Simulation and theory of large-scale cortical networks

Alexander van Meegen

Information
Band / Volume 98
ISBN 978-3-95806-708-0

Mitglied der Helmholtz-Gemeinschaft

JÜLICH
Forschungszentrum

Schriften des Forschungszentrums Jülich
Reihe Information / Information Band / Volume 98

Forschungszentrum Jülich GmbH
Institute of Neurosciences and Medicine (INM)
Computational and Systems Neuroscience (INM-6)
& Theoretical Neuroscience (IAS-6)

Simulation and theory of large-scale cortical networks

Alexander van Meegen

Schriften des Forschungszentrums Jülich
Reihe Information / Information Band / Volume 98

ISSN 1866-1777 ISBN 978-3-95806-708-0

Bibliografische Information der Deutschen Nationalbibliothek.
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der
Deutschen Nationalbibliografie; detaillierte Bibliografische Daten
sind im Internet über http://dnb.d-nb.de abrufbar.

Herausgeber und Vertrieb:	Forschungszentrum Jülich GmbH Zentralbibliothek, Verlag 52425 Jülich Tel.: +49 2461 61-5368 Fax: +49 2461 61-6103 zb-publikation@fz-juelich.de www.fz-juelich.de/zb
Umschlaggestaltung:	Grafische Medien, Forschungszentrum Jülich GmbH
Druck:	Grafische Medien, Forschungszentrum Jülich GmbH
Copyright:	Forschungszentrum Jülich 2023

Schriften des Forschungszentrums Jülich
Reihe Information / Information, Band / Volume 98

D 38 (Diss. Köln, Univ., 2022)

ISSN 1866-1777
ISBN 978-3-95806-708-0

Vollständig frei verfügbar über das Publikationsportal des Forschungszentrums Jülich (JuSER)
unter www.fz-juelich.de/zb/openaccess.

Like the entomologist hunting for brightly coloured butterflies,
my attention was drawn to the flower garden of the grey matter,
which contained cells with delicate and elegant forms,
the mysterious butterflies of the soul,
the beating of whose wings may some day (who knows?)
clarify the secret of mental life.

— Santiago Ramón y Cajal

ABSTRACT

Cerebral cortex is composed of intricate networks of neurons. These neuronal networks are strongly interconnected: every neuron receives, on average, input from thousands or more presynaptic neurons. In fact, to support such a number of connections, a majority of the volume in the cortical gray matter is filled by axons and dendrites. Besides the networks, neurons themselves are also highly complex. They possess an elaborate spatial structure and support various types of active processes and nonlinearities. In the face of such complexity, it seems necessary to abstract away some of the details and to investigate simplified models.

In this thesis, such simplified models of neuronal networks are examined on varying levels of abstraction. Neurons are modeled as point neurons, both rate-based and spike-based, and networks are modeled as block-structured random networks. Crucially, on this level of abstraction, the models are still amenable to analytical treatment using the framework of dynamical mean-field theory.

The main focus of this thesis is to leverage the analytical tractability of random networks of point neurons in order to relate the network structure, and the neuron parameters, to the dynamics of the neurons—in physics parlance, to bridge across the scales from neurons to networks.

More concretely, four different models are investigated: 1) fully connected feedforward networks and vanilla recurrent networks of rate neurons; 2) block-structured networks of rate neurons in continuous time; 3) block-structured networks of spiking neurons; and 4) a multi-scale, data-based network of spiking neurons. We consider the first class of models in the light of Bayesian supervised learning and compute their kernel in the infinite-size limit. In the second class of models, we connect dynamical mean-field theory with large-deviation theory, calculate beyond mean-field fluctuations, and perform parameter inference. For the third class of models, we develop a theory for the autocorrelation time of the neurons. Lastly, we consolidate data across multiple modalities into a layer- and population-resolved model of human cortex and compare its activity with cortical recordings.

In two detours from the investigation of these four network models, we examine the distribution of neuron densities in cerebral cortex and present a software toolbox for mean-field analyses of spiking networks.

ZUSAMMENFASSUNG

Die Großhirnrinde besteht aus komplizierten Netzwerken von Neuronen. Diese neuronalen Netze sind stark untereinander verbunden: Jedes Neuron erhält im Durchschnitt Input von Tausenden oder mehr präsynaptischen Neuronen. Um eine solche Anzahl von Verbindungen zu ermöglichen, ist ein Großteil des Volumens der Substantia grisea mit Axonen und Dendriten gefüllt. Neben den Netzwerken sind auch die Neuronen selbst hochgradig komplex. Sie besitzen eine ausgefeilte räumliche Struktur und unterstützen verschiedene Arten von aktiven Prozessen und Nichtlinearitäten. Angesichts dieser Komplexität erscheint es notwendig, einige Details zu abstrahieren und vereinfachte Modelle zu untersuchen.

In dieser Arbeit werden solche vereinfachten Modelle neuronaler Netze auf verschiedenen Abstraktionsebenen untersucht. Neuronen werden als Punktneuronen modelliert, sowohl ratenbasiert als auch spikebasiert, und Netzwerke werden als blockstrukturierte Zufallsnetzwerke modelliert. Entscheidend ist, dass die Modelle auf dieser Abstraktionsebene immer noch einer analytischen Behandlung im Rahmen der dynamischen Mean-Field-Theorie zugänglich sind.

Das Hauptaugenmerk dieser Arbeit liegt darauf, die analytische Behandelbarkeit von Zufallsnetzwerken aus Punktneuronen zu nutzen, um die Netzwerkstruktur und die Neuronenparameter mit der Dynamik der Neuronen in Beziehung zu setzen - um in der Sprache der Physik eine Brücke von Neuronen zu Netzwerken zu schlagen.

Konkret werden vier verschiedene Modelle untersucht: 1) vollständig verbundene Feedforward-Netzwerke und einfache rekurrente Netzwerke von Ratenneuronen; 2) blockstrukturierte Netzwerke von Ratenneuronen in kontinuierlicher Zeit; 3) blockstrukturierte Netzwerke von spikenden Neuronen; und 4) ein multiskaliges, datenbasiertes Netzwerk von spikenden Neuronen. Wir betrachten die erste Klasse von Modellen im Lichte des überwachten Bayes'schen Lernens und berechnen ihren Kernel im Limes unendlicher Größe. Bei der zweiten Klasse von Modellen verbinden wir die dynamische Mean-Field-Theorie mit der Theorie der großen Abweichungen, berechnen Fluktuationen jenseits des Mean-Fields und führen eine Parameterinferenz durch. Für die dritte Klasse von Modellen entwickeln wir eine Theorie für die Autokorrelationszeit der Neuronen. Schließlich fassen wir Daten mehrerer Modalitäten zu einem schicht- und populationsaufgelösten Modell des menschlichen Kortex zusammen und vergleichen dessen Aktivität mit kortikalen Messungen.

In zwei Abstechern von der Untersuchung dieser vier Netzwerkmodelle untersuchen wir die Verteilung der Neuronendichte in der

Großhirnrinde und stellen eine Software-Toolbox für Mean-Field-Analysen von spikenden Netzwerken vor.

I enjoy the simplistic training and life in marathon.
You run, eat, sleep, walk around - that's how life is.
You don't get complicated.
The moment you get complicated it distracts your mind.

You cannot train alone and expect to run a fast time.
There is a formula: 100% of me is nothing compared to 1% of the whole team.
And that's teamwork. That's what I value.

— Eliud Kipchoge

ACKNOWLEDGMENTS

Hearing Kipchoge—the only human who managed to run the marathon distance below the magical 2h barrier—it seems that conducting research and running a marathon are not so different: in both cases, a lot of success is attributed to a single person while the truly important aspect is the teamwork; in both cases, it seems crucial to not get complicated because it distracts the mind. I'm extremely grateful to all of you who pushed me and kept my mind simple. In particular, I'd like to acknowledge...

...everyone who had to endure my supervision: Hannah Vollenbröker for never letting me loose sight of the brain amidst the jungle of formulas. Bastian Epping for a most convincing demonstration that supervising in the age of zoom can both be highly efficient and a lot of fun. Kai Segadlo for questioning every single detail and thereby of course uncovering mistakes. Jakob Stubenrauch for skillfully managing his supervisors, including myself, and countless discussions about fixed points and more.

...my peers at the INM-6 who were an integral part of my PhD: Anno Kurth for sheer endless discussions (and almost equally long runs, at least in my perception) which profoundly shaped every aspect of this thesis. Aitor Morales-Gregorio for joint projects in which we complemented each other perfectly (like two brushes in a car wash). Alessandra Stella for pistacchio, pizza, and point processes. Christian Keup for not believing a single word I say. Robin Gutzen for color consultations (I still use the blue you suggested). Michael Dick for our joint discovery of the difference between inputs and outputs, and for taking care of the PyMotW and the student representation. Kirsten Fischer for many valuable discussions about GPs, feedback, and more, and for taking care of the student representation. Jari Pronold for being a great travel companion along our journey towards a large-scale model (with a lot of huvi, of course). Agnes Korcsak-Gorzo for joining in on a deep dive into TensorFlow code written by clever people, and for taking care of the book club. Moritz Layer for

remaining inexplicably calm in the face of an endless review process with the worst possible timing (for himself). Javed Lindner for jointly chairing not one but two parallel sessions and discussions about GPs, random matrix theory and diagrams. Alexandre Rene for not being afraid to point out flaws and shortcomings. Since this is growing out of bounds, I'll resort to a plain, alphabetically-ordered list from here on: Jasper Albers, Jan Bauer, Younes Bouhadjar, Simon Essink, Han-Jia Jiang, Sandra Nestler, Tobias Schulte to Brinke, Joanna Senk, Renan Shimoura, Jonas Stapmanns, Lorenzo Tiberi, and Barna Zajzon. It is a privilege to work with you.

...the previous generation of PhD students on whose shoulders I stand: David Dahmen for many discussions and valuable hints. Tobias Kühn for kick-starting the large-deviation project. Pietro Quaglio for setting standards in simplicity and no-bullshit. Jakob Jordan for endless enthusiasm, helpful advice in the beginning, and ASPP. Julia Sprenger for encouraging me to annoy people.

...the wider community of computational neuroscience. All conferences and summer schools that I had the pleasure to attend were incredible experiences. I met old friends and made new ones amidst the scientific discussions. Furthermore, I'm deeply grateful to all of you who gave me valuable advise.

...those who supervised me (in a broad sense): Sacha van Albada for giving me a lot of freedom to shape my projects, for exposing me early on to conferences and summer schools, and for the unrepeatable feat (at least for me) of a red-line proof without red lines. Moritz Helias for unwavering belief in my abilities, for showing me the nuts and bolts of academia, and for providing an endless stream of inputs and ideas. Benjamin Lindner for the best possible preparation for a PhD. Johannes Berg and Martin Nawrot for agreeing to look into this thesis and supporting my tight timeline. Michael Krämer for an unparalleled title, and for allowing me a glimpse into actual physics. Günther Palm for discussing ergodic theory and the role of theorists in neuroscience. Rembrandt Bakker for teaching me the value of always visualizing the data. Sonja Grün and Markus Diesmann for the institute they continuously strive to make a great place for science, and for always being open to constructive feedback.

...my friends and family. Paul Hachmann for countless inputs outside of my comfort zone. Daniel Issing and Niclas Rieger for our joint exploration of the world of physics beyond the dark ages of high school ("knechten"). My (extended) family—Dicken, Hexe, Ulli, Twan, Thomas, Tini, Käthe, Oma, Opa, Gisela, Gregor, Kelly, Tom—because without their support I would not have managed to finish this thesis. Bo and Amy van Meegen; I will not even dare to try and find se appropriate words for you.

Thank you.

Finally, I would like to acknowledge the financial support by the DFG and the Studienstiftung which made this thesis possible in the first place.

Part I

INTRODUCTION

1 STRUCTURE OF THIS THESIS

Establishing relations between the intricate structure of neural networks, in combination with the properties of the neurons, to the emerging network dynamics is the main topic of this thesis. More concretely, the relationship between structure and dynamics is investigated using neural network models with an increasing degree of complexity. This increasing degree of complexity is the organizing principle of this thesis: we start with simple, but analytically well-tractable, models and work our way up to a complex, data-based model which relies almost entirely on simulations.

While the problem is motivated by neuroscience, the approach is shaped by a statistical physics point of view. The introduction is therefore split into two chapters: a summary of the relevant neuroscientific background in Chapter 2 and a summary of the main tools borrowed from statistical physics in Chapter 3.

In Chapter 4, the first chapter in the main part, the network models under investigation are very simple from a neuroscientific point of view: fully-connected feedforward networks (DNNs) and vanilla recurrent networks (RNNs). Owing to their simplicity, we can, for the only time in this thesis, go beyond dynamics and take functional aspects into account. To this end, we use the framework of Bayesian supervised learning and develop a unified field-theoretical perspective on DNNs and RNNs. The main result is a surprising similarity of, but also subtle differences between, the two network models.

The first step towards more complicated models, which we make in Chapter 5, is to replace discrete time steps by dynamics unfolding continuously in time. For block-structured, recurrent networks of rate neurons we combine the field-theoretical approach with large-deviation theory to derive the distribution of network-averaged observables across the network ensemble. This result allows to calculate beyond-mean-field fluctuations of the order parameter. Furthermore, it allows to attack the inverse problem: inferring the statistics of the connectivity from observed dynamics.

Spikes make their first appearance in Chapter 6 where we investigate the single-unit timescale in block-structured, recurrent networks of two types of spiking neuron models. Within our approach based on dynamic mean-field theory, the main technical challenge is the colored noise problem: determining the output statistics of a neuron driven by temporally correlated input. Our analytical solutions to the colored noise problem enable parameter scans to investigate the influence of network parameters on the timescale.

Chapter 7 is a detour from the main track to take a look at a peculiar flower we found along the way: the distribution of neuron densities is, to a surprising degree, lognormal. This holds true within cortical areas as well as across cortical areas and for several mammalian species. To provide an explanation to this curious finding, we propose a simple model of noisy cell division which leads to lognormally distributed neuron densities.

The main part ends with the large-scale, data-based model of human cortex described in Chapter 8. This model can be seen as a framework for data integration: due to the absence of comprehensive data on the structure of human cortical networks, various data modalities are aggregated to arrive at a consistent model. We simulate the resulting model and compare it to activity data from recordings in cortex both on the single-neuron and the area level.

In the final discussion we briefly summarize the main results of the individual chapters and suggest avenues for future work. Finally, the individual chapters are put into the broader context of the overarching theme of this thesis.

A chapter in the appendix is devoted to a software toolbox. Due to the elaborate nature of many analytical results in the context of mean-field theory for spiking networks, their numerical implementation is not straightforward and inherently error prone. To facilitate the use of these methods, we collected their implementations in the toolbox.

2 NEUROSCIENCE

> "The ultimate goal of neural science is to understand how the flow of electrical signals through neural circuits gives rise to mind—to how we perceive, act, think, learn, and remember." (Kandel, Schwartz, Jessell, et al. 2013)

Currently, this "ultimate goal" seems still very far away. But of course, more than a decade of intense research did not pass without significant insights. This chapter summarizes a selection of these insights.

A recurring theme in this thesis is the step from the micro-scale to the macro-scale. In this spirit, we start on the micro-scale with "the basic units of the brain" (Kandel, Schwartz, Jessell, et al. 2013) in Section 2.1: the neurons. The neurons are interconnected into intricate networks on the macro-scale; this is the topic of Section 2.2. Next, we discuss the activity of cortical neurons in the absence of a controlled, external stimulus in Section 2.3. Finally, we leave the realm of the biological reality and discuss models of it in Section 2.4.

2.1 NEURONS

2.1.1 *The Butterflies of the Soul*

Golgi's staining method (Golgi 1873), which stains a random subset of neurons to make them visible under a microscope, enabled Cajal's discovery (Ramón y Cajal 1888) that the brain consists of neurons, giving rise to the *neuron doctrine* (DeFelipe 2015; Yuste 2015). Their

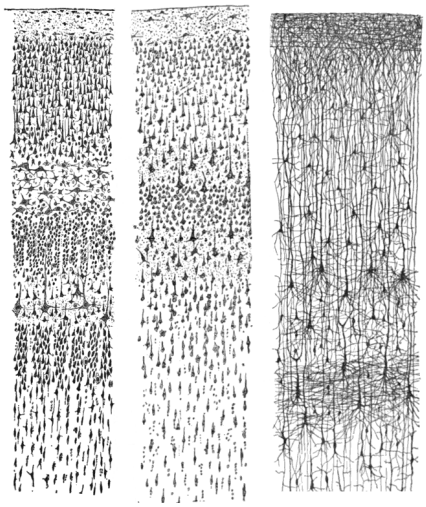

Figure 2.1: Stained sections of human cortex. Left: Visual cortex of a human adult; only cell bodies are visible (Nissl-stained). Middle: Motor cortex of a human adult; only cell bodies are visible (Nissl-stained). Right: Cortex of a 1 1/2 month old infant; cell bodies, dendrites, and axons are visible for a random subset of neurons (Golgi-stained). Drawn by Ramon y Cajal (1899), file obtained from Wikimedia Commons (public domain).

staggering variety caused Cajal to call neurons the "butterflies of the soul" (Ramón y Cajal 1917); a recent census confirms this variety, counting 69 neuron types in human cortex (Hodge et al. 2019).

Of course, Golgi already saw neurons, dendrites, and axons. But he supported the theory that the axons form a continuous network rather than being separated into individual units with an inherent direction (DeFelipe 2015).

Despite their variety, most neurons possess three main elements (Kandel, Schwartz, Jessell, et al. 2013): the *soma* or cell body which is the center of the cell, the *dendrite* which receives input from other neurons, and the *axon* which sends output to other neurons. The intricate structure of the dendrites is well known due to the drawings of Cajal (Figure 2.1 on page 6). In contrast, we are only beginning to understand the even more complex structure of the axons (Winnubst et al. 2019; Peng et al. 2021) because they are very thin, mostly submicron (Liewald et al. 2014) and thus close to or below the diffraction limit of light microscopy, and they can spread across the entire cortex (Figure 2.2 on page 7). Furthermore, the neurons are densely packed: 1 mm^3 of mouse cortex contains approximately 10^5 neurons and 4 km of axons (Braitenberg and Schüz 1998).

In contrast to the heterogeneity of neurons, their means of communication is remarkably homogeneous (Kandel, Schwartz, Jessell, et al. 2013). It relies almost exclusively on stereotyped electrical pulses, the *action potentials* or *spikes*, which are initialized at the soma, more precisely at the axon initial segment, and travel along the axon. The stereotypical nature of the action potentials suggests that all infor-

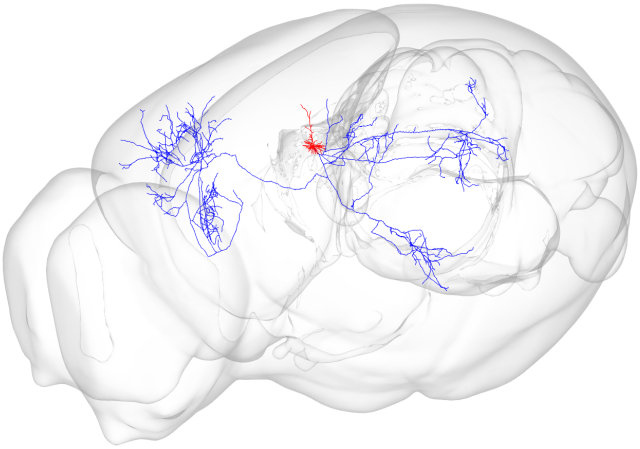

Figure 2.2: Visualization of a single neuron in mouse cortex with its full dendrite (red) and axon (blue). Data from (Winnubst et al. 2019) visualized using the MouseLight viewer by Rembrandt Bakker.

mation is encoded in their timing (Rieke et al. 1997). Propagating the action potential along the axon is an active process; the action potential is regularly regenerated to keep its 100 mV amplitude and 1-2 ms width. This active regeneration accounts for more than a fifth of the brain's energy budget (Laughlin and Sejnowski 2003; Harris, Jolivet, and Attwell 2012).

To reach its target neuron, an action potential has to be transmitted from an axon to a dendrite. This transmission takes place at the *synapses.* Most synapses rely on chemical signaling, with the electrical *gap junctions* being the notable exception to this rule (Kandel, Schwartz, Jessell, et al. 2013). At chemical synapses, pre- and postsynaptic neuron are physically separated by a small space of a few tens of nanometers, the synaptic cleft. To transmit a signal across the synaptic cleft, a *neurotransmitter* is released upon arrival of an action potential at the presynaptic axon terminal. This transmitter diffuses across the cleft and binds to receptors at the postsynaptic cell. The receptors activate ion channels in the postsynaptic cell which leads to a change of membrane conductance and voltage. While this intricate process leads to a delay on the order of a few milliseconds, it allows a graded effect on the postsynaptic cell despite the homogeneity of action potentials.

While diffusion is extremely slow in our macroscopic world, it can be quite effective on cellular length scales (Berg 1993).

There are three main neurotransmitters—glutamate, GABA, and glycine—with corresponding receptors: kainate, AMPA, and NMDA for glutamate, $GABA_A$ and $GABA_B$ for GABA, and glycine receptors (Kandel, Schwartz, Jessell, et al. 2013). Although the effect on the postsynaptic neuron depends on the receptor, not the transmitter, synapses can be organized based on the transmitter into *excitatory* (glutamate)

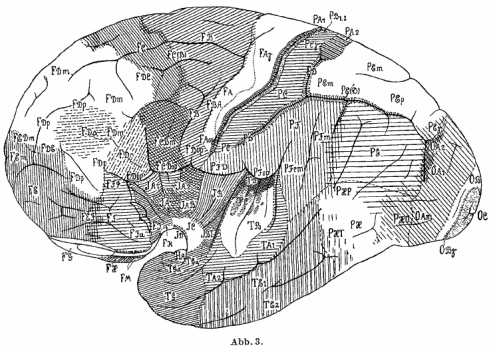

Figure 2.3: Map of cortical areas based on the cytoarchitecture. The first letter of the area name indicates the lobe: temporal (T), occipital (O), parietal (P), and frontal (F). Drawn by von Economo (1929), file obtained from Wikimedia Commons (public domain).

and *inhibitory* (GABA and glycine) because the corresponding receptors typically trigger a depolarization (excitatory) or hyperpolarization (inhibitory) of the postsynaptic cell (Kandel, Schwartz, Jessell, et al. 2013). In combination with *Dale's law*, i.e., neurons release the same set of transmitters at all of their synapses, one can classify not only synapses but also neurons as either excitatory or inhibitory.

2.1.2 *Brain Cartography*

Staining methods revealed not only that the brain consists of neurons, but also that their density and structure varies systematically across cortex. First, neurons are separated into *layers* which are clearly visible already in the drawings by Cajal (Figure 2.1 on page 6). Second, the density of the neurons within the layers as well as the relative size of the layers varies discontinuously. These systematic variations were first compiled by Brodmann (1909) into a map of cortex consisting of distinct *areas*. Roughly two decades later, von Economo and Koskinas (1925) refined Brodmann's work, leading not only to a map that is more finely parcellated (Figure 2.3 on page 8) but also to the only layer-resolved collection of neuron densities across the entire human cortex up to this date.

The early maps of cortex were based on the cytoarchitecture, i.e., the neuron density and the layer structure. This line of work has been continued up to the present day; a recent example is the Julich-Brain atlas of human cortex (Amunts et al. 2020). However, cytoarchitecture is not the only criterion that is used to map cortex. A simple alternative is to use landmarks on the cortical surface as done for example in the widely used Desikan-Killiany parcellation (Desikan et al. 2006).

2.2 CORTICAL NETWORKS

The synaptic connections between the neurons give rise to highly recurrent networks. The structure of these networks—the connectivity—depends on the spatial scale. Below, we will follow a gradient from the large-scale to the small-scale structure.

Before diving into the details, let us start with a rough estimation of the orders a magnitude involved. In a fully reconstructed 1 mm^3 sample of human cortical tissue from the temporal lobe there are 16,087 neurons and 133,7 million synapses (Shapson-Coe et al. 2021), leading to 8,311 synapses per neuron on average. Taking into account that up to a few synapses can connect a single pair of neurons but the vast majority of $> 90\%$ connections are mediated by single synapses (Shapson-Coe et al. 2021), we arrive at an estimate of $O(1{,}000)$ to $O(10{,}000)$ presynaptic neurons. Thus, cortical neurons are highly interconnected.

2.2.1 *Large-Scale Network Structure*

Somewhat counter-intuitively, the cortex-wide network structure between areas is better understood than the local structure within areas. The main reason for this is that on the coarser spatial scale of areas, there are more experimental techniques available.

Tracing

Axons consist of microtubules along which active transport towards the cell body (*retrograde*) and away from the cell body (*anterograde*) takes place. In tracing experiments, this active transport mechanism is exploited to reveal the connectivity. To this end, a tracer, e.g., horseradish peroxidase for retrograde tracing or more recently viral tracers (Nassi et al. 2015; Rockland 2019), is injected in a given area. Importantly, the tracer is designed such that it cannot cross a synapse (or only in a controlled manner, e.g., a predetermined number of crossings). Within a few days, the tracer is transported along the axon towards the source (retrograde) or target (anterograde) areas. Afterwards, the animal is sacrificed, its brain is sectioned, and the tracer is made detectable (for example by staining or immunohistochemistry).

Tracing experiments have been performed since more than half a century with increasing sophistication (Nassi et al. 2015). Since only one, or at most a few, areas can be injected at the same time, these experiments used to provide only very specific information about either sources or targets of a given area. Efforts like the CoCoMac database (Stephan et al. 2001; Bakker, Thomas, and Diesmann 2012) or the Marmoset Brain Connectivity Atlas (Majka et al. 2020) remedy this problem by collecting the results of multiple experiments and bringing them into a coherent structure. In addition, there are

several recent large-scale tracing experiments following standardized protocols and techniques: retrograde tracing in macaque (Markov, Ercsey-Ravasz, Van Essen, et al. 2013; Markov, Vezoli, et al. 2014; Markov, Ercsey-Ravasz, Ribeiro Gomes, et al. 2014), retrograde tracing in mouse (Gămănuţ et al. 2018), and anterograde tracing in mouse (Oh et al. 2014) which was extended to be layer- and cell-type specific (Harris, Mihalas, et al. 2019).

A key finding from the large-scale retrograde tracing experiments is a very dense connectivity on the area level. In mouse, injections into 19 areas showed that 97% of the possible $19(19-1)$ connections are present (Gămănuţ et al. 2018); in macaque, injections into 29 areas showed that 66% of the possible $29(29-1)$ connections are present (Markov, Ercsey-Ravasz, Van Essen, et al. 2013). The dense connectivity comes along with a very high variability of the connection strength, measured for example by the fraction of labeled neurons (*FLN*) in a given source area relative to the total number of labeled neurons in all source areas for a single injection. In both mouse and macaque, the FLN is approximately log-normally distributed and spans five orders of magnitude from below 10^{-5} up to almost 1 (Markov, Ercsey-Ravasz, Van Essen, et al. 2013; Gămănuţ et al. 2018). Intuitively, one might expect that nearby areas have a higher FLN; indeed, this intuition is quantified by the exponential decay of FLN with projection distance in both mouse (decay constant 1.3 mm) and macaque (decay constant 5.3 mm) (Ercsey-Ravasz et al. 2013; Horvát et al. 2016). The extremely wide distribution, and in particular the long left tail, tells a cautionary tale that sophisticated techniques and meticulous procedures are necessary in order to not miss a connection.

Diffusion MRI

Tracing experiments possess the inherent problem that they are invasive. Accordingly, they are not used to investigate the connectivity in human cortex—for obvious reasons. To this end, non-invasive techniques are necessary.

The most prominent non-invasive technique is diffusion MRI in combination with tractography (Jbabdi et al. 2015). Diffusion MRI indicates anisotropies in the diffusion of water molecules caused by the axons. This information about the orientation of the fibres is used by tractography algorithms to construct long-range connections. An inherent problem of the method is that it yields a symmetric connectivity and that it cannot distinguish "crossing" or "kissing" fibres, leading to false positives (Jbabdi et al. 2015; Maier-Hein et al. 2017). Nonetheless, it captures most of the existing connections (Maier-Hein et al. 2017) and is strongly correlated with the connectivity obtained from tracing data: $r = 0.59$ for macaque (Donahue et al. 2016), up to $r = 0.91$ for ferret (Delettre et al. 2019). Note, however, the decrease in correlation from ferret to macaque.

An advantage of diffusion MRI compared to tracing is that it provides data for an individual brain. Large-scale efforts, such as the Human Connectome Project (Van Essen et al. 2013), collected MRI data for a large population of individuals which allows one to study the variability across the population, in particular with respect to neurological diseases.

2.2.2 *Small-Scale Network Structure*

Tracing and diffusion MRI provide insights into the connectivity between entire areas consisting of $O(10^8)$ neurons. The obvious follow-up question is: what does the structure of the network at smaller scales, down to single-synapse resolution, look like?

Electron Microscopy

The straightforward way to answer this question is to reconstruct cortical tissue at a nanometer resolution such that all structures, including axons and synapses, can be identified. This resolution, which is below the diffraction limit of light microscopy, can be achieved by electron microscopy (Gray 1959a; Gray 1959b). While acquiring the data already poses significant challenges, the post-processing—stitching, aligning, and segmenting the images—is an equally heroic task. For example, the 1 mm^3 sample of human cortex mentioned above took 326 days of imaging yielding 2.1 petabyte of raw data, was reconstructed using a sophisticated semi-automated workflow, and required four years of work to complete it (Shapson-Coe et al. 2021).

One of the first major contributions of electron microscopy was to show that dendritic spines are the location of synaptic contacts (Gray 1959a; DeFelipe 2015).

The reward for this effort is a uniquely detailed view on the structure of cortex. The findings of Shapson-Coe et al. (2021) include that there are two times more glia cells than neurons in the volume, that 69% of neurons as well as synapses in the volume are excitatory, that 99.4% of the 133.7 million synapses are between axons and dendrites, that the synapse density is highest in the upper layers, that 74.3% of the volume is occupied by axons and dendrites, and that very rarely an axon forms ten to twenty synapses with a target cell.

Canonical Microcircuits

One drawback of electron microscopic reconstructions is that its results are restricted to a small volume in a specific area (with the notable exception of *C. Elegans* where the whole brain has been reconstructed across multiple stages of development by Witvliet et al. (2021)). However, this problem would be alleviated if basic principles of the network structure are conserved across areas or even species—this is the idea of a stereotypical, or canonical, microcircuit (Douglas, Martin, and Whitteridge 1989; Douglas and Martin 2004; Bastos et al. 2012; Harris and Shepherd 2015).

While there are certain features that seem conserved, e.g., excitatory cells in layer IV receive only little input from excitatory cells in other layers and project to excitatory cells in layers II/III and V, there exists also a significant variability, e.g., layer IV excitatory cells project to layer VI in some species but not in others (Harris and Shepherd 2015). Moreover, it seems necessary to account for the different classes of neurons to capture the features of the microcircuit (Harris and Shepherd 2015). Unfortunately, there is currently no comprehensive quantitative data on the local network structure even in mouse primary visual cortex (Billeh et al. 2020). Thus, while the canonical microcircuit remains a promising candidate for an ordering principle of the local network structure, it remains to be seen to which extent this hypothesis holds true.

2.3 RESTING STATE ACTIVITY

Thus far, we have considered only the structural features of cortex. But to "[...] understand how the flow of electrical signals through neural circuits gives rise to mind [...]" (Kandel, Schwartz, Jessell, et al. 2013), it is necessary to go beyond the static structure and to consider the activity.

Typically, the activity in response to a specific stimulus or task is investigated to create a link between activity and function. Here, we restrict ourselves to *resting state activity* (Deco, Jirsa, and McIntosh 2011) which is (supposedly) intrinsically generated and free from external influences. While this seems considerably less exciting at first sight, it might have the advantage that the absence of external influences allows for a tight link between network structure and activity. For example, the *default mode network*, a group of areas with correlated activity during resting state (Raichle 2015), can be linked to an anatomical backbone using data from tracing experiments (Buckner and DiNicola 2019). Furthermore, resting state activity seems to be altered by neurological disorders like schizophrenia (Buckner and DiNicola 2019).

A wide range of methods exists to measure the activity of neurons, but each one of them is restricted to a specific subset of temporal or spatial scales (Sejnowski, Churchland, and Movshon 2014). Below, we briefly discuss two methods at opposite ends of the spectrum which are frequently used to investigate resting state activity: electrode recordings with single-neuron resolution and high temporal precision, and fMRI imaging with whole-brain coverage but low temporal precision.

2.3.1 *Electrode Recordings*

Electrode recordings measure the electrophysiological properties of neurons at a millisecond scale and they can take place either intra- or extracellularly (see, e.g., the textbook by Brette and Destexhe 2012). For intracellular recordings, the electrode is inserted into the neuron which allows one to record the membrane voltage or the current across the membrane. Extracellular recordings can measure the activity of either single or multiple neurons. For extracellular single-neuron recordings, a microelectrode is positioned close to the cell such that it picks up the changes in the potential caused by currents across the cell membrane during an action potential. Multi-unit recordings involve electrodes with larger tips. These electrodes are arranged either linearly to allow laminar recordings or in a grid. In multi-unit recordings, the signal needs to be post-processed using *spike sorting*—identifying different neurons based on the shape of the action potential—to arrive at parallel recordings of single-neuron activity. Alternatively, the low-pass filtered signal gives rise to the *local field potential*, which is a proxy for the activity of the neurons surrounding the electrode.

Measuring neural activity using electrodes has a long history in neuroscience (Yuste 2015). For example, intracellular recordings were at the heart of Hodgkin and Huxley's Nobel Prize-winning work on the generation of action potentials (Hodgkin and Huxley 1952) and extracellular recordings led to the Nobel Prize-winning discovery by Hubel and Wiesel (1962) of receptive fields of neurons in cat primary visual cortex.

Both intra- and extracellular recordings, as well as other modalities with single-neuron resolution, paint a consistent picture of the resting state activity in cortex (Harris and Thiele 2011): the activity is asynchronous and irregular. Put differently, the cross-correlations between the neurons are low and the variability of the times between spikes, the *inter-spike intervals*, is high. More quantitatively, for resting state activity in macaque motor cortex, the cross-correlations are distributed between -0.2 and 0.2 and vanish on average (Dahmen, Layer, et al. 2022) and the local coefficient of variation is distributed between 0.5 and 1.2 with an average of 0.9 (Dąbrowska et al. 2021).

2.3.2 *fMRI Imaging*

Increased neural activity in a part of the brain triggers a change in the blood flow, the *hemodynamic response*, in order to supply the increased energy demand (for a brief review of fMRI see, e.g., Logothetis 2008). The associated changes in the oxygen level of the blood can be made visible using blood-oxygen-level-dependent (BOLD) imaging because oxygenated and deoxygenated hemoglobin have different magnetic properties. Thus, the BOLD signal provides a proxy of neural activity

across the entire brain. Its drawbacks are the low spatial and temporal resolution on the order of millimeters and seconds, respectively.

The analysis of resting state fMRI activity is in itself a quite prominent field (Raichle 2015). Historically, Biswal et al. (1995) showed that the correlations of BOLD signals between areas in resting state activity, often somewhat confusingly called *functional connectivity*, display striking patterns. The same patterns are also found in the *default mode network*, which was discovered because it exhibits 'negative responses' (compared to baseline) in task related activity (Raichle 2015). This relation between task related and resting state activity spurred significant, and still ongoing, efforts to understand the underlying cause of the default mode network (Buckner and DiNicola 2019). One approach to tackle this problem is to use models to link the underlying network structure with the emerging neural activity (Deco, Jirsa, and McIntosh 2011), which brings us to the next section.

2.4 MODELS

> "What makes a model good? Clearly it must be based on biological reality, but modeling necessarily involves an abstraction of that reality. It is important to appreciate that a more detailed model is not necessarily a better model. A simple model that allows us to think about a phenomenon more clearly is more powerful than a model with underlying assumptions and mechanisms that are obscured by complexity. The purpose of modeling is to illuminate, and the ultimate test of a model is not simply that it makes predictions that can be tested experimentally, but whether it leads to better understanding. No matter how detailed, no model can capture all aspects of the phenomenon being studied. As theoretical neuroscientist Idan Segev has said, borrowing from Picasso's description of art, modeling is the lie that reveals the truth." (Kandel, Schwartz, Jessell, et al. 2013)

Models play a peculiar role in neuroscience. They are always under the tension that is vividly described in the quote above: if they are too simple, they do not capture the relevant phenomenon; if they are too complex, they do not increase our understanding. Somewhere in between is a sweet spot which is hard to achieve—in particular because opinions diverge on the precise location of this sweet spot.

The interplay between theory and simulation adds to this tension (Gerstner, Sprekeler, and Deco 2012). If a model is simple enough, all its consequences can be worked out, and understood, analytically. However, analytically tractable models might be too simple to capture the relevant phenomena, such that more complex models are necessary

which rely on simulations to investigate their implications (Einevoll et al. 2019).

Both tensions—complexity vs. understandability as well as theory vs. simulation—will be a recurring theme not only in this section but throughout the thesis. In this section, we start with models of neurons, then move to models of neural networks, and last mention a few salient properties of cortex which were neglected thus far.

2.4.1 *Neuron Models*

The subdivision of a neuron into dendrite, soma, and axon suggests to model them as input-output devices. Since this is a rather weak constraint, it leaves a lot of room for different implementations of the device. Indeed, the abovementioned challenge—to account for the relevant details and to neglect the irrelevant ones (Herz et al. 2006)—has led to an entire zoo of different neuron models. Below, we introduce two models in detail, the *Generalized Linear Model* (GLM) and the *Leaky Integrate-and-Fire* model (LIF), before zooming out again for a brief tour through the zoo.

Both GLM and LIF model account for spikes. Thus, the output of either model is a spike train

$$x(t) = \sum_k \delta(t - t_k) \tag{2.1}$$

where t_k denotes the time of the k-th spike and $\delta(t)$ is a Dirac delta function. The two models differ fundamentally in their transformation from input to output: the GLM is inherently stochastic, i.e., the exact same input $\eta(t)$ might produce different output spike trains $x(t)$, while the LIF model is deterministic.

Generalized Linear Model

The GLM consists of three distinct steps (Gerstner, Kistler, et al. 2014). In the first step, the input $\eta(t)$ is filtered linearly through a filter $\kappa(t)$ to produce the membrane voltage

$$V(t) = (\kappa * \eta)(t). \tag{2.2}$$

Here, $(a * b)(t) = \int_{-\infty}^{\infty} ds\, a(t-s)b(s)$ denotes a convolution and causality requires $\kappa(t) = 0$ for $t < 0$. For convenience, we also shifted the voltage such that it is zero in the absence of input. The filter $\kappa(t)$ accounts for the effect of an incoming spike on the membrane voltage: considering a single incoming spike at $t = \hat{t}$ with synaptic strength J, the input is $\eta(t) = J\delta(t - \hat{t})$ and (2.2) yields $V(t) = J\kappa(t - \hat{t})$. In the second step, the difference between voltage and threshold θ is transformed through a non-negative nonlinearity $\phi(x) \geq 0$ to produce the firing rate

$$\lambda(t) = c_1 \phi\left(c_2 \left(V(t) - \theta\right)\right). \tag{2.3}$$

The parameters c_1 and c_2 control the overall firing rate and the sensitivity to changes in the voltage, respectively. Typical choices for the nonlinearity include a rectified linear function $\phi(x) = \max(0, x)$ or an exponential function $\phi(x) = \exp(x)$. In the third step, a spike train is generated through an inhomogeneous Poisson process with intensity $\lambda(t)$. In simpler terms, this means that in an infinitesimal time bin dt a spike is generated with probability $\lambda(t)dt$ independent of what happened in any previous (or future) bin.

Although the stochastic nature of the GLM might seem like a disadvantage at first sight, it can be a valuable asset. To fit the model to data, the full toolbox of statistical inference is readily applicable because the likelihood of the GLM is determined by the resulting inhomogeneous Poisson process (Paninski 2004). Accordingly, GLMs are often used to fit to data, see, e.g, Pillow and Latham (2007), Pillow, Shlens, et al. (2008), and Bellec, Wang, et al. (2021).

Leaky Integrate-and-Fire Model

Similarly to the first step of the GLM, the input is filtered linearly to arrive at the membrane voltage. For LIF neurons, these linear filters are typically specified in terms of ordinary differential equations for the membrane voltage V and the synaptic current I,

$$\tau_{\mathrm{m}} \dot{V} = -V + R_{\mathrm{m}} I, \tag{2.4}$$

$$\tau_{\mathrm{s}} \dot{I} = -I + \tau_{\mathrm{s}} \eta(t), \tag{2.5}$$

where τ_{m} and τ_{s} are the membrane and synaptic time constants, respectively, and R_{m} the membrane resistance. The membrane time constant is determined by the membrane resistance R_{m} and the membrane capacitance C_{m} through $\tau_{\mathrm{m}} = R_{\mathrm{m}} C_{\mathrm{m}}$. Similar to the GLM, we shifted the voltage such that it is zero in the absence of input. Sometimes, the *exponential post-synaptic currents* in (2.5) are replaced by *delta synapses* $R_{\mathrm{m}} I(t) = \tau_{\mathrm{m}} \eta(t)$ which correspond to the limit of infinitely short synaptic time constants.

The crucial difference between GLM and LIF neurons is the firing mechanism. LIF neurons emit a spike as soon as the voltage crosses the threshold θ. After the threshold crossing event, the voltage is set to the reset voltage V_{r} where it stays clamped during the refractory period t_{ref}. Once the refractory period is over, the membrane voltage evolves again according to (2.4) until it crosses the threshold again. This "fire-and-reset" mechanism renders the LIF deterministic, in contrast to the stochastic GLM, and highly non-linear.

Although the LIF model is more difficult to fit to data, the vanilla LIF and generalizations thereof were fitted to a large, standardized database of electrophysiological experiments performed at the Allen Institute (Teeter et al. 2018). The authors found that "overall, [the LIF model] had a surprisingly high explained variance of 70% when all neurons were considered." This finding raises the hope that even

strongly simplified neuron models, such as LIF neurons, might capture most of the relevant dynamics displayed by neurons.

The Zoo of Neuron Models

With two concrete examples at hand, let us consider again the entire zoo of different neuron models. To provide a guiding thread, we follow along the five levels introduced by Herz et al. (2006).

On level I, the entire structure of the dendrite shown in Figure 2.1 on page 6 is taken into account. To this end, the neuron is divided into compartments which obey their own dynamical equation. Thus, a single neuron is described by a coupled system of $O(1{,}000)$ dynamical equations. While this allows one to fully take the geometry of the neuron into account, it hardly reduces the complexity. On level II, the situation is remedied by reducing the number of compartments to a few. On level III, this approach is taken to the extreme and the neuron is described by a single compartment. The archetypical single-compartment neuron model is the Hodgkin-Huxley model (Hodgkin and Huxley 1952) which describes how the various ionic currents conspire to produce an action potential in the giant squid axon. Still, the Hodgkin-Huxley model is a four-dimensional, coupled, nonlinear system of differential equations which is hard to understand. To further reduce the complexity, various two-dimensional simplifications of the Hodgkin-Huxley model exist which capture essential qualitative features of the dynamics of the full model; prominent examples are the FitzHugh-Nagumo model and the Morris-Lecar model (Gerstner, Kistler, et al. 2014). Finally, there are a number of models which further reduce the dimensionality to one. The LIF model introduced above belongs to this class together with simpler models, like the Perfect Integrate-and-Fire model (Lapicque 1907; Brunel and van Rossum 2007), as well as more complex ones, like the (adaptive) Exponential Integrate-and-Fire model (Brette and Gerstner 2005).

On level IV, a qualitative change occurs because the quantities described by the mathematical equations become detached from the underlying biophysical processes. Instead, the focus shifts towards taking only the input-output relationship into account. The most prominent model on this level is the GLM, either in the minimal version presented above or in more complicated versions taking, e.g., after-spike currents into account (Gerstner, Kistler, et al. 2014). On level V, the models are entirely detached from the biophysical processes and the goal is fully restricted to capturing the input-output relationship. Typically, this is done using a parameterized distribution $p(x \mid \eta)$. With the advent of deep learning (LeCun, Bengio, and Hinton 2015), level V models gained significant traction because deep learning is essentially a technique to fit complicated parameterized distributions $p(x \mid \eta)$, the various networks, to data (MacKay 2003). Transposing the techniques and networks from deep learning has led to new state-of-

the art models of, among others, tiger salamander retinal ganglion cells (McIntosh et al. 2016) and macaque V1 neurons (Cadena et al. 2019).

Following the above gradient of models further, there is yet another level of models outside the scope of Herz et al. (2006): even more simplified models which are typically employed in theoretical studies of network dynamics. The archetypical model of this level is the binary neuron model employed in associative network models (Amit, Gutfreund, and Sompolinsky 1985) or balanced networks (van Vreeswijk and Sompolinsky 1996; van Vreeswijk and Sompolinsky 1998). Another typical example are rate neurons which fully neglect the discontinuity imposed by the spikes. These rate neuron models are either derived from more complex models (e.g., Wong and Wang 2006; Mastrogiuseppe and Ostojic 2017) or simply chosen to match the qualitative features of the input-output relationship (e.g., Sompolinsky, Crisanti, and Sommers 1988).

2.4.2 *Network Models*

Going beyond single neurons, there exists also a long tradition of models of neural networks (Yuste 2015) starting with the work of McCulloch and Pitts (1943). Such models were built with a wide variety of goals—prominent examples are the associative memory model by Hopfield (1982) or the model of asynchronous, irregular activity by van Vreeswijk and Sompolinsky (1996)—and take into account various levels of detail.

There are several dimensions along which network models can be categorized, for example

- top-down vs. bottom-up models (Gerstner, Sprekeler, and Deco 2012), i.e., either starting from the biological details (bottom-up) or from a high level function (top-down);
- single-neuron vs. population models, i.e., either taking the activity of single neurons into account or describing only the activity of a population of neurons;
- spiking vs. rate-based models, i.e, either taking the spiking (or, more generally, the discontinuous) nature of the communication into account or simplifying it to a continuous signal.

Sticking to the overarching theme of the thesis—starting from the single-neuron level and bridging to the network level—we focus on bottom-up, single-neuron models in the remainder of this introduction.

Random Networks

We saw in Section 2.2 that there is (almost) no comprehensive data on the network structure on the single-neuron level. Thus, single-neuron-

level network models are typically *random networks* with statistics which are in rough agreement with the available data. For example, connections are chosen with a fixed connection probability (*Gilbert model*; Gilbert 1959), a fixed number of connections is randomly assigned to all possible pairs of neurons (*Erdős–Rényi model*; Erdős and Rényi 1959), or a fixed number of presynaptic / postsynaptic neurons is chosen randomly (*fixed indegree / outdegree*).

Beyond the simple models mentioned above, there exists a wide range of more complicated models. Important examples include small-world networks like the model by Watts and Strogatz (1998) or scale-free networks like the model by Barabási and Albert (1999). Furthermore, certain low-order statistics, or *motifs*, seem to be over-represented compared to a Gilbert null model (Song et al. 2005); this can also be taken into account in the model.

All of the above models determine merely the structure of the network, i.e., whether a synapse exists or not. Additional assumptions need to be made to quantify the synaptic weights, e.g., a log-normal distribution (Buzsáki and Mizuseki 2014; Ziv and Brenner 2018).

Balanced Networks

The high number of synapses per neuron, the indegree K, and the asynchronous irregular activity in cortex are seemingly in contradiction to each other: according to the central limit theorem, fluctuations in the input average out in the limit $K \to \infty$ if they are weakly correlated (asynchronous). Thus, the asynchronous irregular state is not self-consistent in the limit $K \to \infty$ because the fluctuations in the input, and hence in the activity, vanish, leading to regular activity which is furthermore the same for all neurons and hence highly synchronous.

There is an elegant resolution of this apparent paradox using the excitatory or inhibitory nature of neurons: *balanced networks* (van Vreeswijk and Sompolinsky 1996; van Vreeswijk and Sompolinsky 1998). In such networks, the inputs are strong, which means that $O(\sqrt{K})$ are sufficient to evoke activity in the postsynaptic neuron. Put differently, the strength of the weights is $O(1/\sqrt{K})$. Thus, the mean input is $O(\sqrt{K})$ and the fluctuations are $O(1)$. We see already that the fluctuations do not vanish; however, the mean activity diverges in the limit $K \to \infty$. Here, the excitatory and inhibitory nature of the neurons enters the picture: both cancel each other approximately leading to an $O(1)$ mean input. Intuitively speaking, there are two very strong forces driving the neuron which annihilate each other on average but give rise to violent fluctuations.

This *balanced state* is self-consistent (and stable) and emerges in a large parameter regime (van Vreeswijk and Sompolinsky 1996; van Vreeswijk and Sompolinsky 1998). The main requirements on the parameters are dominance of inhibition (otherwise recurrent excitation would lead to divergent activity) and excitatory external input (oth-

erwise inhibition dominance would lead to vanishing activity). Thus, asynchronous and irregular activity can be a simple consequence of the large indegree, the presence of both excitation and inhibition, and a random network structure. Note that this is truly a network-level mechanism; there are no fluctuations induced by the external input.

Remarkably, the above arguments for the emergence of a balanced state hardly depend on details such as the neuron model. Indeed, a balanced state emerges not only in the networks of binary neurons investigated by van Vreeswijk and Sompolinsky (1996) but also in spiking networks (Amit and Brunel 1997; Brunel 2000) and rate networks (Sompolinsky, Crisanti, and Sommers 1988; Kadmon and Sompolinsky 2015).

An important detail in the heuristic derivation sketched above is that cross-correlations were implicitly assumed to be vanishing. Understanding this aspect, in particular in light of the shared input that neurons receive, required further investigations (Renart et al. 2010; Tetzlaff et al. 2012; Helias, Tetzlaff, and Diesmann 2014). Furthermore, a direct consequence of the balanced state is a linear input-output relation on the network level (van Vreeswijk and Sompolinsky 1996) which might not seem plausible (Ahmadian and Miller 2021). However, this argument holds only in the limit $K \to \infty$; at finite (but large) K balanced networks support various different nonlinearities (Sanzeni, Histed, and Brunel 2020).

Data-Driven Spiking Networks

Balanced network models are only weakly constrained by anatomical data, e.g., by choosing an appropriate indegree for the neurons (Amit and Brunel 1997; Brunel 2000). The fact that they already yield asynchronous and irregular activity leads to the follow-up questions whether this still holds if one takes additional data into account, and whether this additional data improves the quantitative agreement with the activity statistics in cortex.

"In time, those Unconscionable Maps no longer satisfied, and the Cartographers Guilds struck a Map of the Empire whose size was that of the Empire, and which coincided point for point with it." (Borges, quoted in Abbott 2008)

It seems obvious that a model which takes all details into account also produces realistic dynamics. However, modeling inherently has to neglect features, which renders the question whether it is possible to achieve plausible activity under the given constraints non-trivial. Furthermore, the available data is not yet sufficient to fully specify the models. Thus, the data needs to be supplemented by statistical regularities discovered in other species or modalities (*predictive connectomics*; see Hilgetag et al. 2019; van Albada, Morales-Gregorio, et al. 2022) which adds another layer of complexity and, potentially, errors.

Using data-constrained network models is a research direction which gained traction in the last decade (for a review see Shimoura et al. 2021). The first brain-scale model which was strongly constrained by data is the model of the "mammalian thalamocortical system" by Izhikevich and Edelman (2008). This model is mainly based on a com-

bination of human DTI data and data from cat visual cortex (Binzegger, Douglas, and Martin 2004), it uses the neuron model proposed by Izhikevich (2003), and it includes short-term plasticity as well as long-term spike-time-dependent plasticity. In total, the model comprises 10^6 neurons (but the size was increased to up to 10^{11} according to the supplement) which display, after an initial stimulation, self-sustained activity featuring different rhythms and waves.

Even stronger constraints by data were applied to the model of rat somatosensory cortex by Markram et al. (2015). In this model, the full morphology of approximately 31,000 neurons and their position in space is taken into account. These neurons belong to 55 morphological classes and 11 electrophysiological classes. The resulting activity reproduced features observed either in vitro or in vivo, for example a dependence of the synchronicity on the concentration of extracellular calcium. In follow-up studies, this model was used to investigate the interplay of noise and chaos (Nolte, Reimann, et al. 2019) and the effect of higher-order statistics in the connectivity (Nolte, Gal, et al. 2020).

The most detailed model to date is the model of mouse primary visual cortex developed by Billeh et al. (2020). It is based on the extensive amount of data on mouse cortex collected in particular at the Allen Institute (Tasic et al. 2018; Seeman et al. 2018; Gouwens et al. 2018; Teeter et al. 2018; Siegle et al. 2021). In total, it comprises ~230,000 neurons and it comes in two versions: a version using morphologically detailed multi-compartment neurons and a reduced version based on point neurons. After refinements of the connectivity, the activity of the model quantitatively reproduces features observed in vivo for a large array of different visual stimuli. Recently, the connectivity of the model was further adjusted using gradient-descent based methods to enable the model to solve tasks (Chen, Scherr, and Maass 2021; Scherr and Maass 2021).

In another line of work, the focus is to take the full density of neurons and synapses into account. The first full-density model by Potjans and Diesmann (2014) accounts for 1 mm^2 of generic early sensory cortex. The size of 1 mm^2 was chosen such that the majority of inputs are generated by neurons inside this column. Although all neurons are intrinsically the same, the model displays a systematic variation of firing rates across layers and populations which is also observed in vivo. In a next step, this model was used by Schmidt, Bakker, Hilgetag, et al. (2018) as a blueprint for the within-area connectivity to create a model of all vision-related areas in one hemisphere of macaque cortex. In this model, the connectivity between the areas is determined by tracing data (Markov, Ercsey-Ravasz, Ribeiro Gomes, et al. 2014) and supplemented by predictive connectomics. The resulting activity reproduces features observed in electrophysiological recordings as well as fMRI imaging (Schmidt, Bakker, Shen, et al. 2018).

While the model by Schmidt, Bakker, Hilgetag, et al. (2018) bridges between the single-neuron and area scale, area-level models consist more frequently of a few equations per area. One example are the models by Wang and coworkers (Chaudhuri, Knoblauch, et al. 2015; Mejias et al. 2016; Joglekar et al. 2018) which are based on the same tracing data (Markov, Ercsey-Ravasz, Ribeiro Gomes, et al. 2014). Another example is the Virtual Epileptic Patient—a network model describing the onset of epileptic seizures—developed by Jirsa, Proix, et al. (2017) which is currently being evaluated in an ongoing clinical trial (Wang et al. 2022).

2.4.3 *Neglected Aspects*

Let us conclude this introduction with a selection of phenomena which are omitted in this thesis. This selection is unordered and by no means comprehensive; it focuses only very briefly on particularly salient examples.

Why were these aspects omitted in the first place? "A good theoretical model of a complex system should be like a good caricature [...]" (Frenkel, quoted in Herz et al. 2006). In the caricatures sketched in this thesis, these aspects were neglected to avoid an overflow of complexity. Whether or not the resulting caricatures are still good remains to be judged by the reader.

Plasticity

The key omission is arguably to neglect plasticity: new synapses between neurons can form, old ones can vanish, and existing ones can alter their strength. This is currently believed to be the basis of the remarkable ability of animals, including humans, to learn from experiences and to adapt their behavior (Kandel, Schwartz, Jessell, et al. 2013). A wide variety of mechanisms for plasticity exist which can be loosely grouped into Hebbian plasticity, neuromodulated plasticity, and supervised plasticity (Magee and Grienberger 2020).

Intriguingly, significant changes in the network structure, more concretely the dendritic spines, also occur in the absence of activity (Berry and Nedivi 2017; Ziv and Brenner 2018). These spontaneous changes occur on quite fast timescales, i.e., within a few days (Mongillo, Rumpel, and Loewenstein 2018). The implications of this finding for the synaptic theory of learning are still unclear (Ziv and Brenner 2018).

Dendritic Nonlinearities

Modeling a neuron by a single compartment reduces the dendrites to passive cables. However, they support a range of nonlinear operations—calcium, sodium, and NMDA spikes—as well as backpropagating

action potentials (Stuart and Spruston 2015; Larkum et al. 2022). These additional nonlinearities can significantly increase the computational capacity of single neurons (Poirazi, Brannon, and Mel 2003; Poirazi and Papoutsi 2020). Thus, it might be crucial to take the dendritic nonlinearities into account to understand the full capacity of the brain.

Conversely, if one considers only network dynamics, it is possible to reduce multi-compartment models to point neurons while approximately preserving the dynamics (Rössert et al. 2016; Billeh et al. 2020).

Glial Cells

One hemisphere of human cortex consists of 6.2 billion neurons but even more glial cells (Azevedo et al. 2009; Shapson-Coe et al. 2021). Most glial cells do not produce electric impulses and were thought to be merely supporter cells for the neurons, for example to myelinate axons, but there is evidence that they can also directly affect the activity of neurons (Fields et al. 2014). More generally, it might be necessary to take the wide variety of neuromodulators, and their spatial spread through volume transmission, into account because they bypass the synapse-based connectivity but still alter the activity.

Action-Perception Loop

Brains do not exist in isolation. They are embodied, in the literal sense, and interact with their environment, thereby affecting the sensory input. A prominent theory that takes the interplay between action and perception into account is the free energy principle (Friston 2010), but see the nuanced analysis of the promises and pitfalls of the free energy principle by Buckley et al. (2017). Whether it ultimately turns out to be possible to understand specific aspects of the brain using a reductionist approach which neglects the interaction with the environment is still unclear.

3

TOOLS

> "The constructionist hypothesis breaks down when confronted with the twin difficulties of scale and complexity. The behavior of large and complex aggregates of elementary particles, it turns out, is not to be understood in terms of a simple extrapolation of the properties of a few particles. Instead, at each level of complexity entirely new properties appear [...]." (Anderson 1972)

Even if the microscopic laws of a system are fully determined, it is by no means straightforward to arrive at an understanding of the system's properties on the macroscopic scale. A striking example in physics is a broken symmetry. For example, technically, the microscopic dynamics of a gas is symmetric under time reversal; practically, this symmetry offers little insight into its macroscopic properties because the recurrence time far exceeds all relevant timescales (Boltzmann 1896). The appropriate approach to successfully bridge the scales turned out to be statistical physics (Goldenfeld 1992; Kardar 2007b; Kardar 2007a).

Neuroscience faces a similar problem—one cannot simply extrapolate from neurons to networks. Since this is again a problem involving two distinct scales, re-purposing tools from statistical physics proved to be fruitful (see for example Amit, Gutfreund, and Sompolinsky 1985; Gardner 1988; Sompolinsky, Crisanti, and Sommers 1988). The statistical physics approach is also at the heart of this thesis; this chapter provides a brief introduction to the main tools.

The foundation is based on probability theory, which is briefly summarized in Section 3.1, including an overview of large-deviation theory. Afterwards, stochastic processes, with a focus on a field-theoretical perspective, are in discussed Section 3.2. Next, point processes are introduced in Section 3.3 to treat spikes appropriately. The core of this chapter is Section 3.4 on Dynamic Mean-Field Theory—the main tool to bridge from neurons to networks. Last, the inverse problem is briefly discussed in Section 3.5.

3.1 PROBABILITY THEORY

While probability theory may or may not be the "logic of science" (Jaynes 2003), it is certainly, and rather unsurprisingly, underlying statistical physics (Mezard and Montanari 2009).

We follow (Stratonovich 1967) for this very brief summary of probability theory. Let us consider a collection of random variables $\xi_1, \dots, \xi_n$. The associated n-dimensional probability density is

$$p(x_1, \dots, x_n) = \langle \delta(\xi_1 - x_1) \dots \delta(\xi_n - x_n) \rangle. \tag{3.1}$$

With $p(x_1, \dots, x_n)$, we can compute the average of an arbitrary function $f(\xi_1, \dots, \xi_n)$,

$$\langle f(\xi_1, \dots, \xi_n) \rangle = \int d\xi_1 \cdots \int d\xi_n \, p(\xi_1, \dots, \xi_n) f(\xi_1, \dots, \xi_n). \tag{3.2}$$

One important choice is $f(\xi_1, \dots, \xi_n) = 1$ which leads to the *normalization condition*

$$\int d\xi_1 \cdots \int d\xi_n \, p(\xi_1, \dots, \xi_n) = 1. \tag{3.3}$$

Another important choice is $f(\xi_1, \dots, \xi_n) = \exp[i(k_1\xi_1 + \cdots + k_n\xi_n)]$ which gives rise to the *characteristic function*

$$\phi(k_1, \dots, k_n) = \int d\xi_1 \cdots \int d\xi_n \, p(\xi_1, \dots, \xi_n) e^{i(k_1\xi_1 + \cdots + k_n\xi_n)}. \tag{3.4}$$

From the characteristic function, we can obtain the *moments* using

$$\langle \xi_1 \dots \xi_n \rangle = \frac{1}{i^n} \frac{\partial^n \phi(k_1, \dots, k_n)}{\partial k_1 \dots \partial k_n} \bigg|_{k_1 = \cdots = k_n = 0} \tag{3.5}$$

and the *cumulants* using

$$\langle\langle \xi_1 \dots \xi_n \rangle\rangle = \frac{1}{i^n} \frac{\partial^n \ln \phi(k_1, \dots, k_n)}{\partial k_1 \dots \partial k_n} \bigg|_{k_1 = \cdots = k_n = 0}. \tag{3.6}$$

Furthermore, we can recover the density from the characteristic function

$$p(\xi_1, \dots, \xi_n) = \int \frac{dk_1}{2\pi} \cdots \int \frac{dk_n}{2\pi} \, \phi(k_1, \dots, k_n) e^{-i(k_1\xi_1 + \cdots + k_n\xi_n)}; \tag{3.7}$$

the characteristic function is simply the Fourier transformation of the density.

The density $p(x_1, \dots, x_n)$ contains the information about all random variables $\xi_1, \dots, \xi_n$. In particular, it contains the information about a subset $\xi_1, \dots, \xi_k$, $k < n$, of the random variables. The corresponding density follows by *marginalizing* the density:

$$p(\xi_1, \dots, \xi_k) = \int d\xi_{k+1} \cdots \int d\xi_n p(\xi_1, \dots, \xi_n). \tag{3.8}$$

Furthermore, information about a subset of the random variables $\xi_{k+1}, \dots, \xi_n$ might affect the remaining random variables $\xi_1, \dots, \xi_k$. This is quantified by the *conditional* density

$$p(\xi_1, \dots, \xi_k \,|\, \xi_{k+1}, \dots, \xi_n) = \frac{p(\xi_1, \dots, \xi_n)}{p(\xi_{k+1}, \dots, \xi_n)}. \tag{3.9}$$

Note that $p(\xi_1, \dots, \xi_k \,|\, \xi_{k+1}, \dots, \xi_n)$ is again a probability density; in particular it is normalized due to (3.8). If the the information about $\xi_{k+1}, \dots, \xi_n$ does not affect the information about $\xi_1, \dots, \xi_k$ they are *independent*. In this case the conditional density and the marginal density coincide,

$$p(\xi_1, \dots, \xi_k \,|\, \xi_{k+1}, \dots, \xi_n) = p(\xi_1, \dots, \xi_k). \tag{3.10}$$

In combination with (3.9), we see that the densities of the independent random variables factorize.

3.1.1 *Multivariate Gaussian*

Arguably the most important probability distribution is the Gaussian distribution. Using vector notation $(\boldsymbol{\xi})_i \equiv \xi_i$, its density is

$$p(\boldsymbol{\xi}) = \frac{1}{\sqrt{\det(2\pi \boldsymbol{K})}} \exp\Big(-\frac{1}{2} (\boldsymbol{\xi} - \boldsymbol{\mu})^{\mathsf{T}} \boldsymbol{K}^{-1} (\boldsymbol{\xi} - \boldsymbol{\mu}) \Big) \tag{3.11}$$

where $\mu_i = \langle \xi_i \rangle$ denotes the mean and K^{-1} the inverse of the *covariance matrix* $K_{ij} = \langle\langle \xi_i \xi_j \rangle\rangle$. For convenience, we will use the notation

$$\boldsymbol{\xi} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{K}) \tag{3.12}$$

to designate that $\boldsymbol{\xi}$ is Gaussian distributed with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{K}$.

The characteristic function (3.4) of a multivariate Gaussian is

$$\phi(\boldsymbol{k}) = \exp\left(i \boldsymbol{\mu}^T \boldsymbol{k} - \frac{1}{2} \boldsymbol{k}^T \boldsymbol{K} \boldsymbol{k} \right). \tag{3.13}$$

The observation that the exponent is at most quadratic in $\boldsymbol{k}$, in combination with (3.6), immediately shows that only cumulants up to order two—the mean and the variance—are non-vanishing. Thus, moments

of all orders can be expressed through the first two cumulants (*Isserlis'* or *Wick's theorem*). The finite number of cumulants is unique to the Gaussian (Marcinkiewicz 1939): if the exponent of the characteristic function is a polynomial, it is a polynomial of at most order two (*Marcinkiewicz theorem*).

For the multivariate Gaussian, both marginal and conditional distributions can be calculated analytically (Williams and Rasmussen 2006). Let

$$\boldsymbol{x}, \boldsymbol{y} \sim \mathcal{N}\left(\begin{pmatrix}\boldsymbol{\mu}_x \\ \boldsymbol{\mu}_y\end{pmatrix}, \begin{pmatrix}\boldsymbol{K}_x & \boldsymbol{K}_{xy} \\ \boldsymbol{K}_{xy}^T & \boldsymbol{K}_y\end{pmatrix}\right), \tag{3.14}$$

i.e., $\boldsymbol{x}$ and $\boldsymbol{y}$ are jointly Gaussian with corresponding means and covariance matrices. The marginal distribution of $\boldsymbol{x}$ is

$$\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}_x, \boldsymbol{K}_x) \tag{3.15}$$

which looks almost trivial but is surprisingly cumbersome to show. The conditional distribution of $\boldsymbol{x}$ given $\boldsymbol{y}$ is

$$\boldsymbol{x}|\boldsymbol{y} \sim \mathcal{N}\left(\boldsymbol{\mu}_x + \boldsymbol{K}_{xy}\boldsymbol{K}_y^{-1}(\boldsymbol{y} - \boldsymbol{\mu}_y), \boldsymbol{K}_x - \boldsymbol{K}_{xy}\boldsymbol{K}_y^{-1}\boldsymbol{K}_{xy}^T\right). \tag{3.16}$$

If the marginal distribution of $\boldsymbol{y}$ is degenerate, $\boldsymbol{K}_y^{-1}$ is the generalized inverse of $\boldsymbol{K}_y$.

Additionally, the Gaussian distribution is *stable*: a linear combination of two independent Gaussian random variables is still Gaussian (Papoulis and Pillai 2002). The converse is also true (Cramér 1936): if ξ_1 and ξ_2 are independent and their sum is Gaussian then ξ_1 and ξ_2 are Gaussian (*Cramér's decomposition theorem*).

3.1.2 *Large-Deviation Theory*

The importance of the Gaussian distribution is largely due to the *central limit theorem*: given a sequence of independent and identically distributed (i.i.d.) random variables $\xi_1, \ldots, \xi_N$ with mean μ and variance $\sigma^2 < \infty$, the scaled sample average $\bar{\xi}_N = \sqrt{N}(\frac{1}{N}\sum_i \xi_i - \mu)/\sigma$ is asymptotically Gaussian, $\lim_{N\to\infty} \bar{\xi}_N \sim \mathcal{N}(0, 1)$. While the central limit theorem captures small fluctuations of $O(\sigma/\sqrt{N})$, it does not make a statement about the probability of a rare but large deviations from the sample mean. This is the topic of large-deviation theory (Varadhan 2008; Touchette 2009; Mezard and Montanari 2009; Dembo and Zeitouni 2010).

Let us consider a concrete example by Varadhan (2008): tossing a fair coin $N \gg 1$ times and counting the relative fraction of heads $x \in \{0, 1/N, \ldots, 1\}$. The corresponding probability is determined by the binomial distribution

$$p(x \mid N) = \binom{N}{xN}\frac{1}{2^N}. \tag{3.17}$$

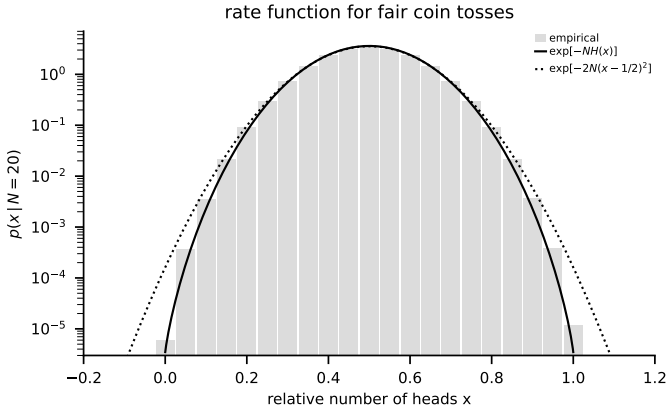

Figure 3.1: Rate function for the relative number of heads x in $N = 20$ tosses of a fair coin. Empirical distribution based (gray bars) based on 10^7 samples. Rate function (solid black line) and its quadratic approximation (dashed black line). Both analytical distributions are normalized.

Using Stirling's approximation $\ln N! = N \ln N - N + O(\ln N)$, we get

$$\ln p(x \mid N) = -NH(x) + O(\ln N), \tag{3.18}$$

$$H(x) = x \ln x + (1-x)\ln(1-x) + \ln 2. \tag{3.19}$$

Put differently, we have $p(x \mid N) \approx \exp[-NH(x)]$ up to algebraic corrections in N. Importantly, we did not put any restrictions on x, thus this results also holds for fluctuations far from the mean value $x = 1/2$ (see Figure 3.1 on page 29). For small fluctuations $\delta x = x - 1/2$, we can expand $H(x) = 2(x-1/2)^2 + O(\delta x^4)$ and we see that we recover a Gaussian $p(x \mid N) \approx \exp[-2N(x-1/2)^2]$ with mean $1/2$ and variance $1/4N$. This is equivalent to the result of the central limit theorem applied to a sum of independent Bernoulli variables with mean $p = 1/2$ and variance $p(1-p) = 1/4$.

In the coin tossing example, we see explicitly how a distribution with exponentially suppressed large fluctuations $p(\xi \mid N) \approx \exp[-NH(\xi)]$ arises and how we recover the result of the central limit theorem by expanding $H(\xi)$ around the mean. More generally, a large deviation result is of the form

$$-\lim_{N\to\infty} \frac{1}{N} \ln p(\xi \mid N) = H(\xi) \tag{3.20}$$

where we assume that the limit exists. Here, the limit $N \to \infty$ ensures that $o(N)$ contributions vanish and the remaining quantity $H(\xi)$ is called the *rate function*. Note that (3.20) is a heuristic statement; the mathematically rigorous statement involves lower and upper bounds on the probability of closed and open sets, respectively (Varadhan

2008). In the following, we stay on this heuristic level. Note furthermore that (3.20) can be used for discrete as well as continuous random variables (Touchette 2009).

Gärtner-Ellis Theorem

In the above example, we used Stirling's approximation which can be derived using a saddle-point (or, more precisely, Laplace) approximation

$$\ln \int dx\, g(x) e^{-Nf(x)} = -N \inf_x f(x) + O(\ln N) \tag{3.21}$$

of $N! = N^{N+1} \int_0^\infty dx\, e^{-N(x-\ln x)}$ (Bender and Orszag 1999). The fact that the saddle-point approximation provides the leading order behavior required in (3.20) suggests a connection between large-deviation theory and the saddle-point approximation—this connection is made precise by the Gärtner-Ellis theorem:

$$H(x) = \sup_{\tilde{x}} [\tilde{x}x - \lambda(\tilde{x})] \tag{3.22}$$

where $\lambda(\tilde{x})$ denotes the *scaled cumulant generating function*

$$\lambda(\tilde{x}) = \lim_{N\to\infty} \frac{1}{N} \ln \langle e^{N\tilde{x}\xi} \rangle. \tag{3.23}$$

We see immediately that $H(x)$ is the Legendre-Fenchel transform of $\lambda(\tilde{x})$. An important caveat of the Gärtner-Ellis theorem is that it only yields the convex hull of $H(x)$ for a multimodal distribution (Touchette 2009).

To highlight the connection to the saddle-point approximation, we follow the heuristic derivation of the Gärtner-Ellis theorem by Touchette (2009). We start from (3.7) and change variables to $ik = N\tilde{x}$, which yields

$$p(x) = \frac{N}{2\pi i} \int_{-i\infty}^{i\infty} d\tilde{x}\, \langle e^{N\tilde{x}\xi} \rangle e^{-N\tilde{x}x}. \tag{3.24}$$

Using the scaled cumulant generating function, we get

$$\begin{aligned} \ln p(x) &= \ln \int_{-i\infty}^{i\infty} d(-i\tilde{x})\, e^{-N[\tilde{x}x - \lambda(\tilde{x})]} + o(N) \\ &= -N \sup_x [\tilde{x}x - \lambda(\tilde{x})] + o(N) \end{aligned} \tag{3.25}$$

where we replaced $\frac{1}{N} \ln \langle e^{N\tilde{x}\xi} \rangle$ by $\lambda(\tilde{x})$ with an $o(N)$ error and employed a saddle-point approximation. The final result looks deceptively simple but a saddle-point approximation involves several technical intricacies, see the excellent examples in (Bender and Orszag 1999, Chapter 6). For the above calculation, it is assumed that $\lambda(\tilde{x})$ is analytic and the Cauchy-Riemann equations were used (Touchette 2009, Appendix

C). Comparing the result with (3.20), we arrive at the Gärtner-Ellis theorem and we see that the supremum in (3.22) can be seen as a consequence of a saddle-point approximation.

The Gärtner-Ellis theorem provides a constructive way to compute rate functions (or at least their convex hull). For example, let us consider again the sample average $x = \frac{1}{N}\sum_{i=1}^{N}\xi_i$ of i.i.d. random variables. The corresponding scaled cumulant generating function is

$$\lambda(\tilde{x}) = \lim_{N\to\infty}\frac{1}{N}\ln\langle e^{\tilde{x}\sum_{i=1}^{N}\xi_i}\rangle = \ln\langle e^{\tilde{x}\xi}\rangle, \tag{3.26}$$

and the rate function is the Legendre-Fenchel transform of the cumulant generating function of ξ (*Cramér's theorem*). Assuming a Bernoulli distribution with $\lambda(\tilde{x}) = \ln(1+e^{\tilde{x}}) - \ln 2$, we recover (3.19).

Sanov's Theorem

The restriction to unimodal distributions of the Gärtner-Ellis theorem seems quite severe. But there is an interesting detour which allows to treat the multimodal case as well. To this end, let us again consider i.i.d. random variables $\xi_1, \ldots, \xi_N$ and the associated *empirical measure*

$$\mu(x) = \frac{1}{N}\sum_{i=1}^{N}\delta(x-\xi_i). \tag{3.27}$$

Intuitively, the empirical measure describes a histogram with 'infinitely narrow' bins and it can be used to calculate arbitrary empirical averages, $\int dx\,\mu(x)f(x) = \frac{1}{N}\sum_{i=1}^{N}f(\xi_i)$. Now we can ask: how is $\mu(x)$ distributed across different realizations of the random variables? We know that sample histograms of i.i.d. random variables roughly look like the underlying distribution $p(\xi)$, thus we might expect that $\mu(x)$ approaches $p(\xi)$ asymptotically. This intuition is made rigorous in *Sanov's theorem*: the rate function of $\mu(x)$ is its Kullback-Leibler divergence with respect to $p(\xi)$,

$$H[\mu] = \int dx\,\mu(x)\ln\frac{\mu(x)}{p(x)}. \tag{3.28}$$

We use the notation $H[\mu]$ to denote a functional, i.e., an object that maps a function $\mu(x)$ to the reals.

Note that $\mu(x)$ converges to $p(x)$ whether or not the latter is multimodal. The reason why this approach circumvents the problem of multimodality is that the distribution of $\mu(x)$—a 'distribution of a distribution'—is unimodal. This step towards considering 'distributions of distributions' is also called level-2 (Touchette 2009).

To provide an intuition about Sanov's theorem, let us sketch a heuristic derivation based on the Gärtner-Ellis theorem. First, we note that

$$\lambda[k] = \lim_{N\to\infty}\frac{1}{N}\ln\langle e^{N\int dx\,k(x)\mu(x)}\rangle = \ln\langle e^{k(\xi)}\rangle \tag{3.29}$$

is the scaled cumulant generating functional of the empirical measure. To proceed, we need the functional derivative

$$\frac{\delta\lambda[k]}{\delta k(x)} = \frac{p(x)e^{k(x)}}{\langle e^{k(\xi)}\rangle} \tag{3.30}$$

where we used the chain rule and $\frac{\delta k(\xi)}{\delta k(x)} = \delta(\xi - x)$. Using this to perform the Legendre-Fenchel transform $\sup_k\{\int dx\, k(x)\mu(x) - \lambda[k]\}$, we arrive at (3.28).

We can obtain the sample mean from the empirical measure using $\int dx\, \mu(x)x = \frac{1}{N}\sum_{i=1}^{N} \xi_i$. Thus, Sanov's theorem is more general than, for example, Cramér's theorem. The relation between the two is given by the *contraction principle*:

$$H(x) = \inf_{\mu \text{ s.t. } \int dy\, \mu(y)y = x} H[\mu]. \tag{3.31}$$

Of course we can replace the sample mean $\int dx\, \mu(x)x$ by an arbitrary empirical average $\int dx\, \mu(x)f(x)$ to go beyond Cramér's theorem.

3.2 STOCHASTIC PROCESSES

Frequently, we encounter random, or noisy, processes that evolve continuously in time. There are plenty of examples for such processes in biology, from the motion of bacteria under the constant bombardment of the surrounding molecules (Purcell 1977; Berg 1993) to the fluctuating membrane potential of a neuron due to the massive synaptic input (Gerstein and Mandelbrot 1964; Shadlen and Newsome 1994). The mathematical description of these random processes leads to the notion of a *stochastic process* or random function $\xi(t)$ (Stratonovich 1967; Risken 1996; Van Kampen 2007; Gardiner 2009).

3.2.1 *Field-Theoretical Formulation*

Stochastic processes can be seen as an infinite-dimensional collection of random variables: for any fixed set of points in time $t_1, \ldots, t_n$ with n arbitrary, the corresponding values of the process $\xi(t_1), \ldots, \xi(t_n)$ are governed by a multivariate distribution $p(\xi(t_1), \ldots, \xi(t_n))$; since there are infinitely many points on a continuous interval, this leads to an 'infinite dimensional distribution'. The natural language for such objects are path integrals (Feynman, Hibbs, and Styer 2010; Kleinert 2009) or, in different words, field theory (Helias and Dahmen 2020).

The first sentence in Kleinert's tome on path integrals (Kleinert 2009) is: "Path integrals deal with fluctuating line-like structures."

Characteristic Functionals

Similar to random variables, stochastic processes can either be specified using the probability distribution functional $p[\xi]$ or the *characteristic functional* (Stratonovich 1967; Van Kampen 2007)

$$\phi[k] = \left\langle \exp\left(i \int_0^T dt\, k(t)\xi(t)\right)\right\rangle \tag{3.32}$$

where $k(t)$ is an arbitrary test function. For stochastic processes, $\phi[k]$ is very convenient because it is normalized, $\phi[0] = 1$, while $p[\xi]$ typically contains an infinite normalization constant. From the characteristic functional, one can obtain either the moments

In a field-theoretical context, this object is usually called (moment) generating functional (Zinn-Justin 1996; Kleinert 2009).

$$\langle \xi(t_1) \dots \xi(t_n)\rangle = \frac{1}{i^n} \frac{\delta^n \phi[k]}{\delta k(t_1) \dots \delta k(t_n)}\bigg|_{k=0} \tag{3.33}$$

or the cumulants

$$\langle\langle \xi(t_1) \dots \xi(t_n)\rangle\rangle = \frac{1}{i^n} \frac{\delta^n \ln \phi[k]}{\delta k(t_1) \dots \delta k(t_n)}\bigg|_{k=0} \tag{3.34}$$

using functional derivatives. Conversely, the characteristic functional can be written as

$$\phi[k] = \exp\left(\sum_{n=1}^{\infty} \frac{i^n}{n!} \int_0^T dt_1 \dots dt_n \, \langle\langle \xi(t_1) \dots \xi(t_n)\rangle\rangle k(t_1) \dots k(t_n)\right) \tag{3.35}$$

because (3.34) defines its expansion in k.

More down to earth—without functionals—a stochastic process can also be defined by the *hierarchy of distributions* $p(\xi(t_1), \dots, \xi(t_n))$ for increasing n (Stratonovich 1967; Van Kampen 2007). These distributions have to be consistent: the marginal of $p(\xi(t_1), \dots, \xi(t_{n+k}))$ has to equal $p(\xi(t_1), \dots, \xi(t_n))$.

A stochastic process is called *stationary* if all its properties (moments, cumulants, marginal distributions) are invariant under a shift in time, e.g.,

$$\langle\langle \xi(t_1 + \tau) \dots \xi(t_s + \tau)\rangle\rangle = \langle\langle \xi(t_1) \dots \xi(t_s)\rangle\rangle.$$

In particular, this implies that the *mean* $\mu(t_1) = \langle\langle \xi(t_1)\rangle\rangle$ is constant, $\mu(t) \equiv \mu$, and that the *correlation function* $C(t_1, t_2) = \langle\langle \xi(t_1)\xi(t_2)\rangle\rangle$ depends only on the time difference, $C(t_1, t_2) \equiv C(t_2 - t_1)$. In some situations, for example for oscillatory processes, the Fourier transform of the stationary correlation function is of interest. The Fourier transformed correlation function is equal to the power spectrum (*Wiener-Khinchin theorem*).

Gaussian Processes

An important, and usually well tractable, stochastic process is the *Gaussian process* (GP) which is defined by vanishing cumulants beyond the second one (Stratonovich 1967; Van Kampen 2007; Rasmussen and Williams 2006):

$$\phi[k] = \exp\left(i\int_0^T dt_1\,\mu(t_1)k(t_1) - \frac{1}{2}\int_0^T dt_1 dt_2\,k(t_1)C(t_1,t_2)k(t_2)\right). \tag{3.36}$$

Alternatively, a GP is defined by Gaussian marginals $p(\xi(t_1),\dots,\xi(t_n))$ for all sets of time points $t_1,\dots,t_n$.

The most common GP is a *Gaussian white noise* with $C(\tau) = 2D\delta(\tau)$. It is called white because its power spectrum is flat, similar to the spectrum of white light.

The generalization to multidimensional GPs is straightforward: one simply needs to replace $\mu(t_1)k(t_1)$ by $\sum_i \mu_i(t_1)k_i(t_1)$ to account for the means of the GPs and $k(t_1)C(t_1,t_2)k(t_2)$ by $\sum_{i,j} k_i(t_1)C_{ij}(t_1,t_2)k_j(t_2)$ to account for their correlations.

3.2.2 *Stochastic Differential Equations*

A typical application of a stochastic process ξ is to model an external influence—for example the thermal motion of molecules surrounding a Brownian particle (Langevin 1908)—which affects a second process x governed by a differential equation

$$\dot{x} = f(x) + g(x)\xi. \tag{3.37}$$

Due to the stochastic term on the right hand side, this is a *stochastic differential equation* (SDE) or Langevin equation (Stratonovich 1967; Risken 1996; Van Kampen 2007; Gardiner 2009).

If $g(x) = \text{const.}$, the SDE is called *additive*, otherwise it is *multiplicative*. For $g(x) \neq 0$, a multiplicative SDE can be transformed into the additive SDE $\dot{y} = \frac{f(y)}{g(y)} + \xi$ with the transformation $dy = \frac{dx}{g(x)}$, or $y(x) = \int^x du\,\frac{1}{g(u)}$; thus, we restrict the following discussion to the additive case. If ξ is a Gaussian white noise, this transformation assumes implicitly the Stratonovich interpretation (Van Kampen 2007). Additionally, in the non-white case, we assume a correlation-free preparation (Hänggi and Jung 1995), i.e., no correlation between x and ξ prior to $t = 0$, throughout.

Probability Density Functional

An SDE specifies the transformation of the stochastic process ξ to the stochastic process x. Thus, the probability density functional of x is

$$p[x] = \left|\frac{\mathcal{D}\xi}{\mathcal{D}x}\right| p_\xi[\xi[x]] \tag{3.38}$$

where $\left|\frac{\mathcal{D}\xi}{\mathcal{D}x}\right|$ denotes the Jacobian of the transformation (3.37). The Jacobian turns out to depend on the discretization parameter λ where $\lambda = 0$ corresponds to the Itô interpretation and $\lambda = 1/2$ to the Stratonovich interpretation (Stratonovich 1989):

All results should only depend on λ in the white noise case; this is indeed the case but highly cumbersome to show (Stratonovich 1989).

$$\left|\frac{\mathcal{D}\xi}{\mathcal{D}x}\right| = \exp\Big(-\lambda\int_0^T dt\, f'(x(t))\Big). \tag{3.39}$$

While setting $\lambda = 0$ looks tempting, $\lambda = 1/2$ ensures the simple transformation property (3.38) under a nonlinear change of variables $y = y(x)$ (Stratonovich 1989). Expressing $p_\xi[\xi]$ using the inverse transform of the characteristic functional, we arrive at

$$p[x] \propto \int \mathcal{D}\tilde{x}\, e^{-i\int_0^T dt\,\tilde{x}(t)[\dot{x}(t)-f(x(t))]-\lambda\int_0^T dt\, f'(x(t))}\phi_\xi[\tilde{x}]. \tag{3.40}$$

If ξ is zero-mean GP (a nonzero mean can be absorbed in $f(x)$), the inverse transform can be performed and we arrive at

$$\begin{aligned} p[x] &\propto \exp\Big(-S[x]-\lambda\int_0^T dt\, f'(x(t))\Big) \qquad (3.41)\\ S[x] &= \frac{1}{2}\int_0^T dt_1 dt_2\,[\dot{x}(t_1)-f(x(t_1))]C^{-1}(t_1,t_2)[\dot{x}(t_2)-f(x(t_2))] \end{aligned}$$

where $\int_0^T C^{-1}(t_1,s)C(s,t_2) = \delta(t_2-t_1)$. For white noise with correlation function $C(\tau) = 2D\delta(\tau)$, the inverse is $C^{-1}(\tau) = \frac{1}{2D}\delta(\tau)$ and we obtain the Onsager-Machlup functional

$$S_{\mathrm{OM}}[x] = \frac{1}{4D}\int_0^T dt\,[\dot{x}(t)-f(x(t))]^2. \tag{3.42}$$

In order to avoid the infinite normalization constant hidden in the proportionality signs, one can consider ratios of probability density functionals akin to Radon-Nikodym derivatives (Stratonovich 1989).

From the probability density functional, we can obtain, for example, the transition probability to start at $x(0) = x_0$ and to end up at $x(T) = x$ using

$$p_T(x\,|\,x_0) = \int_{x(0)=x_0}^{x(T)=x}\mathcal{D}x\, p[x] \tag{3.43}$$

where $\int_{x(0)=x_0}^{x(T)=x}\mathcal{D}x$ denotes a path integral constrained to processes that fulfill the boundary conditions. For weak Gaussian white noise $D \ll 1$, the exponent is dominated by $S_{\mathrm{OM}}[x]$ and a saddle-point, or WKB, approximation yields

$$p_T(x\,|\,x_0) \approx \exp\Big(-\inf_{y\,\mathrm{s.t.}\,y(0)=x_0,y(T)=x} S_{\mathrm{OM}}[y]\Big). \tag{3.44}$$

Note the similarity to the contraction principle (3.31); indeed, the same result can be obtained in the framework of large-deviation theory (Touchette 2009, and references therein).

3.3 POINT PROCESSES

The language of neurons—the spikes—is not continuous. Thus, we need to go beyond continuous-valued stochastic processes to *point process* or random sets of points $\{t_1, \ldots, t_n\}$ (Stratonovich 1967; Van Kampen 2007). For simplicity, we assume throughout that there are no coinciding events, $t_i \neq t_j$ for $i \neq j$.

Generating Functionals

Stochastic processes are fully specified by the characteristic functional (3.32); the analogous quantity for point processes $\{t_1, \ldots, t_s\}$ on a fixed interval $t_i \in [0, T]$ is the *generating functional* (Kuznetsov and Stratonovich 1965; Stratonovich 1967; Van Kampen 2007)

We stick to the names by Stratonovich (1967) although they are somewhat ambiguous (Van Kampen 2007 does not offer a better solution).

$$\ell[k] = \Big\langle \prod_{i=1}^{s} (1 + k(t_i)) \Big\rangle. \tag{3.45}$$

The functional derivatives determine the *distribution functions*

$$f_n(t_1, \ldots, t_n) = \frac{\delta^n \ell[k]}{\delta k(t_1) \ldots \delta k(t_n)}\Big|_{k=0}. \tag{3.46}$$

Performing the functional derivatives of the generating functional, we get $\frac{\delta^n \ell[k]}{\delta k(t_1) \ldots \delta k(t_n)}\Big|_{k=0} = \langle \sum_{i_1 \neq \ldots \neq i_n} \delta(t_1 - t_{i_1}) \ldots \delta(t_n - t_{i_n}) \rangle$ and we see that $f_n(t_1, \ldots, t_n) dt_1 \ldots dt_n$ is the probability to observe an event in each of $(t_1, t_1 + dt)$ to $(t_n, t_n + dt_n)$. Note that the distribution functions are not normalized, for example $\int_0^T dt_1 \, f_1(t_1) = \langle n \rangle$ and $\int_0^T dt_1 dt_2 \, f_2(t_1, t_2) = \langle n(n-1) \rangle$. The derivatives of the logarithm yield the *correlation functions*

$$g_n(t_1, \ldots, t_n) = \frac{\delta^n \ln \ell[k]}{\delta k(t_1) \ldots \delta k(t_n)}\Big|_{k=0} \tag{3.47}$$

which relate to the distribution functions as cumulants relate to moments. Furthermore, the correlation functions define the series expansion of $\ln \ell[k]$ which leads to

$$\ell[k] = \exp\left(\sum_{n=1}^{\infty} \frac{1}{n!} \int_0^T dt_1 \ldots dt_n \, g_n(t_1, \ldots, t_n) k(t_1) \ldots k(t_n) \right). \tag{3.48}$$

We see that the $g_n(t_1, \ldots, t_n)$, or the $f_n(t_1, \ldots, t_n)$, fully specify the generating functional and hence the point process. A point process is called *stationary* if all $f_n(t_1, \ldots, t_n)$, or equivalently all $g_n(t_1, \ldots, t_n)$, are invariant under time shifts $t_i \to t_i + \tau$. For a Poisson process, this means $g_1(t_1) \equiv g_1$, i.e., it is homogeneous.

The most simple example of a generating functional is the case of vanishing $g_n(t_1, \ldots, t_n)$ for $n \geq 2$,

$$\ell[k] = \exp\left(\int_0^T dt_1 \, g_1(t_1) k(t_1) \right). \tag{3.49}$$

The number of events in the interval $[0, T]$ has the characteristic function $\langle e^{isk} \rangle = \ell(e^{ik} - 1)$ and thus $\phi(k) = \exp\left((e^{ik} - 1) \int_0^T dt_1\, g_1(t_1)\right)$ which corresponds to a Poisson distribution with mean $\int_0^T dt_1\, g_1(t_1)$. Hence, (3.49) characterizes an inhomogeneous Poisson process.

Point processes $\{t_1, \ldots, t_s\}$ naturally give rise to the stochastic processes $\xi(t) = \sum_{i=1}^{s} \delta(t - t_i)$. The corresponding characteristic functional is

$$\phi[k] = \ell[e^{ik} - 1] \tag{3.50}$$

which can be seen by inserting the sum of Dirac deltas into (3.32). Taking again the Poisson process as an example, this leads to

$$\phi[k] = \exp\left(\int_0^T dt_1\, g_1(t_1)(e^{ik(t_1)} - 1) \right). \tag{3.51}$$

Thus, by (3.34) the Poisson process is a white process: $\langle\langle \xi(t_1)\xi(t_2) \rangle\rangle = g_1(t_1)\delta(t_2 - t_1)$. Furthermore, we get for a general point process $\langle\langle \xi(t_1)\xi(t_2) \rangle\rangle = g_1(t_1)\delta(t_2 - t_1) + g_2(t_1, t_2)$, i.e., there is always a white component in the correlation function (and all higher cumulants).

3.3.1 *Renewal Processes*

In neuroscience, a frequently used point process is the renewal process (Gerstner, Kistler, et al. 2014) where the intervals between events are drawn independently from a distribution $\rho(\tau)$ (if the events are spikes, $\rho(\tau)$ is the inter-spike interval distribution). Equivalently, a renewal process can be specified by the survival probability $S(\tau) = \int_\tau^\infty ds\, \rho(s)$ or the *hazard function*

$$h(\tau) = \frac{\rho(\tau)}{S(\tau)} = -\frac{d}{d\tau} \ln S(\tau). \tag{3.52}$$

The hazard function $h(\tau)d\tau$ quantifies the conditional probability of an event in $[\tau, \tau + d\tau)$ given that no event took place in $[0, \tau)$. It is a useful quantity because it has a clear interpretation and it only needs to fulfill two basic requirements: it needs to be non-negative (by definition) and its integral needs to diverge (to ensure normalization of $\rho(\tau)$). From (3.52), we see immediately that the hazard function specifies the survival probability by $S(\tau) = \exp\left(-\int_0^\tau ds\, h(s)\right)$ and $\rho(\tau)$ by $\rho(\tau) = h(\tau)S(\tau)$.

The huge advantage of a renewal process is that a single function, e.g., $\rho(\tau)$ or $h(\tau)$, fully specifies it instead of having to consider the entire generating functional or the collection of all $f_n(t_1, \ldots, t_n)$. But this also means that all properties of the point process can be derived from this function. Indeed, the average rate of events is $\nu = \left[\int_0^\infty ds\, \rho(s)s\right]^{-1}$

and the power spectrum follows from (Stratonovich 1967; Gerstner, Kistler, et al. 2014)

$$S(\omega) = \nu \frac{1 - |\tilde{\rho}(\omega)|^2}{|1 - \tilde{\rho}(\omega)|^2} \tag{3.53}$$

where $\tilde{\rho}(\omega) = \int_0^\infty d\tau \, e^{i\omega\tau} \rho(\tau)$.

3.3.2 *Level-Crossing Processes*

Another way to generate a point process is to consider an underlying stochastic process x which (up-)crosses a certain level or threshold θ, akin to a LIF neuron. If the stochastic process generating the level crossings is sufficiently smooth, the distribution functions can be determined using the *Kac-Rice formulae* (Stratonovich 1967; Rainal 1988; Azaïs and Wschebor 2009).

The Kac-Rice formulae follow from a simple consideration. Suppose the process is at time t at position $[x, x + dx)$ below the threshold with velocity $[\dot{x}, \dot{x} + d\dot{x})$. The associated probability is $p_t(x, \dot{x}) dx d\dot{x}$. To up-cross the threshold in the interval $[t, t + dt)$ given a fixed velocity $\dot{x}$, the process needs to be in the interval $[\theta - \dot{x}dt, \theta)$, i.e., a maximum of $dx = \dot{x}dt$ below the threshold. Thus, the probability of a threshold crossing is $p_t(\theta, \dot{x}) \dot{x} dt d\dot{x}$, and we arrive at the first Kac-Rice formula

$$f_1(t_1) = \int_0^\infty d\dot{x} \, p_{t_1}(\theta, \dot{x}) \dot{x} \tag{3.54}$$

where we integrate only over positive velocities to take into account that it is an up-crossing. These arguments generalize for multiple up-crossings to

$$f_n(t_1, \ldots, t_n) = \int_0^\infty d\dot{x}_1 \ldots d\dot{x}_n \, p_{t_1,\ldots,t_n}(\theta, \dot{x}_1, \ldots, \theta, \dot{x}_n) \dot{x}_1 \ldots \dot{x}_n \tag{3.55}$$

where $p_{t_1,\ldots,t_n}(x_1, \dot{x}_1, \ldots, x_n, \dot{x}_n)$ denotes the joint probability to be at times $t_1, \ldots t_n$ at positions $x_1, \ldots, x_n$ with velocities $\dot{x}_1, \ldots, \dot{x}_n$. (3.55) fully determines the point process—in principle. In practice, the integrals quickly become intractable with increasing n.

For a stationary GP with zero mean, Rice (1945) famously derived the up-crossing rate

$$f_1 = \frac{1}{2\pi} \frac{\sigma_{\dot{x}}}{\sigma_x} \exp\Big(- \frac{\theta^2}{2\sigma_{\dot{x}}^2} \Big) \tag{3.56}$$

using the curious property $\langle x\dot{x} \rangle = 0$ of stationary GPs. Here, we see a hallmark of the abovementioned smoothness condition: the variance of the velocity $\sigma_{\dot{x}}$ needs to be finite, excluding for example Ornstein-Uhlenbeck processes. This condition is equivalent to a finite second spectral moment (Azaïs and Wschebor 2009) or a finite second derivative of the correlation function at the origin.

3.4 DYNAMIC MEAN-FIELD THEORY

Dynamic Mean-Field Theory (DMFT) is the central tool in this thesis to bridge between the neuron- and the network-level and it draws heavily on the previously introduced tools.

On an intuitive level, DMFT relies on the approximation of the recurrent input to a neuron by a stochastic process. Importantly, this approximation is not ad-hoc but results from a series of controlled steps. In the spin-glass literature, DMFT was pioneered by Sompolinsky and Zippelius (1982) and it entered neuroscience through the seminal work of Sompolinsky, Crisanti, and Sommers (1988).

Below, we briefly derive the central DMFT result in the most simple setup of a fully connected, zero-mean Gaussian network (for a textbook see Helias and Dahmen 2020). The result can be generalized in various directions, for example taking different inputs into account (Chapter 4 and references therein), using multiple populations (Chapter 5 and references therein), or considering sparse networks (Chapter 6 and references therein).

DMFT describes the single-neuron statistics. An important orthogonal line of work relies on the population activity as the central quantity (see Gerstner, Kistler, et al. 2014); a recent example including finite-size corrections is the work by Schwalger, Deger, and Gerstner (2017). The latter approach is not used in this thesis (except for trivial cases where the distinction is irrelevant), hence it is not further discussed in the remainder of this introduction.

3.4.1 *Model-Independent Formulation*

We follow along the lines of Keup et al. (2021) to derive a DMFT which is independent of the dynamics of the neurons. To this end, we assume that the neuron dynamics are quantified by a distribution $p[x_i \,|\, \eta_i]$ which maps the input η_i to the output x_i. Additionally, we assume that neurons are exclusively coupled through the recurrent input

$$\eta_i(t) = \sum_{j=1}^{N} J_{ij} x_j(t). \tag{3.57}$$

For simplicity, we consider only fully connected Gaussian networks with $J_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, g^2/N)$. We use $\boldsymbol{\eta} = \boldsymbol{J}\boldsymbol{x}$ and $x \cdot y = \int_0^T dt\, x(t)y(t)$ as well as $A \cdot B = \int_0^T dt_1 dt_2\, A(t_1, t_2)B(t_1, t_2)$ to ease the notation.

The neuron dynamics factorize if the inputs $\boldsymbol{\eta}$ are given and we can split the probability density functional of the full system into

$$p[\boldsymbol{x} \,|\, \boldsymbol{J}] = \int \mathcal{D}\boldsymbol{\eta}\, p[\boldsymbol{\eta} \,|\, \boldsymbol{J}, \boldsymbol{x}] \prod_{i=1}^{N} p[x_i \,|\, \eta_i] \tag{3.58}$$

with $p[\boldsymbol{\eta} \mid \boldsymbol{J}, \boldsymbol{x}] = \delta[\boldsymbol{\eta} - \boldsymbol{J}\boldsymbol{x}]$. Now we marginalize the connectivity (we perform the *disorder average*), i.e., we consider the distribution $p[\boldsymbol{x}] = \int d\boldsymbol{J}\, p[\boldsymbol{x} \mid \boldsymbol{J}]p(\boldsymbol{J})$ across the ensemble of connectivities. The marginalization only affects the constraint $\delta[\boldsymbol{\eta} - \boldsymbol{J}\boldsymbol{x}]$; because $\boldsymbol{\eta}$ and $\boldsymbol{J}$ are linearly related and $\boldsymbol{J}$ is Gaussian, $\boldsymbol{\eta}$ is Gaussian as well with $\langle \eta_i(t) \rangle = 0$ and

$$\langle \eta_i(t_1)\eta_j(t_2) \rangle = \delta_{ij} \frac{g^2}{N} \sum_{k=1}^{N} x_k(t_1)x_k(t_2) \equiv \delta_{ij} C[\boldsymbol{x}](t_1, t_2). \tag{3.59}$$

We note that the η_i are independent, hence we arrive at

$$p[\boldsymbol{x}] = \prod_{i=1}^{N} \int \mathcal{D}\eta_i\, p[x_i \mid \eta_i] \mathrm{GP}[\eta_i \mid 0, C[\boldsymbol{x}]]. \tag{3.60}$$

Something remarkable happened here: the system almost factorizes and is only coupled through a single scalar field $C[\boldsymbol{x}]$. To fully decouple the system, we formally introduce this global constraint through a Dirac delta and obtain

$$p[\boldsymbol{x}] = \int \mathcal{D}C\, \delta[C - C[\boldsymbol{x}]] \prod_{i=1}^{N} p[x_i \mid C] \tag{3.61}$$

with the short-hand notation $p[x_i \mid C] \equiv \int \mathcal{D}\eta_i\, p[x_i \mid \eta_i] \mathrm{GP}[\eta_i \mid 0, C]$, i.e., the output of a neuron driven by a zero-mean GP with correlation function C. At this point, we need to specify the type of observables that we are interested in.

We choose arbitrary network-averaged observables, i.e., we consider the *empirical measure of the trajectories*

$$\mu[y] = \frac{1}{N} \sum_{i=1}^{N} \delta[x_i - y]. \tag{3.62}$$

We use the notation $\lambda\{k\}$ to denote an object that maps a functional $k[x]$ to a real number.

The corresponding scaled cumulant generating functional is given by (compare (3.29))

$$\lambda\{k\} = \lim_{N\to\infty} \frac{1}{N} \ln \langle e^{N \int \mathcal{D}y\, k[y]\mu[y]} \rangle = \lim_{N\to\infty} \frac{1}{N} \ln \langle e^{\sum_{i=1}^{N} k[x_i]} \rangle. \tag{3.63}$$

Using (3.61) for the average, expressing $\delta[C - C[\boldsymbol{x}]]$ with the inverse transform of the corresponding characteristic functional $\exp(i\tilde{C} \cdot C[\boldsymbol{x}])$, and rescaling $\tilde{C} \to N\tilde{C}$, we arrive at

$$\lambda\{k\} = \lim_{N\to\infty} \frac{1}{N} \ln \int \mathcal{D}C\mathcal{D}\tilde{C}\, e^{-N(i\tilde{C}\cdot C - \Omega\{k, C, \tilde{C}\})}, \tag{3.64}$$

$$\Omega\{k, C, \tilde{C}\} = \ln \int \mathcal{D}x\, p[x \mid C] e^{ig^2 x\cdot\tilde{C}\cdot x + k[x]}. \tag{3.65}$$

Here, we neglected terms which vanish in the limit $N \to \infty$. This looks very suggestive of a saddle-point approximation:

$$\lambda\{k\} = -i\tilde{C}_k \cdot C_k + \Omega\{k, C_k, \tilde{C}_k\} \tag{3.66}$$

where C_k and $\tilde{C}_k$ are determined self-consistently by the saddle-point equations $i\tilde{C} = \frac{\delta}{\delta C}\Omega\{k, C, \tilde{C}\}$ and $iC = \frac{\delta}{\delta \tilde{C}}\Omega\{k, C, \tilde{C}\}$.

The average empirical measure $\hat{\mu}$ follows by differentiating $\lambda\{k\}$ and evaluating it at zero. Due to the saddle-point equations only the partial derivative w.r.t. k survives and we get

$$\lambda'\{k\} = \frac{p[x \mid C_k] e^{ig^2 x \cdot \tilde{C}_k \cdot x + k[x]}}{\int \mathcal{D}x\, p[x \mid C_k] e^{ig^2 x \cdot \tilde{C}_k \cdot x + k[x]}}. \tag{3.67}$$

To evaluate the expression at $k = 0$, we first evaluate the saddle-point equations at $k = 0$. $\tilde{C}_0 = 0$ is the valid solution; for this choice the remaining saddle-point equation becomes

$$C_0(t_1, t_2) = g^2 \int \mathcal{D}x\, p[x \mid C_0] x(t_1) x(t_2) \equiv g^2 \langle x(t_1) x(t_2) \rangle_{C_0}. \tag{3.68}$$

With this, we arrive at

$$\hat{\mu}[x] = p[x \mid C_0] \equiv \int \mathcal{D}\eta\, p[x \mid \eta] \mathrm{GP}[\eta \mid 0, C_0]. \tag{3.69}$$

In words: the most likely empirical measure corresponds to the single-neuron dynamics driven by a zero-mean GP with self-consistent correlation function determined by (3.68).

Let us briefly recapitulate this result. We determined the average empirical measure $\hat{\mu}[x]$. From $\lambda\{k\}$, we see that its fluctuations are suppressed with $1/N$. Thus, $\hat{\mu}[x]$ is representative for the system, i.e., the system is *self-averaging*. Accordingly, we can use $\hat{\mu}[x]$ to calculate arbitrary population-averaged observables.

3.4.2 *Applications*

Firing Rate Distribution

The empirical measure is a rather abstract quantity. Thus, let us make the above result more concrete with an example: calculating the distribution of firing rates of a fully-connected, zero-mean Gaussian network of GLM neurons with kernel $\kappa(t) = \Theta(t) \exp(-t/\tau)$ and exponential nonlinearity $\lambda(t) = c_1 \exp(V(t) - \theta)$.

First, we need to determine C_0, which means we need to solve the self-consistent colored noise problem posed by (3.68). In the stationary state, the r.h.s. of (3.68) can be solved explicitly for this model (see Chapter 6):

$$C_0(\tau) = g^2 \left(\nu \delta(\tau) + \nu^2 e^{(\tilde{\kappa} * C_0)(\tau)} \right), \tag{3.70}$$

where $\tilde{\kappa}(t) = \frac{\tau}{2} \exp(-|t|/\tau)$ and $\nu \equiv \langle x \rangle_{C_0}$ denotes the network-averaged firing rate which is given by $\nu = c_1 \exp\left(\frac{1}{2} (\tilde{\kappa} * C_0)(0) - \theta \right)$. A numerical solution of (3.70) is straightforward to find using a fixed-point iteration.

With C_0 at hand, we also have the most likely empirical measure $\hat{\mu}[x]$. From $\hat{\mu}[x]$, we can calculate the distribution of firing rates: $p(\nu) = \int \mathcal{D}x\, \hat{\mu}[x]\delta(\nu - \bar{x})$ with $\bar{x} \equiv \lim_{T\to\infty} \frac{1}{T}\int_0^T dt\, x(t)$. Inserting $\hat{\mu}[x] = \int \mathcal{D}\eta\, p[x \mid \eta]\mathrm{GP}[\eta \mid 0, C_0]$, we get

$$p(\nu) = \int \mathcal{D}\eta\, p(\nu \mid \eta)\mathrm{GP}[\eta \mid 0, C_0]. \tag{3.71}$$

Thus, we need to calculate the distribution of firing rates of GLM neurons driven by a stationary GP. Due to the exponential nonlinearity, this leads to a lognormal distribution (see Chapter 6).

Diffusion Approximation

In spiking networks, an elegant method due to Amit and Brunel (1997) allows to circumvent the (typically intractable) colored noise problem if only the firing rate is of interest. For spike trains $x(t)$, $\langle x(t_1)x(t_2)\rangle_{C_0}$ always contains a Dirac delta contribution scaled by the firing rate $\nu(t_1)\delta(t_2 - t_1)$, e.g., the first term on the r.h.s. of (3.70). Neglecting the remaining colored contribution—approximating the presynaptic spike trains as Gaussian white noise—we obtain a self-consistency equation for the firing rate. This approximation is called the *diffusion approximation*.

The diffusion approximation enables one to use the Fokker-Planck equation and its rich toolbox (Risken 1996). For example, it allows not only to determine the self-consistent rate in balanced networks of LIF neurons (Amit and Brunel 1997) but also to investigate oscillatory instabilities which shape the phase diagram (Brunel 2000). A central prerequisite for these results is that the firing rate of an LIF neuron driven by white noise is known analytically (Siegert 1951; Ricciardi 1977; Fourcaud and Brunel 2002); indeed, for white-noise-driven LIF neurons also the linear response function (Brunel and Hakim 1999; Lindner and Schimansky-Geier 2001; Schuecker, Diesmann, and Helias 2015) and the output power spectrum (Lindner, Schimansky-Geier, and Longtin 2002) are known analytically. Thus, the white-noise case is much more amenable to analytical treatment—at the price of self-consistency on the level of the second order statistics.

3.5 INFERENCE

Statistical inference deals with the inverse of the problems considered thus far: determining the properties of the underlying distribution from observed data (see for example MacKay 2003; Gelman et al. 2014). Here, we take a Bayesian point of view.

The probability of the data $\{x\} \equiv \{x_1, \dots\}$ for given parameters θ is determined by the *data distribution* $p(\{x\} \mid \theta)$ which we assume to be known. Using Bayes' theorem, we get

$$p(\theta \mid \{x\}) = \frac{p(\{x\} \mid \theta) p(\theta)}{p(\{x\})}. \tag{3.72}$$

Thus, we immediately get the *posterior* probability of the parameters given the data. The only unknown quantity is the *prior* $p(\theta)$ since the denominator can be expressed as $p(\{x\}) = \int d\theta \, p(\{x\} \mid \theta) p(\theta)$.

The posterior yields the entire distribution of parameters given the data. Often, the goal is more modest: to determine a *point estimate* of the parameters. (3.72) suggests to simply maximize the posterior: $\hat{\theta} = \operatorname{argmax}_\theta p(\theta \mid \{x\})$. If the $\operatorname{argmax}_\theta$ does not affect the prior, we recover the *maximum likelihood estimate* $\hat{\theta} = \operatorname{argmax}_\theta p(\{x\} \mid \theta)$.

If we want to make inferences about a not yet observed datum x_*, we use the data distribution in combination with the posterior:

$$p(x_* \mid \{x\}) = \int d\theta \, p(x_* \mid \theta) p(\theta \mid \{x\}). \tag{3.73}$$

In cases where the data constrains the parameters well, the posterior is sharply peaked and we can use a Laplace approximation, leading to $p(x_* \mid \{x\}) \approx p(x_* \mid \hat{\theta})$.

So far, we assumed the data distribution to be given. However, we typically don't know the underlying distribution. Accordingly, we need to compare different models $p(\{x\} \mid \theta, \mathcal{M})$, i.e., we need to perform *model comparison*. Again using Bayes rule, we get

$$\frac{p(\mathcal{M}_1 \mid \{x\})}{p(\mathcal{M}_2 \mid \{x\})} = \frac{p(\{x\} \mid \mathcal{M}_1)}{p(\{x\} \mid \mathcal{M}_2)} \frac{p(\mathcal{M}_1)}{p(\mathcal{M}_2)} \tag{3.74}$$

where $p(\{x\} \mid \mathcal{M})$ appeared as an innocuous normalization constant in (3.72) and $p(\mathcal{M})$ is the model prior. Thus, we can compare the relative probabilities of the models given the data and reject one of the models if it is much more unlikely.

3.5.1 *Bayesian Supervised Learning*

Supervised learning can be framed as an inference problem (MacKay 2003). In this framework, the model, e.g., a feedforward network, determines the distribution of outputs y for given inputs x and parameters θ, that is $p(y \mid \theta, x)$. The task is to determine the most likely parameters $\hat{\theta}$ giving rise to the dataset of observed inputs $\{x\}$ and outputs $\{y\}$. From Bayes' theorem, we get

$$p(\theta \mid \{y\}, \{x\}) = \frac{p(\{y\} \mid \theta, \{x\}) p(\theta)}{p(\{y\} \mid \{x\})} \tag{3.75}$$

where we used conditional independence of the prior on the input, $p(\theta \mid \{x\}) = p(\theta)$. Thus, we arrive at an inference problem: the 'trained' parameters are determined by $\hat{\theta} = \operatorname{argmax}_\theta p(\theta \mid \{y\}, \{x\})$.

For example, let us assume that the model provides a deterministic mapping of inputs to outputs $y = f_\theta(x)$ and that the output is corrupted by noise $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)$. In addition, let us assume that the prior is also Gaussian, $\theta \sim \mathcal{N}(0, \sigma_\theta^2)$. Taking the logarithm on both sides of (3.75), we get

$$\hat{\theta} = \text{argmin}_\theta \left[\frac{1}{2\sigma_\epsilon^2} \sum_i (y_i - f_\theta(x_i))^2 + \frac{1}{2\sigma_\theta^2} \theta^2 \right]. \tag{3.76}$$

Thus, we recover a quadratic loss function with an L2 regularizer on the parameters. Clearly, different choices of output noise and parameter priors lead to different loss functions and regularizers, respectively.

Gaussian Processes

Does one always have to train the models and obtain a point estimate $\hat{\theta}$? It turns out that there is a neat way around this (MacKay 2003; Rasmussen and Williams 2006): combining the posterior (3.75) with the prediction formula (3.73) yields

$$p(y_* \mid x_*, \{y\}, \{x\}) = \frac{p(y_*, \{y\} \mid x_*, \{x\})}{p(\{y\} \mid \{x\})}. \tag{3.77}$$

Thus, all we need is to condition the *network prior* $p(\{y\} \mid \{x\}) = \int d\theta \, p(\{y\} \mid \theta, \{x\}) p(\theta)$ on the training data—if we manage to obtain the network prior.

Intuitively, the network prior characterizes the distribution of input-output relations of the model obtained from the prior distribution of the parameters. For example, one could initialize a network with many realizations of the weights according to their prior and measure the input-output function.

An easily tractable example is projection into feature space using fixed basis functions $y_n = \sum_k \theta_k \phi_k(x_n)$ with a Gaussian prior $\theta_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_\theta^2)$. Since y relates linearly to the parameters, the network prior is zero-mean Gaussian with $\langle y_n y_m \rangle = \sigma_\theta^2 \sum_k \phi_k(x_n) \phi_k(x_m)$. This multivariate Gaussian can be seen as a marginalization of the zero-mean GP with correlation function (in this context called *kernel*)

$$C(x, x') = \sigma_\theta^2 \sum_k \phi_k(x) \phi_k(x') \tag{3.78}$$

to the data. In this case, we see explicitly how the likelihood in combination with the parameter prior give rise to a GP network prior and the associated kernel. For more complex models, see Chapter 4 and references therein.

The final step (3.77) is straightforward for a GP prior since the conditional of a Gaussian is known analytically, see (3.16). We get

$$p(y_* \mid x_*, \{y\}, \{x\}) = \mathcal{N}(y_* \mid \mu_*, \sigma_*^2) \tag{3.79}$$

with parameters $\mu_* = \sum_{m,n} C(x_*, x_m) C(x_m, x_n)^{-1} y_n$, $\sigma_*^2 = C(x_*, x_*) - \sum_{m,n} C(x_*, x_m) C(x_m, x_n)^{-1} C(x_n, x_*)$. From a computational perspective, the most expensive operation is usually the inversion of the matrix $C(x_m, x_n)$ which scales cubically in the size of the dataset. From a conceptual perspective, GP inference reduces a prediction problem to simple linear algebra!

Part II

PUBLICATIONS & PREPRINTS

4

UNIFIED FIELD THEORY FOR DEEP AND RECURRENT NEURAL NETWORKS

PREAMBLE

We start the main part of this thesis with a focus on rather simple network models: fully-connected, deep, feedforward networks (DNNs) and vanilla recurrent networks (RNNs). Owing to their simplicity, we can go beyond the dynamics and investigate functional aspects. More concretely, we consider supervised learning from a Bayesian perspective (see Section 3.5).

There are two necessary ingredients for Bayesian supervised learning: 1) the likelihood which encodes the input-output relation for given parameters, and 2) a prior on the parameters. For both RNNs and DNNs, the likelihood is readily available—it corresponds to a degenerate distribution which maps inputs to outputs in a deterministic manner. A natural prior on the parameters is an independent Gaussian, akin to commonly used initialization schemes of such networks (Glorot and Bengio 2010; He et al. 2015).

The key challenge is to compute the network prior, i.e., to marginalize the parameter prior. For finite networks, this problem is, in general, intractable. In the limit of infinite network size, however, the situation changes drastically: the network prior of both DNNs and RNNs becomes a Gaussian process (GP). For single-layer networks this is the result of the seminal work by Neal (1996) and Williams (1998); it was extended to deep networks by Lee et al. (2018) and Matthews et al. (2018). For RNNs, this is a recent result by Yang (2019) because weight sharing impedes a straightforward application of the central limit theorem. In both cases, the final step—conditioning on the training data—can be performed analytically owing to the GP limit.

In principle, knowing the infinite-size limit paves the road to a perturbative computation of finite-size corrections (e.g., Naveh et al. 2021; Grosvenor and Jefferson 2022; Roberts, Yaida, and Hanin 2022). However, the above results rely on different techniques such that performing the perturbative expansion is not straightforward. Hence, in this chapter, we unify the calculation of the network prior using the language of field theory—dynamic mean-field theory, to be precise (see Section 3.4)—which comes with a rich toolbox for finite-size corrections (e.g., Zinn-Justin 1996; Moshe and Zinn-Justin 2003). Furthermore, we systematically compare the resulting GPs for DNNs and RNNs which has, to the best of our knowledge, not yet been done.

Author Contributions

This preprint contains material from the bachelor's thesis of Bastian Epping (BE) and the master's thesis of Kai Segadlo (KS) which were both directly supervised by the author (AvM) in collaboration with Dr. David Dahmen (DD), Prof. Michael Krämer (MK), and Prof. Moritz Helias (MH). DD directly supervised the last months of KS thesis due to the paternal leave of AvM.
All calculations were performed by BE, KS and AvM. The codebase was written jointly by BE, KS, and AvM. The numerical experiments were performed, and the corresponding figures were created, by BE and KS. The first draft of the manuscript was written jointly by BE, KS, AvM, and MH. The manuscript was revised by all co-authors. BE, KS, and AvM contributed equally to this manuscript.

Unified Field Theory for Deep and Recurrent Neural Networks

Kai Segadlo[1,2,*], Bastian Epping[2,3,*], Alexander van Meegen[2,4,*], David Dahmen[2], Michael Krämer[3] and Moritz Helias[1,2]

[1] Department of Physics, Faculty 1, RWTH Aachen University, Aachen, Germany
[2] Institute of Neuroscience and Medicine (INM-6) and Institute for Advanced Simulation (IAS-6) and JARA-Institute Brain Structure-Function Relationships (INM-10), Jülich Research Centre, Jülich, Germany
[3] Institute for Theoretical Particle Physics and Cosmology, RWTH Aachen University, Aachen, Germany
[4] Institute of Zoology, University of Cologne, Cologne, Germany
[*] Equal contribution

Abstract. Understanding capabilities and limitations of different network architectures is of fundamental importance to machine learning. Bayesian inference on Gaussian processes has proven to be a viable approach for studying recurrent and deep networks in the limit of infinite layer width, $n \to \infty$. Here we present a unified and systematic derivation of the mean-field theory for both architectures that starts from first principles by employing established methods from statistical physics of disordered systems. The theory elucidates that while the mean-field equations are different with regard to their temporal structure, they yet yield identical Gaussian kernels when readouts are taken at a single time point or layer, respectively. Bayesian inference applied to classification then predicts identical performance and capabilities for the two architectures. Numerically, we find that convergence towards the mean-field theory is typically slower for recurrent networks than for deep networks and the convergence speed depends non-trivially on the parameters of the weight prior as well as the depth or number of time steps, respectively. Our method exposes that Gaussian processes are but the lowest order of a systematic expansion in $1/n$. The formalism thus paves the way to investigate the fundamental differences between recurrent and deep architectures at finite widths n.

Contents

1. Introduction

Deep learning has brought a dramatic improvement of the state-of-the-art in many fields of data science, ranging from speech recognition and translation to visual object classification [1–4]. Any progress in the empirically-driven improvement of algorithms must be accompanied by a profound understanding of why and how deep learning works. Such an understanding is needed to provide guarantees, for example about the accuracy and the robustness of the networks, and will help preventing the frequently reported failures of deep learning, such as its vulnerability to adversarial examples [5].

A common method to obtain analytical insight into deep networks is to study the overparametrized limit in which the width n_ℓ of all layers ℓ tends to infinity. In this limit, it has been shown with mean-field theory that under a Gaussian prior on the weights $W^{(\ell)}$ in each layer, the pre-activations follow a Gaussian process with an iteratively determined covariance [6–9]; in particular, the pre-activations across different layers and across different neurons become independently Gaussian distributed. This approach allows one to investigate learning and prediction in the framework of Bayesian inference [7].

Often, analogies are drawn between deep neural networks (DNNs) and discrete-time recurrent neural networks (RNNs): Unrolling time in RNNs formally converts them to DNNs, however with shared weights $W^{(\ell)} \equiv W \; \forall \, \ell$ across layers of identical size $n_\ell \equiv n \; \forall \, \ell$. This led to parallel developments in terms of training strategies for

both architectures, such as backpropagation [10] and backpropagation through time [11].

There are, however, a number of open issues when applying mean-field theory to deep and recurrent neural networks. First of all, the approximation as a Gaussian process relies on the central limit theorem and is thus strictly valid only in the limit of infinite layer widths $n_\ell \to \infty$. Moreover, due to weight sharing, pre-activations for different points in time are not statistically independent in RNNs; the central limit theorem is thus not applicable and the mean-field approximation becomes uncontrolled. Several studies still find that the mean-field theories of DNNs and RNNs appear to be closely related, culminating in ref. [12] which formulates a variety of network architectures as tensor programs and finds that most common network architectures, under certain conditions on the non-linearities and piors, converge in distribution to a Gaussian process. But the relationship between the Gaussian processes for RNNs and DNNs has so far not been addressed.

Currently, there is no systematic approach that would allow one to simultaneously study DNNs and RNNs and that would be extendable to finite layer width $n_\ell, n < \infty$; the agreement of the mean-field predictions with the performance of finite-size networks is based on numerical evidence so far. Furthermore, in the limit of infinite width the number of trainable parameters of a DNN, $\sum_{\ell=1}^{L} n_{\ell+1} n_\ell \to \infty$, and of an RNN, $n^2 \to \infty$, both tend to infinity and do not enter explicitly in the result of the Gaussian approximation. The Gaussian process thus has limited capability of quantifying the expressivity of neural networks in relation to the required resources, such as the number of trained weights. Studies on finite-size corrections beyond the $n_\ell \to \infty$ limit are so far restricted to DNNs [13–18]. Understanding the limits of the putative equivalence of DNNs and RNNs on the mean-field level requires a common theoretical basis for the two architectures that would extend to finite n and finite n_ℓ.

To overcome these limitations, we here combine the established view of Bayesian inference by Gaussian processes [19] with the emerging methods of statistical field theory applied to neural networks [20–25]. The latter methods have been developed in the field of disordered systems, which are systems with random parameters, such as spin glasses [26–28]. These methods are able to extract the typical behavior of a system with a large number of interacting components. For example, this approach has recently been used to characterize the typical richness, represented by the entropy, of Boolean functions computed in the output layer of DNNs, RNNs, and sparse Boolean circuits [29].

Concretely, in this paper, we present a systematic derivation of the mean-field theories for DNNs and RNNs that is based on the well-established approach of field theory for recurrent networks [20, 24, 30], which allows a unified treatment of the two architectures [29]. This paves the way for extensions to finite n, n_ℓ, enabled by a rich set of systematic methods available in the mathematical physics literature to compute corrections beyond the leading order [31, 32]. Already to leading order, we find that the mean-field theories of DNNs and RNNs are in fact qualitatively different with regard to correlations across layers or time, respectively. The predictive distribution in Bayesian training is therefore in general different between the two architectures. Nonetheless, the structure of the mean-field equations can give rise to the same Gaussian processes kernel in the limit of infinite width for both DNNs and RNNs if the readout in the RNN is taken from a single time step. This finding holds for single inputs, as pointed out in ref. [29], as well as input sequences. Furthermore, for a point-symmetric activation function [29], there is no observable difference between DNNs and RNNs on the mean-

field level if the biases are uncorrelated in time and the input is only supplied in the first time step.

2. Theoretical background

2.1. Bayesian supervised learning

First, we briefly review the Bayesian approach to supervised learning [33]. Let $p(\boldsymbol{y}\,|\,\boldsymbol{x},\boldsymbol{\theta})$ be a model (here DNN or RNN) that maps inputs $\mathbf{x} \in \mathbb{R}^{n_{\text{in}}}$ to outputs $\boldsymbol{y} \in \mathbb{R}^{n_{\text{out}}}$ and that depends on a set of parameters $\boldsymbol{\theta}$. Conventional training of such a model corresponds to finding a particular parameter set $\hat{\boldsymbol{\theta}}$ that maximizes the likelihood $p(\boldsymbol{Y}\,|\,\boldsymbol{X},\boldsymbol{\theta})$ for some given training data $\boldsymbol{D} = \{\boldsymbol{X},\boldsymbol{Y}\}$, with $\boldsymbol{X} \in \mathbb{R}^{n_D \times n_{\text{in}}}$ and $\boldsymbol{Y} \in \mathbb{R}^{n_D \times n_{\text{out}}}$. A prediction for the output $\boldsymbol{y}^*$ caused by an unseen test input $\boldsymbol{x}^*$ is then given by $p(\boldsymbol{y}^*\,|\,\boldsymbol{x}^*,\hat{\boldsymbol{\theta}})$. In the Bayesian view, one instead assumes a prior distribution of parameters $p(\boldsymbol{\theta})$ to obtain, via Bayes' rule, an entire posterior distribution of the parameters

$$p(\boldsymbol{\theta}\,|\,\boldsymbol{Y},\boldsymbol{X}) = \frac{p(\boldsymbol{Y}\,|\,\boldsymbol{X},\boldsymbol{\theta})\,p(\boldsymbol{\theta})}{\int d\boldsymbol{\theta}\,p(\boldsymbol{Y}\,|\,\boldsymbol{X},\boldsymbol{\theta})\,p(\boldsymbol{\theta})}\,. \tag{1}$$

The conditioning on the training data in $p(\boldsymbol{\theta}\,|\,\boldsymbol{Y},\boldsymbol{X})$ can be interpreted as selecting, among all possible parameter sets given by the prior $p(\boldsymbol{\theta})$, those parameter sets that accomplish the mapping $\boldsymbol{X} \to \boldsymbol{Y}$. A Bayesian prediction for some unseen test input $\boldsymbol{x}^*$ correspondingly results from marginalizing the likelihood over the posterior distribution of the parameters

$$p(\boldsymbol{y}^*\,|\,\boldsymbol{x}^*,\boldsymbol{Y},\boldsymbol{X}) = \int d\boldsymbol{\theta}\,p(\boldsymbol{y}^*\,|\,\boldsymbol{x}^*,\boldsymbol{\theta})\,p(\boldsymbol{\theta}\,|\,\boldsymbol{Y},\boldsymbol{X})\,. \tag{2}$$

Inserting Eq. (1) yields the predictive distribution

$$p(\boldsymbol{y}^*\,|\,\boldsymbol{x}^*,\boldsymbol{Y},\boldsymbol{X}) = \frac{p(\boldsymbol{y}^*,\boldsymbol{Y}\,|\,\boldsymbol{x}^*,\boldsymbol{X})}{p(\boldsymbol{Y}\,|\,\boldsymbol{X})} \tag{3}$$

that depends on the model-dependent *network priors*

$$p(\boldsymbol{Y}\,|\,\boldsymbol{X}) = \int d\boldsymbol{\theta}\,p(\boldsymbol{Y}\,|\,\boldsymbol{X},\boldsymbol{\theta})\,p(\boldsymbol{\theta}), \tag{4}$$

$$p(\boldsymbol{y}^*,\boldsymbol{Y}\,|\,\boldsymbol{x}^*,\boldsymbol{X}) = \int d\boldsymbol{\theta}\,p(\boldsymbol{y}^*\,|\,\boldsymbol{x}^*,\boldsymbol{\theta})\,p(\boldsymbol{Y}\,|\,\boldsymbol{X},\boldsymbol{\theta})\,p(\boldsymbol{\theta})\,. \tag{5}$$

The network priors encompass all input-output relationships which are compatible with the prior $p(\boldsymbol{\theta})$ and the model. The difference between the two network priors, Eq. (4) and Eq. (5), is the information on the additional test input $\boldsymbol{x}^*$ and output $\boldsymbol{y}^*$.

Note that the Bayesian approach to supervised learning can also be used for input sequences $\{\mathbf{x}^{(0)},\ldots,\mathbf{x}^{(A)}\}$ with $\mathbf{x}^{(a)} \in \mathbb{R}^{n_{\text{in}}}$. To this end, it is sufficient to replace $\boldsymbol{x} \to \{\mathbf{x}^{(0)},\ldots,\mathbf{x}^{(A)}\}$ and $\boldsymbol{X} \to \{\mathbf{X}^{(0)},\ldots,\mathbf{X}^{(A)}\}$ in the above formulas.

In the following, we use a field theoretic approach to calculate the network priors for both deep and recurrent neural networks. Conditioning on the training data, Eq. (3), then yields the Bayesian prediction of the output.

2.2. Network architectures

Deep feedforward neural networks (DNNs) and discrete-time recurrent neural networks (RNNs) can both be described by a set of pre-activations $\boldsymbol{h}^{(a)} \in \mathbb{R}^{n_a}$ that are

determined by an affine linear transformation

$$\boldsymbol{h}^{(a)} = \boldsymbol{W}^{(a)}\phi(\boldsymbol{h}^{(a-1)}) + \boldsymbol{W}^{(\mathrm{in},a)}\boldsymbol{x}^{(a)} + \boldsymbol{\xi}^{(a)} \tag{6}$$

of activations $\phi(\boldsymbol{h}^{(a-1)}) \in \mathbb{R}^{n_{a-1}}$. The pre-activations are transformed by an activation function $\phi : \mathbb{R} \to \mathbb{R}$ which is applied element-wise to vectors. For DNNs, $\boldsymbol{W}^{(a)} \in \mathbb{R}^{n_a \times n_{a-1}}$ denotes the weight matrix from layer $a-1$ to layer a, and $\boldsymbol{\xi}^{(a)} \in \mathbb{R}^{n_a}$ represents biases in layer a. Inputs $\boldsymbol{x}^{(a)}$ are typically only applied to the first layer such that the input matrices $\boldsymbol{W}^{(\mathrm{in},a)} \in \mathbb{R}^{n_a \times n_{\mathrm{in}}}$ vanish for $a > 0$. For RNNs, the index a denotes different time steps. The weight matrix, input matrix, and biases are static over time, $\boldsymbol{W}^{(a)} \equiv \boldsymbol{W}$, $\boldsymbol{W}^{(\mathrm{in},a)} \equiv \boldsymbol{W}^{(\mathrm{in})}$, and $\boldsymbol{\xi}^{(a)} \equiv \boldsymbol{\xi}$, and couple activities across successive time steps. For both architectures, we include an additional input and output layer

$$\boldsymbol{h}^{(0)} = \boldsymbol{W}^{(\mathrm{in},0)}\boldsymbol{x}^{(0)} + \boldsymbol{\xi}^{(0)}, \tag{7}$$

$$\boldsymbol{y} \quad = \boldsymbol{W}^{(\mathrm{out})}\phi(\boldsymbol{h}^{(A)}) + \boldsymbol{\xi}^{(A+1)}, \tag{8}$$

with $\boldsymbol{W}^{(\mathrm{out})} \in \mathbb{R}^{n_{\mathrm{out}} \times n_A}$, which allow us to set independent input and output dimensions. Here, A denotes the final layer for the DNN and the final time point for the RNN. The set of trainable parameters $\boldsymbol{\theta}$ is the collection of $\boldsymbol{W}^{(\mathrm{in},a)}, \boldsymbol{W}^{(\mathrm{out})}, \boldsymbol{W}^{(a)}$, and $\boldsymbol{\xi}^{(a)}$.

2.3. Parameter priors

We use Gaussian priors for all model parameters, that is for the RNN

$$W_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, n^{-1}g^2), \tag{9}$$

$$W_{ij}^{(\mathrm{in})} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, n_{\mathrm{in}}^{-1}g_0^2), \tag{10}$$

$$\xi_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2), \tag{11}$$

and for the DNN

$$W_{ij}^{(a)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, n_{a-1}^{-1}g_a^2), \tag{12}$$

$$W_{ij}^{(\mathrm{in},a)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, n_{\mathrm{in}}^{-1}g_0^2), \tag{13}$$

$$\xi_i^{(a)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_a^2), \tag{14}$$

as well as

$$W_{ij}^{(\mathrm{out})} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, n_{\mathrm{A}}^{-1}g_{A+1}^2), \tag{15}$$

$$\xi_i^{(A+1)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_{A+1}^2), \tag{16}$$

for both architectures (where $n_A = n$ for the RNN). These priors on the parameters are used to calculate the network prior $p(\boldsymbol{Y} \,|\, \boldsymbol{X})$.

3. Unified field theory for RNNs and DNNs

The network prior $p(\boldsymbol{Y} \,|\, \boldsymbol{X})$, Eq. (4), is a joint distribution of all outputs $\boldsymbol{y}_\alpha \in \boldsymbol{Y}$, each corresponding to a single training input $\boldsymbol{x}_\alpha \in \boldsymbol{X}$. Its calculation is tantamount to a known problem in physics, the replica calculation [31, 34]. Here, each replicon is a copy of the network with the same parameters $\boldsymbol{\theta}$ but a different input $\boldsymbol{x}_\alpha$. For simplicity, in the following we illustrate the derivation of $p(\boldsymbol{y} \,|\, \boldsymbol{x})$ for a single input $\boldsymbol{x} \equiv \boldsymbol{x}^{(a=0)}$ that is presented to the first layer of the DNN or at the first time point

for the RNN, respectively. We present the more cumbersome but conceptually similar general case of multiple inputs, or multiple input sequences, in Appendix 6.1.

The network prior is defined as the probability of the output given the input,

$$p(\boldsymbol{y}\,|\,\boldsymbol{x}) = \int d\boldsymbol{\theta}\, p(\boldsymbol{y}\,|\,\boldsymbol{x},\boldsymbol{\theta})\, p(\boldsymbol{\theta}), \tag{17}$$

marginalized over the parameter prior, where

$$\begin{aligned} p(\boldsymbol{y}\,|\,\boldsymbol{x},\boldsymbol{\theta}) = \int d\boldsymbol{h}^{(0)} \ldots \int d\boldsymbol{h}^{(A)}\, \delta\left(\boldsymbol{y} - \boldsymbol{W}^{(\mathrm{out})}\boldsymbol{\phi}^{(A)} - \boldsymbol{\xi}^{(A+1)}\right) \\ \times \prod_{a=1}^{A} \delta\left(\boldsymbol{h}^{(a)} - \boldsymbol{W}^{(a)}\boldsymbol{\phi}^{(a-1)} - \boldsymbol{\xi}^{(a)}\right) \\ \times\ \delta\left(\boldsymbol{h}^{(0)} - \boldsymbol{W}^{(\mathrm{in})}\boldsymbol{x} - \boldsymbol{\xi}^{(0)}\right), \end{aligned} \tag{18}$$

follows by enforcing the set of equations, Eq. (6) to Eq. (8), using Dirac constraints. Throughout the manuscript, we use the abbreviation $\boldsymbol{\phi}^{(a)} = \phi(\boldsymbol{h}^{(a)})$.

3.1. Marginalization of the parameter prior

From Eq. (18), it follows that the computation of the marginalization of the parameters $\boldsymbol{\theta}$ in Eq. (17) can be reduced to

$$\begin{aligned} p(\boldsymbol{y}\,|\,\boldsymbol{x}) = \int d\boldsymbol{h}^{(0)} \ldots \int d\boldsymbol{h}^{(A)} \\ \times \left\langle \delta\left(\boldsymbol{y} - \boldsymbol{W}^{(\mathrm{out})}\boldsymbol{\phi}^{(A)} - \boldsymbol{\xi}^{(A+1)}\right)\right\rangle_{\boldsymbol{W}^{(\mathrm{out})},\boldsymbol{\xi}^{(A+1)}} \\ \times \left\langle \left\langle \prod_{a=1}^{A} \delta\left(\boldsymbol{h}^{(a)} - \boldsymbol{W}^{(a)}\boldsymbol{\phi}^{(a-1)} - \boldsymbol{\xi}^{(a)}\right)\right\rangle_{\{\boldsymbol{W}^{(a)}\}} \right. \\ \times \left. \left\langle \delta\left(\boldsymbol{h}^{(0)} - \boldsymbol{W}^{(\mathrm{in})}\boldsymbol{x} - \boldsymbol{\xi}^{(0)}\right)\right\rangle_{\boldsymbol{W}^{(\mathrm{in})}} \right\rangle_{\{\boldsymbol{\xi}^{(a)}\}}. \end{aligned} \tag{19}$$

To proceed, it is advantageous to represent the Dirac δ-distributions as Fourier integrals,

$$\delta(\boldsymbol{h}) = \int d\tilde{\boldsymbol{h}}\, \exp(\tilde{\boldsymbol{h}}^{\mathrm{T}}\boldsymbol{h}) \tag{20}$$

with the inner product $\tilde{\boldsymbol{h}}^{\mathrm{T}}\boldsymbol{h} = \sum_k \tilde{h}_k h_k$ and $\int d\tilde{\boldsymbol{h}} = \prod_k \int_{i\mathbb{R}} \frac{d\tilde{h}_k}{2\pi i}$, because it leads to averages of the form $\langle \exp(k\theta)\rangle_\theta$ which are analytically solvable. Using $\langle \exp(k\theta)\rangle_\theta = \exp\left(\frac{1}{2}\sigma^2 k^2\right)$ for $\theta \sim \mathcal{N}(0,\sigma^2)$, the network prior for a single replicon, $p(\boldsymbol{y}\,|\,\boldsymbol{x})$, takes the form (details in Appendix 6.1)

$$p(\boldsymbol{y}\,|\,\boldsymbol{x}) = \int d\tilde{\boldsymbol{y}} \int \mathcal{D}\boldsymbol{h} \int \mathcal{D}\tilde{\boldsymbol{h}}\, \exp\left(\mathcal{S}(\boldsymbol{y},\tilde{\boldsymbol{y}},\boldsymbol{h},\tilde{\boldsymbol{h}}\,|\,\boldsymbol{x})\right), \tag{21}$$

where $\int \mathcal{D}\boldsymbol{h} \equiv \prod_{a=0}^{A}\prod_k \int_{\mathbb{R}} dh_k^{(a)}$ and $\int \mathcal{D}\tilde{\boldsymbol{h}} \equiv \prod_{a=0}^{A}\prod_k \int_{i\mathbb{R}} \frac{d\tilde{h}_k^{(a)}}{2\pi i}$. The exponent $\mathcal{S}$, commonly called *action,* is given by

$$\mathcal{S}(\boldsymbol{y},\tilde{\boldsymbol{y}},\boldsymbol{h},\tilde{\boldsymbol{h}}\,|\,\boldsymbol{x}) = \mathcal{S}_{\mathrm{out}}(\boldsymbol{y},\tilde{\boldsymbol{y}}\,|\,\boldsymbol{h}^{(A)}) + \mathcal{S}_{\mathrm{net}}(\boldsymbol{h},\tilde{\boldsymbol{h}}\,|\,\boldsymbol{x}), \tag{22}$$

where

$$\mathcal{S}_{\mathrm{net}}(\boldsymbol{h},\tilde{\boldsymbol{h}}\,|\,\boldsymbol{x}) := \sum_{a=0}^{A} \tilde{\boldsymbol{h}}^{(a)\mathrm{T}}\boldsymbol{h}^{(a)} + \frac{1}{2}\sum_{a,b=0}^{A} \sigma_a^2\, \tilde{\boldsymbol{h}}^{(a)\mathrm{T}} M_{a,b} \tilde{\boldsymbol{h}}^{(b)}$$

$$+\frac{1}{2}\sum_{a,b=1}^{A}\frac{g_a^2}{n_{a-1}}\tilde{\boldsymbol{h}}^{(a)\mathrm{T}}\boldsymbol{\phi}^{(a-1)\mathrm{T}}M_{a,b}\boldsymbol{\phi}^{(b-1)}\tilde{\boldsymbol{h}}^{(b)}$$
$$+\frac{g_0^2}{2n_{\mathrm{in}}}\tilde{\boldsymbol{h}}^{(0)\mathrm{T}}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}\tilde{\boldsymbol{h}}^{(0)} \tag{23}$$

is the action of the input and the recurrent layer of the RNN or the inner part of the DNN, respectively, and

$$\mathcal{S}_{\mathrm{out}}(\boldsymbol{y},\tilde{\boldsymbol{y}}\,|\,\boldsymbol{h}^{(A)}) := \tilde{\boldsymbol{y}}^{\mathrm{T}}\boldsymbol{y}+\frac{\sigma_{A+1}^2}{2}\tilde{\boldsymbol{y}}^{\mathrm{T}}\tilde{\boldsymbol{y}}+\frac{g_{A+1}^2}{2n_A}\tilde{\boldsymbol{y}}^{\mathrm{T}}\boldsymbol{\phi}^{(A)\mathrm{T}}\boldsymbol{\phi}^{(A)}\tilde{\boldsymbol{y}} \tag{24}$$

is the action for the output layer. Note that $\mathcal{S}$ is diagonal in neuron indices with respect to the explicitly appearing fields $\boldsymbol{h}$ and $\tilde{\boldsymbol{h}}$ and couplings across neurons are only mediated by terms of the form $\propto \boldsymbol{\phi}^{\mathrm{T}}\boldsymbol{\phi}$.

For RNNs, the shared connectivity and biases at different time points imply correlations across time steps; for DNNs, in contrast, the connectivity and biases are realized independently across layers, so that the action decomposes into a sum of $A+2$ individual layers. In Eq. (23), this leads to

$$M_{a,b} = \begin{cases} 1 & \text{RNN} \\ \delta_{a,b} & \text{DNN} \end{cases} \tag{25}$$

which is the only difference between DNN and RNN in this formalism.

3.2. Auxiliary variables

An action that is quadratic in $\boldsymbol{h}$ and $\tilde{\boldsymbol{h}}$ corresponds to a Gaussian and therefore to an analytically solvable integral. However, the post-activations $\boldsymbol{\phi} \equiv \phi(\boldsymbol{h})$ in $\mathcal{S}_{\mathrm{net}}$ and $\mathcal{S}_{\mathrm{out}}$ introduce a non-quadratic part and the terms $\propto \tilde{\boldsymbol{h}}^{\mathrm{T}}\tilde{\boldsymbol{h}}\,\boldsymbol{\phi}^{\mathrm{T}}\boldsymbol{\phi}$ cause a coupling across neurons. To deal with this difficulty, we introduce new auxiliary variables

$$C^{(a,b)} := M_{a,b}\left[\sigma_a^2+\mathbb{1}_{a\geq 1,b\geq 1}\frac{g_a^2}{n_{a-1}}\boldsymbol{\phi}^{(a-1)\mathrm{T}}\boldsymbol{\phi}^{(b-1)}\right]$$
$$+\,\mathbb{1}_{a=0,b=0}\frac{g_0^2}{n_{\mathrm{in}}}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}, \tag{26}$$

where $0 \leq a,b \leq A+1$, a common practice originating from dynamic spin-glass theory [35] and used for random networks [23–25, 36]. The second term $\propto \boldsymbol{\phi}^{\mathrm{T}}\boldsymbol{\phi}$ in Eq. (26) contains the sum of post-activations over all neuron indices. Assuming sufficiently weak correlations among the ϕ_i, we expect the sum to be close to its mean value with decreasing variations as n_a grows; for large n_a the sum is thus close to a Gaussian. This intuition is made precise below by a formal saddle point approximation in C.

Enforcing the auxiliary variables through Dirac-δ constraints, analogous to Eq. (20) (see Appendix 6.1 for details), leads to the prior distribution

$$p(\boldsymbol{y}\,|\,\boldsymbol{x}) = \int d\tilde{\boldsymbol{y}}\,\exp\left(\tilde{\boldsymbol{y}}^{\mathrm{T}}\boldsymbol{y}\right)\left\langle \exp\left(\frac{1}{2}\tilde{\boldsymbol{y}}^{\mathrm{T}}C^{(A+1,A+1)}\tilde{\boldsymbol{y}}\right)\right\rangle_{C,\tilde{C}}, \tag{27}$$

where the distribution $C,\tilde{C} \sim \exp\left(\mathcal{S}_{\mathrm{aux}}(C,\tilde{C})\right)$ is described by the action

$$\mathcal{S}_{\text{aux}}(C,\tilde{C}) := -n\sum_{a,b=0}^{A+1}\nu_{a-1}\,\tilde{C}^{(a,b)}C^{(a,b)} + n\,\mathcal{W}_{\text{aux}}(\tilde{C}\,|\,C). \tag{28}$$

Here, $\nu_a = n_a/n$ with $n_{-1} \equiv n_{\text{in}}$ is the size of a layer in the DNN relative to the size of the RNN. The recurrent part and the input, which decouple in the neurons, are together described for the DNN by

$$\begin{aligned}\mathcal{W}_{\text{aux}}^{\text{DNN}}(\tilde{C}\,|\,C) = &\sum_{a=1}^{A+1}\nu_{a-1}\ln\left\langle e^{\tilde{C}^{(a,a)}\,g_a^2\phi^{(a-1)}\phi^{(a-1)}}\right\rangle_{h^{(a-1)}}\\ &+\nu_{-1}\,\tilde{C}^{(0,0)}\frac{g_0^2}{n_{\text{in}}}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x} + \sum_{a=0}^{A+1}\nu_{a-1}\,\tilde{C}^{(a,a)}\sigma_a^2\end{aligned} \tag{29}$$

with $h^{(a)} \sim \mathcal{N}(0, C^{(a,a)})$ a centered Gaussian with layer-dependent variance $\langle h^{(a)}h^{(a)}\rangle = C^{(a,a)}$ and for the RNN by

$$\begin{aligned}\mathcal{W}_{\text{aux}}^{\text{RNN}}(\tilde{C}\,|\,C) = &\ln\left\langle e^{\sum_{a,b=1}^{A+1}\tilde{C}^{(a,b)}\,g^2\phi^{(a-1)}\phi^{(b-1)}}\right\rangle_{\{h^{(a)}\}}\\ &+\nu_{-1}\,\tilde{C}^{(0,0)}\frac{g_0^2}{n_{\text{in}}}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x} + \sum_{a,b=0}^{A+1}\tilde{C}^{(a,b)}\sigma^2\end{aligned} \tag{30}$$

with $\{h^{(a)}\} \equiv \{h^{(a)}\}_{0\le a\le A}$ and $\{h^{(a)}\}_{0\le a\le A} \sim \mathcal{N}(0, C)$ a centered Gaussian across time with covariance matrix $\langle h^{(a)}h^{(b)}\rangle = C^{(a,b)}$.

3.3. Saddle-point approximation

The factor n in Eq. (28), which stems from the decoupling across neurons, for large n leads to a strongly peaked distribution of C and $\tilde{C}$. Therefore we can use a saddle point approximation to calculate the average over C and $\tilde{C}$ in Eq. (27). In the limit $n \to \infty$ this approximation becomes exact.

We thus search for stationary points of the action, $\frac{\partial}{\partial C^{(a,b)}}\mathcal{S}_{\text{aux}}(C,\tilde{C}) \overset{!}{=} 0$ and $\frac{\partial}{\partial \tilde{C}^{(a,b)}}\mathcal{S}_{\text{aux}}(C,\tilde{C}) \overset{!}{=} 0$, which yields a coupled set of self-consistency equations for the mean values $\overline{C}$ and $\overline{\tilde{C}}$, commonly called mean-field equations: $\overline{\tilde{C}}^{(a,b)} \equiv 0$, which follows from the normalization of the probability distribution [37], and

$$\begin{aligned}\overline{C}^{(a,b)} = &M_{a,b}\left[\sigma_a^2 + \mathbb{1}_{a\ge 1,b\ge 1}\,g_a^2\langle\phi(h^{(a-1)})\phi(h^{(b-1)})\rangle_{h^{(a-1)},h^{(b-1)}}\right]\\ &+\mathbb{1}_{a=0,b=0}\,\frac{g_0^2}{n_{\text{in}}}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}\end{aligned} \tag{31}$$

with $h^{(a-1)}, h^{(b-1)} \sim \mathcal{N}(0, \overline{C})$. Eq. (31) comprises both DNN and RNN; the difference between Eq. (29) and Eq. (30) leads to the appearance of $M_{a,b}$ on the r.h.s. The average on the r.h.s. has to be taken with respect to a theory that only includes two layers or time points. This is due to the marginalization property of the Gaussian distribution of the pre-activations $h^{(a-1)}$, which results from inserting the saddle-point solutions Eq. (31) for $\overline{C}$ and $\overline{\tilde{C}}$. Accordingly, we are left with a closed system

of equations for the saddle-point values $\overline{C}$ that are the layer- or time-dependent correlations. These equations need to be solved recursively from the input $\overline{C}^{(0,0)} = \sigma_0^2 + \frac{g_0^2}{n_{\text{in}}} \boldsymbol{x}^{\text{T}} \boldsymbol{x}$ to the output $\overline{C}^{(A+1,A+1)} = \sigma_{A+1}^2 + g_{A+1}^2 \langle \phi(h^{(A)}) \phi(h^{(A)}) \rangle_{h^{(A)}, h^{(A)}}$.

3.4. Network prior

Computing the Gaussian integral over $\tilde{\boldsymbol{y}}$ in the saddle-point approximation of Eq. (27), one obtains the distribution of the outputs as independent Gaussians across neurons i

$$p(\boldsymbol{y} \,|\, \boldsymbol{x}) = \prod_i p(y_i \,|\, \boldsymbol{x}) = \prod_i \mathcal{N}(y_i \,; 0, \overline{C}^{(A+1,A+1)}) \,. \tag{32}$$

An analogous calculation for multiple input sequences $\{\boldsymbol{x}_\alpha^{(0)}, \ldots, \boldsymbol{x}_\alpha^{(A)}\}$ (see Appendix 6.1) yields the equivalent mean-field equations

$$\begin{aligned} \overline{C}_{\alpha\beta}^{(a,b)} = M_{a,b} \Big[& \sigma_a^2 + \mathbb{1}_{a \geq 1, b \geq 1} \, g_a^2 \langle \phi(h_\alpha^{(a-1)}) \phi(h_\beta^{(b-1)}) \rangle_{h_\alpha^{(a-1)}, h_\beta^{(b-1)}} \\ & + \mathbb{1}_{a \leq A, b \leq A} \, \frac{g_0^2}{n_{\text{in}}} \boldsymbol{x}_\alpha^{(a)\text{T}} \boldsymbol{x}_\beta^{(b)} \Big] \end{aligned} \tag{33}$$

with $h_\alpha^{(a-1)}, h_\beta^{(b-1)} \sim \mathcal{N}(0, \overline{C})$ and $0 \leq a, b \leq A+1$. These lead to the joint network prior

$$\begin{aligned} p(\boldsymbol{Y} \,|\, \{\boldsymbol{X}^{(0)}, \ldots, \boldsymbol{X}^{(A)}\}) &= \prod_i p(\boldsymbol{y}_i \,|\, \{\boldsymbol{X}^{(0)}, \ldots, \boldsymbol{X}^{(A)}\}) \\ &= \prod_i \mathcal{N}(\boldsymbol{y}_i \,; 0, \boldsymbol{K}) \end{aligned} \tag{34}$$

where the covariance matrix is the Gram matrix of the kernel [19],

$$K_{\alpha\beta} = \overline{C}_{\alpha,\beta}^{(A+1,A+1)} \,. \tag{35}$$

Here $\boldsymbol{y}_i$ denotes the i-th row of the output matrix $\boldsymbol{Y}$ that comprises the output of neuron i to all input sequences $\{\boldsymbol{x}_\alpha^{(0)}, \ldots, \boldsymbol{x}_\alpha^{(A)}\}$.

In principle, it is also possible to use independent biases or input weights across time steps in the RNN. This would lead to the respective replacements $M_{a,b}\sigma^2 \to \delta_{a,b}\sigma^2$ and $M_{a,b}\mathbb{1}_{a\leq A, b\leq A} \frac{g_0^2}{n_{\text{in}}} \boldsymbol{x}_\alpha^{(a)\text{T}} \boldsymbol{x}_\beta^{(b)} \to \delta_{a,b}\mathbb{1}_{a\leq A, b\leq A} \frac{g_0^2}{n_{\text{in}}} \boldsymbol{x}_\alpha^{(a)\text{T}} \boldsymbol{x}_\beta^{(b)}$ in Eq. (33).

3.5. Predictive distribution

We split $\boldsymbol{X}, \boldsymbol{Y}$ into training data (indexed by subscript D) and test data (indexed by subscript $*$). The conditioning on the training data via Eq. (3) can here be done analytically because the network priors are Gaussian [19]. For scalar inputs, this yields the predictive distribution

$$p(\boldsymbol{Y}_* \,|\, \boldsymbol{X}_*, \boldsymbol{Y}_D, \boldsymbol{X}_D) = \prod_i \mathcal{N}(\boldsymbol{y}_{*i} \,; \boldsymbol{\mu}_{GP}, \boldsymbol{K}_{GP}) \tag{36}$$

with

$$\boldsymbol{\mu}_{GP} = \boldsymbol{K}_{*D} \boldsymbol{K}_{DD}^{-1} \boldsymbol{y}_D, \qquad \boldsymbol{K}_{GP} = \boldsymbol{K}_{**} - \boldsymbol{K}_{*D} \boldsymbol{K}_{DD}^{-1} \boldsymbol{K}_{*D}^T, \tag{37}$$

which are fully determined by the kernel matrix $\boldsymbol{K} = \begin{pmatrix} \boldsymbol{K}_{DD} & \boldsymbol{K}_{*D}^T \\ \boldsymbol{K}_{*D} & \boldsymbol{K}_{DD} \end{pmatrix}$. For input sequences, it is again sufficient to replace $\boldsymbol{X}_* \to \{\boldsymbol{X}_*^{(0)}, \ldots, \boldsymbol{X}_*^{(A)}\}$ and $\boldsymbol{X}_D \to \{\boldsymbol{X}_D^{(0)}, \ldots, \boldsymbol{X}_D^{(A)}\}$.

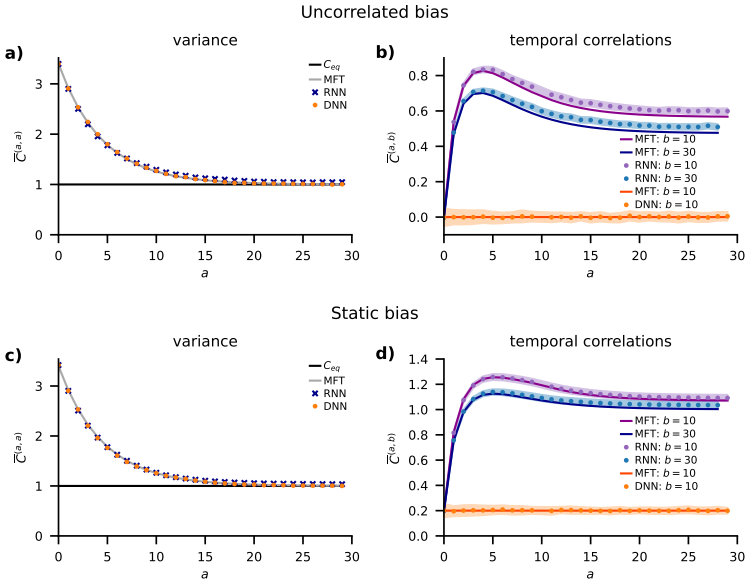

Figure 1. Mean-field theory for DNN and RNN with a single input. **a)** Average variance in mean-field theory $\overline{C}^{(a,a)}$ (Eq. (31); solid gray curve) and estimate $\frac{1}{n_a}\sum_i h_i^{(a)} h_i^{(a)}$ from simulation, averaged over 100 realizations of networks, for biases that are uncorrelated across time/layers (blue crosses RNN; orange dots DNN). **b)** Cross-covariance $\overline{C}^{(a,b)}$ as a function of the hidden layer index a for fixed $b \in \{10, 30\}$ and uncorrelated biases. RNN: Mean-field theory (solid dark blue and dark magenta). Mean (blue / purple dots) and standard error of the mean (light blue / light purple tube) of $\frac{1}{n_a}\sum_i h_i^{(a)} h_i^{(b)}$ estimated from simulation of 100 network realizations. DNN: Mean (orange dots) and standard error of the mean of $\frac{1}{n_a}\sum_i h_i^{(a)} h_i^{(b)}$ estimated from simulation of 100 network realizations. Other parameters $g_0^2 = g^2 = 1.6$, $\sigma^2 = 0.2$, finite layer width $n_a = 2000$, $A = 30$ hidden layers, ReLU activation $\phi(x) = \max(0, x)$ and Gaussian inputs $\boldsymbol{x} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(1, 1)$ with $n_{\text{in}} = 10^5$. **c)** Same as a) but for biases that are static across time/layers. **d)** Same as b) but for the static bias case.

4. Comparison of RNNs and DNNs

Above, we derived the mean-field equations (33) for the kernel matrix $\boldsymbol{K}$ using a field-theoretic approach. Here, we investigate differences in the mean-field distributions of the different network architectures, starting with the kernel and considering the predictive distribution afterwards.

4.1. Kernel

The diagonal elements, $\overline{C}^{(a,a)}$ for the single-input case in Eq. (31) and equivalently $\overline{C}^{(a,a)}_{\alpha,\beta}$ for the multiple-input-sequences case in Eq. (33), are identical for RNNs and DNNs, because $M_{a,a} = 1$ for both architectures. This implies that the equal-time or within-layer statistics, correspondingly, is the same in both architectures. The reason is that the iterations Eq. (31) and Eq. (33) for equal-time points $a = b$ form closed sets of equations; they can be solved independently of the statistics for different time points $a \neq b$. Formally, this follows from the marginalization property of the Gaussian, which implies that any subset of a multivariate Gaussian is Gaussian, too, with a covariance matrix that is the corresponding sector of the covariance matrix of all variables [19]. The precise agreement of this mean-field prediction with the average correlation estimated from direct simulation is shown in Figure 1a and c for the single-input case for both uncorrelated (a) and static biases (c) across time or layers, respectively.

A notable difference between RNN and DNN is that activity in the RNN is correlated across time steps due to the shared weights, even if biases are uncorrelated in time, as shown in Figure 1b. Static biases simply strengthen the correlations across time steps (see Figure 1d). For DNNs, in contrast, cross-layer correlations only arise due to biases that are correlated across layers, because weights are drawn independently for each layer. This is shown in Figure 1b and d: Correlations vanish for DNNs in the uncorrelated bias case (b) and take on the value σ^2, the variance of the bias, in the static bias case (d). Again, the mean-field theory accurately predicts the non-zero correlations across time in the RNN as well as the correlations across layers generated by the correlated biases in the DNN. In the RNN, temporal correlations show a non-trivial interplay due to the shared weights across time. We observe an instability that can build up by this mechanism in finite-size RNNs, even in parameter regimes that are deemed stable in mean-field theory (see Appendix 6.2, Figure 3).

In a particular case, the correlations across time steps also vanish for the RNN: we show by induction that off-diagonal elements vanish for point-symmetric activation functions if inputs are only provided in the initial time step, $\{\boldsymbol{X}^{(0)}, 0, \ldots, 0\} \equiv \boldsymbol{X}$, and the bias is absent, $\sigma = 0$ (or uncorrelated across time steps). Assuming that $\overline{C}^{(a-1,b-1)}_{\alpha,\beta} \overset{a\neq b}{=} 0$, we have

$$\overline{C}^{(a,b)}_{\alpha,\beta} = g^2 \left\langle \phi(h^{(a-1)}_{\alpha}) \right\rangle_{h^{(a-1)}_{\alpha}} \left\langle \phi(h^{(b-1)}_{\beta}) \right\rangle_{h^{(b-1)}_{\beta}} \overset{\phi \text{ odd}}{=} 0 \tag{38}$$

with $h^{(a-1)}_{\alpha} \sim \mathcal{N}(0, \overline{C})$ and $h^{(b-1)}_{\beta} \sim \mathcal{N}(0, \overline{C})$. Hence, if the pre-activations $h^{(a-1)}_{\alpha}, h^{(b-1)}_{\beta}$ at points $a-1$ and $b-1$ are uncorrelated, also $h^{(a)}_{\alpha}, h^{(b)}_{\beta}$ will be uncorrelated. The base case of the induction proof follows from the independence of the input weights $\boldsymbol{W}^{(\text{in})}$ and the recurrent weights $\boldsymbol{W}$: correlations between time point zero and other time points are zero. Therefore, by induction in time, time points will be uncorrelated at any point, meaning that for odd activations ϕ and the considered input layer, the solutions of the mean-field equations are the same for DNNs and RNNs.

4.2. Predictive distribution

Coming back to the general case, we next ask if the different off-diagonal elements of the mean-field equations for RNN and DNN have observable consequences. The answer

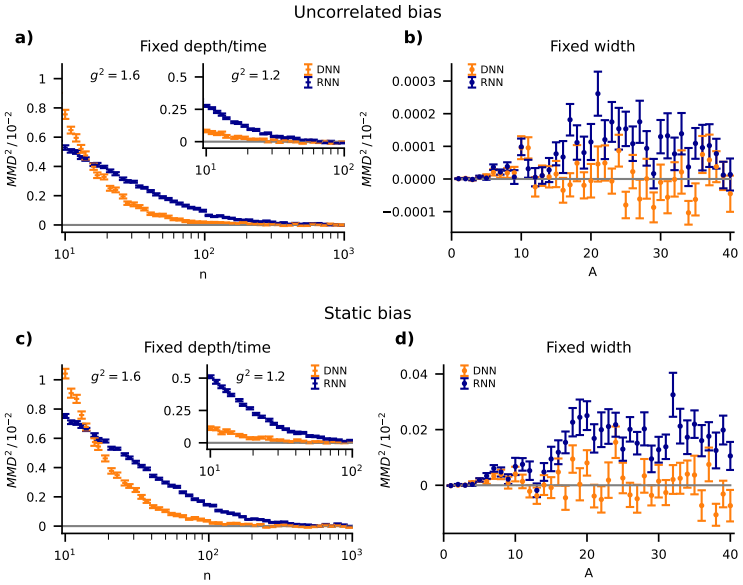

Figure 2. Convergence of RNN and DNN towards the mean-field theory. Maximum mean-discrepancy MMD^2 for a radial basis function kernel with length scale $l = 1/2$ [38] between the empirical distribution of scalar outputs y_α and the Gaussian distribution with covariance matrix $K_{\alpha\beta} = \overline{C}^{(A+1,A+1)}_{\alpha,\beta}$ predicted by MFT Eq. (33). Empirical MMD^2 estimation across 2000 realizations (W, ξ). Average over 40 realizations of $\{\boldsymbol{x}_\alpha\}_{\alpha=1,\dots,10}$, $x_{\alpha,i} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ and $\dim(\boldsymbol{x}_\alpha) = 4$ (error bars showing standard error of the mean). ReLU activation $\phi(x) = \max(0, x)$. **a)** MMD^2 as a function of the width of the network layer n for $g^2 = 1.6$ and $g^2 = 1.2$ (inset), with $A = 15$ and $\sigma^2 = 0.2$ and uncorrelated biases across time/layers. **b)** MMD^2 as a function of the depth or duration A, for width $n = 500$, $g^2 = 1.6$, and $\sigma^2 = 0.2$ and uncorrelated biases. **c)** Same as a) but for biases that are static across time/layers. **d)** Same as b) but for the static bias case.

is no if a linear readout is taken at a single time point or layer A, correspondingly (cf. Eq. (8) for the readout): This is a direct consequence of the identical diagonal elements of the covariance $\overline{C}^{(a,a)}_{\alpha,\beta}$, so that the predictive distribution Eq. (36) for the RNN and the DNN is identical in mean-field theory; the two architectures have the same Gram matrix $K_{\alpha\beta} = \overline{C}^{(A+1,A+1)}_{\alpha,\beta}$ and thus the same predictive distribution Eq. (36). This means that the two architectures have identical computational capabilities in the limit of infinite layer width.

To check how quickly the mean-field theory is approached by finite-size networks, we measure the maximum mean-discrepancy [9, 38] between the Gaussian distribution with covariance matrix $K_{\alpha\beta}$ and the empirical joint distribution of a set of scalar

outputs y_α, Eq. (8), across realizations of W and ξ. The inputs $\boldsymbol{x}_\alpha$ are random patterns presented to the first layer or time step, respectively. We find that convergence is rather fast for both architectures (Figure 2 a and c). For sufficiently deep architectures $A \gg 1$ as well as both uncorrelated and static biases, RNNs systematically show a slower convergence than DNNs, which could be anticipated due to the smaller number of independently drawn Gaussian weights, N^2 versus $A\,N^2$. This observation is in line with the MMD being larger for the RNN than for the DNN for $A \gtrsim 15$ (Figure 2 b and d). This is also consistent with the coherent interplay of shared connectivity and correlated activity across time steps in the RNN (see Appendix 6.2, Figure 3). Overall, we find a faster convergence for uncorrelated biases than for biases that are static over time or layers, respectively.

The temporal correlations present in RNNs become relevant in the case of sequence processing. In such a setting, the network in each time step a receives a time dependent input $\boldsymbol{x}_\alpha^{(a)}$ with a non-trivial temporal correlation structure $\mathbf{x}_\alpha^{(a+\tau)\mathrm{T}}\mathbf{x}_\beta^{(a)}$ that drives the temporal correlations $\overline{C}_{\alpha,\beta}^{(a'+\tau,a')}$ of the RNN activations for $a' \geq a$, see Eq. (33). If the latter are read out in each time step, temporal correlations enter the kernel and thus influence task performance.

We finally note that we here use a separate readout layer. The realization of readout weights as independent Gaussian variables causes vanishing temporal correlations between the readouts and the activity in previous layers or time steps, respectively. For the Gaussian kernel, however, the presence or absence of a readout layer does not make any difference. Alternatively, the readout of n_{out} signals could be taken from an arbitrary choice of n_{out} neurons in the last layer or time step, respectively, leading to the same kernel.

5. Discussion

We present a unified derivation of the mean-field theory for deep (DNN) and recurrent neural networks (RNN) using field-theoretical methods. The derivation in particular yields the Gaussian process kernel that predicts the performance of networks trained in a Bayesian way.

The mean-field theories for the statistics within a layer of the DNN and for the equal-time statistics of the RNN are identical, even if temporally correlated input sequences are supplied to the latter network. The reason is that the mean-field equations (33) form a closed system of equations for this subset of the statistics; they can be solved independently of the correlations across time or layers, respectively. This justifies the 'annealed approximation' [28] for RNNs where the couplings are redrawn at each time step—which corresponds to the DNN-prior. It is also compatible with earlier work [39] which compares simulations of networks with tied weights (RNN) to the mean-field theory for untied weights (DNN). Intriguingly, the equivalence of the equal-time statistics implies that the predictive distributions $p(\boldsymbol{y}^* \,|\, \boldsymbol{x}^*, \boldsymbol{Y}, \boldsymbol{X})$ of DNNs and RNNs are identical, given the readout is taken only from the final layer or the last time step, respectively.

There are qualitative differences between the mean-field theories for the correlations across time in the RNN and across layers of the DNN: correlations across layers vanish in the DNN, while the weight sharing in the RNN generally causes non-trivial correlations across time. For point-symmetric activation functions, these correlations also vanish in the RNN if the bias is absent (or uncorrelated across time

steps) and the input is provided only in the first step. In general, a linear readout from activations that are taken across different time points or layers, respectively, yields different Gaussian process kernels for the RNN compared to the DNN.

Numerically, the convergence of finite-size networks of both architectures to the mean-field theory is generally fast. The RNN converges typically slower than the DNN, at least for long times and correspondingly deep networks. We hypothesize that the temporally correlated activity in the RNN is the cause: The realization of the coupling matrix is the same for all time steps. Also, fluctuations of the activity are coherent over time. Activity and connectivity therefore interact coherently over multiple time steps, so that variations of the connectivity across realizations may cause a corresponding variability of activity. In a DNN, in contrast, both activity and connectivity are uncorrelated across layers, so that variations due to different realizations of the couplings average out over layers.

Identical mean-field theories in the single-input case and for point-symmetric activation functions were already presented in ref. [29] in the context of a characterization of the space of Boolean functions implemented by randomly coupled DNNs and RNNs. Since our work differs on a conceptual level, the implications of the results differ: In the Bayesian inference picture, the equivalent mean-field theories imply identical performance of the trained networks for both architectures at large width; for the characterization of computed Boolean functions, the equivalent mean-field theories imply an equivalent set of functions implemented by any two random instances of the two architectures at large width. The conceptual difference leads to further differences on the technical level: The inputs and outputs considered here include analog values and they are presented not only to the first layer or time step, respectively, but also in a sequential manner at subsequent times or layers. Finally, the disorder average plays a subtle but fundamentally different role in the two works: In ref. [29], the disorder average extracts the typical behavior of any single, sufficiently large, instance of a randomly coupled network. In contrast, in the Bayesian framework considered in this manuscript, the disorder average naturally arises from the marginalization of the parameter prior, i.e., one here considers ensembles of random networks.

The analysis of RNNs and DNNs in this manuscript is based on methods from statistical field theory and our results are formulated in that language [31]. It is worth noting that this field-theoretical approach can be connected to a mathematically more rigorous approach based on large-deviation theory [40].

The main limitation of the presented results is their validity for networks with large widths. There has been previous theoretical work on networks of finite width $n_\ell < \infty$ that is, however, restricted to DNNs: Refs. [13, 14, 17, 41] have presented approaches based on perturbation theory, while refs. [15, 18] employed an Edgeworth expansion. The dynamics of the neural-tangent kernel for deep networks with finite width has been studied in ref. [16]. For specific deep networks of finite width with linear or ReLU activation functions the single-input prior was computed exactly in terms of the Meijer G function in ref. [42]. The formalism proposed here paves the way for a systematic study of generic deep and recurrent networks beyond the leading order in n. Computing finite-size corrections in the presented formalism amounts to calculating fluctuation corrections of the auxiliary fields which is a standard task in field theory [31, 32]. It requires the spectrum of the Hessian $\partial^2 \mathcal{S}\left(C, \tilde{C}\right)/\{\partial C, \partial \tilde{C}\}\big|_{\bar{C},\bar{\tilde{C}}=0}$, evaluated at the mean-field point. Such an approach yields small non-Gaussian corrections to the prior that, moreover, depend on n. The corrections therefore provide quantitative

insight into the limits of the equivalence between RNNs and DNNs at finite widths, and offer a handle to study the capacity of a network in relation to its resources, the number of weights.

Acknowledgments

We would like to thank Bo Li, Alexander Mozeika, and David Saad for bringing their related work to our attention. This work was partially supported by the European Union's Horizon 2020 research and innovation program under Grant agreement No. 945539 (Human Brain Project SGA3), the Helmholtz Association Initiative and Networking Fund under project number SO-092 (Advanced Computing Architectures, ACA), the German Federal Ministry for Education and Research (BMBF Grant 01IS19077A), and the Excellence Initiative of the German federal and state governments (ERS PF-JARA-SDS005).

6. Appendix

6.1. Unified field theory for multiple input sequences

Here, we show the derivation of the mean-field equations with more than one input sequence $\{\boldsymbol{x}_\alpha^{(0)}, \ldots, \boldsymbol{x}_\alpha^{(A)}\}$, the generalization of the derivation presented in the main text. We introduce Greek indices $\alpha \in \{1, \ldots, n_D\}$ for the different input vectors that we also call 'replicas' in the following. Equations for the single-replicon case in the main text can be obtained by setting $n_D = 1$; the non-sequential input case follows by setting $\boldsymbol{x}_\alpha^{(a)} = 0$ for $a > 0$ and all α.

6.1.1. Action and auxiliary variables We start from the parameterized likelihood for multiple replicas

$$
\begin{aligned}
p(\boldsymbol{Y} \,|\, \boldsymbol{X}, \boldsymbol{\theta}) = \prod_{\alpha=1}^{n_D} \Bigg\{ \int \mathcal{D}\boldsymbol{h}_\alpha \, \delta\Big(\boldsymbol{y}_\alpha - \boldsymbol{W}^{(\mathrm{out})} \boldsymbol{\phi}_\alpha^{(A)} - \boldsymbol{\xi}^{(A+1)}\Big) \\
\times \prod_{a=1}^{A} \delta\left(\boldsymbol{h}_\alpha^{(a)} - \boldsymbol{W}^{(a)} \boldsymbol{\phi}_\alpha^{(a-1)} - \boldsymbol{W}^{(\mathrm{in},a)} \boldsymbol{x}_\alpha^{(a)} - \boldsymbol{\xi}^{(a)}\right) \\
\times \delta\left(\boldsymbol{h}_\alpha^{(0)} - \boldsymbol{W}^{(\mathrm{in},0)} \boldsymbol{x}_\alpha^{(0)} - \boldsymbol{\xi}^{(0)}\right) \Bigg\}.
\end{aligned}
$$

Expressing the Dirac distributions as integrals $\delta(x) = \int_{i\mathbb{R}} \frac{d\tilde{x}}{2\pi i} \, e^{\tilde{x}\, x}$, we obtain for the network prior $p(\boldsymbol{Y} \,|\, \boldsymbol{X}) = \int d\boldsymbol{\theta} \, p(\boldsymbol{Y} \,|\, \boldsymbol{X}, \boldsymbol{\theta}) \, p(\boldsymbol{\theta})$ the expression

$$
\begin{aligned}
p(\boldsymbol{Y} \,|\, \boldsymbol{X}) = \prod_{\alpha=1}^{n_D} \left\{ \int d\tilde{\boldsymbol{y}}_\alpha \int D\boldsymbol{h}_\alpha \int D\tilde{\boldsymbol{h}}_\alpha \right\} e^{\tilde{y}_{i,\alpha} y_{i,\alpha} + \sum_{a=0}^{A} \tilde{h}_{i,\alpha}^{(a)} h_{i,\alpha}^{(a)}} \\
\times \left\langle e^{-\tilde{y}_{i,\alpha} W_{ij}^{(\mathrm{out})} \phi_{j,\alpha}^{(A)}} \right\rangle_{\boldsymbol{W}^{(\mathrm{out})}} \left\langle e^{-\sum_{a=1}^{A} \tilde{h}_{i,\alpha}^{(a)} W_{ij}^{(a)} \phi_{j,\alpha}^{(a-1)}} \right\rangle_{\{\boldsymbol{W}^{(a)}\}} \\
\times \left\langle e^{-\sum_{a=0}^{A} \tilde{h}_{i,\alpha}^{(a)} W_{ij}^{(\mathrm{in},a)} x_{j,\alpha}^{(a)}} \right\rangle_{\{\boldsymbol{W}^{(\mathrm{in},a)}\}} \\
\times \left\langle e^{-\sum_{\alpha=1}^{n_D} \sum_{a=0}^{A} \tilde{h}_{i,\alpha}^{(a)} \xi_i^{(a)} - \sum_{\alpha=1}^{n_D} \tilde{y}_{i,\alpha} \xi_i^{(A+1)}} \right\rangle_{\{\boldsymbol{\xi}^{(a)}\}}. \qquad (39)
\end{aligned}
$$

Here, and throughout this section, we use an implicit summation convention for lower indices that appear twice in the exponent, e.g., $\tilde{y}_{i,\alpha} y_{i,\alpha} \equiv \sum_{\alpha=1}^{n_D} \sum_{i=1}^{n_{\text{out}}} \tilde{y}_{i,\alpha} y_{i,\alpha}$, but write the sum over time steps explicitly to avoid ambiguities in their limits. Note that for the DNN, the number of neurons per layer can differ such that formally the upper limits of the implicit sums over neuron indices i or j depends on the layer index a. We also used the independence of the different weight matrices and biases to obtain factorizing expectation values in Eq. (39).

In the following, we compute these expectation values separately, starting with the output weights and biases. These are independent across neurons and we obtain

$$\left\langle \exp\left(-\sum_{\alpha=1}^{n_D} \tilde{y}_{i,\alpha} \xi_i^{(A+1)} \right) \right\rangle_{\boldsymbol{\xi}^{(A+1)}} = \exp\left(\frac{\sigma_{A+1}^2}{2} \sum_{\alpha,\beta=1}^{n_D} \tilde{y}_{i,\alpha} \tilde{y}_{i,\beta} \right),$$

$$\left\langle \exp\left(-\tilde{y}_{i,\alpha} W_{ij}^{(\text{out})} \phi_{j,\alpha}^{(A)} \right) \right\rangle_{\boldsymbol{W}^{(\text{out})}} = \exp\left(\frac{g_{A+1}^2}{2n_A} \tilde{y}_{i,\alpha} \phi_{j,\alpha}^{(A)} \phi_{j,\beta}^{(A)} \tilde{y}_{i,\beta} \right).$$

Now, we calculate the respective averages for the RNN and DNN separately. For a RNN, the weight sharing $\boldsymbol{W}^{(a)} \equiv \boldsymbol{W}$ across time steps a leads to a double sum $\sum_{a,b}$ appearing in the average over the recurrent part

$$\begin{aligned} &\left\langle \exp\left(-\sum_{a=1}^{A} \tilde{h}_{i,\alpha}^{(a)} W_{ij} \phi_{j,\alpha}^{(a-1)} \right) \right\rangle_{\boldsymbol{W}} \\ &\quad = \exp\left(\frac{1}{2} \sum_{a,b=1}^{A} \frac{g^2}{n} \tilde{h}_{i,\alpha}^{(a)} \phi_{j,\alpha}^{(a-1)} \phi_{j,\beta}^{(b-1)} \tilde{h}_{i,\beta}^{(b)} \right), \quad \text{RNN}. \end{aligned}$$

In contrast, for a DNN, the analogous calculation leads to a single sum $\sum_a$

$$\begin{aligned} &\left\langle \exp\left(-\sum_{a=1}^{A} \tilde{h}_{i,\alpha}^{(a)} W_{ij}^{(a)} \phi_{j,\alpha}^{(a-1)} \right) \right\rangle_{\{\boldsymbol{W}^{(a)}\}} \\ &\quad = \prod_{a=1}^{A} \left\langle \exp\left(-\tilde{h}_{i,\alpha}^{(a)} W_{ij}^{(a)} \phi_{j,\alpha}^{(a-1)} \right) \right\rangle_{\boldsymbol{W}^{(a)}} \\ &\quad = \exp\left(\frac{1}{2} \sum_{a=1}^{A} \frac{g_a^2}{n_{a-1}} \tilde{h}_{i,\alpha}^{(a)} \phi_{j,\alpha}^{(a-1)} \phi_{j,\beta}^{(a-1)} \tilde{h}_{i,\beta}^{(a)} \right), \quad \text{DNN}. \end{aligned}$$

The calculation for the inputs and biases is analogous; for the RNN it yields

$$\begin{aligned} &\left\langle \exp\left(-\sum_{a=0}^{A} \sum_{\alpha=1}^{n_D} \tilde{h}_{i,\alpha}^{(a)} \xi_i \right) \right\rangle_{\boldsymbol{\xi}} \\ &\quad = \exp\left(\frac{\sigma^2}{2} \sum_{a,b=0}^{A} \sum_{\alpha,\beta=1}^{n_D} \tilde{h}_{i,\alpha}^{(a)} \tilde{h}_{i,\beta}^{(b)} \right), \quad \text{RNN}, \end{aligned}$$

$$\begin{aligned} &\left\langle \exp\left(-\sum_{a=0}^{A} \tilde{h}_{i,\alpha}^{(a)} W_{ij}^{(\text{in})} x_{j,\alpha}^{(a)} \right) \right\rangle_{\boldsymbol{W}^{(\text{in})}} \\ &\quad = \exp\left(\frac{1}{2} \sum_{a,b=0}^{A} \frac{g_0^2}{n_{\text{in}}} \tilde{h}_{i,\alpha}^{(a)} x_{j,\alpha}^{(a)} x_{j,\beta}^{(b)} \tilde{h}_{i,\beta}^{(b)} \right), \quad \text{RNN}. \end{aligned}$$

For the DNN, we get

$$\begin{aligned}
&\left\langle \exp\left(-\sum_{a=0}^{A}\sum_{\alpha=1}^{n_D}\tilde{h}_{i,\alpha}^{(a)}\xi_i^{(a)}\right)\right\rangle_{\{\boldsymbol{\xi}^{(a)}\}} \\
&\quad=\prod_{a=0}^{A}\left\langle \exp\left(-\sum_{\alpha=1}^{n_D}\tilde{h}_{i,\alpha}^{(a)}\xi_i^{(a)}\right)\right\rangle_{\boldsymbol{\xi}^{(a)}} \\
&\quad=\exp\left(\frac{1}{2}\sum_{a=0}^{A}\sigma_a^2\sum_{\alpha,\beta=1}^{n_D}\tilde{h}_{i,\alpha}^{(a)}\tilde{h}_{i,\beta}^{(a)}\right), \qquad \text{DNN},
\end{aligned}$$

$$\begin{aligned}
&\left\langle \exp\left(-\sum_{a=0}^{A}\tilde{h}_{i,\alpha}^{(a)}W_{ij}^{(\mathrm{in},a)}x_{j,\alpha}^{(a)}\right)\right\rangle_{\{\boldsymbol{W}^{(\mathrm{in},a)}\}} \\
&\quad=\prod_{a=0}^{A}\left\langle \exp\left(-\tilde{h}_{i,\alpha}^{(a)}W_{ij}^{(\mathrm{in},a)}x_{j,\alpha}^{(a)}\right)\right\rangle_{\boldsymbol{W}^{(\mathrm{in},a)}} \\
&\quad=\exp\left(\frac{1}{2}\sum_{a=0}^{A}\frac{g_0^2}{n_{\mathrm{in}}}\tilde{h}_{i,\alpha}^{(a)}x_{j,\alpha}^{(a)}x_{j,\beta}^{(a)}\tilde{h}_{i,\beta}^{(a)}\right), \qquad \text{DNN}.
\end{aligned}$$

For the RNN, the replicas as well as the time steps are coupled by the products $\phi_{j,\alpha}^{(a-1)}\phi_{j,\beta}^{(b-1)}$ and $x_{j,\alpha}^{(a)}x_{j,\beta}^{(b)}$, while for the DNN only products of terms within the same layer occur, $\phi_{j,\alpha}^{(a-1)}\phi_{j,\beta}^{(a-1)}$ and $x_{j,\alpha}^{(a)}x_{j,\beta}^{(a)}$. As we will show below, this leads to different layers in the DNN being uncorrelated, while different time steps in the RNN are correlated.

The products of nonlinearly transformed pre-activations $\phi_{i,\alpha}^{(a)} \equiv \phi(h_{i,\alpha}^{(a)})$ render the integrations in Eq. (39) analytically non-solvable. To find a suitable approximation, we insert auxiliary variables in time (a,b) and in replica space (α,β), which account for the replica and time-step coupling. Introducing these, the system decouples in the neuron indices i. We combine RNN and DNN by defining the auxiliary variables

$$\begin{aligned}
C_{\alpha,\beta}^{(a,b)} = M_{a,b}\Bigg[&\sigma_a^2 + \frac{g_a^2}{n_{a-1}}\mathbb{1}_{a\geq 1,b\geq 1}\,\phi_{i,\alpha}^{(a-1)}\phi_{i,\beta}^{(b-1)} \\
&+ \frac{g_0^2}{n_{\mathrm{in}}}\mathbb{1}_{a\leq A,b\leq A}\,x_{i,\alpha}^{(a)}x_{i,\beta}^{(b)}\Bigg]
\end{aligned} \tag{40}$$

for $0 \leq a,b \leq A+1$ with $M_{a,b}$ defined in Eq. (25), $g_a = g$ for $1 \leq a \leq A$ in RNN, and $n_{-1} \equiv n_{\mathrm{in}}$. The indicator functions $\mathbb{1}_{a\geq 1,b\geq 1}$ and $\mathbb{1}_{a\leq A,b\leq A}$ ensure that the respective terms vanish when they are not present, e.g., the recurrent term $\phi_{i,\alpha}^{(a-1)}M_{a,b}\phi_{i,\beta}^{(b-1)}$ in the first step $a=b=0$. As above, there is an implicit sum over the neuron indices i on the right hand side.

We introduce these auxiliary variables by means of Dirac distributions expressed as Fourier integrals

$$\begin{aligned}
\delta[\text{Eq. (40)}] = &\prod_{\alpha,\beta=1}^{n_D}\prod_{a,b=0}^{A+1}\left\{n_{a-1}\int_{i\mathbb{R}}\frac{d\tilde{C}_{\alpha,\beta}^{(a,b)}}{2\pi i}\right\} \\
&\times\exp\left(-\sum_{a,b=0}^{A+1}n_{a-1}\tilde{C}_{\alpha,\beta}^{(a,b)}\left(C_{\alpha,\beta}^{(a,b)} - \sigma_a^2\,M_{a,b}J_{\alpha,\beta}\right)\right)
\end{aligned}$$

$$\times \exp\left(\sum_{a,b=1}^{A+1} \tilde{C}^{(a,b)}_{\alpha,\beta} g_a^2 \, \phi^{(a-1)}_{i,\alpha} M_{a,b} \phi^{(b-1)}_{i,\beta})\right)$$
$$\times \exp\left(\sum_{a,b=0}^{A} n_{a-1} \tilde{C}^{(a,b)}_{\alpha,\beta} \frac{g_0^2}{n_{\text{in}}} x^{(a)}_{i,\alpha} M_{a,b} x^{(b)}_{i,\beta}\right), \tag{41}$$

where we inserted $J_{\alpha,\beta} = 1$ for all α and β to imply the summation over α, β that accounts for the common biases across replicas. Used in the integrand of Eq. (39), this leads to

$$\begin{aligned} p(\boldsymbol{Y} \,|\, \boldsymbol{X}) = \prod_{\alpha=1}^{n_D} \left\{ \int d\tilde{\boldsymbol{y}}_\alpha \int D\boldsymbol{h}_\alpha \int D\tilde{\boldsymbol{h}}_\alpha \right\} \prod_{\alpha,\beta=1}^{n_D} \left\{ \int D\tilde{C}_{\alpha,\beta} \int DC_{\alpha,\beta} \right\} \\ \times \exp\left(\tilde{y}_{i,\alpha} y_{i,\alpha} + \frac{1}{2} \tilde{y}_{i,\alpha} C^{(A+1,A+1)}_{\alpha,\beta} \tilde{y}_{i,\beta} \right) \\ \times \exp\left(\sum_{a=0}^{A} \tilde{h}^{(a)}_{i,\alpha} h^{(a)}_{i,\alpha} + \sum_{a,b=0}^{A} \frac{1}{2} \tilde{h}^{(a)}_{i,\alpha} C^{(a,b)}_{\alpha,\beta} \tilde{h}^{(b)}_{i,\beta} \right) \\ \times \exp\left(- \sum_{a,b=0}^{A+1} n_{a-1} \tilde{C}^{(a,b)}_{\alpha,\beta} \left(C^{(a,b)}_{\alpha,\beta} - \sigma_a^2 \, M_{a,b} J_{\alpha,\beta} \right) \right) \\ \times \exp\left(\sum_{a,b=1}^{A+1} \tilde{C}^{(a,b)}_{\alpha,\beta} g_a^2 \phi^{(a-1)}_{i,\alpha} M_{a,b} \phi^{(b-1)}_{i,\beta}) \right) \\ \times \exp\left(\sum_{a,b=0}^{A} n_{a-1} \tilde{C}^{(a,b)}_{\alpha,\beta} \frac{g_0^2}{n_{\text{in}}} x^{(a)}_{i,\alpha} M_{a,b} x^{(b)}_{i,\beta} \right) \end{aligned} \tag{42}$$

with $DC_{\alpha,\beta} = \prod_{a,b=0}^{A+1} dC^{(a,b)}_{\alpha,\beta}$, $D\tilde{C}_{\alpha,\beta} = \prod_{a,b=0}^{A+1} \frac{n_{a-1} d\tilde{C}^{(a,b)}_{\alpha,\beta}}{2\pi i}$.

We see in Eq. (42) that there are no auxiliary variables $C^{(a,b)}_{\alpha,\beta}$ that couple the output layer ($a = A+1$, second line) with variables $\boldsymbol{h}^{(a)}_\alpha, \tilde{\boldsymbol{h}}^{(a)}_\alpha$ in the rest of the network ($0 \le a \le A$). This is a consequence of the independence of the priors on the associated weights. We further see in Eq. (42) that no products of variables with different neuron indices appear. The exponential thus factorizes into n_a identical terms for each a. Rearranging the integrations, we obtain

$$p(\boldsymbol{Y} \,|\, \boldsymbol{X}) = \prod_{\alpha=1}^{n_D} \left\{ \int d\tilde{\boldsymbol{y}}_\alpha \right\} e^{\tilde{y}_{i,\alpha} y_{i,\alpha}} \left\langle e^{\frac{1}{2} \tilde{y}_{i,\alpha} C^{(A+1,A+1)}_{\alpha,\beta} \tilde{y}_{i,\beta}} \right\rangle_{\tilde{C},C} \tag{43}$$

where the expectation value is computed with respect to the action

$$\mathcal{S}_{\text{aux}}(C, \tilde{C}) = -n \sum_{a,b=0}^{A+1} \nu_{a-1} \tilde{C}^{(a,b)}_{\alpha,\beta} C^{(a,b)}_{\alpha,\beta} + n \, \mathcal{W}_{\text{aux}}(\tilde{C} \,|\, C) \tag{44}$$

of the auxiliary variables $\tilde{C}^{(a,b)}_{\alpha,\beta}, C^{(a,b)}_{\alpha,\beta}$. This action comprises the nontrivial part of the dynamics of the network in the cumulant generating functional

$$\mathcal{W}_{\text{aux}}(\tilde{C} \,|\, C) = \frac{1}{n} \ln \left\langle e^{\sum_{a,b=1}^{A+1} \tilde{C}^{(a,b)}_{\alpha\beta} g_a^2 \phi^{(a-1)}_{i,\alpha} M_{a,b} \phi^{(b-1)}_{i,\beta}} \right\rangle_{\{h^{(a)}_{i,\alpha}\}}$$

$$+\sum_{a,b=0}^{A}\nu_{a-1}\tilde{C}^{(a,b)}_{\alpha,\beta}\frac{g_0^2}{n_{\text{in}}}x^{(a)}_{i,\alpha}M_{a,b}x^{(b)}_{i,\beta}$$
$$+\sum_{a,b=0}^{A+1}\nu_{a-1}\tilde{C}^{(a,a)}_{\alpha,\beta}\sigma_a^2 M_{a,b}J_{\alpha,\beta} \tag{45}$$

where $\{h^{(a)}_{i,\alpha}\}$ describes the Gaussian statistics of a single pre-activation $h^{(a)}_{i,\alpha}$ with covariance matrix $\langle h^{(a)}_{i,\alpha}h^{(b)}_{i,\beta}\rangle = C^{(a,b)}_{\alpha,\beta}\delta_{i,j}$ across neurons i,j, time steps or layers a,b, and inputs α,β. Here $\nu_a = n_a/n$ denotes the relative layer sizes in the DNN.

To show that

$$\frac{1}{n}\ln\left\langle e^{\sum_{a,b=1}^{A+1}\tilde{C}^{(a,b)}_{\alpha\beta}g_a^2\phi^{(a-1)}_{i,\alpha}M_{a,b}\phi^{(b-1)}_{i,\beta}}\right\rangle_{\{h^{(a)}_{i,\alpha}\}} = O(1) \tag{46}$$

and thus $\mathcal{W}_{\text{aux}}(\tilde{C}\,|\,C) = O(1)$, i.e., that $\mathcal{W}_{\text{aux}}(\tilde{C}\,|\,C)$ does not scale with n, we consider RNN and DNN separately. For the RNN, the result immediately follows because the neurons are uncorrelated, $\langle h^{(a)}_{i,\alpha}h^{(b)}_{i,\beta}\rangle = C^{(a,b)}_{\alpha,\beta}\delta_{i,j}$, which factorizes the expectations and leads to a sum over n identical terms:

$$\frac{1}{n}\ln\left\langle e^{\sum_{a,b=1}^{A+1}\tilde{C}^{(a,b)}_{\alpha\beta}g_a^2\phi^{(a-1)}_{i,\alpha}\phi^{(b-1)}_{i,\beta}}\right\rangle_{\{h^{(a)}_{i,\alpha}\}}$$
$$= \ln\left\langle e^{\sum_{a,b=1}^{A+1}\tilde{C}^{(a,b)}_{\alpha\beta}g_a^2\phi^{(a-1)}_{\alpha}\phi^{(b-1)}_{\beta}}\right\rangle_{\{h^{(a)}_{\alpha}\}}.$$

For the DNN, one first notices that, by definition, $C^{(a,b)}_{\alpha,\beta} = 0$ for $a \neq b$, so different layers decouple in Eq. (45). Formally this can be seen by solving the integrals over the corresponding variables $\tilde{C}^{(a,b)}_{\alpha,\beta}$ with $a \neq b$. This factorization allows us to study each layer separately and decouple the n_a neurons:

$$\frac{1}{n}\ln\left\langle e^{\sum_{a=1}^{A+1}\tilde{C}^{(a,a)}_{\alpha\beta}g_a^2\phi^{(a-1)}_{i,\alpha}\phi^{(a-1)}_{i,\beta}}\right\rangle_{\{h^{(a)}_{i,\alpha}\}}$$
$$= \sum_{a=1}^{A+1}\nu_{a-1}\ln\left\langle e^{\tilde{C}^{(a,a)}_{\alpha\beta}g_a^2\phi^{(a-1)}_{\alpha}\phi^{(a-1)}_{\beta}}\right\rangle_{\{h^{(a)}_{\alpha}\}}.$$

Consequently, for both architectures Eq. (46) holds and $\mathcal{W}_{\text{aux}}(\tilde{C}\,|\,C) = O(1)$.

6.1.2. Saddle-point approximation The action $\mathcal{S}_{\text{aux}}$ in the auxiliary fields scales with the number of neurons n. In the limit $n\to\infty$, a saddle-point approximation of the integrals over $\tilde{C}$ and C appearing in the expectation value in Eq. (43) becomes exact. The saddle points are determined by the stationary points of the action $\mathcal{S}_{\text{aux}}$ as $\frac{\partial}{\partial\tilde{C}^{(a,b)}_{\alpha,\beta}}\mathcal{S}_{\text{aux}}(C,\tilde{C}) \overset{!}{=} 0$ and $\frac{\partial}{\partial C^{(a,b)}_{\alpha,\beta}}\mathcal{S}_{\text{aux}}(C,\tilde{C}) \overset{!}{=} 0$, leading to

$$\overline{\tilde{C}}^{(a,b)}_{\alpha\beta} = 0, \tag{47}$$
$$\overline{C}^{(a,b)}_{\alpha\beta} = M_{a,b}\Big[\sigma_a^2 + \frac{g_0^2}{n_{\text{in}}}\mathbb{1}_{a\le A,b\le A}x^{(a)}_{i,\alpha}x^{(b)}_{i,\beta}$$
$$+ g_a^2\mathbb{1}_{a\ge 1,b\ge 1}\langle\phi(h^{(a-1)}_{\alpha})\phi(h^{(b-1)}_{\beta})\rangle_{\{h^{(a)}_{\alpha}\}\sim\mathcal{N}(0,\overline{C})}\Big] \tag{48}$$

with indices $a,b\in\{0,\dots,A+1\}$. The saddle point $\overline{\tilde{C}}^{(a,b)}_{\alpha\beta} = 0$ is a self-consistent solution because $W_{\text{aux}}(0\,|\,C)\equiv 0$, which is in particular independent of C, so that $\partial W_{\text{aux}}(0\,|\,C)/\partial C^{(a,b)}_{\alpha,\beta}\equiv 0$.

To evaluate the expectation value on the r.h.s. of Eq. (48), we only need the subtensors of $\overline{C}$ formed by the indices that explicitly appear in the expectation due to the marginalization property of the Gaussian. In particular, this means the saddle point equations can be solved iteratively starting from $a = 0$, which requires the starting values $\overline{C}^{(0,a)}_{\alpha\beta} = \overline{C}^{(a,0)}_{\alpha\beta} = M_{a,0}\left[\sigma_0^2 + \frac{g_0^2}{n_{\text{in}}} \sum_{j=1}^{n_{\text{in}}} x^{(a)}_{j,\alpha} x^{(0)}_{j,\beta}\right]$ for the recursion.

After the saddle-point approximation, the conditional probability Eq. (43) simplifies to the factorized Gaussian

$$p(\boldsymbol{Y} \,|\, \boldsymbol{X}) = \prod_{i=1}^{n_{A+1}} p(\boldsymbol{y}_i \,|\, \boldsymbol{X}) \,,$$
$$p(\boldsymbol{y}_i \,|\, \boldsymbol{X}) = \mathcal{N}(0, \overline{C}^{(A+1,A+1)}) \,,$$

with covariance matrix $\langle y_{i,\alpha} y_{i,\beta} \rangle = \overline{C}^{(A+1,A+1)}_{\alpha,\beta}$ across inputs α, β that is determined recursively by Eq. (48), starting from the input covariance $\overline{C}^{(0,0)}_{\alpha\beta} = \sigma_0^2 + \frac{g_0^2}{n_{\text{in}}} \sum_{j=1}^{n_{\text{in}}} x^{(0)}_{j,\alpha} x^{(0)}_{j,\beta}$. The diagonal elements $\overline{C}^{(A+1,A+1)}_{\alpha,\beta}$ thus only depends on the equal-time overlaps $\sum_{i=1}^{n_{\text{in}}} x^{(a)}_{i,\alpha} x^{(a)}_{i,\beta}$ of the inputs with $0 \leq a \leq A$.

6.2. Finite-size instability of RNNs

In the main text Figure 1, the mean-field theory is compared to network simulations with hidden layer width $n = 2000$ and a ReLU nonlinearity for fixed hyperparameters $g^2 = 1.6$, $\sigma^2 = 0.2$. Although this appears to be quite wide already, for the RNN the statistics of the activity in individual networks strongly varies across realizations of weights. The frequency of deviating realizations increases as one approaches $g^2 \to 2$, the instability threshold above which $\overline{C}^{(a,a)}$ diverges with growing a for the ReLU nonlinearity. The instability threshold can be obtained from the MFT solution for a single replicon and $a = b$, Eq. (50): The theory predicts that $g^2 > 2$ will lead to exponential increase of the activity, while $g^2 < 2$ results in finite (but possibly very strong) activity. Beyond this threshold, trajectories of individual neurons diverge towards $\pm\infty$ over time. At finite width and $g^2 < 2$, the activity is typically stable. But for g^2 sufficiently close to 2, the closeness of the instability point is visible in the system. This is observable as a spread of individual neurons' trajectories, each hovering about a non-zero set point. This observation corresponds to a static contribution (independent of Δa) to the time-lagged correlation function, as shown in Figure 3b. The reason for this instability to only occur in the RNN is the coherent interplay of the activity with the connectivity across time: Since the connectivity is identical across all time steps, fluctuations of the activity can be amplified coherently across multiple time steps. Likewise, deviations of the variances $C^{(a,a)}$ are observable in this case (Figure 3a). The effect is suppressed as the network size increases; the mean-field theory then becomes accurate also for values of g^2 close to 2 (Figure 3a,b).

6.3. Details about numerical experiments

For all experiments, we used NumPy [43] and SciPy [44] which are both released under a BSD-3-Clause License. Computations were performed on a CPU cluster. More precisely, the requirements for the experiments are:

- Figure 1 (main): 1h on a single core laptop.

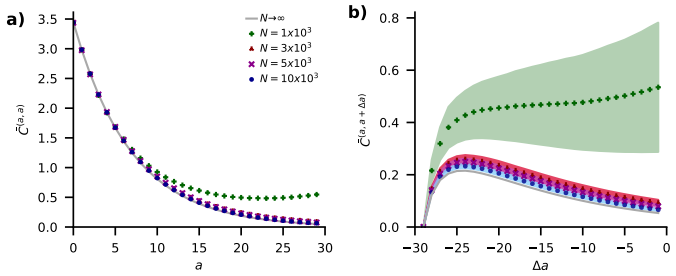

Figure 3. Mean-field theory of the RNN compared to simulation. a Average variance in mean-field theory $\overline{C}^{(a,a)}$ (solid gray curve) and estimate $\frac{1}{n}\sum_i h_i^{(a)} h_i^{(a)}$ from simulation, averaged over 100 realizations of networks with different widths (symbols, see legend). **b** Average cross-covariance in mean-field-theory $\overline{C}^{(a,a+\Delta a)}$ and estimate $\frac{1}{n}\sum_i h_i^{(a)} h_i^{(a+\Delta a)}$ from simulation, averaged over 100 network realizations (mean shown as symbols, same symbol code as in panel a; standard error of the mean shown as tube), as a function of the temporal distance Δa to the hidden layer $a = 30$. Mean-field theory (gray curve). Other parameters: $g^2 = 1.73$, $\sigma^2 = 0$, layer widths $n_a \in \{1, 3, 5, 10\} \cdot 10^3$, $A = 30$ hidden layers, ReLU activation $\phi(x) = \max(0, x)$ and Gaussian input $\boldsymbol{x} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(1, 1)$ with $\dim(\boldsymbol{x}) = 10^5$.

- Figure 2 (main): 50h on a single node with 24 cores of a CPU cluster for each of panel a and c, and 2h on a single node with 24 cores of the CPU cluster for each of panel b and d
- Figure 3 (appendix): 1.5h on a single core laptop.

The code used to produce the figures is stored in a Zenodo archive with the DOI 10.5281/zenodo.5747219.

To solve the mean field theory for a given activation function $\phi(x)$ one needs to calculate the expectation values in Eq. (31), or more general in Eq. (33). These expectation values can be computed analytically for the ReLU activation $\phi(x) = \max(0, x)$ as shown in [45]:

$$\langle \phi(x)\phi(y) \rangle_{x,\, y \sim \mathcal{N}(0,C)} = \frac{1}{2\pi} \nu (\sin\theta + (\pi - \theta)\cos\theta) \,,$$

where

$$\nu = \sqrt{C_{xx} C_{yy}} \,,$$
$$\theta = \cos^{-1}\left(\frac{C_{xy}}{\nu}\right).$$

Inserting this into the MFT equations for multiple replicas (33) results in

$$\overline{C}_{\alpha\beta}^{(a,b)} = M_{a,b}\left[\sigma_a^2 + \mathbb{1}_{a\geq 1, b\geq 1} \frac{g_a^2}{2\pi} \nu_{\alpha\beta}^{(a,b)} (\sin\theta_{\alpha\beta}^{(a,b)} + (\pi - \theta_{\alpha\beta}^{(a,b)})\cos\theta_{\alpha\beta}^{(a,b)}) + \mathbb{1}_{a\leq A, b\leq A} \frac{g_0^2}{n_{\text{in}}} \boldsymbol{x}_\alpha^{(a)\mathrm{T}} \boldsymbol{x}_\beta^{(b)}\right], \tag{49}$$

where

$$\nu_{\alpha\beta}^{(a,b)} = \sqrt{\overline{C}_{\alpha\alpha}^{(a-1,a-1)} \overline{C}_{\beta\beta}^{(b-1,b-1)}} \,,$$

$$\theta_{\alpha\beta}^{(a,b)} = \cos^{-1}\left(\frac{\overline{C}_{\alpha\beta}^{(a-1,b-1)}}{\nu_{\alpha\beta}^{(a,b)}}\right).$$

The special case $a = b$, $\alpha = \beta$, and vanishing external input yields

$$\overline{C}^{(a,a)} = \sigma_a^2 + \frac{g_a^2}{2}\overline{C}^{(a-1,a-1)}. \tag{50}$$

For time or layer independent $g_a \equiv g$ and $\sigma_a \equiv \sigma$, the activity thus increases exponentially in time or over layers for $g^2 > 2$ and converges towards an equilibrium value $C^{(\infty)} = \frac{\sigma^2}{1-g^2/2}$ for $g^2 < 2$.

References

[1] Hinton G E, Osindero S and Teh Y W 2006 *Neural computation* **18** 1527–1554

[2] Krizhevsky A, Sutskever I and Hinton G E 2012 Imagenet classification with deep convolutional neural networks *Advances in Neural Information Processing Systems* pp 1097–1105

[3] Hannun A, Case C, Casper J, Catanzaro B, Diamos G, Elsen E, Prenger R, Satheesh S, Sengupta S, Coates A *et al.* 2014 *arXiv:1412.5567*

[4] LeCun Y, Bengio Y and Hinton G 2015 *Nature* **521** 436–444

[5] Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I and Fergus R 2013 *arXiv:1312.6199*

[6] Neal R M 1996 *Bayesian Learning for Neural Networks* (Springer New York) URL https://doi.org/10.1007/978-1-4612-0745-0

[7] Williams C K I and Barber D 1998 *IEEE Transactions on Pattern Analysis and Machine Intelligence* **12**

[8] Lee J, Sohl-Dickstein J, Pennington J, Novak R, Schoenholz S and Bahri Y 2018 Deep neural networks as gaussian processes *International Conference on Learning Representations* URL https://openreview.net/forum?id=B1EA-M-0Z

[9] Matthews A G d G, Hron J, Rowland M, Turner R E and Ghahramani Z 2018 Gaussian process behaviour in wide deep neural networks *International Conference on Learning Representations* URL https://openreview.net/forum?id=H1-nGgWC-

[10] Rumelhart David E, Hinton Geoffrey E and Williams Ronald J 1986 *Nature* **323** 533–536 URL https://doi.org/10.1038/323533a0

[11] Pearlmutter B 1989 *Neural Comput.* **1** 263–269

[12] Yang G 2019 Wide feedforward or recurrent neural networks of any architecture are gaussian processes *Advances in Neural Information Processing Systems* vol 32

[13] Yaida S 2019 *arXiv:1910.00019*

[14] Dyer E and Gur-Ari G 2019 *arXiv:1909.11304*

[15] Antognini J M 2019 *arXiv:1908.10030*

[16] Huang J and Yau H T 2019 *arXiv:1909.08156*

[17] Halverson J, Maiti A and Stoner K 2020 *arXiv:2008.08601*

[18] Naveh G, Ben-David O, Sompolinsky H and Ringel Z 2020 *arXiv:2004.01190*

[19] Williams C K and Rasmussen C E 2006 *Gaussian Processes for Machine Learning* 1st ed (Cambridge: MIT Press)

[20] Sompolinsky H, Crisanti A and Sommers H J 1988 *Phys. Rev. Lett.* **61**(3) 259–262 URL http://link.aps.org/doi/10.1103/PhysRevLett.61.259

[21] Chow C and Buice M 2015 *J Math. Neurosci* **5** 8

[22] Hertz J A, Roudi Y and Sollich P 2017 *Journal of Physics A: Mathematical and Theoretical* **50** 033001 URL http://stacks.iop.org/1751-8121/50/i=3/a=033001

[23] Martí D, Brunel N and Ostojic S 2018 *Phys. Rev. E* **97**(6) 062314 URL https://link.aps.org/doi/10.1103/PhysRevE.97.062314

[24] Crisanti A and Sompolinsky H 2018 *Phys. Rev. E* **98**(6) 062120 URL https://link.aps.org/doi/10.1103/PhysRevE.98.062120

[25] Schuecker J, Goedeke S and Helias M 2018 *Phys. Rev. X* **8**(4) 041029 URL https://link.aps.org/doi/10.1103/PhysRevX.8.041029

[26] Parisi G 1980 *Journal of Physics A: Mathematical and General* **13** 1101

[27] Sommers H 1987 *Phys. Rev. Lett.* **58** 1268–1271

[28] Fischer K and Hertz J 1991 *Spin glasses* (Cambridge University Press)

[29] Mozeika A, Li B and Saad D 2020 *Phys. Rev. Lett.* **125**(16) 168301 URL https://link.aps.org/doi/10.1103/PhysRevLett.125.168301

[30] Molgedey L, Schuchhardt J and Schuster H 1992 *Phys. Rev. Lett.* **69** 3717

[31] Zinn-Justin J 1996 *Quantum field theory and critical phenomena* (Clarendon Press, Oxford)

[32] Moshe M and Zinn-Justin J 2003 *Physics Reports* **385** 69–228 ISSN 0370-1573 URL http://www.sciencedirect.com/science/article/pii/S0370157303002631

[33] MacKay D J 2003 *Information theory, inference and learning algorithms* (Cambridge university press)

[34] Hertz J, Krogh A and Palmer R G 1991 *Introduction to the Theory of Neural Computation* (Perseus Books) ISBN 0-201-51560-1

[35] Sompolinsky H and Zippelius A 1981 *Phys. Rev. Lett.* **47**(5) 359–362 URL http://link.aps.org/doi/10.1103/PhysRevLett.47.359

[36] Helias M and Dahmen D 2020 *Statistical Field Theory for Neural Networks* vol 970 (Springer International Publishing)

[37] Coolen A C C 2000 *arXiv:cond-mat/0006011*

[38] Gretton A, Borgwardt K M, Rasch M J, Schölkopf B and Smola A 2012 *Journal of Machine Learning Research* **13** 723–773 URL http://jmlr.org/papers/v13/gretton12a.html

[39] Chen M, Pennington J and Schoenholz S 2018 Dynamical isometry and a mean field theory of RNNs: Gating enables signal propagation in recurrent neural networks *Proceedings of the 35th International Conference on Machine Learning* (*Proceedings of Machine Learning Research* vol 80) ed Dy J and Krause A (PMLR) pp 873–882 URL http://proceedings.mlr.press/v80/chen18i.html

[40] van Meegen A, Kühn T and Helias M 2021 *Phys. Rev. Lett.* **127**(15) 158302 URL https://link.aps.org/doi/10.1103/PhysRevLett.127.158302

[41] Roberts D A, Yaida S and Hanin B 2021 *arXiv* 2106.10165 [cs.LG]

[42] Zavatone-Veth J A and Pehlevan C 2021 *arXiv:2104.11734*

[43] Harris C R, Millman K J, van der Walt S J, Gommers R, Virtanen P, Cournapeau D, Wieser E, Taylor J, Berg S, Smith N J, Kern R, Picus M, Hoyer S, van Kerkwijk M H, Brett M, Haldane A, Fernández del Río J, Wiebe M, Peterson P, Gérard-Marchant P, Sheppard K, Reddy T, Weckesser W, Abbasi H, Gohlke C and Oliphant T E 2020 *Nature* **585** 357–362

[44] Virtanen P, Gommers R, Oliphant T E, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, Bright J, van der Walt S J, Brett M, Wilson J, Millman K J, Mayorov N, Nelson A R J, Jones E, Kern R, Larson E, Carey C J, Polat İ, Feng Y, Moore E W, VanderPlas J, Laxalde D, Perktold J, Cimrman R, Henriksen I, Quintero E A, Harris C R, Archibald A M, Ribeiro A H, Pedregosa F, van Mulbregt P and SciPy 10 Contributors 2020 *Nat. Methods* **17** 261–272

[45] Cho Y and Saul L 2009 Kernel methods for deep learning *Advances in Neural Information Processing Systems* vol 22 ed Bengio Y, Schuurmans D, Lafferty J, Williams C and Culotta A (Curran Associates, Inc.) URL https://proceedings.neurips.cc/paper/2009/file/5751ec3e9a4feab575962e78e006250d-Paper.pdf

5

LARGE-DEVIATION APPROACH TO RANDOM RECURRENT NEURONAL NETWORKS: PARAMETER INFERENCE AND FLUCTUATION-INDUCED TRANSITIONS

PREAMBLE

The previous chapter dealt with network models with a discrete time evolution. In this chapter, we take the step to continuous time. We furthermore allow for a block-structure in the connectivity to account, for example, for different populations of neurons. Crucially, the connectivity of the network is still assumed to be random. In contrast to the last chapter, however, the focus shifts mostly to the dynamics.

Random networks with independent weights exhibit a striking property: they are self-averaging (Helias and Dahmen 2020). Put differently, many relevant observables depend only very weakly on the realization of the connectivity. The canonical approach to capture such self-averaging observables analytically is dynamic mean-field theory (DMFT, see Section 3.4).

DMFT was initially formulated by Sompolinsky and Zippelius (1982) and Sompolinsky, Crisanti, and Sommers (1988) in the language of field theory (for a comprehensive introduction see Crisanti and Sompolinsky 2018; Helias and Dahmen 2020). Later, it was subject to rigorous mathematical investigations by Arous and Guionnet (1995) using the framework of large deviation theory (see Section 3.1). However, due to the different approaches, it was not clear if the results of Sompolinsky, Crisanti, and Sommers (1988) and Arous and Guionnet (1995) are consistent.

To address this problem, we reformulate the approach of Arous and Guionnet (1995) in this chapter in a field theoretical language. A crucial step by Arous and Guionnet (1995) was to introduce the empirical measure on the space of trajectories and to consider its distribution across the ensemble of connectivities. We calculate this distribution using field theory, thereby generalizing the result by Arous and Guionnet (1995).

Self-averaging follows immediately from the distribution of the empirical measure because it is sharply peaked, hence all resulting observables attain a value close to the most likely one. Additionally, we use the distribution to determined beyond-mean-field fluctuations of the order parameter. Last, knowledge of the distribution of the em-

pirical measure also allows to address the inverse problem: inferring network statistics from observed trajectories.

Author Contributions

All calculations were performed, the codebase written, and the figures created by the author (AvM) under supervision of Prof. Moritz Helias (MH). The initial concept of comparing dynamical mean-field theory and large deviation theory is due to Dr. Tobias Kühn (TK) and MH; the unification as well as all resulting applications are due to AvM and MH. The first draft of the manuscript was written by AvM and it was jointly revised by AvM, TK, and MH.

Large-Deviation Approach to Random Recurrent Neuronal Networks: Parameter Inference and Fluctuation-Induced Transitions

Alexander van Meegen[1,2,*] Tobias Kühn[1,3,4] and Moritz Helias[1,3]
[1]Institute of Neuroscience and Medicine (INM-6) and Institute for Advanced Simulation (IAS-6) and JARA-Institute Brain Structure-Function Relationships (INM-10), Jülich Research Centre, 52428 Jülich, Germany
[2]Institute of Zoology, University of Cologne, 50674 Cologne, Germany
[3]Department of Physics, Faculty 1, RWTH Aachen University, 52074 Aachen, Germany
[4]Laboratoire de Physique de l'Ecole Normale Supérieure, ENS, Université PSL, CNRS, Sorbonne Université, Université de Paris, F-75005 Paris, France

(Received 21 September 2020; revised 5 July 2021; accepted 19 August 2021; published 7 October 2021)

We here unify the field-theoretical approach to neuronal networks with large deviations theory. For a prototypical random recurrent network model with continuous-valued units, we show that the effective action is identical to the rate function and derive the latter using field theory. This rate function takes the form of a Kullback-Leibler divergence which enables data-driven inference of model parameters and calculation of fluctuations beyond mean-field theory. Lastly, we expose a regime with fluctuation-induced transitions between mean-field solutions.

DOI: 10.1103/PhysRevLett.127.158302

Introduction.—Biological neuronal networks are systems with many degrees of freedom and intriguing properties: their units are coupled in a directed, nonsymmetric manner, so that they typically operate outside thermodynamic equilibrium [1,2]. The primary analytical method to study neuronal networks has been mean-field theory [3–8]. Its field-theoretical basis has been exposed only recently [9,10]. However, to understand the parallel and distributed information processing performed by neuronal networks, the study of the forward problem—from the microscopic parameters of the model to its dynamics—is not sufficient. One additionally faces the inverse problem of determining the parameters of the model given a desired dynamics and thus function. Formally, one needs to link statistical physics with concepts from information theory and statistical inference.

We here expose a tight relation between statistical field theory of neuronal networks, large deviations theory, information theory, and inference. To this end, we generalize the probabilistic view of large deviations theory, which yields rigorous results for the leading-order behavior in the network size N [11,12], to arbitrary single unit dynamics, transfer functions, and multiple populations. We furthermore show that the central quantity of large deviations theory, the rate function, is identical to the effective action in statistical field theory. This link exposes a second relation: Bayesian inference and prediction are naturally formulated within this framework, spanning the arc to information processing. Concretely, we develop a method for parameter inference from transient data for single- and multi-population networks. Lastly, we overcome the inherent limit of mean-field theory—its neglect of fluctuations. We develop a theory for fluctuations of the order parameter when the intrinsic timescale is large and discover a regime with fluctuation-induced transitions between two coexisting mean-field solutions.

First, we introduce the model in its most general form. Then, we develop the theory for a single population. Last, we generalize it to multiple populations.

Model.—We consider block-structured random networks of $N = \sum_\alpha N_\alpha$ nonlinearly interacting units $x_i^\alpha(t)$ driven by an external input $\xi_i^\alpha(t)$. The dynamics of the ith unit in the αth population is governed by the stochastic differential equation

$$\tau_\alpha \dot{x}_i^\alpha(t) = -U'_\alpha\left(x_i^\alpha(t)\right) + \sum_\beta \sum_{j=1}^{N_\beta} J_{ij}^{\alpha\beta} \phi\left(x_j^\beta(t)\right) + \xi_i^\alpha(t). \quad (1)$$

In the absence of recurrent and external inputs, the units undergo an overdamped motion with time constant τ_α in a potential $U_\alpha(x)$. The $J_{ij}^{\alpha\beta}$ are independent and identically Gaussian-distributed random coupling weights with zero mean and population-specific variance $\langle (J_{ij}^{\alpha\beta})^2 \rangle = g_{\alpha\beta}^2/N_\beta$ where the coupling strength $g_{\alpha\beta}$ controls the heterogeneity of the weights. The time-varying external inputs $\xi_i^\alpha(t)$ are independent Gaussian white-noise processes with

zero mean and correlation functions $\langle \xi_i^\alpha(t_1)\xi_j^\beta(t_2)\rangle = 2D_\alpha\delta_{ij}\delta_{\alpha\beta}\delta(t_1-t_2)$. The single-population model corresponds to the one studied in Ref. [4] if the external input vanishes, $D=0$, the potential is quadratic, $U(x)=\frac{1}{2}x^2$, and the transfer function is sigmoidal, $\phi(x)=\tanh(x)$; for $D=\frac{1}{2}$, $U(x)=-\log(A^2-x^2)$, and $\phi(x)=x$ it corresponds to the one in Ref. [11], which is inspired by the dynamical spin glass model of Ref. [13].

Field theory.—The field-theoretical treatment of Eq. (1) employs the Martin-Siggia-Rose-de Dominicis-Janssen path integral formalism [14–17]. We denote the expectation over paths across different realizations of the noise ξ as [[18], Section A.1]

$$\langle\cdot\rangle_{\boldsymbol{x}|\boldsymbol{J}} \equiv \langle\langle\cdot\rangle_{\boldsymbol{x}|\boldsymbol{J},\boldsymbol{\xi}}\rangle_{\boldsymbol{\xi}} = \int \mathcal{D}\boldsymbol{x} \int \mathcal{D}\tilde{\boldsymbol{x}}\cdot e^{S_0(\boldsymbol{x},\tilde{\boldsymbol{x}})-\tilde{\boldsymbol{x}}^\mathsf{T}\boldsymbol{J}\boldsymbol{\phi}(\boldsymbol{x})},$$

where $\langle\cdot\rangle_{\boldsymbol{x}|\boldsymbol{J},\boldsymbol{\xi}}$ integrates over the unique solution of Eq. (1) given one realization $\boldsymbol{\xi}$ of the noise. Here, $S_0(\boldsymbol{x},\tilde{\boldsymbol{x}}) = \tilde{\boldsymbol{x}}^\mathsf{T}[\dot{\boldsymbol{x}}+U'(\boldsymbol{x})] + D\tilde{\boldsymbol{x}}^\mathsf{T}\tilde{\boldsymbol{x}}$ is the action of the uncoupled neurons. We use the shorthand notation $\boldsymbol{a}^\mathsf{T}\boldsymbol{b} = \sum_{i=1}^N \int_0^T dt a_i(t)b_i(t)$.

For large N, the system becomes self-averaging, a property known from many disordered systems with large numbers of degrees of freedom: the collective behavior is stereotypical, independent of the realization J_{ij}. A self-averaging observable has a sharply peaked distribution over realizations of $\boldsymbol{J}$—the observable always attains the same value, close to its average. This, however, only holds for observables averaged over all units, reminiscent of the central limit theorem. These are generally of the form $\sum_{i=1}^N \ell(x_i)$, where ℓ is an arbitrary functional of a single unit's trajectory. It is therefore convenient to introduce the scaled cumulant-generating functional

$$W_N(\ell) := \frac{1}{N}\ln\langle\langle e^{\sum_{i=1}^N \ell(x_i)}\rangle_{\boldsymbol{x}|\boldsymbol{J}}\rangle_{\boldsymbol{J}}, \tag{2}$$

where the prefactor $1/N$ makes sure that W_N is an intensive quantity, reminiscent of the bulk free energy [24]. In fact, we will show that the N dependence vanishes in the limit $N\to\infty$ because the system decouples.

Performing the average over $\boldsymbol{J}$, i.e., evaluating $\langle e^{-\tilde{\boldsymbol{x}}^\mathsf{T}\boldsymbol{J}\boldsymbol{\phi}(\boldsymbol{x})}\rangle_{\boldsymbol{J}}$, and introducing the auxiliary field

$$C(t_1,t_2) := \frac{1}{N}\sum_{i=1}^N \phi(x_i(t_1))\phi(x_i(t_2)) \tag{3}$$

as well as the conjugate field $\tilde{C}$, we can write W_N as [[18], Section A.1]

$$W_N(\ell) = \frac{1}{N}\ln\int\mathcal{D}C\int\mathcal{D}\tilde{C}e^{-NC^\mathsf{T}\tilde{C}+N\Omega_\ell(C,\tilde{C})},$$
$$\Omega_\ell(C,\tilde{C}) := \ln\int\mathcal{D}x\int\mathcal{D}\tilde{x}e^{S_0(x,\tilde{x})+\frac{\sigma^2}{2}\tilde{x}^\mathsf{T}C\tilde{x}+\phi^\mathsf{T}\tilde{C}\phi+\ell(x)}. \tag{4}$$

The effective action is defined as the Legendre transform of $W_N(\ell)$,

$$\Gamma_N(\mu) := \int\mathcal{D}x\mu(x)\ell_\mu(x) - W_N(\ell_\mu), \tag{5}$$

where ℓ_μ is determined implicitly by the condition $\mu = W_N'(\ell_\mu)$ and the derivative $W_N'(\ell)$ has to be understood as a generalized derivative, the coefficient of the linearization akin to a Fréchet derivative [25].

Note that W_N and Γ_N are, respectively, generalizations of a cumulant-generating functional and of the effective action [26] because both map a functional (ℓ or μ) to the reals. For the choice $\ell(x)=j^\mathsf{T}x$, where $j(t)$ is an arbitrary function, we recover the usual cumulant-generating functional of the single unit's trajectory [[18], Section A.4] and the corresponding effective action.

Rate function.—Any network-averaged observable, for which we may expect self-averaging to hold, can likewise be obtained from the empirical measure

$$\mu(y) := \frac{1}{N}\sum_{i=1}^N \delta(x_i-y), \tag{6}$$

since $(1/N)\sum_{i=1}^N \ell(x_i) = \int\mathcal{D}y\mu(y)\ell(y)$. Of particular interest is the leading-order exponential behavior of the distribution of empirical measures $P(\mu) = \langle\langle P(\mu|\boldsymbol{x})\rangle_{\boldsymbol{x}|\boldsymbol{J}}\rangle_{\boldsymbol{J}}$ across realizations of $\boldsymbol{J}$ and $\boldsymbol{\xi}$. This behavior in the large N limit is described by what is known as the rate function

$$H(\mu) := -\lim_{N\to\infty}\frac{1}{N}\ln P(\mu) \tag{7}$$

in large deviations theory [see, e.g., [27]]; $H(\mu)$ captures the leading exponential probability $P(\mu)\overset{N\gg1}{\simeq}\exp[-NH(\mu)]$. For large N, the probability of an empirical measure that does not correspond to the minimum $H'(\bar{\mu})=0$ is thus exponentially suppressed. Put differently, the system is self-averaging and the statistics of any network-averaged observable can be obtained using $\bar{\mu}$.

Similar as in field theory, it is convenient to introduce the scaled cumulant-generating functional of the empirical measure. Because $(1/N)\sum_{i=1}^N \ell(x_i) = \int\mathcal{D}y\mu(y)\ell(y)$ holds for an arbitrary functional $\ell(x_i)$ of the single unit's trajectory x_i, Eq. (2) has the form of the scaled cumulant-generating functional for μ at finite N.

Using a saddle-point approximation for the integrals over C and $\tilde{C}$ in Eq. (4) [[18], Section A.1], we get

$$W_\infty(\ell) = -C_\ell^\mathsf{T}\tilde{C}_\ell + \Omega_\ell(C_\ell,\tilde{C}_\ell). \tag{8}$$

Both C_ℓ and $\tilde{C}_\ell$ are determined self-consistently by the saddle-point equations $C_\ell = \partial_{\tilde{C}}\Omega_\ell(C,\tilde{C})|_{C_\ell,\tilde{C}_\ell}$ and

$\tilde{C}_\ell = \partial_C \Omega_\ell(C, \tilde{C})|_{C_\ell, \tilde{C}_\ell}$ where ∂_C denotes a partial functional derivative.

From the scaled cumulant-generating functional, Eq. (8), we obtain the rate function via a Legendre transformation [28]: $H(\mu) = \int \mathcal{D}x \mu(x) \ell_\mu(x) - W_\infty(\ell)$ with ℓ_μ implicitly defined by $\mu = W'_\infty(\ell_\mu)$. Note that $H(\mu)$ is still convex even if μ itself is multimodal. Comparing with Eq. (5), we observe that the rate function is equivalent to the effective action: $H(\mu) = \lim_{N\to\infty} \Gamma_N(\mu)$. The equation $\mu = W'_\infty(\ell_\mu)$ can be solved for ℓ_μ to obtain a closed expression for the rate function viz. effective action [[18], Section A.2], one main result of our work,

$$H(\mu) = \int \mathcal{D}x \mu(x) \ln \frac{\mu(x)}{\langle \delta(\dot{x} + U'(x) - \eta) \rangle_\eta}, \tag{9}$$

where η is a zero–mean Gaussian process with a correlation function that is determined by $\mu(x)$,

$$C_\eta(t_1, t_2) = 2D\delta(t_1 - t_2) + g^2 \int \mathcal{D}x \mu(x) \phi(x(t_1)) \phi(x(t_2)). \tag{10}$$

For $D = \frac{1}{2}$, $U(x) = -\log(A^2 - x^2)$, and $\phi(x) = x$, Eq. (9) can be shown to be equivalent to the mathematically rigorous result obtained in the seminal work by Ben Arous and Guionnet [[18], Section A.3].

The rate function Eq. (9) takes the form of a Kullback-Leibler divergence. Thus, it possesses a minimum at

$$\bar{\mu}(x) = \langle \delta(\dot{x} + U'(x) - \eta) \rangle_\eta. \tag{11}$$

This most likely measure corresponds to the well-known self-consistent stochastic dynamics that is obtained in field theory [4,9,10,29]. Note that the correlation function of the effective stochastic input η at the minimum depends self-consistently on $\bar{\mu}(x)$ through Eq. (10). However, the rate function $H(\mu)$ contains more information. It quantifies the suppression of departures $\mu - \bar{\mu}$ from the most likely measure and therefore allows the assessment of fluctuations that are beyond the scope of the classical mean-field result.

Parameter inference.—The rate function opens the way to address the inverse problem: given the network–averaged activity statistics, encoded in the corresponding empirical measure μ, what are the statistics of the connectivity and the external input, i.e., g and D?

We determine the parameters using maximum likelihood estimation. Using Eq. (7) and Eq. (9), the likelihood of the parameters is given by

$$\ln P(\mu|g, D) \simeq -NH(\mu|g, D),$$

where $\simeq$ denotes equality in the limit $N \to \infty$ and we made the dependence on g and D explicit. The maximum likelihood estimate of the parameters g and D corresponds to the minimum of the Kullback-Leibler divergence H, Eq. (9), on the right-hand side. Evaluating the derivative of $H(\mu|g, D)$ yields [[18], Section B.1]

$$\partial_a \ln P(\mu|g, D) \simeq -\frac{N}{2} \mathrm{tr}\left((C_0 - C_\eta) \frac{\partial C_\eta^{-1}}{\partial a} \right),$$

where we abbreviated $a \in \{g, D\}$ and defined $C_0(t_1, t_2) \equiv \int \mathcal{D}x \mu(x) (\dot{x}(t_1) + U'(x(t_1)))(\dot{x}(t_2) + U'(x(t_2)))$. The derivative vanishes for $C_0 = C_\eta$. Assuming stationarity, in the Fourier domain this condition reads

$$\mathcal{S}_{\dot{x}+U'(x)}(f) = 2D + g^2 \mathcal{S}_{\phi(x)}(f), \tag{12}$$

where $\mathcal{S}_X(f)$ denotes the network-averaged power spectrum of the observable X. Using non-negative least squares [30], Eq. (12) allows a straightforward inference of g and D (Fig. 1). To determine the transfer function ϕ and the potential U, one can use model comparison techniques [[18], Section B.2]. Using the inferred parameters, we can also predict the future activity of a unit from the knowledge of its recent past [[18], Section B.3].

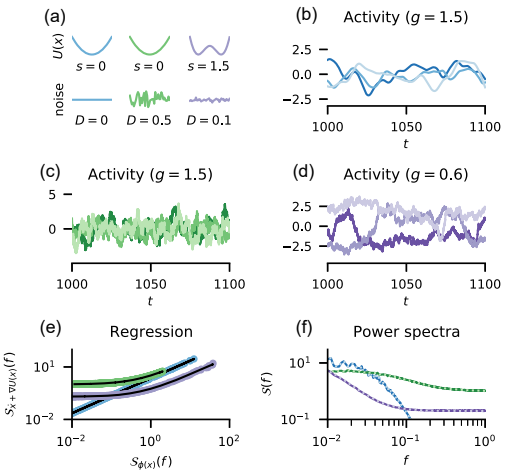

FIG. 1. Maximum likelihood parameter estimation for $\phi(x) = \mathrm{erf}(\sqrt{\pi}x/2)$, potential $U(x) = \frac{1}{2}x^2 + s \ln\cosh x$, and external noise D. (a) Color-coded sketch of potential and noise. (b)–(d) Activity of three randomly chosen units for coupling strengths g indicated in title. (e) Parameter estimation via non-negative least squares regression (black lines) based on Eq. (12). (f) Power spectra on the left- (dark, solid curves) and right-hand sides (light, dotted curves) of Eq. (12) for the inferred parameters. Further parameters: $\tau = 1$, $N = 10\,000$, temporal discretization $dt = 10^{-2}$, simulation time $T = 1000$, time span discarded to reach steady state $T_0 = 100$.

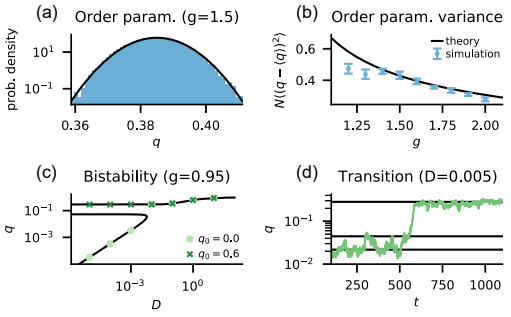

FIG. 2. Order parameter fluctuations for $\phi(x) = \mathrm{erf}(\sqrt{\pi}x/2)$ [(a),(b)] and metastability for $\phi(x) = \mathrm{clip}[\tan(x), -1, 1]$ [(c),(d)]. (a) Temporal order parameter statistics across ten simulations (bars) and theory (solid curve) from Eq. (13). (b) Order parameter variance for 10 realizations of the connectivity with standard error of the mean (symbols) and theory (solid curve) from Eq. (13). (c) Mean order parameter for different initial values q_0 from simulations (symbols) and self-consistent theory (solid curves). (d) Fluctuation-induced bistability of the order parameter for $N = 750$, $g = 0.95$. Further parameters: $T = 5000$ in (a),(d); $U(x) = \frac{1}{2}x^2$ in (a)–(d); other parameters as in Fig. 1.

Fluctuations.—The rate function allows us to go beyond mean-field theory and examine fluctuations of the order parameter. Here, we use the network-averaged variance $q(t) = C(t,t)$ from Eq. (3) as an order parameter and restrict the discussion to the case $U(x) = \frac{1}{2}x^2$.

Figure 2(a) shows the distribution of $q(t)$ across time and across realizations of the connectivity. The fluctuations across realizations of the connectivity can be computed from the curvature of the rate function $I(C)$ that is obtained from (9) by the contraction principle [[18], Section C.1]. In a stationary state and considering only the fluctuations across realizations of the connectivity, for slow recurrent dynamics $\tau_c \gg 1$ we obtain the approximation for the fluctuations of q

$$\langle (q - \langle q \rangle_J)^2 \rangle_J = \frac{\langle (\phi\phi - \langle \phi\phi \rangle_0)^2 \rangle_0}{N[1 - g^2(\langle \phi'' \phi \rangle_0 + \langle \phi' \phi' \rangle_0)]^2}. \quad (13)$$

Here, $\langle fg \rangle_0 \equiv \langle f(x(t))g(x(t)) \rangle_0$ denotes an expectation with respect to the self-consistent measure (11). For vanishing noise, $D = 0$, and $g > 1$, the dynamics are slow and the theory matches the empirical fluctuations very well [Figs. 2(a) and 2(b)]. Deviations in Fig. 2(b) are caused by two effects: For $g \searrow 1$, periodic solutions appear as a finite-size effect; for growing g, the timescale τ_c decreases, eventually violating the assumption $\tau_c \gg 1$ entering Eq. (13). Rate functions like $I(C)$ in general also allow one to estimate the tail probability $\mathbb{P}(q > \theta) \approx \exp[-NI(\theta)]$, which here shows a quadratic decline for large departures [Fig. 2(a)].

When the denominator in Eq. (13) vanishes, fluctuations grow large, indicative of a continuous phase transition. For $\phi'''(0) < 0$ the denominator vanishes for $g \geq 1$ [Fig. 2(b)], in line with the established theory, the breakdown of linear stability of the fixed point $x = 0$ [4]. For $\phi'''(0) > 0$, however, Eq. (13) predicts qualitatively different behavior: the denominator vanishes at $g < 1$, in the linearly stable regime. In fact, we find that this regime features the coexistence of two stable mean-field solutions (Fig. 2(c), [[18], Section C.2]) and fluctuation-driven first-order transitions between them [Fig. 2(d)]. The solution with larger q corresponds to self-sustained activity; the solution with smaller q corresponds to the fixed point $x = 0$ and is stable [[18], Section C.2], in contrast to the case of a threshold-power-law transfer function [6].

Multiple populations.—For multiple populations, any population-averaged observable can be obtained from the empirical measure $\mu^\alpha(y) = (1/N_\alpha)\sum_{i=1}^{N_\alpha} \delta(x_i^\alpha - y)$. The joint distribution of all population-specific empirical measures $\{\mu^\circ\}$ is determined by the rate function [[18], Section D]

$$H(\{\mu^\circ\}) = \sum_\alpha \gamma_\alpha \int \mathcal{D}x \mu^\alpha(x) \ln \frac{\mu^\alpha(x)}{\langle \delta(\tau_\alpha \dot{x} + U'_\alpha(x) - \eta_\alpha) \rangle_{\eta_\alpha}}, \quad (14)$$

where $\gamma_\alpha = N_\alpha/N$ and η_α is a zero-mean Gaussian process with

$$C_\eta^\alpha(t_1, t_2) = 2D_\alpha \delta(t_1 - t_2) + \sum_\beta g_{\alpha\beta}^2 \int \mathcal{D}x \mu^\beta(x) \phi(x(t_1)) \phi(x(t_2)). \quad (15)$$

Again, the rate function can be interpreted as a log-likelihood; its derivative leads to [[18], Section E.1]

$$\mathcal{S}^\alpha_{\tau_\alpha \dot{x} + U'_\alpha(x)}(f) = 2D_\alpha + \sum_\beta g_{\alpha\beta}^2 \mathcal{S}^\beta_{\phi(x)}(f), \quad (16)$$

which generalizes Eq. (12) to multiple populations.

Using Eq. (16), the inferred connectivity $g_{\alpha\beta}$ matches the ground truth well; accordingly, two unconnected populations [Figs. 3(a) and 3(b)] can be clearly distinguished from a more involved network where one population ($\alpha = 1$) is only active due to the recurrent input from the other population [$\alpha = 2$, Figs. 3(c) and 3(d)]. The method can thus distinguish intrinsically generated activity from a case where activity is driven from outside the network. However, inference of a unique set of parameters is only possible if the output spectra $\mathcal{S}^\alpha_{\phi(x)}(f)$ differ sufficiently across α. If the output spectra match closely, Eq. (16) leads to a degenerate set of solutions that satisfy $\sum_\beta g_{\alpha\beta}^2 = \mathrm{const}$ and are all equally likely given the data [[18], Section E.2].

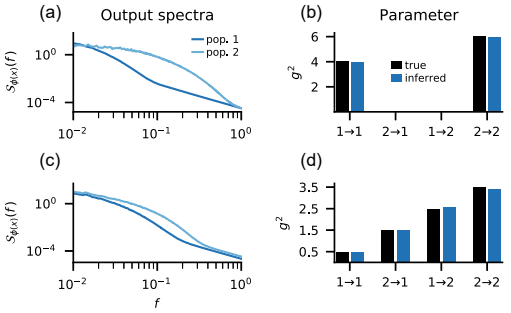

FIG. 3. Maximum likelihood parameter estimation for two populations with different time constants $\tau_1 = 5$, $\tau_2 = 1$. (a) Output power spectra $\mathcal{S}^{\alpha}_{\phi(x)}(f)$ of two unconnected populations $g^2_{12} = g^2_{21} = 0$ with $g^2_{11} = 4$ and $g^2_{22} = 6$. (b) Estimated (blue) and true (black) parameters corresponding to (a). (c) Output power spectra of two connected populations with $g^2_{11} = 0.5$, $g^2_{12} = 1.5$, $g^2_{21} = 2.5$, and $g^2_{22} = 3.5$. (d) Estimated (blue) and true (black) parameters corresponding to (c). Further parameters: $N_1 = N_2 = 5000$, $\phi(x) = \mathrm{erf}(\sqrt{\pi}x/2)$, $U(x) = \frac{1}{2}x^2$, and $D = 0$; simulation parameters as in Fig. 1.

Discussion.—In this Letter, we found a tight link between the field-theoretical approach to neuronal networks and its counterpart based on large deviations theory. We obtained the rate function of the empirical measure for the widely used and analytically solvable model of a recurrent neuronal network [4] by field-theoretical methods. This rate function generalizes the seminal result by Ben Arous and Guionnet [11,12] to arbitrary potentials, transfer functions, and multiple populations. Intriguingly, our derivation elucidates that the rate function is identical to the effective action and takes the form of a Kullback-Leibler divergence, akin to Sanov's theorem for sums of i.i.d. random variables [27,28]. The rate function can thus be interpreted as a distance between an empirical measure, for example given by data, and the activity statistics of the network model. This result allows us to address the inverse problem of inferring the parameters of the connectivity and external input from a set of trajectories and to determine the potential and the transfer function.

We here restricted the analysis to networks with independently drawn random weights with zero mean. Since correlated weights have a profound impact on the dynamics that can be captured using both field theory [31] and large deviations theory [32,33], it is an interesting challenge to extend the analysis in this direction. Likewise, synaptic weights with nonvanishing mean, as they appear in sparsely connected networks, present an interesting extension, because they promote fluctuation-driven states when feedback is sufficiently positive. Motifs are another important deviation from independent weights in biological neural networks are motifs [34], which pose a significant challenge already for the field-theoretical approach [35]. Beyond the weight statistics, we assumed that the dynamics are governed by the first-order differential equation (1). Indeed, the field-theoretical approach can be generalized to a much broader class of dynamics that do not necessarily possess an action [36]; hence, it seems possible to also derive large deviations results for more general dynamics. In this regard, the extension to spiking networks is a particularly interesting but also challenging future direction. Whether the model, Eq. (1), with its current limitations—the independent weights and the first-order dynamics—allows accurate inference of network parameters from cortical recordings is an intriguing question for further research.

The unified description of random networks by statistical field theory and large deviations theory opens the door to established techniques from either domain to capture beyond mean-field behavior. Such corrections are important for small or sparse networks with nonvanishing mean connectivity, to explain correlated neuronal activity, and to study information processing in finite-size networks with realistically limited resources. We here make a first step by computing fluctuation corrections from the rate function. The quantitative theory explains near-critical fluctuations for $g \in [1, 1 + \delta(N)]$ and we discover that expansive gain functions, as found in biology [37], lead to qualitatively different collective behavior than the well-studied contractive sigmoidal ones: The former feature metastable network states with noise-induced first order transitions between them; the latter allow for only a single solution and show second order phase transitions.

We are grateful to Olivier Faugeras and Etienne Tanré for helpful discussions on LDT of neuronal networks, to Anno Kurth for pointing us to the Fréchet derivative, and to Alexandre René, David Dahmen, Kirsten Fischer, and Christian Keup for feedback on an earlier version of the manuscript. This work was partly supported by the Helmholtz young investigator's group VH-NG-1028, European Union Horizon 2020 Grant No. 785907 (Human Brain Project SGA2), the Human Frontier Science Program RGP0057/2016 grant, BMBF Grant "Renormalized Flows" (01IS19077A), and the Excellence Initiative of the German federal and state governments (G:(DE-82)EXS-PF-JARASDS005).

*Corresponding author.
avm@physik.huberlin.de

[1] M. I. Rabinovich, P. Varona, A. I. Selverston, and H. D. I. Abarbanel, Rev. Mod. Phys. **78**, 1213 (2006).
[2] H. Sompolinsky, Phys. Today **41**, No. 12, 70 (1988).
[3] S.-I. Amari, IEEE Trans. **SMC-2**, 643 (1972).
[4] H. Sompolinsky, A. Crisanti, and H. J. Sommers, Phys. Rev. Lett. **61**, 259 (1988).
[5] M. Stern, H. Sompolinsky, and L. F. Abbott, Phys. Rev. E **90**, 062710 (2014).

[6] J. Kadmon and H. Sompolinsky, Phys. Rev. X **5**, 041030 (2015).
[7] J. Aljadeff, M. Stern, and T. Sharpee, Phys. Rev. Lett. **114**, 088101 (2015).
[8] A. van Meegen and B. Lindner, Phys. Rev. Lett. **121**, 258302 (2018).
[9] A. Crisanti and H. Sompolinsky, Phys. Rev. E **98**, 062120 (2018).
[10] J. Schuecker, S. Goedeke, and M. Helias, Phys. Rev. X **8**, 041029 (2018).
[11] G. B. Arous and A. Guionnet, Probab. Theory Relat. Fields **102**, 455 (1995).
[12] A. Guionnet, Probab. Theory Relat. Fields **109**, 183 (1997).
[13] H. Sompolinsky and A. Zippelius, Phys. Rev. Lett. **47**, 359 (1981).
[14] P. Martin, E. Siggia, and H. Rose, Phys. Rev. A **8**, 423 (1973).
[15] H.-K. Janssen, Z. Phys. B **23**, 377 (1976).
[16] C. Chow and M. Buice, J. Math. Neurosci. **5**, 8 (2015).
[17] J. A. Hertz, Y. Roudi, and P. Sollich, J. Phys. A **50**, 033001 (2017).
[18] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevLett.127.158302 for detailed derivations and further information, which includes Refs. [19–23].
[19] J. Stapmanns, T. Kühn, D. Dahmen, T. Luu, C. Honerkamp, and M. Helias, Phys. Rev. E **101**, 042124 (2020).
[20] D. J. MacKay, *Information Theory, Inference and Learning Algorithms* (Cambridge University Press, Cambridge, England, 2003).
[21] G. Matheron, Econ. Geol. **58**, 1246 (1963).
[22] C. K. Williams, Neural Comput. **10**, 1203 (1998).
[23] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright *et al.*, Nat. Methods **17**, 261 (2020).
[24] N. Goldenfeld, *Lectures on Phase Transitions and the Renormalization Group* (Perseus Books, Reading, Massachusetts, 1992).
[25] M. S. Berger, *Nonlinearity and Functional Analysis*, 1st ed. (Elsevier, New York, 1977), ISBN 9780120903504.
[26] J. Zinn-Justin, *Quantum Field Theory and Critical Phenomena* (Clarendon Press, Oxford, 1996).
[27] M. Mezard and A. Montanari, *Information, Physics and Computation* (Oxford University Press, New York, 2009).
[28] H. Touchette, Phys. Rep. **478**, 1 (2009).
[29] M. Helias and D. Dahmen, *Statistical Field Theory for Neural Networks*, vol. 970 (Springer International Publishing, Cham, 2020).
[30] C. L. Lawson and R. J. Hanson, *Solving Least Squares Problems* (SIAM, Philadelphia, 1995).
[31] D. Martí, N. Brunel, and S. Ostojic, Phys. Rev. E **97**, 062314 (2018).
[32] O. Faugeras and J. MacLaurin, Entropy **17**, 4701 (2015).
[33] O. Faugeras, J. MacLaurin, and E. Tanré, arXiv:1901.10248.
[34] S. Song, P. Sjöström, M. Reigl, S. Nelson, and D. Chklovskii, PLoS Biol. **3**, e350 (2005).
[35] D. Dahmen, S. Recanatesi, G. K. Ocker, X. Jia, M. Helias, and E. Shea-Brown, bioRxiv (2020).
[36] C. Keup, T. Kühn, D. Dahmen, and M. Helias, Phys. Rev. X **11**, 021064 (2021).
[37] A. Roxin, N. Brunel, D. Hansel, G. Mongillo, and C. van Vreeswijk, J. Neurosci. **31**, 16217 (2011).

Large Deviations Approach to Random Recurrent Neuronal Networks: Parameter Inference and Fluctuation–Induced Transitions (Supplemental Material)

Alexander van Meegen,[1,2] Tobias Kühn,[1,3,4] and Moritz Helias[1,3]

[1] *Institute of Neuroscience and Medicine (INM-6) and Institute for Advanced Simulation (IAS-6) and JARA-Institute Brain Structure-Function Relationships (INM-10), Jülich Research Centre, Jülich, Germany*
[2] *Institute of Zoology, University of Cologne, 50674 Cologne, Germany*
[3] *Department of Physics, Faculty 1, RWTH Aachen University, Aachen, Germany*
[4] *Laboratoire de Physique de l'ENS, Laboratoire MSC de l'Université de Paris, CNRS, Paris, France*

(Dated: August 23, 2021)

CONTENTS

A. Rate Function (Single Population)

1. Scaled Cumulant Generating Functional

Here, we derive the scaled cumulant generating functional and the saddle-point equations. The first steps of the derivations are akin to the manipulations presented in [1, 2], thus we keep the presentation concise. We interpret the stochastic differential equations governing the network dynamics in the Itô convention. Using the Martin–Siggia–Rose–de Dominicis–Janssen path integral formalism, the expectation $\langle\cdot\rangle_{\boldsymbol{x}|\boldsymbol{J}}$ of some arbitrary functional $G(\boldsymbol{x})$ can be written as

$$\begin{aligned}\left\langle\left\langle G(\boldsymbol{x})\right\rangle_{\boldsymbol{x}|\boldsymbol{J},\boldsymbol{\xi}}\right\rangle_{\boldsymbol{\xi}} &= \int \mathcal{D}\boldsymbol{x}\, \left\langle \delta(\dot{\boldsymbol{x}} + U'(\boldsymbol{x}) + \boldsymbol{J}\phi(\boldsymbol{x}) + \boldsymbol{\xi})\right\rangle_{\boldsymbol{\xi}} G(\boldsymbol{x}) \\ &= \int \mathcal{D}\boldsymbol{x} \int \mathcal{D}\tilde{\boldsymbol{x}}\, e^{S_0(\boldsymbol{x},\tilde{\boldsymbol{x}}) - \tilde{\boldsymbol{x}}^\mathsf{T}\boldsymbol{J}\phi(\boldsymbol{x})} G(\boldsymbol{x}),\end{aligned}$$

where we used the Fourier representation $\delta(x) = \frac{1}{2\pi i}\int_{-i\infty}^{i\infty} e^{\tilde{x}x} d\tilde{x}$ in every timestep in the second step and defined the action

$$S_0(\boldsymbol{x}, \tilde{\boldsymbol{x}}) = \tilde{\boldsymbol{x}}^\mathsf{T}(\dot{\boldsymbol{x}} + U'(\boldsymbol{x})) + D\tilde{\boldsymbol{x}}^\mathsf{T}\tilde{\boldsymbol{x}}.$$

An additional average over realizations of the connectivity $\boldsymbol{J} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, N^{-1}g^2)$ only affects the term $-\tilde{\boldsymbol{x}}^\mathsf{T}\boldsymbol{J}\phi(\boldsymbol{x})$ in the action and results in

$$\langle e^{-\tilde{\boldsymbol{x}}^\mathsf{T}\boldsymbol{J}\phi(\boldsymbol{x})}\rangle_{\boldsymbol{J}} = \int \mathcal{D}C \int \mathcal{D}\tilde{C}\, e^{-N C^\mathsf{T}\tilde{C} + \frac{g^2}{2}\tilde{\boldsymbol{x}}^\mathsf{T} C \tilde{\boldsymbol{x}} + \phi(\boldsymbol{x})^\mathsf{T}\tilde{C}\phi(\boldsymbol{x})},$$

where we introduced the network–averaged auxiliary field

$$C(u, v) = \frac{1}{N}\sum_{i=1}^{N} \phi(x_i(u))\phi(x_i(v))$$

via a Hubbard–Stratonovich transformation. The average over the connectivity and the subsequent Hubbard–Stratonovich transformation decouple the dynamics across units; afterwards the units are only coupled through the global fields C and $\tilde{C}$.

Now, we consider the scaled cumulant generating functional of the empirical density

$$W_N(\ell) = \frac{1}{N}\ln\left\langle\left\langle e^{\sum_{i=1}^N \ell(x_i)}\right\rangle_{\boldsymbol{x}|\boldsymbol{J}}\right\rangle_{\boldsymbol{J}}.$$

Using the above results and the abbreviation $\phi(x) \equiv \phi$, it can be written as

$$W_N(\ell) = \frac{1}{N}\ln\int \mathcal{D}C \int \mathcal{D}\tilde{C}\, e^{-N C^\mathsf{T}\tilde{C} + N\,\Omega_\ell(C,\tilde{C})},$$
$$\Omega_\ell(C, \tilde{C}) = \ln \int \mathcal{D}x \int \mathcal{D}\tilde{x}\, e^{S_0(x,\tilde{x}) + \frac{g^2}{2}\tilde{x}^\mathsf{T} C \tilde{x} + \phi^\mathsf{T}\tilde{C}\phi + \ell(x)},$$

where the N in front of the single–particle cumulant generating functional Ω results from the factorization of the N integrals over x_i and $\tilde{x}_i$ each; thus it is a hallmark of the decoupled dynamics. Next, we approximate the C and $\tilde{C}$ integrals in a saddle–point approximation which yields

$$W_N(\ell) = -C_\ell^\mathsf{T}\tilde{C}_\ell + \Omega_\ell(C_\ell, \tilde{C}_\ell) + O(\ln(N)/N),$$

where C_ℓ and $\tilde{C}_\ell$ are determined by the saddle–point equations

$$C_\ell = \partial_{\tilde{C}}\Omega_\ell(C, \tilde{C})\big|_{C_\ell, \tilde{C}_\ell},$$
$$\tilde{C}_\ell = \partial_C \Omega_\ell(C, \tilde{C})\big|_{C_\ell, \tilde{C}_\ell}.$$

Here, ∂_C denotes a partial functional derivative. In the limit $N \to \infty$, the remainder $O(\ln(N)/N)$ vanishes and the saddle–point approximation becomes exact.

2. *Rate Function*

Here, we derive the rate function from the scaled cumulant generating functional. According to the Gärtner-Ellis theorem [3], we obtain the rate function via the Legendre transformation

$$H(\mu) = \int \mathcal{D}x\, \mu(x)\ell_\mu(x) - W_\infty(\ell_\mu) \tag{1}$$

with ℓ_μ implicitly defined by

$$\mu = W_\infty'(\ell_\mu). \tag{2}$$

Using the Gärtner-Ellis theorem, we implicitly assume that $H(\mu)$ is convex [3]. This is, however, not the same as assuming that μ, or the most likely empirical measure $\bar{\mu}$, is concave. The latter would be a serious restriction as it

would prohibit for example treating the bistable case we investigate in the manuscript. A concave $P(\mu)$, and hence a convex $H(\mu)$, simply corresponds to the situation with a single most likely measure $\bar{\mu}$ but it does not put restrictions on $\bar{\mu}$ itself. In particular, $\bar{\mu}$ may still be bimodal.

Due to the saddle–point equations, the derivative of the cumulant generating functional in Eq. (2) simplifies to $W'_\infty(\ell_\mu) = (\partial_\ell \Omega_\ell)(C_\ell, \tilde{C}_\ell)\big|_{\ell_\mu}$ where the derivative only acts on the ℓ that is explicit in $\Omega_\ell(C_\ell, \tilde{C}_\ell)$ and not on the implicit dependencies through C_ℓ, $\tilde{C}_\ell$. Thus, Eq. (2) yields

$$\mu(x) = \frac{\int \mathcal{D}\tilde{x}\, e^{S_0(x,\tilde{x}) + \frac{g^2}{2}\tilde{x}^\mathsf{T} C_{\ell_\mu} \tilde{x} + \phi^\mathsf{T} \tilde{C}_{\ell_\mu} \phi + \ell_\mu(x)}}{\int \mathcal{D}x \int \mathcal{D}\tilde{x}\, e^{S_0(x,\tilde{x}) + \frac{g^2}{2}\tilde{x}^\mathsf{T} C_{\ell_\mu} \tilde{x} + \phi^\mathsf{T} \tilde{C}_{\ell_\mu} \phi + \ell_\mu(x)}}.$$

Taking the logarithm and using $W_\infty(\ell_\mu) + C_{\ell_\mu}^\mathsf{T} \tilde{C}_{\ell_\mu} = \Omega_{\ell_\mu}(C_{\ell_\mu}, \tilde{C}_{\ell_\mu})$ leads to

$$\ell_\mu(x) = \ln \frac{\mu(x)}{\int \mathcal{D}\tilde{x}\, e^{S_0(x,\tilde{x}) + \frac{g^2}{2}\tilde{x}^\mathsf{T} C_{\ell_\mu} \tilde{x}}} + W_\infty(\ell_\mu) + C_{\ell_\mu}^\mathsf{T} \tilde{C}_{\ell_\mu} - \phi^\mathsf{T} \tilde{C}_{\ell_\mu} \phi.$$

Inserting $\ell_\mu(x)$ into the Legendre transformation (1) yields

$$H(\mu) = \int \mathcal{D}x\, \mu(x) \ln \frac{\mu(x)}{\int \mathcal{D}\tilde{x}\, e^{S_0(x,\tilde{x}) + \frac{g^2}{2}\tilde{x}^\mathsf{T} C_{\ell_\mu} \tilde{x}}} + C_{\ell_\mu}^\mathsf{T} \tilde{C}_{\ell_\mu} - C_\mu^\mathsf{T} \tilde{C}_{\ell_\mu}$$

with

$$C_\mu(u,v) = \int \mathcal{D}x\, \mu(x) \phi(x(u)) \phi(x(v)).$$

Identifying $\mu(x)$ in the saddle–point equation

$$C_{\ell_\mu} = \partial_{\tilde{C}} \Omega_\ell(C, \tilde{C})\big|_{C_{\ell_\mu}, \tilde{C}_{\ell_\mu}} = \frac{\int \mathcal{D}x \int \mathcal{D}\tilde{x}\, \phi\phi e^{S_0(x,\tilde{x}) + \frac{g^2}{2}\tilde{x}^\mathsf{T} C_{\ell_\mu} \tilde{x} + \phi^\mathsf{T} \tilde{C}_{\ell_\mu} \phi + \ell_\mu(x)}}{\int \mathcal{D}x \int \mathcal{D}\tilde{x}\, e^{S_0(x,\tilde{x}) + \frac{g^2}{2}\tilde{x}^\mathsf{T} C_{\ell_\mu} \tilde{x} + \phi^\mathsf{T} \tilde{C}_{\ell_\mu} \phi + \ell_\mu(x)}}$$

yields

$$C_{\ell_\mu}(u,v) = \int \mathcal{D}x\, \mu(x) \phi(x(u)) \phi(x(v))$$

and thus $C_{\ell_\mu} = C_\mu$. Accordingly, the last two terms in the Legendre transformation cancel and we arrive at

$$H(\mu) = \int \mathcal{D}x\, \mu(x) \ln \frac{\mu(x)}{\int \mathcal{D}\tilde{x}\, e^{S_0(x,\tilde{x}) + \frac{g^2}{2}\tilde{x}^\mathsf{T} C_\mu \tilde{x}}} \tag{3}$$

where still $C_\mu(u,v) = \int \mathcal{D}x\, \mu(x) \phi(x(u)) \phi(x(v))$.

In the main text, we use the notation

$$\int \mathcal{D}\tilde{x}\, e^{S_0(x,\tilde{x}) + \frac{g^2}{2}\tilde{x}^\mathsf{T} C_\mu \tilde{x}} = \langle \delta(\dot{x} + U'(x) - \eta) \rangle_\eta$$

with $C_\eta = 2D\delta + g^2 C_\mu$ appearing in the rate function. Indeed, using the Martin–Siggia–Rose–de Dominicis–Janssen formalism, we have

$$\begin{aligned} \langle \delta(\dot{x} + U'(x) - \eta) \rangle_\eta &= \int \mathcal{D}\tilde{x}\, e^{\tilde{x}^\mathsf{T}(\dot{x} + U'(x))} \langle e^{\tilde{x}^\mathsf{T} \eta} \rangle_\eta \\ &= \int \mathcal{D}\tilde{x}\, e^{\tilde{x}^\mathsf{T}(\dot{x} + U'(x)) + \frac{1}{2}\tilde{x}^\mathsf{T} C_\eta \tilde{x}}, \end{aligned}$$

which shows that the two notations are equivalent since $\tilde{x}^\mathsf{T}(\dot{x} + U'(x)) + \frac{1}{2}\tilde{x}^\mathsf{T} C_\eta \tilde{x} = S_0(x,\tilde{x}) + \frac{g^2}{2}\tilde{x}^\mathsf{T} C_\mu \tilde{x}$ for $C_\eta = 2D\delta + g^2 C_\mu$.

3. Equivalence to Ben Arous and Guionnet (1995)

Here, we show explicitly that the rate function we obtained generalizes the rate function obtained by Ben Arous and Guionnet [4], whose limitation to finite temperature and time was lifted later [5]. We start with Theorem 4.1 in [4] adapted to our notation: Define

$$Q(x) \coloneqq \int \mathcal{D}\tilde{x}\, e^{\tilde{x}^\mathsf{T}(\dot{x}+U'(x))+\frac{1}{2}\tilde{x}^\mathsf{T}\tilde{x}}$$

and

$$G(\mu) \coloneqq \int \mathcal{D}x\, \mu(x)\, \ln\left(\langle e^{gy^\mathsf{T}(\dot{x}+U'(x))-\frac{g^2}{2}y^\mathsf{T}y}\rangle_y\right),$$

where $\langle\cdot\rangle_y$ is the expectation value over a zero–mean Gaussian process y with $C_\mu(u,v) = \int \mathcal{D}x\, \mu(x)x(u)x(v)$, written as $\langle\cdot\rangle_y = \int \mathcal{D}y \int \mathcal{D}\tilde{y}\, (\cdot)\, e^{\tilde{y}^\mathsf{T}y+\frac{1}{2}\tilde{y}^\mathsf{T}C_\mu\tilde{y}}$. With the Kullback–Leibler divergence $D_{\mathrm{KL}}(\mu\,|\,Q)$, Theorem 4.1 states that the function

$$\tilde{H}(\mu) = \begin{cases} D_{\mathrm{KL}}(\mu\,|\,Q) - G(\mu) & \text{if } D_{\mathrm{KL}}(\mu\,|\,Q) < \infty \\ +\infty & \text{otherwise} \end{cases}$$

is a good rate function.

Now we relate $\tilde{H}$ to the rate function that is derived above, Eq. (3). Using the Onsager–Machlup action, we can write

$$D_{\mathrm{KL}}(\mu\,|\,Q) = \int \mathcal{D}x\, \mu(x) \ln \frac{\mu(x)}{e^{-S_{\mathrm{OM}}(x)}} + \mathcal{C}$$

with $S_{\mathrm{OM}}(x) = \frac{1}{2}(\dot{x}+U'(x))^\mathsf{T}(\dot{x}+U'(x))$. Next, we transform $gy \to y$, $\tilde{y}/g \to \tilde{y}$ and solve the integral over y in $G(\mu)$:

$$\int \mathcal{D}y\, e^{-\frac{1}{2}y^\mathsf{T}y+y^\mathsf{T}(\dot{x}+U'(x)+\tilde{y})} \propto e^{S_{\mathrm{OM}}[x]+\tilde{y}^\mathsf{T}(\dot{x}+U'(x))+\frac{1}{2}\tilde{y}^\mathsf{T}\tilde{y}}.$$

The Onsager–Machlup action in the logarithm in $D_{\mathrm{KL}}(\mu\,|\,Q)$ and $G(\mu)$ cancel and we arrive at

$$\tilde{H}(\mu) = \int \mathcal{D}x\, \mu(x) \ln \frac{\mu(x)}{\int \mathcal{D}\tilde{y}\, e^{\tilde{y}^\mathsf{T}(\dot{x}+U'(x))+\frac{1}{2}\tilde{y}^\mathsf{T}(g^2C_\mu+\delta)\tilde{y}}}$$

up to an additive constant that we set to zero. Since $C_\mu(u,v) = \int \mathcal{D}x\, \mu(x)x(u)x(v)$, the rate function by Ben Arous and Guionnet is thus equivalent to Eq. (3) with $\phi(x) = x$ and $D = \frac{1}{2}$.

4. Background on Rate Function

Relation to Sompolinsky, Crisanti, Sommers (1988) Here, we relate the approach that we laid out in the main text to the approach pioneered by Sompolinsky, Crisanti, and Sommers [6] (reviewed in [2, 7]) using our notation for consistency. Therein, the starting point is the scaled cumulant–generating functional

$$\hat{W}_N(j) = \frac{1}{N} \ln \left\langle \left\langle e^{j^\mathsf{T}\boldsymbol{x}} \right\rangle_{\boldsymbol{x}|\boldsymbol{J}} \right\rangle_{\boldsymbol{J}},$$

which gives rise to the cumulants of the trajectories. For the linear functional

$$\ell(x) = j^\mathsf{T}x,$$

we have $\sum_{i=1}^N \ell(x_i) = j^\mathsf{T}\boldsymbol{x}$ and thus $W_N(j^\mathsf{T}x) = \hat{W}_N(j)$. Put differently, the scaled cumulant–generating functional of the trajectories $\hat{W}_N(j)$ is a special case of the more general scaled cumulant–generating functional $W_N(\ell)$ we consider in this manuscript. Of course one can start from the scaled cumulant–generating functional of the observable of interest and derive the corresponding rate function. Conversely, we show below how to obtain the rate function of a specific observable from the rate function of the empirical measure.

Contraction Principle Here, we relate the rather general rate function of the empirical measure $H(\mu)$ to the rate function of a particular observable $I(C)$. As an example, we choose the correlation function

$$C(u,v) = \frac{1}{N}\sum_{i=1}^{N}\phi(x_i(u))\phi(x_i(v))$$

because it is a quantity that arises naturally during the Hubbard–Stratonovich transformation. The generic approach to this problem is given by the contraction principle [3]:

$$I(C) = \inf_{\mu \text{ s.t. } C=\int \mathcal{D}x\,\mu(x)\phi\phi} H(\mu).$$

Here, the infimum is constrained to the empirical measures that give rise to the correlation function C, i.e. those that fulfill $C(u,v) = \int \mathcal{D}x\,\mu(x)\phi(x(u))\phi(x(v))$. Writing $H(\mu)$ as the Legendre transform of the scaled cumulant–generating functional, $H(\mu) = \inf_\ell[\int \mathcal{D}x\,\mu(x)\ell(x) - W_\infty(\ell)]$, the empirical measure only appears linearly. Using a Lagrange multiplier $k(u,v)$, the infimum over μ leads to the constraint $\ell(x) = \phi^\mathsf{T} k\phi$ and we arrive at

$$I(C) = \inf_k[k^\mathsf{T}C - W_\infty(\phi^\mathsf{T}k\phi)].$$

Once again, we see how to relate $W_N(\ell)$ to a specific observable—this time for the choice $\ell(x) = \phi^\mathsf{T}k\phi$.

Up to this point, the discussion applies to any observable. For the current example, we can proceed a bit further. With the redefinition $\tilde{C} + k \to \tilde{C}$, we get

$$W_\infty(\phi^\mathsf{T}k\phi) = \operatorname{extr}_{C,\tilde{C}}\left[-C^\mathsf{T}\tilde{C} + C^\mathsf{T}k + \Omega_0(C,\tilde{C})\right],$$
$$\Omega_0(C,\tilde{C}) = \ln\int \mathcal{D}x \int \mathcal{D}\tilde{x}\, e^{S_0(x,\tilde{x}) + \frac{g^2}{2}\tilde{x}^\mathsf{T}C\tilde{x} + \phi^\mathsf{T}\tilde{C}\phi},$$

which made Ω_0 independent of k. Now we can take the infimum over k, leading to

$$I(C) = \operatorname{extr}_{\tilde{C}}\left[C^\mathsf{T}\tilde{C} - \Omega_0(C,\tilde{C})\right]. \tag{4}$$

The remaining extremum gives rise to the condition

$$C = \frac{\int \mathcal{D}x \int \mathcal{D}\tilde{x}\,\phi\phi e^{S_0(x,\tilde{x}) + \frac{g^2}{2}\tilde{x}^\mathsf{T}C_\phi\tilde{x} + \phi^\mathsf{T}\tilde{C}\phi}}{\int \mathcal{D}x \int \mathcal{D}\tilde{x}\, e^{S_0(x,\tilde{x}) + \frac{g^2}{2}\tilde{x}^\mathsf{T}C_\phi\tilde{x} + \phi^\mathsf{T}\tilde{C}\phi}},$$

i.e. a self–consistency condition for the correlation function.

As a side remark, we mention that the expression in the brackets of Eq. (4) is the joint effective action for C and $\tilde{C}$, because for $N \to \infty$, the action equals the effective action. This result is therefore analogous to the finding that the effective action in the Onsager–Machlup formalism is given as the extremum of its counterpart in the Martin–Siggia–Rose–de Dominicis–Janssen formalism [8, Eq.(24)]. The only difference is that here, we are dealing with second order statistics and not just mean values. The origin of this finding is the same in both cases: we are only interested in the statistics of the physical quantity (the one without tilde, x or C, respectively). Therefore we only introduce a source field (k in the present case) for this one, but not for the auxiliary field, which amounts to setting the source field of the latter to zero. This is translated into the extremum in Eq. (4) over the auxiliary variable [8, Appendix 5].

Tail Probability Large deviations results are often stated for the tail probability $\mathbb{P}(x > \theta)$ where θ is in the tail. Since the notion of a tail cannot be unambiguously defined for quantities like the empirical measure or correlation functions, at least not in an obvious way, we here give an example how to relate the rate function of the empirical measure to a tail probability.

First, we use the contraction principle to get a rate function for a scalar quantity, e.g. the order parameter $q = \int \mathcal{D}x\,\mu(x)\phi(x(t))\phi(x(t))$ where t is large but fixed such that the measure becomes stationary:

$$I(q) = \inf_{\mu \text{ s.t. } q=\int \mathcal{D}x\,\mu(x)\phi\phi} H(\mu).$$

Since q is a scalar quantity, one obtains the tail probability as $\ln\mathbb{P}(q > \theta) \simeq -NI(q = \theta)$.

Below, we calculate both the mean and the variance of q. In general, this would not be sufficient to obtain a tail estimate. However, the numerics indicate that the tail is indeed Gaussian (Fig. 3**D**) such that the first two cumulants are indeed sufficient.

B. Inference & Prediction (Single Population)

1. Log–Likelihood Derivative

Here, we calculate the derivatives of the log–likelihood with respect to the parameters g and D. In terms of the rate function, we have

$$\partial_a \ln P(\mu\,|\,g,D) \simeq -N\partial_a H(\mu\,|\,g,D)$$

where a denotes either g or D. The parameters appear only in the cross entropy

$$\partial_a H(\mu) = -\int \mathcal{D}x\,\mu(x)\partial_a \ln\left\langle \delta(\dot{x}+U'(x)-\eta)\right\rangle_\eta$$

through the correlation function $C_\eta(u,v) = 2D\delta(u-v) + g^2\int \mathcal{D}x\,\mu(x)\phi(x(u))\phi(x(v))$. Above, we showed that

$$\left\langle \delta(\dot{x}+U'(x)-\eta)\right\rangle_\eta = \int \mathcal{D}\tilde{x}\, e^{\tilde{x}^\intercal(\dot{x}+U'(x))+\frac{1}{2}\tilde{x}^\intercal C_\eta \tilde{x}}.$$

Because $\tilde{x}$ is at most quadratic in the exponent, the integral is solvable and we get

$$\left\langle \delta(\dot{x}+U'(x)-\eta)\right\rangle_\eta = \frac{e^{-\frac{1}{2}(\dot{x}+U'(x))^\intercal C_\eta^{-1}(\dot{x}+U'(x))}}{\sqrt{\det(2\pi C_\eta)}}.$$

Note that the normalization $1/\sqrt{\det(2\pi C_\eta)}$ does not depend on the potential U. Now we can take the derivatives of $\ln\left\langle \delta(\dot{x}+U'(x)-\eta)\right\rangle_\eta$ and get

$$\partial_a \ln\left\langle \delta(\dot{x}+U'(x)-\eta)\right\rangle_\eta = -\frac{1}{2}(\dot{x}+U'(x))^\intercal \frac{\partial C_\eta^{-1}}{\partial a}(\dot{x}+U'(x)) - \frac{1}{2}\partial_a \mathrm{tr}\ln C_\eta$$

where we used $\ln\det C = \mathrm{tr}\ln C$. With this, we arrive at

$$\partial_a H(\mu) = \frac{1}{2}\mathrm{tr}\left(C_0\frac{\partial C_\eta^{-1}}{\partial a}\right) + \frac{1}{2}\mathrm{tr}\left(\frac{\partial C_\eta}{\partial a}C_\eta^{-1}\right)$$

where the integral over the empirical measure gave rise to $C_0 = \int \mathcal{D}x\,\mu(x)(\dot{x}+U'(x))(\dot{x}+U'(x))$ and we used $\partial_a \ln C = \frac{\partial C}{\partial a}C^{-1}$. Finally, using $\frac{\partial C}{\partial a}C^{-1} = CC^{-1}\frac{\partial C}{\partial a}C^{-1} = -C\frac{\partial C^{-1}}{\partial a}$, we get

$$\partial_a \ln P(\mu\,|\,g,D) \simeq -\frac{N}{2}\mathrm{tr}\left((C_0 - C_\eta)\frac{\partial C_\eta^{-1}}{\partial a}\right)$$

as stated in the main text.

The derivative vanishes for $C_0 = C_\eta$. Assuming stationarity, in Fourier domain this condition reads

$$\mathcal{S}_{\dot{x}+U'(x)}(f) = 2D + g^2\mathcal{S}_{\phi(x)}(f), \tag{5}$$

where $\mathcal{S}_X(f)$ denotes the network–averaged power spectrum of the observable X.

2. Model Comparison

Parameter estimation allows us to determine the statistical properties of the recurrent connectivity g and the external input D. However, this leaves the potential U and the transfer function ϕ unspecified. Here we determine U and ϕ using model comparison techniques [9].

We consider two options to obtain U and ϕ: comparing the mean squared error in Eq. (5) for the inferred parameters and comparing the likelihood of the inferred parameters. For the latter option, we can use the rate function from Eq. (3). Given two choices U_i, ϕ_i, $i \in \{1,2\}$, with corresponding inferred parameters $\hat{g}_i$, $\hat{D}_i$, we have

$$\ln\frac{P(\mu\,|\,U_1,\phi_1,\hat{g}_1,\hat{D}_1)}{P(\mu\,|\,U_2,\phi_2,\hat{g}_2,\hat{D}_2)} \simeq -N(H_1 - H_2) \tag{6}$$

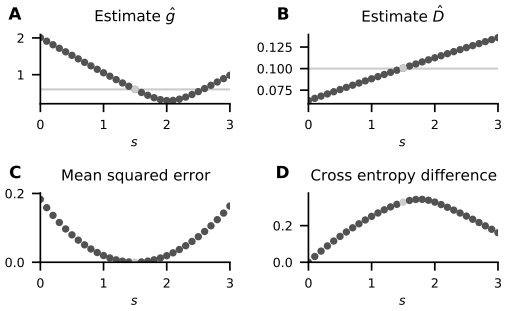

Figure 1. Model comparison for $\phi(x) = \mathrm{erf}(\sqrt{\pi}x/2)$ and $U(x) = \frac{1}{2}x^2 - s\ln\cosh x$. **A**,**B** Maximum likelihood estimates of $\hat{g}$ and $\hat{D}$ for given choices of s. True values of g and D indicated as gray lines; estimates at the true value $s = 1.5$ indicated as gray symbols. **C** Mean squared error between left– and right–hand–side of Eq. (5) for given s. **D** Cross entropy difference between model with $s = 0$ and with given s. Further parameters as in Fig. 1 in the main text.

with $H_i \equiv H(\mu\,|\,U_i, \phi_i, \hat{g}_i, \hat{D}_i)$. The difference $H_1 - H_2$ equals the difference of the minimal cross entropies for the respective choices U_i, ϕ_i. Assuming an infinite observation time, this difference can be expressed as an integral that is straightforward to evaluate numerically (see below).

To illustrate the procedure, we consider the potential

$$U(x) = \frac{1}{2}x^2 - s\ln\cosh x,$$

which is bistable for $s > 1$ [10] and determine s using the mean squared error and the cross entropy difference (see Fig. 1). Parameter estimation yields estimates $\hat{g}$ and $\hat{D}$ that depend on s (Fig. 1**A**,**B**). The mean squared error displays a clear minimum at the true value $s = 1.5$ (Fig. 1**C**) whereas the maximal cross entropy occurs at a value larger than $s = 1.5$ (Fig. 1**D**). The latter effect arises because the cross entropy is dominated by the parameter estimates, thus the mean squared error provides a more reliable criterion in this case.

Cross Entropy Difference Here, we express the cross entropy difference

$$H_1 - H_2 \coloneqq H(\mu\,|\,U_1, \phi_1, \hat{g}_1, \hat{D}_1) - H(\mu\,|\,U_2, \phi_2, \hat{g}_2, \hat{D}_2)$$

in a form that can be evaluated numerically. Using the rate function, we get

$$H_1 - H_2 = \int \mathcal{D}x\,\mu(x)\ln\frac{\langle\delta(\dot{x} + U_2'(x) - \eta_2)\rangle_{\eta_2}}{\langle\delta(\dot{x} + U_1'(x) - \eta_1)\rangle_{\eta_1}}$$

with $C_{\eta_i} = 2\hat{D}_i\delta + \hat{g}_i^2\int\mathcal{D}x\,\mu(x)\phi_i\phi_i$. Again, we use

$$\langle\delta(\dot{x} + U'(x) - \eta)\rangle_\eta = \frac{e^{-\frac{1}{2}(\dot{x}+U'(x))^\mathsf{T}C_\eta^{-1}(\dot{x}+U'(x))}}{\sqrt{\det(2\pi C_\eta)}}$$

to arrive at

$$H_1 - H_2 = -\frac{1}{2}\mathrm{tr}\left(C_1C_{\eta_1}^{-1}\right) - \frac{1}{2}\mathrm{tr}\ln C_{\eta_1} + \frac{1}{2}\mathrm{tr}\left(C_2C_{\eta_2}^{-1}\right) + \frac{1}{2}\mathrm{tr}\ln C_{\eta_2}$$

with $C_i = \int\mathcal{D}x\,\mu(x)(\dot{x} + U_i'(x))(\dot{x} + U_i'(x))$. For stationary correlation functions over infinite time intervals, we can evaluate the traces as integrals over the power spectra:

$$\mathrm{tr}(AB^{-1}) \propto \int_{-\infty}^{\infty}\frac{\tilde{A}(f)}{\tilde{B}(f)}df,$$
$$\mathrm{tr}\ln A \propto \int_{-\infty}^{\infty}\ln(\tilde{A}(f))df.$$

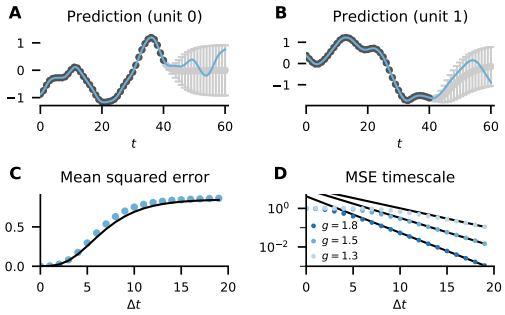

Figure 2. Prediction of the future single–unit activity. **A**,**B** Prediction $\hat{x}$ with uncertainty $\sigma_{\hat{x}}$ (light symbols) for two arbitrary units. Training data (dark symbols) determined by the true trajectory (solid curve). **C** Network–averaged mean squared error ϵ (symbols) and predicted uncertainty $\sigma_{\hat{x}}^2$ (solid curve). **D** The error increases on half of the timescale of the autocorrelation function: $(C_x(0) - \sigma_{\hat{x}}^2)/C_x(0)$ (symbols) decreases asymptotically as $\mathcal{C}\exp(-2\Delta t/\tau_c)$ (lines). Network parameters $\phi(x) = \mathrm{erf}(\sqrt{\pi}x/2)$, $U(x) = \frac{1}{2}x^2$, and $D = 0$; further parameters as in Fig. 1 in the main text.

With this, we get

$$H_1 - H_2 \propto -\frac{1}{2}\int_{-\infty}^{\infty}\frac{\mathcal{S}_{\dot{x}+U_1'(x)}(f)}{2\hat{D}_1 + \hat{g}_1^2\mathcal{S}_{\phi_1(x)}(f)}df - \frac{1}{2}\int_{-\infty}^{\infty}\ln(2\hat{D}_1 + \hat{g}_1^2\mathcal{S}_{\phi_1(x)}(f))df$$
$$+\frac{1}{2}\int_{-\infty}^{\infty}\frac{\mathcal{S}_{\dot{x}+U_2'(x)}(f)}{2\hat{D}_2 + \hat{g}_2^2\mathcal{S}_{\phi_2(x)}(f)}df + \frac{1}{2}\int_{-\infty}^{\infty}\ln(2\hat{D}_2 + \hat{g}_2^2\mathcal{S}_{\phi_2(x)}(f))df.$$

Accordingly, the cross entropy difference can be evaluated with integrals over the respective power spectra that can be obtained using Fast Fourier Transformation.

3. Activity Prediction

If the potential of the model is quadratic, $U(x) \propto \frac{1}{2}x^2$, the measure $\bar{\mu}$ that minimizes the rate function corresponds to a Gaussian process. For Gaussian processes, it is possible to perform Bayes–optimal prediction only based on its correlation function [9, 11]. Denoting the correlation function of the process as C_x (Appendix B4), the prediction is given by

$$\hat{x} = \boldsymbol{k}^{\mathsf{T}}\boldsymbol{K}^{-1}\boldsymbol{x} \tag{7}$$

with $K_{ij} = C_x(t_i, t_j)$, $k_i = C_x(t_i, \hat{t})$, and $x_i = x(t_i)$. Here $\hat{t}$ denotes the time point of the prediction and $\{t_i\}$ a set of time points where the activity is known. The predicted value $\hat{x}$ itself is Gaussian distributed with variance

$$\sigma_{\hat{x}}^2 = \kappa - \boldsymbol{k}^{\mathsf{T}}\boldsymbol{K}^{-1}\boldsymbol{k} \tag{8}$$

where $\kappa = C_x(\hat{t}, \hat{t})$. The variance $\sigma_{\hat{x}}^2$ quantifies the uncertainty associated with the prediction $\hat{x}$.

We use the self-consistent autocorrelation function from Eq. (3) to predict the future activity of two arbitrary units using Eq. (7) and Eq. (8) (Fig. 2**A**,**B**). The network–averaged mean squared error $\epsilon = \frac{1}{N}\sum_{i=1}^{N}(\hat{x}_i - x_i)^2$ is well predicted by Eq. (8) as shown in Fig. 2**C**. The timescale of the error is half of the timescale of the autocorrelation function (Appendix B5). We plot $(C_x(0) - \sigma_{\hat{x}}^2)/C_x(0)$ against an exponential decay $\mathcal{C}\exp(-2\tau/\tau_c)$, where $C_x(\tau)/C_x(0) \sim \exp(-\tau/\tau_c)$, and find a very good agreement (Fig. 2**D**). Since τ_c diverges for $g \searrow 1$ (cf. [6]), the timescale of the error diverges as well.

4. Self–Consistent Correlation Function

Here, we describe how the self–consistent correlation function can be obtained efficiently for quadratic single–unit potentials $U(x) = \frac{1}{2}x^2$. The first part is a brief recapitulation of the approach in [2, 6], the second part specific to the

error function is novel to the best of our knowledge.

For quadratic potentials, the most likely (self–consistent) measure reads

$$\bar{\mu}(x) = \langle \delta(\dot{x} + x - \eta) \rangle_\eta ,$$

corresponding to the Gaussian process $\dot{x} = -x + \eta$, where η is a zero–mean Gaussian process with self–consistent correlation function

$$C_\eta(t_1, t_2) = 2D\,\delta(t_1 - t_2) + g^2 C_\phi(t_1, t_2)$$

with $C_\phi(t_1, t_2) = \int \mathcal{D}x\, \bar{\mu}(x)\, \phi(x(t_1))\phi(x(t_2))$. Using the linearity of the dynamics of x, one obtains an ODE for its stationary autocorrelation function $C_x(\tau)$,

$$\ddot{C}_x = C_x - g^2 C_\phi, \tag{9}$$

with initial conditions $C_x(0) = \sigma_x^2$ and $\dot{C}_x(0) = -D$ [2, 6]. Using Price's theorem, Eq. (9) can be cast into an equation of motion $\ddot{C}_x = -\partial_{C_x} V(C_x, \sigma_x^2)$ in a potential

$$V(C_x, \sigma_x^2) = -\frac{1}{2} C_x^2 + g^2 C_\Phi \tag{10}$$

where $C_\Phi(t_1, t_2) = \int \mathcal{D}x\, \bar{\mu}(x)\, \Phi(x(t_1))\Phi(x(t_2))$ and $\partial_x \Phi(x) = \phi(x)$.

Due to the implicit dependence of C_Φ on C_x and σ_x^2, this is not an initial value problem. To determine σ_x^2, we use energy conservation $\frac{1}{2}\dot{C}_x^2 + V(C_x, \sigma_x^2) = \text{const}$. We restrict ourselves to solutions where $C_x(\tau \to \infty) = 0$ and $\dot{C}_x(\tau \to \infty) = 0$. With this, energy conservation evaluated at $\tau = 0$ and $\tau \to \infty$ yields an equation for σ_x^2:

$$\frac{1}{2} D^2 + V(\sigma_x^2, \sigma_x^2) = V(0, \sigma_x^2). \tag{11}$$

With σ_x^2 determined, Eq. (9) becomes an initial value problem that is straightforward to solve numerically.

Instead of solving Eq. (11) for given D to get σ_x^2, we can use it to answer the inverse question: Given g and a desired activity level σ_x^2, how strong does the external noise D need to be? The answer directly follows from Eq. (11):

$$D(\sigma_x^2) = \sqrt{2(V(0, \sigma_x^2) - V(\sigma_x^2, \sigma_x^2))}. \tag{12}$$

We use Eq. (12) to uncover the multiple self–consistent solutions; they correspond to a non–monotonicity of $D(\sigma_x^2)$. For arbitrary transfer functions, we solve the integrals for C_Φ numerically using an appropriate Gaussian quadrature.

Error Function For the transfer function

$$\phi(x) = \operatorname{erf}(\sqrt{\pi} x/2),$$

we can leverage an analytical expression for C_ϕ [12, Appendix]:

$$C_\phi(\tau) = \frac{2}{\pi} \arcsin\left(\frac{\pi C_x(\tau)}{2 + \pi \sigma_x^2} \right). \tag{13}$$

For convenience, we introduce the scaled correlation function

$$y(\tau) = \frac{\pi C_x(\tau)}{2 + \pi \sigma_x^2}, \qquad C_x(\tau) = \frac{2}{\pi} \frac{y(\tau)}{1 - y_0}.$$

Since y depends linearly on C_x, we get from Eq. (9) an equation of motion for y,

$$\ddot{y} = y - g^2 (1 - y_0) \arcsin(y), \tag{14}$$

with $y(0) \equiv y_0 = \frac{\pi \sigma_x^2}{2 + \pi \sigma_x^2}$ and $\dot{y}(0) = \frac{\pi}{2}(1 - y_0) D$ which again can be rewritten as $\ddot{y} = -\partial_y V(y, y_0)$. Using Eq. (13), we get the explicit expression for the potential

$$V(y, y_0) = -\frac{1}{2} y^2 + g^2 (1 - y_0) \left(\sqrt{1 - y^2} + y \arcsin(y) - 1 \right).$$

We chose the offset of the potential such that $V(0, y_0) = 0$ which reduces Eq. (11) to

$$\frac{\pi^2}{8}(1 - y_0)^2 D^2 + V(y_0, y_0) = 0. \tag{15}$$

We solve Eq. (15) numerically using the Newton–Raphson method implemented in SciPy [13] and Eq. (14) using lsoda from the FORTRAN library odepack through the corresponding SciPy interface.

From Eq. (14) we can determine the timescale of y or equivalently C_x. Since $y(\tau \to \infty) \to 0$, we linearize Eq. (14) for $\tau \gg 0$ to

$$\ddot{y} = (1 - g^2(1 - y_0))y + O(y^3).$$

From here, we can directly read off the timescale:

$$\tau_c = \frac{1}{\sqrt{1 - g^2(1 - y_0)}}. \tag{16}$$

We use Eq. (16) to determine the timescale of the prediction error (see below).

5. *Timescale of Prediction Error*

We here relate the timescale of the prediction error to the timescale of the autocorrelation function $C_x(\tau)/C_x(0) \sim \exp(-\tau/\tau_c)$. The predicted variance in the continuous time limit is determined by the corresponding limit of Eq. (8),

$$\sigma^2_{\hat{x}} = C_x(\hat{t}, \hat{t}) - \int_0^\mathsf{T} \int_0^\mathsf{T} C_x(\hat{t}, u) C_x^{-1}(u, v) C_x(v, \hat{t})\, du\, dv,$$

where T denotes the training interval. Writing $\hat{t} = T + \tau$ and approximating $C_x(T + \tau, u) \approx C_x(T, u)e^{-\tau/\tau_c}$, we get

$$\sigma^2_{\hat{x}} \approx C_x(\hat{t}, \hat{t}) - e^{-2\tau/\tau_c} C_x(T, T),$$

where we used $\int_0^\mathsf{T} C_x^{-1}(u, v) C_x(v, T)\, dv = \delta(u - T)$. Using stationarity $C_x(u, v) = C_x(v - u)$, we arrive at

$$\sigma^2_{\hat{x}}/\sigma^2_x \approx 1 - e^{-2\tau/\tau_c}$$

where $C_x(0) = \sigma^2_x$. Thus, for large τ, the timescale of the prediction error is given by $\tau_c/2$.

C. Fluctuations (Single Population)

1. *Order Parameter Fluctuations*

Here, we derive an expression for the fluctuations of the variance valid for slow dynamics $\tau_c \gg 1$. According to Eq. (16), this is valid for g being of order 1 - in practice, we choose g not too close to 1, however, because of the periodic solutions occurring in finite-size systems in this case [6]. We start with the Legendre transform of the rate function of C, Eq. (4), which is the scaled cumulant generating functional

$$\begin{aligned}
W_\infty(k) &= -C_k^\mathsf{T} \tilde{C}_k + C_k^\mathsf{T} k + \Omega_0(C_k, \tilde{C}_k), \\
\Omega_0(C, \tilde{C}) &= \ln \int \mathcal{D}x \int \mathcal{D}\tilde{x}\, e^{S_0(x,\tilde{x}) + \frac{g^2}{2}\tilde{x}^\mathsf{T} C \tilde{x} + \phi^\mathsf{T} \tilde{C} \phi}, \\
C_k &= \partial_{\tilde{C}} \Omega_0(C, \tilde{C})\big|_{C_k, \tilde{C}_k}, \\
\tilde{C}_k &= k + \partial_C \Omega_0(C, \tilde{C})\big|_{C_k, \tilde{C}_k},
\end{aligned}$$

where we redefined $\phi^\mathsf{T} k \phi \to k$ in the argument of W_∞ to simplify the notation a bit. To determine the fluctuations, we need to calculate the second derivative of the scaled cumulant generating functional $W''(0)$.

We get immediately

$$W'(k) = C_k$$

due to the saddle–point equations. The second derivative is thus simply

$$W''(k) = \frac{dC_k}{dk}.$$

Using the saddle–point equations, we get

$$\left.\frac{dC_k}{dk}\right|_{C_k,\tilde{C}_k} = \frac{dC_k}{dk}^{\mathsf{T}} \partial_C\partial_{\tilde{C}}\Omega_0(C,\tilde{C})\Big|_{C_k,\tilde{C}_k} + \frac{d\tilde{C}_k}{dk}^{\mathsf{T}} \partial_{\tilde{C}}\partial_{\tilde{C}}\Omega_0(C,\tilde{C})\Big|_{C_k,\tilde{C}_k},$$
$$\left.\frac{d\tilde{C}_k}{dk}\right|_{C_k,\tilde{C}_k} = \delta + \frac{dC_k}{dk}^{\mathsf{T}} \partial_C\partial_C\Omega_0(C,\tilde{C})\Big|_{C_k,\tilde{C}_k} + \frac{d\tilde{C}_k}{dk}^{\mathsf{T}} \partial_{\tilde{C}}\partial_C\Omega_0(C,\tilde{C})\Big|_{C_k,\tilde{C}_k}.$$

Evaluated at $k = 0$ where $C_k = C_0$ and $\tilde{C}_k = 0$, we get

$$\left.\frac{dC_k}{dk}\right|_{C_0,0} = \frac{g^2}{2}\left.\frac{dC_k}{dk}\right|_{C_0,0}^{\mathsf{T}} \langle\langle \tilde{x}\tilde{x}, \phi\phi\rangle\rangle_0 + \left.\frac{d\tilde{C}_k}{dk}\right|_{C_0,0}^{\mathsf{T}} \langle\langle \phi\phi, \phi\phi\rangle\rangle_0,$$
$$\left.\frac{d\tilde{C}_k}{dk}\right|_{C_0,0} = \delta + \frac{g^2}{2}\left.\frac{d\tilde{C}_k}{dk}\right|_{C_0,0}^{\mathsf{T}} \langle\langle \phi\phi, \tilde{x}\tilde{x}\rangle\rangle_0,$$

where we dropped $\langle\langle \tilde{x}\tilde{x}, \tilde{x}\tilde{x}\rangle\rangle_0 = 0$. The second equation yields

$$\left.\frac{d\tilde{C}_k}{dk}\right|_{C_0,0} = A^{-1}, \qquad A = \delta - \frac{g^2}{2}\langle\langle \phi\phi, \tilde{x}\tilde{x}\rangle\rangle_0,$$

inserting this in the first we get

$$\left.\frac{dC_k}{dk}\right|_{C_0,0} = A^{-1}\langle\langle \phi\phi, \phi\phi\rangle\rangle_0 B^{-1}, \qquad B = \delta - \frac{g^2}{2}\langle\langle \tilde{x}\tilde{x}, \phi\phi\rangle\rangle_0.$$

We arrive at

$$W''(0) = A^{-1}\langle\langle \phi\phi, \phi\phi\rangle\rangle_0 B^{-1}.$$

To avoid the complication of inverting the operators A and B, which depend on four times, we consider the implicit equation

$$AW''(0)B = \langle\langle \phi\phi, \phi\phi\rangle\rangle_0. \tag{17}$$

Next, we simplify the operators A and B.

First, we note that

$$\begin{aligned}\langle\langle \phi(t_1)\phi(t_2), \tilde{x}(s_1)\tilde{x}(s_2)\rangle\rangle_0 &\equiv \langle \phi(t_1)\phi(t_2)\tilde{x}(s_1)\tilde{x}(s_2)\rangle_0 - \langle \phi(t_1)\phi(t_2)\rangle_0\langle \tilde{x}(s_1)\tilde{x}(s_2)\rangle_0 \\ &= \langle \phi(t_1)\phi(t_2)\tilde{x}(s_1)\tilde{x}(s_2)\rangle_0\end{aligned}$$

because $\langle \tilde{x}\tilde{x}\rangle_0 = 0$. Furthermore, because $\langle\cdot\rangle_0$ is a Gaussian measure, we have

$$\begin{aligned}\langle \phi(t_1)\phi(t_2)\tilde{x}(s_1)\tilde{x}(s_2)\rangle_0 &= \langle \phi''(t_1)\phi(t_2)\rangle_0\langle x(t_1)\tilde{x}(s_1)\rangle_0\langle x(t_1)\tilde{x}(s_2)\rangle_0 \\ &+ \langle \phi'(t_1)\phi'(t_2)\rangle_0\langle x(t_1)\tilde{x}(s_1)\rangle_0\langle x(t_2)\tilde{x}(s_2)\rangle_0 \\ &+ \langle \phi'(t_1)\phi'(t_2)\rangle_0\langle x(t_2)\tilde{x}(s_1)\rangle_0\langle x(t_1)\tilde{x}(s_2)\rangle_0 \\ &+ \langle \phi(t_1)\phi''(t_2)\rangle_0\langle x(t_2)\tilde{x}(s_1)\rangle_0\langle x(t_2)\tilde{x}(s_2)\rangle_0,\end{aligned}$$

which can be derived by expanding $\phi(x(t_1))$ and $\phi(x(t_2))$ as a Taylor series and applying Wick's theorem. The expectation $\langle x(t_1)\tilde{x}(t_2)\rangle_0$ is the response at t_1 to an infinitesimal perturbation at t_2.

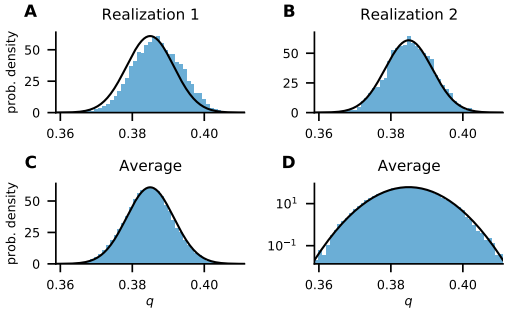

Figure 3. Order parameter fluctuations. **A**,**B** Temporal order parameter statistics for a single realization of random connectivity each. **C**,**D** Temporal order parameter statistics across ten simulations with linear and logarithmic y-axis (panel **D** is identical to Fig. 2**A** in the main text).

For quadratic potentials $U(x) = \frac{1}{2}x^2$, the linear response function is $\langle x(t_1)\tilde{x}(t_2)\rangle_0 = -H(t_1 - t_2)e^{-(t_1-t_2)}$. In particular, its timescale is given by the timescale of the single unit dynamics, i.e. unity in the dimensionless units. In contrast, the timescale of the other expectations $\langle\cdot\rangle_0$ is determined by the timescale of C_x, i.e. τ_c. For $\tau_c \gg 1$, $W''(0)$ hardly changes on the timescale of $\langle x\tilde{x}\rangle_0$, thus we can approximate $\langle x(t_1)\tilde{x}(t_2)\rangle_0 \approx -\delta(t_2 - t_1)$. Because we are only interested in the fluctuations of the variance, we furthermore evaluate Eq. (17) at equal times and consider the stationary case. This turns the contributions to A and B dependent on ϕ and its derivatives into constants and, most notably, renders Eq. (17) independent of time. We therefore suppress the time argument again to arrive at

$$\langle \Delta q^2 \rangle = \frac{\langle\langle \phi\phi, \phi\phi \rangle\rangle_0}{N\left(1 - g^2(\langle \phi''\phi\rangle_0 + \langle \phi'\phi'\rangle_0)\right)^2}$$

as stated in the main text. The factor $1/N$ is due to the definition of the scaled cumulant generating functional, $W_\infty(k) = \lim_{N\to\infty} \frac{1}{N} \ln\left\langle \left\langle e^{NC^\intercal k}\right\rangle_{\boldsymbol{x}|\boldsymbol{J}}\right\rangle_{\boldsymbol{J}}$, where the factor N in the exponent generates a factor N with each derivative of W_∞. Conversely, the derivatives of W_∞ yields the n–th cumulant scaled with $1/N^{n-1}$. Lastly, we used $\langle\langle \phi\phi, \phi\phi\rangle\rangle_0 \equiv \langle \phi\phi\phi\phi\rangle_0 - \langle\phi\phi\rangle_0\langle\phi\phi\rangle_0 = \langle(\phi\phi - \langle\phi\phi\rangle_0)^2\rangle_0$ in the main text.

In the main text, we show the fluctuations of the order parameter across time and realizations of the connectivity in Fig. 2**A**. To supplement this, we show the order parameter fluctuations in Fig. 3 for two realizations of the connectivity (Fig. 3**A**,**B**) and averaged across ten realizations of the connectivity (Fig. 3**C**,**D**). Using a logarithmic y-axis reveals that also the tails are Gaussian.

2. Coexisting Mean–Field Solutions

Here, we determine a regime where two mean–field solutions coexist. We restrict ourselves to quadratic potentials $U(x) = \frac{1}{2}x^2$ and start from Eq. (12),

$$D(\sigma_x^2) = \sqrt{2(V(0,\sigma_x^2) - V(\sigma_x^2,\sigma_x^2))},$$

which determines the necessary external noise to reach a given activity level σ_x^2. Non–monotonicities of $D(\sigma_x^2)$ give rise to multiple solutions since they indicate a case where the same external noise can lead to different activity levels.

We focus on the linearly stable case $g < 1$ with antisymmetric $\phi(x)$ and $\phi'(0) = 1$. For small σ_x^2, we approximate $\Phi(x) = \frac{1}{2}x^2 + \frac{\alpha}{24}x^4 + O(x^6)$. Using Wick's theorem and Eq. (10), we get

$$2(V(0,\sigma_x^2) - V(\sigma_x^2,\sigma_x^2)) = (1-g^2)\sigma_x^4 - \alpha g^2\sigma_x^6 + O(\sigma_x^8).$$

For $g < 1$, the leading order term grows monotonically with σ_x. To introduce a non–monotonicity, the next term has to shrink which implies $\alpha > 0$. This excludes sigmoidal functions like $\phi(x) = \mathrm{erf}(\sqrt{\pi}x/2)$ or $\phi(x) = \tanh(x)$. Thus, we consider non–sigmoidal functions with $\alpha > 1$ that we keep bounded between -1 and 1 by clipping them to the interval $[-1, 1]$.

In the noiseless case $D = 0$, the silent fixed point $\sigma_x^2 = 0$ is one of the two solutions. Using the stability criterion $g^2\langle\phi'(\sigma_x)^2\rangle < 1$ from [14], we get for the transfer function $\phi(x) = \mathrm{clip}(\tan(x), -1, 1)$

$$g^2\langle\phi'(\sigma_x z)^2\rangle = g^2\int_{-\pi/4}^{\pi/4} dz\,\mathcal{N}(z\,|\,0,\sigma_x^2)\,\cos^{-4}(z) \overset{\sigma_x\to 0}{\to} g^2,$$

hence the silent fixed point is stable for $g < 1$.

D. Rate Function (Multiple Populations)

1. Scaled Cumulant Generating Functional

Here, we derive the scaled cumulant generating functional and the saddle-point equations for networks with multiple populations. The steps are similar to the single population case, hence we keep the presentation brief. Throughout, we use greek indices for the populations and latin indices for individual neurons within a given population: x_i^α denotes the trajectory of neuron i of population α, $\boldsymbol{x}^\alpha$ the trajectories of all neurons in population α, and $\boldsymbol{x}$ the trajectories of all neurons. The same convention applies to the connectivity: $J_{ij}^{\alpha\beta}$ governs the connection from neuron j in population β to neuron i in population α, $\boldsymbol{J}^{\alpha\beta}$ the connections from all neurons in population β to all neurons in population α, and $\boldsymbol{J}$ all connections. Furthermore, we denote the size of an individual population by N_α and set $N = \sum_\alpha N_\alpha$.

The expectation $\langle\cdot\rangle_{\boldsymbol{x}|\boldsymbol{J}}$ of some arbitrary functional $G(\boldsymbol{x})$ can again be written as

$$\left\langle\langle G(\boldsymbol{x})\rangle_{\boldsymbol{x}|\boldsymbol{J},\boldsymbol{\xi}}\right\rangle_{\boldsymbol{\xi}} = \prod_\alpha\int \mathcal{D}\boldsymbol{x}^\alpha\,P(\boldsymbol{x}\,|\,\boldsymbol{J})\,G(\boldsymbol{x}),$$

where we introduced

$$\begin{aligned}P(\boldsymbol{x}\,|\,\boldsymbol{J}) &= \prod_\alpha\left\langle\delta(\tau_\alpha\dot{\boldsymbol{x}}^\alpha + U'_\alpha(\boldsymbol{x}^\alpha) + \sum_\beta \boldsymbol{J}^{\alpha\beta}\phi(\boldsymbol{x}^\beta) + \boldsymbol{\xi}^\alpha)\right\rangle_{\boldsymbol{\xi}^\alpha}\\ &= \prod_\alpha\int\mathcal{D}\tilde{\boldsymbol{x}}^\alpha\; e^{\sum_\alpha S_0^\alpha(\boldsymbol{x}^\alpha,\tilde{\boldsymbol{x}}^\alpha) - \sum_{\alpha,\beta}\tilde{\boldsymbol{x}}^{\alpha\mathsf{T}}\boldsymbol{J}^{\alpha\beta}\phi(\boldsymbol{x}^\beta)}.\end{aligned}$$

The action S_0^α now depends on the population,

$$S_0^\alpha(\boldsymbol{y},\tilde{\boldsymbol{y}}) = \tilde{\boldsymbol{y}}^\mathsf{T}(\tau_\alpha\dot{\boldsymbol{y}} + U'_\alpha(\boldsymbol{y})) + D_\alpha\tilde{\boldsymbol{y}}^\mathsf{T}\tilde{\boldsymbol{y}}.$$

The average over realizations of the connectivity $\boldsymbol{J}^{\alpha\beta} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, N_\beta^{-1}g_{\alpha\beta}^2)$ only affects the term $-\sum_{\alpha,\beta}\tilde{\boldsymbol{x}}^{\alpha\mathsf{T}}\boldsymbol{J}^{\alpha\beta}\phi(\boldsymbol{x}^\beta)$. Due to the independence of the entries of $\boldsymbol{J}$, the average factorizes into

$$\langle e^{-\sum_{\alpha,\beta}\tilde{\boldsymbol{x}}^{\alpha\mathsf{T}}\boldsymbol{J}^{\alpha\beta}\phi(\boldsymbol{x}^\beta)}\rangle_{\boldsymbol{J}} = \prod_{\alpha,i}\prod_{\beta,j}\langle e^{-\tilde{x}_i^{\alpha\mathsf{T}}J_{ij}^{\alpha\beta}\phi(x_j^\beta)}\rangle_{J_{ij}^{\alpha\beta}} = \prod_{\alpha,i} e^{\frac{1}{2}\tilde{x}_i^{\alpha\mathsf{T}}\left(\sum_{\beta,j}\frac{g_{\alpha\beta}^2}{N_\beta}\phi(x_j^\beta)\phi(x_j^\beta)^\mathsf{T}\right)\tilde{x}_i^\alpha}.$$

Next, we introduce the population–averaged auxiliary fields

$$C^\alpha(u,v) = \frac{1}{N_\alpha}\sum_{i=1}^{N_\alpha}\phi(x_i^\alpha(u))\phi(x_i^\alpha(v))$$

via Hubbard–Stratonovich transformations:

$$\langle e^{-\sum_{\alpha,\beta}\tilde{\boldsymbol{x}}^{\alpha\mathsf{T}}\boldsymbol{J}^{\alpha\beta}\phi(\boldsymbol{x}^\beta)}\rangle_{\boldsymbol{J}} = \prod_\alpha\int\mathcal{D}C^\alpha\int\mathcal{D}\tilde{C}^\alpha\, e^{-\sum_\alpha N_\alpha C^{\alpha\mathsf{T}}\tilde{C}^\alpha + \sum_\alpha\phi(\boldsymbol{x}^\alpha)^\mathsf{T}\tilde{C}^\alpha\phi(\boldsymbol{x}^\alpha) + \frac{1}{2}\sum_\alpha\tilde{\boldsymbol{x}}^{\alpha\mathsf{T}}(\sum_\beta g_{\alpha\beta}^2 C^\beta)\tilde{\boldsymbol{x}}^\alpha}.$$

As in the single-population case, the average over the connectivity and the subsequent Hubbard–Stratonovich transformation decouple the dynamics across units; afterwards, the units are only coupled through the global fields C^α and $\tilde{C}^\alpha$.

Now, we consider the empirical densities of the populations,

$$\mu^\alpha(y) = \frac{1}{N_\alpha}\sum_{i=1}^{N_\alpha}\delta(x_i^\alpha - y). \tag{18}$$

The corresponding scaled cumulant generating functional is

$$W_N(\{\ell^\circ\}) = \frac{1}{N} \ln \left\langle \left\langle e^{\sum_\alpha \sum_{i=1}^{N_\alpha} \ell^\alpha(x_i)} \right\rangle_{\boldsymbol{x}|\boldsymbol{J}} \right\rangle_{\boldsymbol{J}}, \tag{19}$$

where we introduced one functional ℓ^α for each μ^α and the collection of all ℓ^α, $\{\ell^\circ\}$. Using the above results and the abbreviation $\phi(x) \equiv \phi$, it can be written as

$$W_N(\{\ell^\circ\}) = \frac{1}{N} \ln \prod_\alpha \int \mathcal{D}C^\alpha \int \mathcal{D}\tilde{C}^\alpha \, e^{-\sum_\alpha N_\alpha C^{\alpha\mathsf{T}} \tilde{C}^\alpha + \sum_\alpha N_\alpha \Omega^\alpha_{\ell^\alpha}(\{C^\circ\}, \tilde{C}^\alpha)},$$

where we introduced

$$\Omega^\alpha_\ell(\{C^\circ\}, \tilde{C}) = \ln \int \mathcal{D}x \int \mathcal{D}\tilde{x} \, e^{S_0^\alpha(x,\tilde{x}) + \frac{1}{2}\tilde{x}^\mathsf{T}(\sum_\beta g^2_{\alpha\beta} C^\beta)\tilde{x} + \phi^\mathsf{T}\tilde{C}\phi + \ell(x)}.$$

Again, the N_α in front of the single–particle cumulant generating functionals Ω^α_ℓ results from the factorization of the N_α integrals over x_i^α and $\tilde{x}_i^\alpha$ each; thus it is a hallmark of the decoupled dynamics. Note that $W_N(\{\ell^\circ\})$ is still coupled across populations, because each Ω^α_ℓ depends on the set of all auxiliary fields, $\{C^\circ\}$.

Next, we approximate the C^α and $\tilde{C}^\alpha$ integrals in a saddle–point approximation which yields

$$W_\infty(\{\ell^\circ\}) = -\sum_\alpha \gamma_\alpha C^{\alpha\mathsf{T}}_{\{\ell^\circ\}} \tilde{C}^\alpha_{\{\ell^\circ\}} + \sum_\alpha \gamma_\alpha \, \Omega^\alpha_{\ell^\alpha}(\{C^\circ_{\{\ell^\circ\}}\}, \tilde{C}^\alpha_{\{\ell^\circ\}}), \tag{20}$$

where $\gamma_\alpha = N_\alpha/N$. $C^\alpha_{\{\ell^\circ\}}$ and $\tilde{C}^\alpha_{\{\ell^\circ\}}$ are determined by the saddle–point equations

$$C^\alpha_{\{\ell^\circ\}} = \partial_{\tilde{C}} \Omega^\alpha_{\ell^\alpha}(\{C^\circ\}, \tilde{C})\big|_{\{C^\circ_{\{\ell^\circ\}}\}, \tilde{C}^\alpha_{\{\ell^\circ\}}}, \tag{21}$$

$$\gamma_\alpha \tilde{C}^\alpha_{\{\ell^\circ\}} = \sum_\beta \gamma_\beta \, \partial_{C^\alpha} \Omega^\beta_{\ell^\beta}(\{C^\circ\}, \tilde{C})\Big|_{\{C^\circ_{\{\ell^\circ\}}\}, \tilde{C}^\beta_{\{\ell^\circ\}}}. \tag{22}$$

Here, the asymmetry in the saddle-point equations reflects the fact that Ω^α_ℓ depends on a single $\tilde{C}$ but on all $\{C^\circ\}$.

2. Rate Function

Here, we derive the rate function from the scaled cumulant generating functional for the multi-population case. We obtain the rate function via the Legendre transformation

$$H(\{\mu^\circ\}) = \sum_\alpha \gamma_\alpha \int \mathcal{D}x \, \mu^\alpha(x) \ell^\alpha_{\{\mu^\circ\}}(x) - W_\infty(\{\ell^\circ_{\{\mu^\circ\}}\}) \tag{23}$$

with $\ell^\alpha_{\{\mu^\circ\}}$ implicitly defined by

$$\gamma_\alpha \mu^\alpha = \partial_{\ell^\alpha} W_\infty(\{\ell^\circ\})|_{\{\ell^\circ_{\{\mu^\circ\}}\}}. \tag{24}$$

Due to the saddle–point equations, Eq. (21) and Eq. (22), the derivative of the cumulant generating functional in Eq. (24) simplifies to

$$\partial_{\ell^\alpha} W_\infty(\{\ell^\circ\})|_{\{\ell^\circ_{\{\mu^\circ\}}\}} = \gamma_\alpha \, \partial_{\ell^\alpha} \Omega^\alpha_{\ell^\alpha}(\{C^\circ_{\{\ell^\circ\}}\}, \tilde{C}^\alpha_{\{\ell^\circ\}})\big|_{\{\ell^\circ_{\{\mu^\circ\}}\}},$$

where the derivative only acts on the ℓ^α that is explicit in $\Omega^\alpha_{\ell^\alpha}$ and not on the implicit dependencies through $\{C^\circ_{\{\ell^\circ\}}\}$, $\tilde{C}^\alpha_{\{\ell^\circ\}}$. Thus, Eq. (24) yields

$$\mu^\alpha(x) = \left. \frac{\langle \delta(\tau_\alpha \dot{x} + U'_\alpha(x) - \eta_\alpha) \rangle_{\eta_\alpha} \, e^{\phi^\mathsf{T} \tilde{C}^\alpha_{\{\ell^\circ\}} \phi + \ell^\alpha(x)}}{\int \mathcal{D}x \, \langle \delta(\tau_\alpha \dot{x} + U'_\alpha(x) - \eta_\alpha) \rangle_{\eta_\alpha} \, e^{\phi^\mathsf{T} \tilde{C}^\alpha_{\{\ell^\circ\}} \phi + \ell^\alpha(x)}} \right|_{\{\ell^\circ_{\{\mu^\circ\}}\}}, \tag{25}$$

where we used

$$\int \mathcal{D}\tilde{x} \, e^{\tilde{x}^\mathsf{T}(\tau\dot{x} + U'(x)) + \frac{1}{2}\tilde{x}^\mathsf{T} C_\eta \tilde{x}} = \langle \delta(\tau\dot{x} + U'(x) - \eta) \rangle_\eta$$

to introduce the zero-mean Gaussian process η_α with correlation function

$$C^\alpha_\eta(u,v) = 2D_\alpha\delta(u-v) + \sum_\beta g^2_{\alpha\beta} C^\beta_{\{\ell^\circ_{\{\mu^\circ\}}\}}(u,v).$$

Taking the logarithm of Eq. (25) and using the definition of Ω^α_ℓ leads to

$$\ell^\alpha_{\{\mu^\circ\}}(x) = \ln \frac{\mu^\alpha(x)}{\langle \delta(\tau_\alpha \dot{x} + U'_\alpha(x) - \eta_\alpha)\rangle_{\eta_\alpha}} - \phi^\mathsf{T} \tilde{C}^\alpha_{\{\ell^\circ_{\{\mu^\circ\}}\}}\phi + \Omega^\alpha_{\ell^\alpha}(\{C^\circ_{\{\ell^\circ\}}\}, \tilde{C}^\alpha_{\{\ell^\circ\}})\big|_{\{\ell^\circ_{\{\mu^\circ\}}\}}.$$

Inserting $\ell^\alpha_{\{\mu^\circ\}}(x)$ into the Legendre transformation (23) and using Eq. (20) as $\sum_\alpha \gamma_\alpha \Omega^\alpha_{\ell^\alpha}(\{C^\circ_{\{\ell^\circ\}}\}, \tilde{C}^\alpha_{\{\ell^\circ\}}) - W_\infty(\{\ell^\circ\}) = \sum_\alpha \gamma_\alpha C^{\alpha\mathsf{T}}_{\{\ell^\circ\}} \tilde{C}^\alpha_{\{\ell^\circ\}}$ yields

$$H(\{\mu^\circ\}) = \sum_\alpha \gamma_\alpha \int \mathcal{D}x\, \mu^\alpha(x) \ln \frac{\mu^\alpha(x)}{\langle \delta(\tau_\alpha \dot{x} + U'_\alpha(x) - \eta_\alpha)\rangle_{\eta_\alpha}} - \sum_\alpha \gamma_\alpha C^\mathsf{T}_{\mu^\alpha} \tilde{C}^\alpha_{\{\ell^\circ_{\{\mu^\circ\}}\}} + \sum_\alpha \gamma_\alpha C^{\alpha\mathsf{T}}_{\{\ell^\circ_{\{\mu^\circ\}}\}} \tilde{C}^\alpha_{\{\ell^\circ_{\{\mu^\circ\}}\}},$$

where

$$C_{\mu^\alpha}(u,v) = \int \mathcal{D}x\, \mu^\alpha(x)\phi(x(u))\phi(x(v)).$$

Identifying $\mu^\alpha(x)$ in the saddle–point equation (21) yields

$$C^\alpha_{\{\ell^\circ_{\{\mu^\circ\}}\}}(u,v) = \int \mathcal{D}x\, \mu^\alpha(x)\phi(x(u))\phi(x(v))$$

and thus $C^\alpha_{\{\ell^\circ_{\{\mu^\circ\}}\}} = C_{\mu^\alpha}$. Accordingly, the last two terms in the Legendre transformation cancel and we arrive at

$$H(\{\mu^\circ\}) = \sum_\alpha \gamma_\alpha \int \mathcal{D}x\, \mu^\alpha(x) \ln \frac{\mu^\alpha(x)}{\langle \delta(\tau_\alpha \dot{x} + U'_\alpha(x) - \eta_\alpha)\rangle_{\eta_\alpha}}, \tag{26}$$

where η_α is a zero-mean Gaussian process with correlation function

$$C^\alpha_\eta(u,v) = 2D_\alpha\delta(u-v) + \sum_\beta g^2_{\alpha\beta} \int \mathcal{D}x\, \mu^\beta(x)\phi(x(u))\phi(x(v)). \tag{27}$$

Note that although Eq. (26) is a sum over the populations, the individual terms are still coupled through Eq. (27).

The derivation can be generalized further to population-specific transfer functions $\phi_\alpha(x^\alpha_i)$. Since this would make the notation more complicated without any conceptual changes, we just state the result: The rate function is still given by Eq. (26) but the correlation function of η_α becomes

$$C^\alpha_\eta(u,v) = 2D_\alpha\delta(u-v) + \sum_\beta g^2_{\alpha\beta} \int \mathcal{D}x\, \mu^\beta(x)\phi_\beta(x(u))\phi_\beta(x(v)).$$

In the main text, we state only the slightly less general result for $\phi_\alpha \equiv \phi$.

E. Inference (Multiple Populations)

1. *Log–Likelihood Derivative*

Here, we calculate the derivatives of the log–likelihood with respect to the parameters $g_{\alpha\beta}$ and D_α for the multi-population case. We denote the matrix with elements $g_{\alpha\beta}$ by $\boldsymbol{g}$ and the vector with elements D_α by $\boldsymbol{D}$ and proceed similar to the single population case.

In terms of the rate function, Eq. (26), we have

$$\partial_{a_\alpha} \ln P(\{\mu^\circ\}\,|\,\boldsymbol{g},\boldsymbol{D}) \simeq -N\partial_{a_\alpha} H(\{\mu^\circ\}\,|\,\boldsymbol{g},\boldsymbol{D})$$

where a_α denotes either $g_{\alpha\beta}$ and D_α. The parameters a_α appear only in the cross entropy of population α

$$\partial_{a_\alpha} H(\{\mu^\circ\}) = -\gamma_\alpha \int \mathcal{D}x\, \mu^\alpha(x)\, \partial_{a_\alpha} \ln \langle \delta(\tau_\alpha \dot{x} + U'_\alpha(x) - \eta_\alpha)\rangle_{\eta_\alpha}$$

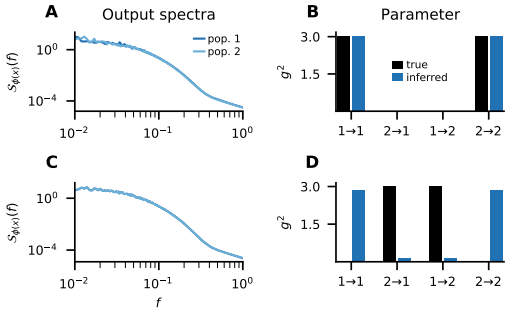

Figure 4. Maximum likelihood parameter estimation for two populations with equal time constants $\tau_1 = \tau_2 = 1$ and equal quadratic row sums $\sum_\beta g^2_{\alpha\beta} = 3 \quad \forall\alpha$. **A** Output power spectra $\mathcal{S}^\alpha_{\phi(x)}(f)$ of two unconnected populations $g^2_{12} = g^2_{21} = 0$ with $g^2_{11} = g^2_{22} = 3$. **B** Estimated (blue) and true (black) parameters corresponding to A. **C** Output power spectra of two connected populations with $g^2_{11} = g^2_{22} = 0$, $g^2_{12} = g^2_{21} = 3$. **D** Estimated (blue) and true (black) parameters corresponding to C. Further parameters as in Fig. 3 in the main text.

through the correlation function $C^\alpha_\eta(u,v) = 2D_\alpha\delta(u-v) + \sum_\beta g^2_{\alpha\beta}\int \mathcal{D}x\,\mu^\beta(x)\phi(x(u))\phi(x(v))$. In the calculation for the log-likelihood derivative for the single population, we showed that

$$\langle\delta(\tau\dot{x} + U'(x) - \eta)\rangle_\eta = \frac{e^{-\frac{1}{2}(\tau\dot{x}+U'(x))^\mathsf{T} C_\eta^{-1}(\tau\dot{x}+U'(x))}}{\sqrt{\det(2\pi C_\eta)}}.$$

With this, we can take the derivatives of $\ln\langle\delta(\tau_\alpha\dot{x} + U'_\alpha(x) - \eta_\alpha)\rangle_{\eta_\alpha}$ and get

$$\partial_{a_\alpha}\ln\langle\delta(\tau_\alpha\dot{x} + U'_\alpha(x) - \eta_\alpha)\rangle_{\eta_\alpha} = -\frac{1}{2}(\tau_\alpha\dot{x} + U'_\alpha(x))^\mathsf{T}\frac{\partial(C^\alpha_\eta)^{-1}}{\partial a_\alpha}(\tau_\alpha\dot{x} + U'_\alpha(x)) - \frac{1}{2}\partial_{a_\alpha}\mathrm{tr}\ln C^\alpha_\eta,$$

where we used $\ln\det C = \mathrm{tr}\ln C$. With this, we arrive at

$$\partial_{a_\alpha}H(\{\mu^\circ\}) = \frac{\gamma_\alpha}{2}\mathrm{tr}\left(C^\alpha_0\frac{\partial(C^\alpha_\eta)^{-1}}{\partial a_\alpha}\right) + \frac{\gamma_\alpha}{2}\mathrm{tr}\left(\frac{\partial C^\alpha_\eta}{\partial a_\alpha}(C^\alpha_\eta)^{-1}\right),$$

where the integral over the empirical measure μ^α gave rise to $C^\alpha_0 = \int\mathcal{D}x\,\mu^\alpha(x)(\tau_\alpha\dot{x} + U'_\alpha(x))(\tau_\alpha\dot{x} + U'_\alpha(x))$ and we used $\partial_a\ln C = \frac{\partial C}{\partial a}C^{-1}$. Finally, using $\frac{\partial C}{\partial a}C^{-1} = CC^{-1}\frac{\partial C}{\partial a}C^{-1} = -C\frac{\partial C^{-1}}{\partial a}$, we get

$$\partial_a\ln P(\{\mu^\alpha\}\,|\,\boldsymbol{g},\boldsymbol{D}) \simeq -\frac{N_\alpha}{2}\mathrm{tr}\left((C^\alpha_0 - C^\alpha_\eta)\frac{\partial(C^\alpha_\eta)^{-1}}{\partial a_\alpha}\right). \tag{28}$$

The derivative vanishes for $C^\alpha_0 = C^\alpha_\eta$.

Assuming stationarity, a Fourier transformation of $C^\alpha_0 = C^\alpha_\eta$ leads to

$$\mathcal{S}^\alpha_{\tau_\alpha\dot{x}+U'_\alpha(x)}(f) = 2D_\alpha + \sum_\beta g^2_{\alpha\beta}\mathcal{S}^\beta_{\phi(x)}(f) \tag{29}$$

as stated in the main text.

2. *Degeneracy of Inference Equation*

Here, we show that parameter inference using Eq. (29) can be degenerate because different models are equally plausible.

If the empirical estimates of the output spectra agree, $\mathcal{S}^\alpha_{\phi(x)}(f) = \mathcal{S}^\beta_{\phi(x)}(f) \equiv \mathcal{S}^\beta_{\phi(x)}(f)$, Eq. (29) reduces to

$$\mathcal{S}^\alpha_{\tau_\alpha\dot{x}+U'_\alpha(x)}(f) = 2D_\alpha + \mathcal{S}_{\phi(x)}(f)\sum_\beta g^2_{\alpha\beta}.$$

Clearly, this leads to a degenerate space of solutions with $\sum_\beta g_{\alpha\beta}^2 = \text{const}$.

For example, we consider the case with $\tau_1 = \tau_2 = 1$ and $\sum_\beta g_{\alpha\beta}^2 = 3$ in Fig. 4. The most likely set of empirical measures for these parameters is $\bar{\mu}^\alpha = \bar{\mu}^\beta$, hence the most likely empirical output spectra agree. Indeed, the empirical output spectra of the two populations agree almost perfectly for a given realization of the connectivity (Fig. 4**A**,**C**), thereby rendering the inference degenerate. Accordingly, for two populations without self-connections, $g_{11}^2 = g_{22}^2 = 0$, $g_{12}^2 = g_{21}^2 = 3$, the parameter inference infers the opposite of two almost unconnected populations (Fig. 4**C**,**D**). Curiously, the inferred parameters agree perfectly with the true parameters if the populations are unconnected (Fig. 4**A**,**B**). This is a finite-size effect: For unconnected networks, the estimates of the output spectra are independent, which leads to different finite-size fluctuations (compare Fig. 4**A** and Fig. 4**C**) such that the inference is not degenerate anymore.

[1] M. Helias and D. Dahmen, *Statistical Field Theory for Neural Networks*, vol. 970 (Springer International Publishing, 2020).
[2] J. Schuecker, S. Goedeke, and M. Helias, Phys. Rev. X **8**, 041029 (2018).
[3] H. Touchette, Physics Reports **478**, 1 (2009).
[4] G. B. Arous and A. Guionnet, Probability Theory and Related Fields **102**, 455 (1995), ISSN 1432-2064.
[5] A. Guionnet, Probability Theory and Related Fields **109**, 183 (1997).
[6] H. Sompolinsky, A. Crisanti, and H. J. Sommers, Phys. Rev. Lett. **61**, 259 (1988).
[7] A. Crisanti and H. Sompolinsky, Phys. Rev. E **98**, 062120 (2018).
[8] J. Stapmanns, T. Kühn, D. Dahmen, T. Luu, C. Honerkamp, and M. Helias, Phys. Rev. E **101**, 042124 (2020).
[9] D. J. MacKay, *Information theory, inference and learning algorithms* (Cambridge university press, 2003).
[10] M. Stern, H. Sompolinsky, and L. F. Abbott, Phys. Rev. E **90**, 062710 (2014).
[11] G. Matheron, Economic Geology **58**, 1246 (1963).
[12] C. K. Williams, Neural Comput. **10**, 1203 (1998).
[13] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, et al., Nat. Methods **17**, 261 (2020).
[14] J. Kadmon and H. Sompolinsky, Phys. Rev. X **5**, 041030 (2015).

6

MICROSCOPIC THEORY OF INTRINSIC TIMESCALES IN SPIKING NEURAL NETWORKS

PREAMBLE

Up to this chapter, the neuron models were rather abstract and, in particular, rate-based. In this chapter, we take spikes into account using either Generalized Linear Models (GLMs) or Leaky Integrate-and-Fire models (LIFs). Due to this additional complexity, the focus is now fully constrained to the dynamics. In terms of the network structure, we stick to block-structured random networks but with the small addition of a non-vanishing mean.

The main motivation for the work presented in this chapter is an intriguing experimental observation by Murray et al. (2014): the intrinsic timescale—the autocorrelation time—increases systematically across the cortical hierarchy. Importantly, this observation is based on single-neuron recordings and not the population activity (although similar trends were recently observed in fMRI data by Manea et al. 2022). This raises the question how network- and neuron-parameters influence the intrinsic timescale.

To investigate this analytically, dynamic mean-field theory is ideally suited because it describes the single-neuron statistics (see Section 3.4). Because the output of the neurons is a point process (see Section 3.3), determining the relation between input- and output-statistics becomes the main challenge.

We address this challenge for GLM neurons by deriving analytical solutions for exponential and error-function nonlinearities. For LIF neurons, we rely on a level-crossing approach and an approximation proposed by Stratonovich (1967). In both cases, the resulting theory yields network-averaged single-neuron correlation functions, which we validate against empirical data from simulations. In a next step, we use the correlation functions to investigate the influence of neuron- and network parameters on the timescale.

Author Contributions

All calculations were performed, the codebase written, and the figures created by the author (AvM) under supervision of Prof. Sacha van Albada (SvA). AvM wrote the first draft of the manuscript and it was revised jointly by AvM and SvA.

Microscopic theory of intrinsic timescales in spiking neural networks

Alexander van Meegen* and Sacha J. van Albada

Institute of Neuroscience and Medicine (INM-6) and Institute for Advanced Simulation (IAS-6) and JARA-Institut Brain Structure-Function Relationships (INM-10), Jülich Research Centre, 52425 Jülich, Germany and Institute of Zoology, University of Cologne, 50674 Cologne, Germany

(Received 2 July 2021; accepted 21 September 2021; published 28 October 2021)

A complex interplay of single-neuron properties and the recurrent network structure shapes the activity of cortical neurons. The single-neuron activity statistics differ in general from the respective population statistics, including spectra and, correspondingly, autocorrelation times. We develop a theory for self-consistent second-order single-neuron statistics in block-structured sparse random networks of spiking neurons. In particular, the theory predicts the neuron-level autocorrelation times, also known as *intrinsic timescales*, of the neuronal activity. The theory is based on an extension of dynamic mean-field theory from rate networks to spiking networks, which is validated via simulations. It accounts for both static variability, e.g., due to a distributed number of incoming synapses per neuron, and temporal fluctuations of the input. We apply the theory to balanced random networks of generalized linear model neurons, balanced random networks of leaky integrate-and-fire neurons, and a biologically constrained network of leaky integrate-and-fire neurons. For the generalized linear model network with an error function nonlinearity, a novel analytical solution of the colored noise problem allows us to obtain self-consistent firing rate distributions, single-neuron power spectra, and intrinsic timescales. For the leaky integrate-and-fire networks, we derive an approximate analytical solution of the colored noise problem, based on the Stratonovich approximation of the Wiener-Rice series and a novel analytical solution for the free upcrossing statistics. Again closing the system self-consistently, in the fluctuation-driven regime, this approximation yields reliable estimates of the mean firing rate and its variance across neurons, the interspike-interval distribution, the single-neuron power spectra, and intrinsic timescales. With the help of our theory, we find parameter regimes where the intrinsic timescale significantly exceeds the membrane time constant, which indicates the influence of the recurrent dynamics. Although the resulting intrinsic timescales are on the same order for generalized linear model neurons and leaky integrate-and-fire neurons, the two systems differ fundamentally: for the former, the longer intrinsic timescale arises from an increased firing probability after a spike; for the latter, it is a consequence of a prolonged effective refractory period with a decreased firing probability. Furthermore, the intrinsic timescale attains a maximum at a critical synaptic strength for generalized linear model networks, in contrast to the minimum found for leaky integrate-and-fire networks.

DOI: 10.1103/PhysRevResearch.3.043077

I. INTRODUCTION

Neural dynamics in the cerebral cortex of awake behaving animals unfolds over multiple timescales, ranging from milliseconds up to seconds and more [1–5]. Such a heterogeneity of timescales in the dynamics is a substrate for temporal processing of sensory stimuli [6] and reflects integration of information over different time intervals [3,4]. Intriguingly, *in vivo* electrophysiological recordings reveal a structure in the autocorrelation timescales of the activity on the level of single neurons [2,7]. This structure could arise from systematic variations in single-neuron or synaptic properties [8,9], from the intricate cortical network structure [10], or from a combination of both [11,12]. Furthermore, timescales may be influenced by the external input to the network, and depend on the chosen measurement procedure [13]. Thus, while these timescales are referred to as *intrinsic timescales*, they are shaped by intrinsic and extrinsic factors alike.

Explaining the timescales of individual neurons embedded in a network poses a theoretical challenge: How to account for a microscopic, neuron-level observable in a macroscopic theory? Clearly, a straightforward coarse-graining of the activity eliminates the microscopic observable of interest [14]. *Dynamic mean-field theory* (DMFT) [15–17] makes microscopic observables accessible because, instead of coarse-graining the activity of the neurons, it coarse-grains their input. Here, the term "dynamic" specifies that the input is approximated as a stochastic process that varies in time, in contrast to the notion of a mean-field theory in physics, which usually describes processes embedded in a constant field. DMFT has led to significant insights into the interrelation between network structure and intrinsic timescales for recurrent networks of (nonspiking) rate neurons [15–23]. In particular, it has been shown that very slow intrinsic timescales emerge close to

*avm@physik.hu-berlin.de

a transition to chaos in autonomous networks [15]. Interestingly, simply adding a noisy input to the network significantly reduces this effect and even leads to a novel dynamical regime [21]. Furthermore, increasing the complexity of the single-neuron dynamics reveals that timescales of slow adaptive currents are not straightforwardly expressed in the network dynamics [22], and leads to yet another dynamical regime termed "resonant chaos" [23]. In combination, these results suggest that the mechanisms shaping the intrinsic timescales in recurrent networks are highly involved.

A characteristic feature of neural communication in the brain is the spike-based coupling [24]: the output of a neuron is a stereotypical pulse, a spike, that is produced once the internal voltage exceeds a threshold and that travels along the axon to the target neurons. Consequently, spiking neural network models have already yielded notable insights into cortical neural dynamics. Prominent examples are the excitatory-inhibitory balance mechanism which dynamically generates strong fluctuations while keeping the activity in a physiological range [25,26] and the mechanism of recurrent inhibitory feedback leading to low cross-correlation between neurons despite the high number of shared inputs [27,28]. From a theoretical perspective, spike-based coupling further increases the complexity of the dynamics. This calls for an extension of DMFT to spiking networks. Following early works where slow synaptic dynamics reduced the spiking networks effectively to rate networks [18,29], this was recently achieved with a model-independent framework [30] (see also the pioneering work [31]).

Perhaps unintuitively, the main obstacle is not the reduction of the recurrent dynamics to the DMFT but the colored noise problem: to obtain the output statistics of the neuron for temporally correlated input statistics. Previous works relied on numerical methods to address the colored noise problem [31–36] because the spiking nonlinearity renders this problem in general analytically intractable. Such a self-consistent numerical scheme already revealed an unexpected minimum instead of a maximum in the intrinsic timescales for spiking networks at a critical coupling strength [37]. However, numerical solutions have the drawback that they lead to noisy estimates of the autocorrelation function, which poses additional challenges on the inference of intrinsic timescales [38] and other dynamical quantities from the neuronal and network parameters. In addition, such a self-consistent numerical scheme is computationally intensive.

In this paper, we use analytical approaches to close the self-consistency equations for spiking networks. First, we transfer the theory for rate networks to one for spiking networks starting from the characteristic functional of the recurrent input. This shows that the first two cumulants (mean and variance) of the connectivity matrix suffice to fully characterize the effective stochastic input, and automatically take the static variabilities (firing rate, indegree) in the network into account. Since it is based on DMFT, the resulting theory indeed accounts for the timescales on the microscopic level, orthogonal to approaches where the activity of a population of neurons is reduced to an effective mesoscopic description [39]. Second, we derive an analytical solution to the colored noise problem for *generalized linear model* (GLM) neurons with exponential and error function nonlinearity. Using these analytical solutions, we validate that the self-consistent DMFT captures both the static second-order statistics, the distribution of firing rates across neurons, and the dynamic second-order statistics, the population-averaged autocorrelation function. Furthermore, we use the theory to investigate the conditions for longer intrinsic timescales, like those observed in *in vivo* electrophysiological recordings [2,7], in a balanced random network of GLM neurons. Due to the analytical tractability, our theory exposes the factors that shape the intrinsic timescale. Third, we derive a numerically efficient analytical approximation for the colored noise problem for *leaky integrate-and-fire* (LIF) neurons in the noise-driven regime based on the Wiener-Rice series and the Stratonovich approximation thereof [40,41]. For a different approach based on a Markovian embedding, which leads to multidimensional Fokker-Planck equations with involved boundary conditions that are solved numerically, see [42]. In contrast, our approximation leads to integrals of which the computationally most involved ones can be solved analytically. Lastly, we use these results to explore the parameter space of a balanced random network of LIF neurons for long timescales, and apply the theory to a more elaborate model with population–specific connection probabilities that are constrained by biological data [43].

We start this manuscript with the derivation of the DMFT equations from the characteristic functional of the recurrent input. The remainder of the results is structured according to the neuron model. First we consider GLM neurons with exponential and error function nonlinearity, respectively, then we turn to LIF neurons. For each neuron model, we begin by deriving the solution or approximation of the colored noise problem. We then describe the numerical method to solve the self-consistent DMFT equations for the given neuron type (GLM or LIF). Subsequently, we use our theory to investigate the timescale in the respective network models.

II. MICROSCOPIC THEORY OF INTRINSIC TIMESCALES

We consider random network topologies where the entries of the matrix $\mathcal{J}$ containing the synaptic strengths, i.e., the amplitudes of evoked post-synaptic currents due to incoming spikes, are independent and identically distributed (i.i.d.). A synapse from neuron j to neuron i exists ($\mathcal{J}_{ij}$ is nonzero) with probability p; each nonzero entry J_{ij} is independently sampled from the distribution of synaptic strengths with mean μ_J and variance $\sigma_J^2 < \infty$:

$$\mathcal{J}_{ij} = \begin{cases} J_{ij} & \text{with probability } p \\ 0 & \text{with probability } 1-p \end{cases}. \tag{1}$$

The connectivity is thus taken to be pairwise Bernoulli, yielding maximally one synapse from a given presynaptic to a given postsynaptic neuron. To account for Dale's law and further heterogeneities, we subdivide the network into populations, e.g., all pyramidal cells in cortical layer V, consisting of statistically identical neurons and denote the population by a Greek superscript. Within this generalization, the entries of $\mathcal{J}$ are still i.i.d. random numbers for a given pair of populations α, β, but $p^{\alpha\beta}$ and the distribution of $J_{ij}^{\alpha\beta}$ can vary for different pairs of populations [Fig. 1(a)]. For example, if I denotes a population of inhibitory interneurons, all $J_{ij}^{\alpha I}$ are negative.

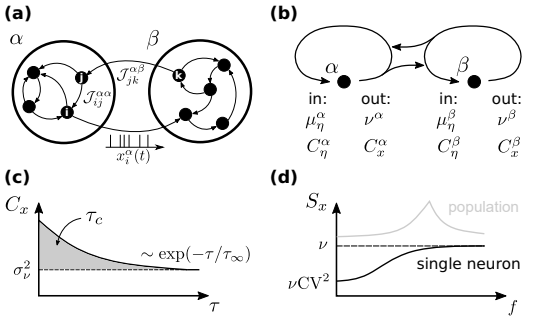

FIG. 1. Illustration of the theory. (a) We consider populations of randomly connected neurons (α, β) that communicate via spike trains $x_i^\alpha(t)$. The neurons of population β are connected to those of population α with connection probability $p^{\alpha\beta}$. (b) The theory reduces a population to a single neuron driven by an effective stochastic input η^α. The first- and second-order statistics μ_η^α and C_η^α of η^α depend self-consistently on the output statistics, ν^α and C_x^α. (c) From the stationary spike train autocorrelation function $C_x(\tau) = \nu\delta(\tau) + \hat{C}_x(\tau)$, we obtain the correlation time τ_c, the asymptotic decay τ_∞, and the variability of the rate across neurons, σ_ν^2. (d) Instead of the stationary autocorrelation function we sometimes consider the power spectrum $S_x(f)$, which saturates at the firing rate, $S_x(f) \overset{f\to\infty}{\to} \nu$, and, for a renewal process, has the zero-frequency limit $S_x(f) \overset{f\to 0}{\to} \nu \mathrm{CV}^2$. Throughout, we consider the population–averaged single-unit statistics (black curve) instead of the statistics of the population-averaged activity (gray curve).

In this manuscript, we focus on the situation where the average number of synapses per neuron, the indegree $K^{\alpha\beta} = p^{\alpha\beta}N^\beta$, is large: $K^{\alpha\beta} \gg 1$ due to a large number of presynaptic neurons $N^\beta \gg 1$ in combination with a moderate connection probability $p^{\alpha\beta}$ on the order of 10%, in agreement with the situation in cortical networks [44]. In line with the theory of balanced networks [45], we assume that neither single spikes are sufficient to cause firing nor coherent input from all presynaptic neurons is necessary. Moreover, we consider networks which are in an asynchronous irregular state exhibited by cortical networks of awake, behaving animals [46].

In the following, we first consider a single population for clarity because the generalization to multiple populations is straightforward.

A. Input statistics

Dynamic mean-field theory reduces the dynamics of the recurrent network to a set of self-consistent stochastic equations. Its core idea is to approximate the recurrent input

$$\eta_i(t) = \sum_{j=1}^{N} \mathcal{J}_{ij} x_j(t) \tag{2}$$

by independent Gaussian processes. In Eq. (2), and throughout this manuscript, $x_j(t) = \sum_n \delta(t - t_{j,n})$ denotes the spikes emitted at times $t_{j,n}$ by neuron j—the spike train of neuron j—which are the output of neuron j and contribute to the input of target neuron i. The sum in Eq. (2) extends over all N neurons, using that $\mathcal{J}_{ij} = 0$ for neurons that are not connected.

1. Gaussian process approximation

Here, we sketch the derivation to expose necessary conditions for the DMFT. For the full treatment of the problem, we refer to the model–independent DMFT developed in Ref. [30], which is applicable to spiking networks.

We start from the deterministic input Eq. (2) and derive its approximation as independent Gaussian processes. To this end, let us consider the characteristic functional of the recurrent input. Because $\eta_i(t)$ is a deterministic quantity, its distribution is a Dirac delta and its characteristic functional, defined by $\Phi_{\boldsymbol{\eta}}[\boldsymbol{u}(t)] = \langle \exp(i \int_0^T \boldsymbol{u}(t)^\intercal \boldsymbol{\eta}(t) dt) \rangle_{\boldsymbol{\eta}}$, is [40,47] (see also Appendix A, Eq. (A2))

$$\Phi_{\boldsymbol{\eta}}[\boldsymbol{u}(t)] = \exp\left(i \int_0^T \sum_{i,j=1}^{N} u_i(t) \mathcal{J}_{ij} x_j(t) dt \right). \tag{3}$$

In Eq. (3), $u_i(t)$ are arbitrary test functions; the derivatives of $\Phi_{\boldsymbol{\eta}}[\boldsymbol{u}(t)]$ with respect to the test functions evaluated at $u_i(t) = 0$ yield the moments of the recurrent input.

Now we assume that the dynamics of the system are, on a statistical level, very similar for any given realization of the connectivity, i.e., we assume that the system is self-averaging. Thus we can consider the average across realizations of $\mathcal{J}$ and neglect the dependence of the spike trains on the realization of $\mathcal{J}$ for this average. For the latter assumption, it is important to keep in mind that we consider the statistics of the entire network: while the spike train of a particular neuron is certainly highly correlated to the realization of the connectivity, self-averaging means that this does not hold for the statistics of the activity across the network. Put differently, the input to the neuron, and hence the neuron itself, "loses its identity" and becomes a statistical representative for an arbitrary neuron in the network.

Under these assumptions, the average of the characteristic functional is

$$\langle \Phi_{\boldsymbol{\eta}}[\boldsymbol{u}(t)] \rangle_{\mathcal{J}} \approx e^{i\langle \mathcal{J} \rangle \sum_{i,j=1}^{N} \int_0^T u_i(t) x_j(t) dt} \times e^{-\frac{1}{2}\langle \Delta \mathcal{J}^2 \rangle \sum_{i,j=1}^{N} (\int_0^T u_i(t) x_j(t) dt)^2},$$

where we used the independence of the $\mathcal{J}_{ij}$, their characteristic function $\langle \exp(ik_{ij}\mathcal{J}_{ij}) \rangle_{\mathcal{J}_{ij}} = \exp(i\langle \mathcal{J} \rangle k_{ij} - \frac{1}{2}\langle \Delta \mathcal{J}^2 \rangle k_{ij}^2 + \dots)$, and neglected the cumulants of $\mathcal{J}_{ij}$ beyond the second-order cumulant (the variance) $\langle \Delta \mathcal{J}^2 \rangle$. Due to the independence of the $\mathcal{J}_{ij}$, the expectation factorizes into a product $\prod_{i,j=1}^{N}$ which leads to the sum $\sum_{i,j=1}^{N}$ in the exponent. Within each factor, the first (second) cumulant leads to a linear (quadratic) term in the exponent. Next, we rewrite the square, $(\int_0^T u_i(t) x_j(t) dt)^2 = \int_0^T \int_0^T u_i(t) u_i(t') x_j(t) x_j(t') dt dt'$, and introduce the network-averaged auxiliary fields

$$\mu_\eta(t) = \langle \mathcal{J} \rangle \sum_{j=1}^{N} x_j(t), \tag{4}$$

$$C_\eta(t, t') = \langle \Delta \mathcal{J}^2 \rangle \sum_{j=1}^{N} x_j(t) x_j(t'). \tag{5}$$

Using the auxiliary fields, the characteristic functional factorizes, $\langle \Phi_{\boldsymbol{\eta}}[\boldsymbol{u}(t)]\rangle_{\mathcal{J}} \approx \prod_{i=1}^{N} \hat{\Phi}_{\eta}[u_i(t)]$, with the individual factors given by

$$\hat{\Phi}_{\eta}[u(t)] = e^{i\int_0^T u(t)\mu_\eta(t)dt - \frac{1}{2}\int_0^T\int_0^T u(t)C_\eta(t,t')u(t')dtdt'},$$

which is the characteristic functional of a Gaussian process with mean $\mu_\eta(t)$ and correlation function $C_\eta(t,t')$ [40,47] (see Appendix A, Eq. (A3)). The factorization $\langle \Phi_{\boldsymbol{\eta}}[\boldsymbol{u}(t)]\rangle_{\mathcal{J}} \approx \prod_{i=1}^{N} \hat{\Phi}_{\eta}[u_i(t)]$ implies that the approximate inputs described by $\hat{\Phi}_{\eta}[u(t)]$ are independent across neurons.

The above sketch of a derivation reveals multiple assumptions we make in the DMFT. First, we assumed self-averaging. This is a necessary assumption if one wants to derive a statement that generalizes beyond a given connectivity matrix to its statistics only. For a broad class of rate networks, one can show rigorously that the statistics of the activity across the network are indeed self-averaging by calculating the distribution of the empirical measure $\frac{1}{N}\sum_{i=1}^{N}\delta[y(t)-x_i(t)]$ across realizations of the connectivity [48,49]. Here, we check this assumption post–hoc by comparison of the theory with simulations for a single realization of the connectivity. Second, we implicitly assumed $\bar{g} := N\langle \mathcal{J}\rangle$ and $g^2 := N\langle \Delta\mathcal{J}^2\rangle$ do not scale with N such that the auxiliary fields remain finite for large networks. Using the mean number of inputs per neuron $K = pN$ and the properties of $\mathcal{J}$, we get

$$\bar{g} = K\mu_J, \quad g^2 = K\big(\sigma_J^2 + (1-p)\mu_J^2\big). \tag{6}$$

Third, we neglected higher cumulants of the input. Using the assumption $J_{ij} = O(1/\sqrt{K})$ leads to $\mu_J = O(1/\sqrt{K})$, $\sigma_J^2 = O(1/K)$ and thus $\bar{g} = O(\sqrt{K})$, $g^2 = O(1)$ as well as $O(1/\sqrt{K})$ for the neglected higher cumulants. Accordingly, in the regime $K \gg 1$, neglecting the contributions from higher cumulants, e.g., due to shot noise effects [35], is justified.

2. Self–consistency problem

Given these assumptions, the recurrent inputs $\eta_i(t)$ can be approximated by independent Gaussian processes, which leads to a coarse-grained description of the dynamics: since all inputs are statistically equivalent, the neurons become statistically equivalent as well and the system reduces to N independent, identical stochastic equations. For $N \gg 1$, we can replace the empirical averages in Eqs. (4) and (5) by ensemble averages such that we arrive at a set of self-consistency equations. This step can be made rigorous using the formalism of Ref. [30], see Eqs. (2) and (3) and Appendix 1 therein.

In the stationary state, the self-consistency equations are given by

$$\mu_\eta = \bar{g}\,\langle x\rangle_\eta, \quad C_\eta(\tau) = g^2\,\langle xx\rangle_\eta(\tau). \tag{7}$$

The averages $\langle x\rangle_\eta \equiv \nu$ and $\langle xx\rangle_\eta(\tau) - \nu^2 \equiv C_x(\tau)$ denote the mean (firing rate) and correlation function of the spike train produced by a neuron driven by the effective stochastic input $\eta(t)$. Since the input thereby appears on both the left-hand and the right-hand sides, this poses a self-consistency problem.

To recapitulate, DMFT approximates the input of a single neuron by an effective Gaussian process with self-consistent statistics [Fig. 1(b)]. Thus the description, albeit stochastic, is still on the level of individual neurons. These individual neurons driven by Gaussian processes form an ensemble with the same statistics across neurons as the original network. In particular, this means that population-averaged quantities, e.g., the autocorrelation function, but also distributions across the neurons, e.g., the distribution of the firing rate, can be computed from the DMFT.

3. Static contribution

The networks we consider are heterogeneous even within a population—each neuron potentially has a different number of presynaptic partners and thus also a different firing rate [50]. On a first glance, DMFT neglects this heterogeneity. However, Eqs. (7) in fact account for such static variabilities: on the right-hand side the second moment of the spike train appears instead of the correlation function. Rewriting $\langle xx\rangle_\eta(\tau) = C_x(\tau) + \nu^2$ reveals a first static component $g^2\nu^2$ of the variability of the effective input due to the firing rate of individual neurons. Moreover, $C_x(\tau\to\infty) \equiv \sigma_\nu^2$ potentially saturates on a plateau which accounts for the variability of the firing rate across neurons [Fig. 1(c)]. To make this explicit, we sometimes rewrite

$$\eta(t) = \zeta + \xi(t), \tag{8}$$

where ζ is a Gaussian random variable with $\mu_\zeta = \bar{g}\nu$, $\sigma_\zeta^2 = g^2(\nu^2+\sigma_\nu^2)$ and $\xi(t)$ a zero-mean Gaussian process with $C_\xi(\tau) = g^2(C_x(\tau) - \sigma_\nu^2)$.

B. Multiple populations

Using the expressions Eqs. (7) for a single population, we can straightforwardly generalize the theory to multiple populations. Due to the independence of the effective inputs in DMFT, both mean and correlation function are a simple sum over the contributions from all populations [18,51]:

$$\mu_\eta^\alpha = \sum_\beta \bar{g}^{\alpha\beta}\nu^\beta, \tag{9}$$

$$C_\eta^\alpha(\tau) = \sum_\beta (g^{\alpha\beta})^2\big(C_x^\beta(\tau) + (\nu^\beta)^2\big), \tag{10}$$

with the corresponding generalizations of Eqs. (6), $\bar{g}^{\alpha\beta} = K^{\alpha\beta}\mu_J^{\alpha\beta}$ and $(g^{\alpha\beta})^2 = K^{\alpha\beta}((\sigma_J^{\alpha\beta})^2 + (1-p^{\alpha\beta})(\mu_J^{\alpha\beta})^2)$. This leads to one stochastic equation per population [Fig. 1(b)]. As before, we can split the input into static and dynamic contributions, $\eta^\alpha(t) = \zeta^\alpha + \xi^\alpha(t)$.

1. External input

We take the sum $\sum_\beta$ to include external populations, e.g., excitatory neurons that drive the network dynamics with homogeneous Poissonian spike trains of rate ν^{ext}. In Eqs. (9) and (10), such an external Poisson input leads to a term $J^{\alpha,\text{ext}}\nu^{\text{ext}}$ and $(J^{\alpha,\text{ext}})^2\nu^{\text{ext}}\delta(\tau)$, respectively. If the network is driven by a constant external input, only Eq. (9) obtains an additional contribution μ_{ext}^α. An external zero-mean, stationary Gaussian process leads to an additional term $C_{\text{ext}}(\tau)$ in Eq. (10).

C. Output statistics

Approximating the input is only the first step. In a second step, the self-consistency problem has to be solved. To this end, the output statistics of a neuron driven by a non-Markovian Gaussian process have to be calculated. In other words, we need a solution for the colored noise problem. The full non-Markovian problem has to be considered because a Markovian approximation neglects the quantity of interest: the temporal correlations. For sufficiently simple rate neurons, the problem is analytically solvable [15,52]; the case of two spiking neuron models is discussed in the following sections. For the remainder of this section, let us assume that we are able to solve the colored noise problem to obtain a self-consistent solution of Eqs. (9) and (10).

1. Timescale

Given a self-consistent solution, we can calculate the intrinsic timescale from the spike-train autocorrelation function $C_x^\alpha(\tau)$. Since $C_x^\alpha(\tau)$ always contains a delta peak [40], we consider only the smooth part of the autocorrelation function $\hat{C}_x^\alpha(\tau) \equiv C_x^\alpha(\tau) - \nu^\alpha\delta(\tau)$. To characterize the timescale, we use the definition of Ref. [40] [Fig. 1(c)]:

$$\tau_c^\alpha = \int_0^\infty \left| \frac{\hat{C}_x^\alpha(\tau) - \hat{C}_x^\alpha(\infty)}{\hat{C}_x^\alpha(0) - \hat{C}_x^\alpha(\infty)} \right| d\tau. \tag{11}$$

Note that the definition of the autocorrelation time is not unequivocal. Other possible definitions include $\tau_c^\alpha = \int_{-\infty}^\infty |\frac{\hat{C}_x^\alpha(\tau)-\hat{C}_x^\alpha(\infty)}{\hat{C}_x^\alpha(0)-\hat{C}_x^\alpha(\infty)}|^2 d\tau$ [37] and $\tau_c^\alpha = \frac{\int_0^\infty \tau|\hat{C}_x^\alpha(\tau)-\hat{C}_x^\alpha(\infty)|d\tau}{\int_0^\infty |\hat{C}_x^\alpha(\tau)-\hat{C}_x^\alpha(\infty)|d\tau}$ [23]. We observed drastic differences between these definitions for empirical correlation functions directly obtained from the simulations. These differences are in part an artifact from the absolute value: the variance of the empirical estimate grows with τ [53]; due to the absolute value these fluctuations add up. The three functional forms carry with them different fluctuations, e.g., the squared fluctuations $|\frac{\hat{C}_x^\alpha(\tau)-\hat{C}_x^\alpha(\infty)}{\hat{C}_x^\alpha(0)-\hat{C}_x^\alpha(\infty)}|^2$ are typically much smaller than $|\frac{\hat{C}_x^\alpha(\tau)-\hat{C}_x^\alpha(\infty)}{\hat{C}_x^\alpha(0)-\hat{C}_x^\alpha(\infty)}| < 1$, and hence lead to different estimates. For theoretically predicted autocorrelations, the difference is less drastic and we choose Eq. (11) because it is the most simple definition. Due to this difficulty, we always use the theoretical prediction of the autocorrelation function to determine the timescale—after checking that it matches the empirical autocorrelation function well apart from fluctuations.

In addition to τ_c^α, we will also consider the asymptotic decay constant [Fig. 1(c)]

$$\hat{C}_x^\alpha(\tau) - \hat{C}_x^\alpha(\infty) \sim \exp\left(-\tau/\tau_\infty^\alpha\right), \tag{12}$$

because in special cases τ_∞^α directly follows from our theory. For a simple exponential autocorrelation function, the timescales in Eqs. (11) and (12) coincide. We work from the assumption that Eq. (12) is a good approximation to Eq. (11) and verify this assumption *post hoc*.

In the literature, there are even more definitions of intrinsic timescales than the ones mentioned above. For example, [2] assume an exponential correlation function and an offset, similar to Eq. (12) but for all time lags and not just asymptotically. In contrast, Ref. [54] determine the timescale by fitting a Lorentzian to the power spectrum after removing oscillatory components. Yet another approach, determining the half width at half maximum of the autocorrelation function, is advocated for in Ref. [55]. To avoid these ambiguities, we use the established definitions, Eqs. (11) and (12), from the stochastic processes literature.

Recently, two new approaches have been proposed to estimate the timescale directly from spiking data [38,56]. While both overcome important challenges, biases in the estimated timescale related to and independent of subsampling, respectively, we do not use them here because they rely on models which implicitly assume (a mixture of) exponential correlation functions: Ref. [56] assumes an autoregressive model and Ref. [38] a mixture of Ornstein-Uhlenbeck processes.

2. Spike train power spectrum

Instead of the autocorrelation function, we sometimes consider the spike train power spectrum [Fig. 1(d)]

$$S_x^\alpha(f) = \int_{-\infty}^\infty e^{2\pi i f\tau} C_x^\alpha(\tau) d\tau. \tag{13}$$

Due to the delta peak in the autocorrelation function, the power spectrum always saturates at the firing rate, $S_x^\alpha(f) \overset{f\to\infty}{\to} \nu^\alpha$. For a renewal process, the zero-frequency limit is $S_x^\alpha(f) \overset{f\to 0}{\to} \nu^\alpha \mathrm{CV}_\alpha^2$ [24], which directly reveals the coefficient of variation of the interspike-interval (ISI) distribution CV_α.

3. Comparison with simulations

In our theory, we consider disorder-averaged quantities and stationary processes. To compare the theory with a single simulation, we assume self-averaging in the sense that the activity distribution across neurons is approximately the same for each network realization. Since neurons with different indegrees have different disorder- and time-averaged inputs, in practice this means that we assume that neurons with comparable indegree have comparable activity statistics in each network realization.

The disorder averages preserve the static variability across neurons, as we consider the same connectivity statistics, and in particular the same indegree distribution, across realizations. Self-averaging works well when each neuron (or at least a sufficiently large proportion of neurons) receives input from a representative sample of the rest of the network.

Under stationarity, distributions across neurons of instantaneous rates at any given time point (but not of instantaneous rates across time points—which we do not consider here) equal distributions of time-averaged rates across neurons. To obtain the rate distributions from the simulations, we use time-averaged rates to reduce the variance of the corresponding estimates. Similarly, we use time averages to compute the single-neuron autocorrelation functions and power spectra.

We focus on the second-order statistics. Since first-order statistics, i.e., the firing rate, scale the power spectra and correlation function [24], we plot $S_x(f)/\nu$ and $C_x(f)/\nu^2$ to eliminate this trivial dependency. Note that a multiplicative factor does not influence the intrinsic timescale, Eq. (11).

III. GENERALIZED LINEAR MODEL NEURONS

First, we consider generalized linear model (GLM) neurons [24,57]. GLM neurons are stochastic model neurons that spike according to an inhomogeneous Poisson process at a rate determined by the synaptic input. Due to their simplicity, GLM neurons are frequently fitted to experimental data [24]; here we consider them because they are analytically tractable.

A. Neuron dynamics

Each neuron generates a spike train according to an inhomogeneous Poisson process with intensity (rate)

$$\lambda_i^\alpha(t) = c_1^\alpha \, \phi\big[c_2^\alpha\big(V_i^\alpha(t) - \theta^\alpha\big)\big], \tag{14}$$

where θ^α denotes the (soft) threshold, $\phi(V)$ is a smooth, nonnegative, monotonically increasing function, and $c_1^\alpha > 0$, $c_2^\alpha > 0$ are free parameters. The voltage is given by a linear filtering of the input

$$V_i^\alpha(t) = \int_{-\infty}^{\infty} \kappa^\alpha(t-s)\eta_i^\alpha(s - d^{\alpha\beta})ds, \tag{15}$$

where $d^{\alpha\beta}$ allows for a transmission delay. For all simulations, we choose a filter with a single exponential with time constant τ_m^α, which corresponds to post-synaptic currents in the form of delta spikes:

$$\kappa^\alpha(t) = \Theta(t)e^{-t/\tau_\mathrm{m}^\alpha}. \tag{16}$$

Here, $\Theta(t)$ denotes the Heaviside function ensuring causality of the filter. We rescale the synaptic weights $J_{ij}^{\alpha\beta}$ and the threshold θ^α using c_2^α such that $c_2^\alpha = 1$ throughout the rest of this section.

1. Colored noise problem

The effective stochastic input $\eta^\alpha(t)$ leads to stochastic voltage dynamics. Because the voltage is given by a convolution, the voltage becomes a Gaussian process with

$$\mu_V^\alpha = \bar{\kappa}^\alpha \mu_\eta^\alpha, \quad C_V^\alpha(\tau) = \int_{-\infty}^{\infty} \tilde{\kappa}^\alpha(\tau - s)C_\eta^\alpha(s)ds, \tag{17}$$

where the filter determines $\bar{\kappa}^\alpha = \int_{-\infty}^{\infty} \kappa^\alpha(t)dt$ and $\tilde{\kappa}^\alpha(t) = \int_{-\infty}^{\infty} \kappa^\alpha(s)\kappa^\alpha(s-t)ds$. For the single-exponential filter that we used in simulations, we have $\bar{\kappa}^\alpha = \tau_\mathrm{m}^\alpha$ and $\tilde{\kappa}^\alpha(t) = \frac{\tau_\mathrm{m}^\alpha}{2} e^{-|t|/\tau_\mathrm{m}^\alpha}$. Note that the transmission delay cancels in the stationary case considered here.

All cumulants of the resulting spike trains $x(t)$ can be obtained from their characteristic functional [40] [see Appendix A, Eq. (A8)]:

$$\Phi_x[u(t)] = \exp\left(\int_0^T (e^{iu(t)} - 1)\lambda(t)dt\right).$$

From here, we temporarily drop the population index for the sake of clarity. Averaging over realizations of the rates yields

$$\langle \Phi_x[u(t)]\rangle_\lambda \approx e^{\int_0^T (e^{iu(t)}-1)\mu_\lambda(t)dt} \times e^{\frac{1}{2}\int_0^T\int_0^T (e^{iu(t)}-1)C_\lambda(t,t')\left(e^{iu(t')}-1\right)dtdt'},$$

where $\mu_\lambda(t)$ denotes the mean of $\lambda(t)$, $C_\lambda(t,t')$ its correlation function, and we neglect terms of $O(u^3)$ since we are only interested in the first and second cumulants. Expanding also $e^{iu(t)} - 1$ to second order in $u(t)$, we can simply read off the stationary cumulants

$$\nu = \mu_\lambda, \quad C_x(\tau) = \mu_\lambda\delta(\tau) + C_\lambda(\tau), \tag{18}$$

in agreement with the result of Ref. [58].

We are left with the task of calculating the first two cumulants of $\lambda(t)$ from μ_V and $C_V(\tau)$, depending on the choice of the nonlinearity $\phi(V)$.

2. Exponential nonlinearity

First, we consider the commonly employed exponential nonlinearity [24]

$$\phi(V) = \exp(V). \tag{19}$$

Both cumulants are straightforward to obtain from the characteristic functional of the voltage. We have [see Appendix A, Eqs. (A4) and (A5)]

$$\begin{aligned}\langle\phi(V(t_1))\rangle_V &= \langle e^{\int_0^T V(t)\delta(t-t_1)dt}\rangle_V \\ &= e^{\mu_V + \frac{1}{2}C_V(0)}, \\ \langle\phi(V(t_1))\phi(V(t_2))\rangle_V &= \langle e^{\int_0^T V(t)[\delta(t-t_1)+\delta(t-t_2)]dt}\rangle_V \\ &= e^{2\mu_V + C_V(0) + C_V(t_2-t_1)},\end{aligned}$$

where we used the stationarity of V. Including the prefactor and the threshold from Eq. (14), we get

$$\mu_\lambda = c_1 \, \exp\big(\mu_V - \theta + \tfrac{1}{2}C_V(0)\big), \tag{20}$$

$$C_\lambda(\tau) = \mu_\lambda^2 \exp(C_V(\tau)) - \mu_\lambda^2. \tag{21}$$

From Eq. (21), it follows that $C_\lambda(\tau)$ has a static part as long as $C_V(\infty) > 0$. Since $C_\eta(\tau)$ contains a static part [see Eqs. (8) and (10)], $C_V(\tau)$ and hence $C_\lambda(\tau)$ and $C_x(\tau)$ indeed also contain a static contribution and saturate on a plateau.

Rate distribution. The rate distribution across neurons is lognormal because the (static) input distribution is Gaussian and the f-I curve is a simple exponential [50]. The theory yields the mean $\nu = c_1 \exp(\mu_V - \theta + \frac{1}{2}C_V(0))$ and variance $\sigma_\nu^2 = C_x(\infty) = \nu^2(e^{C_V(\infty)} - 1)$ of the firing rate. We note that we can obtain the same result from a constant input with mean $\tilde{\mu}_V = \mu_V - \theta + \frac{1}{2}C_V(0) - \frac{1}{2}C_V(\infty)$ and variance across neurons $\tilde{\sigma}_V^2 = C_V(\infty)$. Parameterized in terms of $\tilde{\mu}_V$ and $\tilde{\sigma}_V$, the firing rate distribution is thus

$$p(\nu) = \nu^{-1}\,\mathcal{N}\big(\ln(\nu/c_1)\,|\,\tilde{\mu}_V, \tilde{\sigma}_V^2\big) \tag{22}$$

with the normal distribution $\mathcal{N}(x\,|\,\mu,\sigma^2)$.

3. Error function nonlinearity

A drawback of the exponential function, Eq. (19), is that it allows for infinite rates. Thus we also consider the bounded nonlinearity

$$\phi(V) = \tfrac{1}{2}(1 + \mathrm{erf}(V/\sqrt{2})). \tag{23}$$

The integrals to determine the cumulants can be solved using the table [59] (details in Appendix B 1); the result

is

$$\mu_\lambda = \frac{c_1}{2}(1 + \mathrm{erf}(h/\sqrt{2})), \tag{24}$$

$$C_\lambda(\tau) = c_1\mu_\lambda - 2c_1^2 T(h, a(\tau)) - \mu_\lambda^2, \tag{25}$$

where we again suppressed the population index, abbreviated $h = \frac{\mu_V - \theta}{\sqrt{1+C_V(0)}}$ and $a(\tau) = (\frac{1+C_V(0)-C_V(\tau)}{1+C_V(0)+C_V(\tau)})^{1/2}$, and used Owen's T function $T(h,a) = \frac{1}{2\pi}\int_0^a dx\, \frac{e^{-\frac{1}{2}h^2(1+x^2)}}{1+x^2}$.

Rate distribution. Equivalent to the situation for the exponential nonlinearity, the input distribution across neurons is Gaussian. Again, we consider the equivalent static problem which, in this case, leads to $\tilde{\mu}_V = \frac{\mu_V - \theta}{\sqrt{1+C_V(0)-C_V(\infty)}}$ and $\tilde{\sigma}_V^2 = \frac{C_V(\infty)}{1+C_V(0)-C_V(\infty)}$. Parameterized in terms of $\tilde{\mu}_V$ and $\tilde{\sigma}_V$, the firing rate distribution is

$$p(\nu) = \frac{\mathcal{N}\big(\mathrm{probit}(\nu/c_1)\,|\,\tilde{\mu}_V, \tilde{\sigma}_V^2\big)}{c_1\,\mathcal{N}(\mathrm{probit}(\nu/c_1)\,|\,0,1)}, \tag{26}$$

where $\mathrm{probit}(x)$ denotes the inverse of the standard normal cumulative distribution, i.e., $\mathrm{probit}(\phi(V)) = V$, and we used $\phi'(V) = \mathcal{N}(V\,|\,0,1)$.

4. Numerical solution of the self-consistency problem

We solve the self-consistency problem using a fixed-point iteration [32,35]. To initiate the algorithm, we set $\nu^\alpha = \frac{1}{2}c_1^\alpha$ and $C_\lambda^\alpha(t) = 0$. Next, we determine the input statistics according to Eqs. (9) and (10); then we determine the voltage statistics according to (17). From the voltage statistics, we can obtain the statistics of the rate via Eqs. (20) and (21) [or Eqs. (24) and (25)]. Denoting the rate thus calculated as $\hat{\mu}_{\lambda,n+1}^\alpha$, we then update the rate statistics using incremental steps, $\mu_{\lambda,n+1}^\alpha = \mu_{\lambda,n}^\alpha + \varepsilon(\hat{\mu}_{\lambda,n+1}^\alpha - \mu_{\lambda,n}^\alpha)$ for the mean rate, and similarly for all entries of $C_\lambda^\alpha(t)$. The new firing rate statistics lead via (18) to new spike train statistics. Here, the small update step $\varepsilon < 1$ is crucial because otherwise the fixed-point iteration is numerically unstable. Now we iterate and generate new voltage statistics. With the incremental update and the initialization $\nu^\alpha = \frac{1}{2}c_1^\alpha$, the algorithm quickly converged to the fixed point corresponding to the simulation in the examples we considered. Due to the analytical solutions, the only bottleneck for the numerics is the convolution in Eq. (17), which can be solved efficiently using the fast Fourier transform [60]. Thus, even the parameter scans with 5000 points described in the following run on a laptop in less than two minutes.

B. Balanced random network

As a first application of the theory, we consider a balanced random network of excitatory and inhibitory GLM neurons. The network contains two populations [Fig. 2(a)], $\alpha \in \{E, I\}$, and it is driven by an excitatory external input which we incorporate into an effective threshold $\theta_{\mathrm{eff}} = \theta - \mu_{\mathrm{ext}}$. Here, we use a constant external input rather than a Poisson drive because we are particularly interested in finding long timescales, which might be hindered by the lack of temporal correlation of Poisson spike trains. However, the theory can straightforwardly be applied to Poisson input. Although four times more excitatory cells are present in the network, we typically place it in an inhibition-dominated regime by increasing the synaptic weights of the inhibitory neurons. As well known [26], this settles the network in the balanced state leading to asynchronous irregular activity of the neurons [see, e.g., Fig. 2(b)].

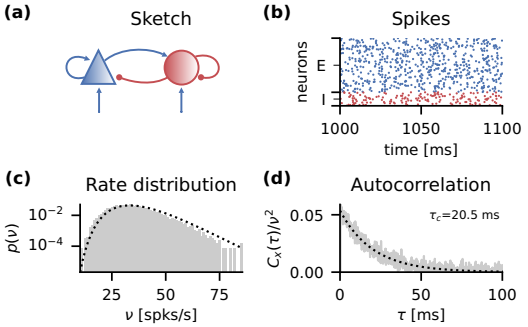

FIG. 2. Balanced random network of GLM neurons with exponential nonlinearity. (a) Sketch of the network with populations of excitatory (blue) and inhibitory (red) neurons. (b) Raster plot of 2% of the excitatory (blue) and inhibitory (red) neurons. (c) Firing rate distribution across all neurons from simulation (gray) and theory (black) using Eq. (22). (d) Population-averaged single-unit autocorrelation function from simulation (gray) and self-consistent theory (black) using Eqs. (20) and (21). Here, we subtracted the static contribution $C_x(\infty)$. Parameters: $N_E = 10000$, $N_I = 2500$, $J_E = 0.25$ mV, $|J_I/J_E| = 4.5$, $p = 0.1$, $\tau_m = 20$ ms, $\theta_{\mathrm{eff}} = 0$ mV, $c_1 = 50\ \mathrm{s}^{-1}$, $c_2 = 0.02\ \mathrm{mV}^{-1}$, and $d = 1.5$ ms.

In line with Brunel's model A [26], we choose identical values for the single-neuron parameters. Since we also choose the same connection probability of 10% for all pairs of populations, both populations receive statistically identical input in the DMFT approximation. Due to identical single-neuron parameters and input statistics, the statistics of the activity is the same for excitatory and inhibitory neurons [see, e.g., Fig. 2(b)]; therefore, we do not distinguish between the populations for the statistics in our plots. In contrast to the network examined by Brunel, we consider the somewhat more involved case of a fixed connection probability between a pair of neurons instead of a fixed number of incoming synapses per neuron (indegree). The fixed connection probability leads to a (binomially) distributed indegree across neurons, such that a strong variability across neurons is present in the network [see, e.g., Fig. 2(c)]. This variability is already present on the level of mean firing rates, i.e., there is static variability in the network.

All simulations were performed using the NEST simulator version 2.20.1 [61]. In all GLM network simulations, we simulated 1 min of biological time with a time step of 0.1 ms and discarded an initial transient of 1 s. For the GLM neurons, we used the "pp_psc_delta" neuron model. To allow for the error function nonlinearity, we modified the "pp_psc_delta" model accordingly.

1. Exponential nonlinearity: absence of long timescales

First, we consider networks with an exponential nonlinearity (Fig. 2). The fixed-point iteration yields a rate distribution

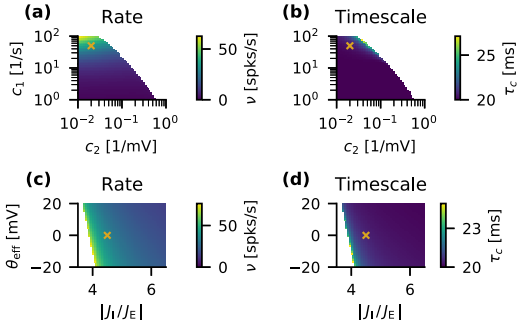

FIG. 3. Parameter scan for a balanced random network of GLM neurons with exponential nonlinearity. [(a) and (b)] Firing rate and intrinsic timescale for varying neuron parameters c_1 and c_2. Parameters used in (c) and (d) and Fig. 2 indicated by orange crosses. [(c) and (d)] Firing rate and intrinsic timescale for varying effective threshold θ_{eff} and relative inhibitory strength $|J_I/J_E|$. Parameters used in (a) and (b) and Fig. 2 indicated by orange crosses. Further parameters as in Fig. 2.

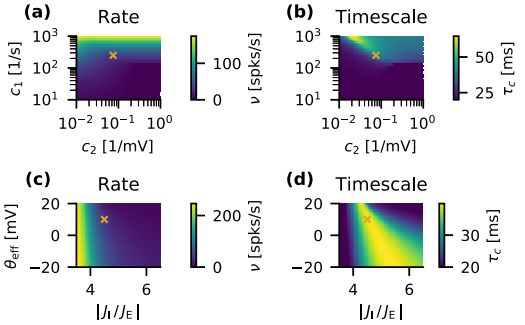

FIG. 4. Parameter scan for a balanced random network of GLM neurons with error function nonlinearity. [(a) and (b)] Firing rate and intrinsic timescale for varying neuron parameters c_1 and c_2. Parameters used in (c) and (d) and Fig. 5 indicated by orange crosses. [(c) and (d)] Firing rate and intrinsic timescale for varying effective threshold θ_{eff} and relative inhibitory strength $|J_I/J_E|$. Parameters used in (a) and (b) and Fig. 5 indicated by orange crosses. Further parameters as in Fig. 2.

and an autocorrelation function that closely match the simulation [Figs. 2(c) and 2(d)]. The theory for the rate distribution [Fig. 2(c)] is slightly biased towards higher rates; a possible cause for this is a finite size effect because the mean inhibitory indegree $K_I = pN_I = 250$ is relatively small. Nonetheless, the theory predicts the autocorrelation function very well [Fig. 2(d)] and yields a timescale $\tau_c \approx \tau_{\mathrm{m}} = 20\,\mathrm{ms}$.

For the parameters in Fig. 2, the intrinsic timescale is close to the membrane time constant. This raises the question whether longer timescales can be achieved in a network of GLM neurons. To answer this question, we employ our theory and perform parameter scans. First, we vary the single-neuron parameters c_1 and c_2 [Figs. 3(a) and 3(b)]. The rate increases monotonically with c_1 while c_2 has as smaller effect up to a certain threshold [Fig. 3(a)]. Beyond this threshold, the rate diverges rapidly to infinity in the threshold iteration [white area in Fig. 3(a)]. The timescale is close to the membrane time constant throughout the nondivergent regime and only increases slightly towards the threshold where the rate diverges [Fig. 3(b)]. Next, we vary the strength of the external input by adjusting the effective threshold θ_{eff} and the inhibition dominance by varying $|J_I/J_E|$ for constant J_E. We find a clear threshold of $|J_I/J_E|$ beyond which the rate diverges [Fig. 3(c)]. Again, this threshold corresponds to the regime where the timescale slowly starts to grow above the membrane time constant.

Put together, these observations suggest that the rate divergence prevents recurrent dynamics with long timescales in balanced random networks of GLM neurons with exponential nonlinearity.

2. *Error function nonlinearity: existence of long timescales*

In the previous section, the rate divergence prevented long timescales. To avoid the divergence, we consider the bounded transfer function Eq. (23) and use our theory for parameter scans (Fig. 4). The effect of the single-neuron parameters c_1 and c_2 is similar to the unbounded case but the rate divergence is absent [Fig. 4(a)]. This allows for a parameter regime with longer timescales up to approximately $3\tau_{\mathrm{m}}$ [Fig. 4(b)]. Similarly, varying θ_{eff} and $|J_I/J_E|$ uncovers a regime with a rate close to the maximum c_1 when the network is not inhibition-dominated [Fig. 4(c)]. Outside the inhibition-dominated regime, we expect that our theory does not yield quantitatively accurate predictions. The effect on the timescale is more subtle: within the inhibition-dominated regime, for any given $|J_I/J_E|$ the timescale displays a maximum whose location depends on the external input [Fig. 4(d)].

What kind of dynamics is displayed by the network at such a local maximum of the timescale? The corresponding spike trains show a strong variability of firing rate across neurons and temporally correlated spikes [Fig. 5(a)]. The rate distribution reveals that all rates between the minimum zero and the maximum c_1 are present, in excellent agreement with the theoretical prediction [Fig. 5(b)]. In the example considered, the empirical estimate of the network–averaged single-unit autocorrelation displays an intrinsic timescale of approximately $2\tau_{\mathrm{m}}$; again, the empirical estimate and the theoretical prediction agree closely [Fig. 5(c)]. From the spike train power spectrum, a high CV > 2 is apparent [Fig. 5(d)]. All of these characteristics agree with the "heterogeneous asynchronous state" uncovered in [62].

3. *Error function nonlinearity: mechanism of timescale*

To uncover the mechanisms that shape the timescale, in particular the local maximum in Fig. 4(d), we develop a theory for the asymptotic timescale τ_∞, Eq. (12). To this end, we use that $\tilde{\kappa}(t) = \frac{\tau_{\mathrm{m}}}{2} e^{-|t|/\tau_{\mathrm{m}}}$ is the fundamental solution to the differential operator $1 - \tau_{\mathrm{m}}^2 \frac{d^2}{dt^2}$, i.e., $(1 - \tau_{\mathrm{m}}^2 \frac{d^2}{dt^2})\tilde{\kappa}(t) = \tau_{\mathrm{m}}^2 \delta(t)$. Thus we can rewrite Eq. (17) into a differential equation:

$$\tau_{\mathrm{m}}^2 \ddot{C}_V = C_V - \tau_{\mathrm{m}}^2 C_\eta$$

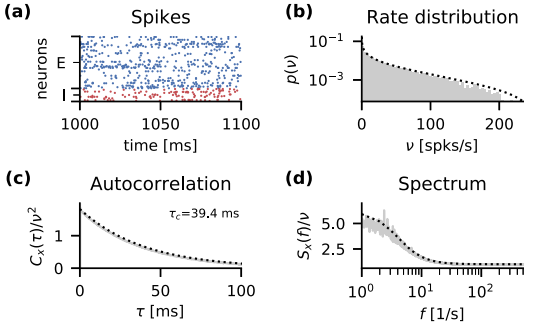

FIG. 5. Balanced random network of GLM neurons with error function nonlinearity. (a) Raster plot of 2% of the excitatory (blue) and inhibitory (red) neurons. (b) Firing rate distributions across all neurons from simulation (gray) and theory (black) using Eq. (26). [(c) and (d)] Population-averaged single-unit autocorrelation function and power spectrum from simulation (gray) and self-consistent theory (black) using Eqs. (24) and (25). As in Fig. 2, we subtracted the static contribution $C_x(\infty)$. Parameters: $c_1 = 250\ \mathrm{s}^{-1}$, $c_2 = 0.075\ \mathrm{mV}^{-1}$, and further parameters as in Fig. 2.

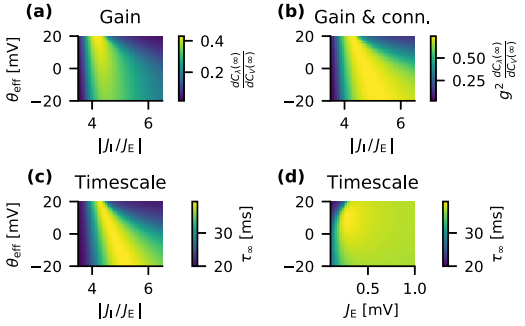

FIG. 6. Mechanisms that shape the asymptotic timescale. (a) Asymptotic gain $\frac{dC_\lambda(\infty)}{dC_V(\infty)}$ of the rate autocorrelation w.r.t. changes in the voltage autocorrelation, Eq. (28). (b) Asymptotic gain multiplied by the second cumulant of the connectivity, g^2. (c) Asymptotic timescale according to Eq. (27), $\tau_\infty = (1 - g^2 \frac{dC_\lambda(\infty)}{dC_V(\infty)})^{-1/2}$, for varying effective threshold θ_{eff} and relative inhibitory strength $|J_I/J_E|$. (d) Same as **c** for varying excitatory synaptic strength J_E with constant $|J_I/J_E|$. Further parameters as in Fig. 5.

where the dependence of C_η on C_V is determined by Eqs. (10), (18), and (25). Next, we rescale time such that $\tau_{\mathrm{m}} = 1$ and linearize this differential equation for small $\Delta_V(\tau) \equiv C_V(\tau) - C_V(\infty)$ to obtain

$$\ddot{\Delta}_V = \left(1 - \frac{dC_\eta(\infty)}{dC_V(\infty)}\right)\Delta_V + O(\Delta_V^2).$$

This allows for an exponential solution with time constant

$$\tau_\infty = \frac{1}{\sqrt{1 - g^2 \frac{dC_\lambda(\infty)}{dC_V(\infty)}}} \tag{27}$$

where we used Eqs. (10) and (18) to derive $\frac{dC_\eta(\infty)}{dC_V(\infty)} = g^2 \frac{dC_\lambda(\infty)}{dC_V(\infty)}$. We see that there are two factors that determine the timescale: the cumulant of the connectivity g^2 and the gain of the rate autocorrelation $\frac{dC_\lambda(\infty)}{dC_V(\infty)}$. For the latter, we obtain from Eq. (25)

$$\frac{dC_\lambda(\infty)}{dC_V(\infty)} = \frac{c_1^2}{2\pi} \frac{\exp\left(-\frac{(\mu_V - \theta_{\mathrm{eff}})^2}{1 + C_V(0) + C_V(\infty)}\right)}{\sqrt{(1 + C_V(0))^2 - C_V(\infty)^2}}. \tag{28}$$

Thus, given a self-consistent autocorrelation C_x and the corresponding voltage statistics from Eq. (17), the asymptotic timescale Eq. (27) can be directly evaluated.

We vary θ_{eff} and $|J_I/J_E|$ in Figs. 6(a)–6(c). First, we plot $\frac{dC_\lambda(\infty)}{dC_V(\infty)}$ alone, which we refer to as the gain [Fig. 6(a)]. Due to the interplay between the exponential suppression $\exp(-\frac{(\mu_V - \theta_{\mathrm{eff}})^2}{1 + C_V(0) + C_V(\infty)})$ and the square root factor $1/\sqrt{(1 + C_V(0))^2 - C_V(\infty)^2} < 1$ in Eq. (28), the gain already exhibits a maximum. The existence of the maximum is mainly determined by the exponential suppression with growing $|\mu_V - \theta_{\mathrm{eff}}|$ in Eq. (28): in both the excitation- and the inhibition-dominated regimes, μ_V is far from the effective threshold θ_{eff}. The precise location of the maximum is not necessarily at $\mu_V = \theta_{\mathrm{eff}}$ as it is also determined by the square root factor. The latter decays reciprocally to $C_V(0)$ and $C_V(\infty)$. Both $C_V(0)$ and $C_V(\infty)$ decay for growing effective threshold and inhibition dominance, which results in a larger square root factor that shifts the maximum towards the upper right and broadens it. The cumulant of the connectivity g^2 grows with $|J_I/J_E|^2$, which further broadens the region of the maximum [Fig. 6(b)]. The resulting asymptotic timescale [Fig. 6(c)] agrees both qualitatively and quantitatively with the intrinsic timescale [Fig. 4(d)]. This is likely due to the single-exponential shape of the autocorrelation function [Fig. 5(c)].

To investigate the interplay of the gain and the connectivity further, we vary the overall synaptic strengths by varying the excitatory weight J_E while keeping $|J_I/J_E|$ fixed at an inhibition-dominated value [Fig. 6(d)]. Increasing J_E in the inhibition-dominated regime shifts μ_V away from the effective threshold and decreases the gain; conversely g^2 grows with J_E^2. This interplay leads to a broad region in parameter space with an increased timescale. However, the exponential decrease of the gain is more pronounced than the quadratic increase of g^2 such that the asymptotic timescale does not continue to grow with J_E but saturates. Thus, although increasing J_E goes together with increased variability across neurons as in the "heterogeneous asynchronous state" described by Ostojic [62], this does not map systematically onto longer single-neuron timescales.

4. Error function nonlinearity: external timescale

Our theory allows arbitrary Gaussian processes as external input. To investigate the influence of an external timescale on the intrinsic timescale, we choose a zero-mean Ornstein-Uhlenbeck process with

$$C_{\mathrm{ext}}(\tau) = \frac{\sigma_{\mathrm{ext}}^2}{\tau_{\mathrm{m}}} \left(\frac{1}{\tau_{\mathrm{m}}} + \frac{1}{\tau_{\mathrm{ext}}}\right) e^{-|\tau|/\tau_{\mathrm{ext}}}. \tag{29}$$

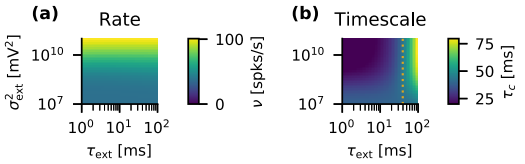

FIG. 7. Influence of colored external input. [(a) and (b)] Firing rate and intrinsic timescale for varying strength $\sigma^2_{\rm ext}$ and timescale $\tau_{\rm ext}$ of an external Ornstein-Uhlenbeck process. Orange line in (b) indicates the intrinsic timescale without external input. Parameters as in Fig. 5.

Here, the scaling factors ensure that the external timescale does not influence the resulting variance of the voltage, $C_V(0) = \int_{-\infty}^{\infty} \tilde{\kappa}(s) C_{\rm ext}(s) ds = \sigma^2_{\rm ext}$ for $\tilde{\kappa}(t) = \frac{1}{2}\tau_{\rm m} e^{-|t|/\tau_{\rm m}}$.

We take the parameters from Fig. 5 where the intrinsic timescale is maximal in the absence of external input. Increasing the strength of the external input $\sigma^2_{\rm ext}$ leads to an increased firing rate [Fig. 7(a)]. As desired, by construction of Eq. (29), the external timescale has a negligible effect on the firing rate at constant $\sigma^2_{\rm ext}$ [Fig. 7(a)]. The effect of the external timescale on the intrinsic timescale is highly intuitive: If $\tau_{\rm ext}$ is smaller than the intrinsic timescale without external input it decreases the intrinsic timescale, and vice versa [Fig. 7(b)]. The strength of this effect grows with the strength of the external input. In the limit of strong external input, the intrinsic timescale approaches the external timescale if $\tau_{\rm ext} > \tau_{\rm m}$; if $\tau_{\rm ext} < \tau_{\rm m}$ the intrinsic timescale approaches the minimum set by the membrane time constant.

IV. LEAKY INTEGRATE-AND-FIRE NEURONS

Considering GLM neurons is a convenient choice due to their analytical tractability. However, their intrinsic stochasticity might fundamentally alter the network dynamics. Thus we consider the frequently used leaky integrate–and–fire neuron model in this section [24]. The synapses are taken to be current-based with an exponential time course. An analytical solution to the colored noise problem for LIF neurons is an open challenge. Here, we focus on the fluctuation-driven regime and employ an approach based on the Wiener–Rice series [41,63,64] and the Stratonovich approximation thereof [40,41]. Below, we briefly introduce both the Wiener–Rice series and its Stratonovich approximation. For a comprehensive and pedagogic introduction to this approach, in particular with a focus on LIF neurons, see Ref. [65] where the approach is used to investigate LIF neurons driven by nonstationary input.

A. Neuron dynamics

The dynamics of individual neurons are governed by

$$\tau^\alpha_{\rm m} \dot{V}^\alpha_i(t) = -V^\alpha_i(t) + I^\alpha_i(t), \tag{30}$$

$$\tau^\alpha_{\rm s} \dot{I}^\alpha_i(t) = -I^\alpha_i(t) + \tau^\alpha_{\rm m} \eta^\alpha_i(t - d^{\alpha\beta}), \tag{31}$$

where V^α_i denotes the membrane voltage, I^α_i the synaptic current, $\tau^\alpha_{\rm m/s}$ the membrane/synaptic time constant, and the voltage is reset to $V^\alpha_{\rm r}$ and held constant during the refractory period $\tau^\alpha_{\rm ref}$ whenever it reaches the threshold θ^α. Threshold crossing triggers a spike which arrives at another neuron after a delay $d^{\alpha\beta}$. We set the resting potential to zero without loss of generality and absorb the membrane resistance into the synaptic current.

1. Effective stochastic dynamics

The effective stochastic input with statistics governed by Eqs. (9) and (10) leads to a stochastic current with

$$\mu^\alpha_I = \tau^\alpha_{\rm m} \mu^\alpha_\eta, \tag{32}$$

$$C^\alpha_I(\tau) = \left(\frac{\tau^\alpha_{\rm m}}{\tau^\alpha_{\rm s}}\right)^2 \int_{-\infty}^{\infty} \tilde{\kappa}^\alpha(\tau - s) C^\alpha_\eta(s) ds, \tag{33}$$

where $\tilde{\kappa}^\alpha(t) = \frac{\tau^\alpha_{\rm s}}{2} e^{-|t|/\tau^\alpha_{\rm s}}$, similar to Eq. (17). Contrary to the GLM neurons, the voltage cannot become a stationary process for LIF neurons due to the fire-and-reset rule. To circumvent this problem, we use the Wiener–Rice series which relates the free process without reset to the spiking statistics.

2. Wiener–Rice series and Stratonovich approximation

We consider a LIF neuron after the refractory period and the voltage dynamics that results if we do not allow for another fire-and-reset. We denote this free voltage $U(t)$. Moreover, we temporarily neglect the static contribution to the input variability and drop the population index. The process starts at $U(0) = V_{\rm r}$ and produces a system of random points $\{t_i\}$ defined by the upcrossings $U(t_i) = \theta, \dot{U}(t_i) > 0$. For this system of random points, the probability that no point falls in the interval $[0, T]$, i.e., the survival probability, is given by [40]

$$S(T) = \exp\left(\sum_{s=1}^{\infty} \frac{(-1)^s}{s!} \int_0^T \cdots \int_0^T g_s(t_1, \ldots, t_s) dt_1 \ldots dt_s\right),$$

where the $g_s(t_1, \ldots, t_s)$ are related to the free upcrossing probabilities $n_s(t_1, \ldots, t_s)$ calculated below, similar to the relation between moments and cumulants. For example, $g_1(t_1) = n_1(t_1)$ and $g_2(t_1, t_2) = n_2(t_1, t_2) - n_1(t_1) n_1(t_2)$. Now we approximate the output process as a renewal process such that the survival probability is sufficient to describe the statistics. Instead of the survival probability, it is more convenient to consider the cumulative hazard $H(T) = -\ln S(T)$ [24], i.e.,

$$H(T) = \sum_{s=1}^{\infty} \frac{(-1)^{s-1}}{s!} \int_0^T \cdots \int_0^T g_s(t_1, \ldots, t_s) dt_1 \ldots dt_s.$$

This can be regarded as a resummation of the Wiener–Rice series in terms of the $g_s(t_1, \ldots, t_s)$ instead of the free upcrossing probabilities $n_s(t_1, \ldots, t_s)$ [41].

Calculating the free upcrossing probabilities $n_s(t_1, \ldots, t_s)$, and thus the $g_s(t_1, \ldots, t_s)$, is tedious. To avoid this difficulty, Stratonovich proposed the approximation [40]

$$H_S(T) = -\int_0^T n_1(t) \frac{\ln\left(1 - \int_0^T Q(t, t') n_1(t') dt'\right)}{\int_0^T Q(t, t') n_1(t') dt'} dt, \tag{34}$$

where $Q(t_1, t_2) = 1 - \frac{n_2(t_1, t_2)}{n_1(t_1) n_1(t_2)}$. Briefly, to derive this approximation, the $g_s(t_1, \ldots, t_s)$ for $s \geqslant 3$ are expressed in terms of $n_1(t_1)$ and $Q(t_1, t_2)$ such that both the symmetry of the time arguments $t_1, \ldots, t_s$ and the equal-time limit $g_s(t_1, \ldots, t_1) =$

$(-1)^{s-1}(s-1)!\,n_1(t_1)^s$ are fulfilled; the resulting approximated $g_s(t_1,\ldots,t_s)$ are inserted into $H(T)$, which leads to a series that can be evaluated and yields Eq. (34). The condition $g_s(t_1,\ldots,t_1)=(-1)^{s-1}(s-1)!\,n_1(t_1)^s$ holds for a system of nonapproaching points where $n_s(t_1,\ldots,t_1)=0$ for $s\geqslant 2$, hence Eq. (34) is an approximation constructed for such a system. Although this seems intuitively reasonable because the voltage dynamics is continuous and differentiable, this condition is violated for LIF neurons with exponential postsynaptic currents [65]. Nonetheless, it yields good results, as shown in the following.

A much simpler alternative to the Stratonovich approximation would be to set $g_s(t_1,\ldots,t_s)=0$ for $s\geqslant 2$, leading to $H(T)=\int_0^T n_1(t)dt$. This approximation is sometimes referred to as the Hertz approximation. In particular, the Hertz approximation leads to a closed expression for the hazard function $h(t)\equiv\frac{d}{dt}H(t)=n_1(t)$. Unfortunately, this approximation is too severe and strongly affects the resulting firing rate. The main difference between the two approximations is the asymptotic saturation of the hazard function. Thus we employ an approximation suggested by Stratonovich for long times [40]: $\int_0^T Q(t,t')n_1(t')dt'\approx n_0\int_0^\infty Q(t,t')dt'\approx n_0\eta$ with $n_0=\lim_{t\to\infty}n_1(t)$ and $\eta=\lim_{t\to\infty}\int_0^\infty Q(t,t')dt'$. Inserting this approximation into Eq. (34) leads to

$$h_S(t)=\frac{\kappa_S}{n_0}n_1(t),\quad \kappa_S=-\frac{1}{\eta}\ln(1-n_0\eta). \tag{35}$$

Equation (35) combines the simplicity of the Hertz approximation with the asymptotic behavior of the Stratonovich approximation. The asymptotic level is given by $\lim_{t\to\infty}h_S(t)=\kappa_S$; to leading order in η we have $\kappa_S=n_0+O(\eta)$, which recovers the Hertz approximation. In the parameter regime we consider, Eq. (35) yields very similar results to Eq. (34) (see Appendix D). In all figures in the main text, we use Eq. (35). Since we approximate the output spike train as a renewal process, the hazard function Eq. (35) fully describes its statistics [24].

From the hazard function, we obtain the firing rate

$$\nu^{-1}=\int_0^\infty e^{-\int_0^T h(t)dt}dT \tag{36}$$

as well as the interspike-interval distribution [24]

$$p(T)=h(T)e^{-\int_0^T h(t)dt}. \tag{37}$$

From the Fourier transform of the interspike-interval distribution $\tilde{p}(f)=\int_0^\infty e^{2\pi i f T}p(T)dT$, we obtain the spike-train power spectrum using [40]

$$S_x(f)=\nu\frac{1-|\tilde{p}(f)|^2}{|1-\tilde{p}(f)|^2}. \tag{38}$$

Thus we are left with the task of calculating $n_1(t_1)$ and $Q(t_1,t_2)$.

3. Free upcrossing probabilities

The free voltage dynamics are governed by Eq. (30)

$$\tau_{\rm m}\dot{U}(t)=-U(t)+I(t),$$

where I is a Gaussian process determined by Eqs. (32) and (33), and the initial condition is $U(0)=V_{\rm r}$. U is a nonstationary Gaussian process due to the initial condition. For a sufficiently smooth Gaussian process, the upcrossing probability is given by the Kac–Rice formulas [40,63,66]

$$n_1(t)=\int_0^\infty \dot{U}_1 p(\theta,\dot{U}_1\,|\,V_{\rm r},\dot{U}_0)d\dot{U}_1,$$

$$n_2(t_1,t_2)=\int_0^\infty\int_0^\infty \dot{U}_2\dot{U}_1 p(\theta,\dot{U}_2;\theta,\dot{U}_1\,|\,V_{\rm r},\dot{U}_0)d\dot{U}_1 d\dot{U}_2,$$

where $p(\theta,\dot{U}_1\,|\,V_{\rm r},\dot{U}_0)$ denotes the probability that the process is at the threshold after time t and has velocity $\dot{U}_1$ given that it started at the reset at $t=0$ with velocity $\dot{U}_0$. Similarly, $p(\theta,\dot{U}_2;\theta,\dot{U}_1\,|\,V_{\rm r},\dot{U}_0)$ denotes the joint probability to be at the threshold at t_1 and t_2 with velocities $\dot{U}_1$ and $\dot{U}_2$. All integrals are over positive velocities only, because we consider upcrossings.

In both equations, we need to specify the distribution of the initial velocity $\dot{U}_0$. Here, it is important to take into account the biased sampling of the initial velocity [67]: at $-\tau^\alpha_{\rm ref}$, the neuron spiked due to an increased input current; hence, the initial velocity $\tau_{\rm m}\dot{U}_0=-V_{\rm r}+I_0$ is likely to be larger than for an I_0 drawn from the stationary current distribution. To keep the integral in Eq. (39) tractable, we assume that I_0 is Gaussian-distributed. To determine the mean and variance of this distribution, we use that the velocity of a stationary process at an upcrossing is Rayleigh-distributed [40] (details in Appendix C).

For $n_2(t_1,t_2)$, we consider only the stationary two-point upcrossing probability, so that it becomes a function of the time difference t_2-t_1 and loses the dependency on the initial velocity. After marginalizing the initial velocity in $n_1(t)$, we obtain

$$n_1(t)=\int_0^\infty \dot{U}_1 p(\theta,\dot{U}_1\,|\,V_{\rm r})d\dot{U}_1, \tag{39}$$

$$n_2(t_2-t_1)=\int_0^\infty\int_0^\infty \dot{U}_2\dot{U}_1 p(\theta,\dot{U}_2;\theta,\dot{U}_1)d\dot{U}_1 d\dot{U}_2, \tag{40}$$

where $n_2(\tau)$ leads to a stationary $Q(\tau)=1-\frac{n_2(\tau)}{n_0^2}$. This makes the integrals in Eq. (34) considerably easier to solve numerically (details in Appendix D).

Since the free dynamics are linear, $p(\theta,\dot{U}_1\,|\,V_{\rm r})$ and $p(\theta,\dot{U}_2;\theta,\dot{U}_1)$ can be obtained analytically. Importantly, the integral in Eq. (39) as well as the double integral in Eq. (40) are analytically solvable using the table [59] (details in Appendixes B 2 and C). The closed-form analytical expression Eq. (B6) for the two-point upcrossing probability of a stationary Gaussian process is a novel result, to the best of our knowledge, and considerably simplifies the numerical evaluation of Eq. (35).

4. Numerical solution of the self-consistency problem

Just as for the GLM networks, we solve the colored noise problem using a fixed-point iteration. To initiate the algorithm, we set the rates to $\nu^\alpha=1/\tau^\alpha_{\rm m}$. We use these rates to calculate the input mean, variance, and spectrum according to Eqs. (9) and (10), beginning with the diffusion approximation $S^\alpha_x(t)=\nu^\alpha$ and $\sigma^\alpha_v=0$ across neurons. Despite assuming initially equal rates across neurons, it is possible to have static input variability both due to distributed indegrees [see Eq. (8)

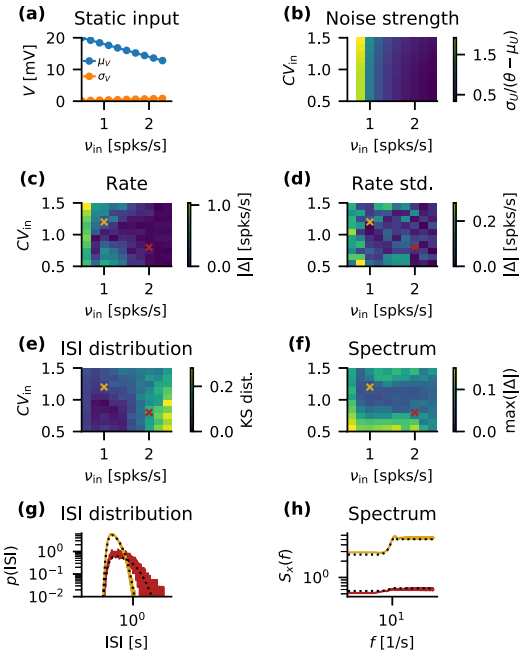

FIG. 8. Colored noise problem for LIF neurons. Comparison between theory, Eq. (35), and LIF neurons driven by Gaussian processes (GPs). (a) Mean (blue) and standard deviation across neurons (orange) of the membrane potential due to the static contribution $\tau_m \zeta$. (b) Noise strength of the effective input measured by the standard deviation of the membrane potential fluctuations relative to the distance to threshold $\sigma_U/(\theta - \mu_U)$. (c) Absolute difference $|\Delta|$ between rate from theory and GP-driven LIF neurons. (d) Same as (c) but for the standard deviation of the rate across neurons. (e) Kolmogorov-Smirnov distance using 2.5-ms bins between ISI distribution from theory and GP-driven LIF neurons. (f) Maximal absolute distance $\max(|\Delta|)$ between power spectra from theory and GP-driven LIF neurons. [(g) and (h)] Example ISI distributions and power spectra from theory (black) and GP-driven LIF neurons (colored) for the parameter values indicated by crosses in (c)–(f). Parameters: $N_E = 40\,000$, $N_I = 10\,000$, $J_E = 0.1$ mV, $|J_I/J_E| = 6.0$, $p = 0.1$, $\tau_m = 20$ ms, $\tau_s = 5$ ms, $\tau_{ref} = 2$ ms, $d = 1.5$ ms, $\theta = 20$ mV, $V_r = 0$ mV, and $\mu_{ext} = 22$ mV.

and Fig. 8(a)] and due to evolution of the rates during the fixed-point iteration. To account for the static variability, we consider an ensemble of inputs $\mu^\alpha + \zeta^\alpha$ and determine the corresponding hazard functions $h_S^\alpha(t \mid \mu^\alpha + \zeta^\alpha)$, Eq. (35), output rates $\nu^\alpha(\mu^\alpha + \zeta^\alpha)$, Eq. (36), ISI distributions $p^\alpha(T \mid \mu^\alpha + \zeta^\alpha)$, Eq. (37), and spectra $S_x^\alpha(f \mid \mu^\alpha + \zeta^\alpha)$, Eq. (38). From this ensemble, we obtain the final output statistics from a numerical average over the ensemble:

$$\nu^\alpha = \int_{-\infty}^{\infty} \nu^\alpha(\mu^\alpha + \zeta^\alpha) \mathcal{N}\big(\zeta^\alpha \mid 0, \sigma_\zeta^\alpha\big)\, d\zeta^\alpha, \tag{41}$$

$$\big(\sigma_\nu^\alpha\big)^2 = \int_{-\infty}^{\infty} [\nu^\alpha(\mu^\alpha + \zeta^\alpha) - \nu^\alpha]^2 \mathcal{N}\big(\zeta^\alpha \mid 0, \sigma_\zeta^\alpha\big)\, d\zeta^\alpha, \tag{42}$$

$$p^\alpha(T) = \int_{-\infty}^{\infty} p^\alpha(T \mid \mu^\alpha + \zeta^\alpha) \mathcal{N}\big(\zeta^\alpha \mid 0, \sigma_\zeta^\alpha\big)\, d\zeta^\alpha, \tag{43}$$

$$S_x^\alpha(f) = \int_{-\infty}^{\infty} S_x^\alpha(f \mid \mu^\alpha + \zeta^\alpha) \mathcal{N}\big(\zeta^\alpha \mid 0, \sigma_\zeta^\alpha\big)\, d\zeta^\alpha. \tag{44}$$

We solve the above Gaussian integrals using Gauss-Hermite quadrature [60]. Gauss-Hermite quadrature of order k solves Gaussian integrals of polynomials up to power k exactly by construction. This allows us to keep the ensemble very small; throughout we use $k = 5$. Finally, we update the statistics using incremental steps, e.g., $\nu_{n+1}^\alpha = \nu_n^\alpha + \varepsilon(\hat{\nu}_{n+1}^\alpha - \nu_n^\alpha)$ for the firing rate, where $\hat{\nu}_{n+1}^\alpha$ denotes the estimated rate based on the input at the previous step. Here, the small update step $\varepsilon < 1$ is crucial because otherwise the algorithm is numerically unstable. Now we iterate and generate new input statistics. Repeated application of this scheme suggests that the self-consistent problem for the type of networks under consideration possesses only a single fixed point to which the algorithm always converges.

B. Balanced random network

First, we consider the same balanced random network as we did for the GLM neurons [Fig. 2(a)]. In particular, we place the network in the inhibition-dominated regime, drive the network with a constant external input, and use identical single-neuron parameters for excitatory and inhibitory neurons. In order to obtain a biologically plausible activity below 10 spks/s, we keep the external input weak to place the network deep in the fluctuation-driven regime. In this regime, the mean input to a neuron is far below threshold and only occasional large fluctuations in the input drive it above the spike threshold [Figs. 8(a) and 8(b)]. If the mean interspike interval exceeds the correlation time of the input, the renewal approximation is admissible. Indeed, since the firing rates are low by construction, even moderate input correlation times are smaller than the inverse firing rate.

1. Colored noise problem

First, we isolate the colored noise problem to gauge the above approximations. To this end, we compare the theory with a population of unconnected LIF neurons driven by independent Gaussian processes (GPs). If the colored noise solution works well for isolated GP-driven LIF neurons, it will also work well for LIF neurons embedded in a balanced random network in the asynchronous irregular regime [35]. The reason for considering a population of neurons is to account for the static input variability that leads to distributed single-neuron firing rates.

We want to investigate the LIF neurons in a regime comparable to that in the balanced random network. However, we do not determine the effective input statistics using network simulation results here, because this would preclude a systematic scan over the parameters of the input, which consists of both external and recurrent network contributions. Instead, we fix the effective external input and determine the statistics of the effective recurrent input in terms of the input spiking statistics

ν_{in}, $\sigma_\nu^{\text{in}} = 0$ across neurons, and

$$S_x^{\text{in}}(f) = \nu_{\text{in}} \frac{1 - \left|\left(1 - 2\pi i \text{CV}_{\text{in}}^2 f/\nu_{\text{in}}\right)^{-1/\text{CV}_{\text{in}}^2}\right|^2}{\left|1 - \left(1 - 2\pi i \text{CV}_{\text{in}}^2 f/\nu_{\text{in}}\right)^{-1/\text{CV}_{\text{in}}^2}\right|^2}, \tag{45}$$

corresponding to a gamma process with rate ν_{in} and CV of the ISI distribution CV_{in}, cf. Eq. (38). This leaves a two-dimensional parameter space spanned by ν_{in} and CV_{in}. From the spiking statistics, we obtain the statistics of the effective input using Eqs. (9) and (10) where $\bar{g}$ and g are determined by the parameters of the balanced random network. Note that although $\sigma_\nu^{\text{in}} = 0$, the static variability of the effective input is nonzero, $\sigma_\zeta > 0$, due to the distributed indegree, see Eq. (8) and Fig. 8(a). Hence, we can compare both the averaged output statistics and the rate variability in the population. For the comparison, we simulate 250 GP-driven LIF neurons for 50 s with a time step of 0.05 ms; we use the same interval and time step for the theory.

Guided by the regime attained in full simulations, we choose $\nu_{\text{in}} \in [0.5, 2.5]$ spks/s and $\text{CV}_{\text{in}} \in [0.5, 1.5]$ (Fig. 8). The network is in the inhibition-dominated regime; thus the mean input decreases with ν_{in} starting from the value that brings the membrane potential on average to threshold [Fig. 8(a)]. In contrast, the static variability increases monotonically with ν_{in} [Fig. 8(a)]. To measure the strength of the dynamic variability, we divide the resulting standard deviation of the free membrane voltage by the distance of the mean free membrane voltage to the threshold, $\sigma_U/(\theta - \mu_U)$. Since the numerator grows with $\sqrt{\nu_{\text{in}}}$ while the denominator grows linearly with ν_{in} in inhibition-dominated networks, the standard deviation relative to the distance to threshold decreases with increasing ν_{in}; in contrast, it slightly increases with increasing CV_{in} [Fig. 8(b)]. For the entire parameter regime, the absolute difference in the firing rate is smaller than 1 spks/s and it is maximal at the brink of the fluctuation-driven regime [Fig. 8(c)]. For the static rate variability, we also consider the absolute difference, which is below 0.3 spks/s throughout the parameter space [Fig. 8(d)]. Next, we compare the ISI distributions using their Kolmogorov-Smirnov distance, i.e., the maximal absolute difference between the cumulative distributions. The Kolmogorov-Smirnov distance is maximal deep in the fluctuation-driven regime where the firing rate is well below 1 spks/s and the estimate of the ISI distribution is noisy [Fig. 8(e)]. Finally, we compare the output spectra using the maximum absolute distance between the scaled spectra $S_x(f)/\nu$. Here, the deviation is below 0.1 in most parts of the parameter space except for low $\text{CV}_{\text{in}} \leqslant 0.6$, high $\text{CV}_{\text{in}} \geqslant 1.3$, and at the brink of the fluctuation-driven regime [Fig. 8(f)]. To give meaning to the quantitative results, we plot two example ISI distributions [Fig. 8(g)] and spectra [Fig. 8(h)]. For the ISI distribution, we see the noisy estimate at low rates. For the spectra, we note that the main difference is a constant offset caused by a small error in the rate, see Eq. (38), while the shape is well matched.

To conclude, the above approximations work well in the fluctuation-driven regime for moderate values $0.6 < \text{CV}_{\text{in}} < 1.3$. Within this regime, the firing rate and its variability across neurons, the ISI distribution, and the power spectra are well predicted. Most importantly for the prediction of the intrinsic timescale, the theory closely predicts the scaled spectrum $S_x(f)/\nu$.

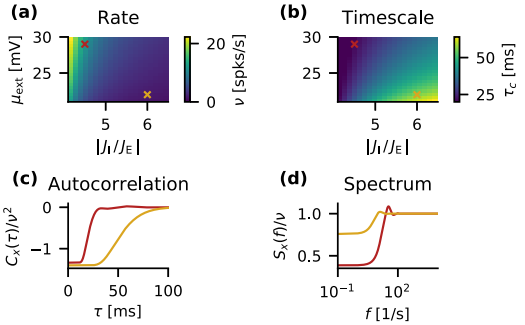

FIG. 9. Parameter scan for a balanced random network of LIF neurons using Eq. (35). [(a) and (b)] Firing rate and intrinsic timescale for varying external input μ_{ext} and relative inhibitory strength $|J_I/J_E|$. [(c) and (d)] Scaled autocorrelation $C_x(\tau)/\nu^2$ and power spectrum $S_x(f)/\nu$ for the parameter values indicated by symbols in (a,b). Further parameters as in Fig. 8.

2. *Timescales in balanced random networks of LIF neurons*

Having established the validity of the theory, we employ it to investigate the intrinsic timescale. It is well known that increasing the overall synaptic strength leads to a network state with long temporal correlations [37,62]. However, this state comes along with giant fluctuations of the membrane potential [68] which are well beyond the physiological regime and which our theory can capture only to a limited extent (in particular, it underestimates the strong increase in low-frequency power observed for strong couplings [37,69]). Hence, we focus on the influence of the external input μ_{ext} and the inhibition dominance $|J_I/J_E|$, in line with our above investigations for GLM neurons. We solve the theory on a $\Delta t = 0.05$ ms grid to a maximum of $T = 10$ s, use an ensemble size of $k = 5$ for the Gauss-Hermite quadrature, and choose an update step $\varepsilon = 0.2$.

We investigate the regime $|J_I/J_E| \in [4.1, 6]$ and $\mu_{\text{ext}} \in [21, 30]$ mV. Within this regime, the rate is below approximately 20 spks/s, increases with μ_{ext}, and decreases with $|J_I/J_E|$ [Fig. 9(a)]. In contrast, the intrinsic timescale decreases with μ_{ext}, increases with $|J_I/J_E|$, and reaches a maximum of approximately 60 ms $= 3\tau_{\text{m}}$ [Fig. 9(b)]. The autocorrelation function reveals that the nature of these longer intrinsic timescales in LIF networks is fundamentally different to the GLM networks above [Fig. 9(c)]: in the GLM networks, the autocorrelation function is positive, which corresponds to an increased probability to spike in succession; in the LIF networks it is negative, which corresponds to a prolonged effective refractory period caused by the fire-and-reset mechanism in combination with the input statistics. Indeed, in the corresponding power spectra and their zero-frequency limit, we see that the CV is well below 1 [Fig. 9(d)]. Hence, the process is more regular than a Poisson process, as opposed to the high irregularity $\text{CV} > 1$ that would go along with bursty spiking.

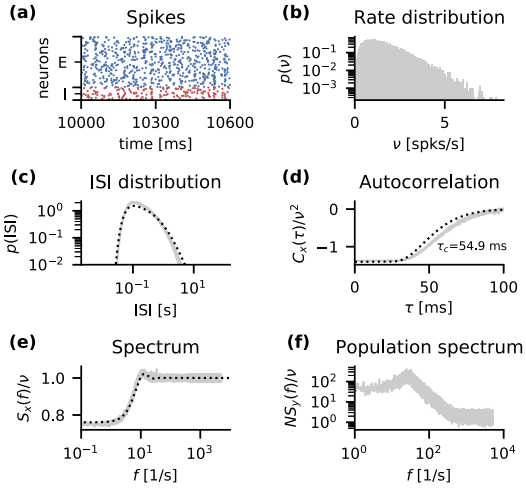

FIG. 10. Balanced random network of LIF neurons. (a) Raster plot of 2% of the excitatory (blue) and inhibitory (red) neurons. (b) Firing rate distribution across all neurons. [(c)–(e)] Population-averaged ISI distribution, population-averaged autocorrelation function, and population-averaged power spectrum from simulation (gray) and theory (black). (f) Power spectrum of the population activity. Parameters as in Fig. 8.

3. Simulation of balanced random network of LIF neurons

We validate the theoretical predictions for the balanced random network of LIF neurons by comparing with a network simulation. To acquire sufficient statistics, we simulate the network for $T = 2.5$ min with time step $\Delta t = 0.1$ ms and discard the first 10 s as an initial transient. After this transient, the network is in an asynchronous irregular state [Fig. 10(a)]. The rates of individual neurons are mostly below 5 spks/s with a peak at around 1 spks/s [Fig. 10(b)]. The theory closely predicts the ISI distribution apart from a slight overestimation of the tail [Fig. 10(c)]. Thus the resulting autocorrelation function is also well matched and the predicted intrinsic timescale of approximately 55 ms is confirmed [Fig. 10(d)]. Also the scaled spectrum is closely reproduced and reveals a $\mathrm{CV}^2 \approx 0.75$ [Fig. 10(e)].

To illustrate the difference between the single-unit and the population statistics, we furthermore plot the power spectrum of the population activity $y(t) = \frac{1}{N}\sum_{i=1}^{N} x_i(t)$ [Fig. 10(f)]. For vanishing cross-correlations, these two spectra would be proportional to each other. Already weak cross-correlations can shape the population spectrum since their contribution is of $O(N^2)$ compared to $O(N)$ contributions from the autocorrelations, leading to the clear differences we see between the single-unit and the population spectrum. A notable difference between the two spectra is the peak around 30 Hz in the population spectrum, contrasting with the roughly 10-Hz peak in the single-unit spectrum. Furthermore, the population spectrum displays increased power at low frequencies compared to high frequencies, while the reverse is true for the single-unit spectrum.

C. Biologically constrained network model

Thus far, we only considered balanced random networks with identical excitatory and inhibitory neurons that reduce to a single effective population. Despite this simplification, these balanced random networks already span a large parameter space. Here, we apply our theory to a multipopulation network model constrained by biological data [43]. Beyond the aspect of multiple populations, this network model allows us to highlight two additional features of our theory that we left out thus far: the possibility to include external Poisson input and distributed synaptic weights. We solve the theory on a $\Delta t = 0.05$ ms grid to a maximum of $T = 10$ s, use an ensemble size of $k = 5$ for the Gauss-Hermite quadrature, choose an update step $\varepsilon = 0.1$, and initialize all populations with a rate of 10 spks/s.

The model represents the neurons under 1 mm^2 of surface of generic early sensory cortex. It comprises eight populations: layers 2/3, 4, 5, and 6 with a population of excitatory cells and inhibitory interneurons for each layer [Fig. 11(a)]. In total, this leads to 77 169 neurons connected via approximately 3×10^8 synapses, with population-specific connection probabilities $p^{\alpha\beta}$ based on an extensive survey of the anatomical and physiological literature. In contrast to the original model, we directly use the connection probabilities to create the connectivity such that the total number of synapses can vary across instantiations of the model, and we draw source and target neurons without replacement, so that multapses are not allowed. Transmission delays follow truncated normal distributions with mean ± standard deviation of 1.5 ± 0.75 ms for excitatory source neurons and 0.75 ± 0.375 ms for inhibitory source neurons, both with a cutoff at 0.1 ms. The synaptic strengths $J_{ij}^{\alpha\beta}$ are normally distributed with $\mu_J^{\alpha I} = -351.2$ pA for inhibitory source neurons and $\mu_J^{\alpha E} = 87.8$ pA for excitatory source neurons except for connections from layer 4 excitatory to layer 2/3 excitatory neurons, which have a mean strength of 175.6 pA. For all synaptic strengths, the standard deviation is fixed to 10% of the mean. The network is driven by external Poisson input with layer-specific rates (for further details see Ref. [43]).

The intrinsic parameters of the neurons do not vary across populations. Shaped by the connectivity, a layer-specific activity arises [Fig. 11(b)] with mean firing rates between 1 and 10 spks/s [Fig. 11(c)] and a standard deviation across neurons between 1 and 5 spks/s [Fig. 11(d)]. While the quantitative agreement is not perfect, our theory captures the specificity of both mean firing rate and its variability across neurons well.

A prominent feature of the model are oscillations on the population level [70] which are already visible in the raster plot of only 2% of the population [Fig. 11(b)]. These oscillations lead to a clear peak at about 80 Hz in the power spectrum of the population activity in all layers [70]. Here, we only show a representative population spectrum [Fig. 11(e)]. These population-level oscillations clearly violate the independence assumption of the effective inputs. Thus they could potentially explain the deviations of the predicted firing rate from the simulation.

For most populations, the peak in the population-level oscillations also manifests itself in the population-averaged single-unit spectra [Figs. 11(f) and 11(g)]. Apart from this

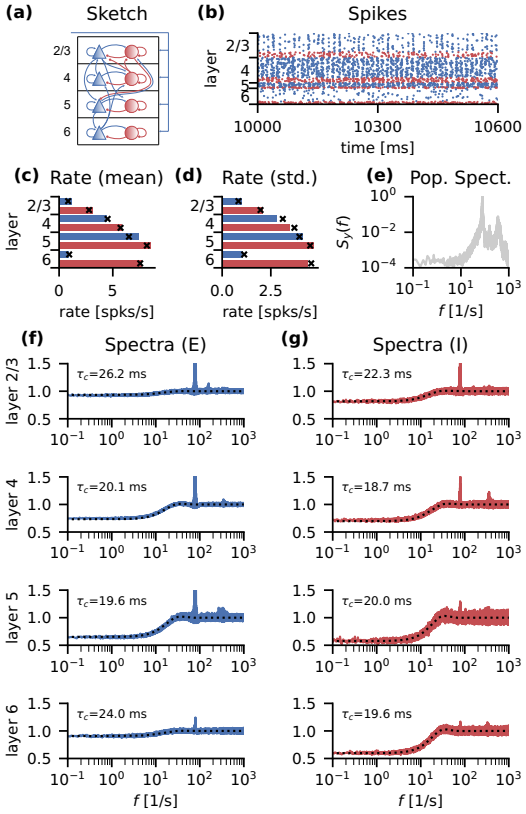

FIG. 11. Multipopulation network of LIF neurons. (a) Sketch of the model, figure adapted from [43]. (b) Raster plot of 2% of the neurons of each population. [(c) and (d)] Neuron-averaged firing rates and their standard deviation from simulations (bars) and theoretical predictions (black crosses). (e) Population spectrum of the layer 4 excitatory population. [(f) and (g)] Spike-train power spectra obtained from simulations (colored) and theory (black) and the corresponding intrinsic timescale. Parameters as specified in [43].

peak, our theory closely captures the shape of all spectra [Figs. 8(f) and 8(g)]. Note that, despite the large heterogeneity of mean rates, the intrinsic timescale is similar across populations. As in the balanced random network, the intrinsic timescales are on the order of magnitude of the membrane time constant (here 10 ms); concretely, the intrinsic timescale is approximately twice as large.

V. DISCUSSION

We developed a self-consistent theory for the second-order statistics, in particular the intrinsic timescales as defined by autocorrelation decay times, in block-structured random networks of spiking neurons in an asynchronous irregular state. Orthogonal to approaches based on the mean activity of a population of neurons, we consider population-averaged single-neuron statistics. To this end, we built on the model-independent dynamic mean-field theory (DMFT) developed in Ref. [30] and applied it to networks of spiking neurons. We sketched the derivation starting from the characteristic functional of the recurrent input, Eq. (3), to expose the inherent assumptions of the DMFT as well as its main result. In particular, we showed that the mean-field equations, Eqs. (9) and (10), where the connectivity matrix enters only through its first two cumulants, account for both (static) interneuron variability and (dynamic) temporal fluctuations. In order to close the self-consistency problem, we derived a novel analytical solution for the output statistics of a generalized linear model (GLM) neuron with error-function nonlinearity driven by a Gaussian process (GP), Eq. (25), and an analytical approximation for the output statistics of a GP-driven leaky integrate-and-fire (LIF) neuron in the fluctuation-driven regime, Eq. (35). These theoretical results yield firing rate distributions, spike-train power spectra, and interspike interval distributions that are close to those obtained from numerical simulations (Figs. 2, 5, and 10) even for a complex, biologically constrained network model (Fig. 11).

The excellent agreement between theory and simulations demonstrates the validity of the DMFT approximation, i.e., the approximation of the recurrent inputs as independent Gaussian processes. The validity of the DMFT approximation is most clearly demonstrated by the networks of GLM neurons, since in that case the DMFT assumption constitutes the only approximation, while the remainder of the solution is exact; while for the LIF networks, additional approximations are made, so that the effects of the DMFT assumption can be less well isolated.

Focusing on balanced random networks, we leveraged our theory to investigate the influence of network parameters on the intrinsic timescale for both GLM (Figs. 3 and 4) and LIF (Fig. 9) neurons. For the former neuron model with error function nonlinearity, our theory unveils that a product of two factors determines the intrinsic timescale [Eq. (27), Fig. 6]: the gain of the rate autocorrelation function with respect to changes in the membrane voltage autocorrelation function for $\tau \to \infty$, Eq. (28), and the variance of the connectivity, Eq. (6). Furthermore, providing a temporally correlated external drive causes the intrinsic timescale to monotonically approach the extrinsic timescale as the input strength is increased (Fig. 7).

For both GLM neurons with error function nonlinearity and LIF neurons, we find parameter regimes where the intrinsic timescale τ_c is longer than the largest time constant of the single-neuron dynamics, the membrane time constant τ_m (Figs. 5 and 9). This demonstrates that the recurrent dynamics shape the intrinsic timescale. Note that we consider a regime where the inverse firing rate ν^{-1} is large compared to τ_m. In contrast, [18,29] consider the opposite regime where slow neuronal timescales lead to effective rate dynamics, and the spiking noise is either left out or treated perturbatively. Our results show that it is possible to obtain longer intrinsic timescales even in a regime where the white component of the spiking noise contributes non-negligibly to the membrane voltage fluctuations. However, the temporal structure that causes the prolonged intrinsic timescale is very different for the two models that we consider: For GLM neurons, the

autocorrelation is positive for a period on the order of τ_c, corresponding to an increased spiking probability. For LIF neurons, the autocorrelation function is negative, corresponding to a prolonged effective refractory period.

Furthermore, LIF networks exhibit a minimum in the intrinsic timescale [37], while the corresponding GLM networks exhibit a maximum (Fig. 6). We hypothesize that this difference is due to the difference in the temporal structure: The minimum in the timescale for LIF networks is caused by a switch from an increased effective refractory period (a negative autocorrelation function for $\tau \to 0$) to an increased probability for another spike (a positive autocorrelation function for $\tau \to 0$). This hypothesis is consistent with the switch from decreased low-frequency power, $S_x(f \to 0) < \nu$, to increased low-frequency power, $S_x(f \to 0) > \nu$, highlighted in Refs. [37,69]. For GLM networks, this switch and hence the minimum is absent. Instead, the more subtle interplay between the gain and the variance of the connectivity leads to the maximum. The presence of a maximum rather than a minimum in the intrinsic timescales renders the GLM networks more similar to networks of rate units [15]. If similar mechanisms are at play as in rate networks, the white spiking noise of the input to the GLM neurons may temper the size of the largest possible timescale [21]. However, due to the inherent stochasticity of GLM neurons, it is unclear whether the maximum occurs at a transition to chaos as it does in rate networks [15].

Considering a more complex block-structured network model that is constrained by biological data [43] exposes limits of our theory: while the theory accurately captures the nonoscillatory components of the power spectra, it misses a high-frequency oscillation (Fig. 11). These high-frequency oscillations are caused by correlated activity on the population level [70]; hence, the peak in the population-averaged single-neuron spectra demonstrates an interplay between single-unit and population-level statistics that was absent in the simpler balanced random network models. By construction, our theory only accounts for population-averaged single-neuron statistics and thus misses the high-frequency peak. It is an interesting challenge to derive a self-consistent theory on both scales simultaneously.

In general, the limits of DMFT when applied to spiking networks merit further investigation. For example, assuming that the network is sparse, $K \ll N$ or $p \ll 1$, is not a necessary condition for a DMFT to apply [18]. Nonetheless, increasing sparsity reduces the pairwise correlations between the neurons [26,62,71] such that DMFT is expected to yield better results. Another important aspect is that for the synaptic weights scaling as $J_{ij} = O(1/\sqrt{K})$, the fluctuations of the mean input $\mu_\eta(t)$ can be $O(1)$, i.e., not scale with $K^{-\alpha}$, $\alpha > 0$, as the network size increases and p is kept constant. In Eq. (7), $\mu_\eta(t)$ and $C_\eta(t, t')$ are replaced by their average, neglecting fluctuations; including the fluctuations of the mean input would lead to an additional term in $C_\eta(t, t')$ [28]. Since these fluctuations of the mean input reflect pairwise correlations, the latter need to be small for the theory to be accurate. The above scaling argument shows that it is nontrivial that the pairwise correlations vanish, even in the large network limit. They only do so for an asynchronous state in which the pairwise correlations are small already for finite networks, e.g., due to a sparse network or due to inhibitory feedback [27]. Conversely, if a network is in an asynchronous irregular state, which has low pairwise correlations by definition, DMFT is expected to yield reliable results.

The heterogeneity of timescales even within a cortical area [72] suggests another interesting extension, namely to calculate the variability of the timescale within a population. This requires calculating the variability of the second-order statistics, which has recently been achieved for linear rate networks [73] but to the best of our knowledge is an open challenge even for simple nonlinear rate networks, let alone for spiking networks.

The microscopic theory presented here enables direct comparisons with experimental measurements of neuron-level intrinsic timescales [2], in contrast to previous works which have considered population rate models [10,74]. It is important to distinguish between neuron-level and population-level autocorrelations, since the latter are shaped by $O(N^2)$ cross-correlations and can therefore differ substantially from neuron-level autocorrelations, as we have illustrated for the balanced random network model [Figs. 10(e) and 10(f)] and the biologically constrained network model [Figs. 11(e)–11(g)].

Establishing a direct link between the connectivity and the emergent intrinsic timescales opens up the possibility of a thorough investigation of the effect of network architecture. Moreover, within our theory, it is possible to account for population-specific intrinsic neuron parameters. Thus the theory also provides an avenue for investigations of the complex interplay between intrinsic parameters [8,9] and the network structure [10]. In this context, an interesting application is clustered networks which feature slow switching between transiently active clusters [75]. In particular, clustered networks with both excitatory and inhibitory clusters [76–78] could be of interest because they robustly give rise to winnerless competition. From a modeler's point of view, uncovering mechanisms shaping intrinsic timescales could be used to fine-tune network models [79–82] to match the experimentally observed hierarchy of timescales [2]. Focusing on computational aspects, diverse timescales strongly enhance the computational capacity of a recurrent network [83–85], and neurons with long intrinsic timescales carry more information in a working memory task [86] (but see [87]). In this light, the results presented here may also contribute to improved understanding of aspects of information processing in the brain.

ACKNOWLEDGMENTS

This work was supported by the European Union's Horizon 2020 Framework Programme for Research and Innovation under Specific Grants Agreement No. 785907 and No. 945539 (Human Brain Project SGA2, SGA3), the Jülich-Aachen Research Alliance (JARA), and DFG Priority Program "Computational Connectomics" (SPP 2041; Project 347572269). A.v.M. would like to thank Moritz Helias for many helpful and inspiring discussions about dynamical mean-field theory and spiking neurons. We further thank Tilo Schwalger for his insights about the Stratonovich approximation and for sharing his unpublished manuscript, and Jasper Albers, Anno Kurth, Alessandra Stella, Christian Keup, David Dahmen, and

the anonymous reviewers for valuable comments on an early version of the manuscript.

APPENDIX A: CHARACTERISTIC FUNCTIONALS

Here, we briefly introduce the characteristic functionals for both types of stochastic processes we consider: Gaussian processes and point processes. We closely follow Stratonovich's book [40], in particular Chaps. I.1. and I.6.

1. Stochastic processes

The characteristic functional of a stochastic process $\xi(t)$ is defined as

$$\Phi_\xi[u(t)] = \langle e^{i\int_0^T u(t)\xi(t)dt}\rangle_\xi$$

where $u(t)$ is an arbitrary test function. In terms of the cumulants $k_r(t_1,\ldots,t_r)$ the characteristic functional can be written as

$$\Phi_\xi[u(t)] = e^{\sum_{s=1}^{\infty}\frac{i^s}{s!}\int_0^T\cdots\int_0^T k_s(t_1,\ldots,t_s)u(t_1)\ldots u(t_s)dt_1\ldots dt_s}. \tag{A1}$$

All properties of a stochastic process are determined by its characteristic functional.

If all cumulants except for the first vanish, the process is deterministic and has the characteristic functional

$$\Phi_\xi[u(t)] = \langle e^{i\int_0^T u(t)\xi(t)dt}\rangle_\xi = e^{i\int_0^T u(t)\xi(t)dt}. \tag{A2}$$

In this case, the first cumulant coincides with the process itself, $k_1(t) = \xi(t)$. If only the first and the second cumulants are nonvanishing, the process is a Gaussian process. The corresponding characteristic functional reads

$$\Phi_\xi[u(t)] = e^{i\int k_1(t_1)u(t_1)dt_1 - \frac{1}{2}\iint u(t_1)k_2(t_1,t_2)u(t_2)dt_1dt_2}. \tag{A3}$$

If the Gaussian process is stationary, $k_1(t_1) = k_1$ and $k_2(t_1,t_2) = k_2(t_2-t_1)$, the characteristic functional simplifies further to $\Phi_\xi[u(t)] = e^{ik_1\int u(t_1)dt_1 - \frac{1}{2}\iint u(t_1)k_2(t_2-t_1)u(t_2)dt_1dt_2}$.

The characteristic functional describes the statistics at all points in time. It is often useful to relate the characteristic functional to the distribution of the values of $\xi(t)$ at fixed points in time, for instance to compute the statistics of the current at upcrossings and after the refractory period, or to obtain marginal activity statistics which, given stationarity, reflect time-averaged activity. To this end, we can use the test functions $u(t) = u_1\delta(t-t_1)$ and $u(t) = u_1\delta(t-t_1) + u_2\delta(t-t_2)$ to obtain

$$\Phi_\xi(u_1) = e^{ik_1(t_1)u_1 - \frac{1}{2}k_2(t_1,t_1)u_1^2}, \tag{A4}$$

$$\Phi_\xi(u_1,u_2) = e^{i(k_1(t_1)u_1 + k_1(t_2)u_2)} \times e^{-\frac{1}{2}(k_2(t_1,t_1)u_1^2 + 2k_2(t_1,t_2)u_1u_2 + k_2(t_2,t_2)u_2^2)}. \tag{A5}$$

These are the characteristic functions of a Gaussian with cumulants determined by k_1 and k_2. Knowing these characteristic functions for all times t_1 and t_2 provides the full picture; this is the marginalization property of Gaussian processes [88].

2. Point processes

The equivalence to the characteristic functional for a point process is the generating functional. For a spike train $\{t_1,\ldots,t_n\}$ (a "system of random points" in Stratonovich's naming) with $t_i \in [0,T]$ for all i, the generating functional is defined by

$$L_T[v(t)] = \left\langle \prod_{j=1}^{n}[1+v(t_j)] \right\rangle.$$

Here, the number of spikes n is itself a random variable because the average is taken with respect to all possible realizations of the spike train [89].

For point processes, the role of the moments is taken by the "distribution functions" $n_r(t_1,\ldots,t_r)$ which denote the probability of having at least one point in each interval $[t_i, t_i+dt]$. The role of the cumulants is taken by the functions $g_r(t_1,\ldots,t_r)$, which are related to the distribution functions as the cumulants of a stochastic process are related to its moments. In terms of the $g_r(t_1,\ldots,t_r)$, the generating functional can be written as [89]

$$L_T[v(t)] = e^{\sum_{s=1}^{\infty}\frac{1}{s!}\int_0^T\cdots\int_0^T g_s(t_1,\ldots,t_s)v(t_1)\ldots v(t_s)dt_1\ldots dt_s}. \tag{A6}$$

The generating functional is directly related to a few useful quantities: The characteristic function of the number of spikes n in the interval $[0,T]$ is given by $\langle e^{inu}\rangle = L_T[e^{iu}-1]$; the probability that no point falls into $[0,T]$, i.e., the survival probability, is given by $L_T[-1]$. The simplest case of a point process where only g_1 is nonvanishing is a Poisson process. The corresponding generating functional reads

$$L_T[v(t)] = \exp\left(\int_0^T g_1(t_1)v(t_1)dt_1\right)$$

with survival probability $S(T) = L_T[-1] = e^{-\int_0^T g_1(t_1)dt_1}$.

The generating functional is directly related to the characteristic functional of the stochastic process $\xi(t) = \sum_{j=1}^{n}\delta(t-t_j)$:

$$\Phi_\xi[u(t)] = \langle e^{i\sum_{j=1}^{n}u(t_j)}\rangle = L_T[e^{iu(t)}-1]. \tag{A7}$$

This relation links the distribution functions n_r through the g_r to the cumulants of the spike train. For example, the characteristic functional of a Poisson spike train is

$$\Phi_\xi[u(t)] = \exp\left(\int_0^T g_1(t_1)(e^{iu(t_1)}-1)dt_1\right). \tag{A8}$$

Note that by convention, $g_1(t)$ is typically called $\lambda(t)$ for a Poisson process—we adopted this convention in the main text, in particular in Eq. (14). Expanding the exponent on the right-hand side of Eq. (A7) to second order in $u(t)$, we obtain the relations

$$k_1(t_1) = g_1(t_1),$$
$$k_2(t_1,t_2) = g_1(t_1)\delta(t_1-t_2) + g_2(t_1,t_2)$$

between the g_r and the first two cumulants of the spike train.

3. Gaussian integrals

We solve several Gaussian integrals using the impressive table by Owen [59]. First, we introduce his notation

$$G(x) = \frac{1}{2}(1 + \mathrm{erf}(x/\sqrt{2})), \quad g(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

for the standard normal CDF and PDF. Furthermore, we need Owen's T function

$$T(h, a) = \frac{1}{2\pi} \int_0^a \frac{e^{-\frac{1}{2}h^2(1+x^2)}}{1+x^2} dx.$$

All formulas were numerically validated using numerical integration routines implemented in SCIPY [90].

a. GLM error function

Here, we derive Eqs. (24) and (25). In the notation of Eq. (23), we have $\phi(x) = G(x)$.

For the mean, we need the expectation $\langle \phi(z) \rangle$ where z is Gaussian with mean μ and variance σ^2. Equivalently, we can calculate $\langle \phi(\mu + \sigma x) \rangle$ where x is standard normal. Expressing the standard normal Gaussian expectations using $g(x)$, we have

$$\langle \phi \rangle = \int_{-\infty}^{\infty} g(x) G(\mu + \sigma x) dx.$$

Using Eq. (10,010.8) from Ref. [59], we get

$$\langle \phi \rangle = G\left(\frac{\mu}{\sqrt{1+\sigma^2}} \right).$$

Equation (24) follows after taking the multiplying factor c_1 and $\mu = \mu_V - \theta$ from Eq. (14) into account.

For the second moment, we need $\langle \phi(z_1)\phi(z_2) \rangle$ were z_1 and z_2 are jointly Gaussian with mean μ, variance σ^2 and correlation coefficient ρ. Equivalently, we can calculate $\langle \phi(\mu + \beta x - \alpha y)\phi(\mu + \beta x + \alpha y) \rangle$ where x and y are standard normal and $\alpha = \sigma\sqrt{(1-\rho)/2}$, $\beta = \sigma\sqrt{(1+\rho)/2}$. Again using $g(x)$ to express the standard normal Gaussian expectations, we get

$$\langle \phi\phi \rangle = \int_{-\infty}^{\infty} g(x) I(x) dx \quad \text{with}$$

$$I(x) = \int_{-\infty}^{\infty} g(y) G(\mu + \beta x - \alpha y) G(\mu + \beta x + \alpha y) dy.$$

Now, we use Eq. (20,010.3) in Ref. [59] for $I(x)$ to obtain

$$\langle \phi\phi \rangle = \int_{-\infty}^{\infty} g(x)(G(a + bx) - 2T(a + bx, c)) dx$$

with $a = \mu/\sqrt{1 + \sigma^2(1-\rho)/2}$, $b = \sigma\sqrt{1+\rho}/\sqrt{2 + \sigma^2(1-\rho)}$, $c = \sqrt{1 + \sigma^2(1-\rho)}$, and Owen's T function $T(h, a)$. For the final integral, we use Eqs. (10,010.8) and (c00,010.1) from Ref. [59] to derive

$$\langle \phi\phi \rangle = G\left(\frac{\mu}{\sqrt{1+\sigma^2}} \right) - 2T\left(\frac{\mu}{\sqrt{1+\sigma^2}}, \sqrt{\frac{1+\sigma^2(1-\rho)}{1+\sigma^2(1+\rho)}} \right).$$

Equation (25) follows after subtracting $\langle \phi \rangle^2$.

b. Free upcrossing probabilities

For the free two-point upcrossing probability, we need integrals of the form

$$I_n(a, b) = \int_0^{\infty} x^n g(x) G(ax + b) dx.$$

For arbitrary n, Eq. (10,01n.4) from Ref. [59] provides the solution

$$I_n(a, b) = \frac{\Gamma((n+1)/2) 2^{(n-1)/2}}{\sqrt{2\pi}} F_{n+1,-b}(\sqrt{n+1}a),$$

where $F_{\nu,\mu}(x)$ denotes the cumulative distribution function of noncentral t-distribution with ν degrees of freedom and noncentrality parameter μ. Analytical expressions for $F_{\nu,\mu}(x)$ in terms of $g(x)$, $G(x)$, and $T(h, a)$ can be found in Ref. [91] (the ones in Ref. [59] contain typos). Using these expressions, the solutions for $n = 0,\ 1,$ and 2 are

$$I_0(a, b) = \frac{1}{2} G(bB) + T(bB, a),$$

$$I_1(a, b) = \frac{1}{\sqrt{2\pi}} G(b) + M_0(a, b),$$

$$I_2(a, b) = I_0(a, b) + M_1(a, b)$$

where we used the shorthand notation $B = 1/\sqrt{1 + a^2}$ and

$$M_0(a, b) = aB\, g(bB)\, G(-abB),$$

$$M_1(a, b) = B^2(-abM_0(a, b) + ag(b)/\sqrt{2\pi}).$$

Since we consider only up to $n = 2$, we are spared the increasingly cumbersome expressions for $n > 2$.

APPENDIX B: FREE UPCROSSING PROBABILITIES

The dynamics of the free membrane voltage and the current for the LIF neuron model are given by

$$\dot{U} = -U + I, \tag{B1}$$

$$\tau_s \dot{I} = -I + \eta, \tag{B2}$$

where we measure time in units of the membrane time constant τ_m, i.e., we set $\tau_m = 1$. Furthermore, we set $\langle \eta \rangle = 0$, i.e., we measure U and I relative to the mean input. Lastly, we define $t = 0$ to be the end of the refractory period, i.e., the time when the free dynamics start evolving.

First, we need the distribution of the voltage and the current. Since η is a Gaussian process, both are Gaussian for arbitrary time arguments. Thus it is sufficient to calculate the first two conditional cumulants. Throughout, we assume a correlation-free preparation [92], i.e., we assume that η and I are uncorrelated prior to $t = 0$.

1. Nonstationary mean and variance of U and I

We need the nonstationary mean and variance of U and I to calculate the free upcrossing probability. For a given initial current and initial voltage, Eqs. (B1) and (B2) lead

to

$$I(t) = I_0 e^{-t/\tau_s} + \frac{1}{\tau_s}\int_0^t e^{-(t-s)/\tau_s}\eta(s)ds,$$
$$U(t) = U_0 e^{-t} + \int_0^t e^{-(t-s)} I(s)ds.$$

This leads immediately to the mean

$$\mu_I(t) = I_0 e^{-t/\tau_s},$$
$$\mu_U(t) = U_0 e^{-t} + \frac{\tau_s}{1-\tau_s} I_0 (e^{-t} - e^{-t/\tau_s}).$$

To obtain the variances numerically, we use that they follow linear differential equations: taking the temporal derivatives of $I(t)^2$, $I(t)U(t)$, and $U(t)^2$, using Eqs. (B1) and (B2), and averaging leads to

$$\frac{\tau_s}{2}\dot{\sigma}_I^2 = -\sigma_I^2 + \sigma_{I\eta}^2,$$
$$\tau_s \dot{\sigma}_{IU}^2 = -(1+\tau_s)\sigma_{IU}^2 + \tau_s\sigma_I^2 + \sigma_{U\eta}^2,$$
$$\frac{1}{2}\dot{\sigma}_U^2 = -\sigma_U^2 + \sigma_{IU}^2.$$

The initial conditions for all of the above differential equations are $\sigma_I^2(0) = \sigma_{IU}^2(0) = \sigma_U^2(0) = 0$. They are straightforward to solve numerically in the order that they appear, but they require two additional quantities:

$$\sigma_{I\eta}^2(t) = \frac{1}{\tau_s}\int_0^t e^{-s/\tau_s} C_\eta(s)ds,$$
$$\sigma_{U\eta}^2(t) = \frac{1}{1-\tau_s}\int_0^t (e^{-s} - e^{-s/\tau_s}) C_\eta(s)ds,$$

which can be numerically computed using a composite trapezoidal rule. If $C_\eta(\tau)$ contains a Dirac delta, $C_\eta(\tau) = \hat{C}_\eta(\tau) + 2D\delta(\tau)$, we have to separate it analytically in $\sigma_{I\eta}^2(t)$:

$$\sigma_{I\eta}^2(t) = \hat{\sigma}_{I\eta}^2(t) + \frac{D}{\tau_s}.$$

Note the factor $1/2$ because we only integrate "half" of the Dirac delta. In $\sigma_{U\eta}^2(t)$, the Dirac delta does not contribute because the integrand vanishes at zero, i.e., $\sigma_{U\eta}^2(t) = \hat{\sigma}_{U\eta}^2(t)$.

Ultimately, we need the cumulants of U and $\dot{U}$ instead of U and I. To relate the respective quantities, we use Eq. (B1). For the initial conditions, we have

$$\dot{U}_0 = I_0 - U_0.$$

The first cumulants are

$$\mu_U(t) = U_0 e^{-t} + (\dot{U}_0 + U_0)A(t),$$
$$\mu_{\dot{U}}(t) = -\mu_U(t) + (\dot{U}_0 + U_0)e^{-t/\tau_s}$$
$$= -U_0 e^{-t} + (\dot{U}_0 + U_0)B(t),$$

where we used Eq. (B1) for $\mu_U(t)$ and abbreviated

$$A(t) = \frac{\tau_s}{1-\tau_s}(e^{-t} - e^{-t/\tau_s}), \quad B(t) = e^{-t/\tau_s} - A(t).$$

The second cumulants do not depend on the initial conditions and we get from Eq. (B1):

$$\sigma_{U\dot{U}}^2(t) = -\sigma_U^2(t) + \sigma_{IU}^2(t),$$
$$\sigma_{\dot{U}}^2(t) = \sigma_U^2(t) - 2\sigma_{IU}^2(t) + \sigma_I^2(t).$$

Finally, we need to marginalize the initial velocity.

We assume that $\dot{U}_0$ is Gaussian distributed with mean $\mu_{\dot{U}_0}$ and variance $\sigma_{\dot{U}_0}^2$. Marginalizing $\dot{U}_0$ again results in a Gaussian distribution because $p(\dot{U}_0)$ and $p(U_1, \dot{U}_1 \mid U_0, \dot{U}_0)$ are Gaussian. Hence, we only need to compute the cumulants. For the mean, we simply have to replace $\dot{U}_0 \to \mu_{\dot{U}_0}$. The second cumulants are

$$\tilde{\sigma}_U^2(t) = \sigma_U^2(t) + \sigma_{\dot{U}_0}^2 A(t)^2,$$
$$\tilde{\sigma}_{U\dot{U}}^2(t) = \sigma_{U\dot{U}}^2(t) + \sigma_{\dot{U}_0}^2 A(t)B(t),$$
$$\tilde{\sigma}_{\dot{U}}^2(t) = \sigma_{\dot{U}}^2(t) + \sigma_{\dot{U}_0}^2 B(t)^2.$$

With this, we can evaluate the mean and the variance numerically from the statistics of $\eta(t)$ and $\dot{U}_0$.

2. Initial velocity distribution

For the distribution of initial velocities, we assume that the voltage has reached a stationary distribution by the time it crosses the threshold. The velocity at an upcrossing of a stationary Gaussian process is Rayleigh distributed [40]. Because at the threshold we have $\dot{U}_{\text{up}} = -\theta + I_{\text{up}}$ (remember that $t = 0$ denotes the end of the refractory period, that the membrane resistance is absorbed into the current, and time is rescaled such that $\tau_m = 1$), the current is also Rayleigh distributed,

$$p(I_{\text{up}}) = \begin{cases} \frac{(I_{\text{up}}-\theta)}{\sigma_I^2} \exp\left(-\frac{(I_{\text{up}}-\theta)^2}{2\sigma_I^2}\right) & \text{for } I_{\text{up}} \geqslant \theta \\ 0 & \text{otherwise} \end{cases},$$

where $\sigma_I^2 = -\ddot{C}_U(0)$ with the stationary autocorrelation $C_U(\tau)$ of the free voltage. We assume that the further development of the current is also stationary, and neglect the conditional dependencies of the transition probability $p(I_0 \mid I_{\text{up}})$ on the threshold crossing beyond I_{up}, e.g., on $\dot{I}_{\text{up}}$ and $\ddot{I}_{\text{up}}$. This transition probability can thus be obtained from the unconstrained ("free") stationary statistics of the current—not conditioned on a threshold crossing—which are Gaussian: $p(I_0 \mid I_{\text{up}}) = p_{\text{free}}(I_0, I_{\text{up}})/p_{\text{free}}(I_{\text{up}})$. The unconstrained joint and instantaneous distributions here function as auxiliary quantities for computing $p(I_0 \mid I_{\text{up}})$. The unconstrained joint distribution is a Gaussian with variance σ_I^2 and covariance $\sigma_I^2 R_I(\tau_{\text{ref}})$ where $R_I(\tau) = -\ddot{C}_U(\tau)/\sigma_U^2$. We derive $C_U(\tau)$ most conveniently by Fourier transforming Eqs. (B1) and (B2), which leads to $S_U(f) = S_\eta(f)/(1+(2\pi f)^2)/(1+(2\pi\tau_s f)^2)$, and using the Wiener-Khinchin theorem to obtain the autocorrelation. From the unconstrained joint and instantaneous distributions, we obtain the transition probability $p(I_0 \mid I_{\text{up}})$, which is again a Gaussian with [88]

$$\tilde{\mu}_I(\tau_{\text{ref}}) = I_{\text{up}} R_I(\tau_{\text{ref}}), \quad \tilde{\sigma}_I^2(\tau_{\text{ref}}) = \sigma_I^2(1 - R_I(\tau_{\text{ref}})^2).$$

Combining this with the Rayleigh-distributed $p(I_{\mathrm{up}})$ yields

$$p(I_0) = \int_\theta^\infty p(I_0 \mid I_{\mathrm{up}}) p(I_{\mathrm{up}}) dI_{\mathrm{up}},$$

which is not a Gaussian anymore. We only calculate the first two cumulants,

$$\hat{\mu}_I(\tau_{\mathrm{ref}}) = \langle \tilde{\mu}_I(\tau_{\mathrm{ref}}) \rangle_{I_0} = \left(\sqrt{\frac{\pi}{2}\sigma_I^2} + \theta \right) R_I(\tau_{\mathrm{ref}}),$$

$$\begin{aligned}\hat{\sigma}_I^2(\tau_{\mathrm{ref}}) &= \tilde{\sigma}_I^2(\tau_{\mathrm{ref}}) + \langle (\tilde{\mu}_I(\tau_{\mathrm{ref}}) - \langle \tilde{\mu}_I(\tau_{\mathrm{ref}}) \rangle_{I_0})^2 \rangle_{I_0} \\ &= \tilde{\sigma}_I^2(\tau_{\mathrm{ref}}) + \frac{4-\pi}{2}\sigma_I^2 R_I(\tau_{\mathrm{ref}})^2,\end{aligned}$$

and neglect the higher cumulants to arrive at a Gaussian approximation. Finally, after the refractory time we have $\dot{U}_0 = -V_{\mathrm{r}} + I_0$. Combining the above equations leads to

$$\mu_{\dot{U}_0} = \left(\sqrt{\frac{\pi}{2}\sigma_I^2} + \theta \right) R_I(\tau_{\mathrm{ref}}) - V_{\mathrm{r}}, \tag{B3}$$

$$\sigma_{\dot{U}_0}^2 = \sigma_I^2 \left(1 - \frac{\pi - 2}{2} R_I(\tau_{\mathrm{ref}})^2 \right), \tag{B4}$$

which determine the Gaussian approximation of the initial velocity distribution.

3. One-point upcrossing probability

Here, we calculate the upcrossing probability Eq. (39),

$$n_1(t) = \int_0^\infty \dot{U}_1 p(\theta, \dot{U}_1 \mid V_{\mathrm{r}}) d\dot{U}_1.$$

Due to the linearity of Eqs. (B1) and (B2), the distribution $p(\theta, \dot{U}_1 \mid V_{\mathrm{r}})$ is a Gaussian with the cumulants we calculated above [92]. Hence, it takes the form

$$p(\theta, \dot{U}_1 \mid V_{\mathrm{r}}) = \frac{1}{\sqrt{\det(2\pi \mathbf{C})}} \exp\left(-\frac{1}{2} (\mathbf{u} - \boldsymbol{\mu})^T \mathbf{C}^{-1} (\mathbf{u} - \boldsymbol{\mu}) \right),$$

where $\mathbf{u}^T = (\theta, \dot{U}_1)$ and the mean and the correlation matrix are given by

$$\boldsymbol{\mu} = \begin{pmatrix} \tilde{\mu}_U(t) \\ \tilde{\mu}_{\dot{U}}(t) \end{pmatrix}, \quad \mathbf{C} = \begin{pmatrix} \tilde{\sigma}_U^2(t) & \tilde{\sigma}_{U\dot{U}}^2(t) \\ \tilde{\sigma}_{\dot{U}U}^2(t) & \tilde{\sigma}_{\dot{U}}^2(t) \end{pmatrix}.$$

Inverting $\mathbf{C}$ leads to

$$\mathbf{C}^{-1} = \frac{1}{\det(\mathbf{C})} \begin{pmatrix} \tilde{\sigma}_{\dot{U}}^2(t) & -\tilde{\sigma}_{U\dot{U}}^2(t) \\ -\tilde{\sigma}_{\dot{U}U}^2(t) & \tilde{\sigma}_U^2(t) \end{pmatrix},$$

$$\det(\mathbf{C}) = \tilde{\sigma}_U^2(t)\tilde{\sigma}_{\dot{U}}^2(t) - \tilde{\sigma}_{U\dot{U}}^4(t).$$

The exponent of $p(\theta, \dot{U}_1 \mid V_{\mathrm{r}})$ takes the form

$$(\mathbf{u} - \boldsymbol{\mu})^T \mathbf{C}^{-1} (\mathbf{u} - \boldsymbol{\mu}) = \frac{1}{\det(\mathbf{C})} \left[a\dot{U}_1^2 - 2b\dot{U}_1 + c^2 \right]$$

with $a = \tilde{\sigma}_U^2(t)$, $b = \tilde{\mu}_{\dot{U}}(t)\tilde{\sigma}_U^2(t) + (\theta - \tilde{\mu}_U(t))\tilde{\sigma}_{U\dot{U}}^2(t)$ and $c^2 = \tilde{\mu}_{\dot{U}}(t)^2\tilde{\sigma}_U^2(t) + 2(\theta - \tilde{\mu}_U(t))\tilde{\mu}_{\dot{U}}(t)\tilde{\sigma}_{U\dot{U}}^2(t) + (\theta - \tilde{\mu}_U(t))^2\tilde{\sigma}_{\dot{U}}^2(t)$.

Putting it together, n_1 is given by

$$n_1(t) = \frac{1}{\sqrt{\det(2\pi \mathbf{C})}} \int_0^\infty \dot{U}_1 \exp\left(-\frac{a\dot{U}_1^2 - 2b\dot{U}_1 + c^2}{2\det(\mathbf{C})} \right) d\dot{U}_1.$$

The integral can be solved in terms of an error function:

$$\begin{aligned}\int_0^\infty \dot{U}_1 e^{-\frac{a\dot{U}_1^2 - 2b\dot{U}_1 + c^2}{2\det(\mathbf{C})}} d\dot{U} &= \frac{\det(\mathbf{C})}{a} e^{-\tilde{c}^2} \\ &+ \frac{\det(\mathbf{C})}{a} e^{-\tilde{c}^2} \sqrt{\pi} \tilde{b} e^{\tilde{b}^2} (1 + \mathrm{erf}(\tilde{b})),\end{aligned}$$

where $\tilde{b} = b/\sqrt{2a\det(\mathbf{C})}$ and $\tilde{c} = c/\sqrt{2\det(\mathbf{C})}$. Thus we get

$$n_1(t) = \frac{\sqrt{\det(\mathbf{C})}}{2\pi\tilde{\sigma}_U^2(t)} e^{-\tilde{c}^2} (1 + \sqrt{\pi}\tilde{b} e^{\tilde{b}^2} (1 + \mathrm{erf}(\tilde{b}))) \tag{B5}$$

for the free upcrossing rate.

4. Stationary correlation function of U and $\dot{U}$

For the stationary two-point upcrossing probability, we need the stationary correlation functions of $U, \dot{U}$, and between U and $\dot{U}$. The power spectrum of U follows from the power spectrum of η using

$$S_U(f) = \frac{S_\eta(f)}{(1 + (2\pi f)^2)(1 + (2\pi f \tau_{\mathrm{s}})^2)}.$$

An inverse Fourier transform leads to the stationary correlation function $C_U(\tau)$. For stationary processes, the formulas

$$C_{U\dot{U}}(\tau) = -C_{\dot{U}U}(\tau) = \dot{C}_U(\tau), \quad C_{\dot{U}}(\tau) = -\ddot{C}_U(\tau)$$

yield the remaining correlation functions. The first formula follows from $\langle U(t)\dot{U}(t+\tau) \rangle = \frac{d}{d\tau}\langle U(t)U(t+\tau) \rangle$ and $\langle \dot{U}(t)U(t+\tau) \rangle = \langle \dot{U}(t-\tau)U(t) \rangle = -\frac{d}{d\tau}\langle U(t-\tau)U(t) \rangle$, the second from $\langle \dot{U}(t)\dot{U}(t+\tau) \rangle = \frac{d}{d\tau}\langle \dot{U}(t)U(t+\tau) \rangle = \frac{d}{d\tau}\langle \dot{U}(t-\tau)U(t) \rangle = -\frac{d^2}{d\tau^2}\langle U(t-\tau)U(t) \rangle$.

5. Stationary two-point upcrossing probability

Here, we calculate the stationary two point upcrossing probability Eq. (40),

$$n_2(\tau) = \int_0^\infty \int_0^\infty \dot{U}_2 \dot{U}_1 p(\theta, \dot{U}_2; \theta, \dot{U}_1) d\dot{U}_1 d\dot{U}_2.$$

The joint density $p(U_2, \dot{U}_2; U_1, \dot{U}_1)$ takes the form

$$p(U_2, \dot{U}_2; U_1, \dot{U}_1) = \frac{1}{\sqrt{\det\left(2\pi\sigma_U^2 \mathbf{C}\right)}} \exp\left(-\frac{1}{2\sigma_U^2} \mathbf{u}^T \mathbf{C}^{-1} \mathbf{u} \right),$$

where $\mathbf{u}^T = (U_1, \dot{U}_1, U_2, \dot{U}_2)$ and $\sigma_U^2 = C_U(0)$. The correlation matrix is given by

$$\mathbf{C} = \begin{pmatrix} 1 & 0 & R(\tau) & \dot{R}(\tau) \\ 0 & -\ddot{R}(0) & -\dot{R}(\tau) & -\ddot{R}(\tau) \\ R(\tau) & -\dot{R}(\tau) & 1 & 0 \\ \dot{R}(\tau) & -\ddot{R}(\tau) & 0 & -\ddot{R}(0) \end{pmatrix},$$

where we introduced $C_U(\tau) = \sigma_U^2 R(\tau)$ and used $\dot{C}_U(0) = 0$ for stationary processes with a differentiable correlation function. Inverting $\mathbf{C}$ is cumbersome and eventually leads to

$$\mathbf{C}^{-1} = \frac{1}{\det(\mathbf{C})} \begin{pmatrix} \alpha & \beta & \gamma & \delta \\ \beta & \epsilon & -\delta & \zeta \\ \gamma & -\delta & \alpha & -\beta \\ \delta & \zeta & -\beta & \epsilon \end{pmatrix} \quad \text{with}$$

$$\alpha = \ddot{R}(0)^2 + \dot{R}(\tau)^2\ddot{R}(0) - \ddot{R}(\tau)^2,$$

$$\beta = R(\tau)\dot{R}(\tau)\ddot{R}(0) - \dot{R}(\tau)\ddot{R}(\tau),$$
$$\gamma = -R(\tau)\ddot{R}(0)^2 + R(\tau)\ddot{R}(\tau)^2 - \dot{R}(\tau)^2\ddot{R}(\tau),$$
$$\delta = \dot{R}(\tau)\ddot{R}(0) - R(\tau)\dot{R}(\tau)\ddot{R}(\tau) + \dot{R}(\tau)^3,$$
$$\epsilon = -\ddot{R}(0) + R(\tau)^2\ddot{R}(0) - \dot{R}(\tau)^2,$$
$$\zeta = \ddot{R}(\tau) - R(\tau)^2\ddot{R}(\tau) + R(\tau)\dot{R}(\tau)^2.$$

The determinant of $\mathbf{C}$ is given by

$$\det(\mathbf{C}) = [1 - R(\tau)^2][\ddot{R}(0)^2 - \ddot{R}(\tau)^2] + \dot{R}(\tau)^2[2\ddot{R}(0) - 2R(\tau)\ddot{R}(\tau) + \dot{R}(\tau)^2].$$

Now, we have to solve the integrals. The exponent of $p(U_2, \dot{U}_2; U_1, \dot{U}_1)$ takes the form

$$\mathbf{u}^T\mathbf{C}^{-1}\mathbf{u} = \frac{1}{\det(\mathbf{C})}\big[\epsilon\big(\dot{U}_1^2 + \dot{U}_2^2\big) + 2\zeta\dot{U}_1\dot{U}_2 + 2(\delta - \beta)\theta(\dot{U}_2 - \dot{U}_1) + 2(\alpha + \gamma)\theta^2\big].$$

With the transformation $v_1 = \frac{1}{\sqrt{2}}(\dot{U}_2 - \dot{U}_1)$ and $v_2 = \frac{1}{\sqrt{2}}(\dot{U}_2 + \dot{U}_1)$, we have $\dot{U}_1^2 + \dot{U}_2^2 = v_1^2 + v_2^2$, $\dot{U}_1\dot{U}_2 = \frac{1}{2}(v_2^2 - v_1^2)$ and thus

$$n_2(\tau) = \frac{e^{-\frac{(\alpha+\gamma)\theta^2}{\sigma_U^2\det(\mathbf{C})}}}{2\sqrt{\det\left(2\pi\sigma_U^2\mathbf{C}\right)}}\int_0^\infty e^{-\frac{(\epsilon+\zeta)v_2^2}{2\sigma_U^2\det(\mathbf{C})}} \times \int_{-v_2}^{v_2}\left(v_2^2 - v_1^2\right)e^{-\frac{(\epsilon-\zeta)v_1^2 + 2\sqrt{2}(\delta-\beta)\theta v_1}{2\sigma_U^2\det(\mathbf{C})}}dv_1dv_2.$$

The substitution $\tilde{v}_i = v_i/\sqrt{2\sigma_U^2\det(\mathbf{C})}$ simplifies the integrals to

$$n_2(\tau) = \frac{\det(\mathbf{C})^{3/2}}{2\pi^2}e^{-\frac{\left((\alpha+\gamma) - \frac{(\beta-\delta)^2}{(\epsilon-\zeta)}\right)\theta^2}{\sigma_U^2\det(\mathbf{C})}}\int_0^\infty e^{-(\epsilon+\zeta)\tilde{v}_2^2} \times \int_{-\tilde{v}_2}^{\tilde{v}_2}\left(\tilde{v}_2^2 - \tilde{v}_1^2\right)e^{-(\epsilon-\zeta)\left(\tilde{v}_1 - \frac{\beta-\delta}{\epsilon-\zeta}\frac{\theta}{\sqrt{\sigma_U^2\det(\mathbf{C})}}\right)^2}d\tilde{v}_1d\tilde{v}_2.$$

The inner integrals over $\tilde{v}_1$ can be solved in terms of error functions:

$$I_0(\tilde{v}_2; a, b) \equiv \int_{-\tilde{v}_2}^{\tilde{v}_2} e^{-a(\tilde{v}_1 - b)^2}d\tilde{v}_1 = \left[\frac{1}{2}\sqrt{\frac{\pi}{a}}\,\mathrm{erf}(\tilde{v}_1)\right]_{\sqrt{a}(b-\tilde{v}_2)}^{\sqrt{a}(b+\tilde{v}_2)},$$

$$I_1(\tilde{v}_2; a, b) \equiv \int_{-\tilde{v}_2}^{\tilde{v}_2} \tilde{v}_1^2 e^{-a(\tilde{v}_1 - b)^2}dx = \left[\frac{1 + 2ab^2}{4a^{3/2}}\sqrt{\pi}\,\mathrm{erf}(\tilde{v}_1)\right]_{\sqrt{a}(b-\tilde{v}_2)}^{\sqrt{a}(b+\tilde{v}_2)} + \left[-\frac{1}{2a^{3/2}}\tilde{v}_1e^{-\tilde{v}_1^2} + \frac{b}{a}e^{-\tilde{v}_1^2}\right]_{\sqrt{a}(b-\tilde{v}_2)}^{\sqrt{a}(b+\tilde{v}_2)},$$

where $a = \epsilon - \zeta$ and $b = \frac{\beta-\delta}{\epsilon-\zeta}\frac{\theta}{\sqrt{\sigma_U^2\det(\mathbf{C})}}$. Some of the outer integrals over $\tilde{v}_2$ can also be solved in terms of error functions:

$$I_2(a, b, c) \equiv -\frac{b}{a}\int_0^\infty e^{-c\tilde{v}_2^2}[e^{-\tilde{v}_1^2}]_{\sqrt{a}(b-\tilde{v}_2)}^{\sqrt{a}(b+\tilde{v}_2)}d\tilde{v}_2 = \frac{b}{a}\sqrt{\frac{\pi}{a+c}}e^{-ab^2 + \frac{a^2b^2}{a+c}}\,\mathrm{erf}\left(\frac{ab}{\sqrt{a+c}}\right),$$

$$I_3(a, b, c) \equiv \frac{1}{2a^{3/2}}\int_0^\infty e^{-c\tilde{v}_2^2}[\tilde{v}_1e^{-\tilde{v}_1^2}]_{\sqrt{a}(b-\tilde{v}_2)}^{\sqrt{a}(b+\tilde{v}_2)}d\tilde{v}_2 = \frac{1}{2a(a+c)}e^{-ab^2} - \frac{bc\sqrt{\frac{\pi}{a+c}}}{2a(a+c)}e^{-ab^2 + \frac{a^2b^2}{a+c}}\,\mathrm{erf}\left(\frac{ab}{\sqrt{a+c}}\right),$$

with $c = \epsilon + \zeta$. The remaining integrals over $\tilde{v}_2$, i.e.

$$I_4(a, b, c) \equiv -\frac{1 + 2ab^2}{4a^{3/2}}\sqrt{\pi}\int_0^\infty e^{-c\tilde{v}_2^2}[\mathrm{erf}(\tilde{v}_1)]_{\sqrt{a}(b-\tilde{v}_2)}^{\sqrt{a}(b+\tilde{v}_2)}d\tilde{v}_2,$$

$$I_5(a, b, c) \equiv \frac{1}{2}\sqrt{\frac{\pi}{a}}\int_0^\infty \tilde{v}_2^2e^{-c\tilde{v}_2^2}[\mathrm{erf}(\tilde{v}_1)]_{\sqrt{a}(b-\tilde{v}_2)}^{\sqrt{a}(b+\tilde{v}_2)}d\tilde{v}_2,$$

can be solved in terms of Owen's T function $T(h, a) = \frac{1}{2\pi}\int_0^a \frac{1}{1+x^2}e^{-\frac{1}{2}h^2(1+x^2)}dx$ (Ref. [59], see Appendix B 2). Combining everything, we obtain

$$n_2(\tau) = \frac{\det(\mathbf{C})^{3/2}}{(2\pi)^2ac}I_{\mathrm{ana}}(\tilde{a}, \tilde{b}, \tilde{c}, \tilde{d}),$$
$$I_{\mathrm{ana}}(\tilde{a}, \tilde{b}, \tilde{c}, \tilde{d}) = e^{-\tilde{d}^2} + \sqrt{\pi}(1 + \tilde{c})\tilde{b}e^{\tilde{b}^2 - \tilde{d}^2}\,\mathrm{erf}\left(\tilde{b}\right) + 2\pi\sqrt{\tilde{c}}(1/\tilde{c} - 2\tilde{a}^2 - 1)e^{\tilde{a}^2 - \tilde{d}^2} \times T(\sqrt{2\tilde{c}}\tilde{b}, 1/\sqrt{\tilde{c}}), \quad \text{(B6)}$$

with $\tilde{a} = \sqrt{a}b = \frac{\beta-\delta}{\sqrt{\epsilon-\zeta}}\frac{\theta}{\sqrt{\sigma_U^2\det(\mathbf{C})}}$, $\tilde{b} = \frac{a}{\sqrt{a+c}}b = \frac{\beta-\delta}{\sqrt{2\epsilon}}\frac{\theta}{\sqrt{\sigma_U^2\det(\mathbf{C})}}$, $\tilde{c} = \frac{c}{a} = \frac{\epsilon+\zeta}{\epsilon-\zeta}$, $\tilde{d} = \sqrt{\alpha+\gamma}\frac{\theta}{\sqrt{\sigma_U^2\det(\mathbf{C})}}$. From $n_2(\tau)$, we obtain

$$Q(\tau) = 1 - \frac{n_2(\tau)}{n_0^2} \quad \text{and} \quad \eta = 2\int_0^\infty Q(\tau)d\tau$$

which allow us to evaluate the Stratonovich approximation.

APPENDIX C: STRATONOVICH APPROXIMATION

Here, we compare the full Stratonovich approximation Eq. (34),

$$H_S(T) = -\int_0^T n_1(t)\frac{\ln\left(1 - \int_0^T Q(t, t')n_1(t')dt'\right)}{\int_0^T Q(t, t')n_1(t')dt'}dt,$$

with its approximation Eq. (35),

$$h_S(t) = \frac{\kappa_S}{n_0}n_1(t), \quad \kappa_S = -\frac{1}{\eta}\ln(1 - n_0\eta).$$

Importantly, both lead to equivalent hazard functions for infinite times [40].

To see this, we need two properties of n_1 and n_2. First, the upcrossing probability saturates at a finite value once the

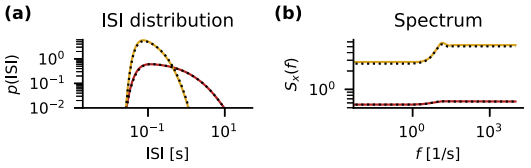

FIG. 12. Comparison of the full Stratonovich approximation with its approximation Eq. (35). (a) ISI distribution from Stratonovich approximation (colors) and Eq. (35) (black). (b) Same for the power spectra. Parameters as in Figs. 8(g) and 8(h).

transient effect of the voltage reset is over, $n_0 = \lim_{t\to\infty} n_1(t)$. Second, $Q(t_1,t_2) = 1 - \frac{n_2(t_1,t_2)}{n_1(t_1)n_1(t_2)}$ decays to zero for $|t_2 - t_1| \to \infty$ because the upcrossings decorrelate, $n_2(t_1,t_2) \to n_1(t_1)n_1(t_2)$. Thus one can approximate $\int_0^T Q(t,t')n_1(t')dt' \approx n_0 \int_0^\infty Q(t,t')dt' \equiv n_0\eta$ for $0 \ll t \ll T$. Then, neglecting the contributions of $\int_0^T Q(t,t')n_1(t')dt' - n_0\eta$ for t close to 0 or T leads to $H_S(T) \approx \int_0^T \frac{\kappa_S}{n_0} n_1(t)dt$. Neglecting these contributions is justified for large T because the integral is dominated by the contributions in between these boundaries. Hence, we arrive at $\lim_{T\to\infty} \frac{d}{dT} H_S(T) = \lim_{t\to\infty} h_S(t) = \kappa_S$ or, in terms of the ISI distribution, $p(T) \sim \exp(-\kappa_S T)$ for $t \to \infty$.

Since the long-time asymptotics are the same, differences can only occur at short times. In Fig. 12, we compare the full Stratonovich approximation with Eq. (35) for two representative examples. Fortunately, both the resulting ISI distributions [Fig. 12(a)] and the power spectra [Fig. 12(b)] agree closely for all times. Solving the full Stratonovich is numerically challenging (see below); thus, we use the simpler and more efficient approximation throughout in the main text.

1. Numerics

Here, we develop a numerical implementation of the Stratonovich approximation that is feasible for long time intervals without excessive demands on the working memory.

For stationary $Q(t,t') = Q(|t'-t|)$, the Stratonovich approximation Eq. (34) reads

$$H_S(T) = -\int_0^T n_1(t) \frac{\ln\left[1 - \int_0^T Q(|t'-t|)n_1(t')dt'\right]}{\int_0^T Q(|t'-t|)n_1(t')dt'} dt.$$

With the definition

$$f(T,t) = \int_0^T Q(|t'-t|)n_1(t')dt',$$

we have

$$H_S(T) = -\int_0^T n_1(t) \frac{\ln[1 - f(T,t)]}{f(T,t)} dt.$$

Since $Q(\tau \to \infty) \to 0$, i.e., it vanishes for long time lags, we can introduce an associated timescale: $Q(\tau) \approx 0$ for all $\tau > \tau_Q$. Similarly, $n_1(t \to \infty) \to n_0$ on the timescale τ_n such that $n_1(t) \approx n_0$ for all $t > \tau_n$.

The main problem in computing $H_S(T)$ is that a large three-dimensional grid is necessary for the three time arguments t, t', and T. To circumvent this problem, we split the domain of integration such that the full grid is only needed in small subdomains. In the remainder of the domain, the integrals can be solved by successive one-dimensional integration.

We consider $f(T,t)$ first. Because $Q(|t'-t|)$ vanishes for $|t'-t| > \tau_Q$, we know that the integrand only contributes in the vicinity of t. Thus we can extend the upper limit to infinity, $f(T,t) \approx f(\infty,t)$ if $t < T - \tau_Q$. Accordingly, we split the integral where possible:

$$H_S^{T\leqslant\tau_Q}(T) = -\int_0^T n_1(t) \frac{\ln[1 - f(T,t)]}{f(T,t)} dt,$$

$$H_S^{T>\tau_Q}(T) \approx -\int_0^{T-\tau_Q} n_1(t) \frac{\ln[1 - f(\infty,t)]}{f(\infty,t)} dt + R^{T>\tau_Q}(T),$$

$$R^{T>\tau_Q}(T) = -\int_{T-\tau_Q}^T n_1(t) \frac{\ln[1 - f(T,t)]}{f(T,t)} dt.$$

The remainder $R^{T>\tau_Q}(T)$ becomes constant for $T > \tau_n + 2\tau_Q$ because $n_1(t) \approx n_0$ in both integrals in this regime and we can set $R^{T>\tau_n+2\tau_Q}(T) \approx R^{T>\tau_n+2\tau_Q}(\tau_n + 2\tau_Q)$. Hence, we only have to calculate the full integral for $H_S^{T\leqslant\tau_Q}(T)$ and for $R^{T>\tau_Q}(T)$ until it is constant.

The remaining integrals in $H_S^{T>\tau_Q}(T)$ can be solved successively. First, we solve the convolution integral

$$f(\infty,t) = \int_0^\infty Q(|t'-t|)n_1(t')dt'$$

using Fourier transformation. Then, we can insert the result in $H_S^{T>\tau_Q}(T)$ and solve the integral over t. All integrals are approximated by their respective Riemann sum.

[1] A. Bernacchia, H. Seo, D. Lee, and X.-J. Wang, A reservoir of time constants for memory traces in cortical neurons, Nat. Neurosci. **14**, 366 (2011).

[2] J. D. Murray, A. Bernacchia, D. J. Freedman, R. Romo, J. D. Wallis, X. Cai, C. Padoa-Schioppa, T. Pasternak, H. Seo, D. Lee *et al.*, A hierarchy of intrinsic timescales across primate cortex, Nat. Neurosci. **17**, 1661 (2014).

[3] C. J. Honey, T. Thesen, T. H. Donner, L. J. Silbert, C. E. Carlson, O. Devinsky, W. K. Doyle, N. Rubin, D. J. Heeger, and U. Hasson, Slow cortical dynamics and the accumulation of information over long timescales, Neuron **76**, 423 (2012).

[4] C. A. Runyan, E. Piasini, S. Panzeri, and C. D. Harvey, Distinct timescales of population coding across cortex, Nature (London) **548**, 92 (2017).

[5] T. Ogawa and H. Komatsu, Differential temporal storage capacity in the baseline activity of neurons in macaque frontal eye field and area V4, J. Neurophysiol. **103**, 2433 (2010).

[6] M. D. Mauk and D. V. Buonomano, The neural basis of temporal processing, Annu. Rev. Neurosci. **27**, 307 (2004).

[7] R. Rossi-Pool, A. Zainos, M. Alvarez, S. Parra, J. Zizumbo, and R. Romo, Invariant timescale hierarchy across the cortical somatosensory network, Proc. Natl. Acad. Sci. USA **118**, e2021843118 (2021).

[8] R. Duarte, A. Seeholzer, K. Zilles, and A. Morrison, Synaptic patterning and the timescales of cortical dynamics, Curr. Opin. Neurobiol. **43**, 156 (2017).

[9] X.-J. Wang, Macroscopic gradients of synaptic excitation and inhibition in the neocortex, Nat. Rev. Neurosci. **21**, 169 (2020).

[10] R. Chaudhuri, K. Knoblauch, M.-A. Gariel, H. Kennedy, and X.-J. Wang, A large-scale circuit mechanism for hierarchical dynamical processing in the primate cortex, Neuron **88**, 419 (2015).

[11] J. M. Huntenburg, P.-L. Bazin, and D. S. Margulies, Large-scale gradients in human cortical organization, Trends Cognit. Sci. **22**, 21 (2018).

[12] A. Goulas, K. Zilles, and C. C. Hilgetag, Cortical gradients and laminar projections in mammals, Trends Neurosci. **41**, 775 (2018).

[13] S. Marom, Neural timescales or lack thereof, Prog. Neurobiol. **90**, 16 (2010).

[14] J. Wilting and V. Priesemann, Between perfectly critical and fully irregular: A reverberating model captures and predicts cortical spike propagation, Cereb. Cortex **29**, 2759 (2019).

[15] H. Sompolinsky, A. Crisanti, and H. J. Sommers, Chaos in Random Neural Networks, Phys. Rev. Lett. **61**, 259 (1988).

[16] A. Crisanti and H. Sompolinsky, Path integral approach to random neural networks, Phys. Rev. E **98**, 062120 (2018).

[17] M. Helias and D. Dahmen, *Statistical Field Theory for Neural Networks* (Springer, Cham, 2020), Vol. 970, p. 203.

[18] J. Kadmon and H. Sompolinsky, Transition to Chaos in Random Neuronal Networks, Phys. Rev. X **5**, 041030 (2015).

[19] C. Huang and B. Doiron, Once upon a (slow) time in the land of recurrent neuronal networks, Curr. Opin. Neurobiol. **46**, 31 (2017).

[20] F. Mastrogiuseppe and S. Ostojic, Intrinsically-generated fluctuating activity in excitatory-inhibitory networks, PLOS Comput. Biol. **13**, e1005498 (2017).

[21] J. Schuecker, S. Goedeke, and M. Helias, Optimal Sequence Memory in Driven Random Networks, Phys. Rev. X **8**, 041029 (2018).

[22] M. Beiran and S. Ostojic, Contrasting the effects of adaptation and synaptic filtering on the timescales of dynamics in recurrent networks, PLOS Comput. Biol. **15**, e1006893 (2019).

[23] S. P. Muscinelli, W. Gerstner, and T. Schwalger, How single neuron properties shape chaotic dynamics and signal transmission in random neural networks, PLOS Comput. Biol. **15**, e1007122 (2019).

[24] W. Gerstner, W. M. Kistler, R. Naud, and L. Paninski, *Neuronal Dynamics. From Single Neurons to Networks and Models of Cognition* (Cambridge University Press, Cambridge, 2014).

[25] D. J. Amit and N. Brunel, Model of global spontaneous activity and local structured activity during delay periods in the cerebral cortex, Cereb. Cortex **7**, 237 (1997).

[26] N. Brunel, Dynamics of sparsely connected networks of excitatory and inhibitory spiking neurons, J. Comput. Neurosci. **8**, 183 (2000).

[27] T. Tetzlaff, M. Helias, G. T. Einevoll, and M. Diesmann, Decorrelation of neural-network activity by inhibitory feedback, PLOS Comput. Biol. **8**, e1002596 (2012).

[28] M. Helias, T. Tetzlaff, and M. Diesmann, The correlation structure of local cortical networks intrinsically results from recurrent dynamics, PLOS Comput. Biol. **10**, e1003428 (2014).

[29] O. Harish and D. Hansel, Asynchronous rate chaos in spiking neuronal circuits, PLOS Comput. Biol. **11**, e1004266 (2015).

[30] C. Keup, T. Kühn, D. Dahmen, and M. Helias, Transient Chaotic Dimensionality Expansion by Recurrent Networks, Phys. Rev. X **11**, 021064 (2021).

[31] C. Fulvi Mari, Random Networks of Spiking Neurons: Instability in the Xenopus Tadpole Moto-Neuron Pattern, Phys. Rev. Lett. **85**, 210 (2000).

[32] J. Hertz, A. Lerchner, and M. Ahmadi, Mean field methods for cortical dynamics, in *Computational Neuroscience: Cortical Dynamics*, edited by P. Érdi, A. Esposito, M. Marinaro, and S. Scarpetta (Springer-Verlag Berlin, Heidelberg, 2004), pp. 71–89.

[33] A. Lerchner, G. Sterner, J. Hertz, and M. Ahmadi, Mean field theory for a balanced hypercolumn model of orientation selectivity in primary visual cortex, Network: Comput. Neural Systems **17**, 131 (2006).

[34] A. Lerchner, C. Ursta, J. Hertz, M. Ahmadi, P. Ruffiot, and S. Enemark, Response variability in balanced cortical networks, Neural Comput. **18**, 634 (2006).

[35] B. Dummer, S. Wieland, and B. Lindner, Self-consistent determination of the spike-train power spectrum in a neural network with sparse connectivity, Front. Comput. Neurosci. **8**, 104 (2014).

[36] R. F. Pena, S. Vellmer, D. Bernardi, A. C. Roque, and B. Lindner, Self-consistent scheme for spike-train power spectra in heterogeneous sparse networks, Front. Comput. Neurosci. **12**, 9 (2018).

[37] S. Wieland, D. Bernardi, T. Schwalger, and B. Lindner, Slow fluctuations in recurrent networks of spiking neurons, Phys. Rev. E **92**, 040901(R) (2015).

[38] R. Zeraati, T. A. Engel, and A. Levina, A flexible Bayesian framework for unbiased estimation of timescales, bioRxiv (2021), doi:10.1101/2020.08.11.245944.

[39] T. Schwalger, M. Deger, and W. Gerstner, Towards a theory of cortical columns: From spiking neurons to interacting neural populations of finite size, PLOS Comput. Biol. **13**, e1005507 (2017).

[40] R. L. Stratonovich, *Topics in the Theory of Random Noise* (Gordon and Breach, New York, 1967).

[41] T. Verechtchaguina, I. M. Sokolov, and L. Schimansky-Geier, First passage time densities in resonate-and-fire models, Phys. Rev. E **73**, 031108 (2006).

[42] S. Vellmer and B. Lindner, Theory of spike-train power spectra for multidimensional integrate-and-fire neurons, Phys. Rev. Research **1**, 023024 (2019).

[43] T. C. Potjans and M. Diesmann, The cell-type specific cortical microcircuit: Relating structure and activity in a full-scale spiking network model, Cereb. Cortex **24**, 785 (2014).

[44] V. Braitenberg and A. Schüz, *Cortex: Statistics and Geometry of Neuronal Connectivity*, 2nd ed. (Springer-Verlag, Berlin, 1998).

[45] C. van Vreeswijk and H. Sompolinsky, Chaos in neuronal networks with balanced excitatory and inhibitory activity, Science **274**, 1724 (1996).

[46] K. D. Harris and A. Thiele, Cortical state and attention, Nat. Rev. Neurosci. **12**, 509 (2011).

[47] R. P. Feynman, A. R. Hibbs, and D. F. Styer, *Quantum Mechanics and Path Integrals: Emended Edition* (Dover, Mineola, 2010).

[48] G. B. Arous and A. Guionnet, Large deviations for Langevin spin glass dynamics, Probab. Theory Relat. Fields **102**, 455 (1995).

[49] A. van Meegen, T. Kühn, and M. Helias, Large-Deviation Approach to Random Recurrent Neuronal Networks: Parameter Inference and Fluctuation-Induced Transitions, Phys. Rev. Lett. **127**, 158302 (2021).

[50] A. Roxin, N. Brunel, D. Hansel, G. Mongillo, and C. van Vreeswijk, On the distribution of firing rates in networks of cortical neurons, J. Neurosci. **31**, 16217 (2011).

[51] J. Aljadeff, M. Stern, and T. Sharpee, Transition to Chaos in Random Networks with Cell-Type-Specific Connectivity, Phys. Rev. Lett. **114**, 088101 (2015).

[52] A. van Meegen and B. Lindner, Self-Consistent Correlations of Randomly Coupled Rotators in the Asynchronous State, Phys. Rev. Lett. **121**, 258302 (2018).

[53] T. S. Grigera, Everything you wish to know about correlations but are afraid to ask, arXiv:2002.01750.

[54] R. Gao, R. L. van den Brink, T. Pfeffer, and B. Voytek, Neuronal timescales are functionally dynamic and shaped by cortical microarchitecture, eLife **9**, e61277 (2020).

[55] M. Golesorkhi, J. Gomez-Pilar, F. Zilio, N. Berberian, A. Wolff, M. C. Yagoub, and G. Northoff, The brain and its time: intrinsic neural timescales are key for input processing, Commun. Biol. **4**, 970 (2021).

[56] J. Wilting and V. Priesemann, Inferring collective dynamical states from widely unobserved systems, Nat. Commun. **9**, 2325 (2018).

[57] T. Toyoizumi, K. R. Rad, and L. Paninski, Mean-field approximations for coupled populations of generalized linear model spiking neurons with Markov refractoriness, Neural Comput. **21**, 1203 (2009).

[58] M. Krumin and S. Shoham, Generation of spike trains with controlled auto- and cross-correlation functions, Neural Comput. **21**, 1642 (2009).

[59] D. B. Owen, A table of normal integrals, Commun. Stat. - Simul. Comput. **9**, 389 (1980).

[60] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes: The Art of Scientific Computing*, 3rd ed. (Cambridge University Press, Cambridge, 2007).

[61] T. Fardet, S. B. Vennemo, J. Mitchell, H. Mørk, S. Graber, J. Hahne, S. Spreizer, R. Deepu, G. Trensch, P. Weidel, J. Jordan, J. M. Eppler, D. Terhorst, A. Morrison, C. Linssen, A. Antonietti, K. Dai, A. Serenko, B. Cai, P. Kubaj *et al.*, NEST 2.20.1 (2020), doi:10.5281/zenodo.4018717.

[62] S. Ostojic, Two types of asynchronous activity in networks of excitatory and inhibitory spiking neurons, Nat. Neurosci. **17**, 594 (2014).

[63] S. O. Rice, Mathematical analysis of random noise, Bell Syst. Tech. J. **24**, 46 (1945), reprinted in [93].

[64] L. M. Ricciardi and S. Sato, A note on first passage time problems for Gaussian processes and varying boundaries, IEEE Trans. Inf. Theory **29**, 454 (1983).

[65] T. Schwalger, Mapping input noise to escape noise in integrate-and-fire neurons: A level-crossing approach, arXiv:2109.07416 [q-bio.NC].

[66] J.-M. Azaïs and M. Wschebor, *Level Sets and Extrema of Random Processes and Fields* (Wiley, Hoboken, 2009).

[67] B. Lindner, Interspike interval statistics of neurons driven by colored noise, Phys. Rev. E **69**, 022901 (2004).

[68] B. Kriener, H. Enger, T. Tetzlaff, H. E. Plesser, M.-O. Gewaltig, and G. T. Einevoll, Dynamics of self-sustained asynchronous-irregular activity in random networks of spiking neurons with strong synapses, Front. Comput. Neurosci. **8**, 136 (2014).

[69] E. Ullner, A. Politi, and A. Torcini, Quantitative and qualitative analysis of asynchronous neural activity, Phys. Rev. Research **2**, 023103 (2020).

[70] H. Bos, M. Diesmann, and M. Helias, Identifying anatomical origins of coexisting oscillations in the cortical microcircuit, PLOS Comput. Biol. **12**, e1005132 (2016).

[71] C. van Vreeswijk and H. Sompolinsky, Chaotic balanced state in a model of cortical circuits, Neural Comput. **10**, 1321 (1998).

[72] S. E. Cavanagh, L. T. Hunt, and S. W. Kennerley, A diversity of intrinsic timescales underlie neural computations, Front. Neural Circuits **14**, 81 (2020).

[73] D. Dahmen, S. Grün, M. Diesmann, and M. Helias, Second type of criticality in the brain uncovers rich multiple-neuron dynamics, Proc. Nat. Acad. Sci. USA **116**, 13051 (2019).

[74] R. Chaudhuri, A. Bernacchia, and X.-J. Wang, A diversity of localized timescales in network activity, eLife **3**, e01239 (2014).

[75] A. Litwin-Kumar and B. Doiron, Slow dynamics and high variability in balanced cortical networks with clustered connections, Nat. Neurosci. **15**, 1498 (2012).

[76] T. Rost, M. Deger, and M. P. Nawrot, Winnerless competition in clustered balanced networks: Inhibitory assemblies do the trick, Biol. Cybern. **112**, 81 (2018).

[77] V. Rostami, T. Rost, A. Riehle, S. J. van Albada, and M. P. Nawrot, Spiking neural network model of motor cortex with joint excitatory and inhibitory clusters reflects task uncertainty, reaction times, and variability dynamics, bioRxiv (2020), doi:10.1101/2020.02.27.968339.

[78] R. Kim and T. J. Sejnowski, Strong inhibitory signaling underlies stable temporal dynamics and working memory in spiking neural networks, Nat. Neurosci. **24**, 129 (2021).

[79] G. Deco and V. K. Jirsa, Ongoing cortical activity at rest: Criticality, multistability, and ghost attractors, J. Neurosci. **32**, 3366 (2012).

[80] M. Garagnani, G. Lucchese, R. Tomasello, T. Wennekers, and F. Pulvermüller, A spiking neurocomputational model of high-frequency oscillatory brain responses to words and pseudowords, Front. Comput. Neurosci. **10**, 145 (2017).

[81] R. Tomasello, M. Garagnani, T. Wennekers, and F. Pulvermüller, A neurobiologically constrained cortex model of semantic grounding with spiking neurons and brain-like connectivity, Front. Comput. Neurosci. **12**, 88 (2018).

[82] M. Schmidt, R. Bakker, C. C. Hilgetag, M. Diesmann, and S. J. van Albada, Multi-scale account of the network structure of macaque visual cortex, Brain Struct. Func. **223**, 1409 (2018).

[83] D. Sussillo, Neural circuits as computational dynamical systems, Curr. Opin. Neurobiol. **25**, 156 (2014).

[84] O. Barak, Recurrent neural networks as versatile tools of neuroscience research, Curr. Opin. Neurobiol. **46**, 1 (2017).

[85] G. Bellec, D. Salaj, A. Subramoney, R. Legenstein, and W. Maass, Long short-term memory and Learning-to-learn in networks of spiking neurons, in *Advances in Neural Information*

Processing Systems 31, edited by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Curran Associates, Inc., Red Hook, 2018), pp. 795–805.

[86] D. F. Wasmuht, E. Spaak, T. J. Buschman, E. K. Miller, and M. G. Stokes, Intrinsic neuronal dynamics predict distinct functional roles during working memory, Nat. Commun. **9**, 3499 (2018).

[87] M. Spitmaan, H. Seo, D. Lee, and A. Soltani, Multiple timescales of neural dynamics and integration of task-relevant signals across cortex, Proc. Nat. Acad. Sci. USA **117**, 22522 (2020).

[88] C. K. Williams and C. E. Rasmussen, *Gaussian Processes for Machine Learning*, 1st ed. (MIT Press, Cambridge, 2006).

[89] P. I. Kuznetsov and R. L. Stratonovich, A note on the mathematical theory of correlated random points, in *Non-Linear Transformations of Stochastic Processes* (Pergamon Press, Oxford, 1965), pp. 101–115.

[90] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey *et al.*, SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python, Nat. Methods **17**, 261 (2020).

[91] D. B. Owen, A special case of a bivariate non-central t-distribution, Biometrika **52**, 437 (1965).

[92] P. Hänggi and P. Jung, Colored Noise in Dynamical Systems, in *Advances in Chemical Physics 89*, edited by I. Prigogine and S. A. Rice (John Wiley & Sons, Hoboken, 1995), pp. 239– 326.

[93] N. Wax (ed.), *Selected Papers on Noise and Stochastic Processes* (Dover, New York, 1954).

7

UBIQUITOUS LOGNORMAL DISTRIBUTION OF NEURON DENSITIES ACROSS MAMMALIAN CEREBRAL CORTEX

PREAMBLE

While the previous chapter followed a gradient of increasing model complexity, this chapter strongly deviates from this organizing principle. Here, we investigate how neurons are distributed across cerebral cortex. Uncovering statistical regularities in the cortical organization is a necessary ingredient for large-scale, data-driven models like the one developed in Chapter 8.

We characterize the distribution of neuron densities both within and across areas for several mammalian species. The main finding is that they are log-normally distributed. More precisely, we cannot exclude the possibility that they are log-normally distributed and we do not find a model which outperforms the log-normal distribution.

This finding calls for an explanation. To this end, we develop a simple model of noisy cell division which leads to log-normal neuron densities within areas, akin to models of noisy synapse formation (e.g. Ziv and Brenner 2018).

Author Contributions

The results presented in this preprint were obtained jointly by Aitor Morales-Gregorio (AMG) and the author (AvM) under supervision of Prof. Sacha van Albada (SvA). AMG collected and curated the data and created the figures; AvM developed the model and focused on the statistical tests. The first draft of this manuscript was written by AMG and AvM and it was jointly revised by AMG, AvM, and SvA. AMG and AvM contributed equally to this manuscript.

Ubiquitous lognormal distribution of neuron densities across mammalian cerebral cortex

Aitor Morales-Gregorio,[1,2,*,†] Alexander van Meegen,[1,2,*] Sacha J. van Albada[1,2]

[1]Institute of Neuroscience and Medicine (INM-6) and Institute for Advanced Simulation (IAS-6) and JARA-Institut Brain Structure-Function Relationships (INM-10), Jülich Research Centre, Jülich, Germany
[2]Institute of Zoology, University of Cologne, Cologne, Germany
∗These authors contributed equally to this work.
[†]To whom correspondence should be addressed; e-mail: a.morales-gregorio@fz-juelich.de

Abstract

Numbers of neurons and their spatial variation are fundamental organizational features of the brain. Despite the large corpus of data available in the literature, the statistical distributions of neuron densities within and across brain areas remain largely uncharacterized. Here, we show that neuron densities are compatible with a lognormal distribution across cortical areas in several mammalian species. We find that this also holds true for uniformly sampled regions across cortex as well as within cortical areas. Our findings uncover a new organizational principle of cortical cytoarchitecture. The ubiquitous lognormal distribution of neuron densities adds to a long list of lognormal variables in the brain.

Introduction

Neurons are not uniformly distributed across the cerebral cortex; their density varies strongly across areas and layers [1]. The neuron density directly affects short-range as well as long-range neuronal connectivity [2, 3]. Elucidating the distribution of neuron densities across the brain therefore provides insight into its connectivity structure and, ultimately, cognitive function. Additionally, statistical distributions are essential for the construction of computational models, which rely on predictive relationships and organizational principles where the experimental data are missing [4, 5]. Previous quantitative studies have provided reliable estimates for cell densities across the cerebral cortex of rodents [6, 7, 8], non-human primates [8, 9, 10, 11, 12, 13], large carnivores [14], and humans [15, 1]. However, to the best of our knowledge, the univariate distribution of neuron densities across and within cortical areas has not yet been statistically characterized. Instead, most studies focus on qualitative and quantitative comparisons across species, areas, or cortical layers. Capturing the entire distribution is necessary because long-tailed, highly skewed distributions are prevalent in the brain [16] and invalidate the intuition—guided by the central limit theorem—that the vast majority of values are in a small region of a few standard deviations around the mean.

Here, we for the first time characterize the distribution of neuron densities ρ across mammalian cerebral cortex. Based on the sample histograms (Figure 1) we hypothesize that ρ follows a lognormal distribution, similar to many other neuroanatomical and physiological variables such as synaptic strengths, axonal widths, and cortico-cortical connection densities [16, 17, 18]. Using neuron density data from mouse (*Mus musculus*), marmoset (*Callithrix jacchus*), macaque (*Macaca mulatta*), human (*Homo sapiens*), galago (*Otolemur garnettii*), owl monkey (*Aotus nancymaae*), and baboon (*Papio cynocephalus anubis*) we confirm this hypothesis for the given species (see Cell density data for a detailed description of the data). Going

beyond the distribution across cortical areas, we furthermore show that neuron densities within most areas of marmoset cortex are also compatible with a lognormal distribution. Moreover, we show that the lognormal distribution can emerge during neurogenesis from a simple cell division model with variability. Finally, we compare with several other distributions and find that none outperform the lognormal distribution as a model of the data within and across cortex.

Results

To test for lognormality, we take the natural logarithm, $\ln(\rho)$, which converts lognormally distributed samples into normally distributed samples (Figure 1B). We then test for normality of $\ln(\rho)$ using the Shapiro-Wilk (SW) test, the most powerful among a number of commonly applied normality tests [19]. Large outliers ($|\text{z-scored } \ln(\rho)| \geq 3$; marked with a red cross in Figure 1C) were excluded from the normality test. The removed outliers are area $\mathrm{V1}$ in macaque and marmoset, which have densities far outside the range for all other areas in both species, and area APir in marmoset, which has a noticeably distinct cytoarchitecture with respect to the rest of the cerebral cortex [9]. We denote different data sets for the same species with subscript indices (see Cell density data). The SW test concludes that the normality hypothesis of $\ln(\rho)$ cannot be rejected for mouse, marmoset, macaque$_1$, human, galago$_1$, owl monkey, and baboon (see Figure 1B). For the data sets macaque$_2$ and galago$_2$ the normality hypothesis is rejected ($p < 0.05$); however, in these data sets, the densities were sampled neither uniformly nor based on a cytoarchitectonic parcellation. The normality hypothesis for the distribution across cytoarchitectonic areas is further supported by Figure 1C, which shows that the relation between theoretical quantiles and ordered samples is almost perfectly linear except for macaque$_2$ and galago$_2$. Next, we test the z-scored $\ln(\rho)$ from the different species and data sets against each other and find that they are pairwise statistically indistinguishable ($\alpha = 0.05$ level; two-sample two-sided Kolmogorov-Smirnov test, see Figure S1 for full test results).

Additionally, we control for cell types in the distributions of the mouse, galago$_1$, owl monkey, and baboon data. In the mouse data, different types of neurons and glia were labeled with specific genetic markers and their respective densities were reported separately for all cell types [7]. In the galago$_1$, owl monkey, and baboon data sets, the total numbers of cells and neurons were reported separately [11]. We show that all subtypes of neurons in the mouse are compatible with a lognormal distribution (Figure S2; SW test on $\ln(\rho)$, $p > 0.05$) while glia are not—with the notable exception of oligodendrocytes. When neurons and glia are pooled together (Figure S2C,D), the distribution of $\ln(\rho)$ still passes the SW normality test, likely due to the distribution being dominated by the neurons. Similar observations are made in the baboon data, where the glia do not pass the lognormality test, but the neurons do. In the cases of galago$_1$ and owl monkey both the neurons and glia pass the lognormality test (Figure S2), which may, however, be partly due to the small number of density samples (N=12 in both cases). Thus, the mouse and baboon data—with large samples sizes (N=42 and N=142, respectively)—suggest that it is the neuron densities that follow a lognormal distribution but not necessarily the glia densities.

Furthermore, we also perform a control test on the different types of staining—Nissl and NeuN—using the macaque$_1$ data. The staining methods differ in their treatment of glia: NeuN tends to label neuronal cell bodies only while Nissl indiscriminately labels both neurons and glia. We show that regardless of staining type the cell densities pass the lognormality test (Figure S3; SW test on $\ln(\rho)$ with $p > 0.05$), suggesting that counting some glia in the cell densities does not confound our analysis of the macaque$_1$ data.

Taken together, the normality test, the quantiles plots, the pairwise tests, the cell type comparison, and the staining method comparison provide compelling evidence that the logarithmized neuron densities are normally distributed across cytoarchitectonic areas. This also holds for uniformly sampled neuron densities (baboon) but not for a sampling that is neither uniform

nor based on a cytoarchitectonic parcellation (macaque_2, galago_2). Thus, the neuron densities are consistent with a lognormal distribution across the different cortical areas, as long as sampling is not irregular.

To investigate whether the lognormal distribution holds within cortical areas, we leverage numerical estimates of neuron density in marmoset [9]. Neurons were counted within $150 \times 150\ \mu m$ counting frames for four strips per cortical area, all originating from the same subject. The neuron densities within the counting frames ρ_s are the within-area samples; their sample distributions in three representative areas (MIP, V2, and V3; Figure 2A) again suggest a lognormal distribution. As before, we test for lognormality by testing $\ln(\rho_s)$ for normality with the SW-test (for full test results see Table S2). At significance level $\alpha = 0.05$, the normality hypothesis is not rejected for 86 out of 116 areas; whereas at $\alpha = 0.001$, this is the case for 112 out of 116 areas (Figure 2B,C). Thus, regardless of the precise significance threshold, the lognormality hypothesis cannot be rejected within most cortical areas in the marmoset cortex.

This finding raises the question how the intricate process of neurogenesis [20] culminates in lognormally distributed neuron densities in almost all areas. A simple model shows that there is no need for a specific regulatory mechanism: assuming that the proliferation of the neural progenitor cells is governed by a noisy rate

$$\lambda(t) = \mu(t) + \xi(t), \tag{1}$$

where $\mu(t)$ denotes the mean rate and $\xi(t)$ is a zero-mean Gaussian process, the resulting population of progenitor cells, and eventually neurons, is lognormally distributed (see Model of progenitor cell division with variability). Thus, the lognormal neuron density distribution within areas could be a hallmark of a cell division process with variability. The model furthermore predicts that the mean and variance of $\ln(\rho)$ increase with proliferation time. Since the proliferation time varies up to twofold between areas [20], mean and variance of $\ln(\rho)$ are correlated

across areas according to the model—indeed, they are significantly correlated in the marmoset data (Pearson $r = 0.32,\ p < 10^{-3}$, Figure S4).

To complement the statistical hypothesis tests on the logarithmic densities, we compared the lognormal model with six other statistical distributions based on the relative likelihood (see Statistical model comparison). We included statistical distributions with support in $\mathbb{R}^+$ since neuron densities cannot be negative: lognormal, truncated normal, inverse normal, gamma, inverse gamma, Lévy, and Weibull. Of those distributions the lognormal, inverse normal, and inverse gamma stand out as the distributions with the highest relative likelihoods, both across the entire cortex and within cortical areas (Figure S5A, Figure S6A). A visual inspection of the fitted distribution reveals that the lognormal, inverse normal, and inverse gamma produce virtually indistinguishable probability densities (Figure S5B, Figure S6C); indeed, the relative likelihoods of the three models are above 0.05 in all cases. This suggests that the data could theoretically be distributed according to either the lognormal, inverse normal, or inverse gamma distribution. However, out of these, the lognormal distribution could arise from a simple model of cell division (equation (1))—while no interpretable mechanisms leading to inverse normal or inverse gamma distributions are known in this context. Thus, the similar likelihood and a simple biophysical explanation together argue for a lognormal rather than an inverse normal or inverse gamma distribution of neuron densities.

Discussion

In conclusion, we show that neuron densities are compatible with a lognormal distribution across cortical areas in multiple mammalian cortices and within most cortical areas of the marmoset, uncovering a previously unexplored organizational principle of cerebral cortex. Furthermore, we propose a simple model, based on a cell division process of the progenitor cells with variability, that accounts for the emerging lognormal distributions within areas. Lastly, we show

that none of an extensive list of statistical models outperform the lognormal distribution. Our results are in agreement with the observation that surprisingly many characteristics of the brain follow lognormal distributions [16]. Moreover, this analysis highlights the importance of characterizing the statistical distributions of brain data because simple summary statistics—such as the mean or standard deviation—lack nuance and are not necessarily a good representation of the underlying distribution.

The distributions of neuron and cell densities in general depend on the underlying spatial sampling. We found that neuron densities follow a lognormal distribution within cytoarchitectonically defined areas, across such areas, and when averaged within small parcels uniformly sampled across cortex, but not when sampled in a highly non-uniform manner not following cytoarchitectonic boundaries. The observation of lognormality both within and across cytoarchitectonic areas as well as across small uniformly sized parcels suggests an interesting topic for further research: uncovering whether the neuron densities obey an invariance principle across scales.

In principle, cortex-wide organizational structures might be by-products of development or evolution that serve no computational function [21]—but the fact that we observe the same organizational principle for several species and across most cortical areas suggests that the lognormal distribution serves some purpose. Heterogeneous neuron densities could assist computation through their association with heterogeneity in other properties such as connectivity and neuronal time constants [4, 22]; indeed, such heterogeneity is known to be a valuable asset for neural computation [23, 24]. Alternatively, localized concentration of neurons in certain areas and regions could also serve a metabolic purpose [25], because centralization supports more efficient energy usage. This is particularly relevant since approximately half of the brain's energy consumption is used to support the communication between neurons [26]. Also from the perspective of cortical hierarchies it makes sense to have few areas with high neuron densities

and many areas with lower neuron densities: Low-density areas contain neurons with large dendritic trees [27] receiving convergent inputs from many neurons in high-density areas lower in the hierarchy. The neurons with extensive dendritic trees in higher areas are involved in different, area-specific abstractions of the low-level sensory information [28, 29]. There is probably not a single factor that leads to lognormal neuron densities in the cortex; further research will be needed to refine our findings and uncover the functional implications.

References

[1] C. von Economo, G. N. Koskinas, L. C. Triarhou, *Atlas of Cytoarchitectonics of the Adult Human Cerebral Cortex* (Karger, 2008).

[2] V. Braitenberg, A. Schüz, *Anatomy of the Cortex: Statistics and Geometry* (Springer-Verlag, Berlin, Heidelberg, New York, 1991).

[3] M. Ercsey-Ravasz, *et al.*, *Neuron* **80**, 184 (2013).

[4] C. C. Hilgetag, S. F. Beul, S. J. van Albada, A. Goulas **3**, 905 (2019).

[5] S. J. van Albada, *et al.*, *arXiv* (2020).

[6] S. Herculano-Houzel, C. Watson, G. Paxinos, *Frontiers in Neuroanatomy* **7** (2013).

[7] C. Erö, M.-O. Gewaltig, D. Keller, H. Markram, *Frontiers in Neuroinformatics* **12**, 84 (2018).

[8] C. J. Charvet, D. J. Cahalane, B. L. Finlay, *Cerebral Cortex* **25**, 147 (2015).

[9] N. Atapour, *et al.*, *Cerebral cortex* **29**, 3836 (2019).

[10] S. F. Beul, C. C. Hilgetag, *NeuroImage* **189**, 777 (2019).

[11] C. E. Collins, D. C. Airey, N. A. Young, D. B. Leitch, J. H. Kaas, *Proceedings of the National Academy of Sciences* **107**, 15927 (2010).

[12] C. E. Collins, *et al.*, *Proceedings of the National Academy of Sciences* **113**, 740 (2016).

[13] E. C. Turner, *et al.*, *Brain, Behavior and Evolution* **88**, 1 (2016).

[14] D. Jardim-Messeder, *et al.*, *Frontiers in Neuroanatomy* **11**, 118 (2017).

[15] C. S. von Bartheld, J. Bahney, S. Herculano-Houzel, *Journal of Comparative Neurology* **524**, 3865 (2016).

[16] G. Buzsáki, K. Mizuseki, *Nature Reviews Neuroscience* **15**, 264 (2014).

[17] N. T. Markov, *et al.*, *Cerebral Cortex* **24**, 17 (2014).

[18] P. A. Robinson, X. Gao, Y. Han, *Biological Cybernetics* **115**, 121 (2021).

[19] N. M. Razali, B. W. Yap, *Journal of Statistical Modeling and Analytics* **2**, 21 (2011).

[20] P. Rakic, *Nature Reviews Neuroscience* **3**, 65 (2002).

[21] A. G. Otopalik, A. C. Sutton, M. Banghart, E. Marder, *eLife* **6**, e23508 (2017).

[22] W. Rall, *Biophysical Journal* **9**, 1483 (1969).

[23] R. Duarte, A. Morrison, *PLOS Computational Biology* **15**, e1006781 (2019).

[24] N. Perez-Nieves, V. C. H. Leung, P. L. Dragotti, D. F. M. Goodman, *Nature Communications* **12**, 5791 (2021).

[25] M. Bélanger, I. Allaman, P. Magistretti, *Cell Metabolism* **14**, 724 (2011).

[26] S. B. Laughlin, T. J. Sejnowski, *Science* **301**, 1870 (2003).

[27] G. N. Elston, M. Rosa, *Cerebral Cortex* **8**, 278 (1998).

[28] S. Kumar, K. E. Stephan, J. D. Warren, K. J. Friston, T. D. Griffiths, *PLoS Computational Biology* **3**, e100 (2007).

[29] S. L. Brincat, M. Siegel, C. von Nicolai, E. K. Miller, *Proceedings of the National Academy of Sciences* **115** (2018).

[30] E. S. Lein, *et al.*, *Nature* **445**, 168 (2007).

[31] H. W. Dong, *The Allen reference atlas: A digital color brain atlas of the C57Bl/6J male mouse.* (John Wiley & Sons inc., 2008).

[32] G. Paxinos, C. R. R. Watson, M. Petrides, M. G. Rosa, H. Tokuno, *The Marmoset Brain in Stereotaxic Coordinates* (2012).

[33] N. G. Van Kampen, *Stochastic Processes in Physics and Chemistry* (North Holland, 2007), third edn.

[34] C. A. Braumann, *Mathematical Biosciences* **206**, 81 (2007).

Acknowledgements

We thank Günther Palm for useful discussions, Alexandre René for useful discussions and help with the Bayesian model comparison, Anno Kurth for discussions about geometric Brownian motion, and Jon Martinez Corral for proofreading an early draft. **Funding:** This work was supported by the European Union Horizon 2020 Framework Programme for Research and Innovation (Human Brain Project SGA2 grant number 785907 and HBP SGA3 grant number 945539) and the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Priority Program (SPP 2041 "Computational Connectomics") [S.J. van Albada: AL 2041/1-1 and

2041/1-2] and Open Access Publication Costs grant 491111487. **Author contributions:** Conceptualization AMG, AvM, SvA; Data curation AMG; Formal Analysis AMG, AvM; Funding acquisition SvA; Investigation AMG, AvM, SvA; Methodology AMG, AvM, SvA; Project administration SvA; Resources SvA; Software AMG, AvM; Supervision SvA; Validation AvM; Visualization AMG; Writing - original draft AMG, AvM; Writing – review & editing AMG, AvM, SvA. **Competing interests:** The authors declare no competing interests. **Data and materials availability:** This work produced no new data and instead relied on a corpus of neuron density data available from the literature, which we gratefully acknowledge; see Cell density data for a detailed description.

Figures

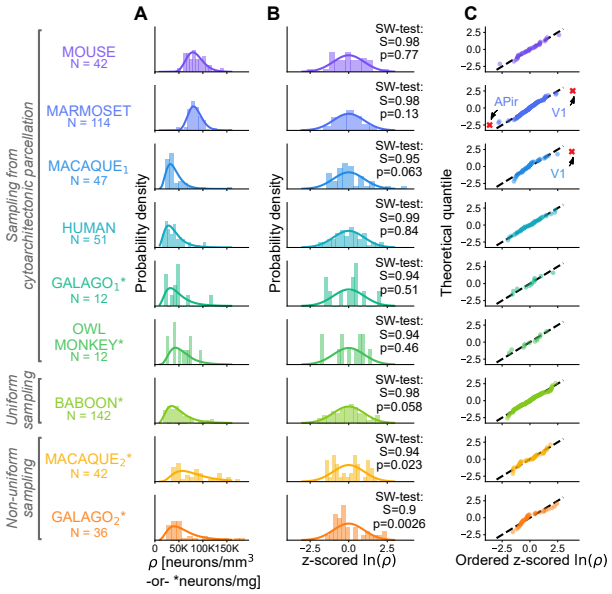

Figure 1: Neuron and cell densities ρ follow a lognormal distribution across cortical areas for multiple species. **A** Histogram of ρ (bars) and probability density function of a fitted lognormal distribution (line). **B** Z-scored $\ln(\rho)$ histogram (bars), standard normal distribution (line), and result of the Shapiro-Wilk normality test. **C** Probability plot of z-scored $\ln(\rho)$. Discarded outliers marked with a red cross.

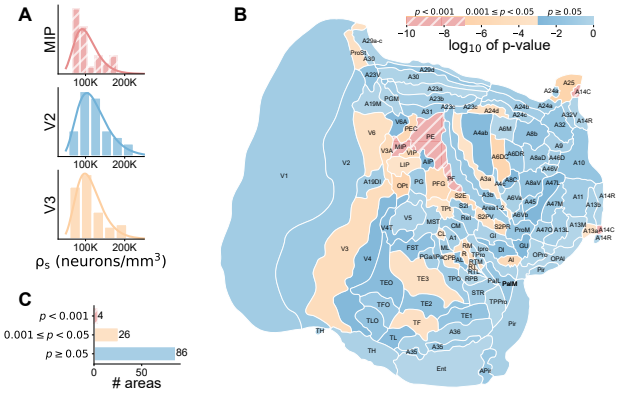

Figure 2: Neuron densities ρ_s follow a lognormal distribution within most areas of marmoset cortex. **A** Sample histograms of ρ_s and fitted lognormal distributions for three areas representing different degrees of lognormality. **B** Log_{10} of p-value of Shapiro-Wilk normality test of $\ln(\rho_s)$ on a flattened representation of the marmoset cortex [9]. **C** Number of areas with p-values in the given significance ranges.

Materials and methods

Cell density data

Estimates of neuron density for the available cortical areas across the mouse (*Mus musculus*), marmoset (*Callithrix jacchus*), macaque (*Macaca mulatta*), human (*Homo sapiens*), galago (*Otolemur garnettii*), owl monkey (*Aotus nancymaae*), and baboon (*Papio cynocephalus anubis*) cerebral cortex were used in this study.

In the cases of mouse, marmoset, macaque$_1$, human, galago$_1$, and owl monkey the data were sampled from standard cytoarchitectonic parcellations; abbreviated names for all areas are listed in Table S1. Note that we use subscript indices to distinguish between different data sets on the same model animal, e.g. macaque$_1$ and macaque$_2$.

Neuron density estimates for the mouse were published in [7], and were measured from Nissl-body-stained slices, where genetic markers were used to distinguish between cell types. The data were provided in the Allen Brain Atlas parcellation [30, 31].

Neuron density estimates for the marmoset cortex were published in [9], and were measured from NeuN-stained slices. The data were provided in the Paxinos parcellation [32]. Neuron densities within each counting frame used in the original publication [9] (their Figure S1) were obtained via personal communication with Nafiseh Atapour, Piotr Majka, and Marcello G. Rosa.

The neuron density estimates in the first macaque data set, macaque$_1$, were previously published in visual form in [10], and were obtained from both Nissl-body- and NeuN-stained brain slices. Counts based on Nissl-body staining were scaled according to a linear relationship with the counts from NeuN staining obtained from selected areas where both types of data were available [10]. The data follow the M132 parcellation [17] and numerical values were provided by Sarah F. Beul and Claus C. Hilgetag via personal communication.

Cell density estimates for the human cortex were previously published in [1], and were

measured from Nissl-body-stained brain slices. The human data therefore most likely reflect combined neuron and glia densities. The data were provided in the von Economo parcellation [1].

Cell and neuron density estimates for galago$_{1\&2}$, owl monkey, baboon, and macaque$_2$ were previously published in [11], and were measured using the isotropic fractionator method. The data are sampled from common parcellation schemes in galago$_1$ and owl monkey, approximately equal-size samples in the baboon, and irregular non-uniform samples in macaque$_2$ and galago$_2$.

Statistical model comparison

In order to assess which model is most compatible with the data, we compared the relative likelihood of different distributions against each other. We included an extensive list of distributions with support on $\mathbb{R}^+$, estimated the distributions' parameters using maximum likelihood, and calculated the Akaike Information Criterion (AIC)

$$AIC = 2k - 2\ln\mathcal{L} \tag{2}$$

where k is the number of estimated parameters of the model and $\mathcal{L}$ is the estimated maximum likelihood. We further compare the models using the relative likelihood ($\mathcal{L}_r$)

$$\mathcal{L}_r = e^{(AIC_{\min} - AIC_i)/2} \tag{3}$$

where $AIC_{\min}$ is the minimum AIC across all models and AIC_i is the AIC for the ith model. Note that the relative likelihood is equal to the relative likelihood if the number of estimated parameters is the same in both models. The relative likelihood indicates the probability that, from among the tested models, the ith model most strongly limits the information loss. We take a significance threshold of $\alpha = 0.05$ on the relative likelihood to determine whether a model is significantly worse than the best possible model.

Model of progenitor cell division with variability

We assume that the proliferation of the neural progenitor cells is governed by a noisy rate

$$\lambda(t) = \mu_{\text{rate}} + \sigma_{\text{rate}}\xi(t), \tag{4}$$

where μ_{rate} denotes the mean rate, $\xi(t)$ is a zero-mean Gaussian white noise process, and σ_{rate} controls the strength of the noise. During proliferation, we assume that the population size of the progenitor cells grows exponentially with rate λ, i.e., it obeys $\frac{d}{dt}N = \lambda N$. Dividing by a reference volume and inserting equation (4), we obtain a stochastic differential equation (SDE) for the density of progenitor cells ρ:

$$\frac{d}{dt}\rho = (\mu_{\text{rate}} + \sigma_{\text{rate}}\xi(t))\rho \tag{5}$$

We here use the Stratonovich interpretation, i.e., we assume that the noise process has a small but finite correlation time before taking the white-noise limit [33].

Working in the Stratonovich interpretation, we can transform the SDE to $\frac{d}{dt}\ln\rho = \mu_{\text{rate}} + \sigma_{\text{rate}}\xi(t)$ with the solution [34]

$$\ln\rho(t) = \ln\rho_0 + \mu_{\text{rate}}t + \sigma_{\text{rate}}\int_0^t \xi(s)ds. \tag{6}$$

Since $\xi(t)$ is Gaussian and equation (6) is linear, $\ln\rho(t)$ is Gaussian and hence $\rho(t)$ is log-normally distributed. The parameters of this lognormal distribution are $\mu(t) = \langle\ln\rho(t)\rangle$ and $\sigma^2(t) = \langle\Delta(\ln\rho(t))^2\rangle$. Using equation (6), $\langle\xi(s)\rangle = 0$, and $\langle\xi(s)\xi(s')\rangle = \delta(s-s')$, we obtain [34]

$$\mu(t) = \ln\rho_0 + \mu_{\text{rate}}t \qquad \text{and} \qquad \sigma^2(t) = \sigma^2_{\text{rate}}t. \tag{7}$$

Thus, the neuron densities resulting from the model of cell division with variability, equation (5), are lognormally distributed with parameters $\mu(t)$ and $\sigma^2(t)$ specified in equation (7). In particular, equation (7) predicts that both parameters increase with the proliferation time t.

The model can be generalized while still leading to a lognormal distribution of neuron densities: 1) The mean rate can be time-dependent, $\mu_{\text{rate}} = \mu_{\text{rate}}(t)$. 2) The noise process can be an arbitrary zero-mean (a non-zero mean can always be incorporated into $\mu_{\text{rate}}(t)$) Gaussian process with correlation function $C_\xi(t, t')$. Both generalizations allow one to incorporate a time dependence of mean and noise strength during the proliferation. Assuming an absence of correlation between noise and neuron density prior to $t = 0$, the above steps lead to the generalized solution

$$\ln \rho(t) = \ln \rho_0 + \int_0^t \mu_{\text{rate}}(s)ds + \int_0^t \xi(s)ds. \tag{8}$$

Here, $\ln \rho(t)$ is still a Gaussian process, because it is a linear transformation of the Gaussian process $\xi(t)$. Due to the marginalization property of Gaussian processes, $\ln \rho(t)$ is normally distributed for any fixed time t with parameters

$$\mu(t) = \ln \rho_0 + \int_0^t \mu_{\text{rate}}(s)ds \qquad \text{and} \qquad \sigma^2(t) = \int_0^t \int_0^t C_\xi(s, s')dsds'. \tag{9}$$

Thus, $\rho(t)$ is lognormally distributed with parameters $\mu(t)$ and $\sigma^2(t)$ specified in equation (9). Note that in equation (9), in contrast to equation (7), $\mu(t)$ and $\sigma^2(t)$ do not necessarily grow linearly with time but may exhibit a more intricate temporal dependence. Nonetheless, equation (9) predicts that $\mu(t)$ and $\sigma^2(t)$ are related through the proliferation time.

Supplementary tables

Table S1: Cortical areas included in this study.

Species	Area abbreviations
Mouse	FRP, MOp, MOs, SSp, SS-n, SSp-bfd, SSp-ll, SSp-m, SSp-ul, SS-tr, SSs, VISC, AUDd, AUDp, AUDpo, AUDv, VISal, VISam, VISl, VISp, VISpl, VISpm, ACAd, ACAv, ACAv, ACAv, ORBl, ORBm, ORBvl, AId, AIp, AIv, RSPagl, RSPd, RSPv, AONd, AONe, AONl, AONm, AONpv, TTd, TTv
Marmoset	A10, A9, A46V, A46D, A8aD, A8b, A8aV, A47L, A47M, A45, A47O, ProM, A11, A13b, A13a, A13L, A13M, OPAl, OPro, Gu, A32, A32V, A14R, A14C, A25, A24a, A24b, A24c, A24d, A6DR, A6Vb, A6Va, A8C, A6M, A6DC, A4c, A4ab, PaIM, AI, PaIL, DI, GI, IPro, TPro, S2PR, A3a, S2PV, A3b, S2I, S2E, A1-2, AuRTL, AuRT, AuRPB, AuRTM, AuR, AuRM, AuAL, AuA1, AuCM, AuCPB, AuML, AuCL, TPPro, STR, TE1, TPO, ReI, TE2, PGa-IPa, TPt, TE3, TEO, Pir, APir, Ent, A36, A35, TF, TL, TH, TLO, TFO, A23c, A23a, A29d, A30, A23b, A29a-c, A23V, ProSt, PF, PE, PFG, A31, AIP, PG, PEC, VIP, LIP, PGM, V6A, OPt, MIP, MST, FST, V5, V4T, A19M, V3A, V4, V6, A19DI, V3, V2, V1
Macaque$_1$	2, 5, 9, 10, 11, 12, 13, 14, 23, 25, 32, 24a, 24c, 24d, 46d, 46v, 7A, 7B, 7m, 8B, 8l, 8m, 8r, 9-46d, 9-46v, DP, ENTO, F1, F2, F3, F4, F5, F6, F7, LIP, MT, OPAI, OPRO, PERI, STPi, TEad, TEav, TEO, TH-TF, V1, V2, V3A, V4
Human	FA, FB, FC, FCBm, FD, FDΔ, FDt, FE, FF, FG, FH, FJ, FK, FL, FM, FN, LA1, LA2, LC1, LC2, LC3, LD, LE1, LE2, IA, IB, OA, OB, OC, PA, PB1, PB2, PC, PD, PE, PF, PG, PH, HA, HB, HC, HD, HE, HF, TA, TB, TC, TD, TE, TF, TG
Galago$_1$ & Owl Monkey	V1, V2, dV3, vV3, S1, M1, A1, MT, premotor, DL, Remain Ctx, Surr Ctx

Table S2: Results of the Shapiro-Wilk test for normality of $\ln(\rho_s)$ in marmoset cortical areas. Values rounded to two significant digits.

Area	S	p-value	Area	S	p-value	Area	S	p-value
V1	0.97	0.39	AI	0.95	0.0043	TH	0.97	0.66
A10	0.95	0.19	PaIL	0.95	0.33	TLO	0.96	0.18
A9	0.98	0.51	DI	0.97	0.098	TFO	0.97	0.26
A46V	0.98	0.56	GI	0.97	0.67	A23c	0.97	0.36
A46D	0.98	0.49	Ipro	0.97	0.66	A23a	0.99	0.98
A8aD	0.97	0.34	TPro	0.97	0.77	A29d	0.95	0.21
A8b	0.96	0.16	S2PR	0.92	0.006	A30	0.98	0.73
A8aV	0.96	0.17	A3a	0.95	0.04	A23b	0.97	0.45
A47L	0.96	0.052	S2PV	0.93	0.014	A29a-c	0.97	0.70
A47M	0.97	0.30	A3b	0.96	0.20	A23V	0.96	0.15
A45	0.96	0.18	S2I	0.97	0.33	ProSt	0.93	0.018
A47O	0.98	0.70	S2E	0.94	0.0046	PF	0.94	0.00083
ProM	0.97	0.21	Area1-2	0.97	0.37	PE	0.94	0.00065
A11	0.97	0.41	AuRTL	0.97	0.40	PFG	0.92	0.0046
A13b	0.96	0.58	AuRT	0.97	0.031	A31	0.97	0.31
A13a	0.91	0.048	AuRPB	0.98	0.89	AIP	0.96	0.063
A13L	0.97	0.45	AuRTM	0.97	0.73	PG	0.99	0.37
A13M	0.99	0.97	AuR	0.98	0.0093	PEC	0.91	0.0032
OPAl	0.99	0.99	AuRM	0.9	0.017	VIP	0.92	0.0044
OPro	0.98	0.75	AuAL	0.94	0.12	LIP	0.95	0.042
GU	0.95	0.058	AuA1	0.98	0.48	PGM	0.98	0.78
A32	0.97	0.20	AuCM	0.97	0.33	V6A	0.95	0.068
A32V	0.96	0.51	AuCPB	0.93	0.037	OPt	0.91	0.0015
A14R	0.98	0.77	AuML	0.97	0.44	MIP	0.9	0.00091
A14C	0.79	5.5e-06	AuCL	0.94	0.045	MST	0.98	0.53
A25	0.89	0.022	TPPro	0.98	0.91	FST	0.95	0.10
A24a	0.96	0.35	STR	0.96	0.44	V5	0.98	0.68
A24b	0.97	0.41	TE1	0.96	0.17	V4T	0.95	0.082
A24c	0.97	0.54	TPO	0.97	0.31	A19M	0.98	0.80
A24d	0.92	0.017	ReI	0.95	0.40	V3A	0.91	0.006
A6DR	0.97	0.23	TE2	0.96	0.15	V4	0.97	0.064
A6Vb	0.97	0.32	PGa/IPa	0.97	0.45	V6	0.96	0.017
A6Va	0.98	0.56	TPt	0.94	0.033	A19DI	0.95	0.074
A8C	0.95	0.055	TE3	0.93	0.026	V3	0.95	0.0076
A6M	0.99	0.98	TEO	0.95	0.087	V2	0.96	0.29
A6DC	0.91	0.002	A36	0.98	0.54	Ent	0.99	0.99
A4c	0.97	0.43	A35	0.97	0.31	APir	0.94	0.24
A4ab	0.96	0.076	TF	0.96	0.021	Pir	0.97	0.53
PaIM	0.93	0.20	TL	0.98	0.084			

Supplementary figures

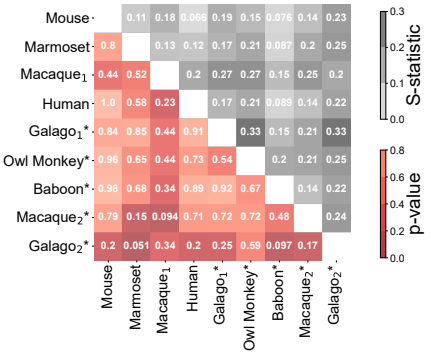

Figure S1: The z-scored log neuron density distributions of the four species are statistically indistinguishable at the 0.05 level based on pairwise Kolmogorov-Smirnov two-sample two-sided tests. P-values and S-statistics displayed below and above the diagonal, respectively.

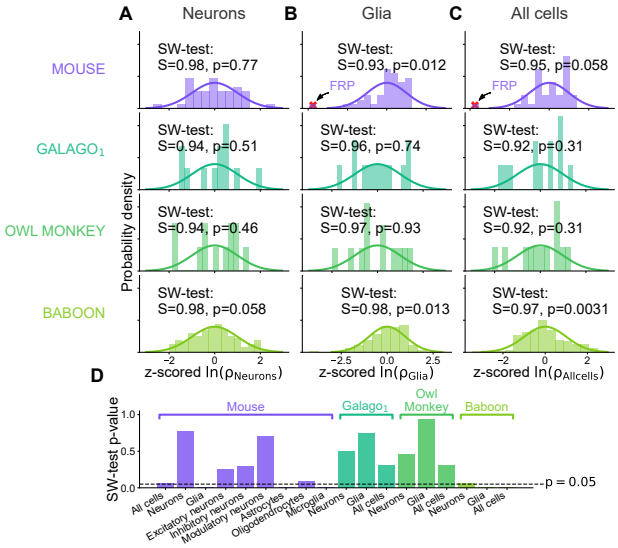

Figure S2: Comparison of neuron and glia lognormality. **A**–**C** Histogram of z-scored log density and result of Shapiro-Wilk test for neurons (**A**), glia (**B**), and all cells combined (**C**). **D** Barplot of p-values resulting from Shapiro-Wilk normality test for all cell types. Panel **A** is equivalent to the data shown in Figure 1.

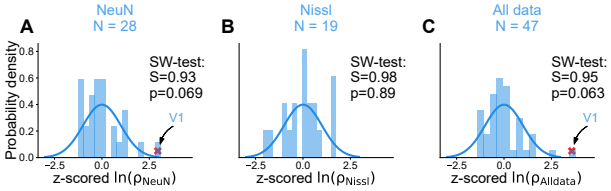

Figure S3: Lognormality of cell densities from different staining types in macaque cortex based on the macaque$_1$ data set. **A**-**C** Histogram of z-scored log density and result of Shapiro-Wilk test for NeuN staining only (**A**), Nissl staining only (**B**) and all measurements combined (**C**). The Nissl data were scaled down based on the linear relationship with the NeuN data [10]. Red crosses indicate outliers ($|\text{z-scored } \ln(\rho)| \geq 3$, which were excluded from the test. Panel **C** is equivalent to the data shown in Figure 1.

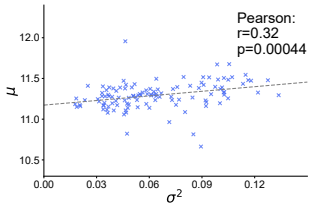

Figure S4: Neuron densities in the marmoset are compatible with our model of progenitor cell division with variability. μ and σ^2 are the mean and variance of $\ln(\rho)$, respectively; and are significantly correlated with each other, as predicted by the model.

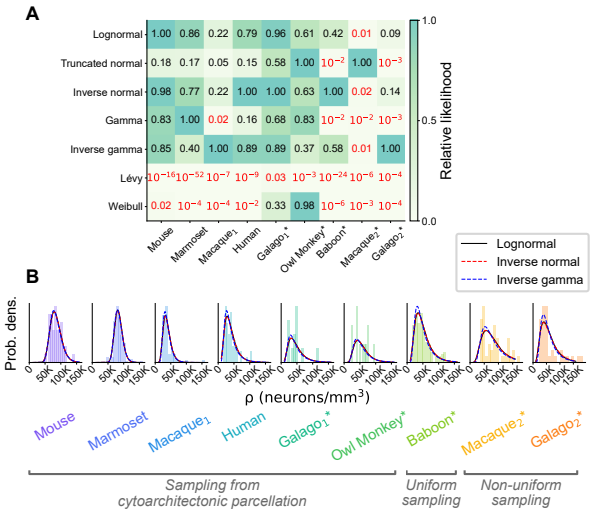

Figure S5: Statistical model comparison across the entire cortex of different animals. **A** Relative likelihood for seven compatible statistical models for all available area-level neuron density data sets; numerical values indicated for each model and animal. The red color indicates a relative likelihood < 0.05 with respect to the model with the highest likelihood. **B** The three best statistical models (according to the relative likelihood) fitted to the neuron density histograms in each animal; the three models produce visually nearly indistinguishable fits.

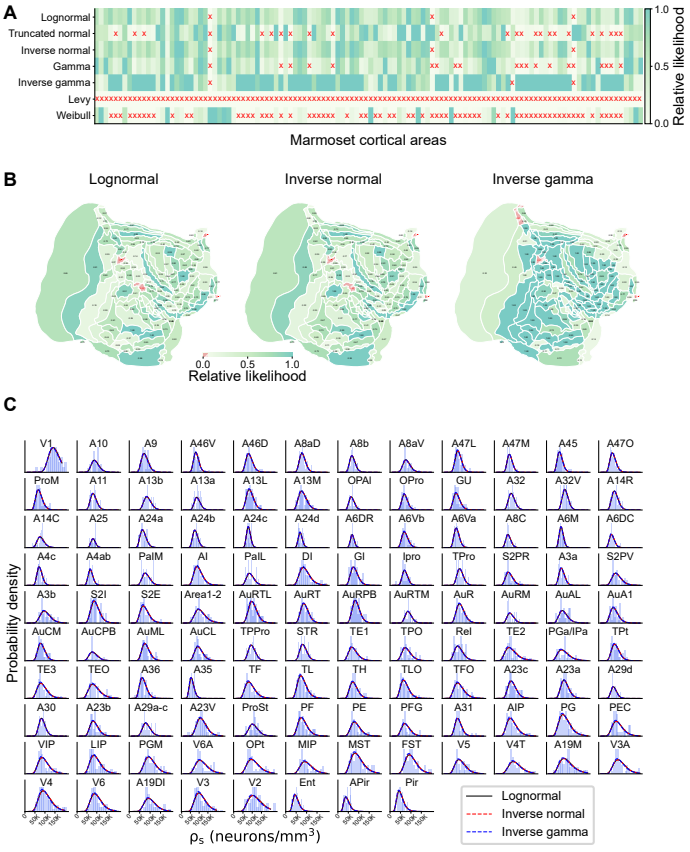

Figure S6: Statistical model comparison within the marmoset cortical areas. **A** Relative likelihood for seven compatible statistical models for all areas of the marmoset; a red cross (x) indicates a relative likelihood < 0.05 with respect to the model with the highest likelihood. **B** Spatial distribution of relative likelihood for the three best statistical models. **C** The three best statistical models fitted to the neuron density histograms in each area of marmoset cortex; the three models produce visually nearly indistinguishable fits.

8

MULTI-SCALE SPIKING NETWORK MODEL OF HUMAN CEREBRAL CORTEX

PREAMBLE

In this final chapter, we return to the gradient of increasing model complexity and consider a large-scale, data-driven, spiking model of human cortex. While the underlying connectivity is still block-structured random, we are leaving the realm of analytically tractable models and rely mostly on simulations. Because of the sheer size of the network, these simulations utilize the highly optimized NEST simulator (Gewaltig and Diesmann 2007) and are run on the JURECA-DC supercomputer at the Jülich Supercomputing Centre.

One of the major contributions of such data-based models is to collect the data and to bring it into a coherent framework (Pulvermüller et al. 2021). This exercise exposes gaps in the current knowledge and might lead to relevant follow-up work. Indeed, the results presented in Chapter 7 originated from an investigation of the relation between laminar origin of connections and cytoarchitecture which enters the model described in this chapter.

Historically, a variety of large-scale, data-driven models have been built (see Section 2.4 and Shimoura et al. 2021). The model presented in this chapter builds on the work by Schmidt, Bakker, Hilgetag, et al. (2018) and Schmidt, Bakker, Shen, et al. (2018). The distinctive feature of the model by Schmidt et al. is that it comprises all scales from single neurons through local circuits to networks of cortical areas. While the model by Schmidt et al. is based on macaque data, we take human data into account.

The final goal is to create a model which reproduces features of cortical activity and which, ideally, makes insightful predictions. For the first goal, we employ both single-neuron electrophysiological recordings and fMRI data to cover the full range from single neurons to all cortical areas. This is the current state of the project: we collected the data and aggregated it into a model which can be simulated and validated against activity data.

A major challenge is to attain a network state with strong inter-area interactions on slow timescales (compared to the single-neuron timescale) while keeping the firing rate in a physiological range. We have not yet achieved this goal. Thus, the manuscript in this chapter is unpublished and, potentially, subject to significant changes and extensions.

Author Contributions

Jari Pronold (JP) and the author (AvM) jointly worked, with equal contributions, on all parts of this project under supervision of Prof. Sacha van Albada (SvA). In addition, the manuscript contains results based on the master's thesis of Hannah Vollenbröker (HV) which was supervised by AvM. The codebase was written jointly, with equal contributions, by AvM and JP. JP focused more on aspects relating to simulations, software optimization, and fMRI data; AvM focused more on aspects relating to mean-field theory, neuron parameters, and spiking data.
Project ideas were developed and refined through joint discussions of AvM, JP, HV, Dr. Renan Shimoura (RS), Dr. Rembrandt Bakker (RB), and SvA. HV and RB contributed to the analysis of human neuron morphologies. Dr. Alexandros Goulas and Prof. Claus Hilgetag provided the DTI data and the cytoarchitectonic data by von Economo and Koskinas. Dr. Mario Senden provided the fMRI data and wrote the corresponding section in the manuscript.
The initial draft of the manuscript was written by AvM and it was jointly revised by AvM, JP, RS, and SvA. AvM and JP contributed equally to the work presented in this manuscript.

Multi-Scale Spiking Network Model of Human Cerebral Cortex

Alexander van Meegen[1,2,†,*], **Jari Pronold**[1,3,†,*], **Hannah Vollenbröker**[1,4], **Renan Shimoura**[1], **Mario Senden**[5,6], **Alexandros Goulas**[7], **Claus C. Hilgetag**[7], **Rembrandt Bakker**[1,8], **Sacha J. van Albada**[1,2,*]

***For correspondence:**
a.van.meegen@fz-juelich.de (AvM);
j.pronold@fz-juelich.de (JP);
s.van.albada@fz-juelich.de (SvA)

[†]These authors contributed equally to this work

[1]Institute of Neuroscience and Medicine (INM-6) and Institute for Advanced Simulation (IAS-6) and JARA-Institute Brain Structure-Function Relationships (INM-10), Jülich Research Centre, 52428 Jülich, Germany; [2]Institute of Zoology, University of Cologne, 50674 Cologne, Germany; [3]RWTH Aachen University, 52062 Aachen, Germany; [4]Heinrich Heine University Düsseldorf, 40225 Düsseldorf, Germany; [5]Department of Cognitive Neuroscience, Faculty of Psychology and Neuroscience, Maastricht University, 6229 ER Maastricht, The Netherlands; [6]Maastricht Brain Imaging Centre, Faculty of Psychology and Neuroscience, Maastricht University, 6229 ER Maastricht, The Netherlands; [7]Institute of Computational Neuroscience, University Medical Center Eppendorf, Hamburg University, 20246 Hamburg, Germany; [8]Donders Institute for Brain, Cognition and Behavior, Radboud University Nijmegen, 6525 EN Nijmegen, The Netherlands

Abstract Although the structure of cortical networks provides the necessary substrate for the neuronal activity, the structure alone does not suffice to understand it. Leveraging the increasing availability of human data, we developed a multi-scale, spiking network model of human cortex to investigate the relationship between structure and dynamics. In this model, each area in one hemisphere of the Desikan-Killiany parcellation is represented by a $1\,\mathrm{mm}^2$ column with a layered structure. The model aggregates data across multiple modalities, including electron microscopy, electrophysiology, morphological reconstructions, and DTI, into a coherent framework. It predicts activity on all scales from single-neuron spiking activity to the area-level functional connectivity. We compared the model activity against human electrophysiological data and human resting state fMRI data. This comparison reveals that further model adjustments are needed to account for the slow fluctuations in spiking activity and the inter-area functional connectivity observed experimentally.

Introduction

The brain is characterized by a multitude of spatial and temporal scales: from the molecular level to whole-brain networks, from sub-millisecond processes to memories that last decades (***Kandel et al., 2000***). Impressive technological advancements have made almost all these scales accessible through specialized techniques, which leads to a comprehensive but fragmented picture (***Sejnowski et al., 2014***). Models have the potential to integrate the diverse data modalities into a unified framework and to bridge across the scales (***Pulvermüller et al., 2021***).

Large-scale, data-driven models at cellular resolution have been constructed for sensory cortex (***Reimann et al., 2013***; ***Markram et al., 2015***; ***Girardi-Schappo et al., 2016***; ***Arkhipov et al., 2018***; ***Billeh et al., 2020***), prefrontal cortex (***Hass et al., 2016***), hippocampus (***Hendrickson et al., 2012***;

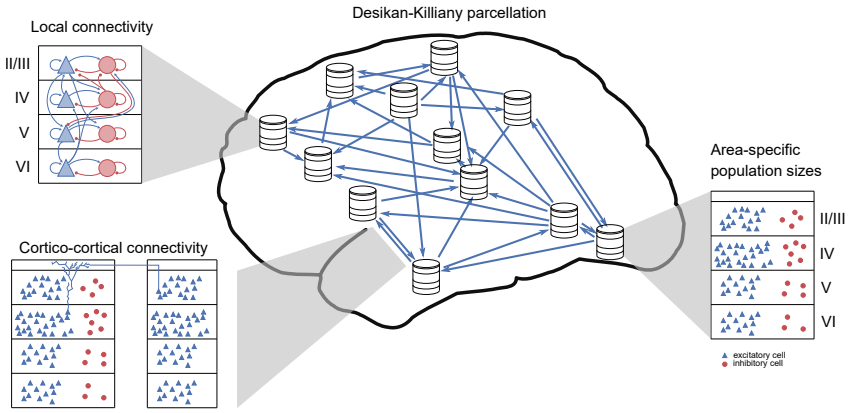

Figure 1. Model overview. The model comprises all 34 areas of the Desikan-Killiany parcellation (***Desikan et al., 2006***) in one hemisphere of human cerebral cortex. Each area is modeled by a column with $1\,\mathrm{mm}^2$ cortical surface. Within each column, the full number of neurons and synapses based on cytoarchitectonic data is included. Both the intrinsic and the cortico-cortical connectivity are layer- and population-specific.

Bezaire et al., 2016; ***Ecker et al., 2020***), and the olfactory bulb (***Migliore et al., 2014***, ***2015***), among others. These models reproduce resting-state activity (e.g. ***Potjans and Diesmann, 2014***; ***Markram et al., 2015***; ***Hass et al., 2016***; ***Bezaire et al., 2016***) and stimulus responses (e.g. ***Arkhipov et al., 2018***; ***Billeh et al., 2020***) on various levels of detail. Advances in the simulation technology for large networks of point neurons (***Einevoll et al., 2019***; ***Jordan et al., 2018***; ***Pronold et al., 2022***) have enabled the step beyond single areas to a multi-area network of vision related areas in macaque cortex (***Schmidt et al. 2018a***,b; see also ***Izhikevich and Edelman 2008*** for a pioneering study).

Due to the lack of available data in comparison with other species, only a single multi-scale human brain network model has been built so far (***Izhikevich and Edelman, 2008***). Leveraging the increasing availability of human data (e.g. ***Mohan et al., 2015***; ***Minxha et al., 2020***; ***Cano-Astorga et al., 2021***; ***Berg et al., 2021***; ***Shapson-Coe et al., 2021***), we build and simulate a model that encompasses the scales from the single-neuron level to the network of areas in one hemisphere of the human brain. The model aggregates data across many scales, from electron microscopy data for the density of synapses (***DeFelipe et al., 2002***; ***Cano-Astorga et al., 2021***) to whole-brain DTI and fMRI data, supplements it through predictive connectomics (e.g. ***Barbas and Rempel-Clower, 1997***; ***Ercsey-Ravasz et al., 2013***; ***Beul et al., 2017***; ***Hilgetag et al., 2019***; ***van Albada et al., 2022***), and yields activity data on all scales from single-neuron spiking activity to area-level correlation patterns.

First, we describe the data integration into a mesoscale connectome and its validation against features that were not explicitly built in. Then, we analyze the spiking activity of the model. Finally, we compare the model's activity to electrophysiological single-neuron spiking statistics and area-level correlation patterns based on fMRI.

Results

Human Mesoscale Connectome

The model comprises all 34 areas of one hemisphere of human cortex in the Desikan-Killiany parcellation (***Desikan et al., 2006***). Each area is modeled by a $1\ \mathrm{mm}^2$ column and the columns are connected through long-range projections (see Fig. 1).

We distinguish two classes of neurons, excitatory and inhibitory, and account for the layered structure of cortex. We determine the number of neurons from their volume density, the layer

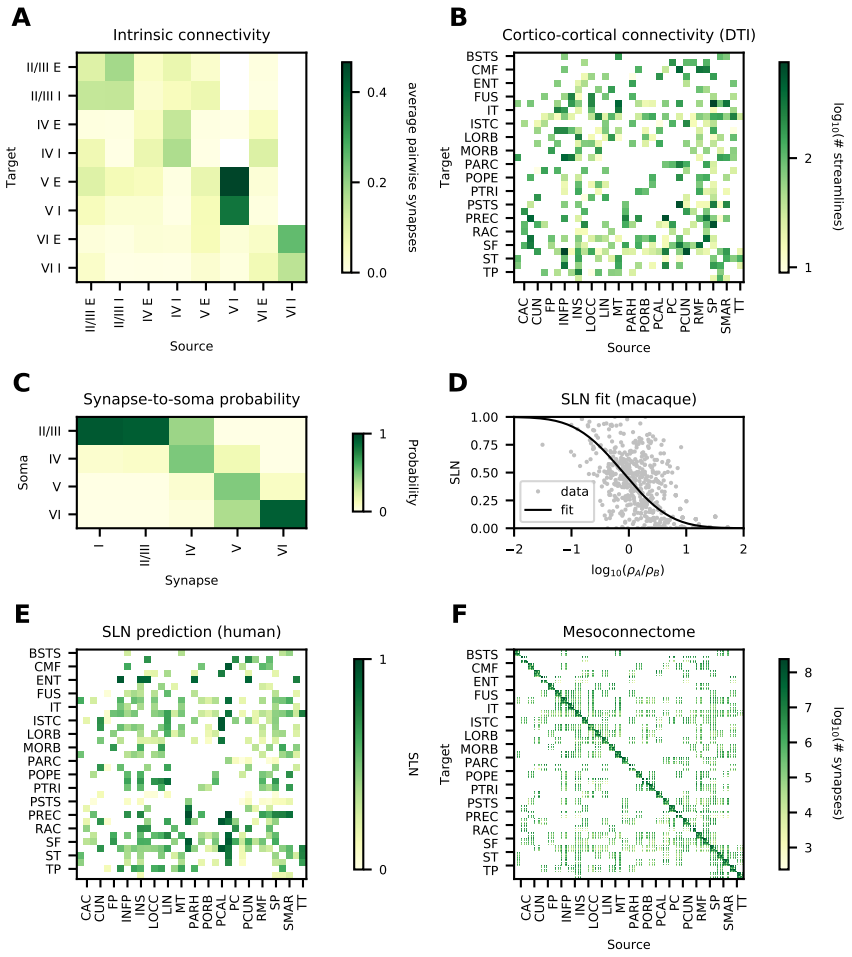

Figure 2. Data and predictive connectomics. (**A**) Within-area connectivity blueprint (average number of synapses per pair of neurons). (**B**) Cortico-cortical connectivity based on DTI (number of streamlines); see Table 1 for acronyms. (**C**) Probability for cortico-cortical synapses in a given layer to be established on neurons with cell body in a given layer, estimated from human neuron morphologies. (**D**) Relation of neuron densities of source area *B* and target area *A* with laminar source pattern (fraction of supragranular labeled neurons, SLN) in macaque. (**E**) Predicted source pattern (SLN) in human. (**F**) Layer- and population-resolved mesoconnectome (number of synapses).

Full name	Acronym	Full name	Acronym
bankssts	BSTS	parsorbitalis	PORB
caudalanteriorcingulate	CAC	parstriangularis	PTRI
caudalmiddlefrontal	CMF	pericalcarine	PCAL
cuneus	CUN	postcentral	PSTS
entorhinal	ENT	posteriorcingulate	PC
fusiform	FUS	precentral	PREC
inferiorparietal	INFP	precuneus	PCUN
inferiortemporal	IT	rostralanteriorcingulate	RAC
isthmuscingulate	ISTC	rostralmiddlefrontal	RMF
lateraloccipital	LOCC	superiorfrontal	SF
lateralorbitofrontal	LORB	superiorparietal	SP
lingual	LIN	superiortemporal	ST
medialorbitofrontal	MORB	supramarginal	SMAR
middletemporal	MT	frontalpole	FP
parahippocampal	PARH	temporalpole	TP
paracentral	PARC	transversetemporal	TT
parsopercularis	POPE	insula	INS

Table 1. All 34 areas in one hemisphere of the Desikan-Killiany parcellation with corresponding acronyms.

thickness, and the surface area of the column (***Von Economo, 2009***; cf. Sec. Neuron number). The parameters of the neurons are derived from the Allen Cell Types Database (https://celltypes.brain-map.org/; cf. Sec. Further Parameters of the Model). At this level of modeling, the connectivity statistics between neurons in both classes and all layers are needed, which are impossible to obtain with the current experimental techniques. Accordingly, we combine available data with predictive connectomics to arrive at a human mesoconnectome at a layer- and population-resolved level.

The lack of data on the connectivity is the main reason for considering only two classes of neurons. While a recent study defines 45 inhibitory and 24 excitatory neuron types in human (***Hodge et al., 2019***), including this diversity would require $69 \times 69 = 4761$ connection probabilities per pair of layers. This is not yet feasible because no connectivity data is available at such a fine granularity; hence, we restrict the model to two classes of neurons, as done in earlier studies (***Potjans and Diesmann, 2014***; ***Schmidt et al., 2018a***,b).

To derive the mesoconnectome, we start from a synapse-centric perspective. We approximate the volume density of synapses $\rho_{\text{synapse}} = 6.6 \times 10^8\ \text{synapses/mm}^3$ (***Cano-Astorga et al., 2021***) as constant across cortex (***DeFelipe et al., 2002***; ***Sherwood et al., 2020***), which allows us to compute the total number of synapses per layer based on their respective thickness (***Von Economo, 2009***). The task that remains is to determine the pre- and post-synaptic neurons of these synapses.

Data Aggregation & Predictive Connectomics

In a first step, we separate the synapses into local (within-area) connections and long-range projections by extrapolating the value of 79% local connections based on tracing data in macaque (***Markov et al., 2011***) to 86% in human using the power-law relation between local connections and total number of neurons (***Herculano-Houzel et al., 2010***; cf. Sec. Fraction of Cortico-Cortical Connections). For the local connections, we use the connection probabilities derived by ***Potjans and Diesmann*** (***2014***) (Fig. 2**A** and Local Connectivity) as a blueprint. The relative connection probabilities across source and target populations are kept constant, and they are only scaled by a constant factor to achieve the desired total number of local synapses in each area. The cortico-cortical connectivity on the area level is specified by DTI data from the Human Connectome Project (***Goulas et al., 2016***, which is based on the data from ***Van Essen et al., 2013***; Fig. 2**B** and Long-range projections). Synapses associated with long-range projections are assigned to postsynaptic neurons

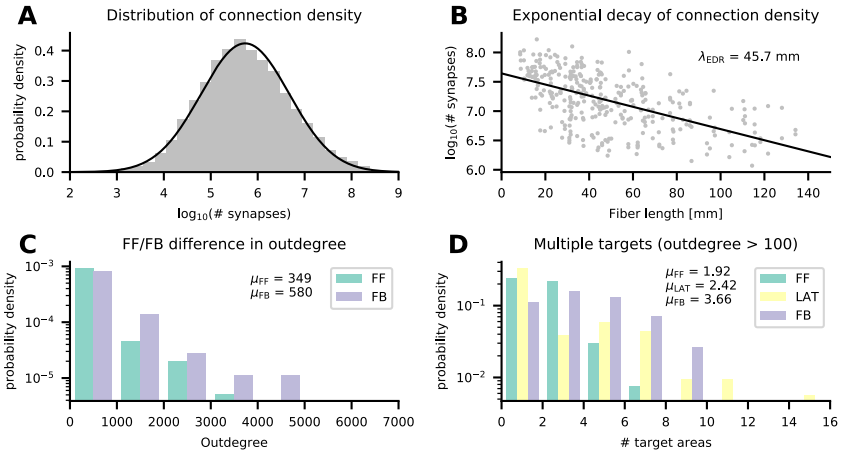

Figure 3. Connectivity validation. (**A**) Histogram of the number of synapses between pairs of populations (gray bars) and a log-normal fit (black line). (**B**) Logarithmic number of synapses between a pair of areas versus distance between these areas (gray symbols) and an exponential fit (black line). (**C**) Average outdegree of a neuron in any given population to any postsynaptic area in either feedforward (FF) or feedback (FB) direction. (**D**) Average number of target areas of a neuron in any given population to any postsynaptic area with average outdegree larger than 100 in either feedforward (FF), lateral (LAT), or feedback (FB) direction.

based on morphological reconstructions of human neurons (***Mohan et al., 2015***; Fig. 2**C** and Long-range projections).

The laminar origin of long-range projections is based on predictive connectomics. Retrograde tracing data in macaque shows that the laminar origin is systematically related to the cytoarchitecture (***Hilgetag et al., 2019***; Fig. 2**D**). Assuming that the same relation also holds in human, we use the fit in combination with the human cytoarchitecture to determine the laminar origin (Fig. 2**E**). For the laminar target, we assume that the relation between laminar origin and target derived in (***Schmidt et al., 2018a***) also holds in human.

Combining this data, we arrive at a human mesoconnectome which specifies the number of synapses between excitatory and inhibitory neurons for all areas in the Desikan-Killiany parcellation on a layer- and population-specific level (Fig. 2**F**).

Connectivity Validation

To validate the derived mesoconnectome, we compare it with anatomical features which were observed in other species but which were not explicitly built in.

The density of connections between areas is highly heterogeneous, spanning five orders of magnitude, and approximately log-normally distributed in mouse (***Gămănuț et al., 2018***), marmoset (***Theodoni et al., 2021***), and macaque (***Ercsey-Ravasz et al., 2013***). Similarly, in our model the number of synapses between pairs of populations span five orders of magnitude (Fig. 3**A**) and they are approximately log-normally distributed. Furthermore, the connection density decays exponentially with distance in mouse (***Horvát et al., 2016***), marmoset (***Theodoni et al., 2021***), and macaque (***Ercsey-Ravasz et al., 2013***). In our model, the number of synapses between pairs of areas also decays exponentially (Fig. 3**B**) with a decay constant of $45.7\,\mathrm{mm}$. Thus, two salient features of tracing data are captured by our model.

From anterograde tracing, it is known that feedback axons arborize more strongly than their feedforward counterparts (***Rockland, 2019***). This suggests a larger outdegree of feedback projections compared to feedforward projections. In our model, the average outdegree from neurons in

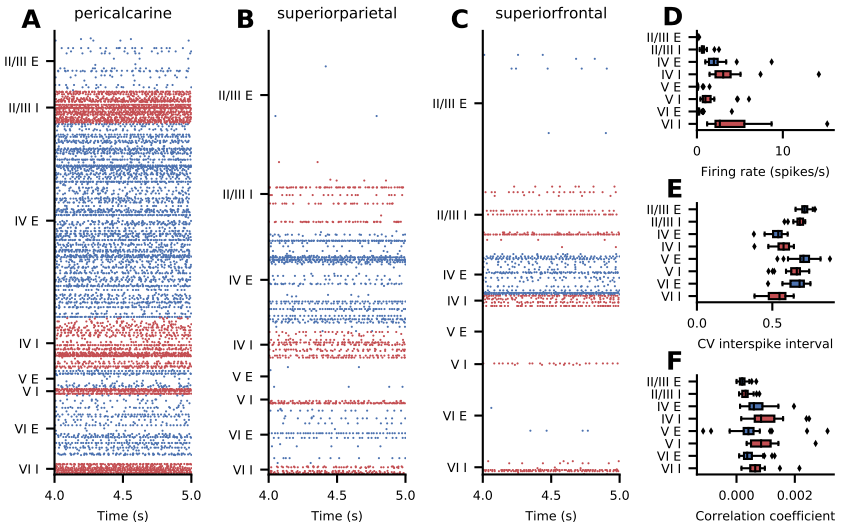

Figure 4. Simulated spiking activity. (**A**-**C**) Raster plots for three representative areas; subsampled to 2.5% of the excitatory (blue) and inhibitory (red) neurons. (**D**-**F**) Layer- and population-resolved distribution of population-averaged statistics across areas; boxes show quartiles, whiskers are within 1.5 times the interquartile range, symbols show outliers outside of the whiskers. (**D**) Firing rate. (**E**) Coefficient of variation of the interspike interval of neurons with at least 10 spikes. (**F**) Correlation coefficient of a random subsample of 2000 neurons.

a given population to a given target area varies systematically between feedforward and feedback projections (Fig. 3**C**); here, feedforward and feedback were classified based on the predicted SLN value (***Schmidt et al., 2018a***): $\mathrm{SLN} > 65\%$ (feedforward), $35\% \leq \mathrm{SLN} \leq 65\%$ (lateral), and $\mathrm{SLN} < 35\%$ (feedback). The average outdegree in our model in the feedforward direction is 349 compared to 580 in the feedback direction.

Finally, fully reconstructed axons (***Winnubst et al., 2019***) suggest that many projecting neurons target multiple areas. To exclude weak connections, we restrict ourselves in the model to connections with an average outdegree larger than 100. Using again the predicted SLN value to separate feedforward, lateral, and feedback connections, we obtain a broad distribution of the number of target areas (Fig. 3**D**). In addition to the larger outdegree in the feedback direction, feedback projections also target more areas: on average 3.66 compared to 2.42 for lateral and 1.92 for feedback projections.

Micro- and Macroscopic Dynamics

Simulated Spiking Activity

The simulated spiking activity of the model is asynchronous and irregular with low firing rates across all areas (Fig. 4). There is a pronounced structure of the activity across populations, layers, and areas (Fig. 4**A**-**C**). Due to the distributed neuron parameters, the activity is furthermore heterogeneous within the populations with some neurons displaying persistent activity and others being inactive (Fig. 4**A**-**C**).

To quantify the spiking activity further, we consider population-averaged statistics (Fig. 4**D**-**F**). The firing rate of the inhibitory neurons is higher than the firing rate of the excitatory neurons with the highest activity in layer VI (Fig. 4**D**). The activity of some excitatory populations is very low, in particular in layers II/III and V (Fig. 4**B**-**D**). In terms of the irregularity of the spike trains, quantified

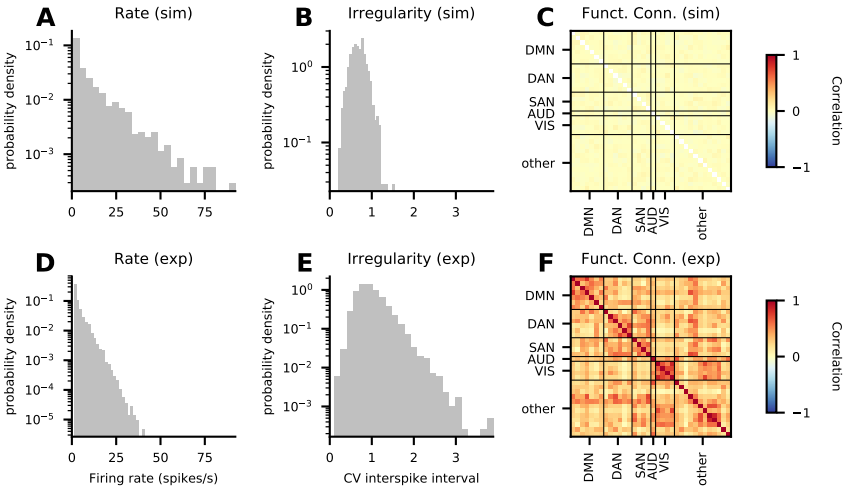

Figure 5. Comparison to experimental activity data. (**A**,**B**) Distribution of simulated firing rates (**A**) and coefficients of variation of the interspike interval (**B**) across neurons in area caudalanteriorcingulate. (**C**) Simulated functional connectivity based on synaptic input currents in the default mode network (DMN), dorsal attention network (DAN), salience network (SAN), auditory network (AUD), visual network (VIS), and the remaining areas (other). (**D**,**E**) Distribution of measured firing rates (**D**) and coefficients of variation of the interspike interval (**E**) across neurons in medial frontal cortex (***Minxha et al., 2020***). (**F**) Experimental functional connectivity of the left hemisphere from fMRI recordings averaged across nineteen subjects.

by the coefficient of variation CV of the interspike interval, all populations are in the regime of $\mathrm{CV} \approx 0.5$ (Fig. 4**E**), i.e., in an intermediate regime between a Poisson process and a periodic process. Lastly, the average pairwise correlation between the neurons is close to zero across all populations (Fig. 4**F**).

Comparison with Experimental Activity Data

We compare the activity of the model with experimental activity data on two levels: on the neuron level, we use the electrophysiological recordings by ***Minxha et al.*** (***2020***) from human medial frontal cortex (cf. Sec. Spiking data); on the cortex level, we use resting state fMRI data from nineteen subjects (cf. Sec. fMRI data).

The electrophysiological data was recorded in dorsal anterior cingulate cortex as well as pre-supplementary motor area; we compare the data with the model activity in area caudalanteriorcingulate (comparison with area superiorfrontal, which comprises pre-supplementary motor area but also more frontal regions, leads to qualitatively similar results). Since the recordings are not layer- or population-specific, we combine the spike trains of all layers and populations in caudalanteriorcingulate for this analysis. For the firing rate, we consider only neurons with at least 0.5 spikes/s; for the irregularity we consider only neurons with at least 10 spikes in the respective interval. Both in our model (Fig. 5**A**) and in the experimental data (Fig. 5**D**), the firing rates are broadly distributed. However, in the experimental data the distribution is less broad with a maximal rate of approximately 50 spikes/s while the model activity reaches to up to 75 spikes/s. Similarly, the CV shows clear differences between recordings and model activity: in the model, the CV is narrowly distributed around a sub-Poissonian average (Fig. 5**B**); in the recordings, the CV is broadly distributed around a Poissonian average (Fig. 5**E**). Thus, the recorded spike trains are more irregular than their simulated counterparts.

To obtain a proxy of the BOLD signal from our model, we use the absolute value of the area-

level synaptic currents (***Schmidt et al., 2018b***). We compute the functional connectivity using the Pearson correlation coefficient of this BOLD proxy (simulation) or the BOLD signal (experiment). To facilitate the comparison, we group the areas into different resting state networks following ***Kabbara et al.*** (***2017***). While the experimental functional connectivity shows a clear structure with increased correlations among the areas in the resting state networks (Fig. 5**F**), the simulated functional connectivity shows only very weak correlation and almost no structure (Fig. 5**C**).

Discussion

We aggregated data across multiple modalities to construct a multi-scale spiking network model of human cortex. This data encompasses, among others, electron microscopy, electrophysiology, morphological reconstructions, and DTI. We filled gaps in the data using statistical regularities found in other species, in particular to determine the laminar origin and target of cortico-cortical connections.

Simulations of the model reveal asynchronous and irregular activity. The activity is heterogeneous across areas, layers, and populations as well as within populations. We compared the model activity against electrophysiological recordings in human medial frontal cortex and human resting state fMRI. On both levels, we observed strong deviations. On the single-neuron level, the firing rate distribution in the model is wider than the observed one while the irregularity is too narrow and too low. On the network level, the activity is hardly correlated which is in stark contrast to the salient structure in the fMRI data.

Addressing these discrepancies is the necessary next step. In a previous multi-scale model of the vision-related areas in macaque, increasing the synaptic weights of cortico-cortical connections led to inter-area correlations in agreement with experimental data (***Schmidt et al., 2018b***). In the current version of the model, however, strengthening the cortico-cortical connections leads to a sudden transition to a high activity state (***Schuecker et al., 2017***). One way to address this problem are scaling factors of specific connections (cf. Sec. Further Parameters of the Model). Using mean-field theory (cf. Sec. Mean-Field Theory), the scaling factors can be adjusted such that the activity does not diverge into the high activity state (for a more principled approach see ***Schuecker et al. 2017***). But in addition to the high activity, the network synchronizes and displays almost perfectly coherent activity. Such oscillations are not captured in the current theory; thus, an approach based on linear response theory akin to the work by ***Bos et al.*** (***2016***) might be necessary. Another way to address this problem is a brute-force parameter search. While this is computationally very expensive, it is not entirely prohibitive because the time required to simulate a second of activity outside of the high activity state is on the order of a minute.

Materials & Methods

Mesoconnectome Construction

Neuron number

The number of neurons per layer follows from multiplying their volume density ρ_{neuron} with the layer thickness h_{layer} and the surface area A_{column} as $N_{\text{neuron}} = \rho_{\text{neuron}} h_{\text{layer}} A_{\text{column}}$. We use the volume density and the layer thickness provided in the seminal work of von Economo and Koskinas (***Von Economo, 2009***). This data distinguishes the layers into finer categories than the ones we use in our model. Therefore, we sum the corresponding "layer thickness overall" and average the corresponding "cell content" values weighted by the relative layer thickness.

Furthermore, the data is provided in the parcellation of von Economo and Koskinas; we use the mapping to the Desikan-Killiany parcellation constructed by ***Goulas et al.*** (***2016***, Table 1). In the given mapping, one or more von Economo and Koskinas areas are assigned to each Desikan-Killiany area. For the layer thicknesses, we take the average across the corresponding areas in the parcellation by von Economo and Koskinas (using that the mapping was constructed based on

cytoarchitectonic similarity, such that the average is across architectonically similar areas); for the volume densities, we weight the average by the relative thickness of the layers.

To separate the neurons in a given layer into inhibitory and excitatory neurons, we use the layer-resolved relative size of the two populations from ***Shapson-Coe et al.*** (***2021***, Supplementary Figure 5); quantitative values were extracted using WebPlotDigitizer (***Rohatgi, 2021***) and the values for layer II and III were averaged. The population sizes follow by multiplying the relative population size with the total number of neurons in the layer determined above.

Fraction of Cortico-Cortical Connections

We separate the $N_{\text{synapse}}^{\text{long-range}}$ cortico-cortical synapses from the $N_{\text{synapse}}^{\text{local}} = N_{\text{synapse}}^{\text{total}} - N_{\text{synapse}}^{\text{long-range}}$ synapses coming from within the area or from subcortical regions (see Data Aggregation & Predictive Connectomics). To determine the fraction of cortico-cortical synapses, we use the scaling rule by ***Herculano-Houzel et al.*** (***2010***)

$$\frac{N_{\text{neuron}}^{\text{long-range}}}{N_{\text{neuron}}^{\text{total}}} \propto \frac{1}{(N_{\text{neuron}}^{\text{total}})^{0.16}}; \tag{1}$$

i.e., the relative number of neurons connected through the white matter decreases with increasing total number of neurons in the gray matter $N_{\text{neuron}}^{\text{total}}$. We determine the proportionality constant using the value $N_{\text{neuron}}^{\text{long-range}}/N_{\text{neuron}}^{\text{total}} = 0.21$ from tracing data in macaque (***Markov et al., 2011***) in combination with $N_{\text{neuron}}^{\text{gray matter}} = 1.4 \times 10^9$ gray matter neurons in macaque (***Collins et al., 2010***). With the number of gray matter neurons in human, $N_{\text{neuron}}^{\text{gray matter}} = 16 \times 10^9$ (***Herculano-Houzel, 2009***), we arrive at the estimate $N_{\text{neuron}}^{\text{long-range}}/N_{\text{neuron}}^{\text{total}} = 0.14$ or $N_{\text{neuron}}^{\text{local}}/N_{\text{neuron}}^{\text{total}} = 0.86$. Finally, we assume that the fraction of neurons connected through the white matter equals the fraction of cortico-cortical synapses.

Local Connectivity

The $N_{\text{synapse}}^{\text{local}}$ local synapses need a further distinction: $N_{\text{synapse}}^{\text{internal}}$ synapses where the presynaptic neuron is part of the simulated column and $N_{\text{synapse}}^{\text{external}}$ synapses where the presynaptic neuron is outside of the simulated column, i.e., in the remainder of the area or in a subcortical region. To split these two categories, we use the spatial connection probability $p(\boldsymbol{x}_1 \mid \boldsymbol{x}_2)$ between a neuron located at $\boldsymbol{x}_1$ and another neuron at $\boldsymbol{x}_2$, which we assume to be a spatially homogeneous three-dimensional exponential distribution $p(\boldsymbol{x}_1 \mid \boldsymbol{x}_2) \propto \exp(-|\boldsymbol{x}_1 - \boldsymbol{x}_2|/\lambda_{\text{conn}})$ with decay constant $\lambda_{\text{conn}} = 160\,\mu\text{m}$ (***Packer and Yuste, 2011***; ***Perin et al., 2011***). From $p(\boldsymbol{x}_1, \boldsymbol{x}_2) = p(\boldsymbol{x}_1 \mid \boldsymbol{x}_2)p(\boldsymbol{x}_2)$ where $p(\boldsymbol{x}_2)$ is assumed to be constant, we obtain the average connection probability P_{internal} within the column as

$$P_{\text{internal}} \propto \int_{\text{column}} d\boldsymbol{x}_1 \int_{\text{column}} d\boldsymbol{x}_2 \, \exp(-|\boldsymbol{x}_1 - \boldsymbol{x}_2|/\lambda_{\text{conn}}) \tag{2}$$

where the proportionality factor is the normalization constant of $p(\boldsymbol{x}_1, \boldsymbol{x}_2)$. We calculate the average connection probability assuming cylindrical columns. In cylindrical coordinates, using $d\boldsymbol{x} = r dr d\phi dz$ and $\int_0^a dx_1 \int_0^a dx_2 \, f(|x_2 - x_1|) = 2\int_0^a dy\,(a-y) f(|y|)$ simplifies this integral to

$$P_{\text{internal}} \propto 4 \int_0^{r_{\text{column}}} dr_1 \int_0^{r_{\text{column}}} dr_2 \int_0^{2\pi} d\phi \int_0^h dz \, r_1 r_2 (2\pi - \phi)(h - z) \exp(-d(r_1, r_2, \phi, z)/\lambda_{\text{conn}}) \tag{3}$$

with $d(r_1, r_2, \phi, z) = \sqrt{r_1^2 - 2r_1 r_2 \cos\phi + r_2^2 + z^2}$, the radius of the column r_{column}, and the total height of the column h. For the probability P_{external} that the postsynaptic neuron is in the column but the presynaptic neuron outside of it, the domain outside of the column has to be integrated: $\int_{\text{column}} d\boldsymbol{x}_1 \to \int_{\neg\text{column}} d\boldsymbol{x}_1$. Approximating the entire area as a cylinder, this leads to the replacement $\int_0^{r_{\text{column}}} dr_1 \to \int_{r_{\text{column}}}^{r_{\text{area}}} dr_1$ where r_{area} is the radius of the larger cylinder, i.e.,

$$P_{\text{external}} \propto 4 \int_{r_{\text{column}}}^{r_{\text{area}}} dr_1 \int_0^{r_{\text{column}}} dr_2 \int_0^{2\pi} d\phi \int_0^h dz \, r_1 r_2 (2\pi - \phi)(h - z) \exp(-d(r_1, r_2, \phi, z)/\lambda_{\text{conn}}). \tag{4}$$

The remaining integrals are solved numerically using the adaptive multidimensional quadrature implemented in SciPy (***Virtanen et al., 2020***). P_{internal} and P_{external} are used to determine the number

of synapses with neurons within and outside of the column, respectively:

$$N_{\text{synapse}}^{\text{internal}} = \frac{P_{\text{internal}}}{P_{\text{internal}} + P_{\text{external}}} N_{\text{synapse}}^{\text{local}}, \tag{5}$$

$$N_{\text{synapse}}^{\text{external}} = \frac{P_{\text{external}}}{P_{\text{internal}} + P_{\text{external}}} N_{\text{synapse}}^{\text{local}}. \tag{6}$$

Note that although we keep r_{column} the same for all areas, both P_{internal} and P_{external} are area-specific because their thickness h and the neuron densities vary.

For the local connectivity within the column, comprising $N_{\text{synapse}}^{\text{internal}}$ synapses, we use the model of ***Potjans and Diesmann* (*2014*)** as a blueprint. More precisely, we use the average number of synapses $p_{\text{B}\to\text{A}}^{\text{PD}}$ between a neuron in source population B and a neuron in target population A. We combine these average numbers of synapses with the number of neurons $N_{\text{neuron}}^{\text{B}}$, $N_{\text{neuron}}^{\text{A}}$ in the pre- and postsynaptic population:

$$N_{\text{synapse}}^{\text{B}\to\text{A}} = \frac{N_{\text{neuron}}^{\text{B}} p_{\text{B}\to\text{A}}^{\text{PD}} N_{\text{neuron}}^{\text{A}}}{\sum_{\text{A,B}} N_{\text{neuron}}^{\text{B}} p_{\text{B}\to\text{A}}^{\text{PD}} N_{\text{neuron}}^{\text{A}}} N_{\text{synapse}}^{\text{internal}}. \tag{7}$$

Eq. (7) keeps the average number of synapses per pair of neurons equal to the respective value in ***Potjans and Diesmann* (*2014*)** by construction.

The $N_{\text{synapse}}^{\text{external}}$ synapses from outside the column are also distributed based on ***Potjans and Diesmann* (*2014*)**. Here, we use the indegrees $K_{\text{ext}\to\text{A}}^{\text{PD}}$ and the number of neurons in the postsynaptic population $N_{\text{neuron}}^{\text{A}}$:

$$N_{\text{synapse}}^{\text{ext}\to\text{A}} = \frac{K_{\text{ext}\to\text{A}}^{\text{PD}} N_{\text{neuron}}^{\text{A}}}{\sum_{\text{A}} K_{\text{ext}\to\text{A}}^{\text{PD}} N_{\text{neuron}}^{\text{A}}} N_{\text{synapse}}^{\text{external}}. \tag{8}$$

Both in Eq. (7) and Eq. (8), we round the final result to obtain an integer number of synapses.

Long-range projections

For the $N_{\text{synapse}}^{\text{long-range}}$ synapses from other cortical areas, we assume that the presynaptic neurons are inside the simulated column in the respective presynaptic area. Thus, we do not distinguish between synapses with simulated and non-simulated presynaptic neurons—all presynaptic neurons of long-range projections are simulated.

We define the area-level connectivity according to processed DTI data from ***Goulas et al.* (*2016*)** which is based on data from the Human Connectome Project (***Van Essen et al., 2013***). For a given target area X, we distribute the synapses among the source areas based on the relative number of streamlines $\text{NoS}_{\text{Y}\to\text{X}}$ in the DTI data,

$$N_{\text{synapse}}^{\text{Y}\to\text{X}} = \frac{\text{NoS}_{\text{Y}\to\text{X}}}{\sum_{Z} \text{NoS}_{\text{Z}\to\text{X}}} N_{\text{synapse}}^{\text{long-range}}. \tag{9}$$

Again, we round the resulting value.

A comprehensive dataset on the layer specificity of the presynaptic neurons based on retrograde tracing is available for macaque (***Markov et al., 2014b***,a). Not only in this species but also in cat, the layer specificity, i.e., the fraction of supragranular labeled neurons SLN, is systematically related to the cytoarchitecture (***van Albada et al., 2022***). For our human model, we assume the same quantitative relationship as in macaque. Fitting a beta-binomial model with a probit link function to the macaque data yields (***Schmidt et al., 2018a***)

$$\text{SLN}(\text{B}\to\text{A}) = \Phi\left(a_0 + a_1 \log(\rho_{\text{neuron}}^{\text{A}}/\rho_{\text{neuron}}^{\text{B}})\right) \tag{10}$$

where $\Phi(x) = \frac{1}{2}\left[1 + \text{erf}\left(x/\sqrt{2}\right)\right]$ denotes the cumulative distribution function of the standard normal distribution and the fitted parameters are $a_0 = -0.152$ and $a_1 = -1.534$. We use the human neuron densities in Eq. (10) to estimate the laminar origin in human. The SLN value allows determining whether the origin is in layer $2/3$ or not. Excluding layer 4, which does not form long-range projections (***Markov et al., 2014b***), the two infragranular layers 5 and 6 still need to be distinguished.

To this end, we simply use the relative size of the two populations to distribute the remaining synapses.

On the target side, anterograde tracing can specify the layer specificity. However, there are no comprehensive datasets of anterograde tracing in non-human primates available to date. Hence, we use the collected data from the CoCoMac database (***Stephan et al., 2001***) which aggregates data across many tracing studies. Relating the target patterns from anterograde tracing to the SLN value reveals three categories of target patterns (***Schmidt et al., 2018a***):

$$\begin{aligned} \mathrm{SLN} > 65\% &: [4] \\ 35\% \leq \mathrm{SLN} \leq 65\% &: [1, 2/3, 4, 5, 6] \\ \mathrm{SLN} < 35\% &: [1, 2/3, 5, 6] \end{aligned}$$

where layer 4 is replaced by $2/3$ in the first case for agranular target areas (***Beul and Hilgetag, 2015***). Using the SLN value to distinguish feedforward ($\mathrm{SLN} > 65\%$), lateral ($35\% \leq \mathrm{SLN} \leq 65\%$), and feedback ($\mathrm{SLN} < 35\%$) connections, this implies that feedforward connections target layer 4, feedback connections avoid layer 4, and lateral connections show no distinct pattern. For the quantitative distribution of the synapses onto the layers included in the respective target pattern, we use the relative thickness of the layer in relation to all layers of the target pattern.

Thus far, we determined the location of the synapse in the target layer. Next, we decide whether the postsynaptic neuron of a synapse in a given layer is excitatory or inhibitory based on the analysis of the data by ***Binzegger et al.*** (***2004***) in ***Schmidt et al.*** (***2018a***, Table S11). To this end, we sum the target probabilities for postsynaptic neurons across all layers separately for excitatory and inhibitory neurons. This yields the probability for a synapse in a given layer to have an excitatory or inhibitory postsynaptic neuron in any layer. However, we take one exception into account: For feedback connections ($\mathrm{SLN} < 35\%$), we fix the fraction of excitatory target cells to 93% (***Schmidt et al., 2018a***) because feedback connections preferentially target excitatory neurons (***Johnson and Burkhalter, 1996***; ***Anderson et al., 2011***).

To finally determine the postsynaptic neuron, we assume that all inhibitory postsynaptic neurons are in the same layer as the synapse. For the excitatory neurons, we take the dendritic morphology into account. Using morphological reconstructions of human pyramidal cells in temporal cortex (***Mohan et al., 2015***), we calculate the layer-resolved length of dendrites for neurons with the soma in a given layer. Assuming a constant density of synapses along the dendrites, the ratio of the length $\ell_{\mathrm{A,B}}$ of dendrites in layer $\mathrm{A} \in [1, 2/3, 4, 5, 6]$ belonging to a neuron with soma in layer $\mathrm{B} \in [2/3, 4, 5, 6]$ to the total length of dendrites in this layer, $\sum_{\mathrm{B}} \ell_{\mathrm{A,B}}$, determines the probability that the postsynaptic cell is in layer B given that the synapse is in layer A: $P(\text{soma in B} \mid \text{synapse in A}) = \ell_{\mathrm{A,B}} / \sum_{\mathrm{B}} \ell_{\mathrm{A,B}}$.

Ultimately, we only need the location of the postsynaptic neuron but not the location of the synapse. Thus, we multiply $P(\text{soma in B} \mid \text{synapse in A})$ with the distribution of the synapses across the layers and marginalize the synapse location.

Further Parameters of the Model

Neuron parameters

We use the leaky integrate-and-fire (LIF) neuron model with exponential postsynaptic currents (***Gerstner et al., 2014***) for all neurons. To determine the parameter values, we analyzed the LIF models from the Allen Cell Types Database (https://celltypes.brain-map.org/; ***Teeter et al., 2018***; ***Berg et al., 2021***) which were fitted to human neurons. For both excitatory and inhibitory cells, we fix the leak and reset potential to $\mathrm{V_L} = \mathrm{V_{reset}} = -70\,\mathrm{mV}$. For the threshold potential $\mathrm{V_{th}}$, the membrane time constant τ_{m}, and the membrane capacitance $\mathrm{C_m}$, we fitted a log-normal distribution using maximum likelihood estimation to the distribution of the respective parameter for all cells in which the LIF model had an explained variance above 0.75 to ensure a good fit of the LIF model. For convenience, we parameterize the log-normal distribution using the mean and the coefficient of

variation CV. The resulting mean threshold potential is $V_{th} = -45\,\mathrm{mV}$ for both excitatory and inhibitory cells with $CV = 0.21$ and $CV = 0.22$ for excitatory and inhibitory cells, respectively. The resulting mean capacitance is $C_m = 220\,\mathrm{pF}$ and $C_m = 100\,\mathrm{pF}$ with $CV = 0.22$ and $CV = 0.34$ for excitatory and inhibitory cells, respectively. To account for the high conductance state in vivo (***Destexhe et al., 2003***), we lower the membrane time constant to $\tau_m = 10\,\mathrm{ms}$ on average with $CV = 0.55$ and $CV = 0.43$ for excitatory and inhibitory cells, respectively. We do not distribute the synaptic time constants, which we fix to $\tau_s = 2\,\mathrm{ms}$ for excitatory and $\tau_s = 5\,\mathrm{ms}$ for inhibitory input (***Fourcaud and Brunel, 2002***), and the absolute refractory period of $t_{ref} = 2\,\mathrm{ms}$.

Synapse parameters

We use static synapses with a transmission probability of $100\,\%$. Excitatory postsynaptic potentials follow a truncated normal distribution with average $0.1\,\mathrm{mV}$ and relative standard deviation of $10\,\%$. The inhibitory postsynaptic potentials also follow a truncated normal distribution with a factor $g = 4$ larger absolute value of the mean and standard deviation. Excitatory (inhibitory) weights are truncated below (above) zero; values outside of the allowed range are redrawn.

We introduce several scaling factors that affect the postsynaptic potentials: First, the synaptic weights of the synapses within a column from layer IV excitatory neurons to layer II/III excitatory neurons are increased twofold in agreement with the blueprint (***Potjans and Diesmann, 2014***). Second, we introduce two scaling factors for the synapses within a column: from layer 5 excitatory neurons to all inhibitory neurons and from all excitatory neurons to all inhibitory neurons. Both scaling factors stabilize the column with respect to cortico-cortical input. For all simulations shown, the first scaling factor is set to 1.8 and the second to 1.2; the scaling factors are multiplied if both apply.

Delays

Within a column, the average delay is $1.5\,\mathrm{ms}$ for excitatory and $0.75\,\mathrm{ms}$ for inhibitory synapses. For the cortico-cortical synapses, we assume a conduction velocity of $3.5\,\mathrm{m/s}$ (***Girard et al., 2001***). Dividing the fiber length between two areas, obtained through tractography (***Goulas et al., 2016***), by the conduction velocity, we get the average delay between the two areas. All delays follow a truncated log-normal distribution with a relative standard deviation of $50\,\%$. Delays are truncated below the resolution of the simulation; values outside of the allowed range are redrawn.

External input

We determined the number of synapses from non-simulated presynaptic neurons in Eq. (8). The postsynaptic potentials follow a truncated normal distribution with average $w_{ext} = 0.1\,\mathrm{mV}$ and relative standard deviation of $10\,\%$. We keep the mean input, measured relative to rheobase, fixed at $\eta_{ext} = 1.1$ and determine the rate of the driving Poisson processes by

$$\nu_{ext}^{A} = \frac{V_{th} - V_L}{\tau_m w_{ext} K_A^{ext}} \eta_{ext} \tag{11}$$

with $K_A^{ext} = N_{synapse}^{ext \to A} / N_{neuron}^{A}$. We further introduce two scaling factors for the postsynaptic potentials arriving at excitatory neurons in layer 5 and 6, respectively. For all simulations shown, the first scaling factor is set to 1.05 and the second to 1.15.

Activity Data

Spiking data

Minxha et al. (***2020***) recorded from thirteen adult epilepsy patients under evaluation for surgical treatment using depth electrodes in medial frontal cortex. In total, they recorded 767 neurons within 320 trials and extracted spikes using a semi-automated spike sorting algorithm. For our analysis, we disregard task related activity and use only the two seconds of activity which were recorded before stimulus onset. The data is publicly available via OSF at http://doi.org/10.17605/OSF.IO/U3KCP.

fMRI data

Participants

Nineteen participants (7 female, age range = 21–33 years, mean age = 25 years) with normal or corrected-to-normal visual acuity took part in this study. All participants provided written informed consent after receiving full information about experimental procedures and were compensated for participation either through monetary reward or course credit. All procedures were conducted with approval from the local Ethical Committee of the Faculty of Psychology and Neuroscience at Maastricht University.

Magnetic resonance imaging

Anatomical and functional images were acquired at Maastricht Brain Imaging Centre (Maastricht University) on a whole-body Magnetom 7T research scanner (Siemens Healthineers, Erlangen, Germany) using a 32-channel head-coil (Nova Medical Inc.; Wilmington, MA, USA). Anatomical data were collected prior to functional data with an MP2RAGE (***Marques et al., 2010***) imaging sequence [240 slices, matrix = 320×320, voxel size = $0.65 \times 0.65 \times 0.65$ mm^3, first inversion time (TI1) = 900 ms, second inversion time (TI2) = 2750 ms, echo time (TE) = 2.51 ms, repetition time (TR) = 5000 ms, first nominal flip angle = 5°, and second nominal flip angle = 3°, GRAPPA = 2]. Functional images were acquired using a gradient-echo echo-planar (***Moeller et al., 2010***) imaging sequence (84 slices, matrix = 186×186, voxel size = $1,6 \times 1.6 \times 1.6$ mm^3, TE = 22 ms, TR = 1500 ms, nominal flip angle = 63°, GRAPPA = 2, multi-band factor = 4). In addition, after the first functional run, they recorded five functional volumes with opposed phase encoding directions to correct for EPI distortions that occur at higher field strengths (***Andersson et al., 2003***).

Participants underwent five functional runs comprising of a resting-state measurement, three individual task measurements and a task-switching paradigm wherein participants repeatedly performed each of the three tasks. With the exception of the task-switching run, which lasted 9.5 min, all functional runs lasted 15 min. Since task-related runs were not included in the present study, they will not be discussed further. However, it is noteworthy that resting-state runs always preceded task-related runs to prevent carry-over effects (***Grigg and Grady, 2010***). Participants were instructed to close their eyes during resting-state runs and otherwise to let their mind wander freely.

Processing of (f)MRI data

Anatomical images were downsampled to $0.8 \times 0.8 \times 0.8$ mm^3 and subsequently automatically processed with the longitudinal stream in FreeSurfer (http://surfer.nmr.mgh.harvard.edu/) including probabilistic atlas-based cortical parcellation according to the Desikan-Killiany (DK) atlas (***Desikan et al., 2006***). Initial preprocessing of functional data was performed in BrainVoyager 20 (version 20.0; Brain Innovation; Maastricht, The Netherlands) and included slice scan time correction and (rigid body) motion correction wherein all functional runs were aligned to the first volume of the first functional run. EPI distortions were then corrected using the COPE ("Correction based on Opposite Phase Encoding") plugin of BrainVoyager that implements a method similar to that described in ***Andersson et al.*** (***2003***) and the 'topup' tool implemented in FSL (***Smith et al., 2004***). The pairs of reversed phase encoding images recorded in the beginning of the scanning session were used to estimate the susceptibility-induced off-resonance field and correct the distortions in the remaining functional runs. This was followed by wavelet despiking (***Patel and Bullmore, 2016***) using the BrainWavelet Toolbox (brainwavelet.org) for MATLAB (2019a, The MathWorks,Natick, MA). Subsequently, high-pass filtering was performed in BrainVoyager with a frequency cutoff of 0.01 Hz and to register functional images to participants' anatomical images. Using MATLAB, functional data were then cleaned further by regressing out a global noise signal given by the first five principal components of signals observed within the cerebrospinal fluid of the ventricles (***Behzadi et al., 2007***). Finally, voxels were uniquely assigned to one of 68 cortical regions of interest (ROIs) and an average blood-oxygen-level-dependent (BOLD) signal for each ROI was obtained as the mean of

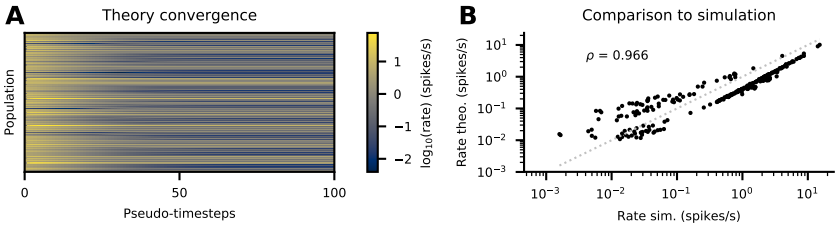

Figure 6. Mean-field theory. (**A**) Firing rates of all populations across the pseudo-timesteps used to find a self-consistent solution. (**B**) Comparison with rates predicted by mean-field theory to empirical rates from a simulation.

the time-series of its constituent voxels.

Mean-Field Theory

While developing the model, it was often beneficial to have a prediction of the activity without performing a computationally demanding full-scale simulation. To this end, we employed the mean-field theory developed in ***Amit and Brunel* (*1997*)** in combination with the extension to exponential post-synaptic currents and multiple synaptic time constants derived in ***Fourcaud and Brunel* (*2002*)**. Within this theory, the input to a neuron in population A is approximated as a Gaussian white noise with mean μ_A and noise intensity σ_A^2. The main assumptions of this theory are that the inputs are uncorrelated, that the temporal structure of the input can be neglected, that a Gaussian approximation of its statistics is valid, that the delays can be neglected, and that the variability of the neuron parameters and synaptic weights can be neglected.

Despite the simplifying assumptions, the theory provides a reliable prediction of the average firing rates if the network is in an asynchronous irregular state (Fig. 6). The remaining deviations are likely a consequence of the neglected variabilities, in particular the distributed neuron parameters. Thus, the theory allows for rapid prototyping without the need for high performance computing resources.

Code & Workflow

The entire workflow of the model, from data preprocessing through the simulation to the final analysis, relies on the Python programming language (***Python Software Foundation, 2008***) version 3.6.5 in combination with NumPy (***Harris et al., 2020***) version 1.14.3, SciPy (***Virtanen et al., 2020***) version 1.1.0, pandas (***Wes McKinney, 2010***) version 0.23.4, Matplotlib (***Hunter, 2007***) version 2.2.2, and seaborn (***Waskom, 2021***) version 0.9.0. All simulations were performed using the NEST simulator (***Gewaltig and Diesmann, 2007***) version 2.20.2 (***Fardet et al., 2021***) on the JURECA-DC supercomputer. The workflow is structured using snakemake (***Köster and Rahmann, 2012***). For the mean-field based analysis, we used the NNMT toolbox (***Layer et al., 2022***).

Acknowledgments

We thank Sebastian Bludau and Timo Dickscheid for helpful discussions about cytoarchitecture and parcellations. Furthermore, we gratefully acknowledge all the shared experimental data, and the effort spent to collect it, which underlies our work.

References

van Albada SJ, Morales-Gregorio A, Dickscheid T, Goulas A, Bakker R, Bludau S, Palm G, Hilgetag CC, Diesmann M. In: Giugliano M, Negrello M, Linaro D, editors. Bringing Anatomical Information into Neuronal Network Models Cham: Springer International Publishing; 2022. p. 201–234. doi: 10.1007/978-3-030-89439-9_9.

Amit DJ, Brunel N. Model of Global Spontaneous Activity and Local Structured Activity During Delay periods in the Cerebral Cortex. Cereb Cortex. 1997 Apr; 7:237–252. https://doi.org/10.1093/cercor/7.3.237, doi: 10.1093/cercor/7.3.237.

Anderson JC, Kennedy H, Martin KAC. Pathways of Attention: Synaptic Relationships of Frontal Eye Field to V4, Lateral Intraparietal Cortex, and Area 46 in Macaque Monkey. J Neurosci. 2011; 31(30):10872–10881. http://www.jneurosci.org/content/31/30/10872.abstract, doi: 10.1523/JNEUROSCI.0622-11.2011.

Andersson JLR, Skare S, Ashburner J. How to correct susceptibility distortions in spin-echo echo-planar images: application to diffusion tensor imaging. Neuroimage. 2003; 20(2):870–888. https://www.sciencedirect.com/science/article/pii/S1053811903003367, doi: 10.1016/S1053-8119(03)00336-7.

Arkhipov A, Gouwens NW, Billeh YN, Gratiy S, Iyer R, Wei Z, Xu Z, Abbasi-Asl R, Berg J, Buice M, et al. Visual physiology of the layer 4 cortical circuit in silico. PLOS Comput Biol. 2018; 14(11):e1006535.

Barbas H, Rempel-Clower N. Cortical structure predicts the pattern of corticocortical connections. Cereb Cortex. 1997; 7(7):635–646. http://cercor.oxfordjournals.org/content/7/7/635.abstract, doi: 10.1093/cercor/7.7.635.

Behzadi Y, Restom K, Liau J, Liu TT. A component based noise correction method (CompCor) for BOLD and perfusion based fMRI. Neuroimage. 2007; 37(1):90–101. https://www.sciencedirect.com/science/article/pii/S1053811907003837, doi: j.neuroimage.2007.04.042.

Berg J, Sorensen SA, Ting JT, Miller JA, Chartrand T, Buchin A, Bakken TE, Budzillo A, Dee N, Ding SL, Gouwens NW, Hodge RD, Kalmbach B, Lee C, Lee BR, Alfiler L, Baker K, Barkan E, Beller A, Berry K, et al. Human neocortical expansion involves glutamatergic neuron diversification. Nature. 2021 Oct; 598(7879):151–158. doi: 10.1038/s41586-021-03813-8.

Beul SF, Barbas H, Hilgetag CC. A predictive structural model of the primate connectome. Sci Rep. 2017; 7(43176):1–12.

Beul SF, Hilgetag CC. Towards a 'canonical' agranular cortical microcircuit. Front Neuroanat. 2015; 8:165. http://www.frontiersin.org/neuroanatomy/10.3389/fnana.2014.00165/abstract, doi: 10.3389/fnana.2014.00165.

Bezaire MJ, Raikov I, Burk K, Vyas D, Soltesz I. Interneuronal mechanisms of hippocampal theta oscillations in a full-scale model of the rodent CA1 circuit. eLife. 2016 Dec; 5. https://doi.org/10.7554/elife.18566, doi: 10.7554/elife.18566.

Billeh YN, Cai B, Gratiy SL, Dai K, Iyer R, Gouwens NW, Abbasi-Asl R, Jia X, Siegle JH, Olsen SR, et al. Systematic integration of structural and functional data into multi-scale models of mouse primary visual cortex. Neuron. 2020; 106(3):388–403.e18. https://www.sciencedirect.com/science/article/pii/S0896627320300672, doi: https://doi.org/10.1016/j.neuron.2020.01.040.

Binzegger T, Douglas RJ, Martin KAC. A Quantitative Map of the Circuit of Cat Primary Visual Cortex. J Neurosci. 2004; 39(24):8441–8453. https://doi.org/10.1523/jneurosci.1400-04.2004, doi: 10.1523/JNEUROSCI.1400-04.2004.

Bos H, Diesmann M, Helias M. Identifying Anatomical Origins of Coexisting Oscillations in the Cortical Microcircuit. PLOS Comput Biol. 2016 Oct; 12(10):e1005132. http://doi.org/10.1371%2Fjournal.pcbi.1005132, doi: 10.1371/journal.pcbi.1005132.

Cano-Astorga N, DeFelipe J, Alonso-Nanclares L. Three-Dimensional Synaptic Organization of Layer III of the Human Temporal Neocortex. Cereb Cortex. 2021; 31(10):4742–4764. doi: 10.1093/cercor/bhab120.

Collins CE, Airey DC, Young NA, Leitch DB, Kaas JH. Neuron densities vary across and within cortical areas in primates. Proc Natl Acad Sci USA. 2010 September; 107(36):15927–15932.

DeFelipe J, Alonso-Nanclares L, Arellano JI. Microstructure of the neocortex: comparative aspects. J Neurocytol. 2002; 31:299–316. https://doi.org/10.1023/A:1024130211265, doi: 10.1023/A:1024130211265.

DeFelipe J, Elston G, Fujita I, Fuster J, Harrison K, Hof P, Kawaguchi Y, Martin K, Rockland K, Thomson A, Wang S, White E, Yuste R. Neocortical circuits: Evolutionary aspects and specificity versus non-specificity of synaptic connections. Remarks, main conclusions and general comments and discussion. J Neurocytol. 2002; 32:387–416. https://doi.org/10.1023/A:1024142513991, doi: 10.1023/A:1024142513991.

Desikan RS, Ségonne F, Fischl B, Quinn BT, Dickerson BC, Blacker D, Buckner RL, Dale AM, Maguire RP, Hyman BT, Albert MS, Killiany RJ. An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. Neuroimage. 2006; 31(3):968–980. https://www.sciencedirect.com/science/article/pii/S1053811906000437, doi: 10.1016/j.neuroimage.2006.01.021.

Destexhe A, Rudolph M, Pare D. The high-conductance state of neocortical neurons in vivo. Nat Rev Neurosci. 2003; 4:739–751.

Ecker A, Romani A, Sáray S, Káli S, Migliore M, Falck J, Lange S, Mercer A, Thomson AM, Muller E, Reimann MW, Ramaswamy S. Data-driven integration of hippocampal CA1 synaptic physiology in silico. Hippocampus. 2020 Jun; 30(11):1129–1145. https://doi.org/10.1002/hipo.23220, doi: 10.1002/hipo.23220.

Einevoll GT, Destexhe A, Diesmann M, Grün S, Jirsa V, de Kamps M, Migliore M, Ness TV, Plesser HE, Schürmann F. The Scientific Case for Brain Simulations. Neuron. 2019; 102(4):735–744. https://www.sciencedirect.com/science/article/pii/S0896627319302909, doi: 10.1016/j.neuron.2019.03.027.

Ercsey-Ravasz M, Markov NT, Lamy C, Essen DCV, Knoblauch K, Toroczkai Z, Kennedy H. A Predictive Network Model of Cerebral Cortical Connectivity Based on a Distance Rule. Neuron. 2013; 80(1):184–197. doi: 10.1016/j.neuron.2013.07.036.

Fardet T, Vennemo SB, Mitchell J, Mørk H, Graber S, Hahne J, Spreizer S, Deepu R, Trensch G, Weidel P, Jordan J, Eppler JM, Terhorst D, Morrison A, Linssen C, Antonietti A, Dai K, Serenko A, Cai B, Kubaj P, et al., NEST 2.20.2. Zenodo; 2021. https://doi.org/10.5281/zenodo.5242954, doi: 10.5281/zenodo.5242954.

Fourcaud N, Brunel N. Dynamics of the firing probability of noisy integrate-and-fire neurons. Neural Comput. 2002; 14:2057–2110. https://doi.org/10.1162/089976602320264015, doi: 10.1162/089976602320264015.

Gămănuţ R, Kennedy H, Toroczkai Z, Ercsey-Ravasz M, Van Essen DC, Knoblauch K, Burkhalter A. The mouse cortical connectome, characterized by an ultra-dense cortical graph, maintains specificity by distinct connectivity profiles. Neuron. 2018; 97(3):698–715.

Gerstner W, Kistler WM, Naud R, Paninski L. Neuronal Dynamics. From Single Neurons to Networks and Models of Cognition. Cambridge: Cambridge University Press; 2014.

Gewaltig MO, Diesmann M. NEST (NEural Simulation Tool). Scholarpedia J. 2007; 2(4):1430. https://doi.org/10.4249/scholarpedia.1430, doi: 10.4249/scholarpedia.1430.

Girard P, Hupé JM, Bullier J. Feedforward and Feedback Connections Between Areas V1 and V2 of the Monkey Have Similar Rapid Conduction Velocities. J Neurophysiol. 2001; 85(3):1328–1331.

Girardi-Schappo M, Bortolotto GS, Gonsalves JJ, Pinto LT, Tragtenberg MHR. Griffiths phase and long-range correlations in a biologically motivated visual cortex model. Sci Rep. 2016 Jul; 6(1). https://doi.org/10.1038/srep29561, doi: 10.1038/srep29561.

Goulas A, Werner R, Beul SF, Säring D, Heuvel Mvd, Triarhou LC, Hilgetag CC. Cytoarchitectonic similarity is a wiring principle of the human connectome. BioRxiv. 2016; doi: 10.1101/068254.

Grigg O, Grady CL. Task-Related Effects on the Temporal and Spatial Dynamics of Resting-State Functional Connectivity in the Default Network. PLOSONE. 2010 10; 5(10):1–12. https://doi.org/10.1371/journal.pone.0013311, doi: 10.1371/journal.pone.0013311.

Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, Wieser E, Taylor J, Berg S, Smith NJ, Kern R, Picus M, Hoyer S, van Kerkwijk MH, Brett M, Haldane A, Fernández del Río J, Wiebe M, Peterson P, Gérard-Marchant P, et al. Array programming with NumPy. Nature. 2020; 585:357–362. doi: 10.1038/s41586-020-2649-2.

Hass J, Hertäg L, Durstewitz D. A detailed data-driven network model of prefrontal cortex reproduces key features of in vivo activity. PLOS Comput Biol. 2016; 12(5):e1004930. doi: 10.1371/journal.pcbi.1004930.

Hendrickson PJ, Yu GJ, Robinson BS, Song D, Berger TW. Towards a large-scale biologically realistic model of the hippocampus. In: *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society* IEEE; 2012. https://doi.org/10.1109/embc.2012.6346990, doi: 10.1109/embc.2012.6346990.

Herculano-Houzel S. The human brain in numbers: a linearly scaled-up primate brain. Front Hum Neurosci. 2009; 3:31. https://doi.org/10.3389/neuro.09.031.2009, doi: 10.3389/neuro.09.031.2009.

Herculano-Houzel S, Mota B, Wong P, Kaas JH. Connectivity-driven white matter scaling and folding in primate cerebral cortex. Proc Natl Acad Sci USA. 2010; 107(44):19008–19013.

Hilgetag CC, Beul SF, van Albada SJ, Goulas A. An Architectonic Type Principle Integrates Macroscopic Cortico-Cortical Connections with Intrinsic Cortical Circuits of the Primate Brain. Netw Neurosci. 2019; 3(4):905–923.

Hodge RD, Bakken TE, Miller JA, Smith KA, Barkan ER, Graybuck LT, Close JL, Long B, Johansen N, Penn O, Yao Z, Eggermont J, Höllt T, Levi BP, Shehata SI, Aevermann B, Beller A, Bertagnolli D, Brouner K, Casper T, et al. Conserved cell types with divergent features in human versus mouse cortex. Nature. 2019; 573(7772):61–68.

Horvát S, Gămănuț R, Ercsey-Ravasz M, Magrou L, Gămănuț B, Van Essen DC, Burkhalter A, Knoblauch K, Toroczkai Z, Kennedy H. Spatial Embedding and Wiring Cost Constrain the Functional Layout of the Cortical Network of Rodents and Primates. PLOS Biol. 2016 07; 14(7):1–30. https://doi.org/10.1371/journal.pbio.1002512, doi: 10.1371/journal.pbio.1002512.

Hunter JD. Matplotlib: A 2D Graphics Environment. Comput Sci Eng. 2007; 9(3):90–95. doi: 10.1109/MCSE.2007.55.

Izhikevich EM, Edelman GM. Large-scale model of mammalian thalamocortical systems. Proc Natl Acad Sci USA. 2008; 105(9):3593–3598. https://doi.org/10.1073/pnas.0712231105, doi: 10.1073/pnas.0712231105.

Johnson RR, Burkhalter A. Microcircuitry of Forward and Feedback Connections Within Rat Visual Cortex. J Comp Neurol. 1996; 368:383–398.

Jordan J, Ippen T, Helias M, Kitayama I, Sato M, Igarashi J, Diesmann M, Kunkel S. Extremely Scalable Spiking Neuronal Network Simulation Code: From Laptops to Exascale Computers. Front Neuroinform. 2018 Feb; 12:2. https://doi.org/10.3389/fninf.2018.00002, doi: 10.3389/fninf.2018.00002.

Kabbara A, EL Falou W, Khalil M, Wendling F, Hassan M. The dynamic functional core network of the human brain at rest. Sci Rep. 2017 Jun; 7(1):2936. doi: 10.1038/s41598-017-03420-6.

Kandel ER, Schwartz JH, Jessel TM. Principles of Neural Science. 4 ed. New York: McGraw-Hill; 2000. ISBN 978-0838577011.

Köster J, Rahmann S. Snakemake—a scalable bioinformatics workflow engine. Biochemistry. 2012; 28(19):2520–2522.

Layer M, Senk J, Essink S, van Meegen A, Bos H, Helias M. NNMT: Mean-Field Based Analysis Tools for Neuronal Network Models. Front Neuroinform. 2022; 16:835657. doi: 10.3389/fninf.2022.835657.

Markov NT, Ercsey-Ravasz MM, Ribeiro Gomes AR, Lamy C, Magrou L, Vezoli J, Misery P, Falchier A, Quilodran R, Gariel MA, Sallet J, Gamanut R, Huissoud C, Clavagnier S, Giroud P, Sappey-Marinier D, Barone P, Dehay C, Toroczkai Z, Knoblauch K, et al. A Weighted and Directed Interareal Connectivity Matrix for Macaque Cerebral Cortex. Cereb Cortex. 2014; 24(1):17–36. doi: 10.1093/cercor/bhs270.

Markov NT, Misery P, Falchier A, Lamy C, Vezoli J, Quilodran R, Gariel MA, Giroud P, Ercsey-Ravasz M, Pilaz LJ, Huissoud C, Barone P, Dehay C, Toroczkai Z, Van Essen DC, Kennedy H, Knoblauch K. Weight Consistency Specifies Regularities of Macaque Cortical Networks. Cereb Cortex. 2011; 21(6):1254–1272.

Markov NT, Vezoli J, Chameau P, Falchier A, Quilodran R, Huissoud C, Lamy C, Misery P, Giroud P, Ullman S, Barone P, Dehay C, Knoblauch K, Kennedy H. Anatomy of hierarchy: Feedforward and feedback pathways in macaque visual cortex. J Comp Neurol. 2014; 522(1):225–259. doi: 10.1002/cne.23458.

Markram H, Muller E, Ramaswamy S, Reimann MW, Abdellah M, Sanchez CA, Ailamaki A, Alonso-Nanclares L, Antille N, Arsever S, Kahou GAA, Berger TK, Bilgili A, Buncic N, Chalimourda A, Chindemi G, Courcol JD, Delalondre F, Delattre V, Druckmann S, et al. Reconstruction and simulation of neocortical microcircuitry. Cell. 2015 Oct; 163(2):456–492. https://doi.org/10.1016/j.cell.2015.09.029, doi: 10.1016/j.cell.2015.09.029.

Marques JP, Kober T, Krueger G, van der Zwaag W, Van de Moortele PF, Gruetter R. MP2RAGE, a self bias-field corrected sequence for improved segmentation and T1-mapping at high field. Neuroimage. 2010; 49(2):1271–1281. https://www.sciencedirect.com/science/article/pii/S1053811909010738, doi: 10.1016/j.neuroimage.2009.10.002.

Wes McKinney. Data Structures for Statistical Computing in Python. In: Stéfan van der Walt, Jarrod Millman, editors. *Proceedings of the 9th Python in Science Conference*; 2010. p. 56 – 61. doi: 10.25080/Majora-92bf1922-00a.

Migliore M, Cavarretta F, Hines ML, Shepherd GM. Distributed organization of a brain microcircuit analyzed by three-dimensional modeling: the olfactory bulb. Front Comput Neurosci. 2014; 8(50):1–14. doi: 10.3389/fncom.2014.00050.

Migliore M, Cavarretta F, Marasco A, Tulumello E, Hines ML, Shepherd GM. Synaptic clusters function as odor operators in the olfactory bulb. Proc Natl Acad Sci USA. 2015 Jun; 112(27):8499–8504. https://doi.org/10.1073/pnas.1502513112, doi: 10.1073/pnas.1502513112.

Minxha J, Adolphs R, Fusi S, Mamelak AN, Rutishauser U. Flexible recruitment of memory-based choice representations by the human medial frontal cortex. Science. 2020; 368(6498):eaba3313.

Moeller S, Yacoub E, Olman CA, Auerbach E, Strupp J, Harel N, Uğurbil K. Multiband multislice GE-EPI at 7 tesla, with 16-fold acceleration using partial parallel imaging with application to high spatial and temporal whole-brain fMRI. Magn Reson Med. 2010; 63(5):1144–1153. https://onlinelibrary.wiley.com/doi/abs/10.1002/mrm.22361, doi: 10.1002/mrm.22361.

Mohan H, Verhoog MB, Doreswamy KK, Eyal G, Aardse R, Lodder BN, Goriounova NA, Asamoah B, B Brakspear AC, Groot C, van der Sluis S, Testa-Silva G, Obermayer J, Boudewijns ZS, Narayanan RT, Baayen JC, Segev I, Mansvelder HD, de Kock CP. Dendritic and axonal architecture of individual pyramidal neurons across layers of adult human neocortex. Cereb Cortex. 2015; 25(12):4839–4853.

Packer AM, Yuste R. Dense, Unspecific Connectivity of Neocortical Parvalbumin-Positive Interneurons: A Canonical Microcircuit for Inhibition? J Neurosci. 2011 Sep; 31(37):13260–13271. https://doi.org/10.1523/jneurosci.3131-11.2011, doi: 10.1523/JNEUROSCI.3131-11.2011.

Patel AX, Bullmore ET. A wavelet-based estimator of the degrees of freedom in denoised fMRI time series for probabilistic testing of functional connectivity and brain graphs. Neuroimage. 2016; 142:14–26. https://www.sciencedirect.com/science/article/pii/S1053811915003523, doi: 10.1016/j.neuroimage.2015.04.052.

Perin R, Berger TK, Markram H. A synaptic organizing principle for cortical neuronal groups. Proc Natl Acad Sci USA. 2011 Mar; 108(13):5419–5424. https://doi.org/10.1073/pnas.1016051108, doi: 10.1073/pnas.1016051108.

Potjans TC, Diesmann M. The Cell-Type Specific Cortical Microcircuit: Relating Structure and Activity in a Full-Scale Spiking Network Model. Cereb Cortex. 2014 Dec; 24(3):785–806. https://doi.org/10.1093/cercor/bhs358, doi: 10.1093/cercor/bhs358.

Pronold J, Jordan J, Wylie BJN, Kitayama I, Diesmann M, Kunkel S. Routing Brain Traffic Through the Von Neumann Bottleneck: Parallel Sorting and Refactoring. Front Neuroinform. 2022; 15. https://www.frontiersin.org/article/10.3389/fninf.2021.785068, doi: 10.3389/fninf.2021.785068.

Pulvermüller F, Tomasello R, Henningsen-Schomers MR, Wennekers T. Biological constraints on neural network models of cognitive function. Nat Rev Neurosci. 2021; 22:488–502.

Python Software Foundation, The Python programming language; 2008. Http://www.python.org.

Reimann MW, Anastassiou CA, Perin R, Hill SL, Markram H, Koch C. A biophysically detailed model of neocortical local field potentials predicts the critical role of active membrane currents. Neuron. 2013 Jul; 79(2):375–390. https://doi.org/10.1016/j.neuron.2013.05.023, doi: 10.1016/j.neuron.2013.05.023.

Rockland KS. What do we know about laminar connectivity? Neuroimage. 2019; 197:772–784.

Rohatgi A, Webplotdigitizer: Version 4.5; 2021. https://automeris.io/WebPlotDigitizer.

Schmidt M, Bakker R, Hilgetag CC, Diesmann M, van Albada SJ. Multi-scale account of the network structure of macaque visual cortex. Brain Struct Funct. 2018 Apr; 223(3):1409–1435. https://doi.org/10.1007/s00429-017-1554-4, doi: 10.1007/s00429-017-1554-4.

Schmidt M, Bakker R, Shen K, Bezgin G, Diesmann M, van Albada SJ. A multi-scale layer-resolved spiking network model of resting-state dynamics in macaque visual cortical areas. PLOS Comput Biol. 2018; 14(10):e1006359. https://doi.org/10.1371/journal.pcbi.1006359, doi: 10.1371/journal.pcbi.1006359.

Schuecker J, Schmidt M, van Albada SJ, Diesmann M, Helias M. Fundamental Activity Constraints Lead to Specific Interpretations of the Connectome. PLOS Comput Biol. 2017 Feb; 13(2):e1005179. https://doi.org/10.1371/journal.pcbi.1005179, doi: 10.1371/journal.pcbi.1005179.

Sejnowski TJ, Churchland PS, Movshon JA. Putting big data to good use in neuroscience. Nat Neurosci. 2014 Nov; 17(11):1440–1441. https://doi.org/10.1038/nn.3839, doi: 10.1038/nn.3839.

Shapson-Coe A, Januszewski M, Berger DR, Pope A, Wu Y, Blakely T, Schalek RL, Li PH, Wang S, Maitin-Shepard J, Karlupia N, Dorkenwald S, Sjostedt E, Leavitt L, Lee D, Bailey L, Fitzmaurice A, Kar R, Field B, Wu H, et al. A connectomic study of a petascale fragment of human cerebral cortex. BioRxiv. 2021; doi: 10.1101/2021.05.29.446289.

Sherwood CC, Miller SB, Karl M, Stimpson CD, Phillips KA, Jacobs B, Hof PR, Raghanti MA, Smaers JB. Invariant synapse density and neuronal connectivity scaling in primate neocortical evolution. Cereb Cortex. 2020; 30(10):5604–5615.

Smith SM, Jenkinson M, Woolrich MW, Beckmann CF, Behrens TEJ, Johansen-Berg H, Bannister PR, De Luca M, Drobnjak I, Flitney DE, Niazy RK, Saunders J, Vickers J, Zhang Y, De Stefano N, Brady JM, Matthews PM. Advances in functional and structural MR image analysis and implementation as FSL. Neuroimage. 2004; 23:S208–S219. https://www.sciencedirect.com/science/article/pii/S1053811904003933, doi: 10.1016/j.neuroimage.2004.07.051, mathematics in Brain Imaging.

Stephan KE, Kamper L, Bozkurt A, Burns GAPC, Young MP, Kötter R. Advanced database methodology for the collation of connectivity data on the macaque brain (CoCoMac). Philos Trans R Soc B. 2001; 356:1159–1186.

Teeter C, Iyer R, Menon V, Gouwens N, Feng D, Berg J, Szafer A, Cain N, Zeng H, Hawrylycz M, Koch C, Mihalas S. Generalized leaky integrate-and-fire models classify multiple neuron types. Nat Commun. 2018; 9:709.

Theodoni P, Majka P, Reser DH, Wójcik DK, Rosa MGP, Wang XJ. Structural Attributes and Principles of the Neocortical Connectome in the Marmoset Monkey. Cereb Cortex. 2021 07; 32(1):15–28. doi: 10.1093/cercor/bhab191.

Van Essen DC, Smith SM, Barch DM, Behrens TE, Yacoub E, Ugurbil K, Consortium WMH, et al. The WU-Minn Human Connectome Project: An overview. Neuroimage. 2013; 80:62–79.

Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, Bright J, van der Walt SJ, Brett M, Wilson J, Millman KJ, Mayorov N, Nelson ARJ, Jones E, Kern R, Larson E, Carey CJ, et al. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. Nat Methods. 2020; 17:261–272. doi: 10.1038/s41592-019-0686-2.

Von Economo C. Cellular Structure of the Human Cerebral Cortex. Karger Medical and Scientific Publishers; 2009. Translated and edited by L.C. Triarhou.

Waskom ML. seaborn: statistical data visualization. Journal of Open Source Software. 2021; 6(60):3021. https://doi.org/10.21105/joss.03021, doi: 10.21105/joss.03021.

Winnubst J, Bas E, Ferreira TA, Wu Z, Economo MN, Edson P, Arthur BJ, Bruns C, Rokicki K, Schauder D, et al. Reconstruction of 1,000 projection neurons reveals new cell types and organization of long-range connectivity in the mouse brain. Cell. 2019; 179(1):268–281.

Part III

DISCUSSION

9 DISCUSSION

9.1 SUMMARY & OUTLOOK

In this thesis, we investigated neural network models of varying complexity. In all cases, a central goal was to push analytical approaches as far as possible to capture the dynamics and function of the models. In the remainder of this chapter, the results presented in the main part are discussed one after the other before the individual results are then woven into a coherent picture.

9.1.1 *Chapter 4*

Summary

In Chapter 4, we investigated Bayesian supervised learning in very wide feedforward and very large recurrent networks (DNNs and RNNs). Using a field-theoretical approach similar to DMFT (see Section 3.4), we derived iterative equations for the kernel of both DNN and RNN in a unified manner. From a DMFT perspective, the calculation corresponds to a M-replica calculation, where M denotes the size of the dataset, with each replicon having a different initial condition determined by the input datum.

We found that the kernels of the two architectures differ: for DNNs, there are no correlations between the layers due to the independent priors of the weights; for RNNs, there are correlations between the time steps due to weight sharing (see also Mozeika, Li, and Saad 2020). Curiously, there is no difference between the kernels in the within-layer or equal-time statistics. This is due to the peculiar structure structure of the iterative equations for the kernel—the within-layer or equal-time statistics can be obtained without the across-layer or across-time statistics. Thus, if the output depends only on the last layer or time step, both architectures have an equivalent performance in the infinite-size limit.

We verified numerically that the empirical distributions approach those predicted by the theory using Maximum Mean Discrepancy. For sufficiently large networks of size $O(1{,}000)$, the correlations predicted by the theory match the empirical ones very well. In particular, both the equal within-layer and equal-time statistics as well as the different across-layer and across-time statistics are well captured.

Outlook

The field-theoretical formulation of the problem paves the way to finite-size corrections (e.g., Zinn-Justin 1996; Moshe and Zinn-Justin 2003; Kleinert 2009). For feedforward networks, there is already a lot of effort devoted to this line of research (e.g., Yaida 2020; Dyer and Gur-Ari 2020; Antognini 2019; Huang and Yau 2020; Halverson, Maiti, and Stoner 2021; Naveh et al. 2021; Roberts, Yaida, and Hanin 2022). In contrast, there is less work on recurrent networks (e.g., Alemohammad et al. 2021; Grosvenor and Jefferson 2022). Investigating the finite-size corrections seems particularly important to further understand the differences between the two architectures.

In terms of the models, our work focused on the most simple feedforward and recurrent networks. The work by Yang (2019) already yields the leading-order kernel for a wide class of models, thus it should be possible to extend our framework. One possible direction is to consider residual networks (He et al. 2016) in which skip connections lead to non-vanishing correlations between layers in feedforward networks. Furthermore, taking the continuum limit, a residual networks becomes a neural ODE (Chen, Rubanova, et al. 2018). Since DMFT originated from continuous-time dynamical systems, this extension seems natural in the context of the field-theoretical approach.

In the long run it will be interesting to see how much insight can be gained by starting from the infinite-size limit and computing corrections (see the textbook on this approach by Roberts, Yaida, and Hanin 2022). Furthermore, it will be valuable to deeply understand the differences between the Bayesian framework used here and gradient-based learning.

9.1.2 *Chapter 5*

Summary

In Chapter 5 we considered block-structured, random networks of rate neurons. We calculated the distribution of the empirical measure of the trajectories across the ensemble of random networks; more precisely its leading-order exponential contribution. This leading-order contribution—the rate function—takes the form of a Kullback-Leibler divergence. The exponential form implies self-averaging: the distribution of the empirical measure is sharply peaked at the maximum with deviations suppressed by $1/N$. Thus, for any given realization of the random network, one obtains an empirical measure which is close to the most likely one. Consequently, by the contraction principle, all network-averaged observables attain a value close to the one described by the most likely empirical measure.

While the rate function is the main theoretical result of this chapter, it is but the starting point for the resulting applications. In the first

application, we used the rate function to investigate the fluctuations around the most likely value, i.e., beyond-mean-field fluctuations. Concretely, we considered the equal-time, network-averaged variance of the single-unit activity; this is an order parameter indicating the transition to a chaotic state (Sompolinsky, Crisanti, and Sommers 1988). For slow single-neuron dynamics close to but above the transition to chaos, we obtained an analytical result for the order parameter fluctuations which matches the empirical result from simulations well. This analytical result led us to a hitherto unknown network state with two stable mean-field solutions and finite-size-fluctuation-driven transitions between these solutions. To obtain this state, a simple modification of the nonlinear interaction is sufficient: replacing the sigmoidal nonlinearity, which has a negative third-order Taylor coefficient, with an expansive nonlinearity with a positive third-order Taylor coefficient.

In a second application, we used the rate function to infer the network statistics from trajectories. From the point of view of inference, the rate function is the (scaled, negative) log-likelihood. Setting the derivative of the rate function to zero, we obtained a necessary condition which needs to hold at the maximum of the likelihood. This condition involves network-averaged power spectra and is linear in the parameters. The spectra are straightforward to obtain from the trajectories, hence the parameters can be extracted using non-negative least squares. This allowed us to successfully infer the parameters both for single- and multi-population networks. We complemented the inference procedure with two methods for model comparison, e.g., to determine the nonlinearity. Furthermore, we used the inferred parameters to predict the future of the single-neuron trajectory.

Outlook

The rate function turned out to be useful for the aforementioned applications. However, these applications do not make full use of our result: the rate function also governs the tails of the distribution far away from the maximum (see Section 3.1). An interesting application where the tails of the distribution, i.e., large fluctuations, become relevant might be the (finite-size-)fluctuation-induced transition between two stable mean-field solutions—only very rare large fluctuations drive the system out of one stable mean-field solution into the other one. To determine the rate of these transitions, for example, the rate function seems to be a natural starting point.

More generally, it might be interesting to extend the rate function beyond the class of models that we considered to arbitrary input-output relationships akin to the model-independent DMFT by Keup et al. (2021). This would allow one to apply the parameter inference procedure also to more general models, e.g., spiking networks.

From a conceptual point of view, we believe that our work advanced the foundations of DMFT in two ways: First, our results show that DMFT can be used to describe any network-averaged observable. Second, we created a link between the field theoretical approach used in the physics community (Crisanti and Sompolinsky 2018; Helias and Dahmen 2020) and the large deviation approach used in the mathematical community (Arous and Guionnet 1995; Guionnet 1997; Faugeras and MacLaurin 2015; Faugeras, MacLaurin, and Tanré 2019). We hope that these links lead to further cross-fertilization between the disciplines.

9.1.3 *Chapter 6*

Summary

In Chapter 6 we linked neuron- and network-parameters to the autocorrelation time of the emerging dynamics in block-structured, spiking networks using DMFT. We focused on two different spiking neuron models: generalized linear model neurons (GLMs) and leaky integrate-and-fire neurons (LIFs). For GLMs, we solved the colored noise problem analytically for exponential and error function nonlinearities. For LIFs, we first considered the voltage dynamics in the absence of the fire-and-reset mechanism and calculated the non-stationary upcrossing probability as well as the stationary correlation function between upcrossings. Together with a renewal approximation and an approximation of the hazard function proposed by Stratonovich (1967), this allowed us to obtain the output statistics of LIF neurons with exponential synapses. A fixed-point iteration using the analytical solution for GLMs or the approximate solution for LIFs yielded self-consistent correlation functions which were successfully validated with simulations. The comparison with simulations furthermore demonstrated that it is necessary to distinguish population-averaged single-neuron statistics from the statistics of the population activity—with our theory capturing only the former.

The second-order statistics contain temporal correlations as well as static contributions, for example caused by the heterogeneity of indegrees across neurons. Both are captured by our theory. An example for a static variability is the distribution of firing rates which matches the analytical prediction for GLMs well (for LIFs, we only determined the variance of the firing rates, not their full distribution). Furthermore, our theory accounts for the effect of temporally correlated external input. The external timescale affects the timescale of the dynamics in a straightforward manner: in the limit of very strong external input, the external timescale determines the timescale of the dynamics if it is above the maximal timescale of the units. If it is below the maximal timescale of the units the latter determines the timescale

of the dynamics. Decreasing the strength of the external input, the timescale of the dynamics approaches the autonomous case.

For balanced networks of excitatory and inhibitory neurons, we leveraged the theory to perform parameter scans of the autocorrelation time. Within the investigated parameter range we found a maximum of a two- to three-fold increase of the autocorrelation compared to the membrane time constant. The mechanisms leading to the increased autocorrelation time differ between the two neuron models: for GLM neurons, temporal correlations in the voltage dynamics lead to bursts of spikes; for LIF neurons, the increased autocorrelation time corresponds to an increased effective refractory period (a dip in the autocorrelation function). In a network model of a cortical column containing four excitatory and four inhibitory populations, the theory still captured the bulk of the spectrum but not a peak in the spectrum due to a fast, population-level oscillation.

Outlook

Capturing such an interplay between population activity, e.g., oscillations, and single-neuron statistics is an interesting avenue for future work. On the technical level, this requires corrections beyond DMFT because the neurons factorize in DMFT.

The timescales uncovered with our theory are still small compared to the timescales observed experimentally (Murray et al. 2014). A simple way to generate longer timescales would be to introduce slow processes on the neuron level, e.g., spike frequency adaptation. But the more interesting question is whether the timescales can be generated on the network level. It is well-known that increasing the synaptic weights yields slow fluctuations (Ostojic 2014; Wieland et al. 2015). However, the increased synaptic weights lead to extreme fluctuations of the membrane potential (Kriener et al. 2014). A potential network mechanism is a clustered connectivity of either the excitatory neurons (Litwin-Kumar and Doiron 2012) or of both excitatory and inhibitory neurons (Rost, Deger, and Nawrot 2018) which leads to slow fluctuations in the cluster activation. Another possibility might be to embed a feedforward structure into the network (Ganguli, Huh, and Sompolinsky 2008; Murphy and Miller 2009; Hennequin, Vogels, and Gerstner 2012).

The autocorrelation time is convenient because it is a single number. Its numerical value, however, depends on many factors which are hard to disentangle in experimental data. For example: Is the activity indeed stationary? Is the bias due to finite recording time (Grigera 2020; Zeraati, Engel, and Levina 2020) relevant? How does binning the spike train affect the autocorrelation time? Conversely, these difficulties imply that it is not straightforward to compare the theoretical autocorrelation time with experimental data. Thus, it is important to investigate the influence of these factors on the estimates of the

autocorrelation time in order to tie a strong link between theory and experiment.

9.1.4 *Chapter 7*

Summary

In Chapter 7 we analyzed the distribution of neuron densities within and across (cytoarchitectonically defined) cortical areas in several mammalian species. Neither a hypothesis test nor a model comparison excluded the possibility that neuron densities are log-normally distributed in the majority of cases.

We proposed a simple model of noisy cell division to account for the log-normal distribution within areas: the rate of cell division λ is a Gaussian process such that the dynamical equation for the density $\dot{\rho} = \lambda\rho$ is a stochastic differential equation leading to a log-normally distributed solution $\rho(t)$.

Outlook

Thus far, our model only accounts for the log-normal distribution within areas. A salient difference between the areas is a broad distribution of development times (Rakic 2002; Cadwell et al. 2019). It would be interesting to extend the model in a way where distributed development times lead to a log-normal distribution of the mean density across areas.

Our finding raises the question whether the long-tailed log-normal distribution of neuron densities is simply a byproduct of the inherent noise in biological processes or whether it serves a purpose, e.g., to facilitate computation (Duarte and Morrison 2019; Perez-Nieves et al. 2021). More generally, heterogeneity is ubiquitous in the brain (Buzsáki and Mizuseki 2014)—an intriguing hypothesis, although it is hard to test, is that this heterogeneity mirrors the heterogeneous environment that we live in.

9.1.5 *Chapter 8*

Summary

In Chapter 8 we built a multi-scale, data-based, spiking model of human cortex. The model is built on, among other modalities, cytoarchitectonic data, neuron morphologies, electrophysiology, and DTI. Gaps in the data were filled using statistical regularities found in other species. We validated the resulting network structure against features from tracing data, in particular the log-normal distribution of connection densities, the exponential decay of connection density with distance, and the differential arborization of feedforward and

feedback axons. To determine the neuron parameters we analyzed LIF models from the Allen Cell Types database which were fitted to human electrophysiology data.

Simulations showed that the model produces asynchronous and irregular activity with low firing rates and multiple levels of heterogeneity in the activity; from distributed firing rates of neurons inside individual population to systematic differences between the areas. However, a comparison of the model's activity with electrophysiological recordings and fMRI data clearly showed that the reproduction is not satisfactory on either level. On the single neuron level, the distribution of firing rates is too broad while the irregularity is too low; on the area level, the correlations between the areas are vanishingly small.

Outlook

The prerequisite for further steps is a good agreement of the model's activity with the experimental data on both levels, spike trains and functional connectivity. A major obstacle is that specific alterations of the model, e.g., increasing the cortico-cortical connection strength or the external input beyond a certain point, lead to a highly synchronous state with firing rates far outside of the biophysiologically plausible regime. Overcoming this obstacle is the main focus of the ongoing work on the model.

Why bother? One possible answer is that in silico, experiments can be performed which are impossible in vivo and in vitro. For example, it is know that stimulating a single neuron can alter behavior (Brecht et al. 2004; Houweling and Brecht 2008). In vivo, it is not possible to investigate how the single-neuron stimulation propagates through the network because the propagation critically depends on the ongoing activity of the network which is not possible to duplicate (at least for the foreseeable future). In silico, in contrast, this is a straightforward exercise: one simply has to fix the random seed.

9.2 SYNTHESIS

Neuroscience exhibits a fascinating collection of paradigms (Parker 2018). While paradigm shifts are often the focus of attention, it is what Kuhn (2012) calls "normal science"—fleshing out all the implications entailed in the current paradigm—that lies at the heart of scientific progress. From such a Kuhnian perspective, this thesis can be seen as normal science under a random network paradigm.

Random networks are not a broadly accepted paradigm despite their frequent use in theoretical studies. There are two immediate objections against studying random networks: cortex is not a random network (e.g., Song et al. 2005) and random networks do not support a function. However, a random network might be a good first-order

approximation for cortical networks such that non-random features can subsequently be studied. This is the overarching theme of Chapter 6 and Chapter 8. In Chapter 6 we investigated the emergent timescale of the dynamics and developed a corresponding theory. This exercise revealed that non-random features are necessary in order to achieve the longer timescales of the dynamics observed in vivo. In future investigations, the theory will ideally prove valuable in determining which non-random features that lead to longer timescales are the most interesting ones. In Chapter 8 we built a multi-scale, data-based, block-structured random network model of human cortex with the goal of reproducing low-order activity statistics across the scales. Once we succeed in reproducing the activity statistics we considered thus far, this paves the way for subsequent refinements of the model in order to match an ever increasing set of observables.

Regarding functional aspects, Chapter 4 shows how closely related random networks and their functional counterparts can be. From the perspective of Bayesian supervised learning, the only difference is a conditioning operation, i.e., conditioning the network prior on the training data. This shifts the focus from function back to ensembles of random networks. How to treat these ensembles of random networks analytically was at the heart of Chapter 5.

Naturally, neither of the two above objections against studying random networks are addressed exhaustively within the scope of this thesis, opening the door to future research. One immediately obvious direction is to extend the methods from Chapter 4 and Chapter 5 to the more complex models considered in the later chapters. For example, how does the kernel corresponding to the model by Potjans and Diesmann (2014) look like? But also further investigation into which features can and, more crucially, cannot be accounted for by simple random networks is needed.

It seems rather unlikely that random networks become a generally accepted paradigm across the many sub-disciplines of neuroscience. Yet, it will be interesting to see how far this simple paradigm can be pushed.

Part IV

APPENDIX

A

NNMT: MEAN-FIELD BASED ANALYSIS TOOLS FOR NEURONAL NETWORK MODELS

PREAMBLE

A theory frequently does not lead to a closed-form analytical solution but requires numerical methods to derive its predictions. Implementing these numerical methods is inherently prone to errors and takes time. This problem is easily solved with a publicly available toolbox with a comprehensive test suite (Riquelme and Gjorgjieva 2021).

We developed such a toolbox for mean-field based methods with a focus on spiking neural networks: NNMT, the Neuronal Network Mean-Field Toolbox. The toolbox is based on the diffusion approximation (see Section 3.4) and comprises, among others, methods to determine the steady state firing rate and linear response properties. It fills a gap because there exists, to the best of our knowledge, no publicly available implementation of these methods.

To demonstrate the use of the toolbox, we reproduced the results of several studies. In the manuscript below, these reproductions are embedded into a coherent description of the toolbox and a discussion of its use and limits.

Author Contributions

Moritz Layer (ML) developed and implemented the current version of the toolbox, including the test suite and the online documentation, under the supervision of Dr. Johanna Senk (JS) and Prof. Moritz Helias (MH). Simon Essink (SE) and the author (AvM) contributed to various parts of the toolbox; in particular, AvM contributed the algorithm described in appendix A.1 of the manuscript. The toolbox is based on earlier work by, among others, Dr. Hannah Bos and Dr. Jannis Schücker. ML (sections 1, 2, 3.1, 3.2.2, 4), JS (section 3.4), SE (section 3.3, appendix A.2), and AvM (section 3.2.1, appendix A.1) wrote the first draft of the manuscript. The examples and figures were created by the authors of the corresponding sections. The manuscript was jointly revised by ML, JS, SE, AvM, and MH.

frontiers | Frontiers in Neuroinformatics

TECHNOLOGY AND CODE
published: 27 May 2022
doi: 10.3389/fninf.2022.835657

NNMT: Mean-Field Based Analysis Tools for Neuronal Network Models

Moritz Layer[1,2]*, *Johanna Senk*[1], *Simon Essink*[1,2], *Alexander van Meegen*[1,3], *Hannah Bos*[1] *and Moritz Helias*[1,4]

[1] *Institute of Neuroscience and Medicine (INM-6) and Institute for Advanced Simulation (IAS-6) and JARA-Institute Brain Structure-Function Relationships (INM-10), Jülich Research Centre, Jülich, Germany,* [2] *RWTH Aachen University, Aachen, Germany,* [3] *Institute of Zoology, Faculty of Mathematics and Natural Sciences, University of Cologne, Cologne, Germany,* [4] *Department of Physics, Faculty 1, RWTH Aachen University, Aachen, Germany*

Mean-field theory of neuronal networks has led to numerous advances in our analytical and intuitive understanding of their dynamics during the past decades. In order to make mean-field based analysis tools more accessible, we implemented an extensible, easy-to-use open-source Python toolbox that collects a variety of mean-field methods for the leaky integrate-and-fire neuron model. The Neuronal Network Mean-field Toolbox (NNMT) in its current state allows for estimating properties of large neuronal networks, such as firing rates, power spectra, and dynamical stability in mean-field and linear response approximation, without running simulations. In this article, we describe how the toolbox is implemented, show how it is used to reproduce results of previous studies, and discuss different use-cases, such as parameter space explorations, or mapping different network models. Although the initial version of the toolbox focuses on methods for leaky integrate-and-fire neurons, its structure is designed to be open and extensible. It aims to provide a platform for collecting analytical methods for neuronal network model analysis, such that the neuroscientific community can take maximal advantage of them.

Keywords: mean-field theory, (spiking) neuronal network, integrate-and-fire neuron, open-source software, parameter space exploration, (hybrid) modeling, python, computational neuroscience

OPEN ACCESS

Edited by:
John David Griffiths, University of Toronto, Canada

Reviewed by:
Caglar Cakan, Technical University of Berlin, Germany
Richard Gast, Max Planck Institute for Human Cognitive and Brain Sciences, Germany

***Correspondence:**
Moritz Layer
m.layer@fz-juelich.de

Received: *14 December 2021*
Accepted: *17 March 2022*
Published: *27 May 2022*

Citation:
Layer M, Senk J, Essink S, van Meegen A, Bos H and Helias M (2022) NNMT: Mean-Field Based Analysis Tools for Neuronal Network Models. Front. Neuroinform. 16:835657. doi: 10.3389/fninf.2022.835657

1. INTRODUCTION

Biological neuronal networks are composed of large numbers of recurrently connected neurons, with a single cortical neuron typically receiving synaptic inputs from thousands of other neurons (Braitenberg and Schüz, 1998; DeFelipe et al., 2002). Although the inputs of distinct neurons are integrated in a complex fashion, such large numbers of weak synaptic inputs imply that average properties of entire populations of neurons do not depend strongly on the contributions of individual neurons (Amit and Tsodyks, 1991). Based on this observation, it is possible to develop analytically tractable theories of population properties, in which the effects of individual neurons are averaged out and the complex, recurrent input to individual neurons is replaced by a self-consistent effective input (reviewed, e.g., in Gerstner et al., 2014). In classical physics terms (e.g., Goldenfeld, 1992), this effective input is called *mean-field*, because it is the self-consistent mean of a *field*, which here is just another name for the input the neuron is receiving. The term *self-consistent* refers to the fact that the population of neurons that receives the effective input is the same that contributes to this very input in a recurrent fashion: the population's output determines its input and vice-versa. The stationary statistics of the effective input therefore can be found in a

self-consistent manner: the input to a neuron must be set exactly such that the caused output leads to the respective input.

Mean-field theories have been developed for many different kinds of synapse, neuron, and network models. They have been successfully applied to study average population firing rates (van Vreeswijk and Sompolinsky, 1996, 1998; Amit and Brunel, 1997b), and the various activity states a network of spiking neurons can exhibit, depending on the network parameters (Amit and Brunel, 1997a; Brunel, 2000; Ostojic, 2014), as well as the effects that different kinds of synapses have on firing rates (Fourcaud and Brunel, 2002; Lindner, 2004; Schuecker et al., 2015; Schwalger et al., 2015; Mattia et al., 2019). They have been used to investigate how neuronal networks respond to external inputs (Lindner and Schimansky-Geier, 2001; Lindner and Longtin, 2005), and they explain why neuronal networks can track external input on much faster time scales than a single neuron could (van Vreeswijk and Sompolinsky, 1996, 1998). Mean-field theories allow studying correlations of neuronal activity (Sejnowski, 1976; Ginzburg and Sompolinsky, 1994; Lindner et al., 2005; Trousdale et al., 2012) and were able to reveal why pairs of neurons in random networks, despite receiving a high proportion of common input, can show low output correlations (Hertz, 2010; Renart et al., 2010; Tetzlaff et al., 2012; Helias et al., 2014), which for example has important implication for information processing. They describe pair-wise correlations in network with spatial organization (Rosenbaum and Doiron, 2014; Rosenbaum et al., 2017; Dahmen et al., 2022) and can be generalized to correlations of higher orders (Buice and Chow, 2013). Mean-field theories were utilized to show that neuronal networks can exhibit chaotic dynamics (Sompolinsky et al., 1988; van Vreeswijk and Sompolinsky, 1996, 1998), in which two slightly different initial states can lead to totally different network responses, which has been linked to the network's memory capacity (Toyoizumi and Abbott, 2011; Schuecker et al., 2018). Most of the results mentioned above have been derived for networks of either rate, binary, or spiking neurons of a linear integrate-and-fire type. But various other models have been investigated with similar tools as well; for example, just to mention a few, Hawkes processes, non-linear integrate-and-fire neurons (Brunel and Latham, 2003; Fourcaud-Trocmé et al., 2003; Richardson, 2007, 2008; Grabska-Barwinska and Latham, 2014; Montbrió et al., 2015), or Kuramoto-type models (Stiller and Radons, 1998; van Meegen and Lindner, 2018). Additionally, there is an ongoing effort showing that many of the results derived for distinct models are indeed equivalent and that those models can be mapped to each other under certain circumstances (Ostojic and Brunel, 2011; Grytskyy et al., 2013; Senk et al., 2020).

Other theories for describing mean population rates in networks with spatially organized connectivity, based on taking a continuum limit, have been developed. These theories, known as neural field theories, have deepened our understanding of spatially and temporally structured activity patterns emerging in cortical networks, starting with the seminal work by Wilson and Cowan (1972, 1973), who investigated global activity patterns, and Amari (1975, 1977), who studied stable localized neuronal activity. They were successfully applied to explain hallucination patterns (Ermentrout and Cowan, 1979; Bressloff et al., 2001), as well as EEG and MEG rhythms (Nunez, 1974; Jirsa and Haken, 1996, 1997). The neural field approach has been used to model working memory (Laing et al., 2002; Laing and Troy, 2003), motion perception (Giese, 2012), cognition (Schöner, 2008), and more; for extensive reviews of the literature, we refer the reader to Coombes (2005), Bressloff (2012), and Coombes et al. (2014).

Clearly, analytical theories have contributed to our understanding of neuronal networks and they provide a plethora of powerful and efficient methods for network model analysis. Comparing the predictions of analytical theories to simulations, experimental data, or other theories necessitates a numerical implementation applicable to various network models, depending on the research question. Such an implementation is often far from straightforward and at times requires investing substantial time and effort. Commonly, such tools are implemented as the need arises, and their reuse is not organized systematically and restricted to within a single lab. This way, not only are effort and costs spent by the neuroscientific community duplicated over and over again, but also are many scientists deterred from taking maximal advantage of those methods although they might open new avenues for investigating their research questions.

In order to make analytical tools for neuronal network model analysis accessible to a wider part of the neuroscientific community, and to create a platform for collecting well-tested and validated implementations of such tools, we have developed the Python toolbox NNMT (Layer et al., 2021), short for Neuronal Network Mean-field Toolbox. We would like to emphasize that NNMT is not a simulation tool; NNMT is a collection of numerically solved mean-field equations that directly relate the parameters of a microscopic network model to the statistics of its dynamics. NNMT has been designed to fit the diversity of mean-field theories, and the key features we are aiming for are modularity, extensibility, and a simple usability. Furthermore, it features an extensive test suite to ensure the validity of the implementations as well as a comprehensive user documentation. The current version of NNMT mainly comprises tools for investigating networks of leaky integrate-and-fire neurons as well as some methods for studying binary neurons and neural field models. The toolbox is open-source and publicly available on GitHub.[1]

In the following, we present the design considerations that led to the structure and implementation of NNMT as well as a representative set of use cases. Section 2 first introduces its architecture. Section 3 then explains its usage by reproducing previously published network model analyses from Schuecker et al. (2015), Bos et al. (2016), Sanzeni et al. (2020), and Senk et al. (2020). Section 4 compares NNMT to other available toolboxes for neuronal network model analysis, discusses its use cases from a more general perspective, indicates current limitations and prospective advancements of NNMT, and explains how new tools can be contributed.

[1] https://github.com/INM-6/nnmt

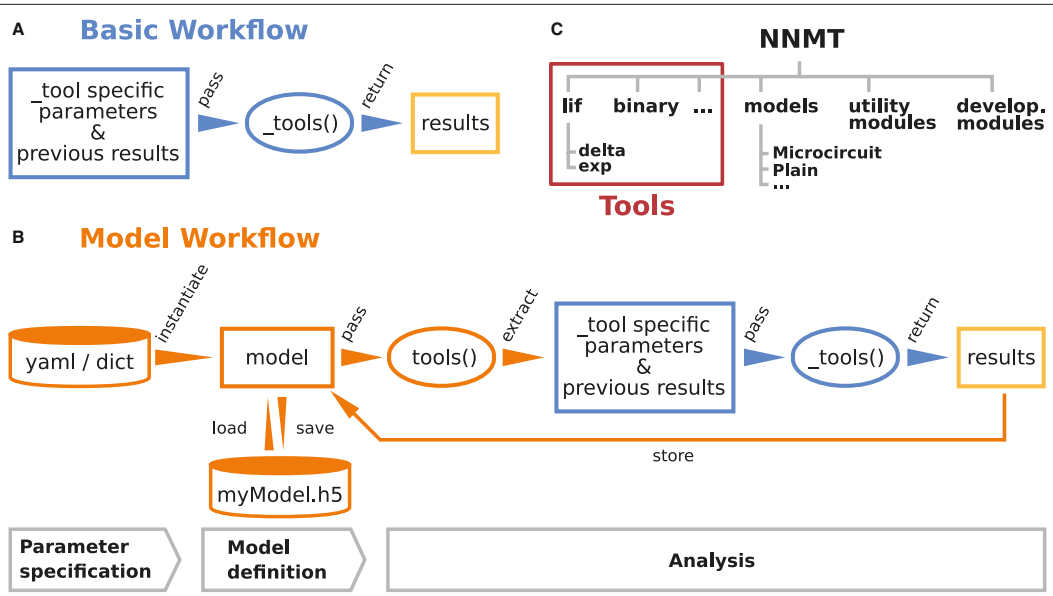

FIGURE 1 | Structure and workflows of the Neuronal Network Mean-field Toolbox (NNMT). **(A)** Basic workflow: individual mean-field based analysis methods are implemented as functions, called `_tools()`, that can be used directly by explicitly passing the required arguments. **(B)** Model workflow: to facilitate the handling of parameters and results, they can be stored in a model class instance, which can be passed to a `tool()`, which wraps the basic workflow of the respective `_tool()`. **(C)** Structure of the Python package. In addition to the tool collection (red frame), containing the `tools()` and the `_tools()`, and pre-defined model classes, the package provides utility functions for handling parameter files and unit conversions, as well as software aiding the implementation of new methods.

```
# basic workflow
result = nnmt.<submodule>.<_tool>(*args, **kwargs)

# model workflow
my_model = nnmt.models.<model>(
    <network_params>, <analysis_params>)
result = nnmt.<submodule>.<tool>(my_model)
```

Listing 1: The two modes of using NNMT: In the basic workflow (top), quantities are calculated by passing all required arguments directly to the underscored tool functions available in the submodules of NNMT. In the model workflow (bottom), a model class is instantiated with parameter sets and the model instance is passed to the non-underscored tool functions which automatically extract the relevant parameters.

2. WORKFLOWS AND ARCHITECTURE

What are the requirements a package for collecting analytical methods for neuronal network model analysis needs to fulfill? To begin with, it should be adaptable and modular enough to accommodate many and diverse analytical methods while avoiding code repetition and a complex interdependency of package components. It should enable the application of the collected algorithms to various network models in a simple and transparent manner. It should make the tools easy to use for new users, while also providing experts with direct access to all parameters and options. Finally, the methods need to be thoroughly tested and well documented.

These are the main considerations that guided the development of NNMT. **Figures 1A,B** illustrate how the toolbox can be used in to two different workflows, depending on the preferences and goals of the user. In the *basic workflow* the individual method implementations called *tools* are directly accessed, whereas the *model workflow* provides additional functionality for the handling of parameters and results.

2.1. Basic Workflow

The core of NNMT is a collection of low-level functions that take specific parameters (or pre-computed results) as input arguments and return analytical results of network properties. In **Figure 1A**, we refer to such basic functions as `_tools()`, as their names always start with an underscore. We term this lightweight approach of directly using these functions the basic workflow. The top part of **Listing 1** demonstrates this usage; for example, the quantity to be computed could be the mean firing rate of a neuronal population and the arguments could be parameters which define neuron model and external drive. While the basic workflow gives full flexibility and direct access to every

parameter of the calculation, it remains the user's responsibility to insert the arguments correctly, e.g., in the right units.

2.2. Model Workflow

The model workflow is a convenient wrapper of the basic workflow (**Figure 1B**). A *model* in this context is an object that stores a larger set of parameters and can be passed directly to a `tool()`, the non-underscored wrapper of the respective `_tool()`. The `tool()` automatically extracts the relevant parameters from the model, passes them as arguments to the corresponding core function `_tool()`, returns the results, and stores them in the model. The bottom part of **Listing 1** shows how a model is initialized with parameters and then passed to a `tool()` function.

Models are implemented as Python classes and can be found in the submodule `nnmt.models`. We provide the class `nnmt.models.Network` as a parent class and a few child classes which inherit the generic methods and properties but are tailored to specific network models; custom models can be created straightforwardly. The parameters distinguish network parameters, which define neuron models and network connectivity, and analysis parameters; an example for an analysis parameter is a frequency range over which a function is evaluated. Upon model instantiation, parameter sets defining values and corresponding units are passed as Python dictionaries or `yaml` files. The model constructor takes care of reading in these parameters, computing dependent parameters from the imported parameters, and converting all units to SI units for internal computations. Consequently, the parameters passed as arguments and the functions for computing dependent parameters of a specific child class need to be aligned. This design encourages a clear separation between a concise set of base parameters and functionality that transforms these parameters to the generic (vectorized) format that the tools work with. To illustrate this, consider the weight matrix of a network of excitatory and inhibitory neuron populations in which all excitatory connections have the same weight and all inhibitory ones another weight. As argument one could pass just a tuple of two different weight values and the corresponding model class would take care of constructing the full weight matrix. This happens in the example presented in Section 3.2.2: The parameter file `network_params_microcircuit.yaml` contains the excitatory synaptic weight and the ratio of inhibitory to excitatory weights. On instantiation, the full weight matrix is constructed from these two parameters, following the rules defined in `nnmt.models.Microcircuit`.

When a `tool()` is called, it checks whether the provided model object contains all required parameters and previously computed results. Then the `tool()` extracts the required arguments, calls the respective `_tool()`, and caches and returns the result. If the user attempts to compute the same property twice, using identical parameters, the `tool()` will retrieve the already computed result from the model's cache and return that value. Results can be exported to an HDF5 file and also loaded.

Using the model workflow instead of the basic workflow comes with the initial overhead of choosing a suitable combination of parameters and a model class, but has the advantages of a higher level of automation with built-in mechanisms for checking correctness of input (e.g., regarding units), reduced redundancy, and the options to store and load results. Both modes of using the toolbox can also be combined.

2.3. Structure of the Toolbox

The structure of the Python package NNMT is depicted in **Figure 1C**. It is subdivided into submodules containing the tools (e.g., `nnmt.lif.exp`, or `nnmt.binary`), the model classes (`nnmt.models`), helper routines for handling parameter files and unit conversions, as well as modules that collect reusable code employed in implementations for multiple neuron models (cf. Section 4.4). The tools are organized in a modular, extensible fashion with a streamlined hierarchy. To give an example, a large part of the currently implemented tools apply to networks of leaky integrate-and-fire (LIF) neurons, and they are located in the submodule `nnmt.lif`. The mean-field theory for networks of LIF neurons distinguishes between neurons with instantaneous synapses, also called delta synapses, and those with exponentially decaying post-synaptic currents. Similarly, the submodule for LIF neurons is split further into the two submodules `nnmt.lif.delta` and `nnmt.lif.exp`. NNMT also collects different implementations for computing the same quantity using different approximations or numerics, allowing for a comparison of different approaches.

Apart from the core package, NNMT comes with an extensive online documentation,[2] including a quickstart tutorial, all examples presented in this paper, a complete documentation of all tools, as well as a guide for contributors.

Furthermore, we provide an extensive test suite that validates the tools by checking them against previously published results and alternative implementations where possible. This ensures that future improvements of the numerics do not break the tools.

3. HOW TO USE THE TOOLBOX

In this section, we demonstrate the practical use of NNMT by replicating a variety of previously published results. The examples presented have been chosen to cover a broad range of common use cases and network models. We include analyses of both stationary and dynamic network features, as mean-field theory is typically divided into two parts: stationary theory, which describes time-independent network properties of systems in a stationary state, and dynamical theory, which describes time-dependent network properties. Additionally, we show how to use the toolbox to map a spiking to a simpler rate model, as well as how to perform a linear stability analysis. All examples, including the used parameter files, are part of the online documentation.[2]

3.1. Installation and Setup

The toolbox can be either installed using pip:

```
pip install nnmt
```

or by installing it directly from the repository, which is described in detail in the online

[2]https://nnmt.readthedocs.io/

documentation. After the installation, the module can be imported:

```
import nnmt
```

3.2. Stationary Quantities

3.2.1. Response Nonlinearities

Networks of excitatory and inhibitory neurons (EI networks, **Figure 2A**) are widely used in computational neuroscience (Gerstner et al., 2014), e.g., to show analytically that a balanced state featuring asynchronous, irregular activity emerges dynamically in a broad region of the parameter space (van Vreeswijk and Sompolinsky, 1996, 1998; Brunel, 2000; Hertz, 2010; Renart et al., 2010). Remarkably, such balance states emerge in inhibition dominated networks for a variety of neuron models if the indegree is large, $K \gg 1$, and the weights scale as $J \propto 1/\sqrt{K}$ (Sanzeni et al., 2020; Ahmadian and Miller, 2021). Furthermore, in a balanced state, a network responds linearly to external input in the limit $K \to \infty$ (van Vreeswijk and Sompolinsky, 1996, 1998; Brunel, 2000; Sanzeni et al., 2020; Ahmadian and Miller, 2021). How do EI networks of LIF neurons respond to external input at finite indegrees? Sanzeni et al. (2020) uncover five different types of nonlinearities in the network response depending on the network parameters. Here, we show how to use the toolbox to reproduce their result (**Figures 2B–F**).

The network consists of two populations, E and I, of identical LIF neurons with instantaneous (*delta*) synapses (Gerstner et al., 2014). The subthreshold dynamics of the membrane potential V_i of neuron i obeys

$$\tau_\mathrm{m} \dot{V}_i = -V_i + RI_i\,, \tag{1}$$

where τ_m denotes the membrane time constant, R the membrane resistance, and I_i the input current. If the membrane potential exceeds a threshold V_th, a spike is emitted and the membrane voltage is reset to the reset potential V_0 and clamped to this value during the refractory time τ_r. After the refractory period, the dynamics continue according to Equation (1). For instantaneous synapses, the input current is given by

$$RI_i(t) = \tau_\mathrm{m} \sum_j J_{ij} \sum_k \delta(t - t_{j,k} - d_{ij})\,, \tag{2}$$

where J_{ij} is the synaptic weight from presynaptic neuron j to postsynaptic neuron i (with $J_{ij} = 0$ if there is no synapse), the $t_{j,k}$ are the spike times of neuron j, and d_{ij} is a synaptic delay (in this example $d_{ij} = d$ for all pairs of neurons). In total, there are N_E and N_I neurons in the respective populations. Each neuron is connected to a fixed number of randomly chosen presynaptic neurons (fixed in-degree); additionally, all neurons receive external input from independent Poisson processes with rate ν_X. The synaptic weights and in-degrees of recurrent and external connections are population-specific:

$$\boldsymbol{J} = \begin{pmatrix} J_\mathrm{EE} & -J_\mathrm{EI} \\ J_\mathrm{IE} & -J_\mathrm{II} \end{pmatrix},\ \boldsymbol{J}_\mathrm{ext} = \begin{pmatrix} J_\mathrm{EX} \\ J_\mathrm{IX} \end{pmatrix},$$
$$\boldsymbol{K} = \begin{pmatrix} K_\mathrm{EE} & K_\mathrm{EI} \\ K_\mathrm{IE} & K_\mathrm{II} \end{pmatrix},\ \boldsymbol{K}_\mathrm{ext} = \begin{pmatrix} K_\mathrm{EX} \\ K_\mathrm{IX} \end{pmatrix}. \tag{3}$$

All weights are positive, implying an excitatory external input.

The core idea of mean-field theory is to approximate the input to a neuron as Gaussian white noise $\xi(t)$ with mean $\langle \xi(t) \rangle = \mu$ and noise intensity $\langle \xi(t)\xi(t') \rangle = \tau_\mathrm{m}\sigma^2\delta(t-t')$. This approximation is well-suited for asynchronous, irregular network states (van Vreeswijk and Sompolinsky, 1996, 1998; Amit and Brunel, 1997b). For a LIF neuron driven by such Gaussian white noise, the firing rate is given by (Siegert, 1951; Tuckwell, 1988; Amit and Brunel, 1997b)

$$\phi(\mu,\sigma) = \left(\tau_\mathrm{r} + \tau_\mathrm{m}\sqrt{\pi} \int_{\widetilde{V}_0(\mu,\sigma)}^{\widetilde{V}_\mathrm{th}(\mu,\sigma)} e^{s^2}(1+\mathrm{erf}(s))\mathrm{d}s \right)^{-1}, \tag{4}$$

where the rescaled reset- and threshold-voltages are

$$\widetilde{V}_0(\mu,\sigma) = \frac{V_0-\mu}{\sigma}\,, \qquad \widetilde{V}_\mathrm{th}(\mu,\sigma) = \frac{V_\mathrm{th}-\mu}{\sigma}\,. \tag{5}$$

The first term in Equation (4) is the refractory period and the second term is the mean first-passage time of the membrane voltage from reset to threshold. The mean and the noise intensity of the input to a neuron in a population $a \in \{E, I\}$, which control the mean first-passage time through Equation (5), are determined by (Amit and Brunel, 1997b)

$$\mu_a = \tau_\mathrm{m}(J_{a\mathrm{E}}K_{a\mathrm{E}}\nu_\mathrm{E} - J_{a\mathrm{I}}K_{a\mathrm{I}}\nu_\mathrm{I} + J_{a\mathrm{X}}K_{a\mathrm{X}}\nu_\mathrm{X})\,, \tag{6}$$
$$\sigma_a^2 = \tau_\mathrm{m}(J_{a\mathrm{E}}^2K_{a\mathrm{E}}\nu_\mathrm{E} + J_{a\mathrm{I}}^2K_{a\mathrm{I}}\nu_\mathrm{I} + J_{a\mathrm{X}}^2K_{a\mathrm{X}}\nu_\mathrm{X})\,, \tag{7}$$

respectively, where each term reflects the contribution of one population, with the corresponding firing rates of the excitatory ν_E, inhibitory ν_I, and external population ν_X. Note that we use the letters $i, j, k, \ldots$ to index single neurons and $a, b, c, \ldots$ to index neuronal populations. Both μ_a and σ_a depend on the firing rate of the neurons ν_a, which is in turn given by Equation (4). Thus, one arrives at the self-consistency problem

$$\nu_a = \phi(\mu_a, \sigma_a)\,, \tag{8}$$

which is coupled across the populations due to Equation (6) and Equation (7).

Our toolbox provides two algorithms to solve Equation (8): (1) Integrating the auxiliary ordinary differential equation (ODE) $\dot{\nu}_a = -\nu_a + \phi(\mu_a, \sigma_a)$ with initial values $\nu_a(0) = \nu_{a,0}$ using `scipy.integrate.solve_ivp` (Virtanen et al., 2020) until it reaches a fixed point $\dot{\nu}_a = 0$, where Equation (8) holds by construction. (2) Minimizing the quadratic deviation $\sum_a \left[\nu_a - \phi(\mu_a,\sigma_a)\right]^2$, using the least squares (LSTSQ) solver `scipy.optimize.least_squares` (Virtanen et al., 2020) starting from an initial guess $\nu_{a,0}$. The ODE method is robust to changes in the initial values and hence a good first choice. However, it cannot find self-consistent solutions that correspond to an unstable fixed point of the auxiliary ODE (note that the stability of the auxiliary ODE does not indicate the stability of the solution). To this end, the LSTSQ method can be used. Its drawback is that it needs a good initial guess, because otherwise the found minimum might be a local one where the quadratic deviation does not vanish, $\sum_a \left[\nu_a - \phi(\mu_a,\sigma_a)\right]^2 > 0$, and which

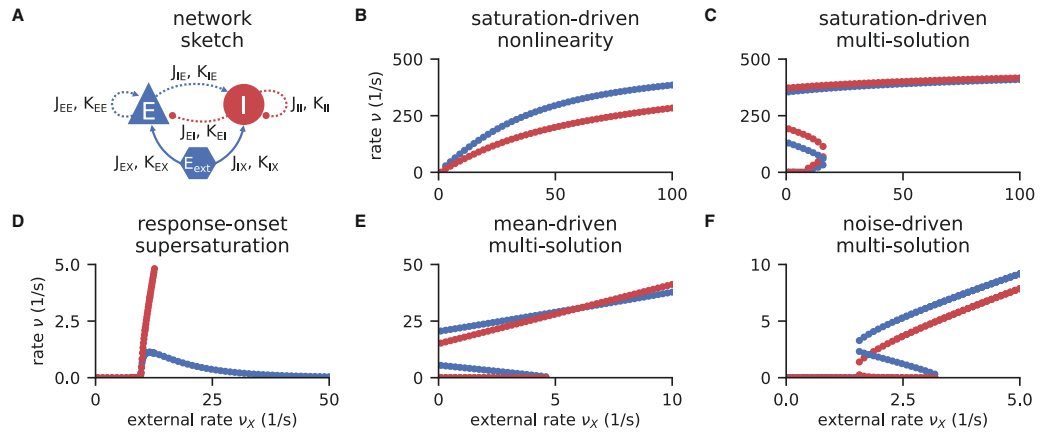

FIGURE 2 | Response nonlinearities in EI-networks. **(A)** Network diagram with nodes and edges according to the graphical notation proposed by Senk et al. (in press). **(B–F)** Firing rate of excitatory (blue) and inhibitory (red) population for varying external input rate ν_X. Specific choices for synaptic weights ($\boldsymbol{J}$, $\boldsymbol{J}_{\text{ext}}$) and in-degrees ($\boldsymbol{K}$, $\boldsymbol{K}_{\text{ext}}$) lead to five types of nonlinearities: **(B)** saturation-driven nonlinearity, **(C)** saturation-driven multi-solution, **(D)** response-onset supersaturation, **(E)** mean-driven multi-solution, and **(F)** noise-driven multi-solution. See Figure 8 in Sanzeni et al. (2020) for parameters.

accordingly does not correspond to a self-consistent solution, $\nu_a \neq \phi(\mu_a, \sigma_a)$. A prerequisite for both methods is a numerical solution of the integral in Equation (4); this is discussed in **Section A.1** in the **Appendix**.

The solutions of the self-consistency problem Equation (8) for varying ν_X and fixed $\boldsymbol{J}$, $\boldsymbol{J}_{\text{ext}}$, $\boldsymbol{K}$, and $\boldsymbol{K}_{\text{ext}}$ reveal the five types of response nonlinearities (**Figure 2**). Different response nonlinearities arise through specific choices of synaptic weights, $\boldsymbol{J}$ and $\boldsymbol{J}_{\text{ext}}$, and in-degrees, $\boldsymbol{K}$ and $\boldsymbol{K}_{\text{ext}}$, which suggests that already a simple EI-network possesses a rich capacity for nonlinear computations. Whenever possible, we use the ODE method and resort to the LSTSQ method only if the self-consistent solution corresponds to an unstable fixed point of the auxiliary ODE. Combining both methods, we can reproduce the first columns of Figure 8 in Sanzeni et al. (2020), where all five types of nonlinearities are presented.

In all cases, we chose appropriate initial values $\nu_{a,0}$ for either method. Note that an exploratory analysis is necessary if the stability properties of a network model are unknown, and potentially multiple fixed points are to be uncovered because there are, to the best of our knowledge, no systematic methods in $d > 1$ dimensions that provide all solutions of a nonlinear system of equations.

In **Listing 2**, we show a minimal example to produce the data shown in **Figure 2B**. After importing the function that solves the self-consistency Equation (8), we collect the neuron and network parameters in a dictionary. Then, we loop through different values for the external rate ν_X and determine the network rates using the ODE method, which is sufficient in this example. In **Listing 2** and to produce **Figure 2B**, we use the basic workflow because only one isolated tool of NNMT (`nnmt.lif.delta._firing_rates()`) is employed, which requires only a few parameters defining the simple EI-network.

```python
import numpy as np
from nnmt.lif.delta import _firing_rates

params = dict(
    # membrane and refractory time constants (in s)
    tau_m=20.*1e-3, tau_r=2.*1e-3,
    # relative reset and threshold potentials (in V)
    V_0_rel=10.*1e-3, V_th_rel=20.*1e-3,
    # recurrent and external weights (in V)
    J=np.array([[0.2, -1.6], [0.2, -1.4]])*1e-3,
    J_ext=np.array([0.2, 0.2])*1e-3,
    # recurrent and external in-degrees
    K=np.array([[400, 100], [400, 100]]),
    K_ext=np.array([1600, 800]),
    # set the method for the fixpoint finder
    fixpoint_method='ODE',
    # initial guess for the firing rate
    nu_0=(0, 0))

# determine self-consistent rates (in 1/s)
nu_ext = np.linspace(1, 100, 50) # external rates (in 1/s)
nu_E, nu_I = np.zeros_like(nu_ext), np.zeros_like(nu_ext)
for i, nu_X in enumerate(nu_ext):
    nu_E[i], nu_I[i] = _firing_rates(nu_ext=nu_X,
                                     **params)
```

Listing 2: Example script to produce the data shown in **Figure 2B** using the ODE method (initial value $\nu_{a,0} = 0$ for population $a \in \{E, I\}$).

3.2.2. Firing Rates of Microcircuit Model

Here we show how to use the model workflow to calculate the firing rates of the cortical microcircuit model by Potjans and Diesmann (2014). The circuit is a simplified point

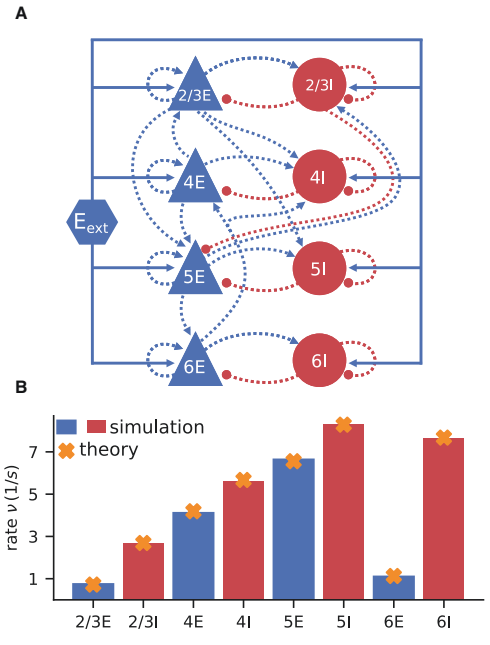

FIGURE 3 | Cortical microcircuit model by Potjans and Diesmann (2014). **(A)** Network diagram (only the strongest connections are shown as in Figure 1 of the original publication). Same notation as in **Figure 2A**. **(B)** Simulation and mean-field estimate for average population firing rates using the parameters from Bos et al. (2016).

neuron network model with biologically plausible parameters, which has been recently used in a number of other works: for example, to study network properties such as layer-dependent attentional processing (Wagatsuma et al., 2011), connectivity structure with respect to oscillations (Bos et al., 2016), and the effect of synaptic weight resolution on activity statistics (Dasbach, Tetzlaff, Diesmann, and Senk, 2021); to assess the performance of different simulator technologies such as neuromorphic hardware (van Albada et al., 2018) and GPUs (Knight and Nowotny, 2018; Golosio et al., 2021); to demonstrate forward-model prediction of local-field potentials from spiking activity (Hagen et al., 2016); and to serve as a building block for large-scale models (Schmidt et al., 2018).

The model consists of eight populations of LIF neurons, corresponding to the excitatory and inhibitory populations of four cortical layers: 2/3E, 2/3I, 4E, 4I, 5E, 5I, 6E, and 6I (see **Figure 3A**). It defines the number of neurons in each population, the number of connections between the populations, the single neuron properties, and the external input. Simulations show that the model yields realistic firing rates for the different populations as observed in particular in the healthy resting-state of early sensory cortex (Potjans and Diesmann, 2014, Table 6).

In contrast to the EI-network model investigated in Section 3.2.1, the neurons in the microcircuit model have exponentially shaped post-synaptic currents: Equation (2) is replaced by Fourcaud and Brunel (2002)

$$\tau_s R \frac{\mathrm{d}I_i}{\mathrm{d}t}(t) = -RI_i(t) + \tau_m \sum_j J_{ij} \sum_k \delta(t - t_{j,k} - d_{ij}), \quad (9)$$

with synaptic time constant τ_s. Note that J_{ij} is a measure in volts here. As discussed in Section 3.2.1, in mean-field theory the second term, representing the neuronal input, is approximated by Gaussian white noise. The additional synaptic filtering leads to the membrane potential (Equation 1) receiving colored noise input. Fourcaud and Brunel (2002) developed a method for calculating the firing rate for this synapse type. They have shown that, if the synaptic time constant τ_s is much smaller than the membrane time constant τ_m, the firing rate for LIF neurons with exponential synapses can be calculated using Equation (4) with shifted integration boundaries

$$\begin{aligned} \widetilde{V}_{\mathrm{cn},0}(\mu,\sigma) &= \widetilde{V}_0(\mu,\sigma) + \frac{\alpha}{2}\sqrt{\frac{\tau_s}{\tau_m}}, \\ \widetilde{V}_{\mathrm{cn,th}}(\mu,\sigma) &= \widetilde{V}_{\mathrm{th}}(\mu,\sigma) + \frac{\alpha}{2}\sqrt{\frac{\tau_s}{\tau_m}}, \end{aligned} \quad (10)$$

with the rescaled reset- and threshold-voltages from Equation (5) and $\alpha = \sqrt{2}|\zeta(1/2)| \approx 2.07$, where $\zeta(x)$ denotes the Riemann zeta function; the subscript cn stands for "colored noise".

The microcircuit has been implemented as an NNMT model (`nnmt.models.Microcircuit`). We here use the parameters of the circuit as published in Bos et al. (2016) which is slightly differently parameterized than the original model (see **Table A1** in the **Appendix**). The parameters of the model are specified in a `yaml` file, which uses Python-like indentation and a dictionary-style syntax. List elements are indicated by hyphens, and arrays can be defined as nested lists. Parameters with units can be defined by using the keys `val` and `unit`, whereas unitless variables can be defined without any keys. **Listing 3** shows an example of how some of the microcircuit network parameters used here are defined. Which parameters need to be provided in the yaml file depends on the model used and is indicated in their respective docstrings.

Once the parameters are defined, a microcircuit model is instantiated by passing the respective parameter file to the model constructor; the units are automatically converted to SI units. Then the firing rates are computed. For comparison, we finally load the simulated rates from Bos et al. (2016):

```
# create the network model using a network parameter yaml
# file
microcircuit = nnmt.models.Microcircuit(
    'network_params_microcircuit.yaml')
# calculate firing rates
firing_rates = nnmt.lif.exp.firing_rates(microcircuit)
# load simulated results
simulated_firing_rates = \
    nnmt.input_output.load_h5('Bos2016_rates.h5')['rates']
```

The simulated rates have been obtained by a numerical network simulation (for simulation details see Bos et al., 2016) in which

```yaml
# membrane time constant
tau_m:
  val: 10.0
  unit: ms

# neuron numbers
N:
  - 20683
  - 5834
  - 21915
```

Listing 3: Some microcircuit network parameters defined in a yaml file. A dictionary-like structure with the keys `val` (value) and `unit` is used to define the membrane time constant, which is the same across all populations. The numbers of neurons in each population are defined as a list. Only the numbers for the first three populations are displayed.

the neuron populations are connected according to the model's original connectivity rule: "random, fixed total number with multapses (autapses prohibited)", see Senk et al. (in press) as a reference for connectivity concepts. The term *multapses* refers to multiple connections between the same pair of neurons and *autapses* are self-connections; with this connectivity rule multapses can occur in a network realization but autapses are not allowed. For simplicity, the theoretical predictions assume a connectivity with a fixed in-degree for each neuron. Dasbach et al. (2021) show that simulated spike activity data of networks with these two different connectivity rules are characterized by differently shaped rate distributions ("reference" in their Figures 3d and 4d). In addition, the weights in the simulation are normally distributed while the theory replaces each distribution by its mean; this corresponds to the case $N_{\text{bins}} = 1$ in Dasbach et al. (2021). Nevertheless, our mean-field theoretical estimate of the average population firing rates is in good agreement with the simulated rates (**Figure 3B**).

3.3. Dynamical Quantities

3.3.1. Transfer Function

One of the most important dynamical properties of a neuronal network is how it reacts to external input. A systematic way to study the network response is to apply an oscillatory external input current leading to a periodically modulated mean input $\mu(t) = \mu + \delta\mu \,\mathrm{Re}\left(\mathrm{e}^{\mathrm{i}\omega t}\right)$ (cf. Equation 6), with fixed frequency ω, phase, and amplitude $\delta\mu$, and observe the emerging frequency, phase, and amplitude of the output. If the amplitude of the external input is small compared to the stationary input, the network responds in a linear fashion: it only modifies phase and amplitude, while the output frequency equals the input frequency. This relationship is captured by the input-output transfer function $N(\omega)$ (Brunel and Hakim, 1999; Brunel et al., 2001; Lindner and Schimansky-Geier, 2001), which describes the frequency-dependent modulation of the output firing rate of a neuron population

$$\nu(t) = \nu + \mathrm{Re}\Big(N(\omega)\,\delta\mu\,\mathrm{e}^{\mathrm{i}\omega t}\Big).$$

Note that in this section we only study the linear response to a modulation of the mean input, although in general, a modulation of the noise intensity (Equation 7) can also be included (Lindner and Schimansky-Geier, 2001; Schuecker et al., 2015). The transfer function $N(\omega)$ is a complex function: Its absolute value describes the relative modulation of the firing rate. Its phase, the angle relative to the real axis, describes the phase shift that occurs between input and output. We denote the transfer function for a network of LIF neurons with instantaneous synapses in linear-response approximation as

$$N(\omega) = \frac{\sqrt{2}\nu}{\sigma} \frac{1}{1+\mathrm{i}\omega\tau_{\mathrm{m}}} \frac{\Phi'_\omega\big|_{\sqrt{2}\widetilde{V}_0}^{\sqrt{2}\widetilde{V}_{\mathrm{th}}}}{\Phi_\omega\big|_{\sqrt{2}\widetilde{V}_0}^{\sqrt{2}\widetilde{V}_{\mathrm{th}}}}, \tag{11}$$

with the rescaled reset- and threshold-voltages $\widetilde{V}_0$ and $\widetilde{V}_{\mathrm{th}}$ as defined in Equation (5) and $\Phi_\omega(x) = e^{\frac{x^2}{4}}\, U\left(\mathrm{i}\omega\tau_{\mathrm{m}} - \frac{1}{2}, x\right)$ using the parabolic cylinder functions $U\left(\mathrm{i}\omega\tau_{\mathrm{m}} - \frac{1}{2}, x\right)$ as defined in (Abramowitz and Stegun, 1974, Section 19.3) and (Olver et al., 2021, Section 12.2). Φ'_ω denotes the first derivative by x. A comparison of our notation and the transfer function given in Schuecker et al. (2015, Equation 29) can be found in **Section A.2.1** in the **Appendix**.

For a neuronal network of LIF neurons with exponentially shaped post-synaptic currents, introduced in Section 3.2.2, Schuecker et al. (2014, 2015) show that an analytical approximation of the transfer function can be obtained by a shift of integration boundaries, akin to Equation (10):

$$N_{\mathrm{cn}}(\omega) = \frac{\sqrt{2}\nu}{\sigma} \frac{1}{1+\mathrm{i}\omega\tau_{\mathrm{m}}} \frac{\Phi'_\omega\big|_{\sqrt{2}\widetilde{V}_{\mathrm{cn,0}}}^{\sqrt{2}\widetilde{V}_{\mathrm{cn,th}}}}{\Phi_\omega\big|_{\sqrt{2}\widetilde{V}_{\mathrm{cn,0}}}^{\sqrt{2}\widetilde{V}_{\mathrm{cn,th}}}}. \tag{12}$$

To take into account the effect of the synaptic dynamics, we include an additional low-pass filter:

$$N_{\mathrm{cn,s}}(\omega) = N_{\mathrm{cn}}(\omega) \frac{1}{1+\mathrm{i}\omega\tau_{\mathrm{s}}}. \tag{13}$$

If the synaptic time constant is much smaller than the membrane time constant ($\tau_{\mathrm{s}} \ll \tau_{\mathrm{m}}$), an equivalent expression for the transfer function is obtained by a Taylor expansion around the original boundaries (cf. Schuecker et al. 2015, Equation 30). The toolbox implements both variants and offers choosing between them by setting the argument `method` of `nnmt.lif.exp.transfer_function` to either `shift` or `taylor`.

Here, we demonstrate how to calculate the analytical "shift version" of the transfer function for different means and noise intensities of the input current (see **Figure 4**) and thereby reproduce Figure 4 in Schuecker et al. (2015).

The crucial parts for producing **Figure 4** using NNMT are shown in **Listing 4** for one example combination of mean and noise intensity of the input current. Instead of using the model workflow with `nnmt.lif.exp.transfer_function`, we here employ the basic workflow, using

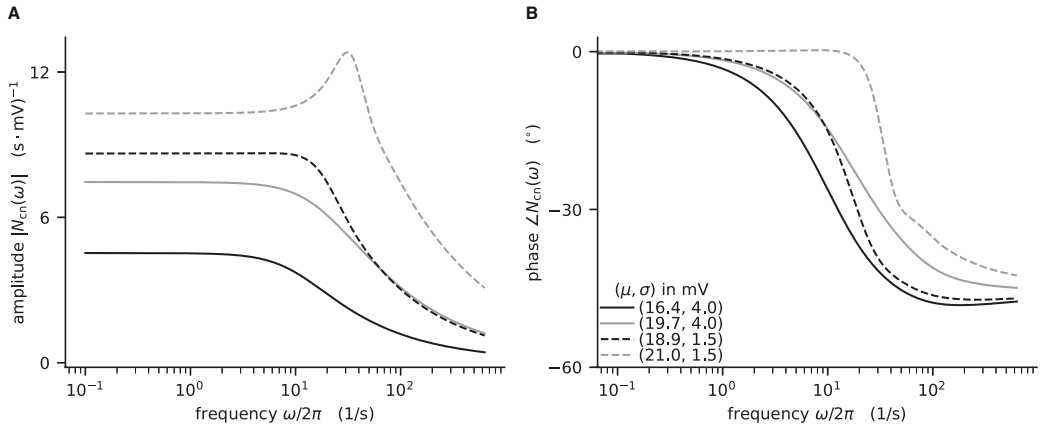

FIGURE 4 | Colored-noise transfer function N_{cn} of LIF model in different regimes. **(A)** Absolute value and **(B)** phase of the "shift" version of the transfer function as a function of the log-scaled frequency. Neuron parameters are set to $V_{th} = 20$ mV, $V_0 = 15$ mv, $\tau_m = 20$ ms, and $\tau_s = 0.5$ ms. For given noise intensities of input current, $\sigma = 4$ mV (solid line) and $\sigma = 1.5$ mV (dashed line), the mean input μ is chosen such that firing rates $\nu = 10$ Hz (black) and $\nu = 30$ Hz (gray) are obtained.

`nnmt.lif.exp._transfer_function` directly. This allows changing the mean input and its noise intensity independently of a network model's structure, but requires two additional steps: First, the necessary parameters are loaded from a `yaml` file, converted to SI units and then stripped off the units using the utility function `nnmt.utils._convert_to_si_and_strip_units`. Second, the analysis frequencies are defined manually. In this example we choose logarithmically spaced frequencies, as we want to plot the results on a log-scale. Finally, the complex-valued transfer function is calculated and then split into its absolute value and phase. **Figure 4** shows that the transfer function acts as a low-pass filter that suppresses the amplitude of high frequency activity, introduces a phase lag, and can lead to resonance phenomena for certain configurations of mean input current and noise intensity.

The replication of the results from Schuecker et al. (2015) outlined here is also used in the integration tests of the toolbox. Note that the implemented analytical form of the transfer function by Schuecker et al. (2015) is an approximation for low frequencies, and deviations from a simulated ground truth are expected for higher frequencies ($\omega/2\pi \gtrsim 100$ Hz at the given parameters).

3.3.2. Power Spectrum

Another frequently studied dynamical property is the power spectrum, which describes how the power of a signal is distributed across its different frequency components, revealing oscillations of the population activity. The power is the Fourier transformed auto-correlation of the population activities (c.f. Bos et al. 2016, Equations 16-18). Linear response theory on top of a mean-field approximation, allows computing the power, dependent on the network architecture, the stationary firing rates, and the neurons' transfer function (Bos et al., 2016). The corresponding analytical expression for the power spectra of population a at angular frequency ω is given by the diagonal elements of the correlation matrix

$$\begin{aligned} P_a(\omega) &= C_{aa}(\omega) \\ &= \left[\left(\mathbf{1} - \widetilde{\mathbf{M}}_{\mathrm{d}}(\omega)\right)^{-1} \operatorname{diag}\left(\boldsymbol{\nu} \oslash \boldsymbol{n}\right) \left(\mathbf{1} - \widetilde{\mathbf{M}}_{\mathrm{d}}(-\omega)\right)^{-\mathrm{T}} \right]_{aa}, \end{aligned} \tag{14}$$

with $\oslash$ denoting the elementwise (Hadamard) division, the effective connectivity matrix $\widetilde{\boldsymbol{M}}_{\mathrm{d}}(\omega) = \tau_{\mathrm{m}} \boldsymbol{N}_{\mathrm{cn,s}}(\omega) \cdot \boldsymbol{J} \odot \boldsymbol{K} \odot \boldsymbol{D}(\omega)$, where the dot denotes the scalar product, while $\odot$ denotes the elementwise (Hadamard) product, the mean population firing rates $\boldsymbol{\nu}$, and the numbers of neurons in each population $\boldsymbol{n}$. The effective connectivity combines the static, anatomical connectivity $\boldsymbol{J} \odot \boldsymbol{K}$, represented by synaptic weight matrix $\boldsymbol{J}$ and in-degree matrix $\boldsymbol{K}$, and dynamical quantities, represented by the transfer functions $N_{\mathrm{cn,s},a}(\omega)$ (Equation (13)), and the contribution of the delays in (Equation 13), represented by their Fourier transformed distributions $D_{ab}(\omega)$ (cf. Bos et al. 2016, Equations 14, 15).

The modular structure in combination with the model workflow of this toolbox permits a step-by-step calculation of the power spectra, as shown in **Listing 5**. The inherent structure of the theory is emphasized in these steps: After instantiating the network model class with given network parameters, we determine the working point, which characterizes the statistics of the model's stationary dynamics. It is defined by the population firing rates, the mean, and the standard deviation of the input to a neuron of the respective population. This is necessary for determining the transfer functions. The calculation of the delay distribution matrix is then required for calculating the effective connectivity and to finally get an estimate of the power spectra.

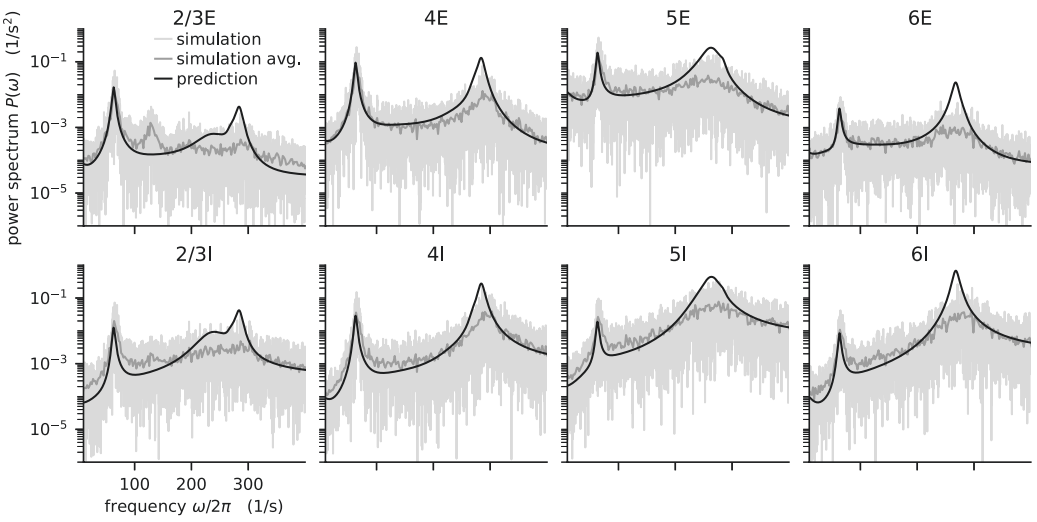

FIGURE 5 | Power spectra of the population spiking activity in the adapted cortical microcircuit from Bos et al. (2016). The spiking activity of each population in a 10 s simulation of the model is binned with 1 ms resolution and the power spectrum of the resulting histogram is calculated by a fast Fourier transform (FFT; light gray curves). In addition, the simulation is split into 500 ms windows, the power spectrum calculated for each window and averaged across windows (gray curves). Black curves correspond to analytical prediction obtained with NNMT as described in **Listing 5**. The panels show the spectra for the excitatory (top) and inhibitory (bottom) populations within each layer of the microcircuit.

Figure 5 reproduces Figure 1E in Bos et al. (2016) and shows the spectra for each population of the adjusted version (see **Table A1** in the **Appendix**) of the microcircuit model.

The numerical predictions obtained from the toolbox largely coincide with simulated data taken from the original publication (Bos et al., 2016) and reveal dominant oscillations of the population activities in the low-γ range around 63 Hz. Furthermore, faster oscillations with peak power around 300 Hz are predicted with higher magnitudes in the inhibitory populations 4I, 5I, and 6I.

The deviation between predicted and simulated power spectra seen at $\sim$ 130 Hz in population 2/3E could be a harmonic of the correctly predicted, prominent 63 Hz peak; a non-linear effect not captured in linear response theory. Furthermore, the systematic overestimation of the power spectrum at large frequencies is explained by the limited validity of the analytical approximation of the transfer function for high frequencies.

3.3.3. Sensitivity Measure

The power spectra shown in the previous section exhibit prominent peaks at certain frequencies, which indicate oscillatory activity. Naturally, this begs the question: which mechanism causes these oscillations? Bos et al. (2016) expose the crucial role that the microcircuit's connectivity plays in shaping the power spectra of this network model. They have developed a method called *sensitivity measure* to directly relate the influence of the anatomical connections, especially the in-degree matrix, on the power spectra.

The power spectrum of the a-th population $P_a(\omega)$ receives a contribution from each eigenvalue λ_b of the effective connectivity matrix, $P_a(\omega) \propto 1/\left(1-\lambda_b(\omega)\right)^2$. Such a contribution consequently diverges as the complex-valued λ_b approaches $1 + 0\mathrm{i}$ in the complex plane, which is referred to as the point of instability. This relation can be derived by replacing the effective connectivity matrix $\widetilde{\mathbf{M}}_\mathrm{d}(\omega)$ in Equation (14) by its eigendecomposition. The sensitivity measure leverages this relationship and evaluates how a change in the in-degree matrix affects the eigenvalues of the effective connectivity and thus indirectly the power spectrum. Bos et al. (2016) introduce a small perturbation α_{cd} of the in-degree matrix, which allows writing the effective connectivity matrix as $\widehat{M}_{ab}(\omega) = (1+\alpha_{cd}\delta_{ca}\delta_{db})\,\widetilde{M}_{ab}(\omega)$, where we dropped the delay subscript d. The sensitivity measure $Z_{b,cd}(\omega)$ describes how the b-th eigenvalue of the effective connectivity matrix varies when the cd-th element of the in-degree matrix is changed

$$Z_{b,cd}(\omega) = \left.\frac{\partial\lambda_b(\omega)}{\partial\alpha_{cd}}\right|_{\alpha_{cd}=0} = \frac{v_{b,c}\widetilde{M}_{cd}u_{b,d}}{\mathbf{v}_b^\mathrm{T}\cdot\mathbf{u}_b}\,, \tag{15}$$

where $\frac{\partial\lambda_b(\omega)}{\partial\alpha_{cd}}$ is the partial derivative of the eigenvalue with respect to a change in connectivity, $\mathbf{v}_b^\mathrm{T}$ and $\mathbf{u}_b$ are the left and right eigenvectors of $\widetilde{\boldsymbol{M}}$ corresponding to eigenvalue $\lambda_b(\omega)$.

The complex sensitivity measure can be understood in terms of two components: $\mathbf{Z}_b^\mathrm{amp}$ is the projection of the matrix $\mathbf{Z}_b$ onto the direction in the complex plane defined by $1-\lambda_b(\omega)$;

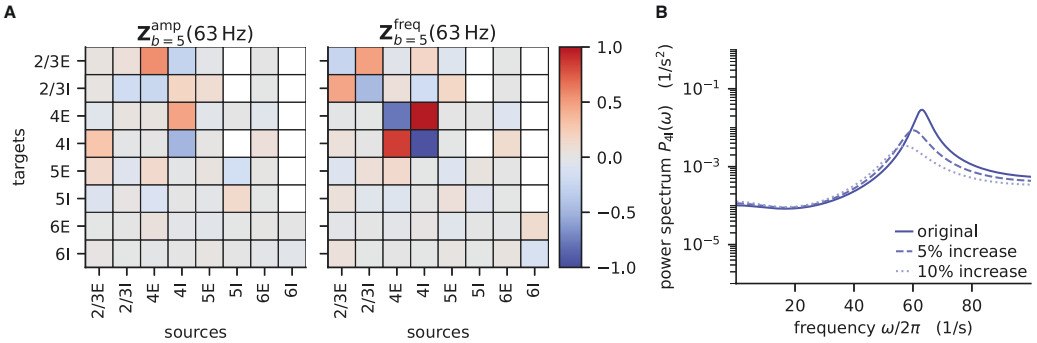

FIGURE 6 | Sensitivity measure at low-γ frequency and corresponding power spectrum of microcircuit with adjusted connectivity. **(A)** Sensitivity measure of one eigenmode of the effective connectivity relevant for low-γ oscillations. The sensitivity measure for this mode is evaluated at the frequency where the corresponding eigenvalue is closest to the point of instability $1 + 0$i in complex plane. $\boldsymbol{Z}^{\text{amp}}_b(\omega)$ (left subpanel) visualizes the influence of a perturbation of a connection on the peak amplitude of the power spectrum. $\boldsymbol{Z}^{\text{freq}}_b(\omega)$ (right subpanel) shows the impact on the peak frequency. Non-existent connections are masked white. **(B)** Mean-field prediction of power spectrum of population 4I with original connectivity parameters (solid line), 5% increase (dashed line) and 10% increase (dotted line) in connections $K_{4\text{I}\to 4\text{I}}$. The increase in inhibitory input to population 4I was counteracted by an increase of the excitatory external input $K_{\text{ext}\to 4\text{I}}$ to maintain the working point.

it describes how, when the in-degree matrix is perturbed, the complex-valued $\lambda_b(\omega)$ moves toward or away from the instability $1 + 0$i, and consequently how the amplitude of the power spectrum at frequency ω increases or decreases. $\mathbf{Z}^{\text{freq}}_b$ is the projection onto the perpendicular direction and thus describes how the peak frequency of the power spectrum changes with the perturbation of the in-degree matrix. For a visualization of these projections, refer to Figure 5B in Bos et al. (2016).

The toolbox makes this intricate measure accessible by supplying two tools: After computing the required working point, transfer function, and delay distribution, the tool `nnmt.lif.exp.sensitivity_measure` computes the sensitivity measure at a given frequency for one specific eigenvalue. By default, this is the eigenvalue which is closest to the instability $1 + 0$i. To perform the computation, we just need to add one line to **Listing 5**:

```
sensitivity_dict = nnmt.lif.exp.sensitivity_measure(
    microcircuit, frequency)
```

The result is returned in form of a dictionary that contains the sensitivity measure and its projections. The tool `nnmt.lif.exp.sensitivity_measure_all_eigenmodes` wraps that basic function and calculates the sensitivity measure for all eigenvalues at the frequency for which each eigenvalue is closest to instability.

According to the original publication (Bos et al., 2016), the peak around 63 Hz has contributions from one eigenvalue of the effective connectivity matrix. **Figure 6** shows the projections of the sensitivity measure at the frequency for which this eigenvalue is closest to the instability, as illustrated in Figure 4 of Bos et al. (2016). The sensitivity measure returns one value for each connection between populations in the network model. For $\mathbf{Z}^{\text{amp}}_b$ a negative value indicates that increasing the in-degrees of a specific connection causes the amplitude of the power spectrum at the evaluated frequency to drop. If the value is positive, the amplitude is predicted to grow as the in-degrees increase. Similarly, for positive $\mathbf{Z}^{\text{freq}}_b$ the frequency of the peak in the power spectrum shifts toward higher values as in-degrees increase, and vice versa. The main finding in this analysis is that the low-γ peak seems to be affected by excitatory-inhibitory loops in layer 2/3 and layer 4.

To decrease the low–γ peak in the power spectrum, one could therefore increase the 4I to 4I connections (cp. **Figure 6A**):

```
# 5 percent increase
K_new = microcircuit.network_params['K'].copy()
K_new[3,3] = 1001 # originally 953
K_ext_new = microcircuit.network_params['K_ext'].copy()
K_ext_new[3] = 2034 # originally 1900
microcircuit_new = microcircuit.change_parameters(
    {'K': K_new, 'K_ext': K_ext_new})
```

and calculate the power spectrum as in **Listing 5** again to validate the change. Note that a change in connectivity leads to a shift in the working point. We are interested in the impact of the modified connectivity on the fluctuation dynamics at the same working point and thus need to counteract the change in connectivity by adjusting the external input. In the chosen example this is ensured by satisfying $J_{4\text{I}\to 4\text{I}}\Delta K_{4\text{I}\to 4\text{I}}\nu_{4\text{I}} = -J_{\text{ext}\to 4\text{I}}\Delta K_{\text{ext}\to 4\text{I}}\nu_{\text{ext}}$, which yields $\Delta K_{\text{ext}\to 4\text{I}} = -\frac{J_{4\text{I}\to 4\text{I}}\Delta K_{4\text{I}\to 4\text{I}}\nu_{4\text{I}}}{J_{\text{ext}\to 4\text{I}}\nu_{\text{ext}}}$.

If several eigenvalues of the effective connectivity matrix influence the power spectra in the same frequency range, adjustments of the connectivity are more involved. This is because a change in connectivity would inevitably affect all eigenvalues simultaneously. Further care has to be taken because the sensitivity measure is subject to the same constraints as the current implementation of the transfer function, which is only valid for low frequencies and enters the sensitivity measure directly.

```python
# load parameters in custom units
params = nnmt.input_output.load_val_unit_dict_from_yaml(
    'Schuecker2015_parameters.yaml')

# convert parameters to SI units
nnmt.utils._convert_to_si_and_strip_units(params)

# define the analysis frequencies
frequencies = np.logspace(
    params['f_start_exponent'],
    params['f_end_exponent'],
    params['n_freqs'])
# add the zero frequency
frequencies = np.insert(frequencies, 0, 0.0)
omegas = 2 * np.pi * frequencies

# extract necessary parameters from params dictionary
mean_input = params['mean_input']
...  # here we leave out similar statements

# calculate the transfer function
transfer_function = nnmt.lif.exp._transfer_function(
    mu, sigma,
    tau_m, tau_s, tau_r,
    V_th_rel, V_0_rel,
    omegas,
    method='shift',
    synaptic_filter=False)

# calculate properties plotted in Schuecker et al. (2015)
absolute_value = np.abs(transfer_function)
phase = np.angle(transfer_function) / 2 / np.pi * 360
```

Listing 4: Example script for computing a transfer function shown in **Figure 4** using the method of shifted integration boundaries.

```python
# create network model microcircuit
microcircuit = nnmt.models.Microcircuit(
    network_params='Bos2016_network_params.yaml',
    analysis_params='Bos2016_analysis_params.yaml')

# calculate working point for exponentially shaped post-
    synaptic currents
nnmt.lif.exp.working_point(microcircuit, method='taylor')
# calculate the transfer function
nnmt.lif.exp.transfer_function(microcircuit,
    method='taylor')
# calculate the delay distribution matrix
nnmt.network_properties.delay_dist_matrix(microcircuit)
# calculate the effective connectivity matrix
nnmt.lif.exp.effective_connectivity(microcircuit)
# calculate the power spectra
power_spectra = nnmt.lif.exp.power_spectra(microcircuit)
```

Listing 5: Example script to produce the theoretical prediction (black lines) shown in **Figure 5B**.

3.4. Fitting Spiking to Rate Model and Predicting Pattern Formation

If the neurons of a network are spatially organized and connected according to a distance-dependent profile, the spiking activity may exhibit pattern formation in space and time, including wave-like phenomena. Senk et al. (2020) set out to scrutinize the non-trivial relationship between the parameters of such a network model and the emerging activity patterns. The model they use is a two-population network of excitatory E and inhibitory I spiking neurons, illustrated in **Figure 7**. All neurons are of type LIF with exponentially shaped post-synaptic currents. The neuron populations are recurrently connected to each other and themselves and they receive additional external excitatory E_{ext} and inhibitory I_{ext} Poisson spike input of adjustable rate as shown in **Figure 7A**. The spatial arrangement of neurons on a ring is illustrated in **Figure 7B** and the boxcar-shaped connectivity profiles in **Figure 7C**.

In the following, we consider a mean-field approximation of the spiking model with spatial averaging, that is a time and space continuous approximation of the discrete model as derived in Senk et al. (2020, Section E. Linearization of spiking network model). We demonstrate three methods used in the original study: First, Section 3.4.1 explains how a model can be brought to a defined state characterized by its working point. The working point is given by the mean μ and noise intensity σ of the input to a neuron, which are both quantities derived from network parameters and require the calculation of the firing rates. With the spiking model in that defined state, Section 3.4.2 then maps its transfer function to the one of a rate model. Section 3.4.3 finally shows that this working-point dependent rate model allows for an analytical linear stability analysis of the network accounting for its spatial structure. This analysis can reveal transitions to spatial and temporal oscillatory states which, when mapped back

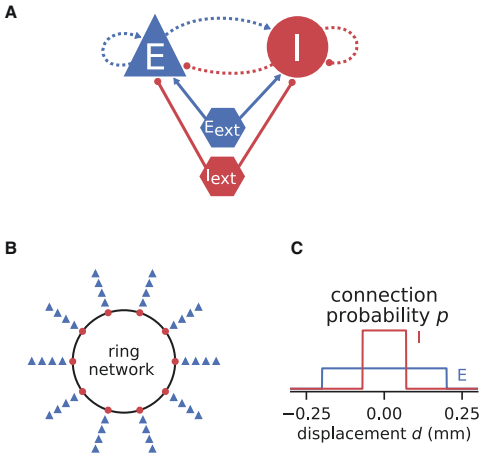

FIGURE 7 | Illustrations of spiking network model by Senk et al. (2020). **(A)** Excitatory and inhibitory neuronal populations randomly connected with fixed in-degree and multapses allowed (autapses prohibited). External excitatory and inhibitory Poisson drive to all neurons. Same notation as in **Figure 2A**. **(B)** One inhibitory and four excitatory neurons per grid position on a one-dimensional domain with periodic boundary conditions (ring network). **(C)** Normalized, boxcar-shaped connection probability with wider excitation than inhibition; the grid spacing is here 10^{-3} mm. For model details and parameters, see Tables II–IV of Senk et al. (2020); the specific values given in the caption of their Figure 6 are used throughout here.

to the parameters of the spiking model, may manifest in distinct patterns of simulated spiking activity after a startup transient.

3.4.1. Setting the Working Point by Changing Network Parameters

With network and analysis parameters predefined in `yaml` files, we set up a network model using the example model class `Basic`:

```
space_model = nnmt.models.Basic(
    network_params='Senk2020_network_params.yaml',
    analysis_params='Senk2020_analysis_params.yaml')
```

Upon initialization the given parameters are automatically converted into the format used by NNMT's tools. For instance, relative spike reset and threshold potentials are derived from the absolute values, connection strengths in units of volt are computed from the post-synaptic current amplitudes in ampere, and all values are scaled to SI units.

We aim to bring the network to a defined state by fixing the working point but also want to explore if the procedure of fitting the transfer function still works for different network states. For a parameter space exploration, we use a method to change parameters provided by the model class and scan through a number of different working points of the network. To obtain the required input for a target working point, we adjust the external excitatory and inhibitory firing rates accordingly; NNMT uses a vectorized version of the equations given in Senk et al. (2020, Appendix F: Fixing the working point) to calculate the external rates needed:

```
# relative to spike threshold (in V)
mu = 10. * 1e-3; sigma = 10. * 1e-3
nu_ext = nnmt.lif.exp.external_rates_for_fixed_input(
    space_model, mu_set=mu, sigma_set=sigma)
space_model = space_model.change_parameters(
    changed_network_params={'nu_ext': nu_ext})
```

The implementation uses only one excitatory and one inhibitory Poisson source to represent the external input rates which typically originate from a large number of external source neurons. These two external sources are connected to the network with the same relative inhibition g as used for the internal connections. The resulting external rates for different choices of (μ, σ) are color-coded in the first two plots of **Figure 8A**. The third plot shows the corresponding firing rates of the neurons, which are stored in the results of the model instance when computing the working point explicitly:

```
nnmt.lif.exp.working_point(space_model)
```

Although the external rates are substantially higher than the firing rates, since a neuron is recurrently connected to hundreds of neurons, the total external and recurrent inputs are of the same order.

3.4.2. Parameter Mapping by Fitting the Transfer Function

We map the parameters of the spiking model to a corresponding rate model by, first, computing the transfer function $\boldsymbol{N}_{\mathrm{cn,s}}$ given in Equation (13) of the spiking model, and second, fitting the simpler transfer function of the rate model, for details see Senk et al. (2020, Section F. Comparison of neural-field and spiking models). The dynamics of the rate model can be written as a differential equation for the linearized activity r_a with populations $a, b \in \{\mathrm{E}, \mathrm{I}\}$:

$$\tau \frac{\mathrm{d}}{\mathrm{d}t} r_a(t) = -r_a(t) + \sum_b w_b r_b(t-d) \tag{16}$$

with the delay d; τ is the time constant and w_b are the unitless weights that only depend on the presynaptic population. The transfer function is just the one of a low-pass filter, $N_{\mathrm{LP}} = 1/(1+\lambda\tau)$, where λ is the frequency in Laplace domain. The tool to fit the transfer function requires that the actual transfer function $\boldsymbol{N}_{\mathrm{cn,s}}$ has been computed beforehand and fits $N_{\mathrm{LP}}\boldsymbol{w}$ to $\tau_{\mathrm{m}}\boldsymbol{N}_{\mathrm{cn,s}} \cdot \boldsymbol{J} \odot \boldsymbol{K}$ for the same frequencies together with τ, $\boldsymbol{w}$, and the combined fit error η:

```
nnmt.lif.exp.transfer_function(space_model)
nnmt.lif.exp.fit_transfer_function(space_model)
```

The absolute value of the transfer function is fitted with non-linear least-squares using the solver `scipy.optimize.curve_fit`. **Figure 8B** illustrates the amplitude and phase of the transfer function and its fit for a few (μ, σ) combinations. The plots of **Figure 8C** show the fitted time constants, the fitted excitatory weight, and the combined fit error. The inhibitory weight is proportional to the excitatory one in the same way as the post-synaptic current amplitudes.

3.4.3. Linear Stability Analysis of Spatially Structured Model With Delay

Sections 3.4.1 and 3.4.2 considered a mean-field approximation of the spiking model without taking space into account. In the following, we assume a spatial averaging of the discrete network depicted in **Figure 7** and introduce the spatial connectivity profiles $p_a(x)$. Changing Equation (16) to the integro-differential equation

$$\tau \frac{\partial}{\partial t} r_a(x,t) = -r_a(x,t) + \sum_b w_b \int_{-\infty}^{\infty} p_b(x-y) r_b(y, t-d)\,\mathrm{d}y \tag{17}$$

yields a neural field model defined in continuous space x. This model lends itself to analytical linear stability analysis, as described in detail in Senk et al. (2020, Section A. Linear stability analysis of a neural-field model). In brief, we analyze the stability of a fixed-point solution to this differential equation system which, together with parameter continuation methods and bifurcation analysis, allows us to determine points in parameter space where transitions from homogeneous steady states to oscillatory states can occur. These transitions are given as a function of a bifurcation parameter, here the constant delay d, which is the same for all connections. The complex-valued, temporal eigenvalue λ of the linearized time-delay system is an indicator for the system's overall stability and can serve as a predictor for temporal oscillations, whereas the spatial oscillations are characterized by the real-valued wave number k. Solutions that relate λ and k with the model parameters are given by a characteristic equation, which in our case reads (Senk et al., 2020, Equation 7):

$$\lambda_B(k) = -\frac{1}{\tau} + \frac{1}{d} W_B\left(c(k)\frac{d}{\tau}\mathrm{e}^{\frac{d}{\tau}}\right), \tag{18}$$

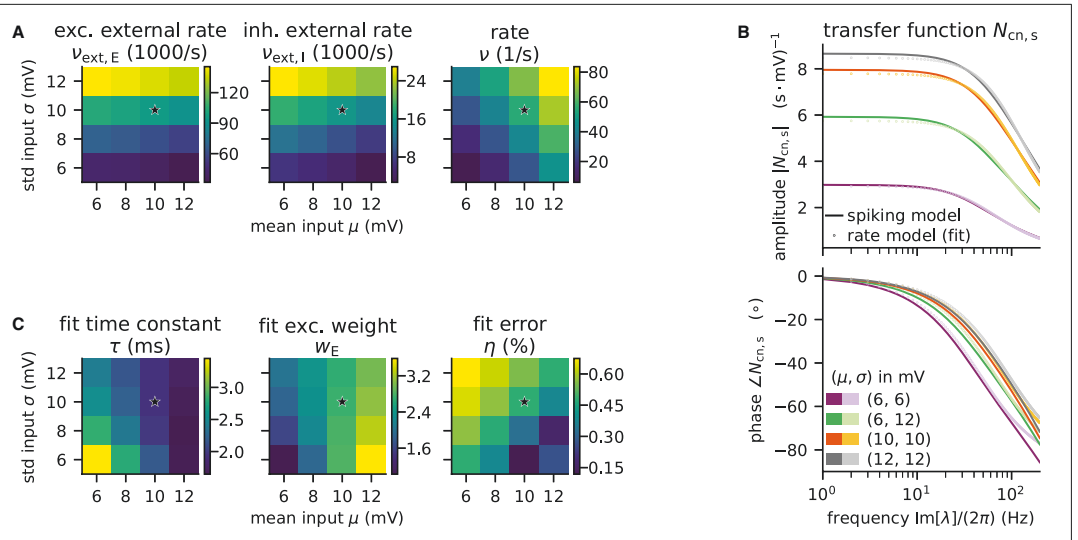

FIGURE 8 | Network parameters and mean-field results from scanning through different working points. Working point (μ, σ) combines mean input μ and noise intensity of input σ. **(A)** External excitatory $\nu_{\text{ext,E}}$ and inhibitory $\nu_{\text{ext,I}}$ Poisson rates required to set (μ, σ) and resulting firing rates ν. **(B)** Transfer function $N_{\text{cn,s}}$ of spiking model and fitted rate-model approximation with low-pass filter for selected (μ, σ) (top: amplitude, bottom: phase). **(C)** Fit results (time constants τ and excitatory weights w_{E}) and fit errors η. The inhibitory weights are $w_{\text{I}} = -gw_{\text{E}}$ with $g = 5$. Star marker in panels **(A)** and **(C)** denotes target working point (10, 10) mV. Similar displays as in Senk et al. (2020, Figure 5).

with the time constant of the rate model τ, the multi-valued Lambert W_B function[3] on branch B (Corless et al., 1996), and the effective connectivity profile $c(k)$, which combines the weights w_b and the Fourier transforms of the spatial connectivity profiles. Note that the approach generalizes from the boxcar-shaped profiles used here to any symmetric probability density function. NNMT provides an implementation to solve this characteristic equation in its `linear_stability` module using the `spatial` module for the profile:

```
import nnmt.spatial as spatial
import nnmt.linear_stability as linstab

connectivity = (
    W_rate * spatial._ft_spatial_profile_boxcar(
        k_wavenumber,
        space_model.network_params['width']))

eigenvalue = (
    linstab._solve_chareq_lambertw_constant_delay(
        branch_nr, tau_rate,
        space_model.network_params['delay'],
        connectivity))
```

Figure 9A shows that the computed eigenvalues come for the given network parameters in complex conjugate pairs. The branch with the largest real part is the principal branch ($B = 0$). Temporal oscillations are expected to occur if the real part of the eigenvalue on the principal branch becomes positive; the oscillation frequency can then be read off the imaginary part of that eigenvalue. In this example, the largest eigenvalue λ^* on the principal branch has a real part that is just above zero. There exists a supercritical Hopf bifurcation and the delay as the bifurcation parameter is chosen large enough such that the model is just beyond the bifurcation point separating the stable from the instable state. The respective wave number k^* is positive, which indicates spatial oscillations as well. The oscillations in both time and space predicted for the rate model imply that the activity of the corresponding spiking model might exhibit wave trains, i.e., temporally and spatially periodic patterns. The predicted propagation speed of the wave trains is given by the phase velocity $\text{Im}[\lambda^*]/k^*$.

To determine whether the results obtained with the rate model are transferable to the spiking model, **Figure 9B** interpolates the analytical solutions of the rate model [$\alpha = 0$, evaluating Equation (18)] to solutions of the spiking model ($\alpha = 1$, accounting for the transfer function $\boldsymbol{N}_{\text{cn,s}}$), which can only be computed numerically. Thus, the parameter α interpolates between the characteristic equations of these two models which primarily differ in their transfer function; for details see Senk et al. (2020, Section F.2 Linear interpolation between the transfer functions). Since the eigenvalues estimated this way show only little differences between rate and spiking model, we conclude that predictions from the rate model should hold also for the spiking model in the parameter regime tested. Following the

[3]The Lambert W_B function is defined as $z = W_B(z)\,e^{W_B(z)}$ for $z \in \mathbb{C}$ and has infinitely many solutions, numbered by the branches B.

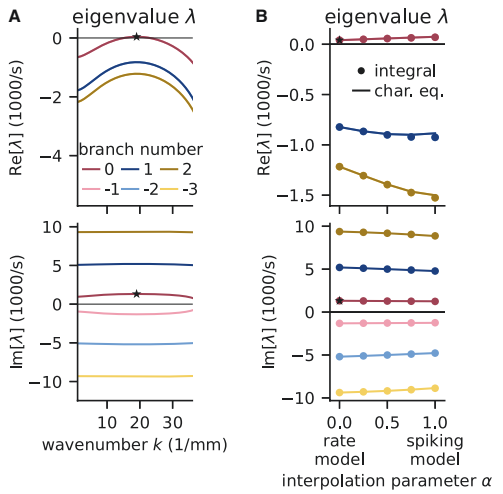

FIGURE 9 | Linear stability analysis of spatially structured network model. **(A)** Analytically exact solution for real (top) and imaginary (bottom) part of eigenvalue λ vs. wavenumber k using rate model derived by fit of spiking model at working point $(\mu, \sigma) = (10, 10)$ mV. Color-coded branches of Lambert W_B function; maximum real eigenvalue (star marker) on principal branch ($B = 0$). **(B)** Linear interpolation between rate ($\alpha = 0$) and spiking model ($\alpha = 1$) by numerical integration of Senk et al. (2020, Equation 30) (solid line) and by numerically solving the characteristic equation in Senk et al. (2020, Equation 29) (circular markers). Star markers at same data points as in **(A)**. Similar displays as in Senk et al. (2020, Figure 6).

argument of Senk et al. (2020), the predicted pattern formation could next be tested in a numerical simulation of the discrete spiking network model. Their Figure 7c for the delay $d = 1.5$ ms shows such results with the same parameters as used here; this figure also illustrates transitions from homogeneous states to oscillatory states by changing the delay (panels b, c, and e).

4. DISCUSSION

Mean-field theory grants important insights into the dynamics of neuronal networks. However, the lack of a publicly available numerical implementation for most methods entails a significant initial investment of time and effort prior to any scientific investigations. In this paper, we present the open-source toolbox NNMT, which currently focuses on methods for LIF neurons but is intended as a platform for collecting standard implementations of various neuronal network model analyses based on mean-field theory that have been thoroughly tested and validated by the neuroscientific community (Riquelme and Gjorgjieva, 2021). As use cases, we reproduce known results from the literature: the non-linear relation between the firing rates and the external input of an E-I-network (Sanzeni et al., 2020), the firing rates of a cortical microcircuit model, its response to oscillatory input, its power spectrum, and the identification of the connections that predominantly contribute to the model's low frequency oscillations (Schuecker et al., 2015; Bos et al., 2016), and pattern formation in a spiking network, analyzed by mapping it to a rate model and conducting a linear stability analysis (Senk et al., 2020).

In the remainder of the discussion, we compare NNMT to other tools suited for network model analysis. We expand on the different use cases of NNMT and also point out the inherent limitations of analytical methods for neuronal network analysis. We conclude with suggestions on how new tools can be added to NNMT and how the toolbox may grow and develop in the future.

4.1. Comparison to Other Tools

There are various approaches and corresponding tools that can help to gain a better understanding of a neuronal network model. There are numerous simulators that numerically solve the dynamical equations for concrete realizations of a network model and all its stochastic components, often focusing either on the resolution of single-neurons, for example NEST (Gewaltig and Diesmann, 2007), Brian (Stimberg et al., 2019), or Neuron (Hines and Carnevale, 2001), or on the population level, for example TheVirtualBrain (Sanz Leon et al., 2013). Similarly, general-purpose dynamical system software like XPPAUT (Ermentrout, 2002) can be used. Simulation tools, like DynaSim (Sherfey et al., 2018), come with enhanced functionality for simplifying batch analysis and parameter explorations. All these tools yield time-series of activity, and some of them even provide the methods for analyzing the generated data. However, simulations only indirectly link a model's parameters with its activity: to gain an understanding of how a model's parameters influence the statistics of their activity, it is necessary to run many simulations with different parameters and analyze the generated data subsequently.

Other approaches provide a more direct insight into a model's behavior on an abstract level: TheVirtualBrain and the Brain Dynamics Toolbox (Heitmann et al., 2018), for example, allow plotting a model's phase space vector field while the parameters can be changed interactively, allowing for exploration of low-dimensional systems defined by differential equations without the need for simulations. XPPAUT has an interface to AUTO-07P (Doedel and Oldeman, 1998), a software for performing numerical bifurcation and continuation analysis. It is worth noting that such tools are limited to models that are defined in terms of differential equations. Models specified in terms of update rules, such as binary neurons, need to be analyzed differently, for example using mean-field theory.

A third approach is to simplify the model analytically and simulate the simplified version. The simulation platform DiPDE[4] utilizes the population density approach to simulate the statistical evolution of a network model's dynamics. Schwalger et al. (2017) start from a microscopic model of generalized integrate-and-fire neurons and derive mesoscopic mean-field population equations, which reproduce the statistical and qualitative behavior of the homogeneous neuronal sub-populations. Similarly, Montbrió et al. (2015) derive a set of non-linear

[4]http://alleninstitute.github.io/dipde

differential equations describing the dynamics of the rate and mean membrane potentials of a population of quadratic integrate-and-fire (QIF) neurons. The simulation platform PyRates (Gast et al., 2019) provides an implementation of this QIF mean-field model, and allows simulating them to obtain the temporal evolution of the population activity measures.

However, mean-field and related theories can go beyond such reduced dynamical equations: they can directly link model parameters to activity statistics, and they can even provide access to informative network properties that might not be accessible otherwise. The spectral bound (Rajan and Abbott, 2006) of the effective connectivity matrix in linear response theory (Lindner et al., 2005; Pernice et al., 2011; Trousdale et al., 2012) is an example of such a property. It is a measure for the stability of the linearized system and determines, for example, the occurrence of slow dynamics and long-range correlations (Dahmen et al., 2022). Another example is the sensitivity measure presented in Section 3.3.3, which directly links a network model's connectivity with the properties of its power spectrum. These measures are not accessible via simulations. They require analytical calculation.

Similarly, NNMT is not a simulator. NNMT is a collection of mean-field equation implementations that directly relate a model's parameters to the statistics of its dynamics or to other informative properties. It provides these implementations in a format that makes them applicable to as many network models as possible. This is not to say that NNMT does not involve numerical integration procedures; solving self-consistent equations, such as in the case of the firing rates calculations in Section 3.2.1 and Section 3.2.2, is a common task, and a collection of respective solvers is part of NNMT.

4.2. Use Cases

In Section 3, we present concrete examples of how to apply some of the tools available. Here, we revisit some of the examples to highlight the use cases NNMT lends itself to, as well as provide some ideas for how the toolbox could be utilized in future projects.

Analytical methods have the advantage of being fast, and typically they only require a limited amount of computational resources. The computational costs for calculating analytical estimates of dynamical network properties like firing rates, as opposed to the costs of running simulations of a network model, are independent of the number of neurons the network is composed of. This is especially relevant for parameter space explorations, for which many simulations have to be performed. To speed up prototyping, a modeler can first perform a parameter scan using analytical tools from NNMT to get an estimate of the right parameter regimes and subsequently run simulations on this restricted set of parameters to arrive at the final model parameters. An example of such a parameter scan is given in Section 3.2.1, where the firing rates of a network are studied as a function of the external input.

Additionally to speeding up parameter space explorations, analytical methods may guide parameter space explorations in another way: namely, by providing an analytical relation between network model parameters and network dynamics, which allows a targeted adjustment of specific parameters to achieve a desired network activity. The prime example implemented in NNMT is the sensitivity measure presented in Section 3.3.3, which provides an intuitive relation between the network connectivity and the peaks of the power spectrum corresponding to the dominant oscillation frequencies. As shown in the final part of Section 3.3.3, the sensitivity measure identifies the connections which need to be adjusted in order to modify the dominant oscillation mode in a desired manner. This illustrates a mean-field method that provides a modeler with additional information about the origin of a model's dynamics, such that a parameter space exploration can be restricted to the few identified crucial model parameters.

A modeler investigating which features of a network model are crucial for the emergence of certain activity characteristics observed in simulations might be interested in comparing models of differing complexity. The respective mappings can be derived in mean-field theory, and one variant included in NNMT, which is presented in Section 3.4, allows mapping a LIF network to a simpler rate network. This is useful to investigate whether spiking dynamics is crucial for the observed phenomenon.

On a general note, which kind of questions researchers pursue is limited by and therefore depends on the tools they have at hand (Dyson, 2012). The availability of sophisticated neural network simulators for example has lead to the development of conceptually new and more complex neural network models, precisely because their users could focus on actual research questions instead of implementations. We hope that collecting useful implementations of analytical tools for network model analysis will have a similar effect on the development of new tools and that it might lead to new, creative ways of applying them.

4.3. Limitations

As a collection of analytical methods, NNMT comes with inherent limitations that apply to any toolbox for analytical methods: it is restricted to network, neuron, and synapse models, as well as observables, for which a mean-field theory exists, and the tools are based on analytical assumptions, simplifications, and approximations, restricting their valid parameter regimes and their explanatory power, which we expand upon in the next paragraphs.

Analytical methods can provide good estimates of network model properties, but there are limitations that must be considered when interpreting results provided by NNMT: First of all, the employed numerical solvers introduce numerical inaccuracies, but they can be remedied by changing hyperparameters such as integration step sizes or iteration termination thresholds. More importantly, analytical methods almost always rely on approximations, which can only be justified if certain assumptions are fulfilled. Typical examples of such assumptions are fast or slow synapses, or a random connectivity. If such assumptions are not met, at least approximately, and the valid parameter regime of a tool is left, the corresponding method is not guaranteed to give reliable results. Hence, it is important to be aware of a tool's limitations, which we aim to document as thoroughly as possible.

An important assumption of mean-field theory is uncorrelated Poissonian inputs. As discussed in Section 3.2.1, asynchronous irregular activity is a robust feature of inhibition

dominated networks, and mean-field theory is well-suited to describe the activity of such models. However, if a network model features highly correlated activity, or strong external input common to many neurons, approximating the input by uncorrelated noise no longer holds and mean-field estimates become unreliable.

In addition to the breakdown of such assumptions, some approaches, like linear response theory, rely on neglecting higher order terms. This restricts the tools' explanatory power, as they cannot predict higher order effects, such as the presence of higher harmonics in a network's power spectrum. Addressing these deficiencies necessitates using more elaborate analyses, and users should be aware of such limitations when interpreting the results.

Finally, a specific limitation of NNMT is that it currently only collects methods for LIF neurons. However, one of the aims of this paper is to encourage other scientists to contribute to the collection, and we outline how to do so in the following section.

4.4. How to Contribute and Outlook

A toolbox like NNMT always is an ongoing project, and there are various aspects that can be improved. In this section, we briefly discuss how available methods could be improved, what and how new tools could be added, as well as the benefits of implementing a new method with the help of NNMT.

First of all, NNMT in its current state is partly vectorized but the included methods are not parallelized, e.g., using multiprocessing or MPI for Python (`mpi4py`). Vectorization relies on `NumPy` (Harris et al., 2020) and `SciPy` (Virtanen et al., 2020), which are thread-parallel for specific backends, e.g., IntelMKL. With the tools available in the toolbox at the moment, run-time only becomes an issue in extensive parameter scans, for instance, when the transfer function needs to be calculated for a large range of frequencies. To further reduce the runtime, the code could be made fully vectorized. Alternatively, parallelization of many tools in NNMT is straightforward, as many of them include `for` loops over the different populations of a network model and `for` loops over the different analysis frequencies. A third option is just-in-time compilation, as provided by `Numba` (Lam et al., 2015).

Another aspect to consider is the range of network models a tool can be applied to. Thus far, the toolbox primarily supports arbitrary block structured networks. Future developments could extend the class of networks to even more general models.

Due to the research focus at our lab, NNMT presently mainly contains tools for LIF neurons in the fast synaptic regime and networks with random connectivity. Nonetheless, the structure of NNMT allows for adding methods for different neuron types, like for example binary (Ginzburg and Sompolinsky, 1994) or conductance-based neurons (Izhikevich, 2007; Richardson, 2007), as well as more elaborate network models. Another way to improve the toolbox is adding tools that complement the existing ones: As discussed in Section 4.3, many mean-field methods only give valid results for certain parameter ranges. Sometimes, there exist different approximations for the same quantity, valid in complementary parameter regimes. A concrete example is the currently implemented version of the transfer function for leaky integrate-and-fire neurons, based on Schuecker et al. (2015), which gives a good estimate for small synaptic time constants compared to the membrane time constant $\tau_s/\tau_m \ll 1$. A complementary estimate for $\tau_s/\tau_m \gg 1$ has been developed by Moreno-Bote and Parga (2006). Similarly, the current implementation of the firing rates of leaky integrate-and-fire neurons, based on the work of Fourcaud and Brunel (2002), is valid for $\tau_s/\tau_m \ll 1$. Recently, van Vreeswijk and Farkhooi (2019) have developed a method accurate for all combinations of synaptic and membrane time constants.

In the following, we explain how such implementations can be added and how using NNMT helps implementing new methods. Clearly, the implementations of NNMT help implementing methods that build on already existing ones. An example is the firing rate for LIF neurons with exponential synapses `nnmt.lif.exp._firing_rates()` which wraps the calculation of firing rates for LIF neurons with delta synapses `nnmt.lif.delta._firing_rates()`. Additionally, the toolbox may support the implementation of tools for other neuron models. As an illustration, let us consider adding the computation of the mean activity for a network of binary neurons (included in NNMT 1.1.0). We start with the equations for the mean input μ_a, its variance σ_a^2, and the firing rates $\boldsymbol{m}$ (Helias et al., 2014, Equations 4, 6, and 7)

$$\begin{aligned} \mu_a(\boldsymbol{m}) &= \sum_b K_{ab} J_{ab} m_b \,, \\ \sigma_a^2(\boldsymbol{m}) &= \sum_b K_{ab} J_{ab}^2 m_b (1 - m_b) \,, \\ m_a(\mu_a, \sigma_a) &= \frac{1}{2} \mathrm{erfc}\left(\frac{\Theta_a - \mu_a}{\sqrt{2}\sigma_a} \right) , \end{aligned} \tag{19}$$

with indegree matrix K_{ab} from population b to population a, synaptic weight matrix J_{ab}, and firing-threshold Θ_a. The sum $\sum_b$ may include an external population providing input to the model. This set of self-consistent equations has the same structure as the self-consistent equations for the firing rates of a network of LIF neurons, Equation (8): the input statistics are given as functions of the rate, and the rate is given as a function of the input statistics. Therefore, it is possible to reuse the firing rate integration procedure for LIF neurons, providing immediate access to the two different methods presented in Section 3.2.1. Accordingly, it is sufficient to implement Equation (19) in a new submodule `nnmt.binary` and apply the solver provided by NNMT to extend the toolbox to binary neurons.

The above example demonstrates the benefits of collecting analytical tools for network model analysis in a common framework. The more methods and corresponding solvers the toolbox comprises, the easier implementing new methods becomes. Therefore, contributions to the toolbox are highly welcome; this can be done via the standard pull request workflow on GitHub (see the "Contributors guide" of the official documentation of NNMT[2]). We hope that in the future, many scientists will contribute to this collection of analytical methods for neuronal network model analysis, such that, at some point, we will have tools from all parts of mean-field theory

of neuronal networks, made accessible in a usable format to all neuroscientists.

DATA AVAILABILITY STATEMENT

Publicly available datasets were used in this study, and the corresponding sources are cited in the main text. The toolbox's repository can be found at https://github.com/INM-6/nnmt, and the parameter files used in the presented examples can be found in the examples section of the online documentation https://nnmt.readthedocs.io/en/latest/.

AUTHOR CONTRIBUTIONS

HB and MH developed and implemented the code base and the initial version of the toolbox. ML, JS, and SE designed the current version of the toolbox. ML implemented the current version of the toolbox, vectorized and generalized tools, developed and implemented the test suite, wrote the documentation, and created the example shown in Section 3.2.2. AM improved the numerics of the firing rate integration (Methods) and created the example shown in Section 3.2.1. SE implemented integration tests, improved the functions related to the `sensitivity_measure`, and created the examples shown in Section 3.3. JS developed and implemented the tools used in Section 3.4 and created the respective example. ML, JS, SE, AM, and MH wrote this article. All authors approved the submitted version.

FUNDING

This project has received funding from the European Union's Horizon 2020 Framework Programme for Research and Innovation under Specific Grant Agreement Nos. 720270 (HBP SGA1), 785907 (HBP SGA2), and 945539 (HBP SGA3), has been partially funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – 368482240/GRK2416, and has been partially funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – 491111487. This research was supported by the Joint Lab "Supercomputing and Modeling for the Human Brain".

ACKNOWLEDGMENTS

We would like to thank Jannis Schuecker, who has contributed to the development and implementation of the code base and the initial version of the toolbox, and Angela Fischer, who supported us designing **Figure 1**. Additionally, we would also like to thank our reviewers for the thorough and constructive feedback, which lead to significant improvements.

REFERENCES

Abramowitz, M., and Stegun, I. A. (1974). *Handbook of Mathematical Functions: With Formulas, Graphs, and Mathematical Tables* (New York: Dover Publications).

Ahmadian, Y., and Miller, K. D. (2021). What is the dynamical regime of cerebral cortex? *Neuron* 109, 3373–3391. doi: 10.1016/j.neuron.2021.07.031

Amari, S.-I. (1975). Homogeneous nets of neuron-like elements. *Biol. Cybern.* 17, 211–220. doi: 10.1007/BF00339367

Amari, S.-I. (1977). Dynamics of pattern formation in lateral-inhibition type neural fields. *Biol. Cybern.* 27, 77–87. doi: 10.1007/bf00337259

Amit, D. J., and Brunel, N. (1997a). Dynamics of a recurrent network of spiking neurons before and following learning. *Netw. Comp. Neural Sys.* 8, 373–404. doi: 10.1088/0954-898x_8_4_003

Amit, D. J. and Brunel, N. (1997b). Model of global spontaneous activity and local structured activity during delay periods in the cerebral cortex. *Cereb. Cortex* 7, 237–252. doi: 10.1093/cercor/7.3.237

Amit, D. J., and Tsodyks, M. V. (1991). Quantitative study of attractor neural network retrieving at low spike rates I: substrate–spikes, rates and neuronal gain. *Network* 2, 259. doi: 10.1088/0954-898X_2_3_003

Bos, H., Diesmann, M., and Helias, M. (2016). Identifying anatomical origins of coexisting oscillations in the cortical microcircuit. *PLOS Comput. Biol.* 12, e1005132. doi: 10.1371/journal.pcbi.1005132

Braitenberg, V. and Schüz, A. (1998). *Cortex: Statistics and Geometry of Neuronal Connectivity, 2nd Edn.* Berlin: Springer-Verlag.

Bressloff, P. C. (2012). Spatiotemporal dynamics of continuum neural fields. *J. Phys. A* 45, 033001. doi: 10.1088/1751-8113/45/3/033001

Bressloff, P. C., Cowan, J. D., Golubitsky, M., Thomas, P. J., and Wiener, M. C. (2001). Geometric visual hallucinations, euclidean symmetry and the functional architecture of striate cortex. *Phil. Trans. R. Soc. B* 356, 299–330. doi: 10.1098/rstb.2000.0769

Brunel, N. (2000). Dynamics of sparsely connected networks of excitatory and inhibitory spiking neurons. *J. Comput. Neurosci.* 8, 183–208. doi: 10.1023/a:1008925309027

Brunel, N., Chance, F. S., Fourcaud, N., and Abbott, L. F. (2001). Effects of synaptic noise and filtering on the frequency response of spiking neurons. *Phys. Rev. Lett.* 86, 2186–2189. doi: 10.1103/physrevlett.86.2186

Brunel, N., and Hakim, V. (1999). Fast global oscillations in networks of integrate-and-fire neurons with low firing rates. *Neural Comput.* 11, 1621–1671. doi: 10.1162/089976699300016179

Brunel, N., and Latham, P. (2003). Firing rate of the noisy quadratic integrate-and-fire neuron. *Neural Comput.* 15, 2281–2306. doi: 10.1162/089976603322362365

Buice, M. A., and Chow, C. C. (2013). Beyond mean field theory: statistical field theory for neural networks. *J. Stat. Mech.* 2013, P03003. doi: 10.1088/1742-5468/2013/03/P03003

Coombes, S. (2005). Waves, bumps, and patterns in neural field theories. *Biol. Cybern.* 93, 91–108. doi: 10.1007/s00422-005-0574-y

Coombes, S., bei Graben, P., Potthast, R., and Wright, J. (2014). *Neural Fields. Theory and Applications.* Berlin; Heidelberg: Springer-Verlag.

Corless, R. M., Gonnet, G. H., Hare, D. E. G., Jeffrey, D. J., and Knuth, D. E. (1996). On the lambert w function. *Adv. Comput. Math.* 5, 329–359. doi: 10.1007/BF02124750

Dahmen, D., Layer, M., Deutz, L., Dąbrowska, P. A., Voges, N., von Papen, M., et al. (2022). Global organization of neuronal activity only requires unstructured local connectivity. *eLife* 11, e68422. doi: 10.7554/eLife.68422.sa0

Dasbach, S., Tetzlaff, T., Diesmann, M., and Senk, J. (2021). Dynamical characteristics of recurrent neuronal networks are robust against low synaptic weight resolution. *Front. Neurosci.* 15, 757790. doi: 10.3389/fnins.2021.757790

DeFelipe, J., Alonso-Nanclares, L., and Arellano, J. (2002). Microstructure of the neocortex: comparative aspects. *J. Neurocytol.* 31, 299–316. doi: 10.1023/A:1024130211265

Doedel, E. J., and Oldeman, B. (1998). *Auto-07p: Continuation and Bifurcation Software.* Montreal, QC: Concordia University Canada

Dyson, F. J. (2012). Is science mostly driven by ideas or by tools? *Science* 338, 1426–1427. doi: 10.1126/science.1232773

Ermentrout, B. (2002). *Simulating, Analyzing, and Animating Dynamical Systems: A Guide to Xppaut for Researchers and Students (Software, Environments, Tools)*. Philadelphia, PA: Society for Industrial and Applied Mathematics.

Ermentrout, G. B., and Cowan, J. D. (1979). A mathematical theory of visual hallucination patterns. *Biol. Cybern.* 34, 137–150. doi: 10.1007/BF00336965

Fourcaud, N., and Brunel, N. (2002). Dynamics of the firing probability of noisy integrate-and-fire neurons. *Neural Comput.* 14, 2057–2110. doi: 10.1162/089976602320264015

Fourcaud-Trocmé, N., Hansel, D., van Vreeswijk, C., and Brunel, N. (2003). How spike generation mechanisms determine the neuronal response to fluctuating inputs. *J. Neurosci.* 23, 11628–11640. doi: 10.1523/JNEUROSCI.23-37-11628.2003

Gast, R., Rose, D., Salomon, C., Möller, H. E., Weiskopf, N., and Knösche, T. R. (2019). Pyrates - a python framework for rate-based neural simulations. *PLoS ONE* 14, e0225900. doi: 10.1371/journal.pone.0225900

Gerstner, W., Kistler, W. M., Naud, R., and Paninski, L. (2014). *Neuronal Dynamics. From Single Neurons to Networks and Models of Cognition*. Cambridge: Cambridge University Press.

Gewaltig, M.-O., and Diesmann, M. (2007). NEST (nEural simulation tool). *Scholarpedia* 2, 1430. doi: 10.4249/scholarpedia.1430

Giese, M. A. (2012). *Dynamic Neural Field Theory for Motion Perception, Vol. 469*. Berlin; Heidelberg: Springer Science & Business Media)

Ginzburg, I., and Sompolinsky, H. (1994). Theory of correlations in stochastic neural networks. *Phys. Rev. E* 50, 3171–3191. doi: 10.1103/PhysRevE.50.3171

Goldenfeld, N. (1992). *Lectures on Phase Transitions and the Renormalization Group*. Reading, MA: Perseus books.

Golosio, B., Tiddia, G., Luca, C. D., Pastorelli, E., Simula, F., and Paolucci, P. S. (2021). Fast simulations of highly-connected spiking cortical models using GPUs. *Front. Comput. Neurosci.* 15, 627620. doi: 10.3389/fncom.2021.627620

Grabska-Barwinska, A., and Latham, P. (2014). How well do mean field theories of spiking quadratic-integrate-and-fire networks work in realistic parameter regimes? *J. Comput. Neurosci.* 36, 469–481. doi: 10.1007/s10827-013-0481-5

Grytskyy, D., Tetzlaff, T., Diesmann, M., and Helias, M. (2013). A unified view on weakly correlated recurrent networks. *Front. Comput. Neurosci.* 7, 131. doi: 10.3389/fncom.2013.00131

Hagen, E., Dahmen, D., Stavrinou, M. L., Lindén, H., Tetzlaff, T., van Albada, S. J., et al. (2016). Hybrid scheme for modeling local field potentials from point-neuron networks. *Cereb. Cortex* 26, 4461–4496. doi: 10.1093/cercor/bhw237

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., et al. (2020). Array programming with NumPy. *Nature* 585, 357–362. doi: 10.1038/s41586-020-2649-2

Heitmann, S., Aburn, M. J., and Breakspear, M. (2018). The brain dynamics toolbox for matlab. *Neurocomputing* 315, 82–88. doi: 10.1016/j.neucom.2018.06.026

Helias, M., Tetzlaff, T., and Diesmann, M. (2014). The correlation structure of local cortical networks intrinsically results from recurrent dynamics. *PLoS Comput. Biol.* 10, e1003428. doi: 10.1371/journal.pcbi.1003428

Hertz, J. (2010). Cross-correlations in high-conductance states of a model cortical network. *Neural Comput.* 22, 427–447. doi: 10.1162/neco.2009.06-08-806

Hines, M. L., and Carnevale, N. T. (2001). NEURON: a tool for neuroscientists. *Neuroscientist* 7, 123–135. doi: 10.1177/107385840100700207

Izhikevich, E. M. (2007). *Dynamical Systems in Neuroscience: The Geometry of Excitability and Bursting*. Cambridge, MA: MIT Press.

Jirsa, V. K., and Haken, H. (1996). Field theory of electromagnetic brain activity. *Phys. Rev. Lett.* 77, 960. doi: 10.1103/PhysRevLett.77.960

Jirsa, V. K., and Haken, H. (1997). A derivation of a macroscopic field theory of the brain from the quasi-microscopic neural dynamics. *Phys. D* 99, 503–526. doi: 10.1016/S0167-2789(96)00166-2

Knight, J. C., and Nowotny, T. (2018). GPUs outperform current HPC and neuromorphic solutions in terms of speed and energy when simulating a highly-connected cortical model. *Front. Neurosci.* 12, 941. doi: 10.3389/fnins.2018.00941

Laing, C. R., and Troy, W. C. (2003). Two-bump solutions of amari-type models of neuronal pattern formation. *Phys. D* 178, 190–218. doi: 10.1016/S0167-2789(03)00013-7

Laing, C. R., Troy, W. C., Gutkin, B., and Ermentrout, B. G. (2002). Multiple bumps in a neuronal model of working memory. *SIAM J. Appl. Math.* 63, 62–97. doi: 10.1137/s0036139901389495

Lam, S. K., Pitrou, A., and Seibert, S. (2015). "Numba: a llvm-based python jit compiler," in *Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC*, Austin, TX, 1–6

Layer, M., Senk, J., Essink, S., van Meegen, A., Bos, H., and Helias, M. (2021). NNMT (1.0.0). *Zenodo*. doi: 10.5281/zenodo.5779548

Lindner, B. (2004). Interspike interval statistics of neurons driven by colored noise. *Phys. Rev. E* 69, 0229011–0229014. doi: 10.1103/PhysRevE.69.022901

Lindner, B., Doiron, B., and Longtin, A. (2005). Theory of oscillatory firing induced by spatially correlated noise and delayed inhibitory feedback. *Phys. Rev. E* 72, 061919. doi: 10.1103/physreve.72.061919

Lindner, B., and Longtin, A. (2005). Effect of an exponentially decaying threshold on the firing statistis of a stochastic integate-and-fire neuron. *J. Theor. Biol.* 232, 505–521. doi: 10.1016/j.jtbi.2004.08.030

Lindner, B., and Schimansky-Geier, L. (2001). Transmission of noise coded versus additive signals through a neuronal ensemble. *Phys. Rev. Lett.* 86, 2934–2937. doi: 10.1103/physrevlett.86.2934

Mattia, M., Biggio, M., Galluzzi, A., and Storace, M. (2019). Dimensional reduction in networks of non-markovian spiking neurons: Equivalence of synaptic filtering and heterogeneous propagation delays. *PLoS Comput. Biol.* 15, e1007404. doi: 10.1371/journal.pcbi.1007404

Montbrió, E., Pazó, D., and Roxin, A. (2015). Macroscopic description for networks of spiking neurons. *Phys Rev X* 5, 021028. doi: 10.1103/PhysRevX.5.021028

Moreno-Bote, R., and Parga, N. (2006). Auto- and crosscorrelograms for the spike response of leaky integrate-and-fire neurons with slow synapses. *Phys. Rev. Lett.* 96, 028101. doi: 10.1103/PhysRevLett.96.028101

Nunez, P. L. (1974). The brain wave equation: a model for the eeg. *Math. Biosci.* 21, 279–297. doi: 10.1016/0025-5564(74)90020-0

Olver, F. W. J., Olde Daalhuis, A. B., Lozier, D. W., Schneider, B. I., Boisvert, R. F., Clark, C. W., et al. (2021). *NIST Digital Library of Mathematical Functions*. Available online at: http://dlmf.nist.gov/

Ostojic, S. (2014). Two types of asynchronous activity in networks of excitatory and inhibitory spiking neurons. *Nat. Neurosci.* 17, 594–600. doi: 10.1038/nn.3658

Ostojic, S., and Brunel, N. (2011). From spiking neuron models to linear-nonlinear models. *PLoS Comput. Biol.* 7, e1001056. doi: 10.1371/journal.pcbi.1001056

Pernice, V., Staude, B., Cardanobile, S., and Rotter, S. (2011). How structure determines correlations in neuronal networks. *PLoS Comput. Biol.* 7, e1002059. doi: 10.1371/journal.pcbi.1002059

Potjans, T. C., and Diesmann, M. (2014). The cell-type specific cortical microcircuit: relating structure and activity in a full-scale spiking network model. *Cereb. Cortex* 24, 785–806. doi: 10.1093/cercor/bhs358

Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. (2007). *Numerical Recipes: The Art of Scientific Computing*, 3rd edn. Cambridge University Press.

Rajan, K., and Abbott, L. F. (2006). Eigenvalue spectra of random matrices for neural networks. *Phys. Rev. Lett.* 97, 188104. doi: 10.1103/PhysRevLett.97.188104

Renart, A., De La Rocha, J., Bartho, P., Hollender, L., Parga, N., Reyes, A., et al. (2010). The asynchronous state in cortical circuits. *Science* 327, 587–590. doi: 10.1126/science.1179850

Richardson, M. J. E. (2007). Firing-rate response of linear and nonlinear integrate-and-fire neurons to modulated current-based and conductance-based synaptic drive. *Phys. Rev. E* 76, 1–15. doi: 10.1103/PhysRevE.76.021919

Richardson, M. J. E. (2008). Spike-train spectra and network response functions for non-linear integrate-and-fire neurons. *Biol. Cybern.* 99, 381–392. doi: 10.1007/s00422-008-0244-y

Riquelme, J. L., and Gjorgjieva, J. (2021). Towards readable code in neuroscience. *Nat. Rev. Neurosci.* 22, 257–258. doi: 10.1038/s41583-021-00450-y

Rosenbaum, R. and Doiron, B. (2014). Balanced networks of spiking neurons with spatially dependent recurrent connections. *Phys. Rev. X* 4, 021039. doi: 10.1103/PhysRevX.4.021039

Rosenbaum, R., Smith, M. A., Kohn, A., Rubin, J. E., and Doiron, B. (2017). The spatial structure of correlated neuronal variability. *Nat. Neurosci.* 20, 107–114. doi: 10.1038/nn.4433

Sanz Leon, P., Knock, S., Woodman, M., Domide, L., Mersmann, J., McIntosh, A., et al. (2013). The virtual brain: a simulator of primate brain network dynamics. *Front. Neuroinform.* 7, 10. doi: 10.3389/fninf.2013.00010

Sanzeni, A., Histed, M. H., and Brunel, N. (2020). Response nonlinearities in networks of spiking neurons. *PLOS Comput. Biol.* 16, e1008165. doi: 10.1371/journal.pcbi.1008165

Schmidt, M., Bakker, R., Hilgetag, C. C., Diesmann, M., and van Albada, S. J. (2018). Multi-scale account of the network structure of macaque visual cortex. *Brain Struct. Func.* 223, 1409–1435. doi: 10.1007/s00429-017-1554-4

Schöner, G. (2008). "Dynamical systems approaches to cognition," in *Cambridge Handbook of Computational Cognitive Modeling*. Cambridge: Cambridge University Press, 101–126.

Schuecker, J., Diesmann, M., and Helias, M. (2014). Reduction of colored noise in excitable systems to white noise and dynamic boundary conditions. *arXiv[Preprint].arXiv:1410.8799*. doi: 10.48550/arXiv.1410.8799

Schuecker, J., Diesmann, M., and Helias, M. (2015). Modulated escape from a metastable state driven by colored noise. *Phys. Rev. E* 92, 052119. doi: 10.1103/PhysRevE.92.052119

Schuecker, J., Goedeke, S., and Helias, M. (2018). Optimal sequence memory in driven random networks. *Phys. Rev. X* 8, 041029. doi: 10.1103/PhysRevX.8.041029

Schwalger, T., Deger, M., and Gerstner, W. (2017). Towards a theory of cortical columns: From spiking neurons to interacting neural populations of finite size. *PLoS Comput. Biol.* 13, e1005507. doi: 10.1371/journal.pcbi.1005507

Schwalger, T., Droste, F., and Lindner, B. (2015). Statistical structure of neural spiking under non-poissonian or other non-white stimulation. *J. Comput. Neurosci.* 39, 29. doi: 10.1007/s10827-015-0560-x

Sejnowski, T. (1976). On the stochastic dynamics of neuronal interaction. *Biol. Cybern.* 22, 203–211. doi: 10.1007/BF00365086

Senk, J., Korvasová, K., Schuecker, J., Hagen, E., Tetzlaff, T., Diesmann, M., et al. (2020). Conditions for wave trains in spiking neural networks. *Phys. Rev. Res.* 2, 023174. doi: 10.1103/physrevresearch.2.023174

Senk, J., Kriener, B., Djurfeldt, M., Voges, N., Jiang, H.-J., Schüttler, L., et al. (in press). Connectivity concepts in neuronal network modeling. *PLOS Comput. Biol*.

Sherfey, J. S., Soplata, A. E., Ardid, S., Roberts, E. A., Stanley, D. A., Pittman-Polletta, B. R., et al. (2018). Dynasim: a matlab toolbox for neural modeling and simulation. *Front. Neuroinform.* 12, 10. doi: 10.3389/fninf.2018.00010

Siegert, A. J. (1951). On the first passage time probability problem. *Phys. Rev.* 81, 617–623. doi: 10.1103/PhysRev.81.617

Sompolinsky, H., Crisanti, A., and Sommers, H. J. (1988). Chaos in random neural networks. *Phys. Rev. Lett.* 61, 259–262. doi: 10.1103/PhysRevLett.61.259

Stiller, J., and Radons, G. (1998). Dynamics of nonlinear oscillators with random interactions. *Phys. Rev. E* 58, 1789. doi: 10.1103/PhysRevE.58.1789

Stimberg, M., Brette, R., and Goodman, D. F. (2019). Brian 2, an intuitive and efficient neural simulator. *eLife* 8, e47314. doi: 10.7554/elife.47314

Tetzlaff, T., Helias, M., Einevoll, G. T., and Diesmann, M. (2012). Decorrelation of neural-network activity by inhibitory feedback. *PLOS Comput. Biol.* 8, e1002596. doi: 10.1371/journal.pcbi.1002596

Toyoizumi, T., and Abbott, L. F. (2011). Beyond the edge of chaos: Amplification and temporal integration by recurrent networks in the chaotic regime. *Phys. Rev. E* 84, 051908. doi: 10.1103/PhysRevE.84.051908

Trousdale, J., Hu, Y., Shea-Brown, E., and Josic, K. (2012). Impact of network structure and cellular response on spike time correlations. *PLoS Comput. Biol.* 8, e1002408. doi: 10.1371/journal.pcbi.1002408

Tuckwell, H. C. (1988). *Introduction to Theoretical Neurobiology, Vol. 2* Cambridge: Cambridge University Press.

van Albada, S. J., Rowley, A. G., Senk, J., Hopkins, M., Schmidt, M., Stokes, A. B., et al. (2018). Performance comparison of the digital neuromorphic hardware SpiNNaker and the neural network simulation software NEST for a full-scale cortical microcircuit model. *Front. Neurosci.* 12, 291. doi: 10.3389/fnins.2018.00291

van Meegen, A., and Lindner, B. (2018). Self-consistent correlations of randomly coupled rotators in the asynchronous state. *Phys. Rev. Lett.* 121, 258302. doi: 10.1103/PhysRevLett.121.258302

van Vreeswijk, C., and Farkhooi, F. (2019). Fredholm theory for the mean first-passage time of integrate-and-fire oscillators with colored noise input. *Phys. Rev. E* 100, 060402. doi: 10.1103/PhysRevE.100.060402

van Vreeswijk, C., and Sompolinsky, H. (1996). Chaos in neuronal networks with balanced excitatory and inhibitory activity. *Science* 274, 1724–1726. doi: 10.1126/science.274.5293.1724

van Vreeswijk, C., and Sompolinsky, H. (1998). Chaotic balanced state in a model of cortical circuits. *Neural Comput.* 10, 1321–1371. doi: 10.1162/089976698300017214

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., et al. (2020). SciPy 1.0: fundamental algorithms for scientific computing in python. *Nat. Methods* 17, 261–272. doi: 10.1038/s41592-019-0686-2

Wagatsuma, N., Potjans, T. C., Diesmann, M., and Fukai, T. (2011). Layer-dependent attentional processing by top-down signals in a visual cortical microcircuit model. *Front. Comput. Neurosci.* 5, 31. doi: 10.3389/fncom.2011.00031

Wilson, H. R., and Cowan, J. D. (1972). Excitatory and inhibitory interactions in localized populations of model neurons. *Biophys. J.* 12, 1 – 24. doi: 10.1016/S0006-3495(72)86068-5

Wilson, H. R., and Cowan, J. D. (1973). A mathematical theory of the functional dynamics of cortical and thalamic nervous tissue. *Kybernetik* 13, 55–80. doi: 10.1007/BF00288786

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

A. APPENDIX

A.1. Siegert Implementation

Here, we describe how we solve the integral in Equation (4) numerically in a fully vectorized manner. The difficulty in Equation (4), $\phi(\mu,\sigma) = 1/[\tau_r + \tau_m\sqrt{\pi} I(\widetilde{V}_0, \widetilde{V}_{th})]$ where $\widetilde{V}_0 = \widetilde{V}_0(\mu,\sigma)$ and $\widetilde{V}_{th} = \widetilde{V}_{th}(\mu,\sigma)$ are determined by either Equation (5) or Equation (10), is posed by the integral

$$I(\widetilde{V}_0, \widetilde{V}_{th}) = \int_{\widetilde{V}_0}^{\widetilde{V}_{th}} e^{s^2}(1 + \mathrm{erf}(s))\mathrm{d}s. \tag{A1}$$

This integral is problematic due to the multiplication of e^{s^2} and $1 + \mathrm{erf}(s)$ in the integrand which leads to overflow and loss of significance.

To address this, we split the integral into different domains depending on the sign of the integration variable. Furthermore, we use the scaled complementary error function

$$\mathrm{erf}(s) = 1 - e^{-s^2}\mathrm{erfcx}(s) \tag{A2}$$

to extract the leading exponential contribution. Importantly, $\mathrm{erfcx}(s)$ decreases monotonically from $\mathrm{erfcx}(0) = 1$ with power law asymptotics $\mathrm{erfcx}(s) \sim 1/(\sqrt{\pi}s)$, hence it does not contain any exponential contribution. For positive s, the exponential contribution in the prefactor of $\mathrm{erfcx}(s)$ cancels the e^{s^2} factor in the integrand. For negative s, the integrand simplifies even further to $e^{s^2}(1 + \mathrm{erf}(-s)) = \mathrm{erfcx}(s)$ using $\mathrm{erf}(-s) = -\mathrm{erf}(s)$. In addition to $\mathrm{erfcx}(s)$, we employ the Dawson function

$$D(s) = e^{-s^2}\int_0^s e^{r^2}\mathrm{d}r \tag{A3}$$

to solve some of the integrals analytically. The Dawson function has a power law tail, $D(s) \sim 1/(2s)$; hence, it also does not carry an exponential contribution. Both $\mathrm{erfcx}(s)$ and the Dawson function are fully vectorized in `SciPy` (Virtanen et al., 2020).

Any remaining integrals are solved using Gauss–Legendre quadrature (Press et al., 2007). By construction, Gauss–Legendre quadrature of order k solves integrals of polynomials up to degree k on the interval $[-1, 1]$ exactly. Thus, it gives very good results if the integrand is well approximated by a polynomial of degree k. The quadrature rule itself is

$$\int_a^b f(s)\mathrm{d}s \approx \frac{b-a}{2}\sum_{i=1}^{k} w_i f\left(\frac{b-a}{2}u_i + \frac{b+a}{2}\right), \tag{A4}$$

where the u_i are the roots of the Legendre polynomial of order k and the w_i are appropriate weights such that a polynomial of degree k is integrated exactly. We use a fixed order quadrature for which Equation (A4) is straightforward to vectorize to multiple a and b. We determine the order of the quadrature iteratively by comparison with an adaptive quadrature rule; usually, a small order $k = O(10)$ already yields very good results for an $\mathrm{erfcx}(s)$ integrand.

Inhibitory Regime

First, we consider the case where lower and upper bound of the integral are positive, $0 < \widetilde{V}_0 < \widetilde{V}_{th}$. This corresponds to strongly inhibitory mean input. Expressing the integrand in terms of $\mathrm{erfcx}(s)$ and using the Dawson function, we get

$$I_{inh}(\widetilde{V}_0, \widetilde{V}_{th}) = 2e^{\widetilde{V}_{th}^2}D(\widetilde{V}_{th}) - 2e^{\widetilde{V}_0^2}D(\widetilde{V}_0) - \int_{\widetilde{V}_0}^{\widetilde{V}_{th}} \mathrm{erfcx}(s)\mathrm{d}s.$$

The remaining integral is evaluated using Gauss–Legendre quadrature, Equation (A4). We extract the leading contribution $e^{\widetilde{V}_{th}^2}$ from the denominator in Equation (4) and arrive at

$$\phi(\mu,\sigma) = \frac{e^{-\widetilde{V}_{th}^2}}{\tau_r e^{-\widetilde{V}_{th}^2} + \tau_m\sqrt{\pi}\left(e^{-\widetilde{V}_{th}^2}I_{inh}(\widetilde{V}_0, \widetilde{V}_{th})\right)}. \tag{A5}$$

Extracting $e^{\widetilde{V}_{th}^2}$ from the denominator reduces the latter to $2\tau_m\sqrt{\pi}D(\widetilde{V}_{th})$ and exponentially small correction terms (remember $0 < \widetilde{V}_0 < \widetilde{V}_{th}$ because $V_0 < V_{th}$), thereby preventing overflow.

Excitatory Regime

Second, we consider the case where lower and upper bound of the integral are negative, $\widetilde{V}_0 < \widetilde{V}_{th} < 0$. This corresponds to strongly excitatory mean input. In this regime, we change variables $s \to -s$ to make the domain of integration positive. Using $\mathrm{erf}(-s) = -\mathrm{erf}(s)$ as well as $\mathrm{erfcx}(s)$, we get

$$I_{exc}(\widetilde{V}_0, \widetilde{V}_{th}) = \int_{|\widetilde{V}_{th}|}^{|\widetilde{V}_0|} \mathrm{erfcx}(s)\mathrm{d}s.$$

Thus, we evaluate Equation (4) as

$$\phi(\mu,\sigma) = \frac{1}{\tau_r + \tau_m\sqrt{\pi}\int_{|\widetilde{V}_{th}|}^{|\widetilde{V}_0|}\mathrm{erfcx}(s)\mathrm{d}s}. \tag{A6}$$

In particular, there is no exponential contribution involved in this regime.

Intermediate Regime

Last, we consider the remaining case $\widetilde{V}_0 \le 0 \le \widetilde{V}_{th}$. We split the integral at zero and use the previous steps for the respective parts to get

$$I_{interm}(\widetilde{V}_0, \widetilde{V}_{th}) = 2e^{\widetilde{V}_{th}^2}D(\widetilde{V}_{th}) + \int_{\widetilde{V}_{th}}^{|\widetilde{V}_0|} \mathrm{erfcx}(s)\mathrm{d}s.$$

Note that the sign of the second integral depends on whether $|\widetilde{V}_0| > \widetilde{V}_{th}$ (+) or not (−). Again, we extract the leading contribution $e^{\widetilde{V}_{th}^2}$ from the denominator in Equation (4) and arrive at

$$\phi(\mu,\sigma) = \frac{e^{-\widetilde{V}_{th}^2}}{\tau_r e^{-\widetilde{V}_{th}^2} + \tau_m\sqrt{\pi}\left(e^{-\widetilde{V}_{th}^2}I_{interm}(\widetilde{V}_0, \widetilde{V}_{th})\right)}. \tag{A7}$$

As before, extracting $e^{\widetilde{V}_{th}^2}$ from the denominator prevents overflow.

Deterministic Limit

The deterministic limit $\sigma \to 0$ corresponds to $|\widetilde{V}_0|, |\widetilde{V}_{\text{th}}| \to \infty$ for both Equation (5) and Equation (10). In the inhibitory and the intermediate regime, we see immediately that $\phi(\mu, \sigma \to 0) \to 0$ due to the dominant contribution $e^{-\widetilde{V}_{\text{th}}^2}$. In the excitatory regime, we use the asymptotics erfcx$(s) \sim 1/(\sqrt{\pi}s)$ to get

$$I(\widetilde{V}_0, \widetilde{V}_{\text{th}}) \to \int_{|\widetilde{V}_{\text{th}}|}^{|\widetilde{V}_0|} \frac{1}{\sqrt{\pi}s} \mathrm{d}s = \frac{1}{\sqrt{\pi}} \ln \frac{|\widetilde{V}_0|}{|\widetilde{V}_{\text{th}}|}.$$

Inserting this into Equation (4) yields

$$\phi(\mu, \sigma) \to \begin{cases} \frac{1}{\tau_{\text{r}} + \tau_{\text{m}} \ln \frac{\mu - V_0}{\mu - V_{\text{th}}}} & \text{if } \mu > V_{\text{th}} \\ 0 & \text{otherwise} \end{cases}, \tag{A8}$$

which is the firing rate of a leaky integrate-and-fire neuron driven by a constant input (Gerstner et al., 2014). Thus, this implementation also tolerates the deterministic limit of a very small noise intensity σ.

TABLE A1 | Microcircuit Parameters.

Symbol	Value (Potjans and Diesmann, 2014)	Value (Bos et al., 2016)	Description
$K_{4E,4I}$	795	675	In-degree from 4I to 4E
$K_{4E,ext}$	2100	1780	External in-degree to 4E
$D(\omega)$	none	truncated Gaussian	Delay distribution
$d_e \pm \delta d_e$	1.5 ± 0.75 ms	1.5 ± 1.5 ms	Mean and standard deviation of excitatory delay
$d_i \pm \delta d_i$	0.75 ± 0.375 ms	0.75 ± 0.75 ms	Mean and standard deviation of inhibitory delay

Parameter adaptions used here are introduced by Bos et al. (2016) compared to original microcircuit model. K_{ij} denotes the in-degrees from population j to population i. The delays in the simulated networks were drawn from a truncated Gaussian distribution with the given mean and standard deviation. The mean-field approximation of the microcircuit by Potjans and Diesmann (2014) assumes the delay to be fixed at the mean value, which is specified in the toolbox by setting the parameter `delay_dist` *to* `none`.

A.2. Transfer Function Notations

In Section 3.3.1 we introduce the analytical form of the transfer function implemented in the toolbox. Schuecker et al. (2015), derive a more general form of the transfer function, which includes a modulation of the variance of the input. Here we compare the notation used in Equation (11) to the notation used in Schuecker et al. (2015, Eq. 29).

Schuecker et al. (2015) define the modulations of input mean and variance as

$$\begin{aligned} \mu(t) &= \mu + \epsilon\mu\, \mathrm{e}^{\mathrm{i}\omega t}, \\ \sigma^2(t) &= \sigma^2 + H\sigma^2\, \mathrm{e}^{\mathrm{i}\omega t}, \end{aligned} \tag{A9}$$

and introduce the transfer function in terms of its influence on the firing rate

$$\nu(t)/\nu_0 = 1 + n(\omega)\, \mathrm{e}^{\mathrm{i}\omega t},$$

where ν_0 is the stationary firing rate. Here the transfer function $n(\omega)$ includes contributions of both the modulation of the mean $n_G(\omega) \propto \epsilon$ and the modulation of the variance $n_H(\omega) \propto H$. We write the modulation of the mean as

$$\mu(t) = \mu + \delta\mu\, \mathrm{e}^{\mathrm{i}\omega t},$$

implying that $\delta\mu$ corresponds to $\epsilon\mu$ in Equation (A9). As we only consider the modulation of the mean, the firing rate can be rewritten as

$$\nu(t) = \nu + N(\omega)\, \delta\mu\, \mathrm{e}^{\mathrm{i}\omega t},$$

where we moved the stationary firing rate ν to the right hand side and included it in the definition of the transfer function $N(\omega)$. In the main text we emphasize that $\mu(t)$ and $\nu(t)$ are physical quantities by only considering the real part of complex contributions. Additionally, we swap the voltage boundaries in Equation (11), introducing a canceling sign change in both the numerator and the denominator. This reformulation was chosen to align the presented formula with the implementation in the toolbox.

BIBLIOGRAPHY

Abbott, L. F. (2008). "Theoretical Neuroscience Rising." In: *Neuron* 60.3, pp. 489–495.

Abramowitz, M. and I. A. Stegun (1974). *Handbook of Mathematical Functions: with Formulas, Graphs, and Mathematical Tables*. New York: Dover Publications.

Ahmadian, Y. and K. D. Miller (2021). "What is the dynamical regime of cerebral cortex?" In: *Neuron* 109.21, pp. 3373–3391.

Alemohammad, S., Z. Wang, R. Balestriero, and R. Baraniuk (2021). "The Recurrent Neural Tangent Kernel." In: *International Conference on Learning Representations*.

Aljadeff, J., M. Stern, and T. Sharpee (2015). "Transition to Chaos in Random Networks with Cell-Type-Specific Connectivity." In: *Phys. Rev. Lett.* 114 (8), p. 088101.

Amari, S.-I. (1972). "Characteristics of random nets of analog neuron-like elements." In: *IEEE Trans. Syst. Man. Cybern.* SMC-2.5, pp. 643–657.

– (1975). "Homogeneous nets of neuron-like elements." In: *Biol. Cybern.* 17.4, pp. 211–220.

– (1977). "Dynamics of Pattern Formation in Lateral-Inhibition Type Neural Fields." In: *Biol. Cybern.* 27.2, pp. 77–87.

Amit, D. J., S. Fusi, and V. Yakovlev (1997). "Paradigmatic Working Memory (Attractor) Cell in IT Cortex." In: *Neural Comput.* 9.5, pp. 1071–1092.

Amit, D. J. and N. Brunel (1997). "Model of Global Spontaneous Activity and Local Structured Activity During Delay periods in the Cerebral Cortex." In: *Cereb. Cortex* 7, pp. 237–252.

Amit, D. J., H. Gutfreund, and H. Sompolinsky (1985). "Storing Infinite Numbers of Patterns in a Spin-Glass Model of Neural Networks." In: *Phys. Rev. Lett.* 55.14, pp. 1530–1533.

Amunts, K., H. Mohlberg, S. Bludau, and K. Zilles (2020). "Julich-Brain: A 3D probabilistic atlas of the human brain's cytoarchitecture." In: *Science* 369.6506, pp. 988–992.

Anderson, J. C., H. Kennedy, and K. A. C. Martin (2011). "Pathways of Attention: Synaptic Relationships of Frontal Eye Field to V4, Lateral Intraparietal Cortex, and Area 46 in Macaque Monkey." In: *J. Neurosci.* 31.30, pp. 10872–10881.

Anderson, P. W. (1972). "More Is Different." In: *Science* 177.4047, pp. 393–396.

Andersson, J. L., S. Skare, and J. Ashburner (2003). "How to correct susceptibility distortions in spin-echo echo-planar images: appli-

cation to diffusion tensor imaging." In: *Neuroimage* 20.2, pp. 870–888.

Antognini, J. M. (2019). "Finite size corrections for neural network Gaussian processes." In: *ArXiv*.

Arkhipov, A. et al. (2018). "Visual physiology of the layer 4 cortical circuit in silico." In: *PLOS Comput. Biol.* 14.11, e1006535.

Arous, G. B. and A. Guionnet (1995). "Large deviations for Langevin spin glass dynamics." In: *Probab. Theory Relat. Fields* 102.4, pp. 455–509.

Atapour, N., P. Majka, I. H. Wolkowicz, D. Malamanova, K. H. Worthy, and M. G. Rosa (2019). "Neuronal distribution across the cerebral cortex of the marmoset monkey (Callithrix jacchus)." In: *Cereb. Cortex* 29.9, pp. 3836–3863.

Azaïs, J.-M. and M. Wschebor (2009). *Level Sets and Extrema of Random Processes and Fields*. John Wiley & Sons.

Azevedo, F. A. C., L. R. B. Carvalho, L. T. Grinberg, J. M. Farfel, R. E. L. Ferretti, R. E. P. Leite, W. J. Filho, R. Lent, and S. Herculano-Houzel (2009). "Equal numbers of neuronal and nonneuronal cells make the human brain an isometrically scaled-up primate brain." In: *J. Comp. Neurol.* 513.5, pp. 532–541.

Bakker, R., W. Thomas, and M. Diesmann (2012). "CoCoMac 2.0 and the future of tract-tracing databases." In: *Front. Neuroinform.* 6, p. 30.

Barabási, A.-L. and R. Albert (1999). "Emergence of scaling in random networks." In: *Science* 286.5439, pp. 509–512.

Barak, O. (2017). "Recurrent neural networks as versatile tools of neuroscience research." In: *Curr. Opin. Neurobiol.* 46, pp. 1–6.

Barbas, H. and N. Rempel-Clower (1997). "Cortical structure predicts the pattern of corticocortical connections." In: *Cereb. Cortex* 7.7, pp. 635–646.

Bastos, A. M., W. M. Usrey, R. a. Adams, G. R. Mangun, P. Fries, and K. J. Friston (2012). "Canonical microcircuits for predictive coding." In: *Neuron* 76.4, pp. 695–711.

Behzadi, Y., K. Restom, J. Liau, and T. T. Liu (2007). "A component based noise correction method (CompCor) for BOLD and perfusion based fMRI." In: *Neuroimage* 37.1, pp. 90–101.

Beiran, M. and S. Ostojic (2019). "Contrasting the effects of adaptation and synaptic filtering on the timescales of dynamics in recurrent networks." In: *PLOS Comput. Biol.* 15.3, e1006893.

Bélanger, M., I. Allaman, and P. J. Magistretti (2011). "Brain energy metabolism: focus on astrocyte-neuron metabolic cooperation." In: *Cell Metabolism* 14.6, pp. 724–738.

Bellec, G., D. Salaj, A. Subramoney, R. Legenstein, and W. Maass (2018). "Long short-term memory and Learning-to-learn in networks of spiking neurons." In: *Advances in Neural Information Processing Systems 31*. Ed. by S. Bengio, H. Wallach, H. Larochelle, K. Grauman,

N. Cesa-Bianchi, and R. Garnett. Curran Associates, Inc., pp. 795–805.

Bellec, G., S. Wang, A. Modirshanechi, J. Brea, and W. Gerstner (2021). "Fitting summary statistics of neural data with a differentiable spiking network simulator." In: *Adv. Neural Inf. Process. Syst.* 34.

Bender, C. M. and S. A. Orszag (1999). *Advanced Mathematical Methods for Scientists and Engineers*. Springer.

Berg, H. C. (1993). *Random Walks in Biology: New and Expanded Edition*. Princeton University Press.

Berg, J. et al. (2021). "Human neocortical expansion involves glutamatergic neuron diversification." In: *Nature* 598.7879, pp. 151–158.

Berger, M. S. (1977). *Nonlinearity and Functional Analysis*. 1st ed. Elsevier.

Bernacchia, A., H. Seo, D. Lee, and X.-J. Wang (2011). "A reservoir of time constants for memory traces in cortical neurons." In: *Nat. Neurosci.* 14.3, pp. 366–372.

Berry, K. P. and E. Nedivi (2017). "Spine dynamics: are they all the same?" In: *Neuron* 96.1, pp. 43–55.

Beul, S. F., H. Barbas, and C. C. Hilgetag (2017). "A predictive structural model of the primate connectome." In: *Sci. Rep.* 7.43176, pp. 1–12.

Beul, S. F. and C. C. Hilgetag (2015). "Towards a 'canonical' agranular cortical microcircuit." In: *Front. Neuroanat.* 8, p. 165.

– (2019). "Neuron density fundamentally relates to architecture and connectivity of the primate cerebral cortex." In: *Neuroimage* 189, pp. 777–792.

Bezaire, M. J., I. Raikov, K. Burk, D. Vyas, and I. Soltesz (2016). "Interneuronal mechanisms of hippocampal theta oscillations in a full-scale model of the rodent CA1 circuit." In: *eLife* 5.

Billeh, Y. N. et al. (2020). "Systematic integration of structural and functional data into multi-scale models of mouse primary visual cortex." In: *Neuron* 106.3, 388–403.e18.

Binzegger, T., R. J. Douglas, and K. A. C. Martin (2004). "A Quantitative Map of the Circuit of Cat Primary Visual Cortex." In: *J. Neurosci.* 39.24, pp. 8441–8453.

Biswal, B., F. Zerrin Yetkin, V. M. Haughton, and J. S. Hyde (1995). "Functional connectivity in the motor cortex of resting human brain using echo-planar MRI." In: *Magn. Reson. Med.* 34.4, pp. 537–541.

Boltzmann, L. (1896). "Entgegnung auf die wärmetheoretischen Betrachtungen des Hrn. E. Zermelo." In: *Ann. Phys.* 293.4, pp. 773–784.

Bos, H., M. Diesmann, and M. Helias (2016). "Identifying Anatomical Origins of Coexisting Oscillations in the Cortical Microcircuit." In: *PLOS Comput. Biol.* 12.10, e1005132.

Braitenberg, V. and A. Schüz (1991). *Anatomy of the Cortex: Statistics and Geometry*. Berlin, Heidelberg, New York: Springer-Verlag.

Braitenberg, V. and A. Schüz (1998). *Cortex: Statistics and Geometry of Neuronal Connectivity*. 2nd. Berlin: Springer-Verlag.

Braumann, C. A. (2007). "Itô versus Stratonovich calculus in random population growth." In: *Math. Biosci.* 206.1, pp. 81–107.

Brecht, M., M. Schneider, B. Sakmann, and T. W. Margrie (2004). "Whisker movements evoked by stimulation of single pyramidal cells in rat motor cortex." In: *Nature* 427.6976, pp. 704–710.

Bressloff, P. C. (2012). "Spatiotemporal dynamics of continuum neural fields." In: *J. Phys. A Math. Theor.* 45.3, p. 033001.

Bressloff, P. C., J. D. Cowan, M. Golubitsky, P. J. Thomas, and M. C. Wiener (2001). "Geometric visual hallucinations, Euclidean symmetry and the functional architecture of striate cortex." In: *Philos. Trans. R. Soc. B* 356.1407, pp. 299–330.

Brette, R. and A. Destexhe (2012). *Handbook of neural activity measurement*. Cambridge University Press.

Brette, R. and W. Gerstner (2005). "Adaptive Exponential Integrate-and-Fire Model as an Effective Description of Neuronal Activity." In: *J. Neurophysiol.* 94.5, pp. 3637–3642.

Brincat, S. L., M. Siegel, C. von Nicolai, and E. K. Miller (2018). "Gradual progression from sensory to task-related processing in cerebral cortex." In: *Proc. Natl. Acad. Sci. USA* 115.30, E7202–E7211.

Brodmann, K. (1909). *Vergleichende Lokalisationslehre der Großhirnrinde in ihren Prinzipien dargestellt auf Grund des Zellenbaues*. Leipzig: Johann Ambrosius Barth.

Brunel, N. (2000). "Dynamics of sparsely connected networks of excitatory and inhibitory spiking neurons." In: *J. Comput. Neurosci.* 8.3, pp. 183–208.

Brunel, N., F. S. Chance, N. Fourcaud, and L. F. Abbott (2001). "Effects of Synaptic Noise and Filtering on the Frequency Response of Spiking Neurons." In: *Phys. Rev. Lett.* 86.10, pp. 2186–2189.

Brunel, N. and V. Hakim (1999). "Fast Global Oscillations in Networks of Integrate-and-Fire Neurons with Low Firing Rates." In: *Neural Comput.* 11.7, pp. 1621–1671.

Brunel, N. and P. Latham (2003). "Firing rate of the noisy quadratic integrate-and-fire neuron." In: *Neural Comput.* 15.10, pp. 2281–2306.

Brunel, N. and M. C. W. van Rossum (2007). "Lapicque's 1907 paper: from frogs to integrate-and-fire." In: *Biol. Cybern.* 97, pp. 337–339.

Buckley, C. L., C. S. Kim, S. Mcgregor, and K. Anil (2017). "The free energy principle for action and perception: A mathematical review." In: *ArXiv*.

Buckner, R. L. and L. M. DiNicola (2019). "The brain's default network: updated anatomy, physiology and evolving insights." In: *Nat. Rev. Neurosci.* 20.10, pp. 593–608.

Buice, M. A. and C. C. Chow (2013). "Beyond mean field theory: statistical field theory for neural networks." In: *J. Stat. Mech. Theory Exp.* 2013.03, P03003.

Buzsáki, G. and K. Mizuseki (2014). "The log-dynamic brain: how skewed distributions affect network operations." In: *Nat. Rev. Neurosci.* 15, pp. 264–278.

Cadena, S. A., G. H. Denfield, E. Y. Walker, L. A. Gatys, A. S. Tolias, M. Bethge, and A. S. Ecker (2019). "Deep convolutional models improve predictions of macaque V1 responses to natural images." In: *PLOS Comput. Biol.* 15.4, e1006897.

Cadwell, C. R., A. Bhaduri, M. A. Mostajo-Radji, M. G. Keefe, and T. J. Nowakowski (2019). "Development and Arealization of the Cerebral Cortex." In: *Neuron* 103.6, pp. 980–1004.

Cain, N., R. Iyer, C. Koch, and S. Mihalas (2016). "The Computational Properties of a Simplified Cortical Column Model." In: *PLOS Comput. Biol.* 12.9, e1005045.

Cano-Astorga, N., J. DeFelipe, and L. Alonso-Nanclares (2021). "Three-Dimensional Synaptic Organization of Layer III of the Human Temporal Neocortex." In: *Cereb. Cortex* 31.10, pp. 4742–4764.

Cavanagh, S. E., L. T. Hunt, and S. W. Kennerley (2020). "A diversity of intrinsic timescales underlie neural computations." In: *Front. Neural Circuits* 14, p. 81.

Charvet, C. J., D. J. Cahalane, and B. L. Finlay (2015). "Systematic, cross-cortex variation in neuron numbers in rodents and primates." In: *Cereb. Cortex* 25.1, pp. 147–160.

Chaudhuri, R., A. Bernacchia, and X.-J. Wang (2014). "A diversity of localized timescales in network activity." In: *eLife* 3, e01239.

Chaudhuri, R., K. Knoblauch, M.-A. Gariel, H. Kennedy, and X.-J. Wang (2015). "A Large-Scale Circuit Mechanism for Hierarchical Dynamical Processing in the Primate Cortex." In: *Neuron* 88.2, pp. 419–431.

Chen, G., F. Scherr, and W. Maass (2021). "Analysis of visual processing capabilities and neural coding strategies of a detailed model for laminar cortical microcircuits in mouse V1." In: *BioRxiv*.

Chen, M., J. Pennington, and S. Schoenholz (2018). "Dynamical Isometry and a Mean Field Theory of RNNs: Gating Enables Signal Propagation in Recurrent Neural Networks." In: *Proceedings of the 35th International Conference on Machine Learning*. Ed. by J. Dy and A. Krause. Vol. 80. Proceedings of Machine Learning Research. PMLR, pp. 873–882.

Chen, R. T., Y. Rubanova, J. Bettencourt, and D. K. Duvenaud (2018). "Neural ordinary differential equations." In: *Advances in neural information processing systems*, pp. 6571–6583.

Cho, Y. and L. Saul (2009). "Kernel Methods for Deep Learning." In: *Adv. Neural Inf. Process. Syst.* Ed. by Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, and A. Culotta. Vol. 22. Curran Associates, Inc.

Chow, C. and M. Buice (2015). "Path Integral Methods for Stochastic Differential Equations." In: *J. Math. Neurosci.* 5, p. 8.

Collins, C. E., D. C. Airey, N. A. Young, D. B. Leitch, and J. H. Kaas (2010). "Neuron densities vary across and within cortical areas in primates." In: *Proc. Natl. Acad. Sci. USA* 107.36, pp. 15927–15932.

Collins, C. E., E. C. Turner, E. K. Sawyer, J. L. Reed, N. A. Young, D. K. Flaherty, and J. H. Kaas (2016). "Cortical cell and neuron density estimates in one chimpanzee hemisphere." In: *Proc. Natl. Acad. Sci. USA* 113.3, pp. 740–745.

Coolen, A. (2001). "Chapter 15 Statistical mechanics of recurrent neural networks II — Dynamics." In: *Neuro-Informatics and Neural Modelling*. Ed. by F. Moss and S. Gielen. Vol. 4. Handbook of Biological Physics. North-Holland, pp. 619–684.

Coombes, S. (2005). "Waves, bumps, and patterns in neural field theories." In: *Biol. Cybern.* 93, pp. 91–108.

Coombes, S., P. bei Graben, R. Potthast, and J. Wright (2014). *Neural Fields. Theory and Applications*. Springer-Verlag Berlin Heidelberg.

Corless, R. M., G. H. Gonnet, D. E. G. Hare, D. J. Jeffrey, and D. E. Knuth (1996). "On the Lambert W function." In: *Adv. Comput. Math.* 5.1, pp. 329–359.

Cramér, H. (1936). "Über eine Eigenschaft der normalen Verteilungsfunktion." In: *Math. Z.* 41.1, pp. 405–414.

Crisanti, A. and H. Sompolinsky (2018). "Path integral approach to random neural networks." In: *Phys. Rev. E* 98 (6), p. 062120.

Dąbrowska, P. A., N. Voges, M. von Papen, J. Ito, D. Dahmen, A. Riehle, T. Brochier, and S. Grün (2021). "On the Complexity of Resting State Spiking Activity in Monkey Motor Cortex." In: *Cereb. Cortex Commun.* 2.3. tgab033.

Dahmen, D., S. Grün, M. Diesmann, and M. Helias (2019). "Second type of criticality in the brain uncovers rich multiple-neuron dynamics." In: *Proc. Natl. Acad. Sci. USA* 116 (26), pp. 13051–13060.

Dahmen, D., M. Layer, et al. (2021). "Global organization of neuronal activity only requires unstructured local connectivity." In: *BioRxiv*, pp. 2020–07.

– (2022). "Global organization of neuronal activity only requires unstructured local connectivity." In: *eLife* 11, e68422.

Dahmen, D., S. Recanatesi, G. K. Ocker, X. Jia, M. Helias, and E. Shea-Brown (2020). "Strong coupling and local control of dimensionality across brain areas." In: *BioRxiv*.

Dasbach, S., T. Tetzlaff, M. Diesmann, and J. Senk (2021). "Prominent characteristics of recurrent neuronal networks are robust against low synaptic weight resolution." In: *ArXiv*, 2105.05002 [q–bio.NC].

Deco, G. and V. K. Jirsa (2012). "Ongoing Cortical Activity at Rest: Criticality, Multistability, and Ghost Attractors." In: *J. Neurosci.* 32.10, pp. 3366–3375.

Deco, G., V. K. Jirsa, and A. R. McIntosh (2011). "Emerging concepts for the dynamical organization of resting-state activity in the brain." In: *Nat. Rev. Neurosci.* 12, pp. 43–56.

DeFelipe, J. et al. (2002). "Neocortical circuits: Evolutionary aspects and specificity versus non-specificity of synaptic connections. Remarks, main conclusions and general comments and discussion." In: *J. Neurocytol.* 32, pp. 387–416.

DeFelipe, J., L. Alonso-Nanclares, and J. Arellano (2002). "Microstructure of the neocortex: comparative aspects." In: *J. Neurocytol.* 31, pp. 299–316.

DeFelipe, J. (2015). "The dendritic spine story: an intriguing process of discovery." In: *Front. Neuroanat.* 9, p. 14.

Delettre, C. et al. (2019). "Comparison between diffusion MRI tractography and histological tract-tracing of cortico-cortical structural connectivity in the ferret brain." In: *Netw. Neurosci.* 3.4, pp. 1038–1050.

Dembo, A. and O. Zeitouni (2010). *Large Deviations Techniques and Applications*. Springer Berlin Heidelberg.

Desikan, R. S. et al. (2006). "An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest." In: *Neuroimage* 31.3, pp. 968–980.

Destexhe, A., M. Rudolph, and D. Pare (2003). "The high-conductance state of neocortical neurons in vivo." In: *Nat. Rev. Neurosci.* 4, pp. 739–751.

Donahue, C. J. et al. (2016). "Using Diffusion Tractography to Predict Cortical Connection Strength and Distance: A Quantitative Comparison with Tracers in the Monkey." In: *J. Neurosci.* 36.25, pp. 6758–6770.

Dong, H. W. (2008). *The Allen reference atlas: A digital color brain atlas of the C57Bl/6J male mouse*. John Wiley & Sons Inc.

Douglas, R. J. and K. A. C. Martin (2004). "Neuronal Circuits of the Neocortex." In: *Annu. Rev. Neurosci.* 27, pp. 419–451.

Douglas, R. J., K. A. C. Martin, and D. Whitteridge (1989). "A Canonical Microcircuit for Neocortex." In: *Neural Comput.* 1.4, pp. 480–488.

Duarte, R. and A. Morrison (2019). "Leveraging heterogeneity for neural computation with fading memory in layer 2/3 cortical microcircuits." In: *PLOS Comput. Biol.* 15.4, e1006781.

Duarte, R., A. Seeholzer, K. Zilles, and A. Morrison (2017). "Synaptic patterning and the timescales of cortical dynamics." In: *Curr. Opin. Neurobiol.* 43, pp. 156–165.

Dummer, B., S. Wieland, and B. Lindner (2014). "Self-consistent determination of the spike-train power spectrum in a neural network with sparse connectivity." In: *Front. Comput. Neurosci.* 8, p. 104.

Dyer, E. and G. Gur-Ari (2020). "Asymptotics of Wide Networks from Feynman Diagrams." In: *International Conference on Learning Representations*.

Ecker, A. et al. (2020). "Data-driven integration of hippocampal CA1 synaptic physiology in silico." In: *Hippocampus* 30.11, pp. 1129–1145.

Einevoll, G. T. et al. (2019). "The Scientific Case for Brain Simulations." In: *Neuron* 102.4, pp. 735–744.

Elston, G. N. and M. Rosa (1998). "Morphological variation of layer III pyramidal neurones in the occipitotemporal pathway of the macaque monkey visual cortex." In: *Cereb. Cortex* 8.3, pp. 278–294.

Ercsey-Ravasz, M., N. T. Markov, C. Lamy, D. C. V. Essen, K. Knoblauch, Z. Toroczkai, and H. Kennedy (2013). "A Predictive Network Model of Cerebral Cortical Connectivity Based on a Distance Rule." In: *Neuron* 80.1, pp. 184–197.

Erdős, P. and A. Rényi (1959). "On random graphs." In: *Publ. Math.* 6, pp. 290–297.

Ermentrout, G. B. and J. D. Cowan (1979). "A mathematical theory of visual hallucination patterns." In: *Biol. Cybern.* 34, pp. 137–150.

Erö, C., M.-O. Gewaltig, D. Keller, and H. Markram (2018). "A cell atlas for the mouse brain." In: *Front. Neuroinform.* 12, p. 84.

Faugeras, O. and J. MacLaurin (2015). "Asymptotic Description of Neural Networks with Correlated Synaptic Weights." In: *Entropy* 17.7, pp. 4701–4743.

Faugeras, O., J. MacLaurin, and E. Tanré (2019). "The meanfield limit of a network of Hopfield neurons with correlated synaptic weights." In: *ArXiv*.

Feynman, R. P., A. R. Hibbs, and D. F. Styer (2010). *Quantum Mechanics and Path Integrals: Emended Edition*. Dover Publications.

Fields, R. D. et al. (2014). "Glial biology in learning and cognition." In: *Neuroscientist* 20.5, pp. 426–431.

Fischer, K. and J. Hertz (1991). *Spin glasses*. Cambridge University Press.

Fourcaud, N. and N. Brunel (2002). "Dynamics of the firing probability of noisy integrate-and-fire neurons." In: *Neural Comput.* 14, pp. 2057–2110.

Fourcaud-Trocmé, N., D. Hansel, C. van Vreeswijk, and N. Brunel (2003). "How spike generation mechanisms determine the neuronal response to fluctuating inputs." In: *J. Neurosci.* 23, pp. 11628–11640.

Friston, K. (2010). "The free-energy principle: a unified brain theory?" In: *Nat. Rev. Neurosci.* 11, pp. 127–138.

Fulvi Mari, C. (2000). "Random networks of spiking neurons: instability in the Xenopus tadpole moto-neuron pattern." In: *Phys. Rev. Lett.* 85, pp. 210–213.

Gămănuț, R., H. Kennedy, Z. Toroczkai, M. Ercsey-Ravasz, D. C. Van Essen, K. Knoblauch, and A. Burkhalter (2018). "The mouse cortical connectome, characterized by an ultra-dense cortical graph, maintains specificity by distinct connectivity profiles." In: *Neuron* 97.3, pp. 698–715.

Ganguli, S., D. Huh, and H. Sompolinsky (2008). "Memory traces in dynamical systems." In: *Proc. Natl. Acad. Sci. USA* 105.48, pp. 18970–18975.

Gao, R., R. L. van den Brink, T. Pfeffer, and B. Voytek (2020). "Neuronal timescales are functionally dynamic and shaped by cortical microarchitecture." In: *eLife* 9, e61277.

Garagnani, M., G. Lucchese, R. Tomasello, T. Wennekers, and F. Pulvermüller (2017). "A Spiking Neurocomputational Model of High-Frequency Oscillatory Brain Responses to Words and Pseudowords." In: *Front. Comput. Neurosci.* 10, p. 145.

Gardiner, C. (2009). *Stochastic Methods: A Handbook for the Natural and Social Sciences*. 4th. Berlin, Heidelberg: Springer.

Gardner, E. (1988). "The space of interactions in neural network models." In: *J. Phys. A Math. Gen.* 21.1, p. 257.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin (2014). *Bayesian data analysis*. Vol. 2. Chapman & Hall/CRC Boca Raton, FL, USA.

Gerstein, G. L. and B. Mandelbrot (1964). "Random walk models for the spike activity of a single neuron." In: *Biomed. Pharmacol. J.* 71, pp. 41–68.

Gerstner, W., W. M. Kistler, R. Naud, and L. Paninski (2014). *Neuronal Dynamics. From Single Neurons to Networks and Models of Cognition*. Cambridge: Cambridge University Press.

Gerstner, W., H. Sprekeler, and G. Deco (2012). "Theory and simulation in neuroscience." In: *Science* 338.6103, pp. 60–65.

Gewaltig, M.-O. and M. Diesmann (2007). "NEST (NEural Simulation Tool)." In: *Scholarpedia J.* 2.4, p. 1430.

Giese, M. A. (2012). *Dynamic neural field theory for motion perception*. Vol. 469. Springer Science & Business Media.

Gilbert, E. N. (1959). "Random Graphs." In: *Ann. Math. Stat.* 30.4, pp. 1141–1144.

Ginzburg, I. and H. Sompolinsky (1994). "Theory of correlations in stochastic neural networks." In: *Phys. Rev. E* 50 (4), pp. 3171–3191.

Girard, P., J. M. Hupé, and J. Bullier (2001). "Feedforward and Feedback Connections Between Areas V1 and V2 of the Monkey Have Similar Rapid Conduction Velocities." In: *J. Neurophysiol.* 85.3, pp. 1328–1331.

Girardi-Schappo, M., G. S. Bortolotto, J. J. Gonsalves, L. T. Pinto, and M. H. R. Tragtenberg (2016). "Griffiths phase and long-range correlations in a biologically motivated visual cortex model." In: *Sci. Rep.* 6.1.

Glorot, X. and Y. Bengio (2010). "Understanding the difficulty of training deep feedforward neural networks." In: *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*. Ed. by Y. W. Teh and M. Titterington. Vol. 9. Proceedings of Machine Learning Research. Chia Laguna Resort, Sardinia, Italy: PMLR, pp. 249–256.

Goldenfeld, N. (1992). *Lectures on phase transitions and the renormalization group*. Reading, Massachusetts: Perseus books.

Golesorkhi, M., J. Gomez-Pilar, F. Zilio, N. Berberian, A. Wolff, M. C. Yagoub, and G. Northoff (2021). "The brain and its time: intrinsic neural timescales are key for input processing." In: *Commun. Biol.* 4.1, pp. 1–16.

Golgi, C. (1873). "Sulla struttura della sostanza grigia del cervello." In: *Gaz. Med. Ital. Lomb.* 33, pp. 244–46.

Golosio, B., G. Tiddia, C. D. Luca, E. Pastorelli, F. Simula, and P. S. Paolucci (2021). "Fast Simulations of Highly-Connected Spiking Cortical Models Using GPUs." In: *Front. Comput. Neurosci.* 15, p. 627620.

Goulas, A., R. Werner, S. F. Beul, D. Säring, M. v. d. Heuvel, L. C. Triarhou, and C. C. Hilgetag (2016). "Cytoarchitectonic similarity is a wiring principle of the human connectome." In: *BioRxiv*.

Goulas, A., K. Zilles, and C. C. Hilgetag (2018). "Cortical Gradients and Laminar Projections in Mammals." In: *Trends Neurosci.* 41.11, pp. 775–788.

Gouwens, N. W., J. Berg, D. Feng, S. A. Sorensen, H. Zeng, M. J. Hawrylycz, C. Koch, and A. Arkhipov (2018). "Systematic generation of biophysically detailed models for diverse cortical neuron types." In: *Nat. Commun.* 9.1, p. 710.

Grabska-Barwinska, A. and P. Latham (2014). "How well do mean field theories of spiking quadratic-integrate-and-fire networks work in realistic parameter regimes?" In: *J. Comput. Neurosci.* 36.3, pp. 469–81.

Gray, E. G. (1959a). "Electron microscopy of synaptic contacts on dendrite spines of the cerebral cortex." In: *Nature* 183, pp. 1592–1593.

Gray, E. G. (1959b). "Axo-somatic and axo-dendritic synapses of the cerebral cortex: an electron microscope study." In: *J. Anat.* 93.Pt 4, p. 420.

Gretton, A., K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola (2012). "A Kernel Two-Sample Test." In: *J. Mach. Learn. Res.* 13.25, pp. 723–773.

Grigera, T. S. (2020). "Everything you wish to know about correlations but are afraid to ask." In: *ArXiv*, p. 2002.01750.

Grigg, O. and C. L. Grady (2010). "Task-Related Effects on the Temporal and Spatial Dynamics of Resting-State Functional Connectivity in the Default Network." In: *PLOSONE* 5.10, pp. 1–12.

Grosvenor, K. T. and R. Jefferson (2022). "The edge of chaos: quantum field theory and deep neural networks." In: *SciPost Phys.* 12 (3), p. 81.

Grytskyy, D., T. Tetzlaff, M. Diesmann, and M. Helias (2013). "A unified view on weakly correlated recurrent networks." In: *Front. Comput. Neurosci.* 7, p. 131.

Guionnet, A. (1997). "Averaged and quenched propagation of chaos for spin glass dynamics." In: *Probab. Theory Relat. Fields* 109, pp. 183–215.

Hagen, E., D. Dahmen, M. L. Stavrinou, H. Lindén, T. Tetzlaff, S. J. van Albada, S. Grün, M. Diesmann, and G. T. Einevoll (2016). "Hybrid scheme for modeling local field potentials from point-neuron networks." In: *Cereb. Cortex* 26.12, pp. 4461–4496.

Halverson, J., A. Maiti, and K. Stoner (2021). "Neural networks and quantum field theory." In: *Machine Learning: Science and Technology* 2.3, p. 035002.

Hänggi, P. and P. Jung (1995). "Colored noise in dynamical systems." In: *Adv. Chem. Phys.* 89, pp. 239–326.

Hannun, A. et al. (2014). "Deep speech: Scaling up end-to-end speech recognition." In: *ArXiv*.

Harish, O. and D. Hansel (2015). "Asynchronous Rate Chaos in Spiking Neuronal Circuits." In: *PLOS Comput. Biol.* 11.7, e1004266.

Harris, C. R., K. J. Millman, et al. (2020). "Array programming with NumPy." In: *Nature* 585, pp. 357–362.

Harris, J. J., R. Jolivet, and D. Attwell (2012). "Synaptic Energy Use and Supply." In: *Neuron* 75.5, pp. 762–777.

Harris, J. A., S. Mihalas, et al. (2019). "Hierarchical organization of cortical and thalamic connectivity." In: *Nature* 575.7781, pp. 195–202.

Harris, K. D. and G. M. G. Shepherd (2015). "The neocortical circuit: themes and variations." In: *Nat. Neurosci.* 18.2, pp. 170–181.

Harris, K. D. and A. Thiele (2011). "Cortical state and attention." In: *Nat. Rev. Neurosci.* 12, pp. 509–523.

Hass, J., L. Hertäg, and D. Durstewitz (2016). "A detailed data-driven network model of prefrontal cortex reproduces key features of in vivo activity." In: *PLOS Comput. Biol.* 12.5, e1004930.

He, K., X. Zhang, S. Ren, and J. Sun (2015). "Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification." In: *ArXiv*, p. 1502.01852.

– (2016). "Deep Residual Learning for Image Recognition." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Helias, M. and D. Dahmen (2020). *Statistical Field Theory for Neural Networks*. Springer International Publishing, p. 203.

Helias, M., T. Tetzlaff, and M. Diesmann (2014). "The correlation structure of local cortical networks intrinsically results from recurrent dynamics." In: *PLOS Comput. Biol.* 10.1, e1003428.

Hendrickson, P. J., G. J. Yu, B. S. Robinson, D. Song, and T. W. Berger (2012). "Towards a large-scale biologically realistic model of the hippocampus." In: *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, pp. 4595–4598.

Hennequin, G., T. Vogels, and W. Gerstner (2012). "Non-normal amplification in random balanced neuronal networks." In: *Phys. Rev. E* 86, p. 011909.

Herculano-Houzel, S. (2009). "The human brain in numbers: a linearly scaled-up primate brain." In: *Front. Hum. Neurosci.* 3, p. 31.

Herculano-Houzel, S., B. Mota, P. Wong, and J. H. Kaas (2010). "Connectivity-driven white matter scaling and folding in primate cerebral cortex." In: *Proc. Natl. Acad. Sci. USA* 107.44, pp. 19008–19013.

Herculano-Houzel, S., C. Watson, and G. Paxinos (2013). "Distribution of neurons in functional areas of the mouse cerebral cortex reveals quantitatively different cortical zones." In: *Front. Neuroanat.* 7, p. 35.

Hertz, J., A. Lerchner, and M. Ahmadi (2004). "Computational Neuroscience: Cortical Dynamics." In: vol. 3146. LNCS. Springer-Verlag Berlin Heidelberg. Chap. Mean field methods for cortical dynamics, pp. 71–89.

Hertz, J. (2010). "Cross-Correlations in High-Conductance States of a Model Cortical Network." In: *Neural Comput.* 22, pp. 427–447.

Hertz, J., A. Krogh, and R. G. Palmer (1991). *Introduction to the Theory of Neural Computation*. Cambridge, MA, USA: Perseus Books.

Hertz, J. A., Y. Roudi, and P. Sollich (2017). "Path integral methods for the dynamics of stochastic and disordered systems." In: *J. Phys. A* 50.3, p. 033001.

Herz, A. V. M., T. Gollisch, C. K. Machens, and D. Jaeger (2006). "Modeling Single-Neuron Dynamics and Computations: A Balance of Detail and Abstraction." In: *Science* 314, pp. 80–84.

Hilgetag, C. C., S. F. Beul, S. J. van Albada, and A. Goulas (2019). "An Architectonic Type Principle Integrates Macroscopic Cortico-Cortical Connections with Intrinsic Cortical Circuits of the Primate Brain." In: *Netw. Neurosci.* 3.4, pp. 905–923.

Hinton, G. E., S. Osindero, and Y.-W. Teh (2006). "A fast learning algorithm for deep belief nets." In: *Neural Comput.* 18.7, pp. 1527–1554.

Hodge, R. D. et al. (2019). "Conserved cell types with divergent features in human versus mouse cortex." In: *Nature* 573.7772, pp. 61–68.

Hodgkin, A. L. and A. F. Huxley (1952). "A Quantitative Description of Membrane Current and Its Application to Conduction and Excitation in Nerve." In: *J. Physiol.* 117, pp. 500–544.

Honey, C. J. et al. (2012). "Slow Cortical Dynamics and the Accumulation of Information over Long Timescales." In: *Neuron* 76.2, pp. 423–434.

Hopfield, J. J. (1982). "Neural networks and physical systems with emergent collective computational abilities." In: *Proc. Natl. Acad. Sci. USA* 79, pp. 2554–2558.

Horvát, S. et al. (2016). "Spatial Embedding and Wiring Cost Constrain the Functional Layout of the Cortical Network of Rodents and Primates." In: *PLOS Biol.* 14.7, pp. 1–30.

Houweling, A. and M. Brecht (2008). "Behavioural report of single neuron stimulation in somatosensory cortex." In: *Nature* 451.7174, pp. 65–68.

Huang, C. and B. Doiron (2017). "Once upon a (slow) time in the land of recurrent neuronal networks..." In: *Curr. Opin. Neurobiol.* 46, pp. 31–38.

Huang, J. and H.-T. Yau (2020). "Dynamics of deep neural networks and neural tangent hierarchy." In: *International Conference on Machine Learning*. PMLR, pp. 4542–4551.

Hubel, D. H. and T. N. Wiesel (1962). "Receptive Fields, Binocular Interaction, and Functional Architecture in the Cat's Visual Cortex." In: *J. Physiol.* 160, pp. 106–154.

Huntenburg, J. M., P.-L. Bazin, and D. S. Margulies (2018). "Large-scale gradients in human cortical organization." In: *Trends. Cogn. Sci.* 22.1, pp. 21–31.

Hunter, J. D. (2007). "Matplotlib: A 2D Graphics Environment." In: *Comput. Sci. Eng.* 9.3, pp. 90–95.

Izhikevich, E. (2003). "Simple model of spiking neurons." In: *IEEE Trans. Neural Netw.* 14.6, pp. 1569–1572.

Izhikevich, E. M. (2007). *Dynamical Systems in Neuroscience: The Geometry of Excitability and Bursting*. MIT Press.

Izhikevich, E. M. and G. M. Edelman (2008). "Large-scale model of mammalian thalamocortical systems." In: *Proc. Natl. Acad. Sci. USA* 105.9, pp. 3593–3598.

Janssen, H.-K. (1976). "On a Lagrangean for classical field dynamics and renormalization group calculations of dynamical critical properties." In: *Z. Phys. B* 23.4, pp. 377–380.

Jardim-Messeder, D. et al. (2017). "Dogs have the most neurons, though not the largest brain: trade-off between body mass and number of neurons in the cerebral cortex of large carnivoran species." In: *Front. Neuroanat.*, p. 118.

Jaynes, E. T. (2003). *Probability theory: The logic of science*. Cambridge university press.

Jbabdi, S., S. N. Sotiropoulos, S. N. Haber, D. C. Van Essen, and T. E. Behrens (2015). "Measuring macroscopic brain connections in vivo." In: *Nat. Neurosci.* 18.11, p. 1546.

Jirsa, V., T. Proix, et al. (2017). "The Virtual Epileptic Patient: Individualized whole-brain models of epilepsy spread." In: *Neuroimage* 145, pp. 377–388.

Jirsa, V. K. and H. Haken (1996). "Field theory of electromagnetic brain activity." In: *Phys. Rev. Lett.* 77.5, p. 960.

– (1997). "A derivation of a macroscopic field theory of the brain from the quasi-microscopic neural dynamics." In: *Physica D* 99.4, pp. 503–526.

Joglekar, M. R., J. F. Mejias, G. R. Yang, and X.-J. Wang (2018). "Inter-areal balanced amplification enhances signal propagation in a large-scale circuit model of the primate cortex." In: *Neuron* 98.1, pp. 222–234.

Johnson, R. R. and A. Burkhalter (1996). "Microcircuitry of Forward and Feedback Connections Within Rat Visual Cortex." In: *J. Comp. Neurol.* 368, pp. 383–398.

Jordan, J., T. Ippen, M. Helias, I. Kitayama, M. Sato, J. Igarashi, M. Diesmann, and S. Kunkel (2018). "Extremely Scalable Spiking Neuronal Network Simulation Code: From Laptops to Exascale Computers." In: *Front. Neuroinform.* 12, p. 2.

Kabbara, A., W. EL Falou, M. Khalil, F. Wendling, and M. Hassan (2017). "The dynamic functional core network of the human brain at rest." In: *Sci. Rep.* 7.1, p. 2936.

Kadmon, J. and H. Sompolinsky (2015). "Transition to Chaos in Random Neuronal Networks." In: *Phys. Rev. X* 5 (4), p. 041030.

Kandel, E. R., J. H. Schwartz, and T. M. Jessel (2000). *Principles of Neural Science*. 4th ed. ISBN 978-0838577011. New York: McGraw-Hill.

Kandel, E. R., J. H. Schwartz, T. M. Jessell, S. A. Siegelbaum, A. Hudspeth, and S. Mack (2013). *Principles of Neural Science*. 5th ed. New York: McGraw-Hill.

Kardar, M. (2007a). *Statistical Physics of Fields*. Cambridge, England: Cambridge University Press.

– (2007b). *Statistical Physics of Particles*. Cambridge, England: Cambridge University Press.

Keup, C., T. Kühn, D. Dahmen, and M. Helias (2021). "Transient Chaotic Dimensionality Expansion by Recurrent Networks." In: *Phys. Rev. X* 11.2.

Kim, R. and T. J. Sejnowski (2021). "Strong inhibitory signaling underlies stable temporal dynamics and working memory in spiking neural networks." In: *Nat. Neurosci.* 24.1, pp. 129–139.

Kleinert, H. (2009). *Path Integrals in Quantum Mechanics, Statistics, Polymer Physics, and Financial Markets*. 5th ed. World Scientific.

Knight, J. C. and T. Nowotny (2018). "GPUs Outperform Current HPC and Neuromorphic Solutions in Terms of Speed and Energy When Simulating a Highly-Connected Cortical Model." In: *Front. Neurosci.* 12, pp. 1–19.

Köster, J. and S. Rahmann (2012). "Snakemake—a scalable bioinformatics workflow engine." In: *Biochemistry* 28.19, pp. 2520–2522.

Kriener, B., H. Enger, T. Tetzlaff, H. E. Plesser, M.-O. Gewaltig, and G. T. Einevoll (2014). "Dynamics of self-sustained asynchronous-irregular activity in random networks of spiking neurons with strong synapses." In: *Front. Comput. Neurosci.* 8, p. 136.

Krizhevsky, A., I. Sutskever, and G. E. Hinton (2012). "Imagenet classification with deep convolutional neural networks." In: *Adv. Neural Inf. Process. Syst.* Ed. by F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger. Vol. 25. Curran Associates, Inc., pp. 1097–1105.

Krumin, M. and S. Shoham (2009). "Generation of Spike Trains with Controlled Auto- and Cross-Correlation Functions." In: *Neural Comput.* 21.6, pp. 1642–1664.

Kuhn, T. S. (2012). *The structure of scientific revolutions*. Chicago, IL: University of Chicago Press.

Kumar, S., K. E. Stephan, J. D. Warren, K. J. Friston, and T. D. Griffiths (2007). "Hierarchical processing of auditory objects in humans." In: *PLOS Comput. Biol.* 3.6, e100.

Kuznetsov, P. I. and R. L. Stratonovich (1965). "A note on the mathematical theory of correlated random points." In: *Non-Linear Transformations of Stochastic Processes*. Elsevier, pp. 101–115.

Laing, C. R. and W. C. Troy (2003). "Two-bump solutions of Amari-type models of neuronal pattern formation." In: *Physica D* 178.3-4, pp. 190–218.

Laing, C. R., W. C. Troy, B. Gutkin, and B. G. Ermentrout (2002). "Multiple Bumps in a Neuronal Model of Working Memory." In: *SIAM J. Appl. Math.* 63, pp. 62–97.

Langevin, P. (1908). "Sur la théorie du mouvement brownien." In: *Comptes-rendus de l'Académie des sciences* 146, pp. 530–533.

Lapicque, L. (1907). "Recherches quantitatives sur l'excitation electrique des nerfs traitee comme une polarization." In: *J. Physiol. Pathol. Gen.* 9, pp. 620–635.

Larkum, M. E., J. Wu, S. A. Duverdin, and A. Gidon (2022). "The Guide to Dendritic Spikes of the Mammalian Cortex In Vitro and In Vivo." In: *Neuroscience* 489, pp. 15–33.

Laughlin, S. B. and T. J. Sejnowski (2003). "Communication in neuronal networks." In: *Science* 301, pp. 1870–1874.

Lawson, C. L. and R. J. Hanson (1995). *Solving Least Squares Problems*. SIAM.

Layer, M., J. Senk, S. Essink, A. van Meegen, H. Bos, and M. Helias (in press). "NNMT: Mean-field based analysis tools for neuronal network models." In: *Front. Neuroinform.*

LeCun, Y., Y. Bengio, and G. Hinton (2015). "Deep learning." In: *Nature* 521.7553, pp. 436–444.

Lee, J., J. Sohl-Dickstein, J. Pennington, R. Novak, S. Schoenholz, and Y. Bahri (2018). "Deep Neural Networks as Gaussian Processes." In: *International Conference on Learning Representations*.

Lein, E. S. et al. (2007). "Genome-wide atlas of gene expression in the adult mouse brain." In: *Nature* 445.7124, pp. 168–176.

Lerchner, A., G. Sterner, J. Hertz, and M. Ahmadi (2006). "Mean field theory for a balanced hypercolumn model of orientation selectivity in primary visual cortex." In: *Netw. Comput. Neural Syst.* 17, pp. 131–150.

Lerchner, A., C. Ursta, J. Hertz, M. Ahmadi, P. Ruffiot, and S. Enemark (2006). "Response variability in balanced cortical networks." In: *Neural Comput.* 18.3, pp. 634–659.

Liewald, D., R. Miller, N. Logothetis, H.-J. Wagner, and A. Schüz (2014). "Distribution of axon diameters in cortical white matter: an

electron-microscopic study on three human brains and a macaque." In: *Biol. Cybern.* 108.5, pp. 541–557.

Lindner, B. (2004). "Interspike interval statistics of neurons driven by colored noise." In: *Phys. Rev. E* 69, pp. 0229011–0229014.

Lindner, B., B. Doiron, and A. Longtin (2005). "Theory of oscillatory firing induced by spatially correlated noise and delayed inhibitory feedback." In: *Phys. Rev. E* 72.6, p. 061919.

Lindner, B. and A. Longtin (2005). "Effect of an exponentially decaying threshold on the firing statistis of a stochastic integate-and-fire neuron." In: *J. Theor. Biol.* 232, pp. 505–521.

Lindner, B. and L. Schimansky-Geier (2001). "Transmission of noise coded versus additive signals through a neuronal ensemble." In: *Phys. Rev. Lett.* 86, pp. 2934–2937.

Lindner, B., L. Schimansky-Geier, and A. Longtin (2002). "Maximizing spike train coherence or incoherence in the leaky integrate-and-fire model." In: *Phys. Rev. E* 66.3, p. 031916.

Litwin-Kumar, A. and B. Doiron (2012). "Slow dynamics and high variability in balanced cortical networks with clustered connections." In: *Nat. Neurosci.* 15.11, pp. 1498–1505.

Logothetis, N. K. (2008). "What we can do and what we cannot do with fMRI." In: *Nature* 453, pp. 869–878.

MacKay, D. J. (2003). *Information theory, inference and learning algorithms*. Cambridge university press.

Magee, J. C. and C. Grienberger (2020). "Synaptic Plasticity Forms and Functions." In: *Annu. Rev. Neurosci.* 43.1, pp. 95–117.

Maier-Hein, K. H. et al. (2017). "The challenge of mapping the human connectome based on diffusion tractography." In: *Nat. Commun.* 8, p. 1349.

Majka, P. et al. (2020). "Open access resource for cellular-resolution analyses of corticocortical connectivity in the marmoset monkey." In: *Nat. Commun.* 11.1, pp. 1–14.

Manea, A. M., A. Zilverstand, K. Ugurbil, S. R. Heilbronner, and J. Zimmermann (2022). "Intrinsic timescales as an organizational principle of neural processing across the whole rhesus macaque brain." In: *eLife* 11, e75540.

Marcinkiewicz, J. (1939). "Sur une propriété de la loi de Gauss." In: *Math. Z.* 44.1, pp. 612–618.

Markov, N. T., M. M. Ercsey-Ravasz, A. R. Ribeiro Gomes, et al. (2014). "A Weighted and Directed Interareal Connectivity Matrix for Macaque Cerebral Cortex." In: *Cereb. Cortex* 24.1, pp. 17–36.

Markov, N. T., P. Misery, et al. (2011). "Weight Consistency Specifies Regularities of Macaque Cortical Networks." In: *Cereb. Cortex* 21.6, pp. 1254–1272.

Markov, N. T., M. Ercsey-Ravasz, D. C. Van Essen, K. Knoblauch, Z. Toroczkai, and H. Kennedy (2013). "Cortical High-Density Counterstream Architectures." In: *Science* 342.6158.

Markov, N. T., J. Vezoli, et al. (2014). "Anatomy of hierarchy: Feedforward and feedback pathways in macaque visual cortex." In: *J. Comp. Neurol.* 522.1, pp. 225–259.

Markram, H. et al. (2015). "Reconstruction and simulation of neocortical microcircuitry." In: *Cell* 163.2, pp. 456–492.

Marom, S. (2010). "Neural timescales or lack thereof." In: *Prog. Neurobiol.* 90.1, pp. 16–28.

Marques, J. P., T. Kober, G. Krueger, W. van der Zwaag, P.-F. Van de Moortele, and R. Gruetter (2010). "MP2RAGE, a self bias-field corrected sequence for improved segmentation and T1-mapping at high field." In: *Neuroimage* 49.2, pp. 1271–1281.

Martin, P., E. Siggia, and H. Rose (1973). "Statistical Dynamics of Classical Systems." In: *Phys. Rev. A* 8.1, pp. 423–437.

Martí, D., N. Brunel, and S. Ostojic (2018). "Correlations between synapses in pairs of neurons slow down dynamics in randomly connected neural networks." In: *Phys. Rev. E* 97 (6), p. 062314.

Mastrogiuseppe, F. and S. Ostojic (2017). "Intrinsically-generated fluctuating activity in excitatory-inhibitory networks." In: *PLOS Comput. Biol.* 13.4, e1005498.

Matheron, G. (1963). "Principles of geostatistics." In: *Econ. Geogr.* 58.8, pp. 1246–1266.

Matthews, A. G. d. G., J. Hron, M. Rowland, R. E. Turner, and Z. Ghahramani (2018). "Gaussian Process Behaviour in Wide Deep Neural Networks." In: *International Conference on Learning Representations*.

Mattia, M., M. Biggio, A. Galluzzi, and M. Storace (2019). "Dimensional reduction in networks of non-Markovian spiking neurons: Equivalence of synaptic filtering and heterogeneous propagation delays." In: *PLOS Comput. Biol.* 15.10. Ed. by B. Ermentrout, e1007404.

Mauk, M. D. and D. V. Buonomano (2004). "The Neural Basis of Temporal Processing." In: *Annu. Rev. Neurosci.* 27, pp. 307–340.

McCulloch, W. S. and W. Pitts (1943). "A logical calculus of the ideas immanent in neural nets." In: *Bull. Math. Biol.* 5, pp. 115–137.

McIntosh, L., N. Maheswaranathan, A. Nayebi, S. Ganguli, and S. Baccus (2016). "Deep learning models of the retinal response to natural scenes." In: *Adv. Neural Inf. Process. Syst.* 29.

McKinney, W. (2010). "Data Structures for Statistical Computing in Python." In: *Proceedings of the 9th Python in Science Conference*. Ed. by S. van der Walt and J. Millman, pp. 56–61.

Mejias, J. F., J. D. Murray, H. Kennedy, and X.-J. Wang (2016). "Feedforward and feedback frequency-dependent interactions in a large-scale laminar network of the primate cortex." In: *Sci. Adv.* 2.11, e1601335.

Mezard, M. and A. Montanari (2009). *Information, physics and computation*. Oxford University Press.

Migliore, M., F. Cavarretta, M. L. Hines, and G. M. Shepherd (2014). "Distributed organization of a brain microcircuit analyzed by three-

dimensional modeling: the olfactory bulb." In: *Front. Comput. Neurosci.* 8.50, pp. 1–14.

Migliore, M., F. Cavarretta, A. Marasco, E. Tulumello, M. L. Hines, and G. M. Shepherd (2015). "Synaptic clusters function as odor operators in the olfactory bulb." In: *Proc. Natl. Acad. Sci. USA* 112.27, pp. 8499–8504.

Minxha, J., R. Adolphs, S. Fusi, A. N. Mamelak, and U. Rutishauser (2020). "Flexible recruitment of memory-based choice representations by the human medial frontal cortex." In: *Science* 368.6498, eaba3313.

Moeller, S., E. Yacoub, C. A. Olman, E. Auerbach, J. Strupp, N. Harel, and K. Uğurbil (2010). "Multiband multislice GE-EPI at 7 tesla, with 16-fold acceleration using partial parallel imaging with application to high spatial and temporal whole-brain fMRI." In: *Magn. Reson. Med.* 63.5, pp. 1144–1153.

Mohan, H. et al. (2015). "Dendritic and axonal architecture of individual pyramidal neurons across layers of adult human neocortex." In: *Cereb. Cortex* 25.12, pp. 4839–4853.

Molgedey, L., J. Schuchhardt, and H. Schuster (1992). "Suppressing chaos in neural networks by noise." In: *Phys. Rev. Lett.* 69.26, p. 3717.

Mongillo, G., S. Rumpel, and Y. Loewenstein (2018). "Inhibitory connectivity defines the realm of excitatory plasticity." In: *Nat. Neurosci.* 21.10, pp. 1463–1470.

Montbrió, E., D. Pazó, and A. Roxin (2015). "Macroscopic Description for Networks of Spiking Neurons." In: *Phys. Rev. X* 5 (2), p. 021028.

Moreno-Bote, R. and N. Parga (2006). "Auto- and crosscorrelograms for the spike response of leaky integrate-and-fire neurons with slow synapses." In: *Phys. Rev. Lett.* 96, p. 028101.

Moshe, M. and J. Zinn-Justin (2003). "Quantum field theory in the large N limit: a review." In: *Phys. Rep.* 385, pp. 69–228.

Mozeika, A., B. Li, and D. Saad (2020). "Space of Functions Computed by Deep-Layered Machines." In: *Phys. Rev. Lett.* 125 (16), p. 168301.

Murphy, B. K. and K. D. Miller (2009). "Balanced Amplification: A New Mechanism of Selective Amplification of Neural Activity Patterns." In: *Neuron* 61.4, pp. 635–648.

Murray, J. D. et al. (2014). "A hierarchy of intrinsic timescales across primate cortex." In: *Nat. Neurosci.* 17.12, pp. 1661–1663.

Muscinelli, S. P., W. Gerstner, and T. Schwalger (2019). "How single neuron properties shape chaotic dynamics and signal transmission in random neural networks." In: *PLOS Comput. Biol.* 15.6, e1007122.

Nassi, J. J., C. L. Cepko, R. T. Born, and K. T. Beier (2015). "Neuroanatomy goes viral!" In: *Front. Neuroanat.* 9, p. 80.

Naveh, G., O. Ben David, H. Sompolinsky, and Z. Ringel (2021). "Predicting the outputs of finite deep neural networks trained with noisy gradients." In: *Phys. Rev. E* 104 (6), p. 064301.

Neal, R. M. (1996). *Bayesian Learning for Neural Networks*. Springer New York.

Nolte, M., E. Gal, H. Markram, and M. W. Reimann (2020). "Impact of higher order network structure on emergent cortical activity." In: *Netw. Neurosci.* 4.1, pp. 292–314.

Nolte, M., M. W. Reimann, J. G. King, H. Markram, and E. B. Muller (2019). "Cortical reliability amid noise and chaos." In: *Nat. Commun.* 10.1, pp. 1–15.

Nunez, P. L. (1974). "The brain wave equation: a model for the EEG." In: *Math. Biosci.* 21.3-4, pp. 279–297.

Ogawa, T. and H. Komatsu (2010). "Differential Temporal Storage Capacity in the Baseline Activity of Neurons in Macaque Frontal Eye Field and Area V4." In: *J. Neurophysiol.* 103.5, pp. 2433–2445.

Oh, S. W. et al. (2014). "A mesoscale connectome of the mouse brain." In: *Nature* 508.7495, pp. 207–214.

Ostojic, S. (2014). "Two types of asynchronous activity in networks of excitatory and inhibitory spiking neurons." In: *Nat. Neurosci.* 17, pp. 594–600.

Ostojic, S. and N. Brunel (2011). "From Spiking Neuron Models to Linear-Nonlinear Models." In: *PLOS Comput. Biol.* 7.1, e1001056.

Otopalik, A. G., A. C. Sutton, M. Banghart, and E. Marder (2017). "When complex neuronal structures may not matter." In: *eLife* 6, e23508.

Owen, D. B. (1965). "A Special Case of a Bivariate Non-Central t-Distribution." In: *Biometrika* 52.3/4, pp. 437–446.

– (1980). "A table of normal integrals." In: *Commun. Stat. Simul. Comput.* 9.4, pp. 389–419.

Packer, A. M. and R. Yuste (2011). "Dense, Unspecific Connectivity of Neocortical Parvalbumin-Positive Interneurons: A Canonical Microcircuit for Inhibition?" In: *J. Neurosci.* 31.37, pp. 13260–13271.

Paninski, L. (2004). "Maximum likelihood estimation of cascade point-process neural encoding models." In: *Netw. Comput. Neural Syst.* 15.4, pp. 243–262.

Papoulis, A. and S. U. Pillai (2002). *Probability, Random Variables, and Stochastic Processes*. 4th. Boston: McGraw-Hill.

Parisi, G. (1980). "The order parameter for spin glasses: A function on the interval 0-1." In: *J. Phys. A Math. Gen.* 13.3, p. 1101.

Parker, D. (2018). "Kuhnian revolutions in neuroscience: the role of tool development." In: *Biology & Philosophy* 33.3, p. 17.

Patel, A. X. and E. T. Bullmore (2016). "A wavelet-based estimator of the degrees of freedom in denoised fMRI time series for probabilistic testing of functional connectivity and brain graphs." In: *Neuroimage* 142, pp. 14–26.

Paxinos, G., C. Watson, M. Petrides, M. Rosa, and H. Tokuno (2012). *The Marmoset Brain in Stereotaxic Coordinates*. Elsevier Academic Press.

Pearlmutter, B. A. (1989). "Learning State Space Trajectories in Recurrent Neural Networks." In: *Neural Comput.* 1.2, pp. 263–269.

Pena, R. F., S. Vellmer, D. Bernardi, A. C. Roque, and B. Lindner (2018). "Self-Consistent Scheme for Spike-Train Power Spectra in Heterogeneous Sparse Networks." In: *Front. Comput. Neurosci.* 12, p. 9.

Peng, H. et al. (2021). "Morphological diversity of single neurons in molecularly defined cell types." In: *Nature* 598.7879, pp. 174–181.

Perez-Nieves, N., V. C. Leung, P. L. Dragotti, and D. F. Goodman (2021). "Neural heterogeneity promotes robust learning." In: *Nat. Commun.* 12.1, pp. 1–9.

Perin, R., T. K. Berger, and H. Markram (2011). "A synaptic organizing principle for cortical neuronal groups." In: *Proc. Natl. Acad. Sci. USA* 108.13, pp. 5419–5424.

Pillow, J. and P. Latham (2007). "Neural characterization in partially observed populations of spiking neurons." In: *Adv. Neural Inf. Process. Syst.* 20.

Pillow, J. W., J. Shlens, L. Paninski, A. Sher, A. M. Litke, E. J. Chichilnisky, and E. P. Simoncelli (2008). "Spatio-temporal correlations and visual signalling in a complete neuronal population." In: *Nature* 454, pp. 995–999.

Poirazi, P., T. Brannon, and B. W. Mel (2003). "Pyramidal Neuron as Two-Layer Neural Network." In: *Neuron* 37.6, pp. 989–999.

Poirazi, P. and A. Papoutsi (2020). "Illuminating dendritic function with computational models." In: *Nat. Rev. Neurosci.* 21.6, pp. 303–321.

Potjans, T. C. and M. Diesmann (2014). "The Cell-Type Specific Cortical Microcircuit: Relating Structure and Activity in a Full-Scale Spiking Network Model." In: *Cereb. Cortex* 24.3, pp. 785–806.

Press, W. H., S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery (2007). *Numerical Recipes: The Art of Scientific Computing*. 3rd. Cambridge University Press.

Pulvermüller, F., R. Tomasello, M. R. Henningsen-Schomers, and T. Wennekers (2021). "Biological constraints on neural network models of cognitive function." In: *Nat. Rev. Neurosci.* 22, pp. 488–502.

Purcell, E. M. (1977). "Life at low Reynolds number." In: *Am. J. Phys.* 45.1, pp. 3–11.

Python Software Foundation (2008). *The Python programming language*. http://www.python.org.

Rabinovich, M. I., P. Varona, A. I. Selverston, and H. D. Abarbanel (2006). "Dynamical principles in neuroscience." In: *Rev. Mod. Phys.* 78, p. 1213.

Raichle, M. E. (2015). "The brain's default mode network." In: *Annu. Rev. Neurosci.* 38, pp. 433–447.

Rainal, A. J. (1988). "Origin of Rice's formula." In: *IEEE Trans. Inf. Theory* 34.6, pp. 1383–1387.

Rakic, P. (2002). "Neurogenesis in adult primate neocortex: an evaluation of the evidence." In: *Nat. Rev. Neurosci.* 3.1, pp. 65–71.

Rall, W. (1969). "Time constants and electrotonic length of membrane cylinders and neurons." In: *Biophys. J.* 9.12, pp. 1483–1508.

Ramon y Cajal, S. (1899). *Comparative study of the sensory areas of the human cortex*. Clark University.

Ramón y Cajal, S. (1888). "Estructura de los centros nerviosos de las aves." In: *Rev. Trim. Histol. Norm. Pat.* 1, pp. 1–10.

– (1917). *Recuerdos de mi vida*. Imprenta y Librería de N. Moya.

Rasmussen, C. and C. Williams (2006). *Gaussian Processes for Machine Learning*. Adaptive Computation and Machine Learning. Cambridge, MA, USA: MIT Press, p. 248.

Razali, N. M. and Y. B. Wah (2011). "Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests." In: *Journal of Statistical Modeling and Analytics* 2.1, pp. 21–33.

Reimann, M. W., C. A. Anastassiou, R. Perin, S. L. Hill, H. Markram, and C. Koch (2013). "A biophysically detailed model of neocortical local field potentials predicts the critical role of active membrane currents." In: *Neuron* 79.2, pp. 375–390.

Renart, A., J. De La Rocha, P. Bartho, L. Hollender, N. Parga, A. Reyes, and K. D. Harris (2010). "The asynchronous State in Cortical Circuits." In: *Science* 327, pp. 587–590.

Ricciardi, L. M. (1977). *Diffusion Processes and Related Topics on Biology*. Berlin: Springer-Verlag.

Ricciardi, L. M. and S. Sato (1983). "A note on first passage time problems for Gaussian processes and varying boundaries." In: *IEEE Trans. Inf. Theory* 29.3, pp. 454–457.

Rice, S. O. (1945). "Mathematical Analysis of Random Noise." In: *Bell Syst. Tech. J.* 24.1. reprinted in Wax 1954, pp. 46–156.

Richardson, M. J. E. (2007). "Firing-rate response of linear and non-linear integrate-and-fire neurons to modulated current-based and conductance-based synaptic drive." In: *Phys. Rev. E* 76.021919, pp. 1–15.

– (2008). "Spike-train spectra and network response functions for non-linear integrate-and-fire neurons." In: *Biol. Cybern.* 99, pp. 381–392.

Rieke, F., D. Warland, R. de Ruyter van Steveninck, and W. Bialek (1997). *Spikes: Exploring the Neural Code*. Cambridge, MA: The MIT Press.

Riquelme, J. L. and J. Gjorgjieva (2021). "Towards readable code in neuroscience." In: *Nat. Rev. Neurosci.* 22.5, pp. 257–258.

Risken, H. (1996). *The Fokker-Planck Equation*. Springer Verlag Berlin Heidelberg.

Roberts, D. A., S. Yaida, and B. Hanin (2022). *The Principles of Deep Learning Theory*. Cambridge University Press.

Robinson, P. A., X. Gao, and Y. Han (2021). "Relationships between lognormal distributions of neural properties, activity, criticality, and connectivity." In: *Biol. Cybern.* 115.2, pp. 121–130.

Rockland, K. S. (2019). "What do we know about laminar connectivity?" In: *Neuroimage* 197, pp. 772–784.

Rohatgi, A. (2021). *Webplotdigitizer: Version 4.5*.

Rosenbaum, R. and B. Doiron (2014). "Balanced Networks of Spiking Neurons with Spatially Dependent Recurrent Connections." In: *Phys. Rev. X* 4.2, p. 021039.

Rosenbaum, R., M. A. Smith, A. Kohn, J. E. Rubin, and B. Doiron (2017). "The spatial structure of correlated neuronal variability." In: *Nat. Neurosci.* 20.1, pp. 107–114.

Rössert, C. et al. (2016). "Automated point-neuron simplification of data-driven microcircuit models." In: *ArXiv* 1604.00087 [q-bio.NC].

Rossi-Pool, R., A. Zainos, M. Alvarez, S. Parra, J. Zizumbo, and R. Romo (2021). "Invariant timescale hierarchy across the cortical somatosensory network." In: *Proc. Natl. Acad. Sci. USA* 118.3.

Rost, T., M. Deger, and M. P. Nawrot (2018). "Winnerless competition in clustered balanced networks: inhibitory assemblies do the trick." In: *Biol. Cybern.* 112.1, pp. 81–98.

Rostami, V., T. Rost, A. Riehle, S. J. van Albada, and M. P. Nawrot (2020). "Spiking neural network model of motor cortex with joint excitatory and inhibitory clusters reflects task uncertainty, reaction times, and variability dynamics." In: *BioRxiv*.

Roxin, A., N. Brunel, D. Hansel, G. Mongillo, and C. van Vreeswijk (2011). "On the distribution of firing rates in networks of cortical neurons." In: *J. Neurosci.* 31.45, pp. 16217–16226.

Rumelhart David, E., E. Hinton Geoffrey, and J. Williams Ronald (1986). "Learning representations by back-propagating errors." In: *Nature* 323, pp. 533–536.

Runyan, C. A., E. Piasini, S. Panzeri, and C. D. Harvey (2017). "Distinct timescales of population coding across cortex." In: *Nature* 548, pp. 92–96.

Sanzeni, A., M. H. Histed, and N. Brunel (2020). "Response nonlinearities in networks of spiking neurons." In: *PLOS Comput. Biol.* 16.9, e1008165.

Scherr, F. and W. Maass (2021). "Analysis of the computational strategy of a detailed laminar cortical microcircuit model for solving the image-change-detection task." In: *BioRxiv*.

Schmidt, M., R. Bakker, C. C. Hilgetag, M. Diesmann, and S. J. van Albada (2018). "Multi-scale account of the network structure of macaque visual cortex." In: *Brain Struct. Funct.* 223.3, pp. 1409–1435.

Schmidt, M., R. Bakker, K. Shen, G. Bezgin, M. Diesmann, and S. J. van Albada (2018). "A multi-scale layer-resolved spiking network model of resting-state dynamics in macaque visual cortical areas." In: *PLOS Comput. Biol.* 14.10, e1006359.

Schuecker, J., M. Diesmann, and M. Helias (2014). "Reduction of colored noise in excitable systems to white noise and dynamic boundary conditions." In: *ArXiv*, p. 1410.8799.

– (2015). "Modulated escape from a metastable state driven by colored noise." In: *Phys. Rev. E* 92 (5), p. 052119.

Schuecker, J., S. Goedeke, and M. Helias (2018). "Optimal Sequence Memory in Driven Random Networks." In: *Phys. Rev. X* 8 (4), p. 041029.

Schuecker, J., M. Schmidt, S. J. van Albada, M. Diesmann, and M. Helias (2017). "Fundamental Activity Constraints Lead to Specific Interpretations of the Connectome." In: *PLOS Comput. Biol.* 13.2, e1005179.

Schwalger, T. (2021). "Mapping input noise to escape noise in integrate-and-fire neurons: a level-crossing approach." In: *Biol. Cybern.* 115.5, pp. 539–562.

Schwalger, T., M. Deger, and W. Gerstner (2017). "Towards a theory of cortical columns: From spiking neurons to interacting neural populations of finite size." In: *PLOS Comput. Biol.* 13.4, e1005507.

Schwalger, T., F. Droste, and B. Lindner (2015). "Statistical structure of neural spiking under non-Poissonian or other non-white stimulation." In: *J. Comput. Neurosci.* 39, p. 29.

Seeman, S. C. et al. (2018). "Sparse recurrent excitatory connectivity in the microcircuit of the adult mouse and human cortex." In: *eLife* 7, e37349.

Sejnowski, T. J., P. S. Churchland, and J. A. Movshon (2014). "Putting big data to good use in neuroscience." In: *Nat. Neurosci.* 17.11, pp. 1440–1441.

Sejnowski, T. (1976). "On the stochastic dynamics of neuronal interaction." In: *Biol. Cybern.* 22.4, pp. 203–211.

Senk, J., K. Korvasová, J. Schuecker, E. Hagen, T. Tetzlaff, M. Diesmann, and M. Helias (2020). "Conditions for wave trains in spiking neural networks." In: *Phys. Rev. Res.* 2.2.

Senk, J., B. Kriener, et al. (2021). "Connectivity Concepts in Neuronal Network Modeling." In: *ArXiv*, 2110.02883 [q–bio.NC].

Shadlen, M. N. and W. T. Newsome (1994). "Noise, neural codes and cortical organization." In: *Curr. Opin. Neurobiol.* 4.4, pp. 569–579.

Shapson-Coe, A. et al. (2021). "A connectomic study of a petascale fragment of human cerebral cortex." In: *BioRxiv*.

Sherwood, C. C., S. B. Miller, M. Karl, C. D. Stimpson, K. A. Phillips, B. Jacobs, P. R. Hof, M. A. Raghanti, and J. B. Smaers (2020). "Invariant synapse density and neuronal connectivity scaling in primate neocortical evolution." In: *Cereb. Cortex* 30.10, pp. 5604–5615.

Shimoura, R. O., R. F. O. Pena, V. Lima, N. L. Kamiji, M. Girardi-Schappo, and A. C. Roque (2021). "Building a model of the brain: from detailed connectivity maps to network organization." In: *Eur. Phys. J. Spec. Top.* 230.14-15, pp. 2887–2909.

Siegert, A. J. (1951). "On the first passage time probability problem." In: *Phys. Rev.* 81.4, pp. 617–623.

Siegle, J. H. et al. (2021). "Survey of spiking in the mouse visual system reveals functional hierarchy." In: *Nature* 592.7852, pp. 86–92.

Smith, S. M. et al. (2004). "Advances in functional and structural MR image analysis and implementation as FSL." In: *Neuroimage* 23. Mathematics in Brain Imaging, S208–S219.

Sommers, H. (1987). "Path-Integral Approach to Ising Spin-Glass Dynamics." In: *Phys. Rev. Lett.* 58.12, pp. 1268–1271.

Sompolinsky, H. and A. Zippelius (1982). "Relaxational dynamics of the Edwards-Anderson model and the mean-field theory of spin-glasses." In: *Phys. Rev. B* 25.11, pp. 6860–6875.

Sompolinsky, H., A. Crisanti, and H. J. Sommers (1988). "Chaos in Random Neural Networks." In: *Phys. Rev. Lett.* 61 (3), pp. 259–262.

Sompolinsky, H. and A. Zippelius (1981). "Dynamic Theory of the Spin-Glass Phase." In: *Phys. Rev. Lett.* 47 (5), pp. 359–362.

Sompolinsky, H. (1988). "Statistical mechanics of neuronal networks." In: *Phys. Today* 41.12, pp. 70–80.

Song, S., P. Sjöström, M. Reigl, S. Nelson, and D. Chklovskii (2005). "Highly nonrandom features of synaptic connectivity in local cortical circuits." In: *PLOS Biol.* 3.3, e68.

Spitmaan, M., H. Seo, D. Lee, and A. Soltani (2020). "Multiple timescales of neural dynamics and integration of task-relevant signals across cortex." In: *Proc. Natl. Acad. Sci. USA* 117.36, pp. 22522–22531.

Stapmanns, J., T. Kühn, D. Dahmen, T. Luu, C. Honerkamp, and M. Helias (2020). "Self-consistent formulations for stochastic nonlinear neuronal dynamics." In: *Phys. Rev. E* 101 (4), p. 042124.

Stephan, K., L. Kamper, A. Bozkurt, G. Burns, M. Young, and R. Kötter (2001). "Advanced database methodology for the collation of connectivity data on the macaque brain (CoCoMac)." In: *Philos. Trans. R. Soc. B* 356, pp. 1159–1186.

Stern, M., H. Sompolinsky, and L. F. Abbott (2014). "Dynamics of random neural networks with bistable units." In: *Phys. Rev. E* 90 (6), p. 062710.

Stiller, J. and G. Radons (1998). "Dynamics of nonlinear oscillators with random interactions." In: *Phys. Rev. E* 58.2, p. 1789.

Stratonovich, R. L. (1989). "Some Markov methods in the theory of stochastic processes in nonlinear dynamical systems." In: *Noise in Nonlinear Dynamical Systems*. Ed. by F. Moss and P. V. E. McClintock. Vol. 1. Cambridge University Press, pp. 16–71.

Stratonovich, R. L. (1967). *Topics in the Theory of Random Noise*. New York: Gordon and Breach.

Stuart, G. J. and N. Spruston (2015). "Dendritic integration: 60 years of progress." In: *Nat. Neurosci.* 18.12, pp. 1713–1721.

Sussillo, D. (2014). "Neural circuits as computational dynamical systems." In: *Curr. Opin. Neurobiol.* 25, pp. 156–163.

Szegedy, C., W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus (2013). "Intriguing properties of neural networks." In: *ArXiv*.

Tasic, B. et al. (2018). "Shared and distinct transcriptomic cell types across neocortical areas." In: *Nature* 563, pp. 72–78.

Teeter, C. et al. (2018). "Generalized leaky integrate-and-fire models classify multiple neuron types." In: *Nat. Commun.* 9, p. 709.

Tetzlaff, T., M. Helias, G. T. Einevoll, and M. Diesmann (2012). "Decorrelation of Neural-Network Activity by Inhibitory Feedback." In: *PLOS Comput. Biol.* 8.8. Ed. by N. Brunel, e1002596.

Theodoni, P., P. Majka, D. H. Reser, D. K. Wójcik, M. G. P. Rosa, and X.-J. Wang (2021). "Structural Attributes and Principles of the Neocortical Connectome in the Marmoset Monkey." In: *Cereb. Cortex* 32.1, pp. 15–28.

Tomasello, R., M. Garagnani, T. Wennekers, and F. Pulvermüller (2018). "A Neurobiologically Constrained Cortex Model of Semantic Grounding With Spiking Neurons and Brain-Like Connectivity." In: *Front. Comput. Neurosci.* 12, p. 88.

Touchette, H. (2009). "The large deviation approach to statistical mechanics." In: *Phys. Rep.* 478 (1–3), pp. 1–69.

Toyoizumi, T. and L. F. Abbott (2011). "Beyond the edge of chaos: Amplification and temporal integration by recurrent networks in the chaotic regime." In: *Phys. Rev. E* 84.5, p. 051908.

Toyoizumi, T., K. R. Rad, and L. Paninski (2009). "Mean-Field Approximations for Coupled Populations of Generalized Linear Model Spiking Neurons with Markov Refractoriness." In: *Neural Comput.* 21, pp. 1203–1243.

Trousdale, J., Y. Hu, E. Shea-Brown, and K. Josic (2012). "Impact of network structure and cellular response on spike time correlations." In: *PLOS Comput. Biol.* 8.3, e1002408.

Turner, E. C., N. A. Young, J. L. Reed, C. E. Collins, D. K. Flaherty, M. Gabi, and J. H. Kaas (2016). "Distributions of cells and neurons across the cortical sheet in Old World macaques." In: *Brain Behav. Evol.* 88.1, pp. 1–13.

Ullner, E., A. Politi, and A. Torcini (2020). "Quantitative and qualitative analysis of asynchronous neural activity." In: *Phys. Rev. Res.* 2.2, p. 023103.

Van Essen, D. C., S. M. Smith, D. M. Barch, T. E. Behrens, E. Yacoub, K. Ugurbil, W.-M. H. Consortium, et al. (2013). "The WU-Minn Human Connectome Project: An overview." In: *Neuroimage* 80, pp. 62–79.

Van Kampen, N. G. (2007). *Stochastic Processes in Physics and Chemistry, Third Edition (North-Holland Personal Library)*. 3rd ed. North Holland.

Van Albada, S. J., A. Morales-Gregorio, T. Dickscheid, A. Goulas, R. Bakker, S. Bludau, G. Palm, C.-C. Hilgetag, and M. Diesmann (2022). "Bringing Anatomical Information into Neuronal Network Models." In: *Computational Modelling of the Brain: Modelling Approaches to Cells,*

Circuits and Networks. Ed. by M. Giugliano, M. Negrello, and D. Linaro. Cham: Springer International Publishing, pp. 201–234.

Van Albada, S. J., A. G. Rowley, J. Senk, M. Hopkins, M. Schmidt, A. B. Stokes, D. R. Lester, M. Diesmann, and S. B. Furber (2018). "Performance Comparison of the Digital Neuromorphic Hardware SpiNNaker and the Neural Network Simulation Software NEST for a Full-Scale Cortical Microcircuit Model." In: *Front. Neurosci.* 12, p. 291.

Van Meegen, A., T. Kühn, and M. Helias (2021). "Large-Deviation Approach to Random Recurrent Neuronal Networks: Parameter Inference and Fluctuation-Induced Transitions." In: *Phys. Rev. Lett.* 127 (15), p. 158302.

Van Meegen, A. and B. Lindner (2018). "Self-Consistent Correlations of Randomly Coupled Rotators in the Asynchronous State." In: *Phys. Rev. Lett.* 121.25, p. 258302.

Van Vreeswijk, C. and H. Sompolinsky (1998). "Chaotic Balanced State in a Model of Cortical Circuits." In: *Neural Comput.* 10.6, pp. 1321–1371.

Van Vreeswijk, C. and F. Farkhooi (2019). "Fredholm theory for the mean first-passage time of integrate-and-fire oscillators with colored noise input." In: *Phys. Rev. E* 100.6, p. 060402.

Van Vreeswijk, C. and H. Sompolinsky (1996). "Chaos in Neuronal Networks with Balanced Excitatory and Inhibitory Activity." In: *Science* 274, pp. 1724–1726.

Varadhan, S. (2008). "Large deviations." In: *Ann. Probab.* 36.2, pp. 397–419.

Vellmer, S. and B. Lindner (2019). "Theory of spike-train power spectra for multidimensional integrate-and-fire neurons." In: *Phys. Rev. Res.* 1.2, p. 023024.

Verechtchaguina, T., I. M. Sokolov, and L. Schimansky-Geier (2006). "First passage time densities in resonate-and-fire models." In: *Phys. Rev. E* 73.3, p. 031108.

Virtanen, P. et al. (2020). "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python." In: *Nat. Methods* 17, pp. 261–272.

Von Bartheld, C. S., J. Bahney, and S. Herculano-Houzel (2016). "The search for true numbers of neurons and glial cells in the human brain: A review of 150 years of cell counting." In: *J. Comp. Neurol.* 524.18, pp. 3865–3895.

Von Economo, C. (2009). *Cellular Structure of the Human Cerebral Cortex*. Translated and edited by L.C. Triarhou. Karger Medical and Scientific Publishers.

Von Economo, C. (1929). "Der Zellaufbau der Grosshirnrinde und die progressive Cerebration." In: *Ergebnisse der Physiologie* 29.1, pp. 83–128.

Von Economo, C. F. and G. N. Koskinas (1925). *Die Cytoarchitektonik der Hirnrinde des erwachsenen Menschen*. J. Springer.

Wagatsuma, N., T. C. Potjans, M. Diesmann, and T. Fukai (2011). "Layer-dependent attentional processing by top-down signals in a visual cortical microcircuit model." In: *Front. Comput. Neurosci.* 5, p. 31.

Wang, H. E. et al. (2022). "Virtual Epileptic Patient (VEP): Data-driven probabilistic personalized brain modeling in drug-resistant epilepsy." In: *MedRxiv*.

Wang, X.-J. (2020). "Macroscopic gradients of synaptic excitation and inhibition in the neocortex." In: *Nat. Rev. Neurosci.* 21.3, pp. 169–178.

Waskom, M. L. (2021). "seaborn: statistical data visualization." In: *Journal of Open Source Software* 6.60, p. 3021.

Wasmuht, D. F., E. Spaak, T. J. Buschman, E. K. Miller, and M. G. Stokes (2018). "Intrinsic neuronal dynamics predict distinct functional roles during working memory." In: *Nat. Commun.* 9.1, pp. 1–13.

Watts, D. J. and S. H. Strogatz (1998). "Collective dynamics of small-world networks." In: *Nature* 393, pp. 440–442.

Wax, N., ed. (1954). *Selected Papers on Noise and Stochastic Processes*. New York: Dover Publications.

Wieland, S., D. Bernardi, T. Schwalger, and B. Lindner (2015). "Slow fluctuations in recurrent networks of spiking neurons." In: *Phys. Rev. E* 92.4, p. 040901.

Williams, C. K. I. and D. Barber (1998). "Bayesian classification with Gaussian processes." In: *IEEE Trans. Pattern Anal. Mach. Intel.* 20.12, pp. 1342–1351.

Williams, C. K. (1998). "Computation with infinite neural networks." In: *Neural Comput.* 10.5, pp. 1203–1216.

Williams, C. K. and C. E. Rasmussen (2006). *Gaussian Processes for Machine Learning*. 1st. Cambridge: MIT Press.

Wilson, H. R. and J. D. Cowan (1972). "Excitatory and inhibitory interactions in localized populations of model neurons." In: *Biomed. Pharmacol. J.* 12.1, pp. 1–24.

– (1973). "A Mathematical Theory of the Functional Dynamics of Cortical and Thalamic Nervous Tissue." In: *Kybernetik* 13.2 (2), pp. 55–80.

Wilting, J. and V. Priesemann (2018). "Inferring collective dynamical states from widely unobserved systems." In: *Nat. Commun.* 9.1, p. 2325.

– (2019). "Between Perfectly Critical and Fully Irregular: A Reverberating Model Captures and Predicts Cortical Spike Propagation." In: *Cereb. Cortex* 29.6, pp. 2759–2770.

Winnubst, J. et al. (2019). "Reconstruction of 1,000 projection neurons reveals new cell types and organization of long-range connectivity in the mouse brain." In: *Cell* 179.1, pp. 268–281.

Witvliet, D. et al. (2021). "Connectomes across development reveal principles of brain maturation." In: *Nature* 596.7871, pp. 257–261.

Wong, K.-F. and X.-J. Wang (2006). "A Recurrent Network Mechanism of Time Integration in Perceptual Decisions." In: *J. Neurosci.* 26.4, pp. 1314–1328.

Yaida, S. (2020). "Non-Gaussian processes and neural networks at finite widths." In: *Proceedings of The First Mathematical and Scientific Machine Learning Conference*. Ed. by J. Lu and R. Ward. Vol. 107. Proceedings of Machine Learning Research. Princeton University, Princeton, NJ, USA: PMLR, pp. 165–192.

Yang, G. (2019). "Wide Feedforward or Recurrent Neural Networks of Any Architecture are Gaussian Processes." In: *Adv. Neural Inf. Process. Syst.* Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett. Vol. 32. Curran Associates, Inc.

Yuste, R. (2015). "From the neuron doctrine to neural networks." In: *Nat. Rev. Neurosci.* 16.8, pp. 487–497.

Zavatone-Veth, J. A. and C. Pehlevan (2021). "Exact marginal prior distributions of finite Bayesian neural networks." In: *Adv. Neural Inf. Process. Syst.* Ed. by A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan.

Zeraati, R., T. A. Engel, and A. Levina (2020). "Estimation of autocorrelation timescales with Approximate Bayesian Computations." In: *BioRxiv*.

Zinn-Justin, J. (1996). *Quantum field theory and critical phenomena*. Clarendon Press, Oxford.

Ziv, N. E. and N. Brenner (2018). "Synaptic tenacity or lack thereof: spontaneous remodeling of synapses." In: *Trends Neurosci.* 41.2, pp. 89–99.

OFFIZIELLE ERKLÄRUNG

Hiermit versichere ich an Eides statt, dass ich die vorliegende Dissertation selbstständig und ohne die Benutzung anderer als der angegebenen Hilfsmittel und Literatur angefertigt habe. Alle Stellen, die wörtlich oder sinngemäß aus veröffentlichten und nicht veröffentlichten Werken dem Wortlaut oder dem Sinn nach entnommen wurden, sind als solche kenntlich gemacht. Ich versichere an Eides statt, dass diese Dissertation noch keiner anderen Fakultät oder Universität zur Prüfung vorgelegen hat; dass sie - abgesehen von unten angegebenen Teilpublikationen und eingebundenen Artikeln und Manuskripten - noch nicht veröffentlicht worden ist sowie, dass ich eine Veröffentlichung der Dissertation vor Abschluss der Promotion nicht ohne Genehmigung des Promotionsausschusses vornehmen werde. Die Bestimmungen dieser Ordnung sind mir bekannt. Darüber hinaus erkläre ich hiermit, dass ich die Ordnung zur Sicherung guter wissenschaftlicher Praxis und zum Umgang mit wissenschaftlichem Fehlverhalten der Universität zu Köln gelesen und sie bei der Durchführung der Dissertation zugrundeliegenden Arbeiten und der schriftlich verfassten Dissertation beachtet habe und verpflichte mich hiermit, die dort genannten Vorgaben bei allen wissenschaftlichen Tätigkeiten zu beachten und umzusetzen. Ich versichere, dass die eingereichte elektronische Fassung der eingereichten Druckfassung vollständig entspricht.

Köln, 2022

Alexander van Meegen

TEILPUBLIKATIONEN

Layer, M., J. Senk, S. Essink, A. van Meegen, H. Bos, and M. Helias (2022). "NNMT: Mean-Field Based Analysis Tools for Neuronal Network Models." In: *Front. Neuroinform.* 16, p. 835657.

Morales-Gregorio, A., A. van Meegen, and S. J. van Albada (2022). "Ubiquitous lognormal distribution of neuron densities across mammalian cerebral cortex." In: *bioRxiv*.

Segadlo, K., B. Epping, A. van Meegen, D. Dahmen, M. Krämer, and M. Helias (2021). "Unified Field Theory for Deep and Recurrent Neural Networks." In: *arXiv*.

Van Albada, S. J., J. Pronold, A. van Meegen, and M. Diesmann (2021). "Usage and Scaling of an Open-Source Spiking Multi-Area Model of Monkey Cortex." In: *Brain-Inspired Computing*. Ed. by K. Amunts, L. Grandinetti, T. Lippert, and N. Petkov. Cham: Springer International Publishing, pp. 47–59.

Van Meegen, A., T. Kühn, and M. Helias (2021). "Large-Deviation Approach to Random Recurrent Neuronal Networks: Parameter Inference and Fluctuation-Induced Transitions." In: *Phys. Rev. Lett.* 127 (15), p. 158302.

Van Meegen, A. and S. J. van Albada (2021). "Microscopic theory of intrinsic timescales in spiking neural networks." In: *Phys. Rev. Research* 3 (4), p. 043077.

Band / Volume 85
Dynamical and statistical structure of spatially organized neuronal networks
M. Layer (2022), xiii, 167 pp
ISBN: 978-3-95806-651-9

Band / Volume 86
Persistent firing and oscillations in the septo-hippocampal systemand their relation to locomotion
K. Korvasová (2022), 111 pp
ISBN: 978-3-95806-654-0

Band / Volume 87
Sol-Gel-Synthese, Tintenstrahldruck und Blitzlampentemperung von Tantaloxid-Dünnschichten zur pH-Messung
C. D. Beale (2022), xlix, 339 pp
ISBN: 978-3-95806-656-4

Band / Volume 88
Diversity of chiral magnetic solitons
V. Kuchkin (2022), xiv, 155 pp
ISBN: 978-3-95806-665-6

Band / Volume 89
Controlling the electrical properties of oxide heterointerfaces through their interface chemistry
M.-A. Rose (2022), vi, 162 pp
ISBN: 978-3-95806-667-0

Band / Volume 90
Modeling and Suppressing Unwanted Parasitic Interactions in Superconducting Circuits
X. Xu (2022), 123, XVIII pp
ISBN: 978-3-95806-671

Band / Volume 91
Activating molecular magnetism by controlled on-surface coordinationI. Cojocariu (2022), xi, 169 pp
ISBN: 978-3-95806-674-8

Band / Volume 92
Computational study of structural and optical properties of two-dimensional transition-metal dichalcogenides with implanted defects
S. H. Rost (2023), xviii, 198 pp
ISBN: 978-3-95806-682-3

Band / Volume 93
DC and RF characterization of bulk CMOS and FD-SOI devices at cryogenic temperatures with respect to quantum computing applications
A. Artanov (2023), xv, 80, xvii-liii pp
ISBN: 978-3-95806-687-8

Band / Volume 94
HAXPES study of interface and bulk chemistry of ferroelectric HfO_2 capacitors
T. Szyjka (2023), viii, 120 pp
ISBN: 978-3-95806-692-2

Band / Volume 95
A brain inspired sequence learning algorithm and foundations of a memristive hardware implementation
Y. Bouhadjar (2023), xii, 149 pp
ISBN: 978-3-95806-693-9

Band / Volume 96
Characterization and modeling of primate cortical anatomy and activity
A. Morales-Gregorio (2023), ca. 260 pp.
ISBN: 978-3-95806-698-4

Band / Volume 97
Hafnium oxide based memristive devices as functional elements of neuromorphic circuits
F. J. Cüppers (2023), vi, ii, 214 pp
ISBN: 978-3-95806-702-8

Band / Volume 98
Simulation and theory of large-scale cortical networks
A. van Meegen (2023), ca. 250 pp
ISBN: 978-3-95806-708-0

Information
Band / Volume 98
ISBN 978-3-95806-708-0

Mitglied der Helmholtz-Gemeinschaft

JÜLICH
Forschungszentrum