengths are omitted. Adapted from Sporns et al. 2010, Figure 1 [109].

A staple technique of graphical model inference is the *Bayesian inference methods*. These use Bayes formulas [68] as a framework for finding the likelihood of a possible state on an element; in this context inferring the system parameters which are most probable in the production of an observed distribution of output states. A method referred to as "fitting" a model of a system to the observed information (Figure 4).

The *maximum entropy model* or *pairwise equilibrium Ising model* provides a viable model in the study of networks and has become popular in problems of inference [80], due to its large and well studied inverse methods. The inverse Ising methods are shown to be highly effective in network reconstruction, particularly when paired with *Bayesian model selection methods* [87] [66] [42]. The

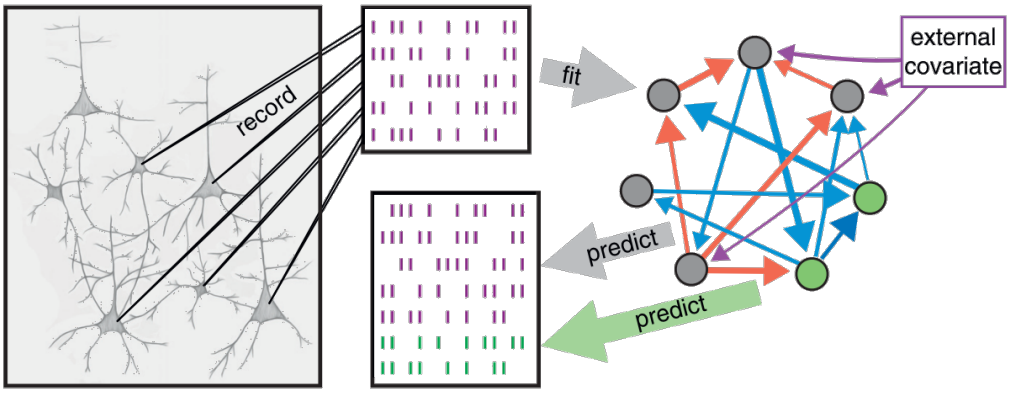


Figure 4
Illustration of the general process of recording biophysiological spike data (from neurons in this example), and fitting this data to a statistical model of network connectivity. This figure also refers to *unknown variables* unobserved neurons or elements in the system whose effects can be indirectly inferred. Adapted from Roudi et al. 2015, Figure 1 [99].

Ising model provides a parametric framework to describe a network model, while a Bayesian criteria adds additional constraints to an inverse solution.

The primary goal of any model reconstruction technique is to accurately reconstruct a large diversity of network models without: **a.** *over-fitting* to one type or types of networks, and **b.** retaining the highest level of detail as possible in reconstructing of the original model. *Model selection criteria* can also be described as penalization modules, reducing the model parameters to only the most essential elements in a process called *Occam fitting* [28] [69]. This follows the principle of Occam's Razor: the model which best describes the observed data, will be the simplest model.

The inverse problem is a computationally expensive one. Inferring an exact solution for the *structural connectivity* of a Hopfield network becomes infeasible for networks with more than some tens of neurons. The maximum entropy model is a preferable statistical representation of the neural network, modeling the *functional connectivity* of a network (see functional connectivity inset and Figure 5). In this model the network is represented by the pairwise interactions between nodes, mapping the weighted values between each set of two nodes. This approach has been shown to be effective in correlating neural data [24] [116].

Types of network inference schemes are generally split into two classes: *parametric* and *non-parametric* models, here we focus on a parametric approach, i.e using the parameters of the Ising model. An exhaustive review of current inference techniques can be found in Abril et al. 2018 [70] and Gardella et al. 2018 [40]. Both provide excellent overviews of the mathematical models being applied in connectivity inference, and the challenges associated to each, in the context of neural recording data.

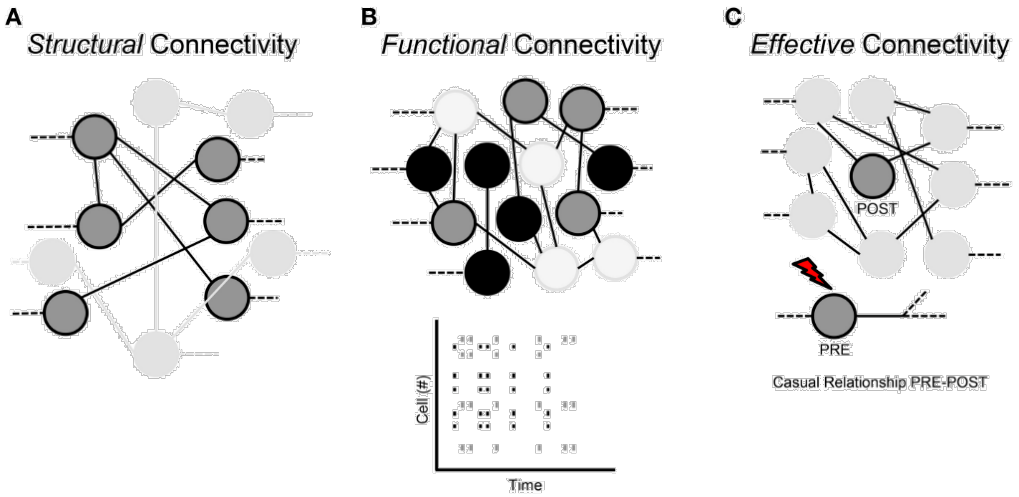


Figure 5
Definitions in interpretation of connectivity. **(A)** The structural connectivity, the physical structure of neurons. **(B)** The functional connectivity, where we can see the mutual activity between nodes within an observed spike train . **(C)** The effective connectivity, where the influence one neuron has on another is observed by stimulation of one and the respondent activity of an affected cell. Adapted from Poli et al. 2015, figure 3 [89].

Functional Connectivity

The human connectome is a comprehensive structural description of the network of elements and connections forming the human brain. Cortical areas are neither completely connected with each other nor randomly linked, instead their structure shows a specific and intricate organization [109]. Friston 1994 distinguished two types of interconnections as maps of*functional* and *effective connectivity* [39], a third interpretation later considered alongside these is *Structural connectivity* [110].

Structural or Anatomical connectivity is the physical makeup of the neural connectome. The physical interactions via electrical or chemical synapses which determine the mapping of a neuronal communication network. This ranges over multiple spatial scales as the connections can be located both in local neuronal circuits and in long-range communications linking other sub-networks [89] [18].

Effective connectivity describes the causal effects of one neuronal unit on the other by direct means, once any indirect means have been discounted [70]. In a highly interconnected system, the dominant source of correlations between two neurons will always be through the multitude of indirect paths involving other neurons [115] [44]. The "effectiveness" being any observable interactions between two neurons, which alters their activity. This can be inferred by inducing perturbations in the network or observing the temporal order of neuronal activities [41].

Functional connectivity is the statistical representation of a network where dependence and independence between neuronal units obtained by measurements of neuronal activity [70]. By measuring the correlation between spikes coming from different neurons over some time series, predictions can be made about the activity of one of the two neurons based on the activity of the other neuron [89]. Functional connections is considered a subset of the structural connectivity as the properties of a single neuron are dependent on their anatomical connections [111]. Functional connectivity is evaluated among all the elements of a system, regardless whether these elements are connected by direct structural links [41]. Functional connectivity is shown to be effective at reproducing a network structure and is particularly useful for understanding models with hidden nodes [20] [32] [99] [11].

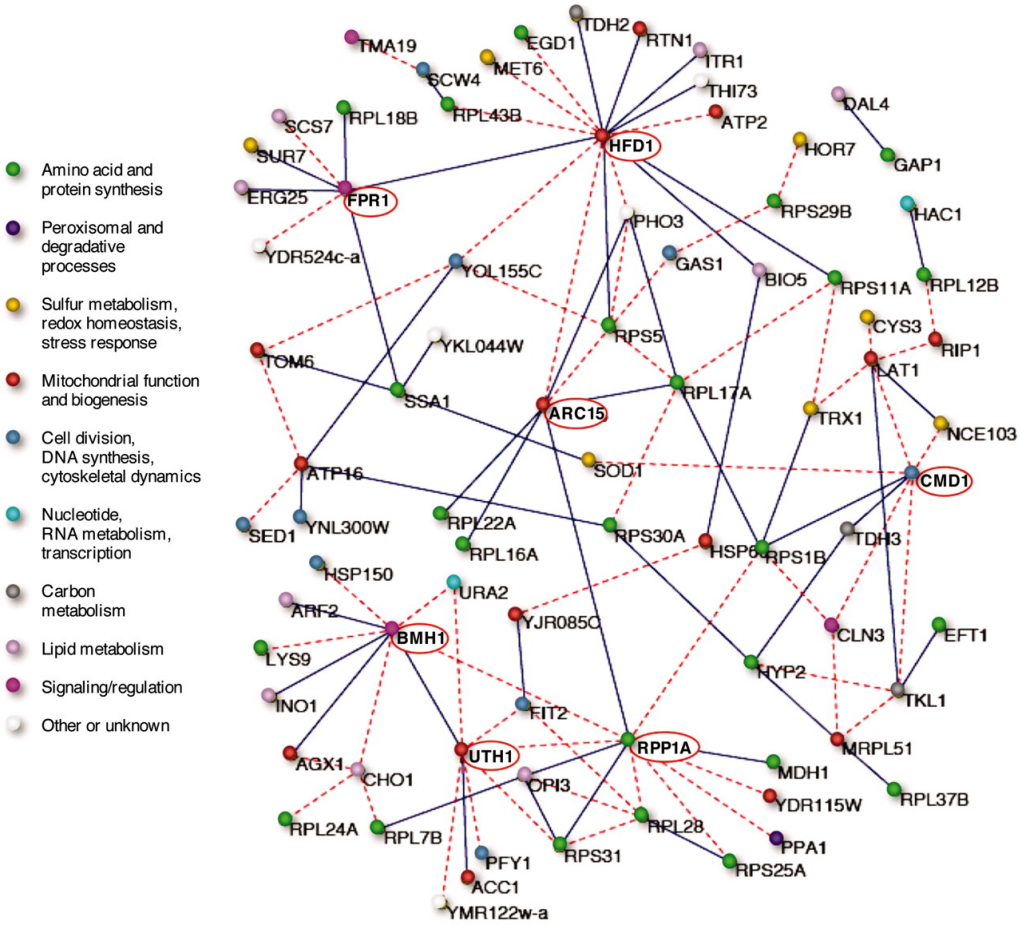


Figure 6
A maximum entropy model used to form an undirected graph of interaction in gene expression behavior from a pool of 582 genes. Pictured are the 110 strongest interactions after thresholding weakly correlated interactions in a full network of 169,071 interactions. Nodes are identified by gene names and color-coded to indicate the cell process in which they participate. Positive in interactions correspond to the solid blue edges, while negative interactions correspond to dashed red edges. Adapted from Lezon et al. 2006 [65].

1.3 Statistical Physics in Biology

Modeling biological processes with Markov models has been found to be an effective tool, with statistical solutions having been utilized effectively in sub-cellular problems such as the interactions of multi-molecular chains in protein folding structures [57] [25] [71] and gene expression patterns [65] [7] (Figure 6), all the way to the macroscale, modeling animal collective behavior [22] (Figure 1).

It has been suggested that biological systems consist of integrated elements poised at a point of self-organized *criticality* [76], an equilibrium between constraints of entropy and energy [79] [11]. Systems of nonlinear dynamics studied in statistical physics exhibit similar properties, where self-organized systems regularly balance between complexity and chaos.

In the Hopfield model the dynamics of the neural network can be imagined as motion on the energy surface, an abstract 2D plane where a multi-dimensional problem is reduced to a flat surface, and levels of energy pock its landscape with hills of high energy and basins of low energy. On this surface, local minima of energy where the system can "settle" result from the competition between positive and negative interactions directed by the Ising parameters, these stable attractor states can represent stored memories or brain states [55] [56]. As will be demonstrated in Section 2, probability distributions of system states are localized to an attractor basin of minimal energy in the region of phase-state space defined by a configuration of Ising parameters.

1.3.1 Statistical Mechanics in Neuroscience

Advances in morphogenetic neuro-engineering have created novel methods of direct imaging and neural ensemble recording. *In vitro* neural cell cultivation, the growth of neural cells on specially designed substrates, makes it possible to study the activity of neural circuits at finer resolutions. Another well established approach, involves growing monolayer neural ensembles from dissociated neural tissue or stem cells. These express fundamental traits of brain networks, such as self-organization, spontaneous network formation and interactivity, are reproduced in these models [118]. Neuronal activity produced by these ensembles are then recorded by microelectrode arrays (MEAs) or Optical and Optogenetic imaging. Current commercially available MEAs can provide 60–120 electrodes with 100–500µm inter-electrode spacing or up to thousands of microelectrodes (4000–10,000) and high-density MEAs with a spatial resolution in the tens of micrometers (Figure 7) [89] [38] [6].

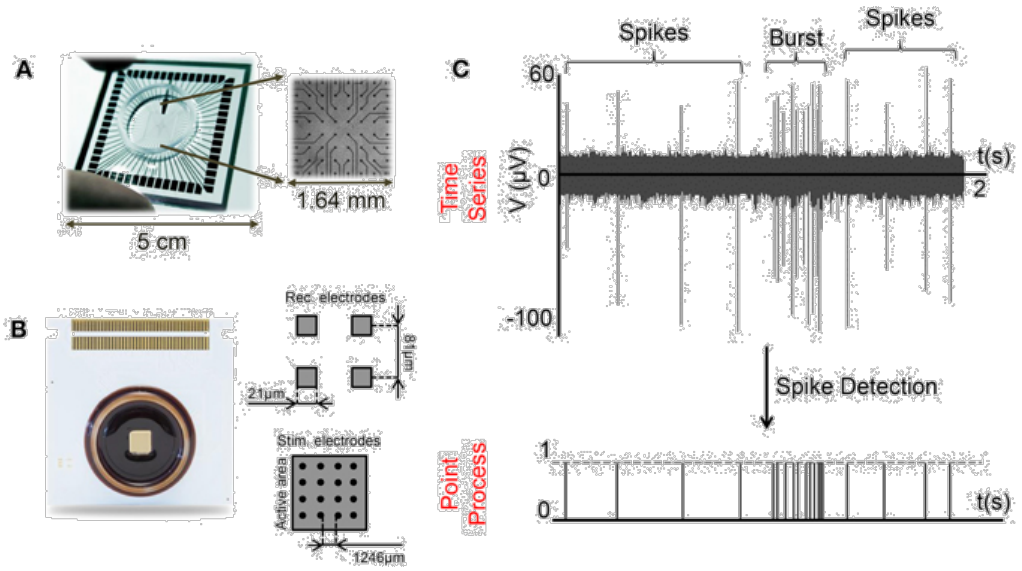


Figure 7
Multi Electrode Arrays (MEAs). **(A & B)** Example MEAs, **(A)** has 60 individual recording sites and **(B)** with 4096 recording sites. The continuous time series spike data from these presents a mix of bursting and spike activity which must be discretized into binary data; as done here in a serial point process [6]. Adapted from Poli et al. 2015 [89].

Alternatively, *in vivo* spike train data may be obtained by use of neuropixel probes (Figure 8), which can be placed in target lobes of an animal and provide recording data from thousands of individual node sites along a single probe shank [112]. These also provide opportunity to record across multiple lobes and layers of live neuronal tissue, allowing for monitoring of communication comparisons between brain sections while an animal responds to stimuli [60].

Both these options present an opportunity to use spike data in creating novel models of brain activity and structure, but this also comes with the challenge of processing and interpreting datasets of such large dimensions [16] [17]. These techniques will only increase in resolution and data dimension as methods further develop and refine.

Another exciting possibility is the statistical analysis of electrophysiological data generated by neuronal cells cultivated with specific neuroanatomical conditions or pathologies [67]. Statistical inference of neural activity in these ensembles could allow for insight into the functional connectome structure of these networks and comparison with the maps of functional connectivity in healthy neuronal networks. Because functional connectivity is the effective statistical connections between nodes, treatments could be developed with the intent of restoring a functional connectome map in a damaged neural connectome with a quantitative metric.

Any number of unique neural ensembles could be classified and compared with others based on neural activity. Two papers from Valderhaug and coauthors [119] [117], use this approach to investigate both structural and functional changes of *in vitro* human tissue derived neural networks monitored by MEAs. These studies captured the network activity of healthy neural networks and made comparisons with neural networks that had introduced pathological conditions consistent with Parkinson's disease. Functional connectivity was done by analysis of electrophysiological recordings, while structural connectivity was obtained by optical analysis.

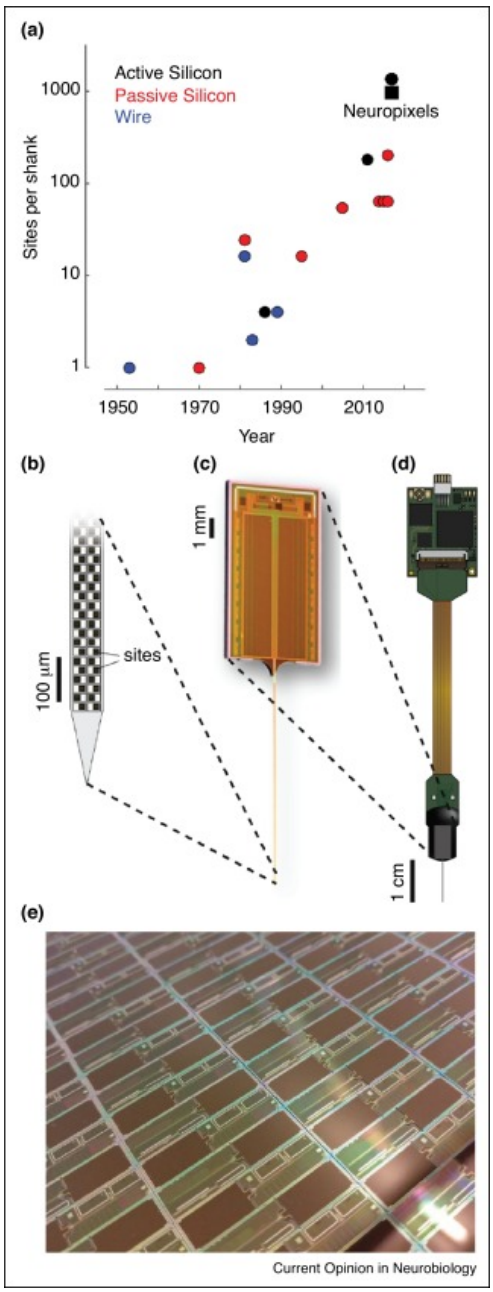


Figure 8
Overview of the growth in electrode technology.
(A) Density growth of electrodes per shank over the years.
(B-D) Schematic of the Neuropixel probe.
(B) The tip with electrodes arranged in a dense checkerboard pattern.
(C) The printed CMOS element, including the shank as well as circuitry implementing amplification, multiplexing, and digitization.
(D) The packaged device with flex cable and headstage for interfacing and further multiplexing.
(E) Picture of neuropixel probes on a CMOS wafer.
Adapted from Steinmetz et al. 2018 [112].

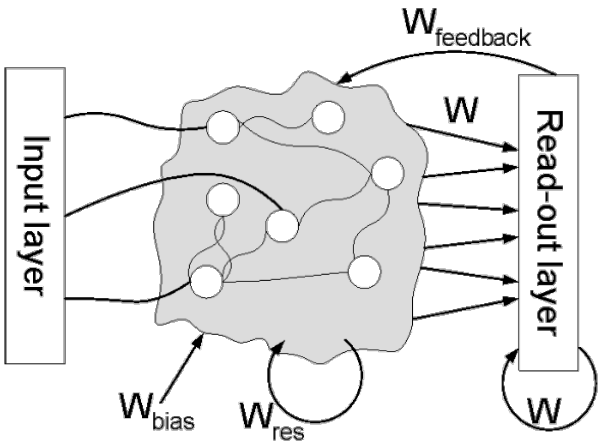


Figure 9
Schematic of a reservoir computer where the middle "blob" is the reservoir: some self-organizing system with desired properties. In our example an MEA. This is perturbed by stimulation from an input layer which the reservoir will self-organize in response to, effectively processing the input data. This is then attached to an output layer, typically some sort of directed artificial neural network which can be trained to the desired task and even back-propagate to the reservoir. Adapted from Schrauwen et al. 2007 [106].

In vitro neural networks grown on MEAs have also been studied for potential application in *biological reservoir computing* (Figure 9) [62] [49] [90]. *Reservoir computers* are computational modules which rely on some self-adjusting, dynamic system which can self-organize in order to simplify complex, nonlinear data. Much like biological systems, they are found to be most useful when poised at criticality, an "edge of chaos" between order and disorder [91]. The computational capacity of *in vitro* neural networks has been studied for use in simple computational tasks [49] [90]. Aaser et al. 2017 [1] uses a biological neural network paired with an artificial neural net interpretation layer in simulated guidance tasks. These methods interpreted the output activity of the *in vitro* network by means an artificial neural network interpretive layer. A means of inferring the functional connectivity of a neural culture could give extra depth to the capabilities of *in vitro* neural reservoir computing.

1.4 Proposal and Building on Bulso 2019

Bulso et al. 2019 [19] introduces a novel Bayesian selection criteria based on the concept of *Minimum Description Length (MDL)*, an information theory implementation on Occam's razor. The MDL principle is the ansatz: "Choose the model that gives the shortest description of data [96]." Other model selection techniques based on this principle [8] [96] [97] [78] precede the Bulso et al. 2019 novel MDL criterion. However, the novel criterion proposed uniquely implements frequency distribution of unique spike-words in localizing the family of possible models and may show an advantage over the classical methods in regimes of high informational entropy. This is paired with logistic regression technique analogous to an inverse Ising technique known as the pseudo-log-likelihood. In this thesis we test the ability of the Bulso et al. 2019 novel MDL criterion to reconstruct the structure of an Ising network model in a variety of network topologies, conditions, and observation sample sets.

1.5 Paper Structure

This paper will construct the basis of Ising network models and build the methods employed both their construction as well as the inference problem. This will span across the disciplines of graph theory, statistical thermodynamics, information theory, and Bayesian model selection. Throughout we'll ground these methods to their mathematical motivations and material parallels with focus on the neuroscientific regime. Furthermore, we will build upon these mechanics underlying the model selection method used. Once motivations, background, theory, and methodology has been exhausted, we will demonstrate their implementation in the experiment and analyze the results. We will finish by addressing the experimental results and discussing the findings, ending with projections of future directions in which to continue.

2 The Ising model

2.1 The Ising Model

The *Ising model* is a parametric model adopted from statistical mechanics. It was originally used as a model of dynamics in a ferromagnetic lattice, whereby the spin of each polar moment is influenced by the field of magnetic energy exerted on it by neighboring magnetic moments. The Ising model has since made the transition as a model for Markov network state statistics. Its well-studied properties provide sufficient statistics for problems of inverse system dynamics [40] [80] and its binary properties allow application of information theory concepts to the inference problem [48] [69].

The Ising model is a network as a system of interacting nodes which produce a distribution of binary variables $\{-1, +1\}$. The individual binary state on each node is the *spin* σ, with the spin state of each node i influenced by the spin of its neighboring node j (Figure 10). The *connection strength* J_{ij} determines the level of interaction between two nodes, when the node j express their spin on i or vice-versa as $J_{ij} = J_{ji}$. Each node is also influenced by its own bias h_i, which influences its own spin activity. This is referred to as the *external field*, or simply, the *bias*. The field of effect exerted on a node by its neighbors is measured as the surrounding energy $E(\sigma)$ as expressed by the *energy function*, the Ising Hamiltonian,

$$E(\sigma_i) = -\sum_{i<j} J_{ij}\sigma_i\sigma_j - \sum_i h_i\sigma_i, \tag{1}$$

where σ is the spin $\sigma_i \in \{-1, +1\}$, exhibited by the nodes.

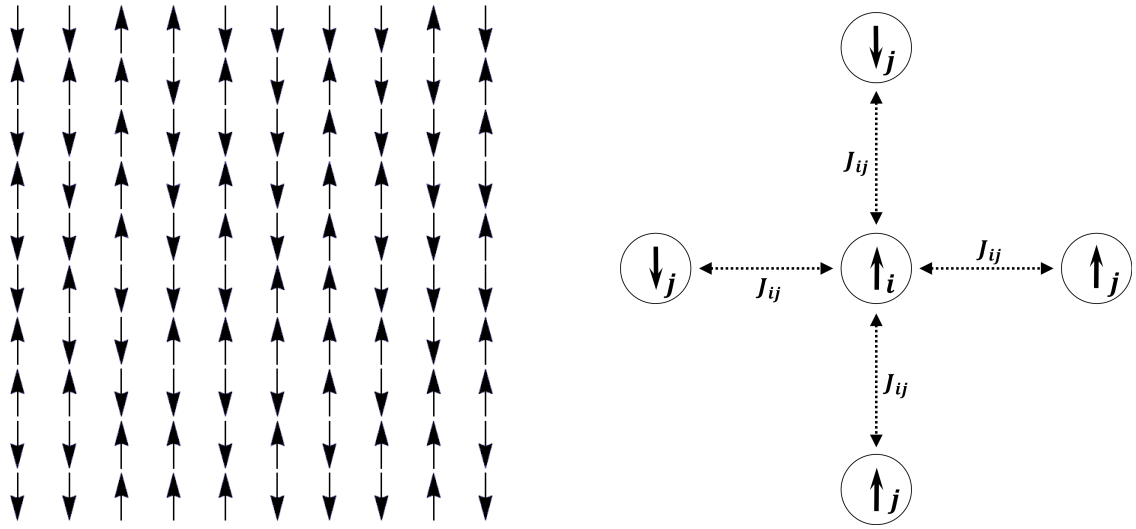


Figure 10
(Left) A lattice of polar spin moments in an two-dimensional Ising model. The up arrows represent a positive spin $\sigma = +1$, and down arrows represent a negative spin $\sigma = -1$. **(Right)** Illustration of nearest neighbor interactions, where the node in the middle, i is being acted upon by its nearest neighbors j with connection strengths J_{ij}.

This field of effect is calculated for all nodes in the system at each moment in time, with the level of energy exerted on the nodes dictating their spin state. Simulating this dynamic process over a network of spins is the *Glauber Dynamics* of the Ising system (Figure 12) [45]. Each discrete time step in the Glauber process is given as $t = [1, ..., T]$ where T is the total number of time steps observed. The spin state over the network updates at each new time step in the process $t+1$. The spin configuration of the network is represented as the spike-word vector $\vec{s} = \{\sigma_1, \sigma_2 ..., \sigma_N\}$ for N total nodes in the system. The total process output of time-discrete network spin-states observed over the time T is the spike-train array $\hat{S}$ (Figure 11) denoted,

$$\hat{S} = \{\vec{s}^{\,1}, \vec{s}^{\,2}, ..., \vec{s}^{\,t}, ..., \vec{s}^{\,T}\}. \tag{2}$$

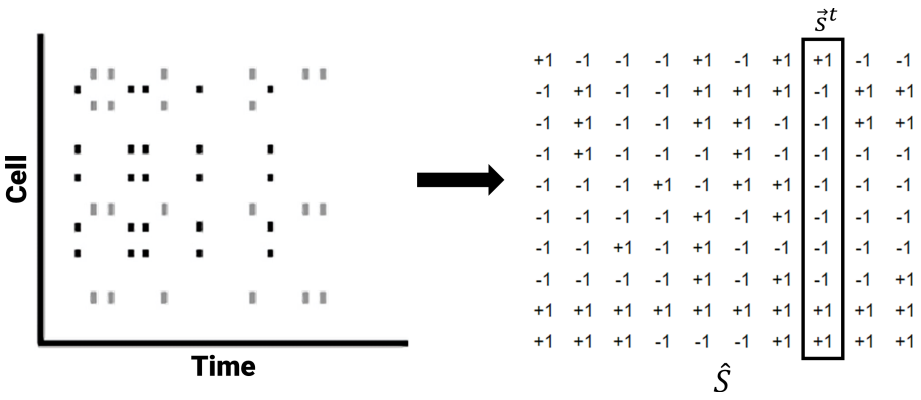


Figure 11
Biophysical spike data converted into its Ising Glauber interpretation. The array $\hat{S}$ is analogous to the *spike-train* observed in biophysical contexts. Likewise, the discrete spin state of the network $\vec{s}$ is a *spike-word* where each spike-word or combined pattern of spike-words can encode some information.

The process of generating the Glauber dynamics as the *forward Ising*, contrasting with its inversion, the *inverse Ising problem*. The inverse method begins with observing the spike-train, a distribution of spin-states output by the function of a system. The Ising network capable of producing a particular distribution $\vec{s}$, is defined by a similarly unique configuration of the parameters (J_{ij}, h_i). Because the Glauber dynamics produces a stochastic output, if we are provided sufficient observation samples, we may infer the parameter configuration with the highest probability of producing the samples [100] [5].

Figure 12
Interactive model of Ising Glauber dynamics for a 2D lattice, generated from a simple Gibbs sampling implementation [61]. Here the positive spins are represented as in white and the negative spins represented in black. The Gibbs sampling is initiated with some randomness with a set pairwise interaction strength and external bias for all nodes. The equilibrium state of the Glauber dynamics can be seen in the Turing pattern visualization. This pattern of self-sustaining equilibrium dynamics can be better seen in the animation available in the online version.

The probability of a spin on a node $P(\sigma_i = \pm 1)$ is given by the *Gibbs-Boltzmann Distribution*,

$$P(\sigma) = \frac{e^{E(\sigma)}}{Z}, \tag{3}$$

where the partition function $Z = \sum_S e^{E(\vec{S})}$, is the normalization factor. In the minimal example of a system containing a single node, there are two possible states of the network, $(+1, -1)$, every additional node added to this system grows the complexity of this probability exponentially 2^N where N is the total number of nodes.

The probability of the spin state $\vec{s}$ for a system of nodes i and their interacting nodes j is,

$$P(\vec{s}) = \frac{1}{Z} \exp \left[\sum_{i<j} J_{ij} \sigma_i \sigma_j + \sum_i h_i \sigma_i \right]. \tag{4}$$

Assuming the system states observed at all time steps are *independent and identically distributed* (i.e. probability is independent of the previous time state as opposed to how it would be in a Generalized Linear Model, see inset: *Ising Model in Biodata*), the probability of a spike-train configuration is,

$$P(\hat{S}) = \prod_t^T \frac{e^{E(\vec{s})}}{Z} = \frac{1}{Z^t} \exp \sum_t \left[\sum_{i<j} J_{ij} \sigma_i \sigma_j + \sum_i h_i \sigma_i \right]. \tag{5}$$

The exact inference of the parameters J_{ij} and h_i quickly becomes a computationally intractable problem.

The pairwise equilibrium Ising model assumes the system of interactions has settled in a Gibbs-equilibrium steady state, essentially gravitated into a basin of low-energy on the hyper-plane of phase-state space. In this attractor state, the output distribution, the activity of the neuronal population, abides by a stochastic behavior, with a particular pattern of output states. In this model the connection strengths between nodes is a symmetric weighted edge $J_{ij} = J_{ji}$ as thus what is being inferred is the pairwise activity between the nodes. The Ising *expectation values* (m_i, m_j, c_{ij}), are the minimal sufficient statistics required to infer the network interaction parameters, where the *magnetization* m_i is the average spin on a node over all observations, m_i,

$$\langle \sigma_i \rangle \equiv \frac{1}{T} \sum_t \sigma_i^t, \tag{6}$$

and the *pair correlation* c_{ij}, is the mean correlated spin over the observations, c_{ij},

$$\langle \sigma_i \sigma_j \rangle \equiv \frac{1}{T} \sum_t \sigma_i^t \sigma_j^t, \tag{7}$$

The *pair covariance* (or pairwise connected correlation) χ_{ij},

$$\chi_{ij} = \langle \sigma_i \sigma_j \rangle - \langle \sigma_i \rangle \langle \sigma_i \rangle, \tag{8}$$

is also an important metric as we will show shortly. Maximizing the Ising probability function with respect to the minimal sufficient statistics (i.e. the expectation values) reduces the computational complexity of the problem, while still returning the parameter reigon of highest likelihood (Figure 14).

Ising Model in Biodata

The Hopfield model [53] [3] adapted the concept of a neural network to an Ising model and it has since gained traction as a parametric model for contextualizing neural activity [104] (Figure 13) and biological data where connected systems may not be directly observable [81] [99]. However, as biological data is typically continuous in nature, it must first be discretized if used with an Ising model. So, neural electrophysiological recordings require some method of binning the continuous spike data into the time bins t, and thresholding neuron spiking activity into binary representations $\sigma_i^t = \pm 1$.

Knowing the mechanics of neuronal communication one might use a direct inference method, the full inversion of the Glauber dynamics, taking the probability of a spin state as dependent on the previous spin state $P(\vec{s}^{\,t}|\vec{s}^{\,t-1})$. This is the Generalized Linear Model (GLM) [99], which also considers the direction of effect between nodes, but is computationally difficult.

The equilibrium Ising or maximum entropy model, models a network at a *Boltzmann equilibrium* (somewhat analogous to the critical state of a neural ensemble) and constructs the *functional connectivity*, a statistical summary of the network's structure [89] [24]. The maximum entropy model is shown to give a closer reconstruction of a network when compared to models which treat the neuron firing rate as disconnected from other neurons in the network (Figure 15) [100] [104].

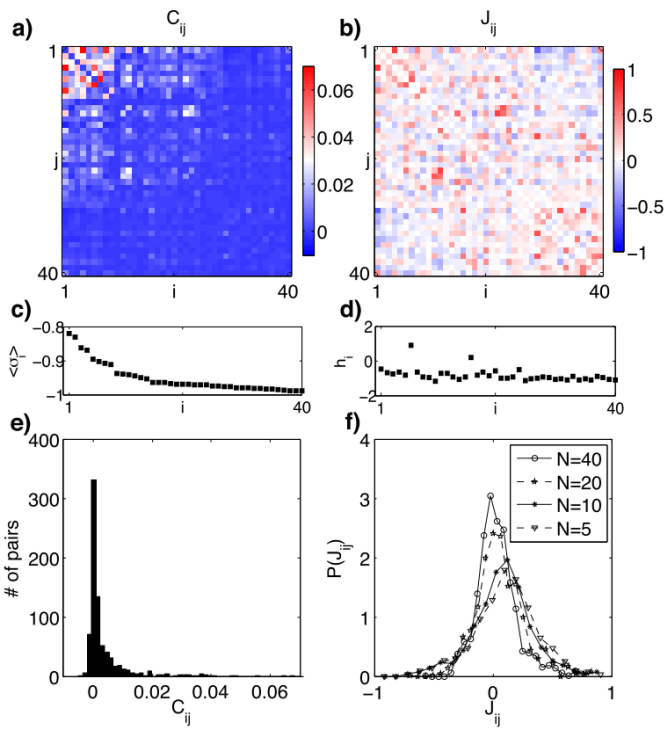


Figure 13
Here the expectation values and corresponding Ising terms have been computed from real neural datasets [93] [105]. The left column of figures shows the correlation data taken from the data and the right column represents the effective Ising values inferred from the maximum entropy model. Note the distribution of the Ising connection strength J_{ij} is a Gaussian distribution with a mean set about 0. The neurons are ordered by descending mean spike rate.
(a) The pair covariance $\chi_{ij} = \langle \sigma_i \sigma_j \rangle - \langle \sigma_i \rangle \langle \sigma_i \rangle$ for the neurons. **(b)** The inferred pairwise connection strength between neurons J_{ij}; note that the interactions are spread more uniformly throughout the network than the pairwise connected correlations. **(c)** The mean magnetization m_i of the individual neurons. **(d)** The bias h_i of individual neurons. The intrinsic tendency of the neuron towards spiking or silence. **(e)** The histogram of correlations. **(f)** The inferred connection distributions for sub-networks of varying sizes. Adapted from Tkacik et al. 2009, figure 1 [115].

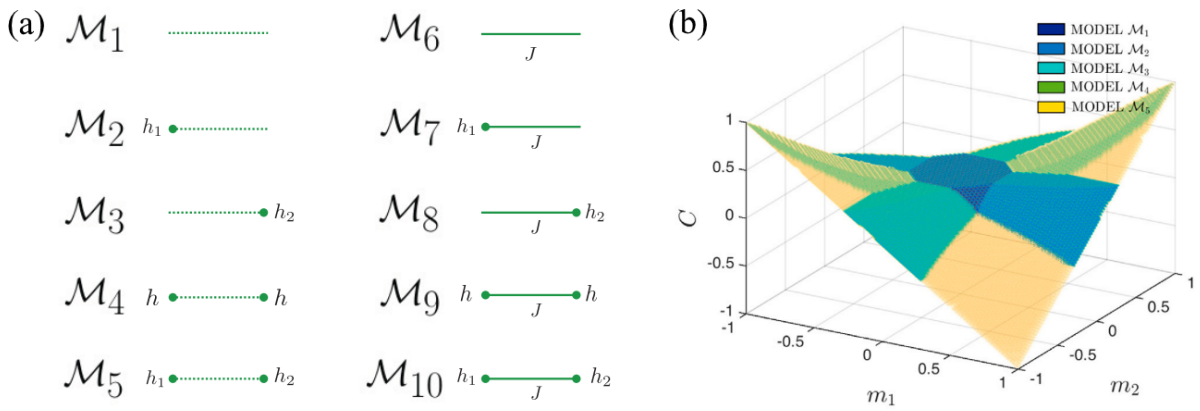


Figure 14
Maximizing the Ising probability function (Equation 4) for the expectation values by saddle point optimization will return the parameters of highest likelihood [78]. An example of the connectome model space for the minimal cluster ($N = 2$) as dimensionalized by the expectation values is exhaustively in Bulso et al 2016 [20]. **(a)** Ten models for a system of the minimal cluster where the network size is $N = 2$. Here the presence of a non-zero connection J is represented by a solid line and the presence of a nodal bias h is represented by a dot. Note a difference is made between models $\mathcal{M}_4$ & $\mathcal{M}_5$ as well as $\mathcal{M}_9$ & $\mathcal{M}_{10}$ as in the case where $h_1 = h_2$ the nodes are conditioned by the same bias, effectively reducing the inference problem to a different model space. **(b)** The model space for the first five models which have no interaction with each other. This space is indexed by the expectation values $\{m_i, m_j, c_{ij}\}$ and the regions of highest likelihood for the respective models are represented by color. Adapted from Bulso et al. 2016, figures 1 & 2 [20].

2.2 Maximum Entropy

The Maximum Entropy principle [59] (maxent) states that among all distributions compatible with a set of measured observables, one should choose the distribution with maximum entropy [59]. In this context this is the informational entropy which is used as a measure of ignorance when selecting a distribution (see inset: *Entropy*). By this principle, it is preferable to select a distribution which does not add any additional biases or extra constraints to the set of possible distributions.

A Gibbs-equilibrium distribution is at maximum entropy when its expectation values match the observed data. Thus, for the distribution of a spin $P(\sigma)$ indexed by some parameters $\theta = (J_{ij}, h_i)$, the expectation values of the distribution will approach the same mean values of the observed spike train [45] [81],

$$\sum_{\sigma} p_\theta(\sigma)\sigma_i = \langle \sigma_i \rangle_{\text{observed}} , \tag{9}$$

$$\sum_{\sigma} p_\theta(\sigma)\sigma_i\sigma_j = \langle \sigma_i\sigma_j \rangle_{\text{observed}} , \tag{10}$$

the parameters θ are then maximized within this constraint. Here we use the word *indexed* to describe the configuration of Ising parameters capable of producing a particular distribution of output spin states when introduced to the probability function (Equation 4).

Entropy

Entropy is the measure of possible state configurations for a system or how much information can be encoded into a system. In the case where we are attempting to find a specific configuration state of the system while otherwise uninformed about it, we can think of the entropy as a measure of uncertainty. The higher the entropy, the more possible system configurations, the more difficult it will be to find a specific configuration. We measure the entropy of a probability distribution P as,

$$S[P] = -\sum P(s) \ln P(s). \tag{11}$$

Entropy in this sense is the *expected value* of the possible system states s. To measure the difference between two distributions, we use the Kullback-Leibler (KL) divergence to find this distance between the two. However, it is not a very informative measure as it reaches zero when the distributions are equal, and for all other values can only tell us that they are dissimilar [100]. So if we have our original distribution P_{true} and the probability distribution which we have inferred P_{recon} the KL divergence between the two is measured as,

$$\begin{aligned} KL\left(P_{true} \mid P_{recon}\right) &= \sum_s P_{true}(s) \ln \frac{P_{true}(s)}{P_{recon}(s)} \\ &= \sum_s P_{true}(s) \ln P_{recon}(s) + \sum_s P_{true}(s) \ln P_{true}(s) \\ &= -L(J,h) + \sum_s P(s) \ln P(s). \end{aligned} \tag{12}$$

$L(J,h)$ is the likelihood function which we will optimize to find our most probable model parameters [80].

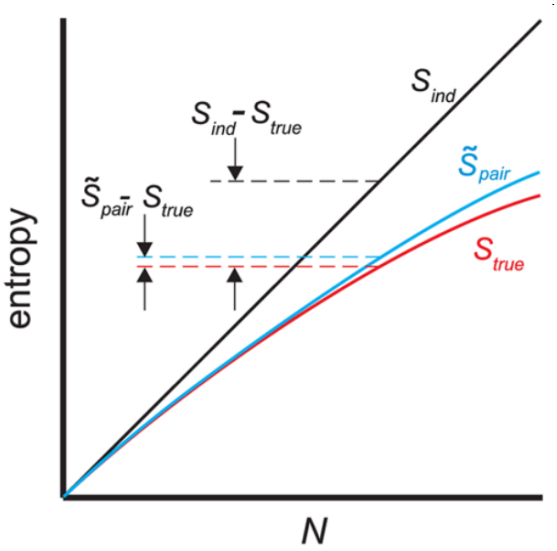


Figure 15
Schematic comparison plot of the entropy for an independent model of disconnected spins S_{ind} (black line), entropy of a pairwise model $\tilde{S}_{pair}$ (cyan line), and S_{true} with respect to size of the network N. The maximum entropy pairwise model is closer to the true distribution as $\tilde{S}_{pair}$ approaches S_{true}. This is shown by the normalized distance measure $\Delta_N = \frac{S_{\text{maxent}} - S_{\text{true}}}{S_{\text{ind}} - S_{\text{true}}}$ Adapted from Roudi et al. 2009 fig 3 [100].

2.3 Maximum Log-Likelihood

The log-likelihood function $L_{\hat{S}}(\theta)$ is the probability $P(\hat{S} \mid \theta)$ of the set of observed outputs $\hat{S}$ as a function of the parameters $\theta = (J_{ij}, h_i)$,

$$\begin{aligned} L_{\hat{S}}(\theta) &= \frac{1}{T} \ln P(\hat{S} \mid \theta) \\ &= \sum_{i<j} J_{ij} \frac{1}{T} \sum_t \sigma_i^t \sigma_j^t + \sum_i h_i \frac{1}{T} \sum_i \sigma_i^t - \ln Z(\theta) \\ &= \sum_{i<j} J_{ij} \left\langle \sigma_i \sigma_j \right\rangle^{\hat{S}} + \sum_i h_i \left\langle \sigma_i \right\rangle^{\hat{S}} - \ln Z(\theta). \end{aligned} \tag{13}$$

The log-likelihood only needs the first and second moments of interaction (magnetizations and pair correlations) as these are considered sufficient statistics to determine the model parameters. It becomes inconvenient to extend calculations beyond pairwise correlations as computational complexity increases, but it can still be done efficiently in some cases [98].

In order to maximize the likelihood, we calculate its derivatives with respect to,

$$\begin{aligned} \frac{\partial L_{\hat{S}}}{\partial h_i}(J,h) &= \left\langle \sigma_i \right\rangle_{\hat{S}} - \left\langle \sigma_i \right\rangle_{max}, \\ \frac{\partial L_{\hat{S}}}{\partial J_{ij}}(J,h) &= \left\langle \sigma_i \sigma_j \right\rangle_{\hat{S}} - \left\langle \sigma_i \sigma_j \right\rangle_{max}. \end{aligned} \tag{14}$$

This can now be set into a convex optimization algorithm to find the region in parameter space with maximum likelihood for the expectation values. An exact maximization approach is a Boltzmann learning gradient-descent algorithm,

$$\begin{aligned} h_i^{n+1} &= h_i^n + \eta \frac{\partial L_D}{\partial h_i} \left(J^n, h^n\right), \\ J_{ij}^{n+1} &= J_{ij}^n + \eta \frac{\partial L_D}{\partial J_{ij}} \left(J^n, h^n\right), \end{aligned} \tag{15}$$

where for some number of update steps n and the learning parameter η determines the step size of each iteration of the algorithm. This quickly runs into the problem of computational costs for the exact maximization. While expectation value calculations average over all spin configurations, the partition function must sum over the terms at each step, making exact maximization infeasible for networks larger than a few tens of nodes [101].

Approximate methods are used to sidestep these limitations. Sampling methods such as Monte Carlo methods are excellent options as they can provide an exact answer if given a sufficient amount of time. However, a "sufficiently long time" grows exponentially with the size of the network [98]. Alternative approximate approaches available are the mean-field equations and the Pseudo-Log-Likelihood.

2.4 Approximate Approaches

2.4.1 Naive Mean Field

The mean-field approach considers a simple approximation of an system by averaging over its general field of effect, reducing many the degrees of freedom in the system to a smaller set averaged variables. There are no interactions between the constituents of the system, just the combined average effect they exert. This absence of interactions is called the mean-field assumption [84]. The field which effects on a single spin arises from the local field h_i as well as the mean field from

all spins coupled to the node being affected. An average over these fields gives a "mean effective field" [80].

The simplest approximation is the naïve mean field (nMF) [113],

$$\tilde{h}_i = \tanh^{-1}(m_i) - \sum_j J_{ij}^{MF} m_j. \tag{16}$$

This is the derivative of the mean-field free energy with respect to the magnetization $m_i = \langle \sigma_i \rangle$. Likewise the second order derivative gives us the inverse susceptibility (i.e. inverse correlation) matrix,

$$\left(\chi^{-1}\right)_{ij} = -J_{ij}^{MF}, \tag{17}$$

for $i \neq j$ and $\chi_{ij} = \langle \sigma_i \rangle \langle \sigma_i \rangle - \langle \sigma_i \sigma_j \rangle$. If the magnetizations and pair correlations are known, then the coupling matrix J_{ij}^{MF} can be approximated and subsequently the bias h_i.

Roudi et al. 2009 [101] derives a nMF approximation for a system of independent spins which is shown to perform well in small model sizes. This technique is expanded on in the methods section, where it is used to test the forward Ising sampling implementation.

2.4.2 Thouless-Anderson-Palmer (TAP) Equations

The TAP equations [114] are an extension of the nMF [113], overcoming the limits of the nMF in approximation for large populations with high firing rates by adding the *Onsager term* which can be derived from the Plefka expansion among other approaches [88]. Essentially these take into account the second-hand effect a node has on itself through the energy it exerted on its neighbors. These are given as,

$$\tanh^{-1} m_i = h_i + \sum_{j \neq i} J_{ij} m_j - \sum_{j \neq i} J_{ij}^2 m_i \left(1 - m_j^2\right). \tag{18}$$

Differentiation with respect to m_j $(i \neq j)$ then gives,

$$\left(\chi^{-1}\right)_{ij} = -J_{ij} - 2 m_i m_j J_{ij}^2. \tag{19}$$

Solving this quadratic equation gives the TAP reconstruction,

$$J_{ij}^{TAP} = \frac{-2\left(\chi^{-1}\right)_{ij}}{1 + \sqrt{1 - 8\left(\chi^{-1}\right)_{ij} m_i m_j}}, \tag{20}$$

in the solution for the mean-field reconstruction when the magnetizations are zero. The magnetic fields can again be found by differentiating the Gibbs free energy.

$$h_i = \operatorname{artanh}(m_i) - \sum_{j \neq i} J_{ij}^{TAP} m_j + m_i \sum_{j \neq i} \left(J_{ij}^{TAP}\right)^2 \left(1 - m_j^2\right). \tag{21}$$

The TAP equations are shown to effectively reconstruct parameters as network volume increases. When applied to spike trains from populations of up to 200 neurons, the inversion of TAP equations was shown to give remarkably accurate results [98].

2.4.3 Pseudo Log-Likelihood

The *Pseudo Log-Likelihood* [13], which we will contract simply to *Pseudo-Likelihood* (PLH), is an alternative to the log-likelihood. The regular likelihood function becomes computationally

expensive as the partition function Z scales exponentially with the sum of 2^N terms and requires re-evaluation many times during the maximization of the likelihood. The pseudo-likelihood replaces the log-likelihood with a series of logistic regressions on the node variables [20], scaling polynomially with the size of the network N and number of samples T. This is still magnitudes more efficient and approaches an exact inference of the model parameters in the limit of infinite sample size [100].

The key feature of the PLH is it reduces dependency on model parameters by splitting the Hamiltonian energy function into two parts, with the first part dependent only on the immediate node and includes all couplings to spin σ_i, while the second part sums the energy over all other nodes, and excludes couplings with σ_i. The Hamiltonian becomes,

$$E_{pl}(\sigma_i) = E_i(\sigma_i) + E_{\setminus i}\left(\sigma_{\setminus \sigma_i}\right). \tag{22}$$

Given sufficient sampling size T, the average expectation values will match those of the standard LLH. The separation of these variables is possible because the statistical effect of σ_i on the other nodes $\vec{s}_{\setminus i}$ is given by the parameters (h_i, J_{ij}). We modify the partition function,

$$Z(J,h) = \sum_{\sigma \setminus \sigma_i} 2\cosh\left(h_i + \sum_{j \neq i} J_{ij}\sigma_j\right) e^{-E_{ji}(\sigma \setminus \sigma_i)}, \tag{23}$$

it now only sums over spin i reducing our computational complexity. Differentiating with respect to the parameters to yield our expectation values,

$$\begin{aligned} \langle \sigma_i \rangle &= \left\langle \tanh\left(h_i^{PL} + \sum_{k \neq i} J_{ik}^{PL}\sigma_k\right)\right\rangle, \\ \langle \sigma_i \sigma_j \rangle &= \left\langle \sigma_j \tanh\left(h_i^{PL} + \sum_{k \neq i} J_{ik}^{PL}\sigma_k\right)\right\rangle, \end{aligned} \tag{24}$$

These are the Callen identities [21]. While the expectation values on the right-hand-sides are an average over the spins except for σ_i, they approach exact values with sufficient sampling. Most importantly, the average over all 2^{N-1} states is replaced with an average over all configurations of the samples [80].

Substituting the average over all states for an average over data corresponds to a probability distribution which is a series of logistic regression models. Writing this new distribution function as a logistic regression where the probability of spin σ_i is conditional on all the other spins $\{\sigma_j\}_{j \neq i}$ is given as,

$$P(\sigma_i | \sigma_{j \neq i}\ , (J_{i*}, h_i)) = \frac{e^{\sigma_i(\sum_{j \neq i} J_{ij}\sigma_j\ +\ h_i)}}{2cosh(\sum_{j \neq i} J_{ij}\sigma_j + h_i)}. \tag{25}$$

We obtain the *normalized pseudo-likelihood* of a node in our system by taking the mean of the spin distributions over the space of observations t,

$$L_{PL}^{i}\left(J_{i*}, h_i\right) = \frac{1}{T}\sum_{t} \ln P\left(\sigma_i^t \mid \sigma_{j \neq i}^t\right). \tag{26}$$

In the limit of infinite samples, maximizing this function returns the parameter vector $\theta = (J_{ij}, h_i)$ of highest likelihood in relation to the node n_i. Extending this to every other node in our system n_j returns an asymmetric coupling matrix $J_{ij} \neq J_{ji}$ due to statistical variance when taking the likelihood of J_{ji}. This is compensated for by taking the average of the two values, $\frac{1}{2}\left(J_{ij} + J_{ji}\right)$, returning the inferred coupling matrix to an equilibrium state.

The PLH is shown to be further effective when paired with some regularization method in order to reduce bias. The most common regularization term being the ℓ_1 regularization which will be expanded on in the Section 3 [37].

Nguyen et al. 2017 [80], creates a PLH variant of the mean field and TAP equations by replacing expression in the Callen identities with their PLH counterparts, essentially replacing the local spin fields with their mean values. Their resulting equation for the PLH-Mean Field J,

$$J_{ik}^{\mathrm{PLH-MF}} = \left[1 - m_i^2\right] \sum_{j \neq i} \chi_{ij} \times \left[\left(\boldsymbol{\chi}_{\setminus i} \right)^{-1} \right]_{jk}, \tag{27}$$

where $\chi_{\setminus i}$ is the submatrix of the correlation matrix with row and column i removed. This can be expanded to the second order to obtain a TAP variant as well. In the methods section this is expanded on further and compared in effectiveness to other approximate methods for various network configurations.

3 Bayesian Model Selection

3.1 A Discrete Definition

In this section we must refine the definition of *model* to mean the network model $\mathcal{M}_i$ i.e. the graph of the network connectome. These models are indexed by the Ising parameter configuration θ which corresponds to a probability distribution $P(\hat{S}|\mathcal{M}_i)$. Inferring this model from a point in the space of outputs $\hat{S}$ for the distribution function, is a search in the space of probable models, or a *model neighborhood*. Like the energy phase-state space, this neighborhood is a visualization of a hyper-dimensional space, a *manifold*, to which some models or model families are local. This may be thought of as a continuous volume made of "points" in the space, each point a particular configuration of the model $\mathcal{M}_i$ indexed by the Ising parameters (J_{ij}, h_i) much in the same way a point or volume in three dimensional space is "indexed" by the coordinates (x, y, z). Figure 14 illustrates a model neighborhood of a network of two nodes.

The task of the inference problem, is *model selection*, the ranking and fitting the models within this space which best support the observed spike-train $\hat{S}$ for the probability distribution $P(\hat{S}|\mathcal{M}_i)$.

3.2 Bayesian Techniques

There are two layers to *Bayesian model selection*. The first is to assume a model is true and can be fit to the data, i.e. a direct inference of the parameters which best explain the given data $\hat{S}$ exists. The second is weighing the potential models by some method and ranking them by their ability to target the data distribution. There is no perfect model selection method. While more complex methods can better fit a certain set or sets of data, they are prone to *over-fitting* data to specific model families. Alternatively, a coarse-grain model selection approach may be able to fit more models, but often fail in recovering network detail. Regarding the probability of recovering the model $\mathcal{M}_i$ by the Bayes formula, the central question of this process is framed as such:

$P(\mathcal{M}_i|\hat{S})$, "What is the probability of finding the model of the network $\mathcal{M}_i$ given the observations $\hat{S}$?"

Expressed by Bayes formula,

$$P(\mathcal{M}_i|\hat{S}) = \frac{P(\hat{S}|\mathcal{M}_i)\, P(\mathcal{M}_i)}{\mathcal{Z}(\hat{S})}. \tag{28}$$

We define the terms,

- $P(\hat{S}|\mathcal{M}_i)$ is the ***likelihood*** and the point of focus to the likelihood functions built in section two on Ising inference. This is a data-dependent term, from which evidence is built for our manifold of probable models.
- $P(\mathcal{M}_i)$ is the ***prior*** probability of the model, or the probability of the model in the absence of the observation data.
- $\mathcal{Z}(\hat{S})$ is the ***evidence*** , or our normalization: $\mathcal{Z}(\hat{S}) = \sum_J P(\hat{S}|\mathcal{M})\, P(\mathcal{M})$. The probability on the space of all possible models given the observation. The evidence can be momentarily ignored while we build the space of probable models.

The posterior probability of each prospective model is:

$$P(\mathcal{M}_i|\hat{S}) \propto P(\hat{S}|\mathcal{M}_i)P(\mathcal{M}_i), \tag{29}$$

where the prior is assumed uniform over the model space and so attention is focused on the likelihood term $P(\hat{S}|\mathcal{M}_i)$.

When beginning with no information other than the observed data, the prior of a prospective model $\mathcal{M}_i$ is defined by the parameters θ as conditioned by the observed spins $\hat{S}$. To evaluate the likelihood, the likelihood function must be integrated over all parameter configurations which fit to the the constraints of the graph [87],

$$P(\hat{S}|\mathcal{M}_i) \propto P(\hat{S}|\theta, \mathcal{M}_i)P(\theta|\mathcal{M}_i). \tag{30}$$

The term $P(\hat{S}|\theta, \mathcal{M}_i)$ is again the likelihood term (in this paper it is the pseudo-likelihood function $L^*(\theta)$), and the prior $P(\theta|\mathcal{M}_i)$ become our *evidence*:

$$P(\hat{S}|\mathcal{M}_i) \ \propto \ \int d\theta \ e^{TL^*(\theta)} \ P(\theta|\mathcal{M}_i). \tag{31}$$

Given no prior information about the prospective model, an *uninformative prior* must be used. The simple solution is to treat all models indexed by the parameters as equally likely. That is, an unbiased probability distribution which assumes all prior parameters are just as equally as likely across the model manifold. This causes a significant problem as a uniform prior can assign wildly different probability masses to the same subset of parameters since two different parameter values can index very similar distributions [78].

3.3 Model Selection Criteria

We pair the prior with an *Occam factor* $P(\theta|\mathcal{M}_i)\alpha_{\theta|\hat{S}}$. This Occam factor is a measure of uncertainty on the data given, the ratio of the information accessible about the target model's parameter space, and the factor by which the model space is constrained once data is observed. An Occam factor will more strongly penalize a complex model with a high number of parameters and a high degree of possible models, opting instead for a simpler model, seeking a balance in model complexity while minimizing misfit [69].

The implementation of complexity penalization on the graphical model is *model selection criteria*, a complexity term which penalizes the likelihood based on the parameters which define the size of the space, which is not only the Ising parameters, but by network size, sample rate, or informational content [28]. We illustrate popular criteria solutions in the inset: *Selection Criteria*. Most of the criteria used here do not penalize the model based on the Ising parameters and instead use the uniform prior. However, in the MDL complexity terms we introduce below, an uninformed prior called the *Jeffery's prior* (the very last term under the integral in Equation 35) constricts the model space, based on the parameters found by the likelihood function.

Selection Criteria

ℓ_1 **- Regularization:** The simplest of the criteria and one which can be easily paired with the others, is the attachment of a regularization parameter λ which is typically set small [51], and allows elimination of the smallest, and presumably least significant, connections in the network [80].

$$\ell_1 = -L[\theta] + \lambda \sum_{ij} |J_{ij}| \,. \tag{32}$$

This has been shown to pair well with the PLH [75] [85] and optimization of the criteria in high-dimensional regimes (large network and large sample size) can return an exact recovery of initial network topology [94]. A standard way of optimizing the regularization parameter λ is by *cross-validating* against a part of the observation data originally withheld to determine the effectiveness of the criteria [69].

Two standard selection criteria are the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). Both act as complexity penalizations but with differing advantages. The *AIC-BIC dilemma* [30] refers to the problem where the trade off between these terms is one between predictive quality and consistency.

Akaike Information Criteria (AIC): The AIC is a criterion proposed by Akaike (1974) [2], and it approaches from an information theory perspective, it attempts to approximate the out-sample prediction loss by the sum of the in-sample prediction loss and a correction term [31]. Given a finite number of models the AIC tends to select the optimal one for prediction. However, it loses consistency within regimes of larger N and T [48].

$$\mathrm{AIC}_m = -2L(\theta) + 2N. \tag{33}$$

Bayesian Information Criterion (BIC): Relatively similar to the AIC however its strength lies in its consistency, penalizing models with a large amount of parameters ($N \gg 1$) [107]. The BIC selects the smallest model containing the target distribution [48]. The key difference from the AIC being that it scales the penalization factor logarithmically with the size of the sample data [31] [19].

$$\mathrm{BIC}_m = -2L(\theta) + N \log(T). \tag{34}$$

3.4 Minimum Description Length

The Minimum Description Length (MDL) principle acts as is an informational implementation of Occam's razor. The idea stems from algorithmic coding theory, and proposes the best model to describe some data is the one that encodes the data with the greatest compression of the data description. That is, if the probability distribution function takes some set of input parameters which encode a description of the output distribution, then the smallest set of input parameters which can encode that description is the most viable model [47]. While the AIC and BIC penalization factor scales with the network size and/or sample rate size, an MDL based approach prunes parameters unnecessary to the generation of the observed output space.

One model selection criterion conceived of this concept was a proposed MDL modification of Rissanen's stochastic complexity criterion titled the *Predictive MDL* (PDML) [97], which integrates into the rearranged Bayesian formula as:

$$\log P(\bar{\sigma}_i \mid \bar{\sigma}_j, \theta) = T\ell\,(\theta^*) - \frac{n^*}{2} \log \frac{T}{2\pi} - \log \int d\theta \sqrt{\det F(\theta)}. \tag{35}$$

The two new terms which make up this criterion are referred to as the *geometric complexity* [78],

$$C_{Geometric} = -\frac{n^*}{2} \log \frac{T}{2\pi} - \log \int d\theta \sqrt{\det F(\theta)}. \tag{36}$$

The first term, coincidental to the BIC, increases logarithmically with the sample size T while the latter term is independent of T. Meaning, the effects of the latter term diminish as sample

size grows because the Fischer Information matrix $F(\theta)$, which acts as metric of distance in the distribution space of the Riemann manifold [78], will gradually decrease in impact respectively to the number of non-zero parameters n^*. This effectively reduces the whole criterion to a measure equivalent to the BIC.

The Fischer Information matrix is the matrix of expectation values for the Hessian matrix of the likelihood $H_{i,j}(\theta) = -\partial^2_{\theta_i,\theta_j} L^*(\theta)$ with respect to our model distribution $P(\bar{\sigma_i} \mid \bar{\sigma_j}, \theta)$, such that:

$$F_{i,j}(\theta) = -\mathcal{E}\left[\ H_{i,j}\ \right], \tag{37}$$

$$F_{i,j}(\theta)\ = -\sum_{\theta} P(\bar{\sigma_i} \mid \bar{\sigma_j}, \theta)\ \left(\partial^2_{\theta_i,\theta_j} L^*(\theta)\right). \tag{38}$$

However the Hessian is not dependent on the probability of σ_i, so the Fischer information matrix is the same as the Hessian.

The penalty terms in the PMDL is the *intrinsic complexity* of our target family of models. Rissanen showed that as the network size increases, the PMDL is the length in bits of the shortest possible code describing the output generated by a target model family. This suggests the model parameter configuration which best minimizes the PMDL (and thus maximizes the probability) gives the parameters which generalize best.

3.4.1 The Bulso et al. 2019 MDL Criterion

Using the PMDL as a basis, Bulso et al. (2019) [19] focused on the latter term of the *Geometric Complexity*:

$$C_{GC} = \log \int d\theta \sqrt{\det F(\theta)}. \tag{39}$$

In logistic regression models, such as the PLH, the elements of the Fisher Information matrix can be expressed,

$$F_{i,j}(\theta) = \sum_{\mu} \nu\left(\vec{s}^{\,\mu}\right) \cosh^{-2}\left(\theta \cdot \vec{s}^{\,\mu}\right) \vec{s}_i^{\,\mu} \vec{s}_j^{\,\mu}, \tag{40}$$

where $\nu\left(\vec{s}^{\,\mu}\right)$ is the frequency of observing a unique spike-word configuration $\vec{s}^{\,\mu}$ in the data, with the size of the spike-word "dictionary" being $\mu = 1, \ldots, 2^n$. Deriving the lower and upper theoretical boundaries on the latter term of the geometric complexity, Bulso et al. 2019 proposed the novel MDL-entropy (MDLent) based criterion,

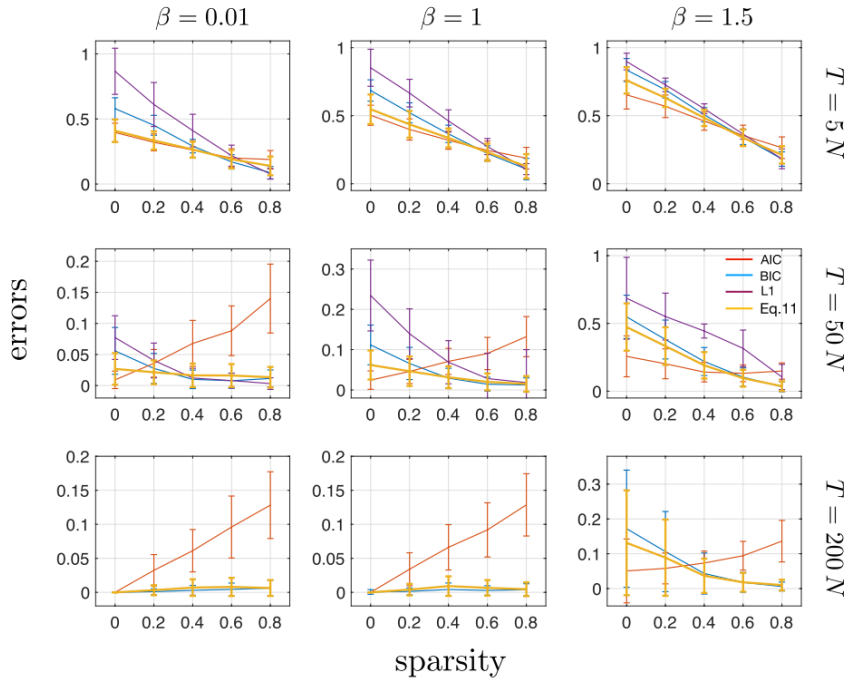
$$C_{Bulso} = -\frac{n^*}{2} - \frac{n^*}{2} \log\left(\frac{T S_{n^*}}{n^* S_N}\right) + \log n^*. \tag{41}$$

Here n^* is the number of non-zero parameters of the parameter vector θ for the model, while N stays the total number of nodes. The term $\mathcal{S}$ is the Shannon bitwise entropy of our of our spike-word frequency distribution (for all spike-words, $\vec{s}^{\,\forall}$)

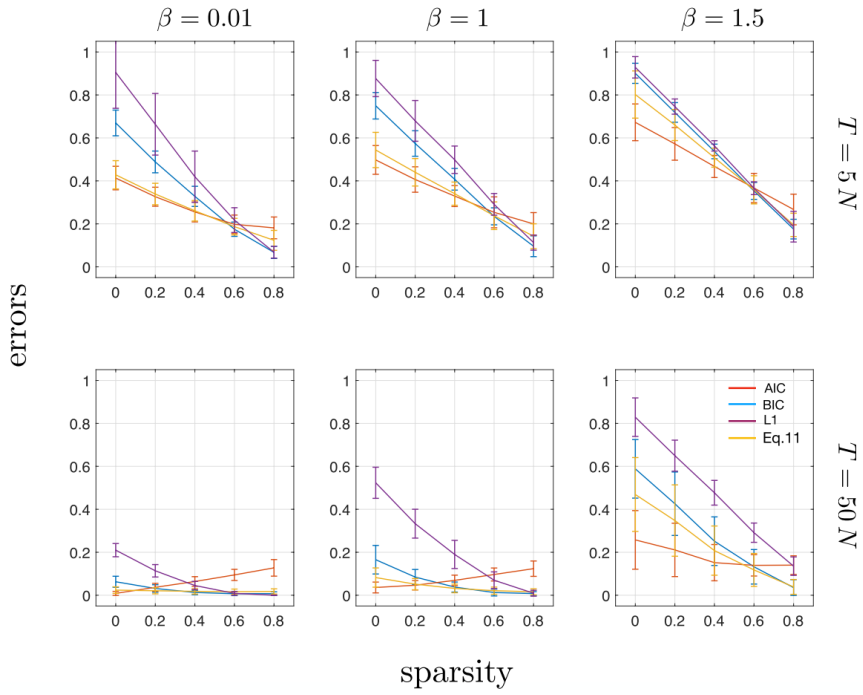
$$\mathcal{S}_{n^*}(\vec{\boldsymbol{s}}^{\,\forall}) = -\sum_{\boldsymbol{\mu}} \nu(\vec{\boldsymbol{s}}^{\,\boldsymbol{\mu}}) \log_2 \nu(\vec{\boldsymbol{s}}^{\,\boldsymbol{\mu}}). \tag{42}$$

Likewise, $\mathcal{S}_N$ is the entropy of the full nodal set. The criterion scales with the entropy distribution to localize on the model distributions capable of producing the observed spike-word frequencies. In practicality, this term trends towards an AIC-like penalty term in fully-connected graphs with low observed samples $n \approx T \approx N$ and a BIC-like term in sparse networks : $C \to \frac{n^*}{2} \log(T)$ as $\mathcal{S}_{n^*} \to n^*$.

In comparison to other criteria, the novel MDLent criterion showed a general BIC-like trend, however was also able to match the AIC reconstruction rate in sparse networks, where the BIC method tends to show weakness (Figure 16). This was observed across two large sized networks, $N = 50, 100$.



(a) $N = 50$



(b) $N = 100$

Figure 16
Subplots of the mean misclassification error of the different criteria including the novel MDLent term, versus levels of network sparsity for a network of size $N = 50$; error bars represent the mean standard deviation. The misclassification error was averaged from 100 sample trials. The superplot columns represent the spin-glass model beta values β used to adjust network couplings strengths; the rows are variations of sample rate sizes T taken with respect to the network size. The performance of the selection terms AIC (red), BIC (blue), ℓ_1 regularization (violet, and only for the first two values of T), and the novel MDLent criterion (yellow). All inferences were done using a logistic regression based likelihood method (akin to the PLH). Adapted from Bulso et al. 2019 [19].

4 Methods

4.1 Network Regimes and Glauber Dynamics

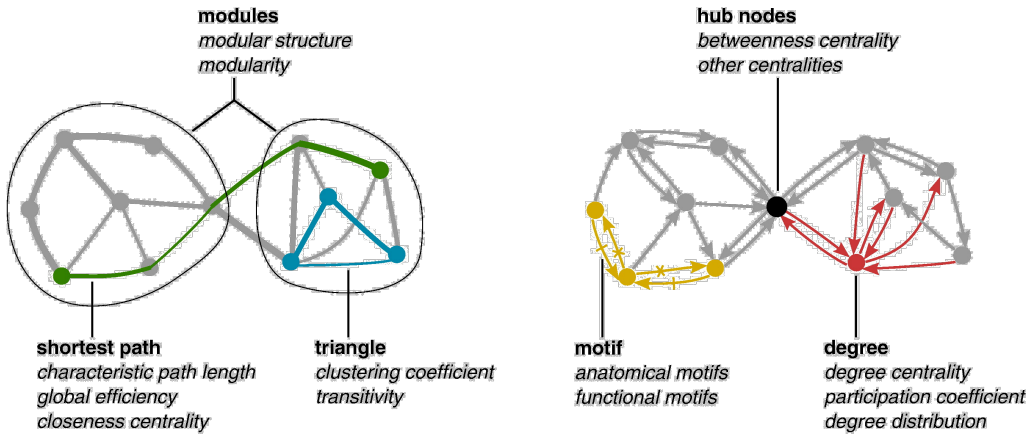


Figure 17
Illustration of different measures of network topology. Here we mainly refer to the *degree of connectivity* or *coordination number*: the average number of connections each node in the network has. We will also refer to the concepts of *hub nodes*: nodes which lie at the intersection of multiple *paths of shortest lengths* both of which are used describe the integration and segregation of a network connectome. Adapted from Rubinov et al. 2010 [103].

4.1.1 Model Topologies

To generate a ground truth for our inference methods, we needed to first create a forward Ising implementation, allowing control over the experimental conditions. This required us first making a selection of the topologies for the intended connectome structures.

We start with a symmetric $N \times N$ *adjacency matrix* where the entries of 0 and 1 define the presence of an *edges* K between the nodes n. Depending on the graph structure we want, we define the probability distribution of edges for the nodes. For example, the *Random Graph* or *Erdős–Rényi* [36] assigns the edges randomly to each node with a weighted probability. In our implementation, the probability of an edge between two nodes $P(K) = \frac{\mathcal{C}}{N-1}$ where our *coordination number* $\mathcal{C}$ is the average number of edges per node (or *degree*, Figure 17) in the graph (pre-selected as a density measure) and $N-1$ the total number of other nodes that can be connected to (no self-connections) which keeps the network at some level of sparsity (and not fully-connected) as long as $\mathcal{C} \neq (N-1)$ [36] [63].

Tested Connectome Topologies The following connectome topologies were used: Cayley tree (CT), Erdős–Rényi (ER), & small world (SW). Topology descriptions can be found in the inset: *Topologies*. These were chosen for their scaling levels of network structure and trade-offs between rigidness and randomness. The Cayley Tree with its fixed structure and consistent node degree provides a baseline metric for the inference method, as it consistently proves to be the easiest topology to infer among the methods tried here. Opposite to this is the random graph with an entirely probabilistic structure and node degree distribution which tests the generalizability, or an inference method's tendency to overfit to a single structure type. The Watts-Strogatz SW topology overlaps features between both, with the model keeping a fixed node degree while also maintaining an aspect of variability and change within the network structure.

Topologies

The last section covered connectivity types we wish to infer. Here we apply graph theory to describe the underlying network topologies. Multiple graphical model types exist, and in the neuro-anatomical context graphs are used to describe the connectome of neuronal structures. The Ising pairwise model is a weighted undirected network, the connectome *model* $\mathcal{M}$ is an undirected network. For testing we will use several established graph topologies. The most straight forward example we've just described, is the *Erdős–Rényi* or *random graph*, as it has no particular structure to its topology. However, real world networks and self-organized assemblies are not randomly arranged but instead have ordered and hierarchical structures.

Cayley Tree: A Cayley tree [23] is a simple undirected graph with a "tree like" structure and a consistent number of branches $\mathcal{C}$ at every node with no closed loops, or *cycles* [12]. It is recursively constructed by designating a "seed node" as the zeroth generation *hub* of the lattice that "branches out" (creates unique edges) to $\mathcal{C}$ number of new nodes. This "first generation" of hub nodes in turn branches out to another $\mathcal{C} - 1$ nodes for any specified number of generations (Figure 18). These are useful in inference problems as they have exact solutions in the Ising model via the *Bethe-Peierls approximation* [82] [34]. Tree structures provide an important baseline for testing model reconstructing as they avoid many of the problems associated with complex structures (loops, density) while also providing insight into dependencies of the parameter distribution at reduced computational cost [28].

Small World Networks: Small world networks are networks that are more clustered than random networks, yet the average *path length* (number of hops between any two nodes) is similar to those in random networks. The seminal example being real-life social networks [74] [4] where indirect relationships between people often follows paths which cluster around "hub nodes" (Figure 19). Plenty of other examples and variations exist [52]. These network topologies combine features found in segregated network modules of specialized functionality, into a larger, sparser, cross-connected network of such modules. The idealized version features high amounts of both segregation and integration [103].

Many naturally occurring self-organized networks are of this variety. In neuroscience we find many examples of highly self-connected neuronal modules which make local and long distance communications in the greater neural network [67] [122].

Our tests use a simple *Watts-Strogatz small world graph model* [121], which organizes itself first as a *ring model* where each node connects to $\mathcal{C}$ of its nearest neighbors. Each edge in the graph then rewires to a random node with the probability $P(K)$, excluding duplicate edges or self-connections. The graph begins as a ring lattice, so when $P(K) = 0$ there are no rewires and the graph stays a ring lattice, when $P(K) = 1$, every edge rewires and the topology is a random graph. As with the random graph, we set the probability to scale with the size of the network and the selected coordination number $P(K) = \frac{\mathcal{C}}{N-1}$.

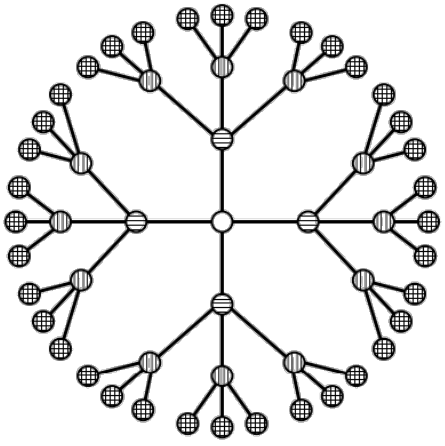


Figure 18
Recursive Bethe lattice for coordination number of $\mathcal{C} = 4$. The Bethe lattice is an infinite graph where each node has the same number of edges and there is only a single path between any two nodes. The Cayley tree is a finite portion of the Bethe lattice. Adapted from Eckstein et al. 2005 [33].

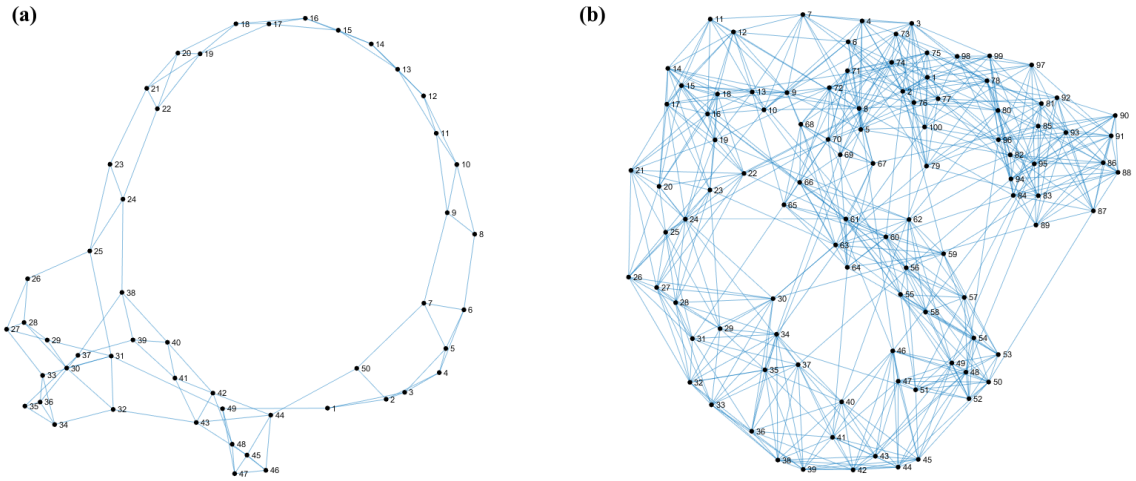


Figure 19
Nodal plot of a Watts-Strogatz small world network. On the left is a sparse small world network ($N = 50,\ \mathcal{C} = 4$) with low probability of edge reconnection, making for a network with few hub nodes. On the right is a larger, denser, small world network ($N = 100,\ \mathcal{C} = 12$) with a slightly higher chance of reconnections, increasing the number of hubs in the network.

4.1.2 Connection Strength Distributions

The value distribution on the parameter connection strengths J and bias h, is the next key concern while building a synthetic network. Similarly to the Sherrington-Kirkpatrick (SK) spin-glass model, we adopt the *inverse temperature* β [83] which adjusts the localization of the phase space distribution over the manifold, by controlling the average strength of pairwise connections [19] [75]. Changing this parameters diversifies the distributions of data in the observation samples.

We normalize β to scale with our network density such that $\frac{\beta}{\sqrt{\mathcal{C}}}$ where $\mathcal{C}$ is the coordination number as defined above and β is the initial inverse temperature setting. In this way we are able to reliably compare results from networks of the same β but different densities.

The distributions can also be set differently between the two Ising parameters. For instance, in Nguyen et al. 2017, their Ising inference tests used a fully connected model and took connection strengths J_{ij} from a Gaussian distribution centered at 0 with a standard deviation of $\frac{\beta}{\sqrt{N}}$ [80] (the coordination number in a fully connected SK-model is the size of the network N); while the bias strength h_i was drawn uniformly from the interval $[-0.3\beta,\ +0.3\beta]$.

While the distribution of our experimental values for h initially followed the same distribution used for our coupling strength distribution in a weakly connected model, we use a disconnected external field $h = 0$ in interest of simplicity.

Tested Ising Distributions For the purposes of building and testing the experiment we used a simple *Double Delta distribution*, with a split mean at $\mu_o = \pm 1$, making the connection strengths $J_{ij} = \pm\frac{\beta}{\sqrt{\mathcal{C}}}$ where $\frac{1}{\sqrt{\mathcal{C}}}$ acts as a normalization factor for the density of the network. We chose to normalize by the coordination number $\mathcal{C}$ instead of the network size N as our networks are generally not fully connected. The biological analog to the positive and negative connection strengths would be excitatory and inhibitory neurons (more specifically excitatory and inhibitory *connections* between neurons) respectively. Future variation in these experiments may adjust the total number of negative or positive connections to emulate excitatory/inhibitory networks.

In our final iteration of the experiment we opted for a *split-mean normal distribution* (Figure 20) where the connection strength has a split mean with a Gaussian distribution around both means and a standard deviation σ set to scale with network by the normalization factor $\sigma_o = \frac{1}{\sqrt{\mathcal{C}}}$. Thus,

allowing for a more even and varying distribution of the connection strengths scaling along with networks sizes and regimes.

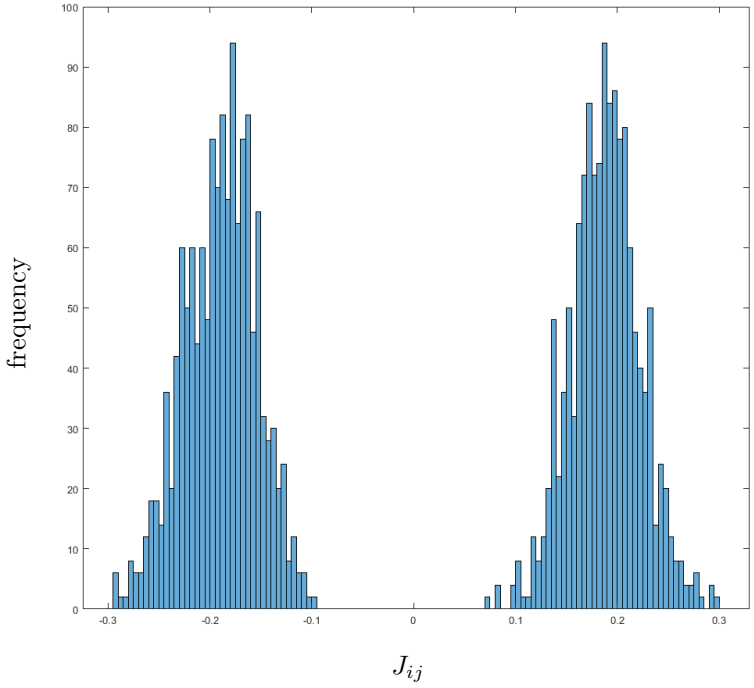


Figure 20
Histogram of connection strength distributions J_{ij} for a large, sparse, small world network with mid-range beta ($N = 100,\ \mathcal{C} = 28,\ \beta = 1$).

4.1.3 Network Regimes

We define the *regime* here as the initial network and experiment conditions. We split this into two general regimes, the *small network* (small N) and *large network* (large N), over which the other conditions are varied. The small N regime consists of some few tens of neurons (Table 1). In this regime we choose to use only small sets of observations sizes ($T = 10N, 15N, 20N$) or the *low-rate* regime. The large N regime (Table 2) has larger networks $N \geq 50$ along with *high-rate* sample sets ($T = 100N, 200N$). The network density is also tracked by the coordination number in the network structure ($\mathcal{C} = 2, 4, 8, 12, 28$) and use it as our point of reference for network density or sparsity.

Small Network Regime	**Parameters**
Coordination Number:	$\mathcal{C} = 2, 4, 8$
Network Size:	N = 10, 15, 20
Distribution Localization:	$\beta = 0.3, 0.7, 1, 1.3, 1.6$
Observation Samples:	T = 10N, 15N, 20N
Trials:	100

Table 1: Small Network Regime Table

Large Network Regime	**Parameters**
Coordination Number:	$\mathcal{C} = 8, 12, 28, 50, 70, 90$
Network Size:	N = 50, 80, 100
Distribution Localization:	$\beta = 0.3, 0.7, 1, 1.3, 1.6$
Observation Samples:	T = 10N, 30N, 50N, 100N, 200N
Trials:	100

Table 2: Large Network Regime Table

4.1.4 Metropolis Hastings Algorithm

For each set of experiment conditions a Gibbs sampling implementation was used simulate the Ising Glauber dynamics and produce a a binary Markov chain. i.e. a synthetic spike-train. The exact Gibbs sampling process which directly replicates the Ising Glauber dynamics as described in Section two. Each update in the system state is calculated from the probability $P(\sigma_i^{t+1}|\sigma_i^t)$ and the probability of the spin state on a node $\sigma_i = \pm 1$ is taken at each new time step $t + 1$.

This exact method is computationally very expensive and slow. A more effective implementation is probabilistic sampling, i.e. Markov Chain Monte Carlo (MCMC) methods. These are guaranteed to produce samples for a target probability distribution, but they must first "burn-in", a period where the Markov process runs long enough to converge at the equilibrium point of the target distribution [69]. This time required to reach the burn-in state becomes a major drawback.

We can append a learning rule to a probabilistic variant of the MCMC algorithm known as a *Metropolis Hastings MC* (MHMC). The algorithm probabilistically samples the Markov process and forces the Markov chain to converge to an equilibrium point of minimal energy. This begins by calculating the total energy of a system in its state by evaluating the energy function:

$$E = -\frac{1}{2}\sum_{i<j} J_{ij}\sigma_i\sigma_j + \sum_i h_i\sigma_i, \tag{43}$$

where a factor of half has been applied to account for the spins being double counted. The MHMC algorithm then minimizes this value by an *adaptive rejection* sampling process [43].

At each time-step in the chain, the algorithm proposes a new system state, selecting a random node in the initial system and flipping its spin to the opposite value. The algorithm then compares the difference in magnitude of the initial energy state versus the energy of the proposed system. If the proposed system state has an energy lower than the previous $\Delta E \leq 0$, it is automatically accepted as the next state in the Markov chain. However, if the magnitude is greater than that of the initial state ΔE the proposed system state is auto-rejected and the previous spin state remains the current one. There is included, however, a probability the new state will be accepted regardless $P = \frac{1}{e^{-\Delta E}}$ [72] [73] (Figure 21).

This process is allowed to run for a large number of steps (in our simulations we use $T \times N \times 10$ steps) at which point the MHMC algorithm will converge at the desired equilibrium [92] [69]. This process is then continued for another $T \times N \times 10$ steps, where every tenth system state is sampled as part of the simulated spike-train $\hat{S}$.

MH Algorithm

- 1: Compute E^t for current system state $\bar{s}^t$
- 2: Select random node i in our system and change the spin (*i.e.* : -1 $\rightarrow$ +1 , +1 $\rightarrow$ -1). This is our *proposal state* $t + 1$
- 3: Take difference of both state energies: $\Delta E = E^t - E^{t+1}$
  If: $\Delta E < 0$
  Or if: $u <$ exp$[-\Delta E] >$ Where u is a uniform random value between $(0, 1)$
  Then, update the proposal state to be the current state: $\bar{s}^{t+1} \rightarrow \bar{s}^t$
- 4: *Else If:* $\Delta E \geq 0$ Then keep current state $\bar{s}^t$

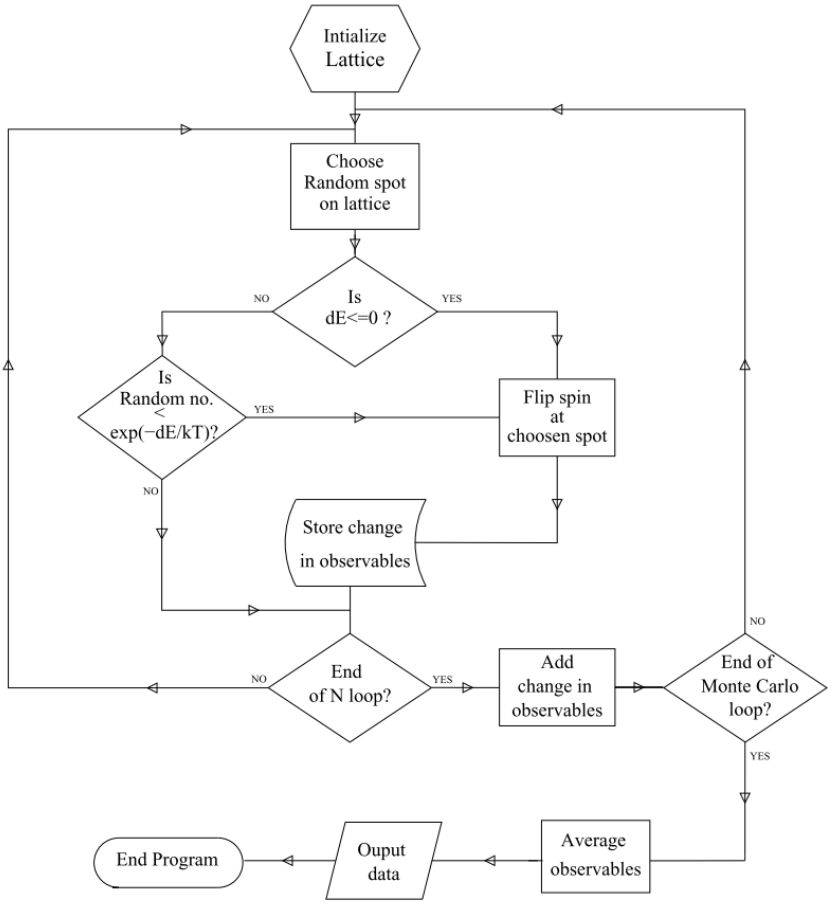


Figure 21
Flowchart of the Metropolis-Hasting algorithm. Adapted from Kotze 2008 [58].

4.2 Inverse Ising with Approximate methods

4.2.1 Sanity Checks

We ensure the forward Ising implementation converges to the target distribution by comparing the expectation values of the output Markov chain against those generated by mean field approximate methods described in section two. Certain derivations of the mean field equations can be incredibly accurate even in the *low-rate* regimes [102]. Approximating the expectation values from the initial Ising parameters with well established solutions provides us a basis of comparison. We use the mean field approximations for the *Independent-Pair* and *Field of Disconnected Spins*, (Figure 22).

Roudi et al. 2009 [102] makes an approximation for a network of independent-pairs. In this topology every node i is connected to a single other node j. The coupling strength J_{ij}^{pair} for all spin configurations of the two neuron system is approximated,

$$J_{ij}^{\text{pair}} = \frac{1}{4} \ln \left[\frac{(1 + m_i + m_j + \tilde{\chi}_{ij})\,(1 - m_i - m_j + \tilde{\chi}_{ij})}{(1 - m_i + m_j - \tilde{\chi}_{ij})\,(1 + m_i - m_j - \tilde{\chi}_{ij})} \right], \tag{44}$$

where $\tilde{\chi}_{ij} = \chi_{ij} + m_i m_j$.

We first produce a spike-train using the forward implementation for a network of independent pairs J_{ij}^{pair} . We then calculate the expectation values from our generated outputs. If our Glauber dynamics are correctly simulated, then we should expect these generated expectation values to correlate with the expectation values calculated from our approximate solutions.

Inserting the coupling strengths J_{ij}^{pair} into equation Equation 44 and letting $m_i = 0$ (where external field $h_i = 0$ removes dependencies on the magnetization $m_i \equiv tanh(h_i)$) then solving to produce our estimated pair covariance $\chi_{ij} = \tanh(J_{ij})$ [95].

We likewise produce a spike-train for the configuration of disconnected spins, where nodes are disconnected $J_{ij}\text{pair} = 0$ and set the external field to some level of influence e.g. $h_i = \beta$. Here we can simply use the TAP equations (Equation 21) with any dependencies on J_{ij} being removed. The estimated magnetizations become $m_i \equiv tanh(h_i)$.

Expectation values:
Magnetization : $m_i \equiv \langle \sigma_i \rangle$
Pair Correlation : $c_{ij} \equiv \langle \sigma_i \sigma_j \rangle$
Pair Covariance : $\chi_{ij} \equiv c_{ij} - m_i m_j$

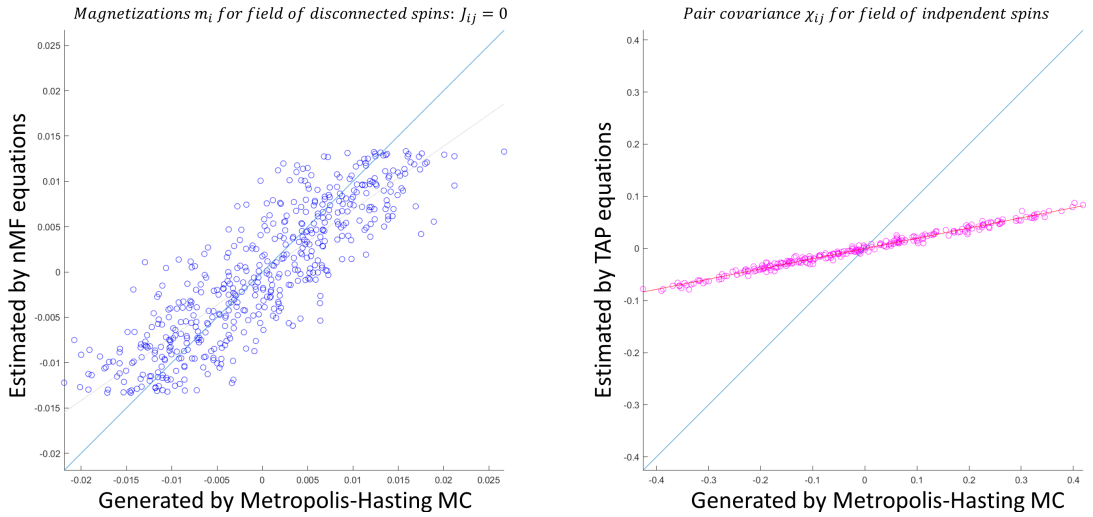


Figure 22
Comparing expectation values: **(Left)** Comparison of the magnetization values of a disconnected coupling matrix with a Gaussian distribution on the bias field h_i. Estimated by the Mean Field $m_i \equiv tanh(h_i)$ on the x-axis and the values generated by the adaptive rejection Sampling technique along the y-axis. **(Right)** Comparison of the pair covariance values for a system of independent spins (Gaussian distribution and no bias). Estimated values given by the TAP approximation $\chi_{ij} = \tanh(J_{ij})$ along the x-axis, and simulated data along the y-axis. Both distributions are taken for a system of $N = 500$ & $\beta = 1$ where T = 100N.

4.2.2 Inverse Ising of a Gaussian Distribution

After having verified the implementation of the Metropolis-Hasting algorithm matches the estimated results in the simple cases (as described above), we tested the approximate inverse Ising methods. Using a similar method to Nguyen et al. 2017 [80], we compared the approximate Ising methods across differing coupling strength parameters, network sizes, and spike train sample sizes (Figure 23 and accompanying inset).

We see the nMF perform most favorably in networks where $\beta \sim 0.5 - 0.9$ and with large network and sample sizes. The TAP equations largely followed the same trends as the nMF, particularly in smaller networks, but diverges slightly from the nMF with marginally better reconstruction scores at $\beta = 0.9$.

The PL-MF has the most unique reconstruction performance of of the methods tested. Its strength appears to lie in consistency at lower sample sizes even in smaller networks, perhaps performing best at small network size and sample sizes. However, it shows weakness in networks of higher connectivity strength. Figure 23 shows it dropping off in accuracy at a lower beta value in comparison to the other methods, reaching a point of diminishing returns around $\beta \approx 0.4$ and increasing in error afterwards. However, its climb in inaccuracy trends differently, plateauing in error as the beta increases, whereas the others rise exponentially in error after reaching their optimal β. The PLH will also tend to overestimate a connection strength when in error, whereas the other methods tend to underestimate.

As expected the strongest inference method was the PLH under Boltzmann learning, showing better reconstruction error across all parameters but particularly improving its error rate as T increases. While its trend initially follows the same one as the nMF and TAP, it noticeably performs better in all network configurations and sample sizes for $\beta \geq 0.6$ but below $\beta \leq 1.8$.

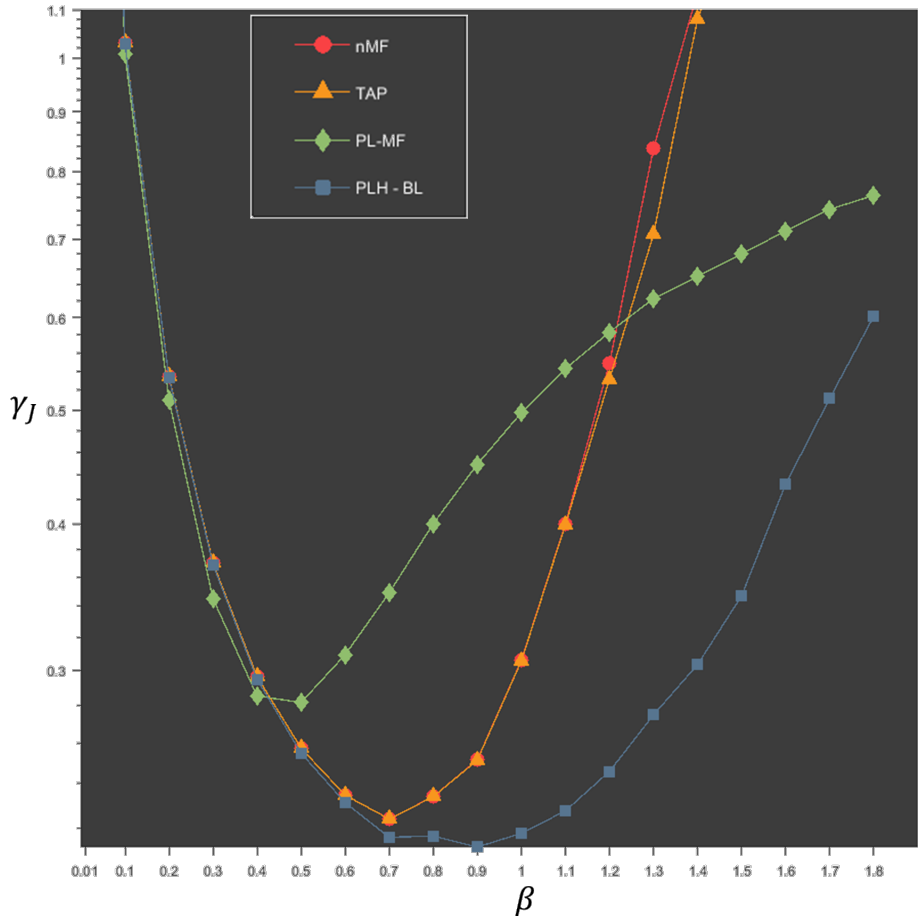


Figure 23
Replication of figure 5 from Nguyen et al. 2017 [80]. Here we have compared the Inverse Ising methods which we have tested across multiple values of β and have plotted the mean Relative Reconstruction Error γ_J from 100 realizations of the network. Our version has added the PL-MF variant and the PLH with a Boltzmann learning algorithm (PLH-BL in legend). Our results followed trends similar to the ones found in Nguyen et al 2017. See companion inset for more information.

Companion to Figure 23:
The inverse Ising methods: nMF, TAP, PL-MF, and the exact PLH inference, are used to infer the parameters for an Ising model in a fully-connected graph of no bias $h_i = 0$ and coupling strengths J_{ij} drawn from a Gaussian distribution with a mean centered at zero $\mu_0 = 0$ and standard deviation $\sigma_0 = \frac{\beta}{\sqrt{N}}$ for the parameters:

β	T	N
0.01, ... ,1.8	100,000	150

Full tests (not shown) also included regimes of $T = [1,000, 10,000]$ & $N = [50, 100]$. We score the quality of the reconstructions using the relative reconstruction error (RRE):

$$\gamma_J = \sqrt{\frac{\sum_{i<j} \left(J^*_{ij} - J^0_{ij}\right)^2}{\sum_{i<j} \left(J^0_{ij}\right)^2}}, \tag{45}$$

where J^*_{ij} is the reconstructed coupling matrix and J^0_{ij} is the original matrix.

4.3 Model Selection Tests

4.3.1 Replicating Bulso et al. 2019

Our intent is to replicate the model selection tests done in Bulso et al. 2019 [19] comparing the MDLent novel model selection criteria against other criteria. Additionally we would like to expand the ground truth network model and Ising parameter distributions to a wider array of network topologies and conditions to be inferred. Bulso et al. 2019 utilized a random graph with a double delta distribution of Ising model parameters where $J_{ij} = \frac{\beta}{N}$ and $h_i = 0.1$ for all parameters (and a zero diagonal $J_{ii} = 0$) and normalized by $\frac{1}{\sqrt{C}}$ for networks of size $N = 50, 100$ at $T = 5N, 50N, 200N$ at various levels of sparsity and β. We expand our tests to the regimes shown above, with the key difference of adding the two additional topologies and the split mean distribution.

We focus on the performance of the BIC and novel MDLent criteria in our results, as their reconstruction error rates were less distinct from each other in the 2019 paper and we are interested to observe any divergences in performance across any of the additional experimental conditions.

Our implementation of model selection criteria uses the PLH to calculate the likelihood and connection strengths J_{ij} as inferred from the simulated Glauber dynamics of our Metropolis-Hastings algorithm implementation. The fully connected model of each node's connections is then recurrently decimated (see inset: *Optimal Brain Damage*) to a fully disconnected model; the criteria then make selections from the decimated set of models. The reconstructed matrix is then compared to the ground truth adjacency matrix and scored.

Optimal Brain Damage
For larger network sizes it becomes increasingly difficult to appraise and rank all possible model configurations. A common solution to the "search and evaluate" problem is to apply a decimation technique [64]. This technique creates a configuration of the nodal connections by starting with a fully-connected topology and "pruning" connections, thus we can "walk" the model space by selecting a model configuration, evaluating it, then selecting the next model to be assessed. This can be done by a "random walk", via random connection decimation, or even in reverse, starting with a disconnected graph and step-wise adding connections instead. We use the method from Decelle et al. 2014 [29] where connections with the lowest Ising connection strengths are recurrently decimated until the matrix is reduced to a disconnected graph. This implementation of the random walk decimates a single connection per step in the walk, but the decimation can be done by any fraction of the total connections, which may be a more preferable method in larger networks. After recursively decimating $N - 1$ models, with our last model of $n^* = 1$, we may then choose the best model from the models indexed based on criteria score.

Another possible method, and one that may be relevant to the question of symmetric connections, is a model walk process used in Pensar et al. 2017 [87] which applied an inclusive OR-gate postprocess to their first model reconstruction, considering any of these recovered connections as part of the candidate set of possible edges before reapplying the selection walk to the model subspace.

4.3.2 Symmetrizing the Reconstructed Graph

Because the form of our model is one of pairwise interaction, the initial undirected graph is symmetrical $J_{ij} = J_{ji}$, however the reconstructed graph returned by the selection methods is largely asymmetrical. We want to consider the selection methods' performance in respect to reconstructing a graph of pairwise connections and so we applied a layer of postprocessing which would symmetricize the reconstructed graph. There are two options to this step: applying a "generous" inclusive OR-gated function repairing any asymmetries $K_{ij} = K_{ji} = 1$, or a "conservative" AND-gated function pruning asymmetrical connections so that $K_{ij} = K_{ji} = 0$. This showed interesting properties in early experiments, sometimes providing additional accuracy to reconstruction error in certain Ising distributions and showing a certain consistency across conditions (Figure 24). Unfortunately

it could not be fully implemented due to time constraints. See appendix C for more figures.

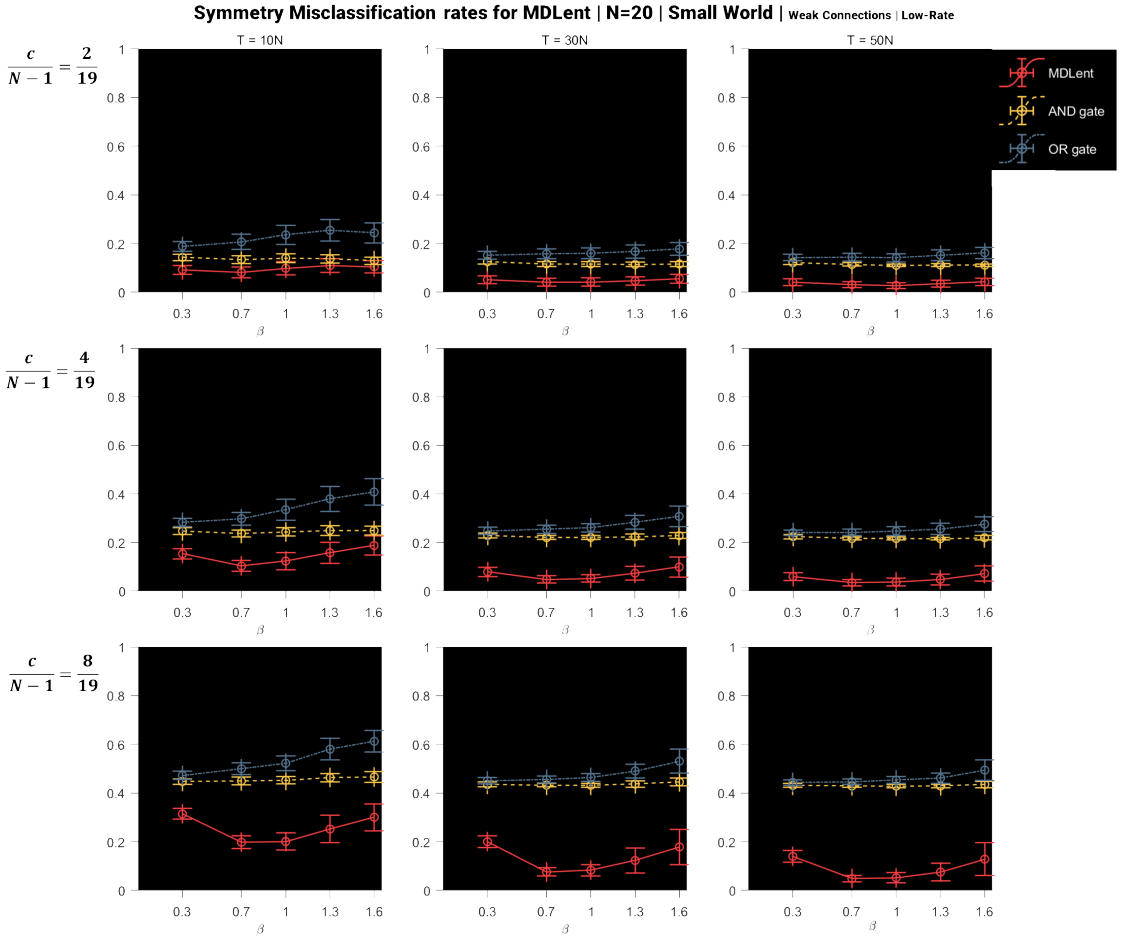


Figure 24
Sample analysis of the symmetry gating feature for the novel MDLent criteria. Here we can see the misclassification error (hamming distance) of the novel MDLent criteria and the symmetry gated variations of its reconstructed adjacency matrix. This was done for the small world, low-rate dataset. The individual subplots show the mean misclassification error taken from over 100 realizations of the network, over the beta values, error bars show mean standard deviation of the error rate. The greater plot compares observation sample rate and the density by the coordination number. We can see a jump in error with the symmetry gated methods, with a slight advantage going to the exclusive-AND gated method, especially in lower observation sample sets, though this may be due to the overwhelming sparsity of the original connectome. Other than that, they largely follow a similar trend to the original criteria.

4.3.3 Reconstruction Scoring

We tracked performance of the model criteria by three metrics derived from the *false positives (FP)*, *false negatives (FN)*, *true positives (TP)*, and *true negatives (TN)*. All metrics were taken by averaging the scores returned from 100 realizations of the network. The standard deviation of the trials was used to measure mean accuracy.

The misclassification error rate (or Hamming distance) is the sum difference between the original adjacency graph of the network and the reconstructed connectivity graph, i.e. the percent of connections which were incorrectly inferred, by dividing the net sum of the erroneously inferred connections by the size of the adjacency matrix (minus the diagonal of non-interacting elements):

$$Misclassification\ Error\ Rate = \frac{FP + FN}{N^2 - N} \ . \tag{46}$$

The Receiver Operating Characteristic (ROC) plots the True Positive Rate (TPR):

$$TPR = \frac{TP}{TP + FN}, \tag{47}$$

i.e. the ratio of correctly inferred connections out of the set of ground truth connections. And the False Positive Rate (FPR):

$$FPR = \frac{FP}{FP + TN}, \tag{48}$$

i.e. the number of incorrectly inferred connections out of the set of sparse connections.

Lastly, we consider the false-positive and false-negative occurrence scores. These are the fraction of false-positive and false-negative errors returned by the model selection criteria results.

4.4 Implementing our methods

Our simulation and inference methods were coded and executed in a MATLAB2020a/b environment. A sample of the top-level script can be found in the appendices and a github repository containing the full code is available in the online version of this paper. Evaluation of low-rate regimes were performed on a standard four core Intel CPU personal computer with Ubuntu OS. Evaluations of large regimes were performed on the NTNU IDUN High Performance Computing cluster [108]. Each evaluation of an experiment regime was split into individual jobs and assigned to one of the cluster's more than 70 nodes. Each node contains two Intel Xeon cores and each job utilized up to 27 gigabytes of the total 128 gigabytes of main memory. IDUN's storage is provided by two storage arrays and a Lustre parallel distributed file system. Due to limitation in the MATLAB Parallel Computing Toolbox, our evaluations were limited to a single core per node. Future improvements may consider porting our code implementation to a more parallelizable solution to better utilize the highly parallel traits of the PLH.

5 Results

5.1 Model Selection Results

We have reconstructed the connectome for the Cayley tree, random graph, and small world topologies from the Glauber dynamics of an Ising network, using a model selection process which implements the BIC and novel MDLent criteria. This model selection process was repeated for a diverse set of network conditions and sample rates, which we have binned into two large categories based on network size: the small and large network regimes. We compare the reconstruction performance of the BIC and MDLent criteria in the three topologies with respect to these categories.

5.2 Cayley Tree Topology

Reconstruction of the Cayley tree topology establishes a baseline of criteria performance as reconstruction of a tree topology is a simpler task and reveals many of the same trends in the criteria performance which will be seen in other topologies, such as distribution of the error rates with respect to the inverse temperature parameter β. Figure 25 illustrates how criteria performances stays within the same error, with mean values overlapping, but the BIC and MDLent mean reconstruction rates diverge as network density increases. As expected, the global error decreases in regimes of larger sample rates and network size. The reconstruction rates for the Cayley tree show consistent low global reconstruction error, which decreases with network size as can be seen in the large network ($N = 100$) at low-rate (Figure 26), by comparing with the smaller network regimes (Figure 25). There is also a cleaner divergence in criteria performance in the large network regimes, with the BIC maintaining a consistently lower mean error than the MDLent.

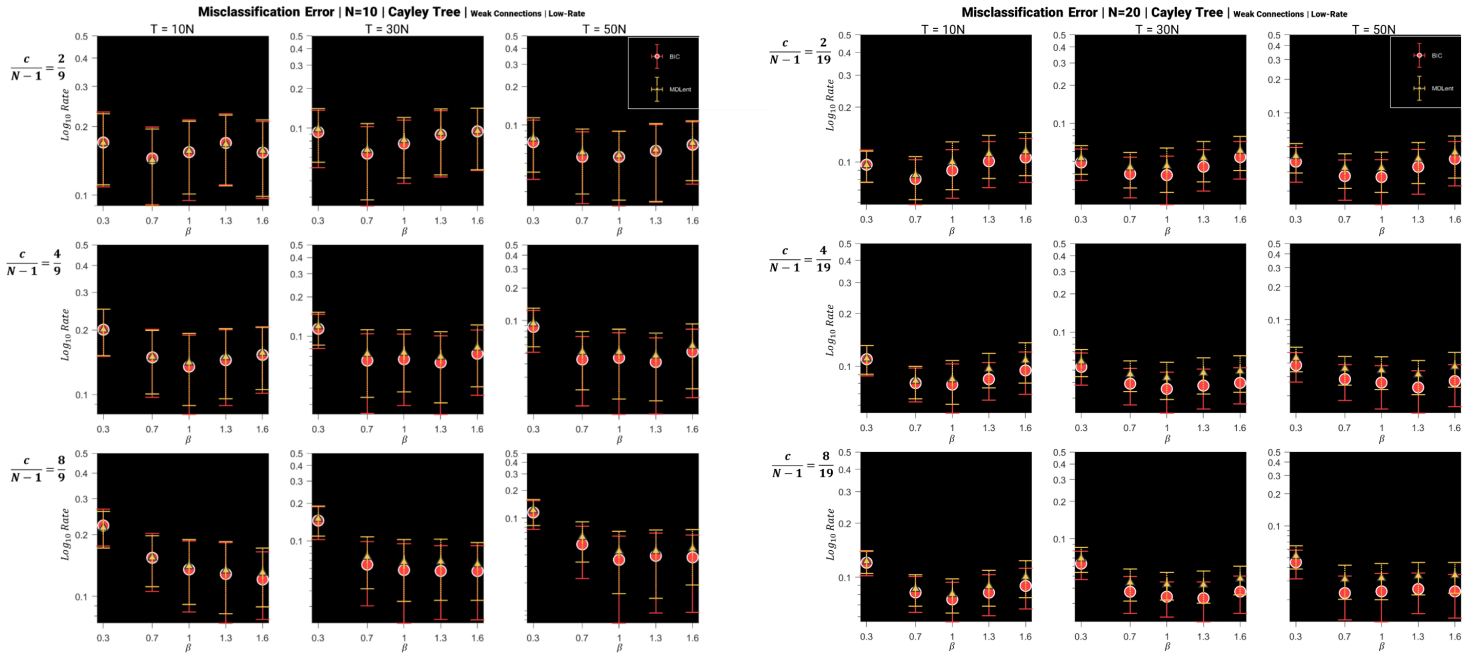


Figure 25
Misclassification error (Equation 46) for Cayley tree in small and large network size regimes. **Left** $N = 10$. **Right** $N = 20$. Subplot y-axes measured in log-scale. BIC is represented by red circular marker. MDLent is represented by yellow triangular marker. Error bars represent standard deviations. Plots are grouped into columns by sample rates T and rows by network density.

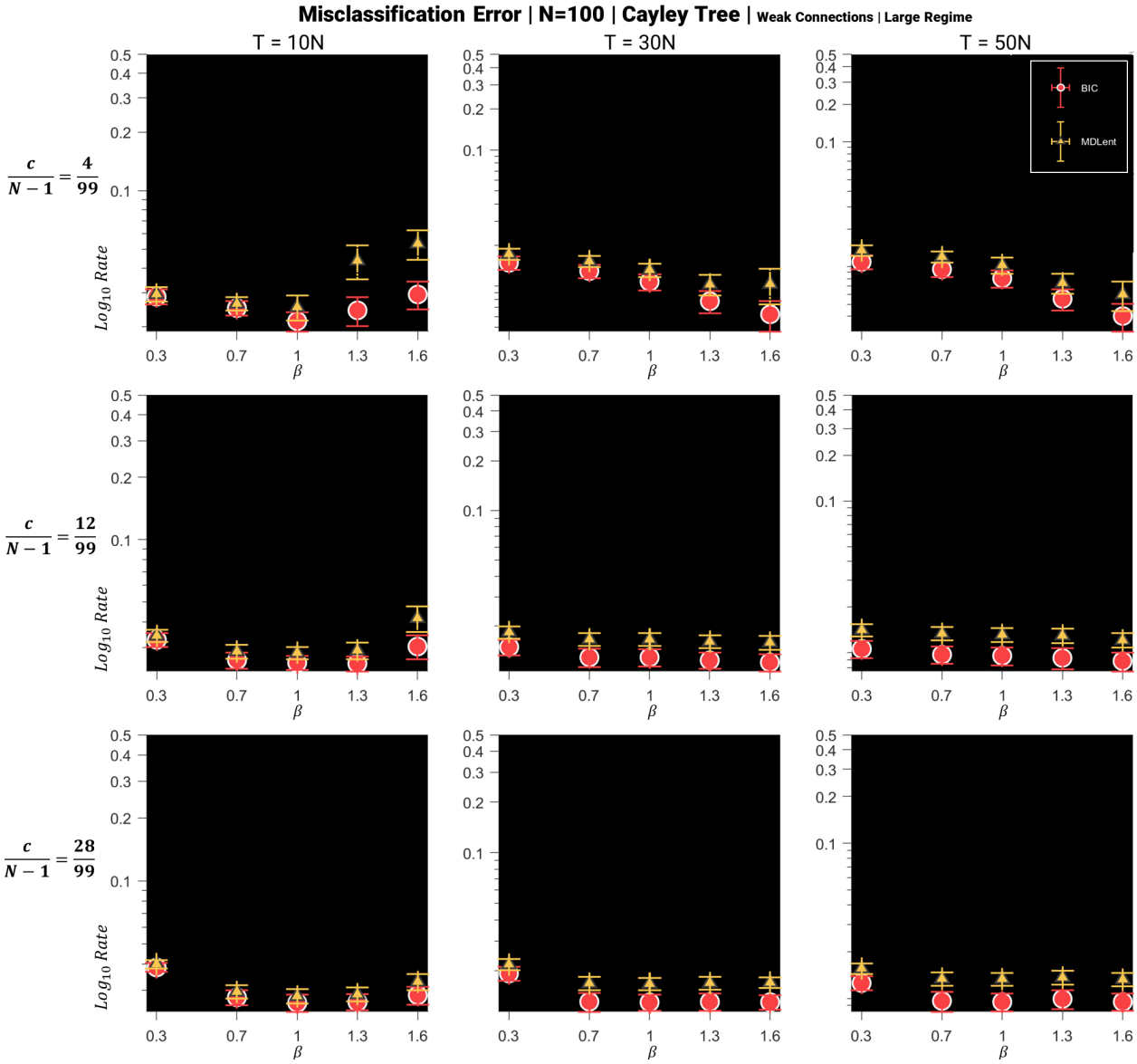


Figure 26
Misclassification error (Equation 46) for Cayley Tree topology, $N = 100$ regime. Subplot y-axes measured in log-scale. BIC is represented by red circular marker. MDLent is represented by yellow triangular marker. Error bars represent standard deviations.

Figure 27 shows the cause of increased MDLent reconstruction error: an increase in false positives. The MDLent FP score diverges from the BIC's scoring higher as density and sample rate increases. There is no divergence between criteria FN scores in the small network regime, but the MDLent begins to score lower than the BIC in the large network as $\beta > 0.3$, and only in the denser network.

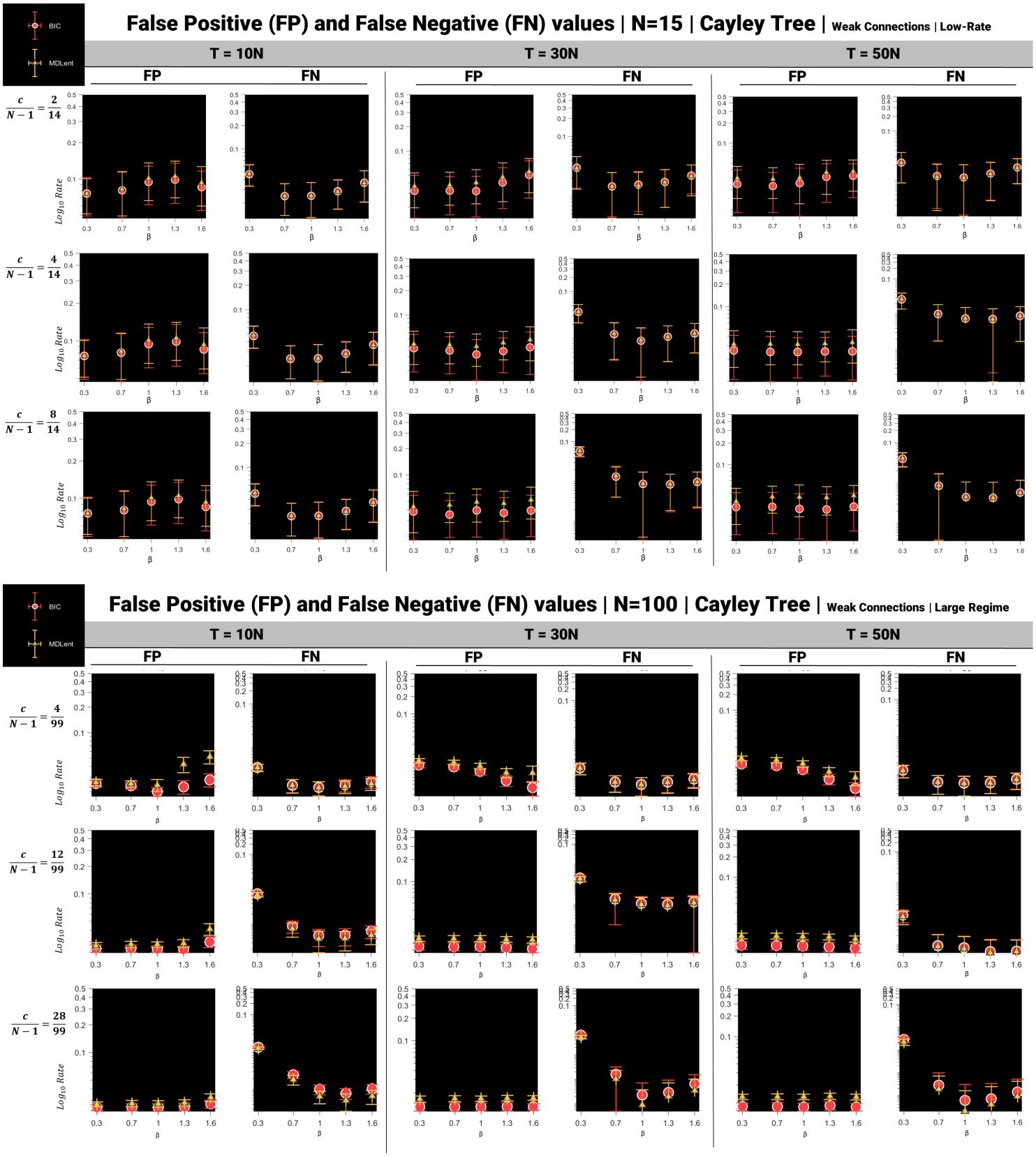


Figure 27
False positives and false negatives in the Cayley tree small network regimes. **Top**: $N = 15$. **Bottom**:$N = 100$. Subplots show total occurrence of FP or FN per network regime. Subplots are divided into 6 columns, split into three groups of sample rates T. Subplot y-axes measured in log-scale. BIC is represented by red circular marker. MDLent is represented by yellow triangular marker. Error bars represent standard deviations.

5.3 Random Graph and Small World Topologies

Reconstruction metrics in the random graph regime parallel those in the small world regimes (Figure 28). Distribution of BIC and MDLent performance metrics are the same in the respective regimes of the small world and random graph topologies. This equivalence appears to hold in all metrics, as can be seen in the FP-FN scores (Figure 29, Figure 30) and ROC (Figure 31). We can safely say results from the small world regimes directly reflect results in the random graph regimes.

5.3.1 Small Network Regimes: Misclassification, FP-FN, ROC

In the small network regime for the random graph and small world topologies, overlap in criteria performance continues. While there is larger variance in global misclassification scores between changes in β (Figure 28a and b) than was seen in the Cayley tree reconstruction, there is still little divergence between criteria reconstruction rates in all densities and sample rates. Criteria reconstruction rates don't begin to widely diverge until the large network regime (Figure 28c and d). The largest difference in performances in the small network regime occurs in the lowest sample rate $T = 10N$, where the MDLent FPR increases relative to the BIC as $\beta > 0.3$ (Figure 31a and b). Figure 29 shows the cause to be an increase in MDLent FPs as β increases.

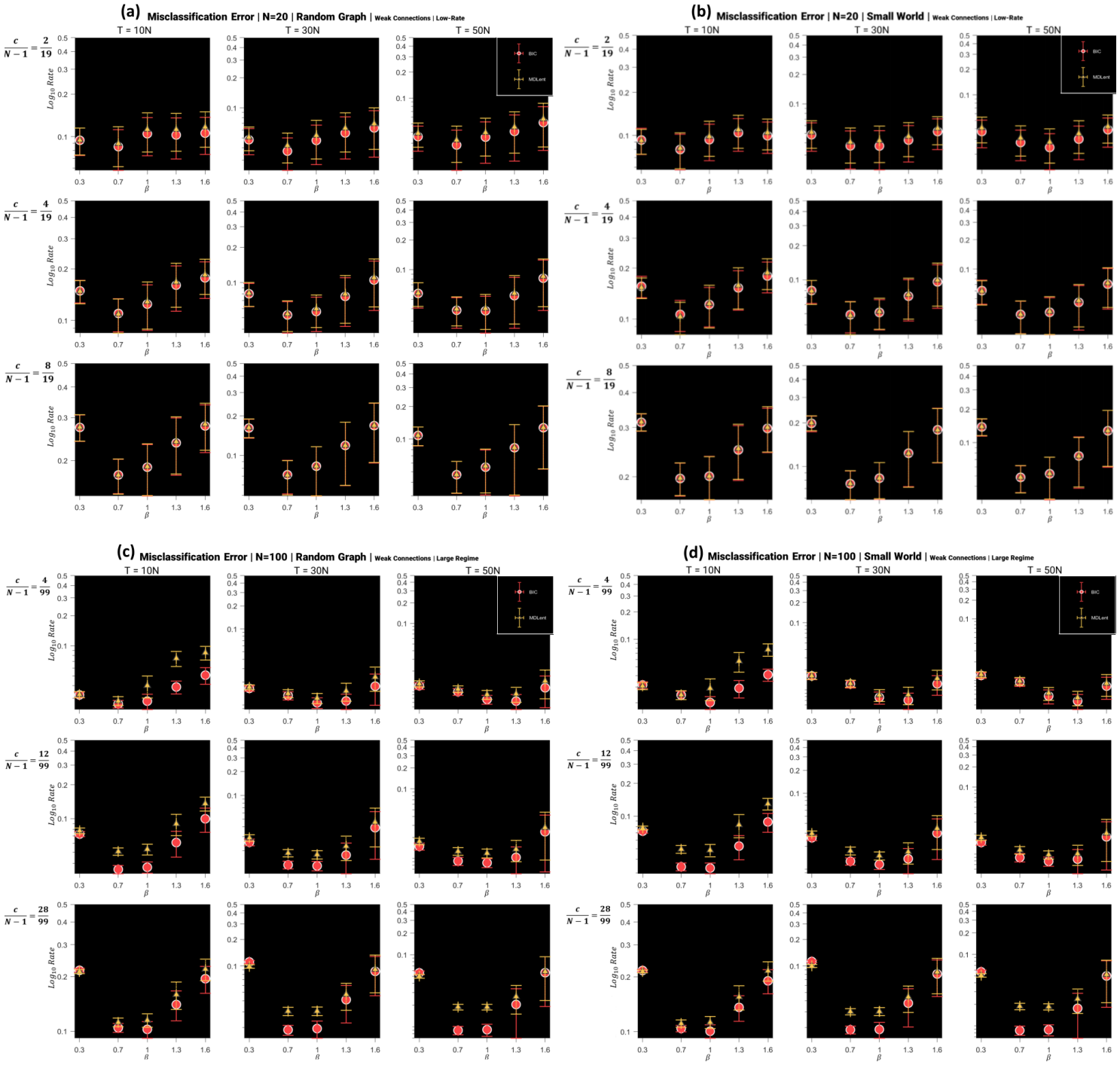


Figure 28

Misclassification comparison for random graph and small world topologies in small $N = 20$ and large network $N = 100$ regimes. Subplot y-axes measured in log-scale. BIC is represented by red circular marker. MDLent is represented by yellow triangular marker. Error bars represent standard deviations.

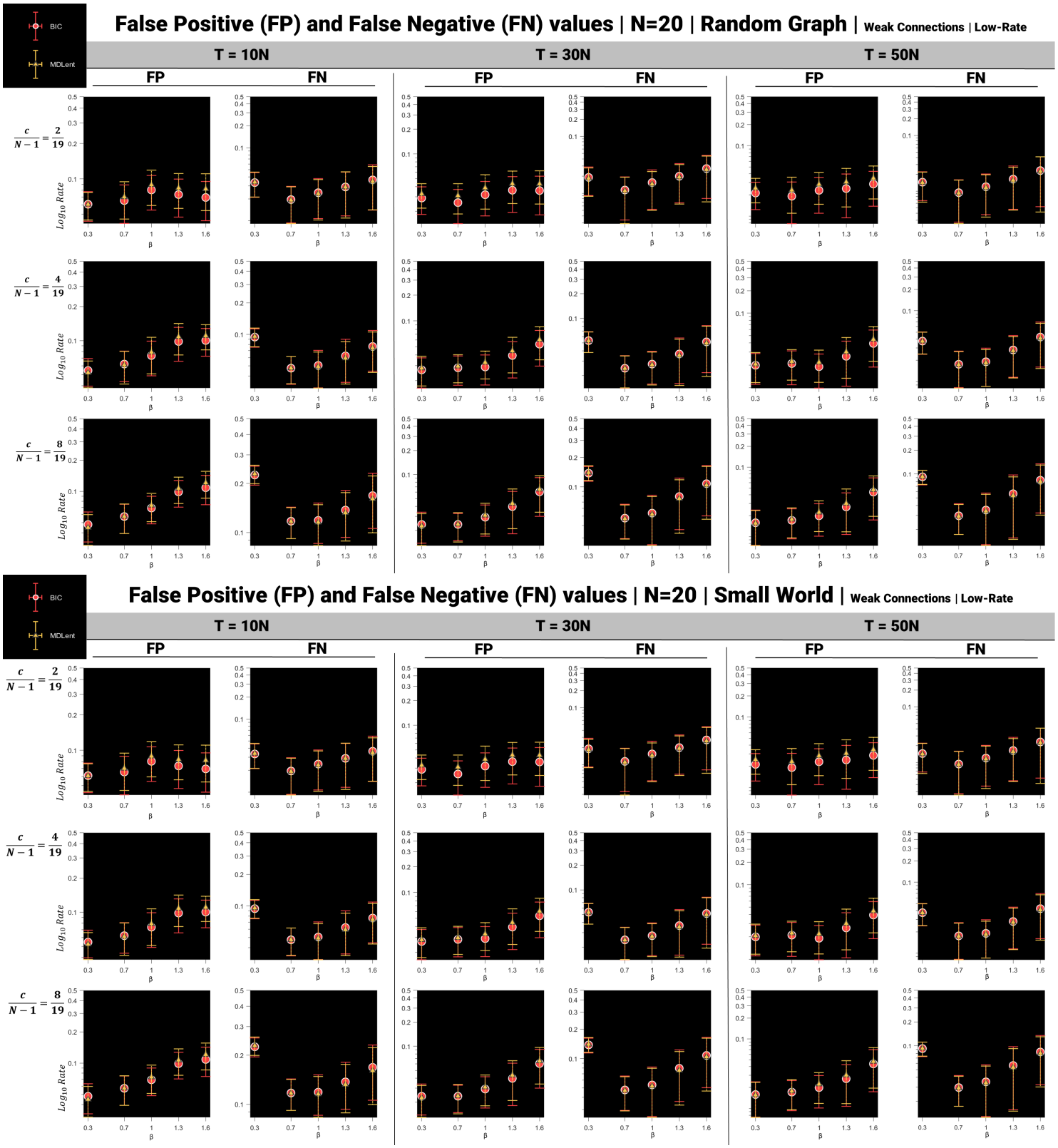
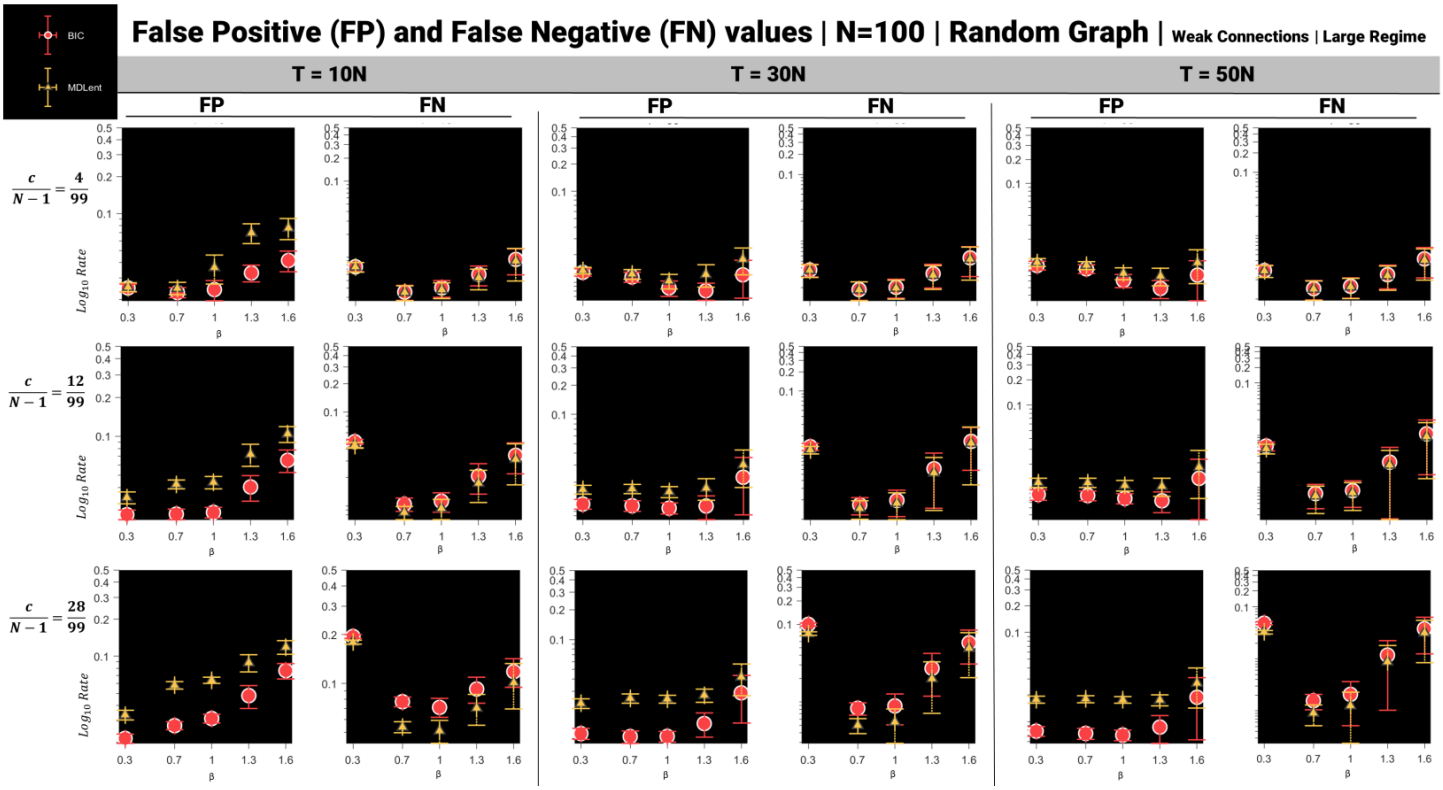


Figure 29
False positives and false negatives for the random graph **(Top)** and small world **(Bottom)** topologies in the small network $N = 20$ regime. Subplot y-axes measured in log-scale. BIC is represented by red circular marker. MDLent is represented by yellow triangular marker. Error bars represent standard deviations.



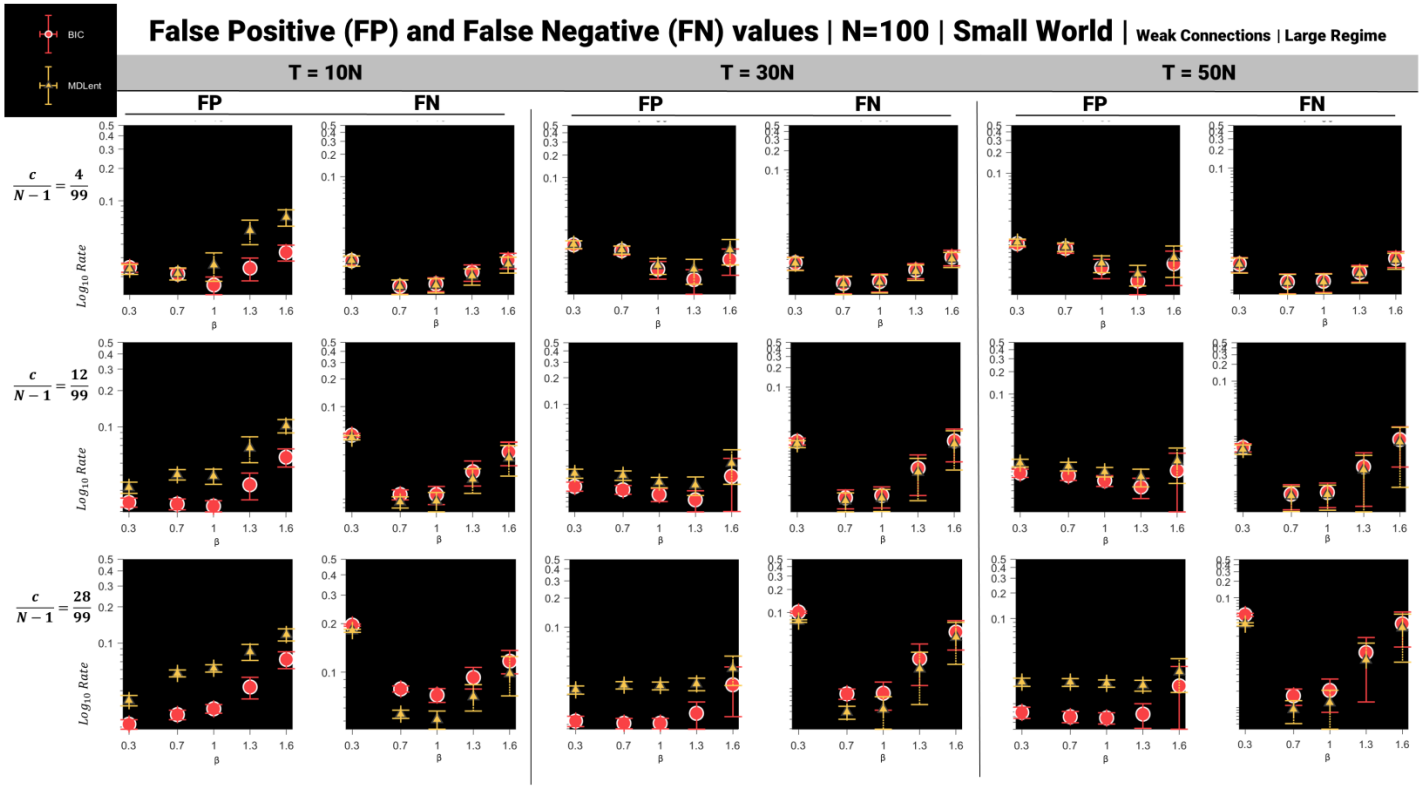


Figure 30
False positives and false negatives for the random graph **(Top)** and small world **(Bottom)** topologies in the large network $N = 100$ regime. Subplot y-axes measured in log-scale. BIC is represented by red circular marker. MDLent is represented by yellow triangular marker. Error bars represent standard deviations.

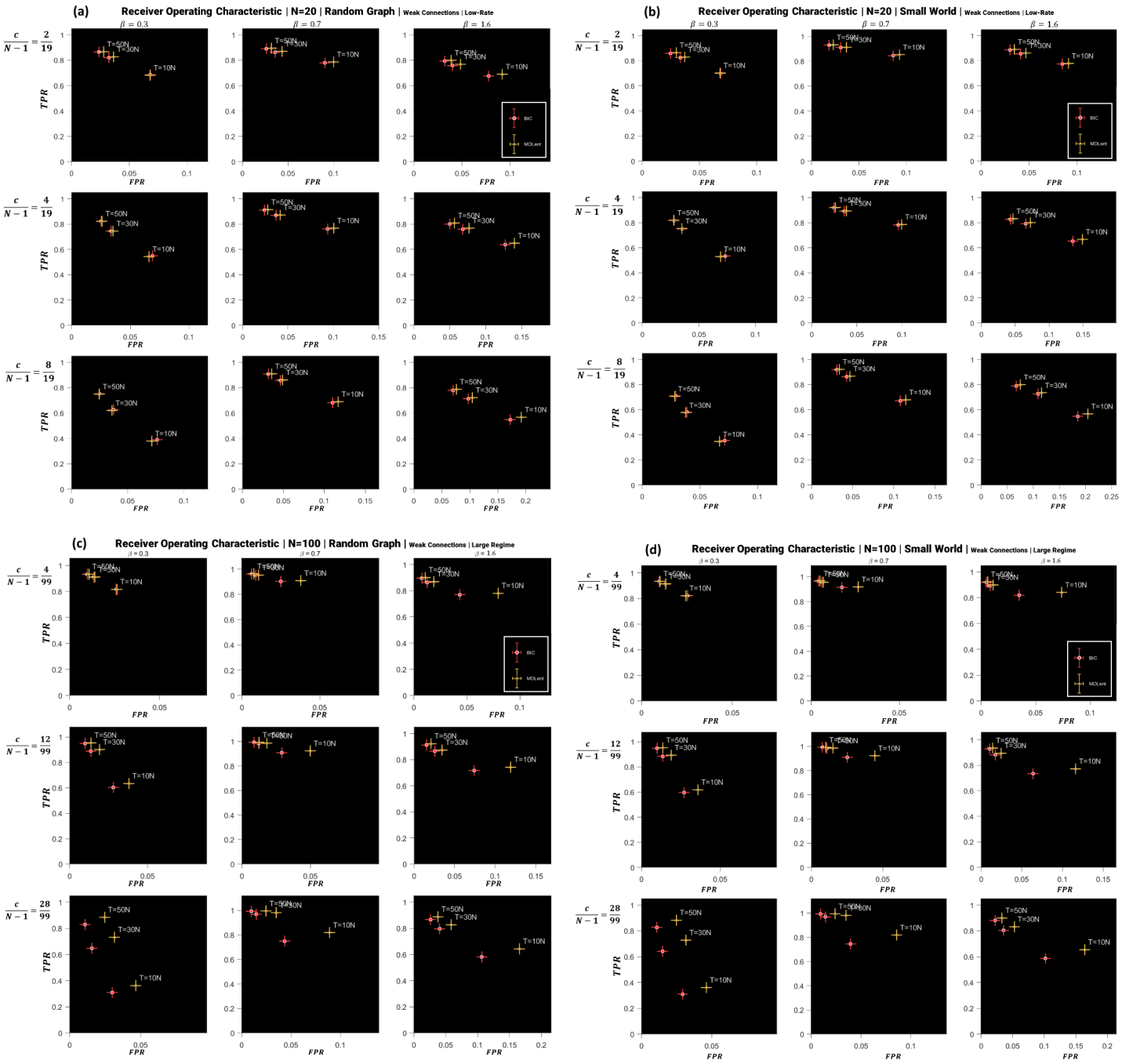


Figure 31
ROC (Equation 47) comparison for random graph and small world topologies in small $N = 20$ and large networks $N = 100$ regimes. BIC is represented by red circular marker. MDLent is represented by yellow triangular marker. Error bars are removed for clarity.

5.3.2 Large Network Regimes - Misclassification Error

In the large network regime ($N > 50$), major differences in criteria performance appear. In the low-rate regimes (Figure 32), global misclassification rates suddenly increases with network density increases. Likewise, the MDLent reconstruction error shows better performance relative to the BIC as the network density increases. This only appears in the low-rate $T < 50N$, as global reconstruction errors rapidly decrease and re-converge as sample rate increases. Comparing the performance in a smaller network $N = 50$ and a larger network at similar levels of network density, shows a similar performance between criteria in the high-rate and convergence of criteria reconstruction error as sample rate increase (Figure 33). Comparing reconstruction error at low and high densities in the largest network (Figure 34) also shows the MDLent performing better relative to the BIC in a low-rate regime, especially in higher densities where global reconstruction error rapidly increases. A figure of misclassification scores for all densities in a large network ($N = 100$) is available in the appendix, Figure A2.

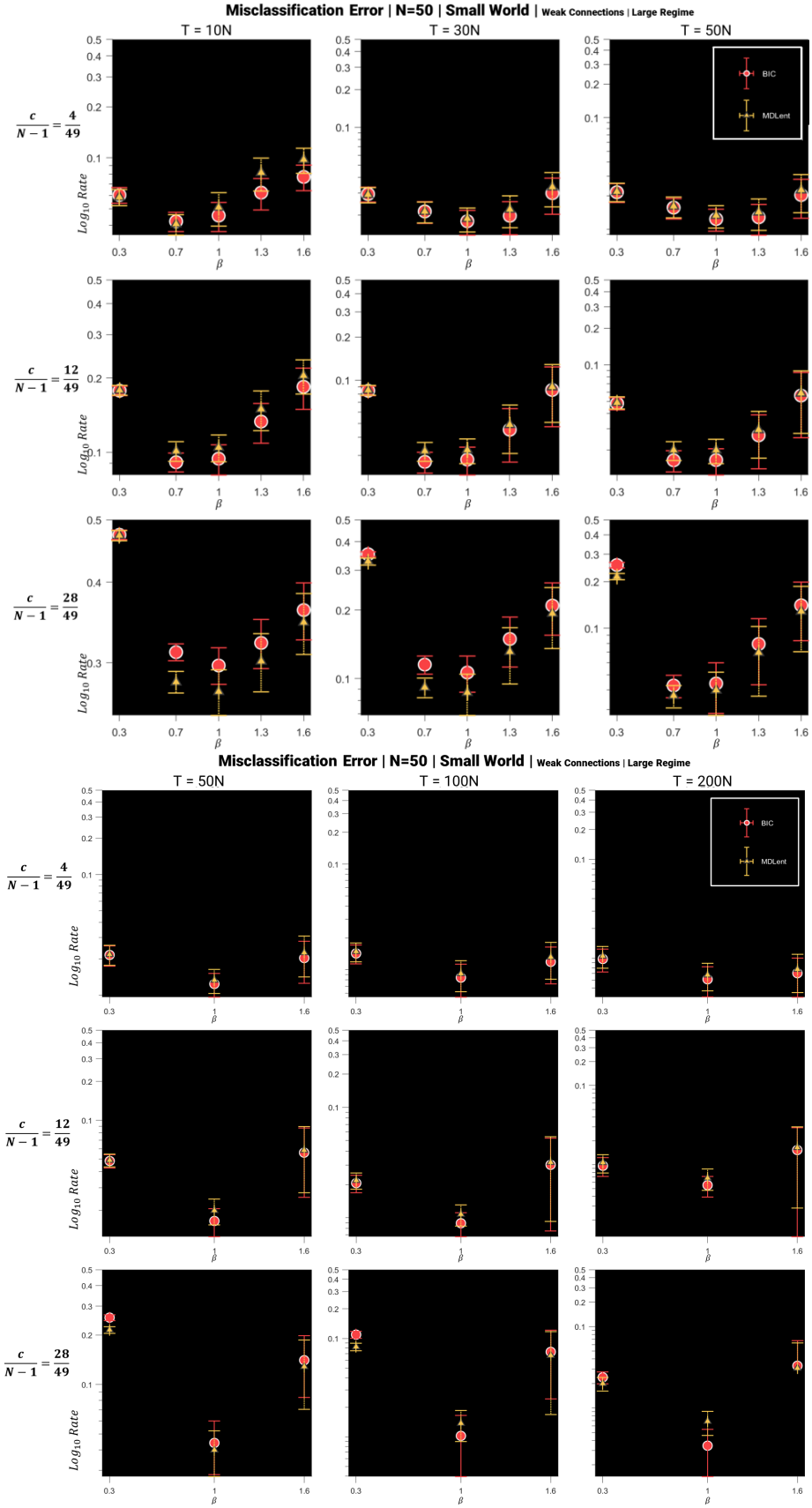


Figure 32
Misclassification error (Equation 46) for small world topology, $N = 50$ at low and high sample rates. Select beta for the high-rate. Subplot y-axes measured in log-scale. BIC is represented by red circular marker. MDLent is represented by yellow triangular marker. Error bars represent standard deviations.

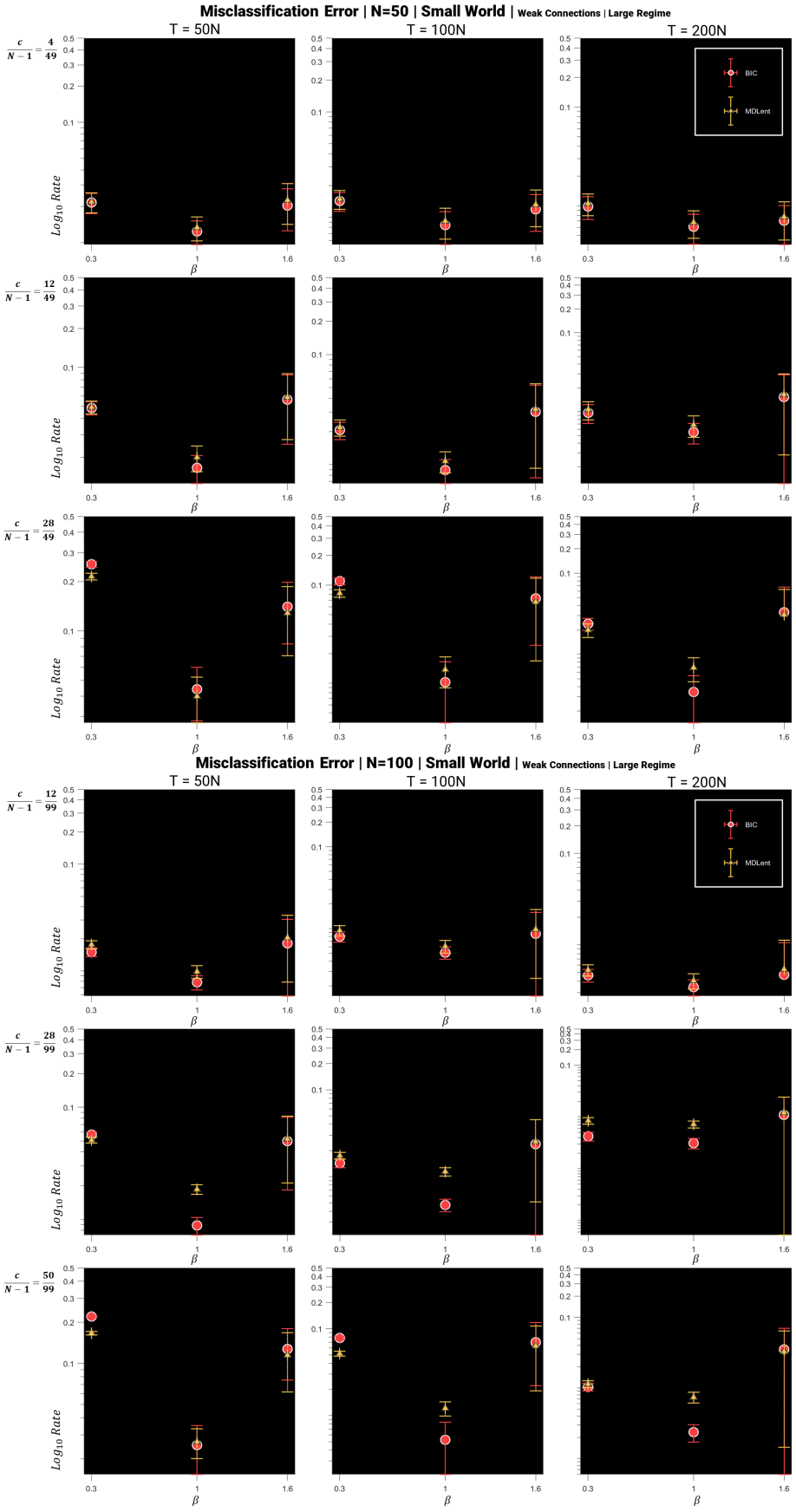


Figure 33
Misclassification for $N = [50, 100]$ for high sample rates with similar densities. **(Top)** $N = 50$, $density = [0.8, 0.24, 0.57]$. **(Bottom)** $N = 100$, $density = [0.12, 0.28, 0.5]$. Subplot y-axes measured in log-scale. BIC is represented by red circular marker. MDLent is represented by yellow triangular marker. Error bars represent standard deviations.

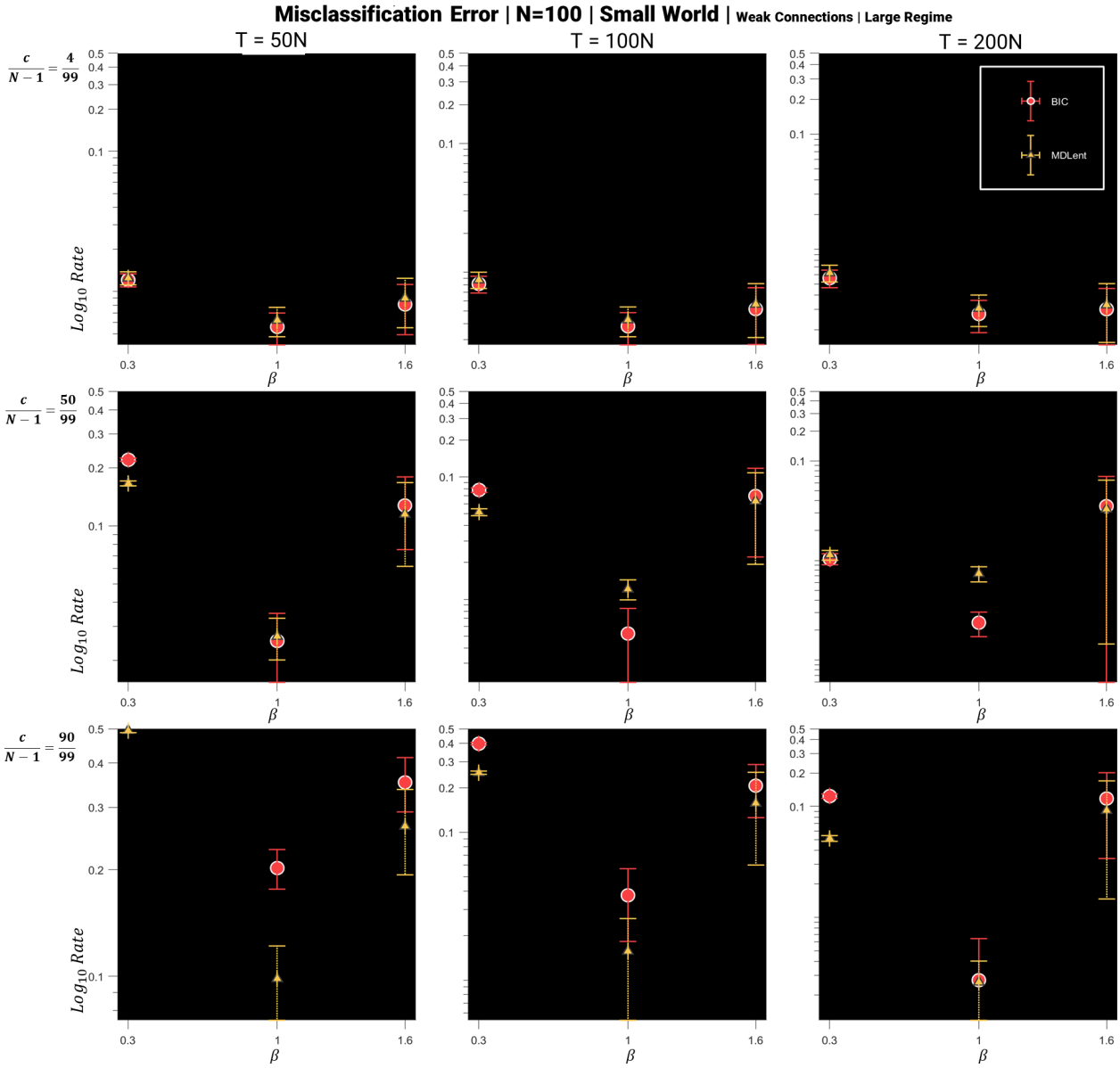


Figure 34
Misclassification error (Equation 46) for $N = 100$ regime at $\mathcal{C} = [4, 50, 90]$. Misclassification for all densities, available in appendix fig:multidens. Subplot y-axes measured in log-scale. BIC is represented by red circular marker. MDLent is represented by yellow triangular marker. Error bars represent standard deviations. Figure of misclassification scores for all densities in a large network ($N = 100$) is available in the appendix, Figure A2.

5.3.3 Large Network Regimes - False Positive - False Negative Rates

In the small world FP-FN figures ($N = 100$, Figure 30 and $N = 50$, Figure 35) the MDLent scores a higher total of false positives relative to the BIC. However, the global FP score decreases as network density increases and the performance difference between criteria is slight. The area of worst performance for both criteria occurs in dense networks where global FN rate increases, this is mediated as sample rate increases. However, the FN score for both criteria ranges between $\approx 0.2 - 0.45$ in the weakest regime $\beta = 0.3$. The MDLent actually performs better relative the BIC in FN score, its relative performance increasing with the network density $\beta > 0.3$, but the gap in performance closes as sample rate increases (Figure 35).

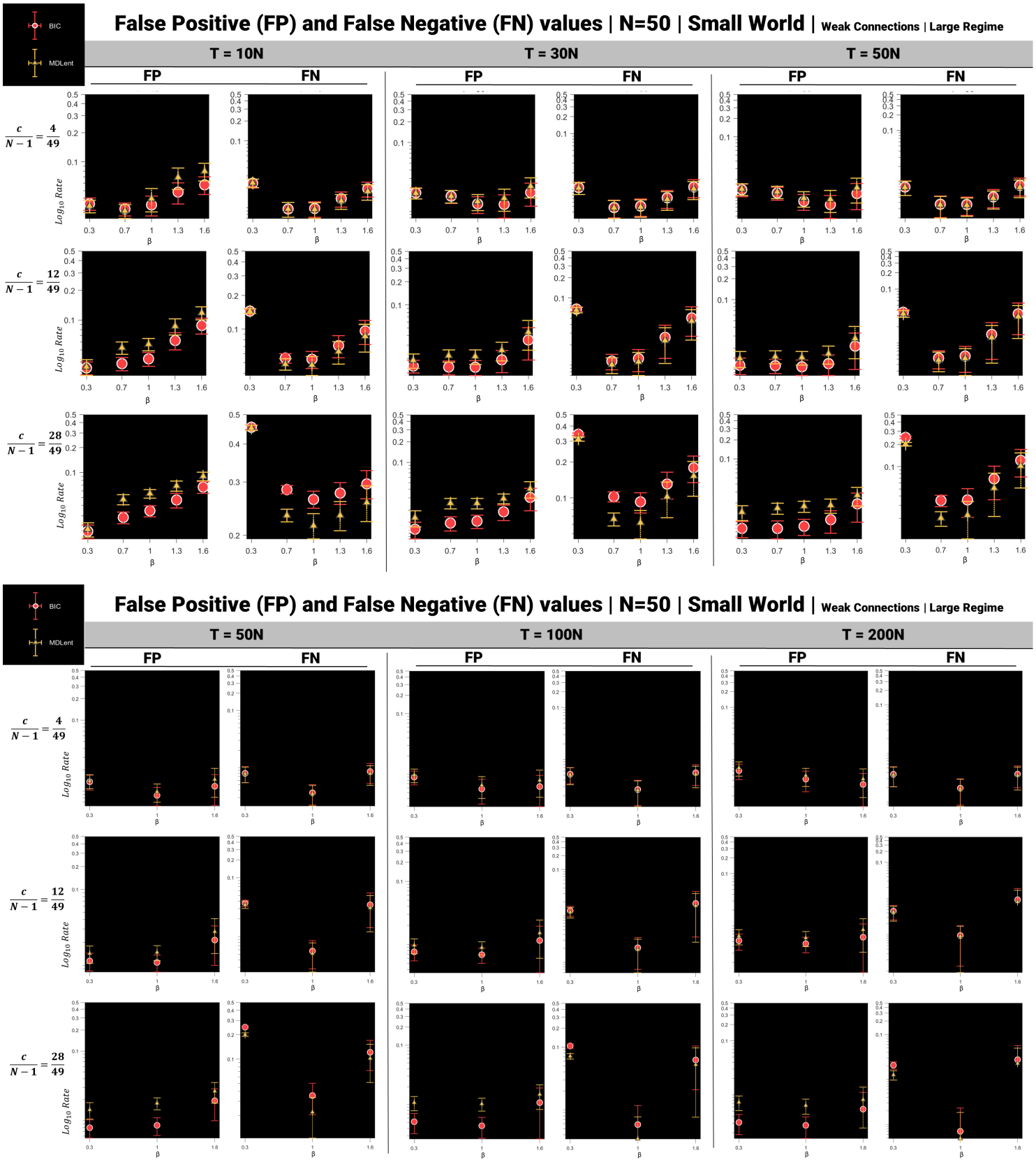


Figure 35
False positives and false negatives for small world topology $N = 50$ at low-rate **(Top)** and high-rate **(Bottom)** regimes. Subplot y-axes measured in log-scale. BIC is represented by red circular marker. MDLent is represented by yellow triangular marker. Error bars represent standard deviations.

5.3.4 Large Network Regimes - ROC: TPR v. FPR

Figure 36 gives the clearest depiction of the the global performance with respect to the sample rate, with both criteria increasing in performance approaching a $TPR \approx 1$ and $FPR < 0.01$ in the largest sample size,$T = 200N$. The MDLent TPR score matches or outperforms the BIC in all regimes, especially ones of higher density. There is a global increase in TPR as β increases and the MDLent TPR score increases relative to the BIC in smaller sample rates. The MDLent FPR score also increases relative to the BIC as the network density increases the largest margin of which is < 0.08.

Figure 37 compares ROC scores in regimes of similar density in different network sizes, there are no changes in global TPR in respect to network size, but a decrease in BIC FPR. A figure of ROC for all densities in a large network ($N = 100$) is available in the appendix, Figure A3.

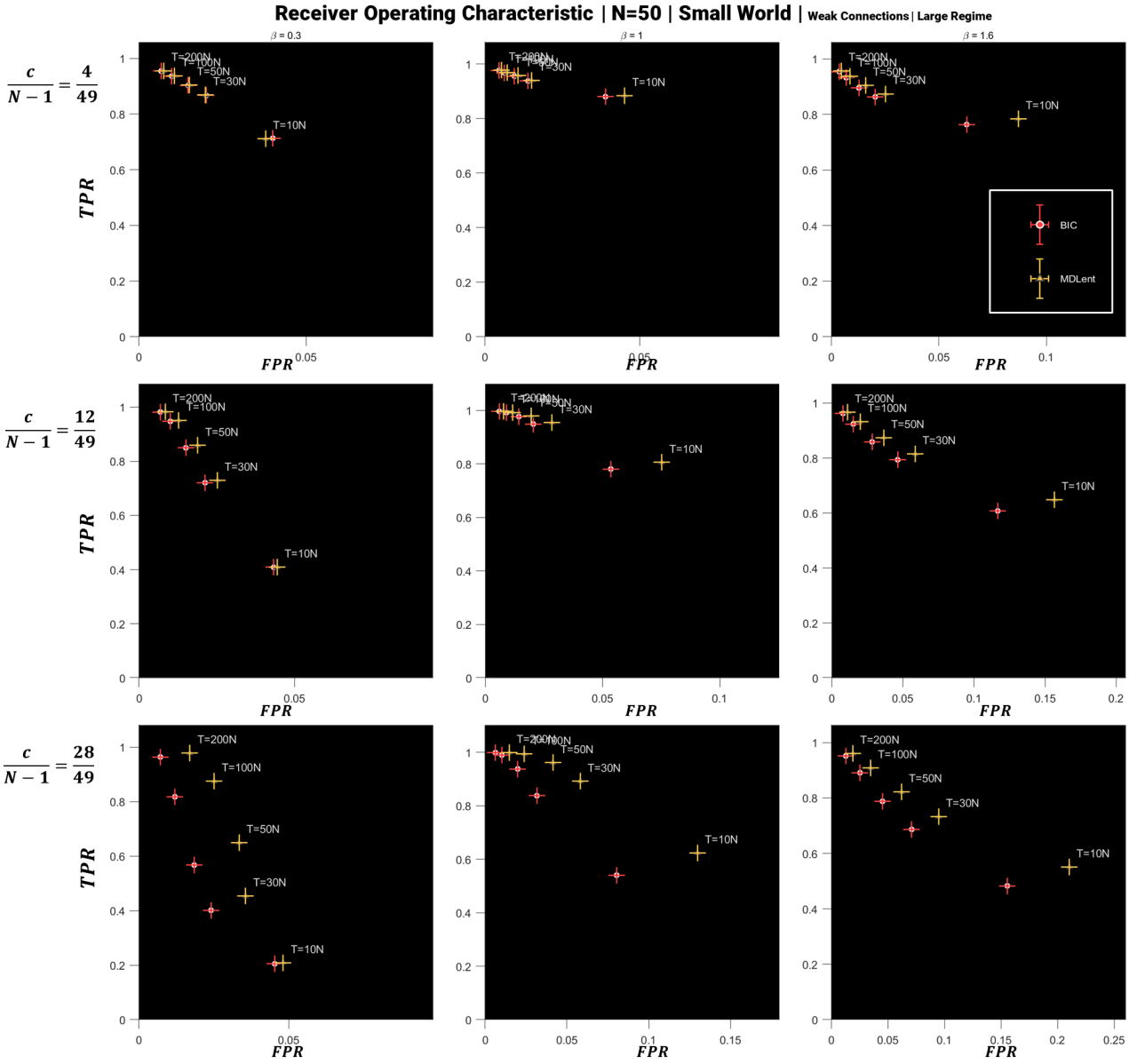


Figure 36
ROC (Equation 47) Small World Topography for $N = 50$ for all sample rates. BIC is represented by red circular marker. MDLent is represented by yellow triangular marker. Error bars removed for clarity.

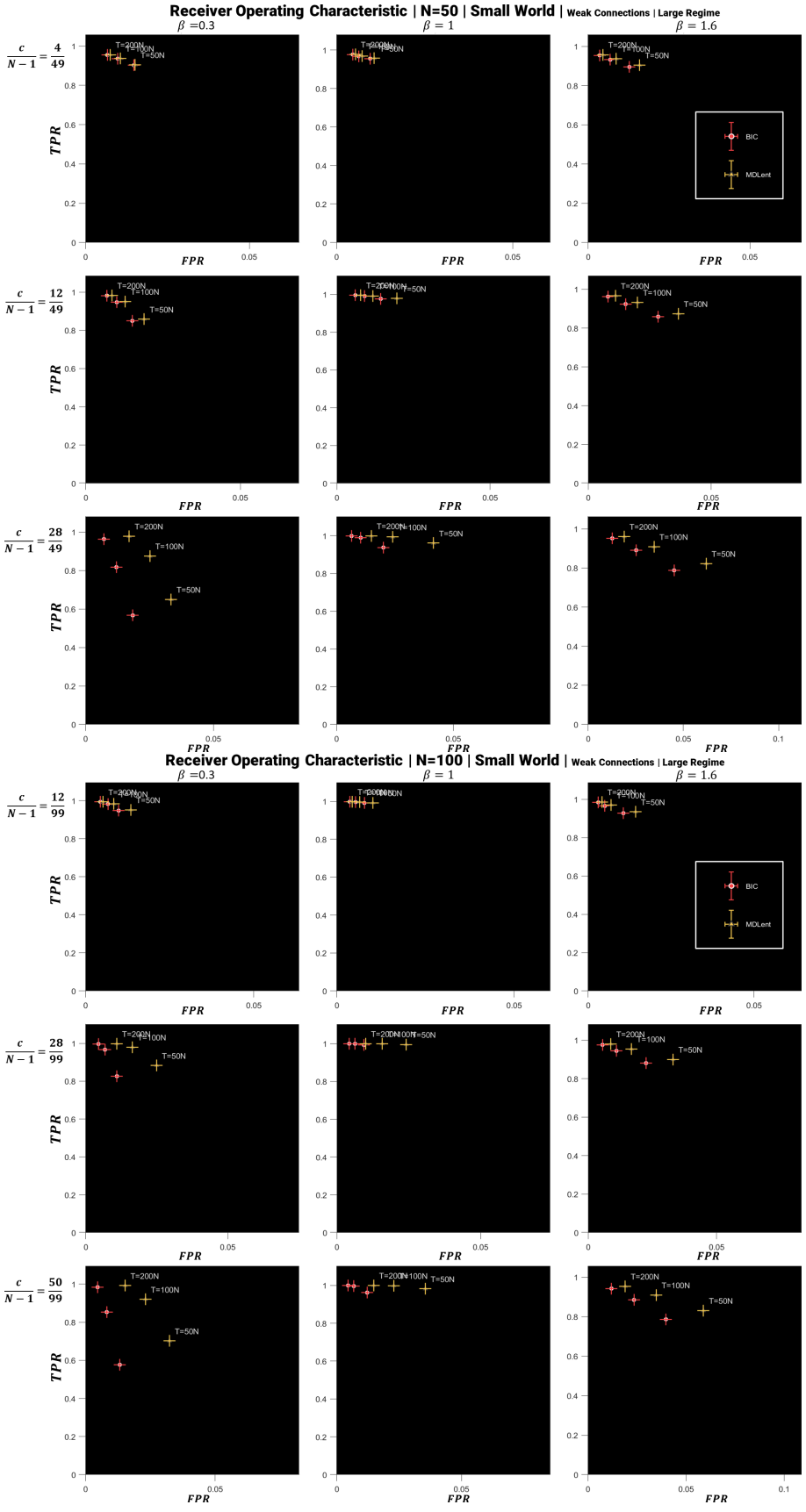


Figure 37
ROC (Equation 47) for $N = [50, 100]$ for high sample rates with similar densities.
(Left) $N = 50$ $density = [0.8, 0.24, 0.57]$.
(Right) $N = 100$ $density = [0.12, 0.28, 0.5]$. BIC is represented by red circular marker. MDLent is represented by yellow triangular marker. Error bars removed for clarity.

6 Discussion and Future Directions

6.1 Discussion

In this project we have studied the equilibrium Ising network model and its inverse solutions in order to implement them into a greater framework of Bayesian model selection, and test the performance of the novel MDL selection criteria introduced in Bulso et al. 2019 [19] in the inference of Ising model networks with a pseudo-log-likelihood implementation. We compare this performance against another well established model selection criteria, the BIC. In addition to the random graph topology used in the 2019 paper we also test the selection criteria in two additional topologies; the Cayley tree for its reliability in inference problems and the Watts-Strogatz small world network as a test of performance in naturally occurring topologies. We also use a different distribution of Ising parameters, opting for a normalized split mean Gaussian distribution with no external bias.

Recovery of the Cayley tree network topology regimes returned very low error across all conditions tested and the results displayed trends similar to those found in the other topological regimes, however less pronounced and more homogeneous between regimes. The BIC and MDLent criteria error rates hardly diverged except in the low density and low-rate regimes of the large network where the MDLent returned a higher error rate than the BIC. The "flattened" and unique distributions of reconstruction scores in the Cayley tree regimes contrasts with the high variability in performance between inverse temperatures β in the random and small world graphs. In the small world and random graph regimes the MDLent performance was similar to the BIC in sparse networks and in high-rate regimes, but showed a performance advantage in regimes of limited observation samples and higher density. The MDLent criteria tended towards overestimating connections in larger and denser regimes, scoring a consistently higher false positive rate. However, it also showed increased accuracy in inferring sparse connections with a reduced false negative rate. This led to an even or increased TPR in the MDLent with respect to the BIC in all tests.

The MDLent followed the same reconstruction performance as the BIC (with a *slight* increase in its margin of error) except in the dense and low-rate regimes, where its reconstruction performance was better and closer to that of the AIC. This reflects the results in Bulso et al. 2019 for the random graph model with samples generated from a flat distribution of Ising inputs. This was not entirely reflected in the Cayley tree results, in which the MDLent never performed better than the BIC even in low rate or dense connections, but general trends could still be observed. Surprisingly, the criteria reconstruction metrics in the small world model regimes were near exact to the reconstruction results returned from the random graph regimes. We were able to confirm this wasn't caused by an error in the methods, and the results showed a minute enough difference to prove separate topological adjacency graphs were used in the forward process. This similarity between results was confirmed to be consistent across all regimes tested for the two topologies and in the false positive and false negative rates. This would suggest the small world and random graph models may have been too similar in shared topological properties, such as node degree or clustering coefficient. The Watts-Strogatz small world model may have had too much overlap with the random graph, as it resolves to a random graph as network density increases. Future experiments will need to better track topological properties and perhaps a more distinct topological model could be tested in the future such as a scale-free or power-law network topology.

The overall performance of the MDLent across the network conditions and sample-rates tested, proves it is the best choice of the model selection methods in the diverse set of experimental conditions tried, meaning it would be a preferable choice of selection method in situations where underlying ground truths of the model are unknown or poorly informed such as biological systems. As we show here, its use with a maximum entropy model could prove further benefits in inferring models containing "hidden variables" [11] [99], however more experimentation will be required to know if the MDLent can be reasonably applied in these settings. As we discovered and as is mentioned in Bulso et al. 2019, the increase in incorrectly inferred connections by the MDLent would require additional processing steps; the authors of the 2019 paper even suggest using the MDLent criteria as a first step in an implementation followed by a more complex algorithm [19]. This would be a required measure in recovering the graph of functional connectivity. Our quick examination of the pairwise reconstruction (symmetrization) of the recovered graph (Section 4.3.2

and Appendix C) showed a much higher misclassification rate that may not be suitable in actual applications. This could be rectified by an additional post-processing step on the subspace of models from the recovered graph, e.g. by cross-validation (as suggested by the Bulso et al. 2019) or by a model walk of the subspace in the recovered model [87].

The concept of applying minimum description length methods to model selection is one that is still being explored [26] [48] and could be quite useful in biological networks where evolutionary constraints have maximized systems for high informational transfer at minimal energy expenditure. The novel MDLent method of penalizing the model space based on the encoded informational content should continue to be expanded on. Because the MDLent model space localization depends on the observed distribution frequency of code-words, it may be useful to apply pre-processing filters to identify input samples of minimal informational value, such as applying a multiscale relevance [27] implementation, or a simple cutoff ranking in the Zipf's law order [71] of the unique code-words. These minimally informative samples could then either be thrown out, or have an error correction method applied to reinterpret them as a maximally informative sample of nearest hamming distance.

If not for time we would have preferred to test other Ising parameter distributions and reproduce an Ising model more closely based on maximum entropy models of neural recordings which appear to show a single mean Gaussian distribution in their connection strengths with some bias (Figure 13) [115] [10], or networks of unevenly distributed inhibitory/excitatory connections. The intention being to eventually apply these methods to *in-vitro* and *in-vivo* neural recording data. This will be a topic covered in any follow up research.

The Bulso et al. 2019 novel MDL criterion, is a highly useful method of Bayesian model selection which seemingly bridges the AIC-BIC dilemma, proving highly accurate while recovering a larger diversity of models in comparison to other model selection criteria we've tested here. Future exploration of the MDLent should test its use over other network conditions with additional processing methods attached, especially if used in the recovery of network functional connectivity.

Bibliography

[1] Peter Aaser et al. 'Towards Making a Cyborg : A Closed-Loop Reservoir-Neuro System'. In: *Ecal* September (2017), pp. 430–437. ISSN: 978-0-262-34633-7. DOI: 10.7551/ecal_a_072.

[2] Hirotugu Akaike. 'A New Look at the Statistical Model Identification'. In: *IEEE Transactions on Automatic Control* 19.6 (1974), pp. 716–723. DOI: 10.1109/TAC.1974.1100705.

[3] Daniel J Amit and Daniel J Amit. *Modeling Brain Function: The World of Attractor Neural Networks*. Cambridge university press, 1992.

[4] Valerio Arnaboldi et al. 'Chapter 2 - Human Social Networks'. In: *Online Social Networks*. Ed. by Valerio Arnaboldi et al. Computer Science Reviews and Trends. Boston: Elsevier, 2015, pp. 9–35. ISBN: 978-0-12-803023-3. DOI: 10.1016/B978-0-12-803023-3.00002-3.

[5] E. Aurell, C. Ollion and Y. Roudi. 'Dynamics and Performance of Susceptibility Propagation on Synthetic Data'. In: *The European Physical Journal B* 77.4 (Oct. 2010), pp. 587–595. DOI: 10.1140/epjb/e2010-00277-0.

[6] Claudio Babiloni et al. 'Fundamentals of Electroencefalography, Magnetoencefalography, and Functional Magnetic Resonance Imaging'. eng. In: *International Review of Neurobiology* 86 (2009), pp. 67–80. ISSN: 0074-7742. DOI: 10.1016/S0074-7742(09)86005-4.

[7] Marc Bailly-Bechet et al. 'Inference of Sparse Combinatorial-Control Networks from Gene-Expression Data: A Message Passing Approach'. In: *BMC Bioinformatics* 11.1 (June 2010), pp. 355–355. DOI: 10.1186/1471-2105-11-355.

[8] Vijay Balasubramanian. *Statistical Inference, Occam's Razor and Statistical Mechanics on The Space of Probability Distributions*. 1996. DOI: 10.1162/neco.1997.9.2.349.

[9] M. Ballerini et al. 'Interaction Ruling Animal Collective Behavior Depends on Topological Rather than Metric Distance: Evidence from a Field Study'. In: *Proceedings of the National Academy of Sciences of the United States of America* 105.4 (Jan. 2008), pp. 1232–1237. DOI: 10.1073/pnas.0711437105.

[10] John Barton and Simona Cocco. 'Ising Models for Neural Activity Inferred via Selective Cluster Expansion: Structural and Coding Properties'. In: *Journal of Statistical Mechanics: Theory and Experiment* 2013.3 (2013). DOI: 10.1088/1742-5468/2013/03/P03002.

[11] Claudia Battistin, Benjamin Dunn and Yasser Roudi. 'Learning with Unknowns: Analyzing Biological Data in the Presence of Hidden Variables'. In: *Current Opinion in Systems Biology* 1 (2017), pp. 122–128. DOI: 10.1016/j.coisb.2016.12.010.

[12] A. J. Berlinsky and A. B. Harris. 'The Cayley Tree'. In: Springer, Cham, 2019, pp. 345–370. DOI: 10.1007/978-3-030-28187-8_14.

[13] Julian Besag. *Spatial Interaction and the Statistical Analysis of Lattice Systems*. Tech. rep. 2. 1974, pp. 192–236.

[14] Eva Bianconi et al. 'An Estimation of the Number of Cells in the Human Body'. In: *Annals of Human Biology* 40.6 (Nov. 2013), pp. 463–471. DOI: 10.3109/03014460.2013.807878.

[15] Stanislav S Borysov, Yasser Roudi and Alexander V Balatsky. 'US Stock Market Interaction Network as Learned by the Boltzmann Machine'. In: *The European Physical Journal B* 88.12 (2015), p. 321.

[16] Alessio P. Buccino et al. 'Combining Biophysical Modeling and Deep Learning for Multielectrode Array Neuron Localization and Classification'. In: *Journal of Neurophysiology* 120.3 (2018), pp. 1212–1232. DOI: 10.1152/jn.00210.2018.

[17] Alessio P. Buccino et al. 'Independent Component Analysis for Fully Automated Multi-Electrode Array Spike Sorting'. In: *Conference proceedings : ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual Conference* 2018 (2018), pp. 2627–2630. ISSN: 9781538636466. DOI: 10.1109/EMBC.2018.8512788.

[18] P. A. Buchs and D. Muller. 'Induction of Long-Term Potentiation Is Associated with Major Ultrastructural Changes of Activated Synapses'. en. In: *Proceedings of the National Academy of Sciences* 93.15 (July 1996), pp. 8040–8045. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.93.15.8040.

[19] Nicola Bulso, Matteo Marsili and Yasser Roudi. *On the Complexity of Logistic Regression Models*. Tech. rep. 2019, pp. 1–50.

[20] Nicola Bulso, Matteo Marsili and Yasser Roudi. 'Sparse Model Selection in the Highly Under-Sampled Regime'. In: *Journal of Statistical Mechanics: Theory and Experiment* 2016.9 (Sept. 2016), pp. 093404–093404. DOI: 10.1088/1742-5468/2016/09/093404.

[21] H.B. Callen. 'A Note on Green Functions and the Ising Model'. en. In: *Physics Letters* 4.3 (Apr. 1963), p. 161. ISSN: 00319163. DOI: 10.1016/0031-9163(63)90344-5.

[22] Andrea Cavagna et al. 'New Statistical Tools for Analyzing the Structure of Animal Groups'. In: *Mathematical Biosciences* 214.1-2 (July 2008), pp. 32–37. DOI: 10.1016/j.mbs.2008.05.006.

[23] Professor Cayley. 'Desiderata and Suggestions: No. 2. The Theory of Groups: Graphical Representation'. In: *American Journal of Mathematics* 1.2 (1878), p. 174. ISSN: 00029327. DOI: 10.2307/2369306.

[24] Simona Cocco et al. 'Functional Networks from Inverse Modeling of Neural Population Activity'. In: *Current Opinion in Systems Biology* 3 (June 2017), pp. 103–110. DOI: 10.1016/j.coisb.2017.04.017.

[25] Simona Cocco et al. 'Inverse Statistical Physics of Protein Sequences: A Key Issues Review'. In: *Reports on Progress in Physics* 81.3 (Jan. 2018), pp. 032601–032601. DOI: 10.1088/1361-6633/aa9965.

[26] Ryan Cubero, Matteo Marsili and Yasser Roudi. 'Minimum Description Length Codes Are Critical'. In: *Entropy* 20.10 (Oct. 2018), pp. 755–755. DOI: 10.3390/e20100755.

[27] Ryan John Cubero. 'Statistical Criticality Arises in Most Informative Representations Recent Citations Multiscale Relevance and Informative Encoding in Neuronal Spike Trains'. In: (2019). DOI: 10.1088/1742-5468/ab16c8.

[28] Daphne Koller. *Probabilistic Graphical Models : Principles and Techniques*. 2009, p. 1270. ISBN: 978-0-262-01319-2.

[29] Aurélien Decelle and Federico Ricci-Tersenghi. 'Pseudolikelihood Decimation Algorithm Improving the Inference of the Interaction Network in a General Class of Ising Models'. In: (2014). DOI: 10.1103/PhysRevLett.112.070603.

[30] Jie Ding, Vahid Tarokh and Yuhong Yang. 'Bridging AIC and BIC: A New Criterion for Autoregression'. In: *IEEE Transactions on Information Theory* 64.6 (2018), pp. 4024–4043. DOI: 10.1109/TIT.2017.2717599.

[31] Jie Ding, Vahid Tarokh and Yuhong Yang. 'Model Selection Techniques: An Overview'. In: *IEEE Signal Processing Magazine* 35.6 (2018), pp. 16–34. DOI: 10.1109/MSP.2018.2867638.

[32] Benjamin Dunn, Maria Mørreaunet and Yasser Roudi. 'Correlations and Functional Connections in a Population of Grid Cells'. In: (Apr. 2014). DOI: 10.1371/journal.pcbi.1004052.

[33] Martin Eckstein et al. *Hopping on the Bethe Lattice: Exact Results for Densities of States and Dynamical Mean-Field Theory*. Tech. rep. 2005.

[34] T P Eggarter. *PHYSICAL RE VIEW B Cayley t Ees, the Ismg Problem, and the Thermodymtmc H˜t*. Tech. rep.

[35] L. El Ghaou. *GRAPHICAL MODEL OF SENATE VOTING. http://www.eecs.berkeley.edu/˜elghaoui/StatNews/Ex_senate.Html*.

[36] Paul Erd¨s and Alfréd Rènyi. 'On Random Graphs I'. In: *Publicationes Mathematicae* 6 (1959), pp. 290–297.

[37] Silvio Franz, Federico Ricci-Tersenghi and Jacopo Rocchi. *A Fast and Accurate Algorithm for Inferring Sparse Ising Models via Parameters Activation to Maximize the Pseudo-Likelihood*. Tech. rep. 2019.

[38] U. Frey et al. 'Microelectronic System for High-Resolution Mapping of Extracellular Electric Fields Applied to Brain Slices'. en. In: *Biosensors and Bioelectronics* 24.7 (Mar. 2009), pp. 2191–2198. ISSN: 0956-5663. DOI: 10.1016/j.bios.2008.11.028.

[39] Karl J. Friston. 'Functional and Effective Connectivity in Neuroimaging: A Synthesis'. en. In: *Human Brain Mapping* 2.1-2 (1994), pp. 56–78. ISSN: 1097-0193. DOI: 10.1002/hbm.460020107.

[40] Christophe Gardella, Olivier Marre and Thierry Morra. 'Modeling the Correlated Activity of Neural Populations: A Review'. In: 2733 (2018), pp. 2709–2733. DOI: 10.1162/NECO.

[41] Matteo Garofalo et al. 'Evaluation of the Performance of Information Theory-Based Methods and Cross-Correlation to Estimate the Functional Connectivity in Cortical Networks'. In: *PLoS ONE* 4.8 (Aug. 2009). Ed. by Olaf Sporns, e6482–e6482. DOI: 10.1371/journal.pone.0006482.

[42] Stuart Geman and Donald Geman. *Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images*. Tech. rep. 6. 1984, pp. 721–721.

[43] W R Gilkst, N G Best and K K C Tan. *Adaptive Rejection Metropolis Sampling within Gibbs Sampling*. Tech. rep. 4. 1995, pp. 455–472.

[44] Iris Ginzburg and Haim Sompolinsky. 'Theory of Correlations in Stochastic Neural Networks'. In: *Physical Review E* 50 (1994), pp. 3171–3191.

[45] Roy J. Glauber. 'Time-Dependent Statistics of the Ising Model'. In: *Journal of Mathematical Physics* 4.2 (Feb. 1963), pp. 294–307. DOI: 10.1063/1.1703954.

[46] A. Grabowski and R. A. Kosiński. 'Ising-Based Model of Opinion Formation in a Complex Network of Interpersonal Interactions'. en. In: *Physica A: Statistical Mechanics and its Applications* 361.2 (Mar. 2006), pp. 651–664. ISSN: 0378-4371. DOI: 10.1016/j.physa.2005.06.102.

[47] Peter Grunwald. 'A Tutorial Introduction to the Minimum Description Length Principle'. In: *arXiv:math/0406077* (June 2004). arXiv: math/0406077.

[48] Peter Grünwald and Teemu Roos. *Minimum Description Length Revisited*. Tech. rep. 2019.

[49] Kristine Heiney et al. 'Assessment and Manipulation of the Computational Capacity of in Vitro Neuronal Networks through Criticality in Neuronal Avalanches'. In: *2019 IEEE Symposium Series on Computational Intelligence, SSCI 2019*. Institute of Electrical and Electronics Engineers Inc., Dec. 2019, pp. 247–254. ISBN: 978-1-72812-485-8. DOI: 10.1109/SSCI44817.2019.9002693. arXiv: 1907.13118.

[50] John Hertz, Yasser Roudi and Joanna Tyrcha. 'Ising Models for Inferring Network Structure from Spike Data'. In: (June 2011). arXiv: 1106.1752.

[51] John A. Hertz, Yasser Roudi and Joanna Tyrcha. 'Ising Models for Inferring Network Structure from Spike Data'. In: *Principles of Neural Coding* (2013), pp. 527–546. ISSN: 9781439853313. DOI: 10.1201/b14756.

[52] Henry Hexmoor. 'Chapter 1 - Ubiquity of Networks'. In: *Computational Network Science*. Ed. by Henry Hexmoor. Emerging Trends in Computer Science and Applied Computing. Boston: Morgan Kaufmann, 2015, pp. 1–14. ISBN: 978-0-12-800891-1. DOI: 10.1016/B978-0-12-800891-1.00001-9.

[53] J J Hopfield. *Neural Networks and Physical Systems with Emergent Collective Computational Abilities (Associative Memory/Parallel Processing/Categorization/Content-Addressable Memory/Fail-Soft Devices)*. Tech. rep. 1982, pp. 2554–2558.

[54] J. J. Hopfield et al. 'Neural Networks and Physical Systems with Emergent Collective Computational Abilities.' In: *Proceedings of the National Academy of Sciences* 79.8 (Apr. 1982), pp. 2554–2558. DOI: 10.1073/pnas.79.8.2554.

[55] John J Hopfield and David W Tank. '"Neural" Computation of Decisions in Optimization Problems'. In: *Biological cybernetics* 52.3 (1985), pp. 141–152.

[56] John J Hopfield and David W Tank. 'Computing with Neural Circuits: A Model'. In: *Science* 233.4764 (1986), pp. 625–633.

[57] Sheng You Huang and Xiaoqin Zou. 'Statistical Mechanics-Based Method to Extract Atomic Distance-Dependent Potentials from Protein Structures'. In: *Proteins: Structure, Function and Bioinformatics* 79.9 (Sept. 2011), pp. 2648–2661. DOI: 10.1002/prot.23086.

[58] Jacques Kotze. *Introduction to Monte Carlo Methods for an Ising Model of a Ferromagnet*. Tech. rep. 2008.

[59] E. T. Jaynes. 'Information Theory and Statistical Mechanics'. In: *Physical Review* 106.4 (May 1957), pp. 620–630. DOI: 10.1103/PhysRev.106.620.

[60] James J. Jun et al. 'Fully Integrated Silicon Probes for High-Density Recording of Neural Activity'. en. In: *Nature* 551.7679 (Nov. 2017), pp. 232–236. ISSN: 1476-4687. DOI: 10.1038/nature24636.

[61] Andrej Karpathy. *Gibbs Sampling on Ising Model*. https://cs.stanford.edu/people/karpathy/visml/ising_exampl

[62] Zoran Konkoli et al. 'Reservoir Computing with Computational Matter'. In: *Natural Computing Series*. Springer Verlag, 2018, pp. 269–293. DOI: 10.1007/978-3-319-65826-1_14.

[63] BY László Erd et al. 'SPECTRAL STATISTICS OF ERD OS-RÉNYI GRAPHS I: LOCAL SEMICIRCLE LAW'. In: *The Annals of Probability* 41.3B (2013), pp. 2279–2375. DOI: 10.1214/11-AOP734.

[64] Yann LeCun, John Denker and Sara Solla. 'Optimal Brain Damage'. In: *Advances in neural information processing systems* 2 (1989), pp. 598–605.

[65] Timothy R. Lezon et al. 'Using the Principle of Entropy Maximization to Infer Genetic Interaction Networks from Gene Expression Patterns'. In: *Proceedings of the National Academy of Sciences of the United States of America* 103.50 (Dec. 2006), pp. 19033–19038. DOI: 10.1073/pnas.0609152103.

[66] Andrey Y. Lokhov et al. 'Optimal Structure and Parameter Learning of Ising Models'. In: *Science Advances* 4.3 (Mar. 2018). DOI: 10.1126/sciadv.1700791.

[67] Chun Yan Luo et al. 'Functional Connectome Assessed Using Graph Theory in Drug-Naive Parkinson's Disease'. en. In: *Journal of Neurology* 262.6 (June 2015), pp. 1557–1567. ISSN: 1432-1459. DOI: 10.1007/s00415-015-7750-3.

[68] David J C Mackay. 'A Practical Bayesian Framework'. In: 472.1 (1992), pp. 448–472.

[69] David J. C. MacKay. *Information Theory, Inference and Learning Algorithms*. Vol. 100. Cambridge, 2008. DOI: 10.2277/0521642981.

[70] Ildefons Magrans de Abril, Junichiro Yoshimoto and Kenji Doya. 'Connectivity Inference from Neural Recording Data: Challenges, Mathematical Bases and Research Directions'. In: *Neural Networks* 102 (June 2018), pp. 120–137. DOI: 10.1016/j.neunet.2018.02.016.

[71] Matteo Marsili, Iacopo Mastromatteo and Yasser Roudi. 'On Sampling and Modeling Complex Systems'. In: *Journal of Statistical Mechanics: Theory and Experiment* 2013.09 (Sept. 2013), P09003–P09003.

[72] Luca Martino, Jesse Read and David Luengo. 'Independent Doubly Adaptive Rejection Metropolis Sampling within Gibbs Sampling'. In: *IEEE Transactions on Signal Processing* (2015). DOI: 10.1109/TSP.2015.2420537.

[73] Renate Meyer, Bo Cai and François Perron. 'Adaptive Rejection Metropolis Sampling Using Lagrange Interpolation Polynomials of Degree 2'. In: *Computational Statistics and Data Analysis* (2008). DOI: 10.1016/j.csda.2008.01.005.

[74] Stanley Milgram. 'The Small World Problem'. In: *Psychology today* 2.1 (1967), pp. 60–67.

[75] Andrea Montanari and Jose Pereira. 'Which Graphical Models Are Difficult to Learn?' In: *Advances in Neural Information Processing Systems* 22 (2009), pp. 1303–1311.

[76] Thierry Mora and William Bialek. 'Are Biological Systems Poised at Criticality?' In: *J Stat Phys* 144 (2011), pp. 268–302. DOI: 10.1007/s10955-011-0229-4.

[77] Kevin P. Murphy. *Machine Learning: A Probabilistic Perspective*. en. Adaptive Computation and Machine Learning Series. Cambridge, MA: MIT Press, 2012. ISBN: 978-0-262-01802-9.

[78] In Jae Myung, Vijay Balasubramanian and Mark A. Pitt. 'Counting Probability Distributions: Differential Geometry and Model Selection'. In: *Proceedings of the National Academy of Sciences of the United States of America* 97.21 (Oct. 2000), pp. 11170–11175. DOI: 10.1073/pnas.170283897.

[79] Hiroya Nakao and Alexander S Mikhailov. 'Turing Patterns in Network-Organized Activator–Inhibitor Systems'. In: (2010). DOI: 10.1038/NPHYS1651.

[80] H. Chau Nguyen, Riccardo Zecchina and Johannes Berg. 'Inverse Statistical Problems: From the Inverse Ising Problem to Data Science'. In: *Advances in Physics* 66.3 (July 2017), pp. 197–261. DOI: 10.1080/00018732.2017.1341604.

[81] Tomoyuki Obuchi, Simona Cocco and Rémi Monasson. 'Learning Probabilities from Random Observables in High Dimensions: The Maximum Entropy Distribution and Others'. In: *Journal of Statistical Physics* 161.3 (Mar. 2015), pp. 598–632. DOI: 10.1007/s10955-015-1341-7.

[82] M Ostilli. *Cayley Trees and Bethe Lattices, a Concise Analysis for Mathematicians and Physicists*. Tech. rep. 2012.

[83] Dmitry Panchenko. 'Introduction to the SK Model'. In: *Current Developments in Mathematics* 2014.1 (Nov. 2014), pp. 231–291.

[84] Thomas Parr, Noor Sajid and Karl J. Friston. 'Modules or Mean-Fields?' In: *Entropy* 22.5 (2020), pp. 1–25. ISSN: 10994300. DOI: 10.3390/E22050552.

[85] Johan Pensar et al. 'High-Dimensional Structure Learning of Binary Pairwise Markov Networks: A Comparative Numerical Study'. en. In: *arXiv:1901.04345 [cs, stat]* (July 2019). DOI: 10.1016/j.csda.2019.06.012. arXiv: 1901.04345 [cs, stat].

[86] Johan Pensar et al. *High-Dimensional Structure Learning of Binary Pairwise Markov Networks: A Comparative Numerical Study*. Tech. rep.

[87] Johan Pensar et al. 'Marginal Pseudo-Likelihood Learning of Discrete Markov Network Structures'. In: *Bayesian Analysis* 12.4 (2017), pp. 1195–1215. DOI: 10.1214/16-BA1032.

[88] T Plefka. 'Convergence Condition of the TAP Equation for the Infinite-Ranged Ising Spin Glass Model'. In: *Journal of Physics A: Mathematical and General* 15.6 (June 1982), pp. 1971–1978. ISSN: 0305-4470, 1361-6447. DOI: 10.1088/0305-4470/15/6/035.

[89] Daniele Poli, Vito P. Pastore and Paolo Massobrio. 'Functional Connectivity in in Vitro Neuronal Assemblies'. In: *Frontiers in Neural Circuits* 9.October (2015), pp. 1–14. ISSN: 1662-5110. DOI: 10.3389/fncir.2015.00057.

[90] Sidney Pontes-Filho et al. 'A General Representation of Dynamical Systems for Reservoir Computing'. In: (July 2019). arXiv: 1907.01856.

[91] Sidney Pontes-Filho et al. 'A Neuro-Inspired General Framework for the Evolution of Stochastic Dynamical Systems: Cellular Automata, Random Boolean Networks and Echo State Networks towards Criticality'. In: *Cognitive Neurodynamics* 14.5 (Oct. 2020), pp. 657–674. ISSN: 18714099. DOI: 10.1007/s11571-020-09600-x.

[92] James Gary Propp and David Bruce Wilson. 'Exact Sampling with Coupled Markov Chains and Applications to Statistical Mechanics'. en. In: *Random Structures & Algorithms* 9.1-2 (1996), pp. 223–252. ISSN: 1098-2418. DOI: 10.1002/(SICI)1098-2418(199608/09)9:1/2⟨223::AID-RSA14⟩3.0.CO;2-O.

[93] Jason L. Puchalla et al. 'Redundancy in the Population Code of the Retina'. eng. In: *Neuron* 46.3 (May 2005), pp. 493–504. ISSN: 0896-6273. DOI: 10.1016/j.neuron.2005.03.026.

[94] Pradeep Ravikumar, Martin J Wainwright and John D Lafferty. 'High-Dimensional Ising Model Selection Using L1-Regularized Logistic Regression'. In: *The Annals of Statistics* 38.3 (2010), pp. 1287–1319. DOI: 10.1214/09-AOS691.

[95] Federico Ricci-Tersenghi. *The Bethe Approximation for Solving the Inverse Ising Problem: A Comparison with Other Inference Methods*. Tech. rep. 2012.

[96] Jorma Rissanen. 'Stochastic Complexity and the MDL Principle'. In: *Econometric Reviews* 6.1 (1987), pp. 85–102. DOI: 10.1080/07474938708800126.

[97] Jorma J Rissanen. 'Fisher Information and Stochastic Complexity'. In: 42.1 (1996), pp. 40–47.

[98] Yasser Roudi, Erik Aurell and John A. Hertz. 'Statistical Physics of Pairwise Probability Models'. In: *Frontiers in Computational Neuroscience* 3.NOV (2009), pp. 1–15. DOI: 10.3389/neuro.10.022.2009.

[99] Yasser Roudi, Benjamin Dunn and John Hertz. 'Multi-Neuronal Activity and Functional Connectivity in Cell Assemblies'. In: *Current Opinion in Neurobiology* 32 (2015), pp. 38–44. ISSN: 0959-4388. DOI: 10.1016/j.conb.2014.10.011.

[100] Yasser Roudi, Sheila Nirenberg and Peter E. Latham. 'Pairwise Maximum Entropy Models for Studying Large Biological Systems: When They Can Work and When They Can't'. In: *PLoS Computational Biology* 5.5 (May 2009), pp. 1000380–1000380. DOI: 10.1371/journal.pcbi.1000380.

[101] Yasser Roudi, Joanna Tyrcha and John Hertz. 'Ising Model for Neural Data: Model Quality and Approximate Methods for Extracting Functional Connectivity'. In: (). DOI: 10.1103/PhysRevE.79.051915.

[102] Yasser Roudi, Joanna Tyrcha and John Hertz. *The Ising Model for Neural Data: Model Quality and Approximate Methods for Extracting Functional Connectivity*. Tech. rep. 2009.

[103] Mikail Rubinov and Olaf Sporns. 'Complex Network Measures of Brain Connectivity: Uses and Interpretations'. en. In: *NeuroImage*. Computational Models of the Brain 52.3 (Sept. 2010), pp. 1059–1069. ISSN: 1053-8119. DOI: 10.1016/j.neuroimage.2009.10.003.

[104] Elad Schneidman et al. 'Weak Pairwise Correlations Imply Strongly Correlated Network States in a Neural Population'. In: *Nature* 440.7087 (2006), pp. 1007–1012. ISSN: 1476-4687 (Electronic). DOI: 10.1038/nature04701.

[105] Elad Schneidman et al. 'Weak Pairwise Correlations Imply Strongly Correlated Network States in a Neural Population'. In: *Nature* 440.7087 (Apr. 2006), pp. 1007–1012. DOI: 10.1038/nature04701.

[106] Benjamin Schrauwen, David Verstraeten and Jan Van Campenhout. 'An Overview of Reservoir Computing: Theory, Applications and Implementations'. In: *Proceedings of the 15th European Symposium on Artificial Neural Networks. p. 471-482 2007*. 2007, pp. 471–482.

[107] Gideon Schwarz. *Estimating the Dimension of a Model*. Tech. rep. 2. 1978, pp. 461–464.

[108] Magnus Själander et al. 'EPIC: An Energy-Efficient, High-Performance GPGPU Computing Research Infrastructure'. In: *CoRR* abs/1912.05848 (2019).

[109] Olaf Sporns. 'The Human Connectome: A Complex Network'. In: *Annals of the new york academy of sciences* 1224.1 (Apr. 2011), pp. 109–125. ISSN: 17496632. DOI: 10.1111/j.1749-6632.2010.05888.x.

[110] Olaf Sporns and Giulio Tononi. 'Classes of Network Connectivity and Dynamics'. In: *Complexity* 7.1 (Sept. 2001), pp. 28–38. ISSN: 1076-2787. DOI: 10.1002/cplx.10015.

[111] Olaf Sporns and Giulio Tononi. 'Classes of Network Connectivity and Dynamics'. In: *Complexity* 7.1 (Sept. 2001), pp. 28–38. ISSN: 1076-2787. DOI: 10.1002/cplx.10015.

[112] Nicholas A. Steinmetz et al. 'Challenges and Opportunities for Large-Scale Electrophysiology with Neuropixels Probes'. In: *Current Opinion in Neurobiology* 50 (2018), pp. 92–100. DOI: 10.1016/j.conb.2018.01.009.

[113] Toshiyuki Tanaka. *Mean-Field Theory of Boltzmann Machine Learning*. Tech. rep. 1998.

[114] D. J. Thouless, P. W. Anderson and R. G. Palmer. 'Solution of 'Solvable Model of a Spin Glass''. en. In: *Philosophical Magazine* 35.3 (Mar. 1977), pp. 593–601. ISSN: 0031-8086. DOI: 10.1080/14786437708235992.

[115] Gasper Tkacik et al. 'Spin Glass Models for a Network of Real Neurons'. In: (Dec. 2009).

[116] Gasper Tkacik et al. 'Thermodynamics for a Network of Neurons: Signatures of Criticality'. In: *Proceedings of the National Academy of Sciences* 112.37 (July 2014), pp. 11508–11513.

[117] Vibeke Devold Valderhaug et al. 'Criticality as a Measure of Developing Proteinopathy in Engineered Human Neural Networks'. In: *bioRxiv* (May 2020), p. 2020.05.03.074666. DOI: 10.1101/2020.05.03.074666.

[118] Vibeke Devold Valderhaug et al. 'Formation of Neural Networks with Structural and Functional Features Consistent with Small-World Network Topology on Surface-Grafted Polymer Particles'. In: *Royal Society Open Science* 6.10 (Oct. 2019), p. 191086. ISSN: 2054-5703. DOI: 10.1098/rsos.191086.

[119] Vibeke Devold Valderhaug et al. 'Structural and Functional Alterations Associated with the LRRK2 G2019S Mutation Revealed in Structured Human Neural Networks'. In: *bioRxiv* (May 2020), p. 2020.05.02.073726. DOI: 10.1101/2020.05.02.073726.

[120] Ludwig Von Bertalanffy, George Braziller and New York. *General System Theory Foundations, Development, Applications Revised Edition*. Tech. rep.

[121] Duncan J. Watts and Steven H. Strogatz. 'Collective Dynamics of 'Small-World' Networks'. en. In: *Nature* 393.6684 (June 1998), pp. 440–442. ISSN: 1476-4687. DOI: 10.1038/30918.

[122] Shan Yu et al. 'A Small World of Neuronal Synchrony'. In: *Cerebral Cortex December* 18 (2008), pp. 2891–2901. DOI: 10.1093/cercor/bhn047.

Appendix

A Additional Figures

Figure A1 compares the ROC (Equation 47) for multiple topologies in the small network regime and gives an overall picture of the criteria performance across the regimes. We see how the MDLent and BIC performance in the CT remains largely static except in the lowest beta and sample rate. There is also a similar pattern but clear difference between performances in the CT topology and the other two topologies which show very similar scores.

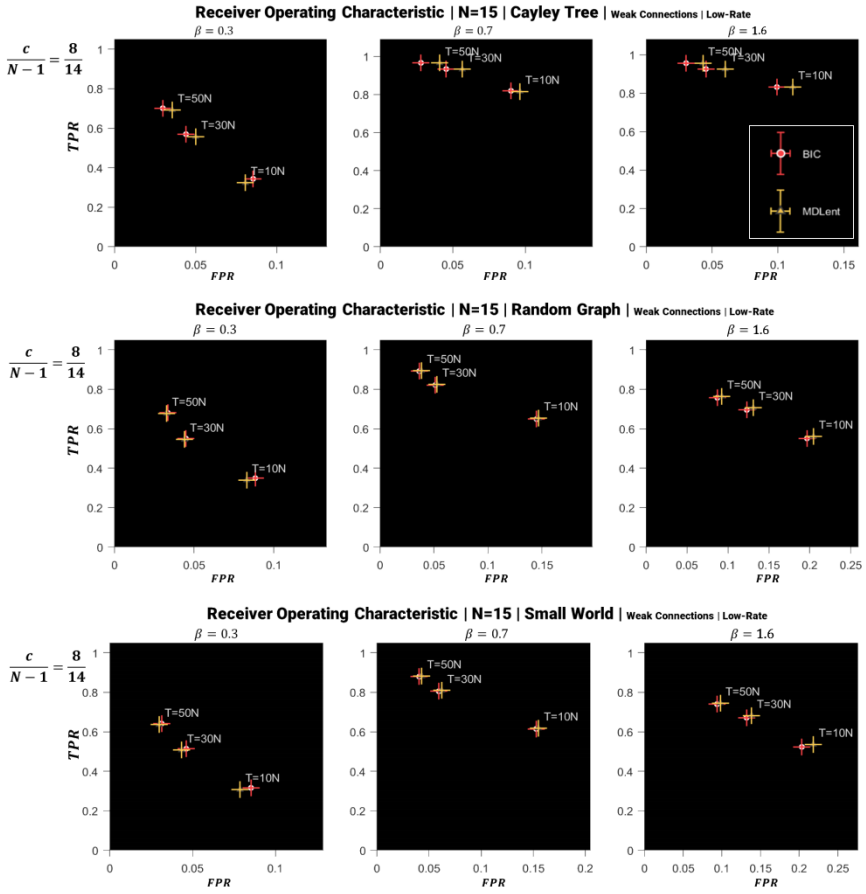


Figure A1
Receiver Operating Characteristic (ROC (Equation 47)) for multiple topologies at same coordination number ($\mathcal{C} = 8$ and size ($N = 15$) in the low-rate regime. Plot of the True Positive Rate over the False Positive Rate at multiple sample rates ($T = 10N, 30N, 50N$). Plots are grouped in columns by $\beta = [0.3, 0.7, 1.6]$.

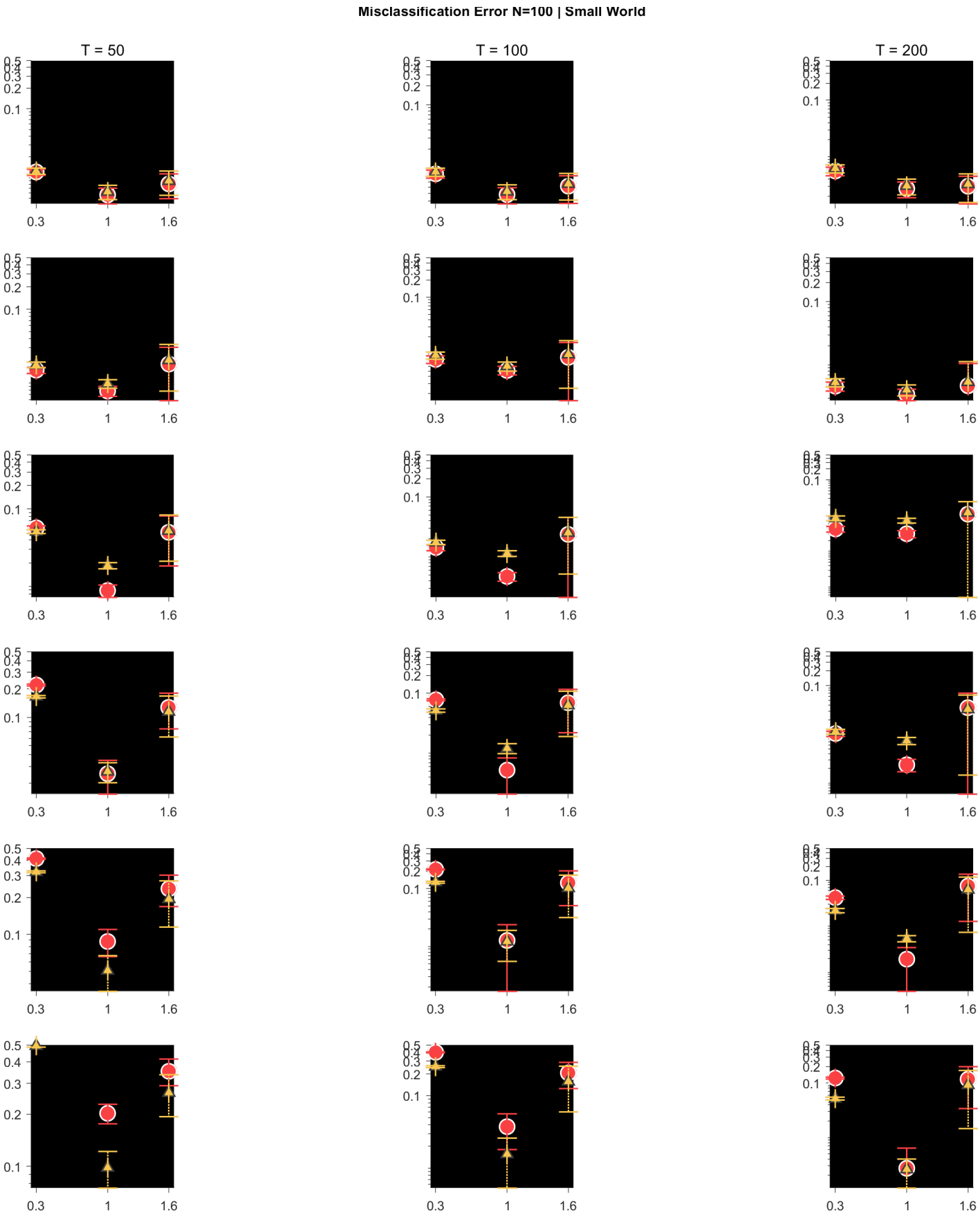


Figure A2
Misclassification Error for $N = 100$ at all densities (by rows descending $\mathcal{C} = 4, 12, 28, 50, 70, 90$) and multiple inputs of the inverse temperature ($\beta = 0.3, 1, 1.6$)

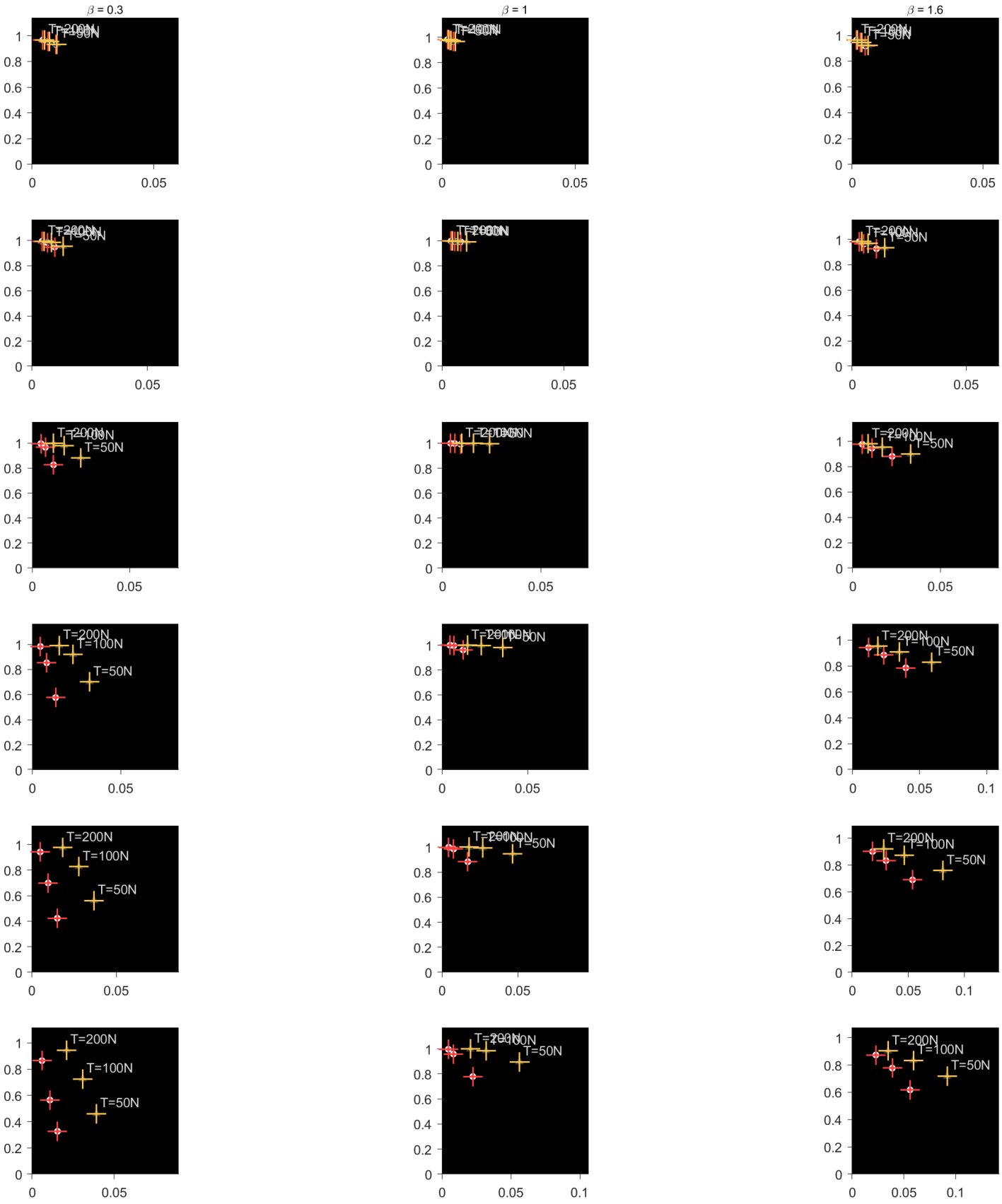


Figure A3
ROC (Equation 47) figure for large network $N = 100$ at all densities (by rows descending $\mathcal{C} = 4, 12, 28, 50, 70, 90$) and multiple inputs of the inverse temperature ($\beta = 0.3, 1, 1.6$)

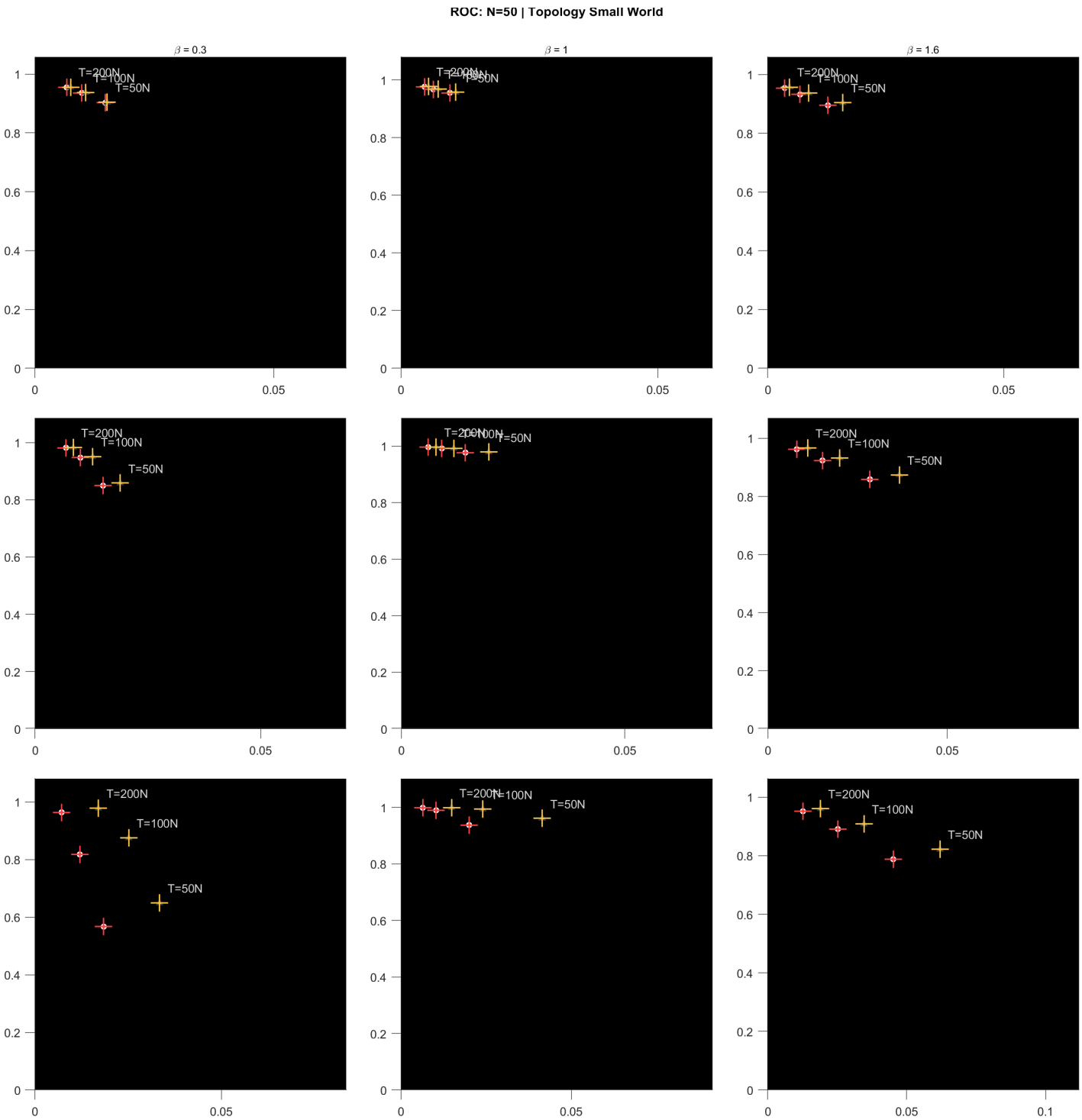


Figure A4
ROC (Equation 47) $N = 50$ regime at $\mathcal{C} = 4, 12, 28$

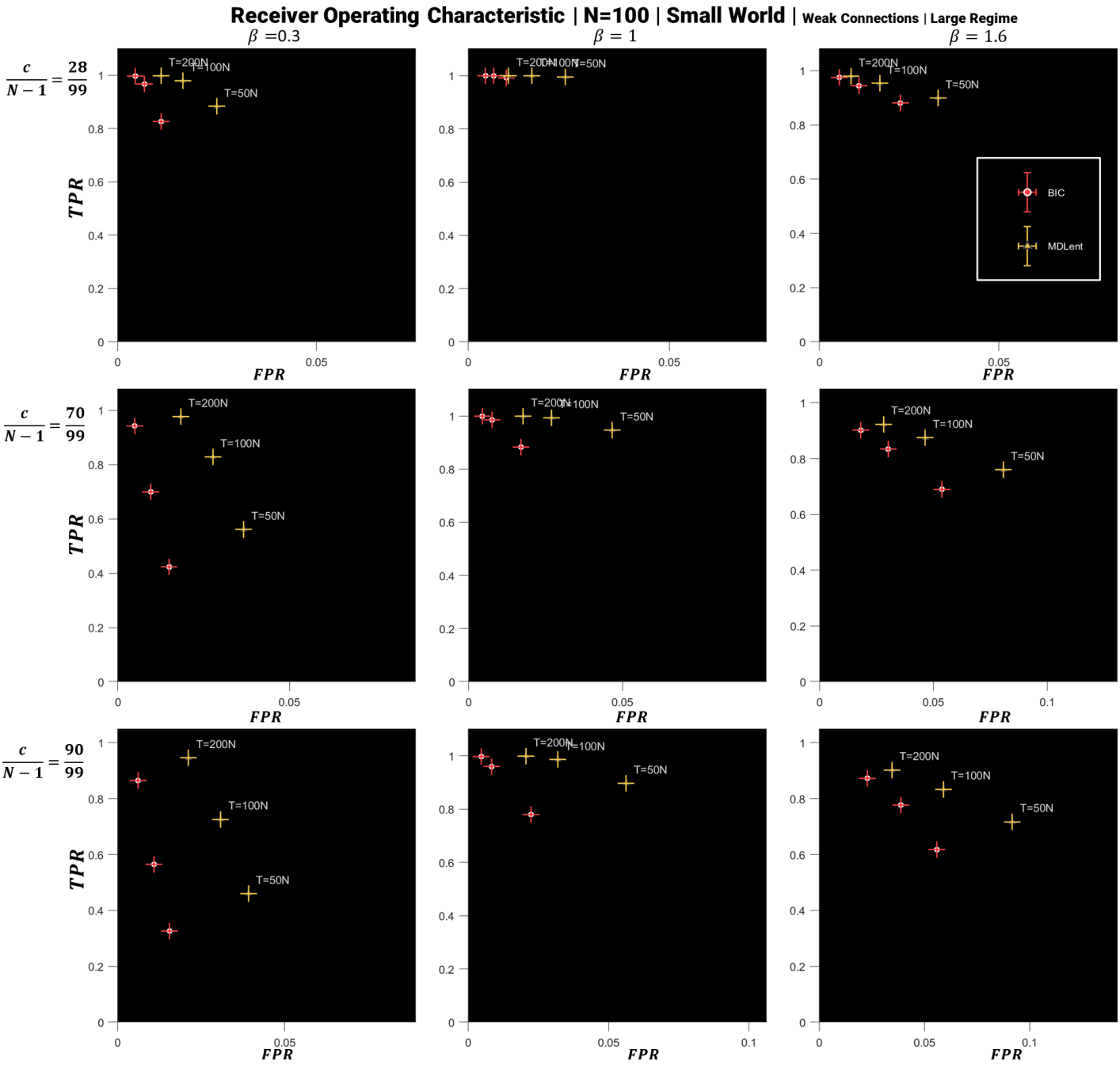


Figure A5
ROC (Equation 47) for $N = 100$ regime at $\mathcal{C} = 28, 70, 90$ The ROC scores overall appear the same to those in the smaller network $N = 50$ (fig. 36 or (fig **??**) for version with exact same T values), both criteria return similar TPR at similar density and sample rates and the MDLent FPR appears to slightly increase in the larger network size. We do get to see in this large regime the performance at various levels of density. Figure for other levels of density ($\mathcal{C} = 4, 12, 28, 50, 70, 90$ by rows descending) can be found in the appendix, A3 .

B Sample Code

Full code base can be found in my github repository: https://github.com/MichaTarlton/Inv.Is.Models.git

B.1 MATLAB main script

```
%%fi150121.m
%% Model Selection of Ising Model
%% Nicola Bulso is largely to thank for this code

%% Inputs:
% Jobname: Name of current batch job submitted to IDUN
% Intbeta: Realization of the experimental regime, passed in by the SLURM array
    ↪ number


%% Outs
% Multiallstruct: contains the results of this realization of the regimes
% Parameters
% Statvecs



%clear all;
function fi1501(jobname,intbeta)

totaltime = tic;

disp(['Beta number: ',num2str(intbeta)])

%%% File Structure and Storage
%cd(cd)
%addpath(genpath('E:\GitHub\Inv.Is.Models\Mike_Code_4'));
%savepath

addpath(genpath('/lustre1/home/michaeta/Mike_Code_4'));
cd('/lustre1/home/michaeta');

%%% Storage

time = datestr(now,'HHMM-ddmmmyy');
disp(time)

%% for rng
hpct=clock();

seed=hpct(6) * 1000; % Seed with the second part of the clock array.

rng(seed);

%rng('shuffle','philox')
%rndy1 = num2str(randi([1 99],1))
%rng('shuffle','philox')
%rndy2 = num2str(randi([1 99],1))
%rng('shuffle','philox')
%rndy3 = num2str(randi([1 99],1))
```

```
%rng('shuffle','philox')
%rndy4 = num2str(randi([1 999],1))

rndy = num2str(randi([100 999],1))
%dirname = [time,'-',rndy]
dirname = [time,'-',num2str(jobname),'_',num2str(intbeta)]

mkdir(cd,dirname);
cd(dirname);
disp(cd)

topdir = cd;


%%%% Parameters

    %% Trials
    jn = 100; %| number of Trials

    h_on = 0 % h field genereation

    Tvec = [10,30,50,100,200]

    Nvec = [100,200,300,400,500];

    betavecint = [0.3,0.5,0.7,1,1.3,1.6]

    %% Sparsity measure, used in old "sk" distribution method
    sprsvec = [0];
    sprs = 0; % only here as temp measure

    %% Coordination number
    %coordvec = [1,2,4,8,12,16];



    % select topology
    topovec = {1,3,4,5,6};

%%% for distribtions, see TCS.m for details
    %couplings = 4; % "SK" or Mike's Gaussian
    %couplings = 5; % double mean gauss
    %couplings = 1; %---Gaussian
    %couplings = 2; %---Delta Function
    couplings = 3; %---Double Delta Function
    J0 = 1; %---"The Mean"


%% For displaying and monitoring the number of trials that are being run
jta = 1; % Our measure of how many trials are run so far
jtatot = length(coordvec)*length(betavec)*length(Tvec)*length(Nvec)*length(
    ↪ topovec)*jn

runs = 1; % for indexing the trials ran for sprs, beta, T , N
          % keep out of the trials loop

OverStruct = struct;
```

```
OverStruct.Nvec  = Nvec ;
OverStruct.Tvec  = Tvec ;
OverStruct.betavec  = betavec ;
OverStruct.topologies = topovec;
OverStruct.jn   = jn  ;
OverStruct.sprsvec = sprsvec;
OverStruct.h_on  = h_on ;
OverStruct.topdir  = topdir ;
OverStruct.time  = time ;
save([overdir,'/',num2str(jobname),'_',num2str(intbeta),'-OverStruct.mat'],'
    ↪ OverStruct','-v7.3');



%create a local cluster object
%distcomp.feature( 'LocalUseMpiexec', false ) % highly experimental here
pc = parcluster('local')
%pc = parcluster('threads')
% explicitly set the Job Storage Location to the temp directory that was created
    ↪ in your sbatch script
mkdir(cd,'scratch')
parscratch = [topdir,'/scratch']
pc.JobStorageLocation = parscratch
parpool(pc,20)


for Ti = 1:length(Tvec)


    tic
    for Ni = 1:length(Nvec)
        N = Nvec(Ni);
        T = Tvec(Ti).*N;

        %for Si = 1:length(sprsvec)
        for Ci = 1:length(coordvec)

            %sprs = sprsvec(Si);
            c = coordvec(Ci);


            for Bi = 1:length(betavec)

                beta = betavec(Bi);

                cd(topdir)


               name = ['T',num2str(Ti),'N',num2str(N),'St',num2str(Ci),'Bt',
                   ↪ num2str(Bi)];


                OverStruct.list(runs).name = name;
                OverStruct.list(runs).T = T;
                OverStruct.list(runs).N = N;
                OverStruct.list(runs).beta = beta;
                OverStruct.list(runs).sprsvec = sprsvec;
                OverStruct.list(runs).coordvec = coordvec;
```

```
OverStruct.list(runs).topology = topovec;

OverStruct.list(runs).topology = topovec;
OverStruct.list(runs).couplings = couplings;
OverStruct.list(runs).c = c;

for tp = 1:length(topovec)

    topo = topovec{tp};

    %%%Forward Ising Topologies and Distributions
    JHnorm = struct;

    parfor trn = 1:jn
        % call rng for reproducibility
        rng(trn);

        [Adj,J,hfield] = TCS2(tp,N,c,couplings,beta,J0,sprs,h_on);
        JHnorm(trn).Adjset= Adj;
        JHnorm(trn).Jtopo = J;
        JHnorm(trn).Htopo = hfield;

    end

    OverStruct.list(runs).Jcontru(tp).topo = {JHnorm.Adjset};
    OverStruct.list(runs).Jtru(tp).topo = {JHnorm.Jtopo};
    OverStruct.list(runs).htru(tp).topo = {JHnorm.Htopo};


    %%% Part 2, generate samples (or spike train) S_hat
   SStruct = Met_Hast_norm(T,N,jn,JHnorm,sprs,time,beta);

    %%% Part 3, inference and model select
    %% Bulso Likelihood Estimator

    [LLH,statvecs,stats,jta] = PBLLH4(T,N,tp,beta,c,h_on,SStruct,
        ↪ JHnorm,jta,jtatot);

    %OverStruct.list(runs).BLLH(tp).topo = LLH;
    OverStruct.list(runs).BLLH(tp).statvecs = statvecs;
    OverStruct.list(runs).BLLH(tp).stats = stats;



    %seed=hpct(6) * 1000; % Seed with the second part of the clock
        ↪ array.

    %rng(seed);
    %rndy2 = num2str(randi([100 999],1))
    overdir = [time(1:12),'-OverStructs_',num2str(intbeta)];
    mkdir(cd,overdir);
    disp(overdir)




end
```

```
                runs = runs + 1; %stays out of trial loop, for measuring the runs
                    ↪ per other parameters
            end


        end
    end
    toc
end
delete(gcp('nocreate'))
save([overdir,'/',time(1:12),'-OverStruct_final.mat'],'OverStruct','-v7.3');
endtime = toc(totaltime);
disp(['Total Time: ', num2str(endtime./3600)])
end
% comp will set what tpe of figures we want
% comp = 1; % 1. Perconerr v beta
%
% figstor = multiallgraph3(OverStruct,comp,topdir,time);
% save([overdir,'\',time(1:12),'-figstor.mat'],'figstor','-v7.3');
```

B.2 SLURM batch job

```
#!/bin/bash
#SBATCH -J SWWC2212 # Sensible name for the job
#SBATCH -N 1
#SBATCH --account=mh-kin
#SBATCH -t 01-10:00:00 # Upper time limit for the job (DD-HH:MM:SS)
#SBATCH -p CPUQ
#SBATCH --mem=27G    # Set to 110g to secure a dedicated node, does affect
    ↪ priority queuing. Normally only need 27G
#SBATCH -c 20    # cores


SBATCH --array=1-135  # 135 different regimes
echo $SLURM_ARRAY_JOB_ID
echo $SLURM_ARRAY_TASK_ID
echo $PWD
SCRATCH_DIRECTORY=/home/michaeta/$SLURM_ARRAY_JOB_ID
mkdir -p $SCRATCH_DIRECTORY
echo $SCRATCH_DIRECTORY


module load MATLAB/2020b

matlab -nodisplay -nodesktop -nosplash -r "fi2212_SWWC1($SLURM_ARRAY_JOB_ID,
    ↪ $SLURM_ARRAY_TASK_ID)" ## Can't use -r if passing input into matlab
```

C Results on Symmetrized Graphs

Figures for the misclassification results returned by the pairwise form (symmetrized) of the MDLent reconstructed graphs for the small network, low-rate regimes. While the original reconstructed graph is asymmetrical, here we have applied a post-processing step to ensure the symmetry of the graph. This comes in two fashions of symmetrizing, applying a "generous" inclusive OR-gated function repairing any asymmetries $K_{ij} = K_{ji} = 1$, or a "conservative" AND-gated function pruning asymmetrical connections so that $K_{ij} = K_{ji} = 0$. Unfortunately this could not be fully-featured due to time constraints. Future experiments may want to reapply a model selection walk from the subspace of the symmetrized graph, as done in Pensar et al. 2017 [87]. Similar results were found in the large network regimes.

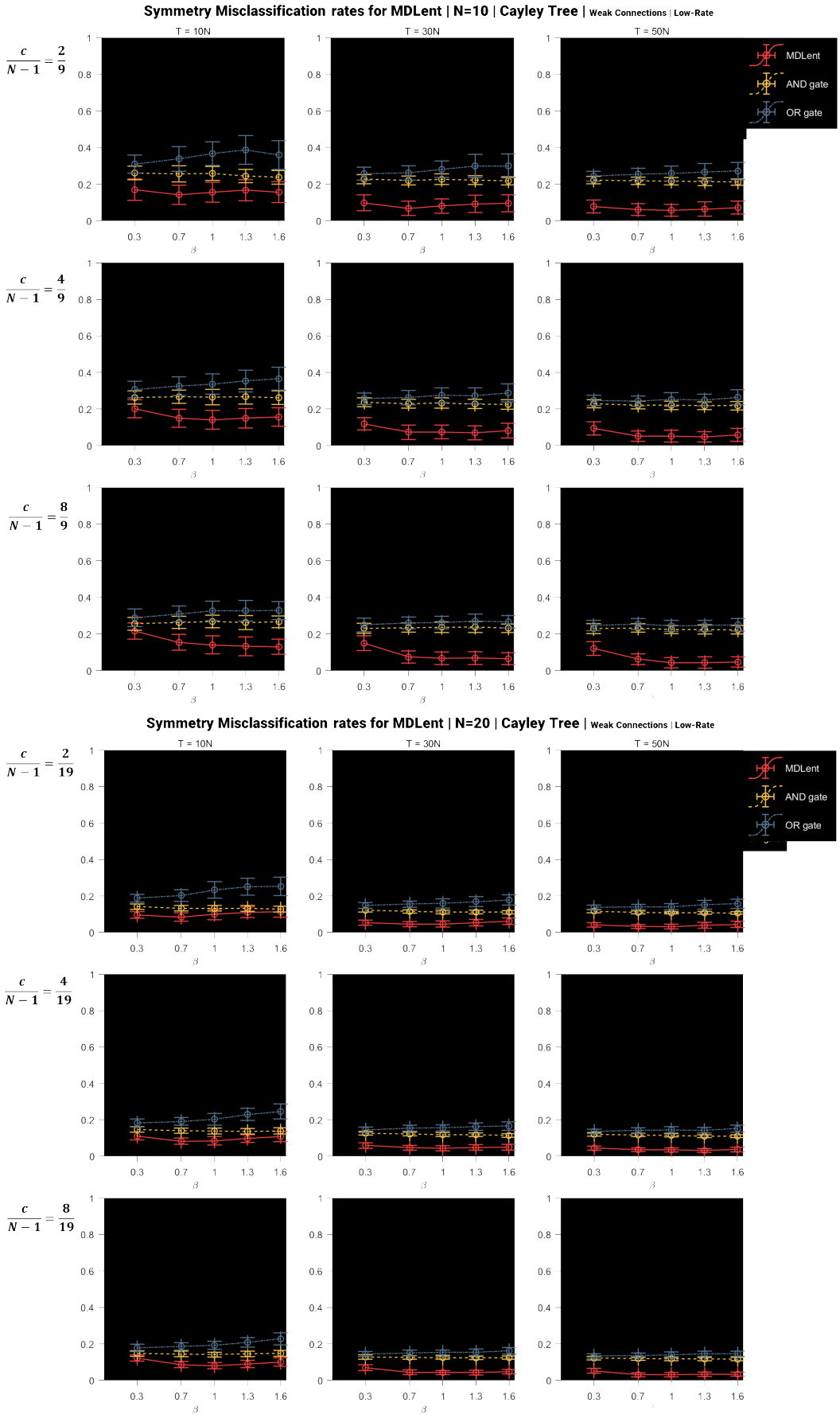


Figure C1
MDLent misclassification error of symmetric graphs for Cayley tree topology in small network regime at low sample rates. **(Top)** $N = 10$. **(Bottom)** $N = 20$. Rows are arranged by network density and columns by sample rate. Original misclassification error for MDLent reconstructed graph represented by the red line. AND-gated misclassification error represented by the yellow line. OR-gated graph misclassification error represented by the red line. Error bars represent standard deviations.

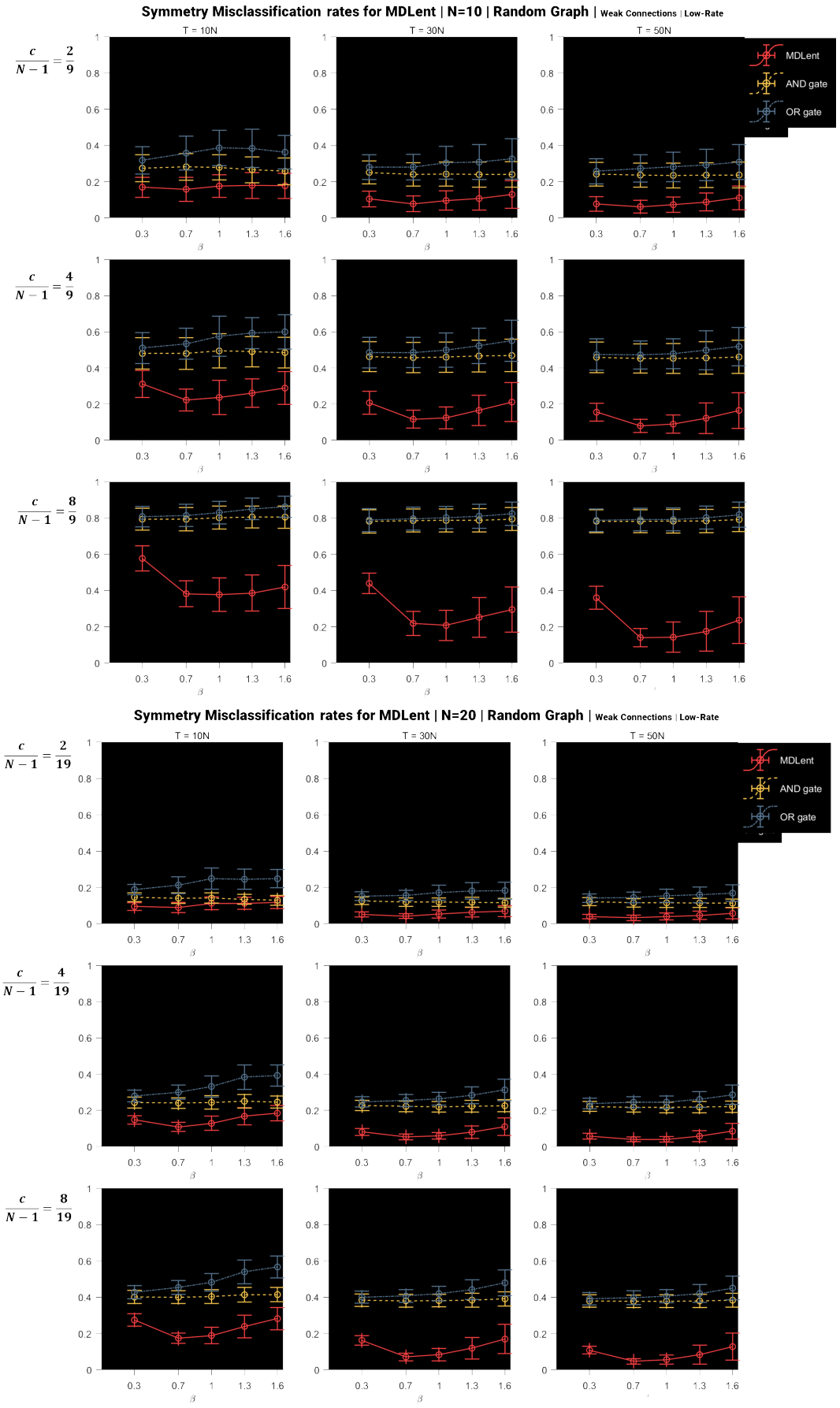


Figure C2
MDLent misclassification error of symmetric graphs for random graph topology in small network regime at low sample rates. **(Top)** $N = 10$. **(Bottom)** $N = 20$. Rows are arranged by network density and columns by sample rate. Original misclassification error for MDLent reconstructed graph represented by the red line. AND-gated misclassification error represented by the yellow line. OR-gated graph misclassification error represented by the red line. Error bars represent standard deviations.

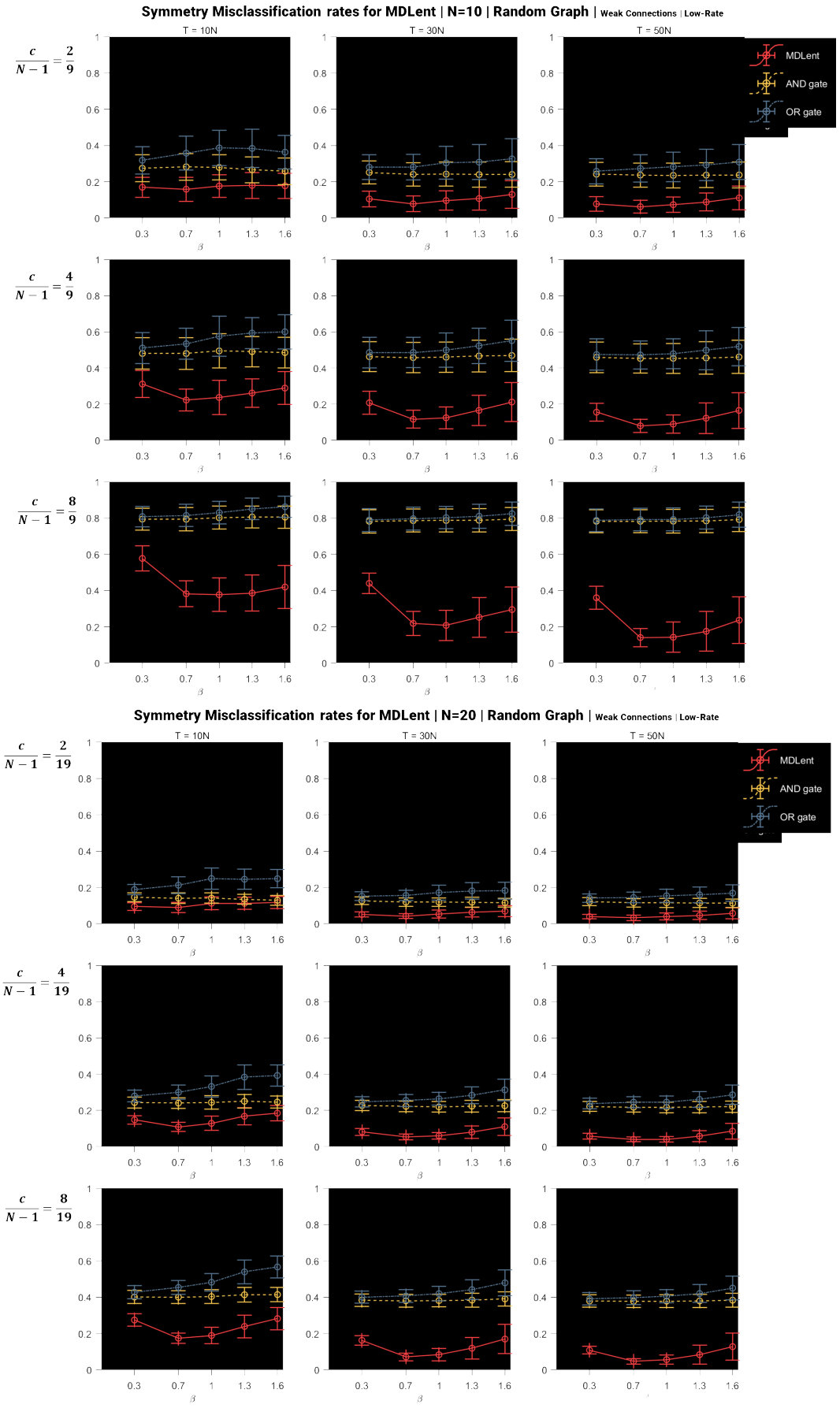
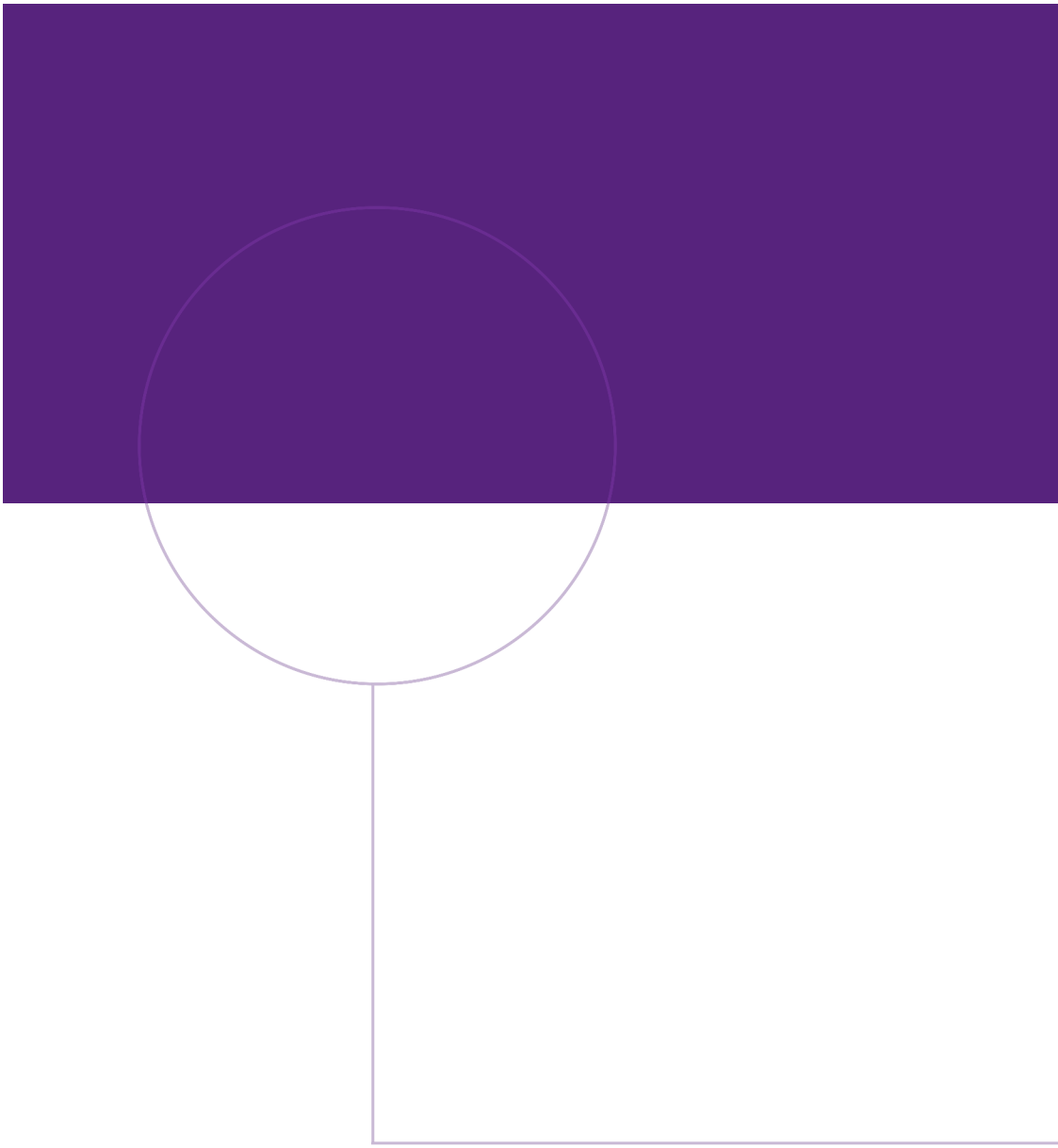


Figure C3
MDLent misclassification error of symmetric graphs for small world topology in small network regime at low sample rates. **(Top)** $N = 10$. **(Bottom)** $N = 20$. Rows are arranged by network density and columns by sample rate. Original misclassification error for MDLent reconstructed graph represented by the red line. AND-gated misclassification error represented by the yellow line. OR-gated graph misclassification error represented by the red line. Error bars represent standard deviations.



NTNU

Norwegian University of
Science and Technology